Personalizing Robot Assistance under Uncertainty about the Human

by

Shen Li

B.S. Computer Science, Pennsylvania State University, 2015
B.S. Psychology, Pennsylvania State University, 2015
M.S. Robotics, Carnegie Mellon University, 2017

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN AUTONOMOUS SYSTEMS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2025

Authored by: Shen Li
Department of Aeronautics and Astronautics
July 15, 2025

Certified by: Julie A. Shah
Department Head, Department of Aeronautics and Astronautics
H.N. Slater Professor in Aeronautics and Astronautics, Thesis Supervisor

Accepted by: Jonathan P. How
Ford Professor of Engineering
Chair, Graduate Program Committee

THESIS COMMITTEE

Julie A. Shah
Department Head and H.N. Slater Professor in Aeronautics and Astronautics
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology

Thesis Committee Members

Aude Billard
Full Professor and Head of the Learning Algorithms and Systems Laboratory
School of Engineering
École Polytechnique Fédérale de Lausanne

Dylan Hadfield-Menell
Bonnie and Marty (1964) Tenenbaum Career Development Assistant Professor
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Na (Lina) Li
Winokur Family Professor of Electrical Engineering and Applied Mathematics
School of Engineering and Applied Sciences
Harvard University

Thesis Readers

Tariq Iqbal
Assistant Professor
Department of Computer Science and Department of Systems and Information Engineering
University of Virginia

Vaibhav Unhelkar
Assistant Professor
Department of Computer Science
Rice University

Personalizing Robot Assistance under Uncertainty about the Human

by

Shen Li

Submitted to the Department of Aeronautics and Astronautics
on July 15, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN AUTONOMOUS SYSTEMS

ABSTRACT

Robots have the potential to enhance human well-being by assisting with daily activities, particularly for older adults and people with disabilities. One example is robot-assisted dressing, where a robot helps a person put on clothing. However, no two individuals are alike. Each person has unique preferences, behaviors, and needs, making personalization essential for effective assistance. A central challenge is that robots often operate under uncertainty about the human they are helping. This uncertainty may involve the person's preferences, hidden physical states, or reactions to assistance. If not properly addressed, such uncertainty can lead to ineffective, undesired, or even unsafe outcomes.

This thesis asks: *How should a robot behave when it is uncertain about the human?* To answer this, I present a unified framework for uncertainty-aware personalization in human-robot interaction, spanning three core components of robot intelligence: *preference learning*, *state estimation*, and *motion planning*. I propose methods that (1) *reduce uncertainty* using implicit cognitive signals, (2) *represent and respect uncertainty* through set-based state estimation, and (3) *act under uncertainty* using relaxed safety constraints.

First, I introduce an approach that uses response time, a subtle yet informative cognitive signal, as implicit feedback for preference learning. While traditional methods rely solely on binary choices, I developed the first algorithm that integrates both choices and response times to infer not just what a person prefers, but how strongly they feel about those preferences. Theoretical analysis reveals that response times significantly reduce uncertainty about user preferences, especially when users have strong preferences. In simulation studies, this method decreased misidentification of the most preferred option by up to 55 %, enabling faster and more accurate personalization without extra user input.

Second, I address the problem of estimating hidden human states during physical interaction. For example, in dressing, parts of the body, such as the elbow, may be occluded. I introduce the first set-based estimator that represents and respects uncertainty from human behavior and sensing models trained on limited data. Instead of outputting a point estimate, the method constructs a geometric set, such as a 3D box, guaranteed with high probability to contain the true human state. In dressing experiments, the estimator achieved 92 % inclusion using significantly smaller boxes than prior methods, balancing reliability and precision, supporting safe and responsive physical assistance.

Third, I consider how a robot should plan motion when it is uncertain about future human behavior. Traditional safety constraints typically prohibit any contact, which can cause the robot to freeze when uncertainty is high. I propose a more flexible definition of safety that allows either collision avoidance or low-impact contact. Integrated into a learning-based control framework, this approach enables efficient motion while maintaining safety. In dressing tasks, it reduced task time by 78 % without compromising safety.

Together, these contributions show how robots can reduce, represent and respect, and act under uncertainty to personalize their assistance. This thesis lays a foundation for robots that not only respond to commands, but also understand and adapt to the nuanced, evolving, and uncertain nature of human behavior.

Thesis supervisor: Julie A. Shah
Title: Department Head, Department of Aeronautics and Astronautics
H.N. Slater Professor in Aeronautics and Astronautics

Acknowledgments

> *"All of the patents, publications, presentations, and personal technical achievements can be amazing: They can literally save lives and bring immense delight, win us world acclaim, fill our shelves with awards, tally up clicks online, and even make our resumes impressively long. However, they all pale in comparison to something that is even more joy-giving: Achieving deeply satisfying, personally-significant human relationships."*

—Rosalind Picard [1]

A PhD is a long journey filled with success and failure, joy and doubt, frustration and discovery. I could not have made it through alone. This thesis would not have been possible without the care, support, and love from so many people. As my friend Joe Kim once said after defending his dissertation: "What matters most isn't just the research; it's the people." Words can't fully capture what each of you means to me, but I hope these reflections show a fraction of my gratitude.

To my advisor, Julie Shah, thank you for saving me, trusting me, and giving me the freedom to succeed. In the spring of 2017, after my PhD applications were unsuccessful, I felt lost. Then came your email offering me a research engineer position in your lab. That message changed everything. You saw potential in me before I saw it in myself, and opened the door to the work and people I deeply cherish today.

When I first arrived at MIT, one of my earliest projects was deploying a collaborative robot demo for Honda. The day before every sponsor meeting, you would check on our progress. I don't think you ever saw a full demo run that worked, but you never got mad. You would say, "That's okay. Keep trying." Because you believed in me, I began to believe in myself, too. Each time, the final demo worked. That trust was transformative. I learned from you that great mentorship isn't about telling someone where to go, but about giving them the trust and space to find their own path, and being there when it gets hard.

You also gave me the freedom to pursue the kind of research I cared about. I wanted to build something useful, like the autonomous wheelchair I worked on as an undergraduate, but also grounded in strong theoretical foundations. She encouraged me to explore new classes and collaborations, even when my project diverged from the lab's core direction. That freedom led to our work on response times, a paper that integrated insights from psychology and offered new theoretical insights. Only recently did I realize that my first exposure to response times came from a psychology lab I joined as an undergraduate, where I studied sleep deprivation and vigilance. It's funny how life sometimes loops back on itself. So, thank you, Julie. You gave me the rarest gift in academia: the freedom to fail, the freedom to change direction, and the encouragement to keep going. Because of you, I didn't just become a researcher; I became the kind of researcher I wanted to be.

To Lina Li, thank you for believing in my ideas and supporting me through uncertainty. When I was exploring a theoretical formulation of human-robot interaction, I felt lost. I remember the joy I felt when I got your email saying you'd be on my committee, and even more so when you offered to meet regularly. The path was not smooth. We explored directions that didn't pan out, despite involving your brilliant students, Yuyang Zhang and Zhaolin Ren. But you kept showing up. You offered feedback on my slides for proposal defense, even when it wasn't your responsibility. Eventually, our efforts clicked and became our NeurIPS paper. From you, I learned that mentorship is not about instant answers; it's about patience, asking the right questions, and walking through the messy middle together.

To Dylan Hadfield-Menell, thank you for setting an example of rigor, interdisciplinarity, and creativity. Your work on cooperative inverse reinforcement learning was one of the most insightful papers I have read. I admired how you integrated theory with user studies and drew connections across recommender systems, economics, and human-robot interaction. Your creativity extended beyond research to presentation. When I shared my response-time paper with you, you suggested introducing the intuition first, presenting the method as if it arose naturally, and only then revealing its roots in decades of psychology research. That twist in storytelling was brilliant. You shaped how I think about research and communication. While I may not have reached that level, the response time paper felt like a step in that direction. Thank you for setting a reward function that pushed me to grow.

To Aude Billard, thank you for your lasting influence on my research and professional growth. In one committee meeting, as I discussed inverse reinforcement learning, you smiled and said, "I don't optimize. I just try to be feasible." That simple sentence stayed with me. Especially under pressure, it reminded me that navigating constraints can matter more than chasing perfection. You also gave me a concrete "constraint": don't let flashy slides and animations bury the core ideas. Your honest feedback pushed me to prioritize clarity over flair. At our next meeting, I presented with fewer slides and greater focus, and our discussion became more productive as a result. You helped me see that, in research as in life, clarity and feasibility often matter more than complexity. Thank you for your wisdom, candor, and high standards.

To Vaibhav Unhelkar, thank you for teaching me what true mentorship looks like. When I made a serious mistake, forgetting to record audio data right before a paper deadline, you stayed calm. "Let's see if we can save it," you said. We did, and the paper was published. That moment taught me that support and problem-solving matter more than blame. For years, you worked behind me in the lab, steady and focused. I'd often glance back to see you typing in your green-themed terminal or playing a quick chess game to relax. That image stayed with me. In a field driven by rapid change and pressure, you showed me that the most powerful force is consistency in effort, support, and belief in one another.

To Tariq Iqbal, thank you for being someone I could always count on. In 2017, we worked together on the Honda demo, you led the algorithm work, and I managed the system side. Even though it wasn't your responsibility, you helped me again and again. I remember you spending 1 hour commuting home to care for your son, then returning late around 9 p.m. to debug with me. Sometimes we worked until the next morning. When the robot finally ran, you'd say, "Don't touch anything before the demo!" That advice stuck with me. You demonstrated to me that true mentorship isn't just about helping with work. It's about being the kind of person someone can count on, no matter what.

To Nadia Figueroa, thank you for reminding me of the joy in robotics. In 2020, during COVID, we set up a new robot together. I expected it to be a tedious and mechanical task. But you were thrilled. I still remember the excitement in your eyes as you looked at the robot, packed in its box, already imagining how to bring it to life. Your enthusiasm was infectious. In research, we're often trained to focus on results, but you reminded me of the magic of loving the process.

To Theodoros Stouraitis, thank you for making remote research not just productive, but joyful. During COVID, we met every Friday, usually for much longer than planned. We brainstormed, debated, and made real progress. I remember one time we started chatting at noon and didn't stop until 7 p.m., with your girlfriend, Stamatina, trying to pull you away for dinner while you stayed just a bit longer to finish our discussion. I think I talked about research more during COVID than ever before, and that's thanks to you. Thank you, Theo, for reminding me that even in isolation, ideas can bring people together.

To Yuyang Zhang, thank you for your insight and patience. You knew every theory, both the ones I had heard of and the ones I hadn't. I worried I'd fall behind, but you always listened, waited, and helped shape my rough ideas. You asked questions like "What's the intuition here?" that pushed me to really understand the math. You taught me that theory isn't just about solving, but about understanding. Insight is the soul of good research.

To all my collaborators, including those I haven't yet mentioned by name, such as Ankit Shah, Chris Fourie, Felix Wang, Daehyung Park, Yoonchang Sung, Zhaolin Ren, Claire Liang, Sarah Chung, Brad Hayes, Nick Roy, Michael Gienger, and Sethu Vijayakumar, I wish I had space to share a story about each of you. Each of you brought something unique: your curiosity, your ideas, your questions, and your feedback. Thank you for the arguments, the late nights, and the moments when things finally clicked. If a thesis is a mountain, then each of you laid a stepping stone along the path. Some helped build the foundation, some helped me climb, and some reminded me to pause and enjoy the view.

To the Interactive Robotics Group (IRG), thank you for being a home where I could learn, grow, laugh, struggle, and always feel supported. Before I joined MIT in 2017, I imagined everyone here would be brilliant, and maybe a little intimidating. After joining IRG, I found I was only half right. Yes, everyone was brilliant, but you were also kind, humble, and warm. From beer nights to camping trips, from late-night debugging to joyful weddings, you made this journey deeply human. I especially thank Stefanos Nikolaidis, Claudia Perez D'Arpino, Pem Lasota, Ramya Ramakrishnan, Joe Kim, Vaibhav Unhelkar, Ankit Shah, Lindsay Sanneman, Yi-Shiuan Tung, Kailah Cabral, Yilun Zhou, Thavishi Illandara, Chris Fourie, Tariq Iqbal, Serena Booth, Mycal Tucker, Bilkit Githinji, Sarah Chung, Josh Creamer, Theodoros Stouraitis, Felix Wang, Andi Peng, Katya Arquilla, Nadia Figueroa, Nicole La, Alex Cuellar, Alex Forsey, Naomi Schurr, Valerie Chen, Timothy Holder, Eoin Kenny, Matt Boyd, Josh Rountree, Samir Wadhwania, Ruaridh Mon-Williams, Claire Liang, Dimos Kontogiorgos, Mike Hagenow, Abriana Stewart-Height, Anthony Favier, Pulkit Verma, Maria Ramos Gonzalez, and Jason Liu. IRG, you taught me that a lab isn't just about research. It's about growing together as humans.

To all my friends, thank you for never giving up on me, even when I was too buried in research to spend as much time together as I wished. You made Boston feel like home. Some of you welcomed me like family when I first arrived, including Feiyi Wang, Bailing Zhang, Jingxian Zhang, Xiaolin Wu, Ang Gao, and Quan Cai. Some I hadn't seen since high school, but when we reconnected, it felt like no time had passed. That was true for friends like Yijie Ren, Munan Hou, and Yingbei Wang. Some of you are chasing academic or startup dreams with relentless focus, and your drive inspired me to keep going. I especially thought of friends such as Yujie Wang, Yiou Wang, Rachel Holladay, Wenhao Luo, and Yifan Hou. Others brought joy and laughter into everyday life, such as Shiheng Jiang, Yun Wu, Luna Zhang, and Fango Lin. Every time we hung out, I couldn't help but wish my life could be as fun as yours. During my most stressful times, friends such as Daqing Yi, Yanglan Ou, Anqi Li, Rui Zhu, and Yiming Liu were always there, offering to walk, talk, call, or simply listen. From all of you, I've learned that friendship is the quiet force that keeps everything else moving, especially through the hardest chapters.

To my parents, thank you for supporting me with your quiet strength and unwavering love. You taught me the value of reflection, how you write monthly and yearly reports to evaluate and replan. That habit stayed with me. It's why I put so much thought into my term evaluations with Julie, and why I spent so much time writing these acknowledgments. You showed me that progress comes from consistency and reflection.

Whenever you visit, you fill my home with care. Dad cooks every day, and when you're back in China, you send me recipe videos so I can cook for myself. That's also the secret behind the tofu I bring to IRG Thanksmas every year. When you're not cooking, you're staying active: Mom teaches me yoga, and Dad takes me to the gym. This year, we even ran the Boston 5K together with Reina. And as anyone who has noticed my wardrobe upgrades knows, your fashion advice always finds its way into my closet when you're here. You've given me more than I can express. You may not be engineers, but you've engineered my life with love, resilience, and quiet strength.

Now, I want to thank you, Reina. Life often feels like a sine wave with ups and downs. But if you zoom in, it's not smooth at all. It's more like a jagged roller coaster. It is probably like a Weierstrass function: continuous everywhere, but differentiable nowhere. That's how life has felt at times. And through so many of those bumps and twists, you were there, quietly nudging my path upward, both intellectually and emotionally.

But your support wasn't easy to give. You've been on your own challenging journey. You completed your master's at Northeastern and faced many struggles applying to PhD programs. Your curve had plenty of dips, too. When we were both at our low points, we coped in the best way we knew: by eating. We'd order delivery with more than two entrées, and more than once, restaurants gave us four sets of chopsticks, assuming we were hosting friends. But no. It was just us.

Despite the challenges, this journey with you has been filled with love, growth, and laughter. That's why I proposed last year, and just last month, we got married at the MIT Chapel. In math, continuity keeps a function unbroken. In life, it's you who keep us from falling apart. Life, like the Weierstrass function, is full of hidden bumps. But love is the continuity that carries us forward.

Finally, to everyone who played a part in this journey, thank you. This thesis may bear my name, but it carries the mark of every person who helped me become who I am.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"One size does not fit all."

—Proverb

Robots have the potential to enhance human well-being by assisting with daily activities. This is especially critical as we face an aging global population and a growing number of people living with disabilities. Providing accessible and personalized support is becoming a major societal challenge. My research aims to build robots that adapt their assistance to individual users in everyday life.

The global population aged 65 and older is rapidly increasing and will soon surpass the number of children under 14 [2]. Simultaneously, one in six people live with significant disabilities, making healthcare up to six times less accessible for them [3]. These challenges are compounded by a projected shortage of 14.5 million healthcare workers worldwide [4]. Robots offer a promising solution to support older adults and people with disabilities in their daily lives. Beyond elderly care and healthcare, robots can also assist the broader population by automating household chores and routine tasks, freeing up time for more meaningful activities [5].

Personalization is key. Each person has unique preferences and behaviors, shaped by their comfort, physical capabilities, and interaction style. For a robot to provide meaningful assistance, it must adapt to the specific needs of the individual. For example, in robot-assisted dressing (fig. 1.1), one person may prefer the robot to move slowly to feel safe [6, 7], while another may prefer faster motion for efficiency. Some users may actively move their arm to help guide the clothing, reflecting a high preference for autonomy [8–10], whereas others may remain passive due to discomfort or limited mobility [11]. To deliver safe, comfortable, and effective assistance, robots must personalize their behavior to align with these individual differences [12].

Personalization is fundamentally challenging because robots rarely have a complete understanding of the human they are assisting. Acting without this understanding can lead to ineffective or even unsafe assistance. A central reason is that robots must operate under *uncertainty* about the user's preferences, behaviors, and needs, which stems from two key sources. First, there is often a scarcity of data: when a robot is deployed with a new user, or when a user's condition changes, the robot may have very limited interaction history to draw from [11]. Second, even when data is available, human behavior can be noisy or ambiguous.

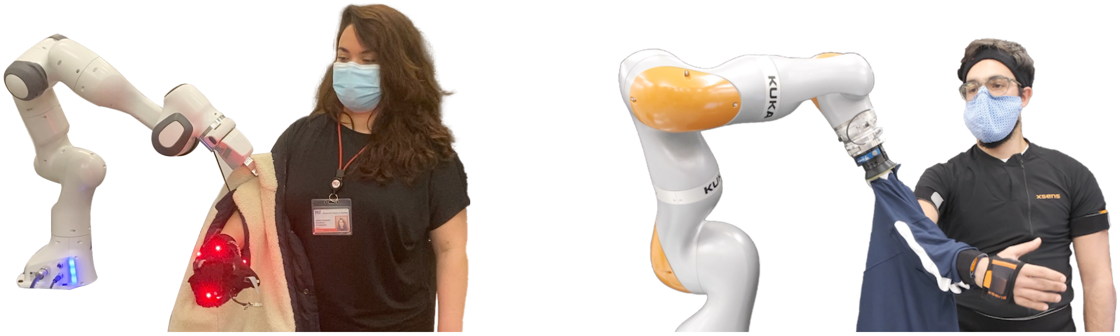


Figure 1.1: This thesis focuses its real-world deployment on robot-assisted dressing [25, 26], though its core contributions apply to broader human-robot interaction tasks.

For example, if a person moves more actively during dressing, they might be signaling a desire for greater autonomy, or they might simply be in a hurry. Without explicitly accounting for such uncertainty, the robot may act on incorrect assumptions, leading to unsafe or ineffective assistance [13].

Human-robot interaction research has developed many models of human preferences and behaviors, along with adaptive algorithms applied in domains such as education [14], factory assembly [15–17], autonomous driving [18], and assistive tasks like handover [19], shared autonomy [8, 9, 20], exoskeletons [21], prostheses [22], dressing [10, 23], and feeding [12, 24]. However, most prior work does not explicitly quantify or manage the robot's uncertainty about its understanding of human preferences and behavior, which could limit the robot's ability to make reliable decisions when data is scarce or ambiguous.

To address this challenge, this thesis asks a fundamental question:

How should a robot behave when it is uncertain about the human?

Rather than avoiding uncertainty, I propose embracing it as a design principle for learning, estimation, and planning. This thesis presents a unified framework that enables robots to leverage uncertainty to offer personalized, safe, and effective assistance. Across three key contributions:

1. When the robot is *learning* about human preferences, I develop an algorithm that incorporates implicit cognitive signals like response time to *reduce uncertainty* about preference strength, enabling more efficient personalization.

2. When the robot is *estimating* hidden human states, I develop an algorithm that *explicitly represents and respects uncertainty* about human motion during interaction and how that motion is perceived through the robot's sensors, leading to more robust estimation.

3. When the robot is *planning* its motion to physically assist the human, I redefine safety to allow low-impact contact. This relaxes the overly conservative constraint of avoiding all contact, which can cause the robot to freeze under uncertainty. By permitting safe contact, the robot is able to *act safely under uncertainty* about future human motion and provide both safe and efficient assistance.

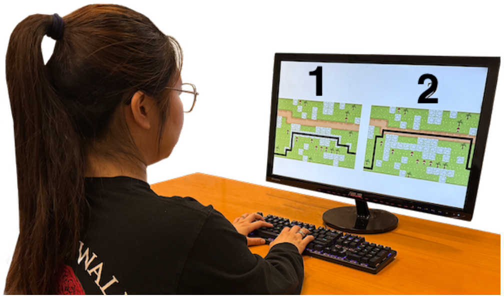


Figure 1.2: In preference learning, the robot asks a user to compare two robot trajectories. In addition to observing the user's choice, this thesis incorporates the user's *response time* (how long they take to decide) as a signal of preference strength. By modeling this implicit cognitive feedback, the robot can reduce uncertainty about user preferences and accelerate learning [30].

Each contribution focuses on a different aspect of robot decision-making, including learning, estimation, and planning, and together form a foundation for reliable, uncertainty-aware robot personalization. While robot-assisted dressing [10, 23, 25–29] is used as the primary application, the ideas generalize to other collaborative human-robot tasks.

Below, I summarize each contribution in more detail.

1.1 Reducing Uncertainty Using Cognitive Feedback for Fast Preference Learning

Understanding a person's preferences, such as which clothes to wear or which robot trajectory to follow, is fundamental to providing personalized assistance. A widely used approach asks users to compare two options, e.g., different robot speeds or paths, and uses their choices to infer preferences. This framework, illustrated in fig. 1.2, is common in robotics [31, 32] and has also been applied in domains such as recommender systems [33–36] and large language model alignment [37–41]. As discussed in the related work chapter (section 2.1.2), these methods typically rely on binary choices, which provide limited information, especially about how strongly the user prefers one option over another.

This thesis's first contribution [30] proposes using cognitive feedback, specifically, the human's response time during comparisons, as an additional source of information. Psychology research shows that response time reflects hesitation and preference strength [42]. The key insight is that strong preferences often lead to faster decisions, while weak preferences result in longer deliberation.

Building on the drift-diffusion model from psychology [42, 43], I developed a preference learning algorithm that integrates both binary choices (explicit comparative feedback, section 2.1.2) to identify a user's preferred option and response times (implicit feedback, section 2.2) to assess the strength of those preferences. Theoretical analysis shows that incorporating response time reduces uncertainty in the estimated preferences and improves learning efficiency, especially when the user has strong opinions. Empirical results in simulated recommender systems show that this method reduces the misidentification rate of the most

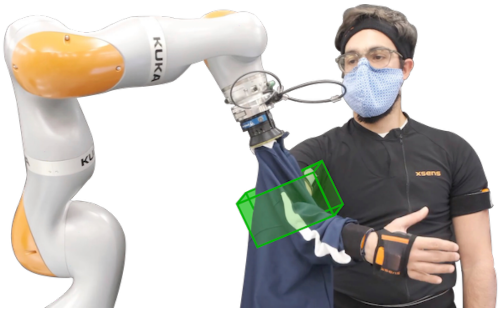


Figure 1.3: In robot-assisted dressing, parts of the human body (such as the elbow) may be occluded from sensors. This thesis introduces a state estimation algorithm that explicitly models uncertainty in both human motion and sensor observations. By representing and respecting this uncertainty, the robot produces more reliable estimates, which is critical for safe physical assistance [26].

preferred option by up to 55 % compared to choice-only baselines.

This work considers the simplified setting, where each comparative query presents a pair of static options, such as visualized robot paths [31] or food items [44–46]. While this abstraction sidesteps the complexity of physical robot execution, it captures the core challenge of learning from limited binary feedback. Looking ahead, this work lays the foundation for future extensions to robotics settings where each query involves dynamic, time-extended trajectories, such as a dressing robot executing two motion strategies that the user experiences and evaluates [32, 47, 48]. See section 7.2 for more.

Takeaway. Human feedback isn't just about what people choose. *How long* they take to choose reveals how strongly they feel. By leveraging this cognitive signal, robots can learn preferences more efficiently under limited feedback.

This work is published in Li* et al. [30] and presented in detail in chapter 4.

1.2 Representing and Respecting Uncertainty for Robust State Estimation

In physical assistance tasks, robots often need to estimate aspects of the human body that are hidden or only partially observable. For example, during dressing, the person's elbow is frequently occluded by clothing. As shown in fig. 1.3, the robot must infer its position from indirect sensor data, such as interaction forces, which is an essential step for planning safe assistance.

Accurate estimation requires two components: a dynamics model describing how the human arm moves during interaction, and an observation model capturing how that motion is perceived through the robot's sensors. These models are often trained from data, as human behavior can vary widely across individuals due to differences in autonomy preference, motor ability, or rehabilitation status. But when training data is scarce or noisy, the resulting models may be inaccurate, leading to errors in estimating the human state, and potentially causing discomfort or harm.

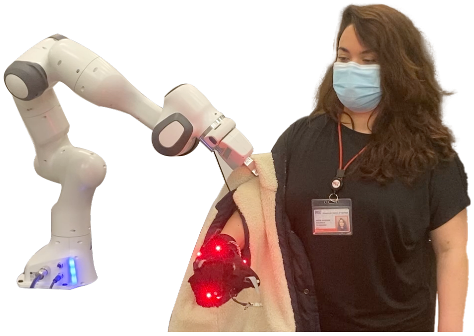


Figure 1.4: When assisting a person during dressing, the robot must plan safe motions under uncertainty about human behavior. Traditional safety constraints often prohibit any contact, leading to robot freezing [49]. This thesis proposes a relaxed safety definition that allows either collision avoidance or low-impact contact, enabling the robot to continue assisting effectively even under uncertainty [25].

To ensure safety under these conditions, this thesis proposes a set-based estimation algorithm that explicitly models and respects the robot's uncertainty about the learned dynamics and observation models. The method builds on techniques from control theory, which represent uncertainty using geometric bounds. Specifically, the robot maintains a bounding box (see green box in fig. 1.3) that is guaranteed to contain the true elbow position, even when models are imperfect. Unlike prior approaches that rely on hand-crafted models, this method learns from data and adjusts the size of the bounding box dynamically, based on the calibrated uncertainty.

Empirical results from robot-assisted dressing show that this method produces tight and reliable estimates: the resulting ∼9cm × 9cm × 9cm boxes contained the true elbow position 92 % of the time. In comparison, baseline methods either produced boxes that contained the elbow only 68 % of the time, or up to 81 % of the time, but required much larger boxes of ∼18cm × 15cm × 19cm. These results demonstrate that explicitly accounting for uncertainty in both human behavior and robot sensing leads to more reliable estimation, enabling robots to act safely even when human states are only partially known.

Takeaway. When the human state is hidden or uncertain, the robot must estimate conservatively, by explicitly modeling what it does not know, to ensure safety in physical interaction.

This work is published in Li* et al. [26] and presented in detail in chapter 5.

1.3 Acting under Uncertainty by Allowing Contact for Safe and Efficient Planning

Planning robot motion in close proximity to the human body requires safety guarantees. Traditionally, safety is defined as strict collision avoidance [50], which prohibits any physical contact between the robot and the human. However, when the robot is uncertain about how the human might move, due to scarce noisy data, this constraint can cause the robot to freeze, waiting to reduce uncertainty before acting [49]. This is especially problematic in physically assistive settings like dressing, as shown in fig. 1.4, where the robot must move garments around the human arm through tight spaces, and some level of contact is often inevitable.

To address this challenge, this thesis's third contribution [25] proposes a more flexible safety definition: the robot is safe if it either avoids contact or ensures that any contact has low physical impact [51]. This two-pronged safety constraint allows robots to move even when the future human motion is uncertain, as long as any resulting contact remains within a safe threshold.

I implement this safety constraint within a learning-based model predictive control framework [52]. The robot learns a model of human motion from online data and calibrates its uncertainty about that model. It then predicts all plausible future human motions and optimizes its own actions to remain within the safety constraint.

Empirical results from a robot-assisted dressing study showed that the proposed two-pronged safety constraint enabled the robot to complete the task efficiently, even with limited data about human arm motion. In contrast, a strict collision-avoidance constraint caused the robot to frequently freeze, waiting to collect more human data and reduce uncertainty before acting. With the two-pronged safety constraint, the robot completed the task in 25 seconds, achieving a 78 % reduction in task time, while maintaining safety. These results demonstrate that when human motion is uncertain, allowing low-impact contact enables robots to assist more efficiently without compromising safety.

Takeaway. Safety does not have to mean inaction. By allowing safe contact, robots can remain effective even when uncertain about how the human will move.

This work is published in Li et al. [25] and presented in detail in chapter 6.

Together, these contributions provide a coherent approach to uncertainty-aware personalization. Rather than ignoring what robots do not know, this thesis shows how to reduce uncertainty through richer human feedback, respect uncertainty in estimation, and act under uncertainty during physical assistance. These capabilities enable safer interaction, faster learning, and more human-aligned assistance.

The rest of this thesis builds on this central idea. Chapter 2 reviews prior personalization approaches in human-robot interaction and highlights their limitations in managing uncertainty. Chapter 3 introduces the technical foundations. Chapters 4 to 6 present the three key contributions in detail. Finally, Chapter 7 reflects on the implications of this work and outlines directions for future research.

Chapter 2

Related Work

"If I have seen further, it is by standing on the shoulders of giants."

—Isaac Newton

Robots that assist people must adapt to each user's unique preferences and behaviors. This process of personalization is central to making robot assistance safe and effective in real-world settings. To personalize their behavior, robots must learn about humans, either by gathering human feedback directly or by observing how humans interact with the system.

The literature on personalized robotics can be broadly categorized based on the source of human feedback:

- *Explicit feedback*, such as user-provided ratings [53], comparisons [31], and demonstrations [54], is deliberately communicated to the robot.
- *Implicit feedback*, such as human motion [55, 56] and eye movements [57], is passively observed and interpreted as a reflection of human intent.
- *Transfer learning* approaches aim to leverage data from previous users to quickly personalize to a new user.

This chapter reviews key work across these three areas. Section 2.1 surveys personalization methods based on explicit human feedback. Section 2.2 discusses how robots use implicit cues to infer user preferences, with an emphasis on modeling uncertainty. Section 2.3 reviews learning-to-personalize approaches across multiple users. Each section concludes by identifying key limitations that motivate this thesis's contributions.

2.1 Personalization with Explicit Human Feedback

Explicit human feedback refers to information that users deliberately provide to guide the robot's behavior, such as numerical evaluations [53], comparisons [31], or demonstrations [54]. This type of feedback is often more informative than passive implicit human feedback, as it directly reflects the user's preferences. However, it can also be cognitively demanding to provide, especially in real-time or interactive settings.

Researchers have developed many algorithms to interpret and optimize robot behavior based on different forms of explicit feedback. This section reviews three major categories of explicit feedback: (1) evaluative feedback, (2) comparative feedback, and (3) demonstrative and corrective feedback. We also highlight how prior work typically relies on a single feedback modality, without leveraging additional implicit signals that may accompany explicit feedback. This gap motivates the first contribution of this thesis: combining binary comparisons with response times to reduce uncertainty and improve preference learning efficiency.

2.1.1 Evaluative Feedback

A common form of explicit feedback is a numerical score indicating how desirable a robot action or trajectory is [58]. This evaluative feedback allows users to directly shape the robot's behavior based on their preferences.

A well-known framework in this category is TAMER (Training an Agent Manually via Evaluative Reinforcement), which enables robots to learn from real-time human-provided scalar rewards [53]. Variants of TAMER interpret feedback as reward functions or Q-values [53, 59–63], or advantage functions [64].

Instead of scoring individual actions, some systems ask users to rate entire trajectories [20], a strategy that reduces cognitive load but sacrifices granularity. These ratings have been used in both empirical studies [20] and theoretical analyses [65, 66].

However, subjective ratings can be inconsistent and biased. For instance, users may unintentionally create positive feedback loops, reinforcing suboptimal behaviors [64]. The feedback is also often sparse, since providing real-time ratings can interrupt the task at hand.

2.1.2 Comparative Feedback

In many tasks, it is difficult for users to provide absolute ratings or specify desired actions. Instead, it can be more natural for users to express their preferences by comparing options. This form of comparative feedback is cognitively lighter and more reliable than evaluative feedback [32].

A large body of work has explored reward learning from pairwise comparisons. In these settings, the robot presents two or more trajectory options and asks the user which one they prefer. These preferences are then used to learn a reward function or preference model. For example, Christiano et al. [67] proposed a deep learning framework that learns from trajectory comparisons, and Sadigh et al. [31] developed an active learning algorithm that selects informative comparisons to efficiently infer reward functions.

Subsequent work has expanded pairwise-comparison-based reward learning in many directions. Some methods improve query efficiency, for example, by generating batch queries [68], comparing more than two options at a time [69], or avoiding queries that are too obvious to be informative [70]. Others extend pairwise-comparison-based learning to handle more complex reward structures, including nonstationary [71], nonlinear [72], or multimodal reward functions [73]. Several approaches incorporate richer forms of explicit feedback, such as numerical ratings [74], corrective actions [75], or feature-level queries [76].

Pairwise-comparison-based learning has also been combined with off-policy reinforcement learning and unsupervised pretraining [77], as well as meta-learning for few-shot generalization [78]. Additionally, pairwise-comparison-based learning has also been extended to ranking-based learning. Brown et al. [79] proposed a method that learns from a batch of ranked trajectories, which was later extended to automatically generate rankings [80, 81], scale to high-dimensional tasks [82], and integrate with offline reinforcement learning [83]. Finally, pairwise-comparison-based reinforcement learning has received increasing theoretical attention, with recent work developing sample-efficient algorithms for learning from comparisons [84–87].

In robot personalization, comparative feedback has been used in various applications. For instance, Akbarzadeh, Lobarinas, and Kehtarnavaz [88] used pairwise comparisons to personalize hearing aid parameters. In exoskeleton control, Tucker et al. [32] developed a Bayesian optimization algorithm that learns user preferences from comparisons, and extended it to higher-dimensional settings [47] and dynamic locomotion tasks [48]. Further developments incorporated safety constraints [89–91].

2.1.3 Demonstrative and Corrective Feedback

Demonstrations and physical corrections offer a rich form of explicit feedback. Instead of rating or comparing options, users directly teach the robot by showing desired behaviors through example trajectories, labeled actions, or real-time physical guidance.

These feedback modes are central to imitation learning, learning from demonstration, and apprenticeship learning [54]. In offline settings, users provide demonstrations of expert behavior, which the robot imitates using supervised learning or inverse reinforcement learning (IRL) [92–94]. In interactive settings, users may label correct actions at visited states [95] or intervene in real time to correct behavior through physical input [96–107].

In robot personalization, such feedback has been used to adapt robot behavior to individual users. For example, in robot-assisted dressing, Gao, Chang, and Demiris [27] proposed a system that adapts its trajectory based on physical corrections from the user; this was later extended to preserve natural posture [108]. Gopinath, Jain, and Argall [8] used verbal feedback to optimize shared-control parameters. Demonstrations have also been used to guide policy optimization. Cakmak et al. [109] collected examples of good and bad policies in simulation, enabling the robot to find solutions closer to the good examples and farther from the bad ones. Batzianoulis et al. [110] learned human reward functions from suboptimal trajectories, using signals from brain-computer interfaces to evaluate performance.

While demonstrations and corrections can be highly informative, they impose significant user burden. Demonstrations require task expertise and motor control, while physical corrections demand sustained attention and engagement. As such, these forms of feedback may not be practical in high-load or assistive settings, where users have limited bandwidth or physical capacity.

2.1.4 Summary

Across a wide range of applications, explicit human feedback has proven effective for personalizing robot behavior. Each type of feedback comes with trade-offs:

- *Evaluative feedback* (section 2.1.1) is informative, but often not reliable [32], and suffers from issues, e.g., positive reward cycles [64].

- *Comparative feedback* (section 2.1.2) is more robust and intuitive but requires generating and evaluating multiple behaviors.

- *Demonstrative and corrective feedback* (section 2.1.3) provides rich information but demands high user effort and expertise [58].

Most prior work treats these feedback signals independently and aims to extract as much information as possible from each explicit response. However, they often overlook implicit behavioral signals that naturally accompany explicit feedback, such as response times and hesitation. These signals can reveal how confident or strongly the user feels about their choices, providing information that is otherwise lost when using only the binary or numeric label.

This thesis's first key contribution, introduced in chapter 4, proposes to integrate human response time (an implicit signal) with binary comparisons (an explicit signal) to more efficiently reduce uncertainty and learn user preferences. By combining these two modalities through a cognitively grounded decision-making model, the robot can extract richer information from each query without requiring additional effort from the user.

2.2 Personalization with Implicit Human Feedback

Unlike explicit feedback, which is deliberately communicated by the user, implicit feedback arises passively during interaction. Human behaviors, such as motion, gaze, timing, and physiological signals, often reflect internal preferences, intentions, or discomfort, even if they are not consciously conveyed. By interpreting these signals, robots can personalize assistance without increasing the user's cognitive or physical workload.

Broadly, prior work on learning from implicit feedback falls into three categories:

- *Objective metric-based personalization*, which treats human behavior or physiology (e.g., metabolic cost and keystroke patterns) as implicit indicators of task quality or user preference, and optimizes robot assistance to improve these metrics.

- *Black-box behavioral model-based personalization* [13], which learns mappings from past interactions to future human behavior, and uses these models to predict human motion and adapt robot planning accordingly.

- *Theory-of-Mind model-based personalization* [13], which assumes that humans act approximately rationally to optimize latent reward functions, and uses these models to infer those reward functions and plan robot behavior accordingly.

This section reviews each approach and examines its role in robot personalization. A common limitation across these methods is the lack of uncertainty quantification: most do not explicitly represent the robot's uncertainty about the learned human model, i.e., how human intentions and preferences relate to observed implicit feedback. As a result, these systems can be fragile when data is scarce, ambiguous, or noisy. In contrast, this thesis's second and third contributions develop algorithms that explicitly represent, respect, and act under uncertainty, enabling safer and more effective personalization in real-world human-robot interaction.

2.2.1 Objective Metric-Based Personalization

One approach to learning from implicit feedback is to treat measurable human behavior or physiology as an objective signal reflecting task success, user comfort, or preference. These metrics are not explicitly provided by the user, but arise naturally during interaction and can be passively monitored, making them particularly valuable in assistive settings where user effort must be minimized.

In physical assistance, researchers have used physiological signals to guide personalization. For example, Slade et al. [111] measured metabolic cost using respirometry equipment to personalize exoskeleton parameters for more efficient walking. Behavioral metrics such as keystroke patterns (e.g., backspace frequency) have been used to infer typing difficulty and personalize assistive typing systems [112]. In physical activity coaching, Hochberg et al. [113] used the number of active minutes per day to optimize robot behavior and encourage exercise. Similarly, other works use signals from body-worn sensors to evaluate how effective a robot's assistance is [21, 114–117].

Objective feedback is also widely used in intelligent tutoring systems, where student performance, particularly answer correctness, serves as a clear and interpretable signal of learning progress [118]. Many works in this domain model personalization as a sequential decision-making problem, using frameworks such as multi-armed bandits [118–124], contextual bandits [125], or reinforcement learning [126–128] to optimize teaching strategies based on observed student behavior. These systems adaptively select content or feedback that best supports each learner's needs, based on their ongoing implicit performance signals.

Objective evaluative feedback is often more reliable than subjective ratings, particularly in domains with clear performance metrics or access to physiological signals from wearable sensors. However, such feedback is typically limited to specific applications and may not capture the full spectrum of user preferences, for example, low metabolic cost does not necessarily indicate comfort or a sense of autonomy. Moreover, these approaches often assume that the observed metrics are clean and directly reflect user intent, overlooking the uncertainty in how external behavior maps to internal goals or preferences. In contrast, this thesis develops estimation and control algorithms that explicitly model and respond to this uncertainty, enabling more robust, adaptive, and user-aligned robot assistance.

2.2.2 Black-box Behavioral Model-Based Personalization

Black-box human models aim to predict human behavior directly from interaction history, without assuming a specific cognitive structure. These models are typically trained on datasets of human behavior and deployed within learning or control pipelines to adapt robot assistance accordingly.

The black-box approach offers modeling flexibility and has been widely adopted in real-world human-robot collaboration scenarios. However, because these models often do not represent uncertainty, they can produce unreliable predictions when human behavior deviates from the training data due to noisy measurements and distribution shifts.

Black-box models come in several forms, which we review below:

- *Markovian policies*, where human actions depend only on the current state or short interaction history.
- *Hierarchical policies*, which include both high-level modes (e.g., intentions) and low-level motion policies.
- *Trajectory models*, which learn structured models of full human trajectories.
- *Hierarchical trajectory models*, which incorporate latent high-level modes within trajectory modeling frameworks.

In what follows, we review representative examples of each model class, highlight their strengths and limitations, and discuss how they relate to this thesis.

Markovian Policy

One of the simplest forms of black-box human modeling assumes that the human follows a Markovian policy: that is, their next action depends only on the current state or a short history of past interactions. These policies are typically learned directly from data and integrated into robot planning pipelines.

For example, Nikolaidis et al. [129] proposed a cross-training framework where a robot learns a Markovian human policy during interaction and uses it to improve coordination in shared tasks. Chen et al. [130] incorporated human trust dynamics into a Markovian policy and proposed to plan robot trajectories under partial observability of trust. Other works model human behavior as Markovian policies and deploy online model-free reinforcement learning [14, 131–140], offline model-free reinforcement learning [138, 141], and PD control design [142], to personalize robot assistance.

While these methods enable personalization, they generally do not explicitly represent uncertainty about the human model. As a result, when data is scarce or noisy, human predictions can be unreliable, leading to undesirable or unsafe robot behavior. In contrast, this thesis emphasizes the importance of uncertainty-aware estimation and planning: explicitly modeling the robot's uncertainty about human behavior improves robustness and safety in real-time interaction.

A related line of research focuses on tracking reference trajectories generated by humans. In many physical human-robot interaction tasks, such as rehabilitation or dressing, the human limb moves along a trajectory that the robot must follow, despite not knowing the trajectory in advance. To address this, several works [143–149] model the human as an impedance controller with unknown stiffness, damping, and reference trajectory. The robot then learns to track this latent trajectory using adaptive control, with formal guarantees on tracking error stability. Similar approaches have been applied to robotic knee prostheses, where the goal is to imitate natural human gait or track the motion of the intact limb using adaptive controllers [150, 151].

Like this thesis, these methods handle uncertainty by ensuring stability in the presence of unknown parameters. However, they are primarily designed for tasks where the robot is expected to track a human-generated trajectory. In contrast, this thesis considers more general and flexible human-robot interactive tasks, where the robot may need to estimate latent human states, reason about uncertainty in human states and behavior, and compute actions that completes interactive tasks while ensuring human safety. These broader settings are addressed through the optimal control formulations introduced in this thesis's second and third key contributions (chapters 5 and 6).

Hierarchical Policy

Hierarchical policies provide a richer representation of human behavior by decomposing it into high-level modes (e.g., intentions, subgoals) and low-level motion policies. This structure captures the intuition that humans operate at multiple levels of abstraction: for instance, a person may decide to reach for a cup (high-level intention) and then execute a reaching motion (low-level behavior). These models are particularly useful for capturing complex or multimodal behaviors.

Early work often assumed that the set of high-level human modes was known in advance. For example, Nikolaidis, Hsu, and Srinivasa [152] modeled humans as switching between predefined modes, such as adaptive or stubborn, and used a bounded-memory probabilistic automaton to capture transitions between these modes based on recent interaction history. These modes were combined with low-level Markovian motion policies and used for real-time inference and planning. This framework was later extended to shared autonomy [153] and verbal communication scenarios [154].

Park, Park, and Manocha [155] proposed a hierarchical human policy in which both high-level mode transitions and low-level motion trajectories were learned offline via supervised learning. At runtime, the robot performs online Bayesian inference to estimate the current mode, predict future human behavior, and plan accordingly to ensure safety. Similarly, a line of works [57, 156–158] modeled humans using a high-level static latent intent variable alongside low-level Markovian policies, and applied online inference to adapt robot behavior based on the inferred intent. Other approaches [159, 160] jointly learn human mode transitions and motion policies by combining offline reinforcement learning with latent representation learning, enabling personalized robot policies under partially observed human modes.

Later research moved toward learning high-level modes directly from data, rather than specifying them a priori. For example, Unhelkar and Shah [161] formalized human behavior as a hierarchical probabilistic graphical model, with both mode transitions and motion policies modeled as Markovian. They applied Bayesian nonparametrics for parameter estimation and integrated this model into a partially observable Markov decision process framework for decision-making, both with [56] and without [55] verbal communication. Gopinath, Javaremi, and Argall [162] proposed a probabilistic model to capture the gap between the measured and human-intended physical actions in shared autonomy, using supervised learning to estimate parameters and correcting robot commands accordingly.

More recent approaches integrate learning latent high-level modes into end-to-end reinforcement learning. Xie et al. [163] developed an online multi-agent reinforcement learning algorithm that learns a human-mode encoder from interaction history and jointly trains the robot's policy conditioned on this latent state. This work was later extended to actively stabilize the human's behavior [164] and to account for nonstationary human policies during long-term interaction [165].

These hierarchical models offer expressive representations of human behavior and can capture long-term structure in interaction. However, like simpler black-box policies, they typically do not quantify epistemic uncertainty in the learned models. When trained on limited or biased data, the mode transitions or motion predictions may be unreliable, potentially leading to undesirable or unsafe robot behavior. This thesis takes a complementary perspective: rather than relying solely on accurate prediction, it explicitly represents uncertainty in human state estimation and robot planning. By reasoning about what the robot does not know, the system can make more conservative estimates and avoid overconfident decisions under uncertainty.

Trajectory Model

Trajectory models represent human behavior as complete motion trajectories, rather than state-action mappings. Unlike Markovian policies that model human decisions step-by-step, trajectory models capture the temporal structure of entire motion sequences and are often trained to match expert demonstrations or interaction patterns.

A well-known example is Interaction Primitives [166], which use Dynamic Movement Primitives to represent human-robot joint trajectories. Given partial observations of an ongoing trajectory, the system infers the phase of the interaction and predicts future human motion. This approach was later extended with probabilistic movement primitives [167], Bayesian filtering [168], and multi-modal sensor fusion [169].

Trajectory models are especially useful for predicting motion in structured tasks such as collaborative assembly or handover, where typical trajectories are smooth, repeatable, and easily aligned. However, most trajectory models rely on supervised learning and do not explicitly account for uncertainty in prediction. When test-time trajectories deviate from training data due to novel user behavior, occlusions, or sensor noise, the models may extrapolate poorly. Moreover, they offer limited interpretability and do not typically model human preferences or intent. In contrast, this thesis adopts a reinforcement learning and control-theoretic perspective where human motion is estimated at each time step, with uncertainty explicitly modeled and propagated. By using uncertainty as a core design principle, the proposed methods offer robustness in settings with limited training data and ambiguous sensory input.

Hierarchical Trajectory Model

Hierarchical trajectory models extend trajectory-based approaches by incorporating latent high-level modes, such as user intent or task phase, into the trajectory generation process. These models aim to jointly capture the structure of motion over time and the higher-level decision patterns that influence it. As such, they are particularly useful for modeling human behavior in complex, multimodal, or context-dependent tasks.

Early approaches assume the high-level modes are predefined. For example, Koppula and Saxena [170] introduced Anticipatory Temporal Conditional Random Fields, which model human trajectories in terms of latent object affordances and task goals. Mainprice and Berenson [171] used a Gaussian mixture model to represent possible trajectories under different modes, and inferred the human's current mode by matching observed motion. These models were extended for real-time inference [172] and for combining multiple prediction strategies [173].

Other work learns the high-level structure directly from data. Schmerling et al. [174] used conditional variational autoencoders to learn distributions over future human trajectories conditioned on past motion and static human response modes. This method was later extended to multi-agent trajectory forecasting [175, 176] and incorporated into control systems with safety assurance [177].

These models offer strong predictive performance in structured environments and can capture multimodal distributions over future motion. However, they typically operate as black boxes and do not represent uncertainty in the learned structure. This limits robustness when behavior falls outside the training distribution or when user intent is ambiguous. In contrast, this thesis emphasizes explicitly modeling the robot's uncertainty about human motion, particularly in physically assistive tasks where human movement may be occluded or human data is scarce. By maintaining uncertainty-aware estimates and designing controllers that can safely act under this uncertainty, the proposed methods extend beyond purely predictive models to support reliable real-time interaction.

2.2.3 Theory-of-Mind Model-Based Personalization

Theory-of-Mind models incorporate structure from cognitive science and behavioral economics to model humans as approximately rational agents. Instead of directly predicting behavior from data, these models assume that human actions arise from optimizing a latent reward function, which captures the user's goals, preferences, or intentions. The robot then infers this hidden reward function by observing the human's behavior.

This framework provides a principled way to interpret human actions and enables the robot to generalize beyond previously observed behavior. Theory-of-Mind models are especially useful in settings where humans act strategically, adapt over time, or interact with the robot as a partner rather than a passive system.

A common pipeline for Theory-of-Mind model-based personalization involves three steps: first, the human policy is modeled as the result of optimizing an unknown reward function; second, this reward function is inferred using inverse reinforcement learning (IRL); and third, the learned model is used for robot planning, either to assist the human or to interact with them strategically.

Theory-of-Mind-based personalization methods vary in their assumptions about human rationality and structure. In the next subsections, we review three key classes:

- *Rational models*, which assume that humans are optimal in decision-making.

- *Boltzmann noisily rational models*, which assume that humans are bounded-rational in decision-making.

- *Hierarchical Theory-of-Mind models*, which introduce additional latent structure, such as roles, intents, or reasoning levels.

While Theory-of-Mind models offer interpretability and generalization, many prior works do not quantify the robot's uncertainty about the learned human reward functions. As a result, robots may act with uncalibrated confidence, leading to undesirable or unsafe assistance. This thesis addresses this gap by explicitly modeling uncertainty during estimation and control, enabling safer and more adaptive behavior in physically assistive tasks.

Rational Model

A foundational Theory-of-Mind assumption is that humans act rationally: they choose actions that optimize an internal reward function, given their knowledge of the environment. Under this view, personalization becomes a problem of inferring the user's reward function from observed behavior and adapting the robot's actions accordingly.

This idea underlies a wide range of inverse reinforcement learning (IRL) approaches. For instance, in assistive shared autonomy, Dragan and Srinivasa [178] proposed estimating the human's goal based on real-time joystick inputs and blending robot assistance with user control. In autonomous driving, Sadigh et al. [179] modeled other drivers as rational agents optimizing unknown reward functions, and used this model to generate robot behavior that anticipates and influences human responses.

While assuming perfect rationality simplifies modeling, it often fails to capture real-world human behavior, which could be biased. To address this, later works introduce bounded-rational alternatives, which will be reviewed in the next section.

Some efforts have explored online reward learning in continuous control settings. For example, Li et al. [180] modeled human-robot collaboration as a linear-quadratic team game, where both agents optimize a common quadratic cost function. Their algorithm simultaneously learns this cost from human movements while computing optimal robot actions. A follow-up work [181] proposed a more scalable actor-critic variant. These methods incorporate the robot's uncertainty about the human's cost functions implicitly by ensuring the stability of the learning process.

While these works support personalization through reward inference, they are typically limited to trajectory tracking tasks and often assume full observability of the human state. In contrast, this thesis addresses real-time physical assistance scenarios, such as robot-assisted dressing, where the robot must not only generate its own motion plans but also estimate hidden human states from scarce and noisy data. In such settings, effective assistance requires the robot to estimate, plan, and act cautiously under uncertainty. To address this, the second and third contributions of this thesis explicitly model and reason under uncertainty about human behavior, improving both safety and efficiency.

Boltzmann Noisily Rational Model

The Boltzmann noisily rational model relaxes the assumption of perfect human rationality by introducing stochasticity in decision-making. Rather than always choosing the optimal action, humans are modeled as selecting actions probabilistically, with higher-probability actions yielding higher expected reward. This captures bounded rationality and variability in human behavior.

Formally, the human policy is modeled as a Boltzmann distribution over actions:

$$\pi_H(a \mid s; \beta) \propto \exp\left(\beta Q_H(s,a)\right) \quad [182, \text{eq. } (3)],$$

where $Q_H(s,a)$ is the expected value of taking action a in state s. And β is the rationality coefficient, where higher values imply more deterministic, goal-directed behavior.

This model has been used in a range of human-robot interaction tasks. For example, Fridovich-Keil et al. [182] assumed a known reward function (see Fridovich-Keil et al. [182, section 3.3]) and used Bayesian inference to learn the human's rationality coefficient online. The robot then planned its actions using a confidence-aware motion prediction framework. While this work captures uncertainty about rationality, it assumes the human's reward function is known.

Other works extend the model further. Tian et al. [183] incorporated human learning dynamics, where the human updates their internal model of the environment over time. This allows the robot to reason about nonstationary and multimodal human behavior. However, it does not model the robot's uncertainty about these learned components, which may lead to overconfident behavior when the training data is limited or mismatched at test time.

More closely aligned with this thesis, Hu, Nakamura, and Fisac [184] introduced latent intent parameters into the human reward function and used Bayesian inference to estimate both intent and rationality. This allowed the robot to plan conservatively under uncertainty, ensuring safety during interaction. Hu and Fisac [185] generalized the model further by expressing human behavior as an unknown linear combination of known Boltzmann policies, enabling robots to adapt to richer behavioral patterns.

While these works address uncertainty in the human reward model and emphasize safe planning, they assume full observability of the human state. In contrast, this thesis focuses on physical assistance tasks where human motion is only partially observable, such as during dressing, where key body parts may be occluded. This thesis's second contribution addresses this challenge through robust estimation of hidden human states under uncertainty. Building on this, the third contribution introduces a relaxed human safety constraint that allows either collision avoidance or low-impact contact, enabling the robot to act effectively even under uncertainty. Together, these contributions complement prior work by addressing both estimation and planning challenges in physically assistive settings where human behavior is uncertain and partially observed.

Hierarchical Theory-of-Mind Models

Hierarchical Theory-of-Mind models extend basic rationality frameworks by introducing structured latent variables, such as roles, goals, modes, or reasoning levels, that influence human behavior. These models aim to better capture the diversity, ambiguity, and strategic depth of human decision-making in interaction with robots.

One common approach is to assume that each human belongs to a discrete "type" or follows a particular strategy. For example, Nikolaidis et al. [186] clustered demonstrations from different users into behavioral types and learned a separate reward function for each. During interaction, the robot inferred the user's type online and adapted its behavior accordingly.

Other hierarchical models embed latent structure directly into the reward function. Schwarting et al. [18] introduced the concept of social value orientation (SVO), modeling humans as optimizing a weighted combination of their own reward and the robot's reward. The robot maintained a belief over these weights to plan socially compliant driving behavior. SVO has since been extended to competitive multi-agent interactions [187].

Goal-conditioned latent structure is also common. Le et al. [188] used a hierarchical approach in which high-level rewards were learned via IRL and low-level motion was modeled by a goal-conditioned recurrent policy. Wang et al. [189] similarly trained goal-conditioned human motion policies offline, then inferred human goals online via IRL to guide robot decision-making.

A separate line of work augments Boltzmann-rational models with latent modes. For instance, Tian et al. [190] modeled human role switching between leader and follower modes, using online reinforcement learning to infer both the human's rationality and current role while planning safe robot actions.

More sophisticated hierarchical reasoning frameworks, such as level-k and cognitive hierarchy models from behavioral economics, have also been adapted to robotics. These models assume that humans perform limited iterations of strategic reasoning: for example, a level-1 user's best response to a level-0 partner [191]. Such models have been used in autonomous driving settings [192–195], where robots learn to infer the human's reasoning level and rationality while ensuring safety during interaction.

ToM-based personalization enables robots to interact with strategic or nonstationary humans. This is particularly relevant in cooperative tasks with asymmetric information, where either the human learns from the robot [183, 196–199], the robot learns from the human [200, 201], or both learn simultaneously [202]. Other applications include competitive multi-agent games [203] and strategic driving scenarios such as intersections and merges [18, 179, 184, 195, 204, 205].

While these hierarchical models offer strong representational power, they often assume that the structure or latent parameters (e.g., goals, roles, SVO weights) are either known or can be inferred with confidence. In practice, incorrect assumptions or scarce human data can lead to misidentification of modes or strategies, resulting in unsafe or ineffective behavior.

This thesis takes a complementary approach. Rather than leveraging strong structural assumptions about human reasoning, it focuses on uncertainty-aware estimation and control. By explicitly modeling the robot's uncertainty over hidden human states and planning conservatively under that uncertainty, the proposed methods enable robust assistance even when human behavior is uncertain and partially observed.

2.3 Personalization with Transfer Learning

While most personalization methods require robots to learn from scratch for each user, transfer learning aims to accelerate this process by leveraging data from prior users. These approaches seek to extract patterns or representations that generalize across individuals, enabling faster adaptation with less interaction data.

For example, OhnBar, Kitani, and Asakawa [206] proposed a transfer learning framework that first learns history-dependent black-box models from multiple users and transfers them to personalize to a new user. Rudovic et al. [207] used facial and body movement data from multiple children with autism to train a supervised model that predicts affective states during robot-assisted therapy, using demographic features for personalization. Rudovic et al. [208] further developed a reinforcement learning approach to quickly tailor robot policies to new users with minimal interaction data.

Other work has focused on learning latent user representations. For instance, He et al. [209] and Schrum et al. [210] proposed methods that jointly learn an encoder mapping human trajectories to a latent space of user types and a corresponding set of type-conditioned robot policies. These systems support zero-shot or few-shot adaptation by inferring the latent type of a new user and deploying the appropriate policy. Similarly, Huang, Luo, and Liu [211] introduced a meta-learning framework that can quickly adapt to a new user using gradient-based updates derived from user-provided numerical feedback at meta-test time.

In parallel, the field of ad hoc teamwork [212] explores how to design agents that adapt to new partners without prior coordination. This problem can be formalized as a stochastic Bayesian game [213, 214] or as an interactive partially observable Markov decision process [215], where agents reason about hidden states and others' models. Barrett and Stone [216] proposed an online reinforcement learning approach that trains with diverse teammates and selects policies based on a belief over the new partner's type. Related areas include zero-shot coordination [217–220], which trains agents to coordinate with unseen partners without adaptation, and convention transfer [221], which adapts the agent's partner-specific convention behavior for each human user while reusing the same rule-dependent behavior.

These transfer-based methods are well-suited to scenarios where large-scale offline data is available across users. However, they typically rely on strong assumptions about cross-user generalization and may struggle when encountering user behaviors that fall outside the training distribution.

By contrast, this thesis focuses on personalization in real-time, one-on-one human-robot interaction, where the robot must adapt on the fly with minimal prior data. Rather than transferring across users, the proposed methods operate under uncertainty about the current user's behavior and preferences. They emphasize robustness to data scarcity, making them complementary to transfer learning approaches and better suited to safety-critical, assistive settings such as robot-assisted dressing.

Chapter Summary

This chapter reviewed prior work on robot personalization through explicit feedback (section 2.1), implicit behavioral modeling (section 2.2), and transfer learning (section 2.3). Each line of research offers powerful tools for adapting robot behavior to human users, but common limitations remain: many approaches rely on costly human feedback, do not represent uncertainty about human behavior and preferences, or assume access to large-scale multi-user data. In contrast, this thesis focuses on real-time personalization from limited interaction with a single user, where uncertainty is unavoidable. By explicitly modeling and responding to this uncertainty, the proposed methods aim to enable robots that are not only personalized but also safe, robust, and effective in real-world safety-critical human-robot interaction.

Chapter 3

Technical Background

"All models are wrong, but some are useful."

—George Box

In this chapter, we provide nomenclature, background on intervals, ellipsoids, zonotopes, Gaussian Process (GP), and a high-probability bound for Gaussian random variables.

3.1 Nomenclature

In this thesis, calligraphic uppercase symbols denote sets, such as $\mathcal{A}$, uppercase symbols denote matrices, such as A, and lowercase symbols denote scalars or vectors. In addition, $[n]$ denotes the set $\{1, \ldots, n\}$. For a scalar random variable x, the expectation and variance are denoted by $\mathbb{E}[x]$ and $\mathbb{V}[x]$, respectively. The function $\mathrm{sgn}(y)$ denotes the sign of y.

3.2 Intervals, Zonotopes, and Ellipsoids

Intervals, zonotopes, and ellipsoids are high-dimensional geometry shapes. Due to their appealing geometric and computational properties, they are widely used in the robust control community to compute reachable sets. In other words, these shapes intuitively and conveniently parameterize sets that bound the future states that a control system can possibly reach at a given time, given some control input, based on all possible realizations of system noises and uncertainties. In this section, we briefly review these shapes with some of their properties.

3.2.1 Set Operations

The Minkowski sum between two sets, $\mathcal{X}$ and $\mathcal{Y}$, is defined as $\mathcal{X} \oplus \mathcal{Y} \coloneqq \{x + y \colon x \in \mathcal{X}, y \in \mathcal{Y}\}$. Affine transformation for a set, $\mathcal{X}$, is defined as $b \oplus A\mathcal{X} \coloneqq \{b + Ax \colon x \in \mathcal{X}\}$.

3.2.2 Interval

An interval along the real line, denoted by $[a, b]$, is defined as $\{x \in \mathbb{R} \colon a \leq x \leq b\}$. A box in $\mathbb{R}^n$ is a vector of intervals, $([a_1, b_1], \ldots, [a_n, b_n])^\top$, which is defined as follows:

$$\left\{x \coloneqq [x_1, \ldots, x_n]^\top \in \mathbb{R}^n \colon \forall i = 1, \ldots, n \colon a_i \leq x_i \leq b_i\right\}.$$

A zero-centered box with radius $r \in \mathbb{R}^n$ is denoted by $[0 \pm r] \coloneqq ([-r_1, r_1], \ldots, [-r_n, r_n])^\top \subset \mathbb{R}^n$. For more details about intervals, please refer to Alamo, Bravo, and Camacho [222], Rego et al. [223], and Althoff [224].

3.2.3 Zonotope

Zonotopes are convex polytopes that are centrally symmetric [225]. Formally, a zonotope is a set $\mathcal{Z} \subset \mathbb{R}^n$ that is defined as follows:

$$\mathcal{Z} \coloneqq \{c_\mathcal{Z} + G_\mathcal{Z}\xi \colon \xi \in \mathbb{R}^{n_\xi}, \|\xi\|_\infty \leq 1\},$$

where $c_\mathcal{Z} \in \mathbb{R}^n$ denotes the zonotope's center and $G_\mathcal{Z} \in \mathbb{R}^{n \times n_\xi}$ denotes the zonotope's generator matrix. Each column of $G_\mathcal{Z}$ is called a "generator", and ξ contains all generator variables. Sometimes, when convenient, we denote $\mathcal{Z}$ by $\mathcal{Z}(G_\mathcal{Z}, c_\mathcal{Z})$ to explicitly expose the generator matrix and the center. Additionally, given a zonotope $\mathcal{Z}$, we let $(\mathcal{Z})_c$ denote its center and let $(\mathcal{Z})_G$ denote its generator matrix.

Zonotopes are closed under affine transformations and Minkowski sums, both of which can be computed exactly. Formally, $A \cdot \mathcal{Z}(G, c) \oplus b \coloneqq \mathcal{Z}(AG, Ac+b)$ and $\mathcal{Z}_1(G_1, c_1) \oplus \mathcal{Z}_2(G_2, c_2) = \mathcal{Z}([G_1 \;\; G_2], c_1 + c_2)$. For more details about zonotopes, please refer to Alamo, Bravo, and Camacho [222], Rego et al. [223], and Althoff [224].

3.2.4 Ellipsoid

An ellipsoid is a surface that can be obtained from a sphere by deforming it by an affine transformation [226]. Formally, an ellipsoid is a set $\mathcal{E} \subset \mathbb{R}^n$ that is defined as follows:

$$\mathcal{E} \coloneqq \left\{x \in \mathbb{R}^n \colon (x - c_\mathcal{E})^\top Q_\mathcal{E}^{-1}(x - c_\mathcal{E}) \leq 1\right\},$$

where $c_\mathcal{E} \in \mathbb{R}^n$ denotes the ellipsoid's center and $Q_\mathcal{E} \in \mathbb{R}^{n \times n}$ is a symmetric positive definite shape matrix. Sometimes, when convenient, we denote $\mathcal{E}$ by $\mathcal{E}(c_\mathcal{E}, Q_\mathcal{E})$ to explicitly expose the center and the shape matrix.

Ellipsoids are invariant under affine subspace transformations, such that, for $A \in \mathbb{R}^{n \times r}$ with full column rank and $b \in \mathbb{R}^r$, we have that $A \cdot \mathcal{E}(c, Q) \oplus b \coloneqq \mathcal{E}(c+b, AQA^\top)$ [52]. The Minkowski sum between two ellipsoids is, in general, not an ellipsoid anymore, but can be bounded by an ellipsoid. Formally, for all constant $c > 0$, we have that $\mathcal{E}(c_1, Q_1) \oplus \mathcal{E}(c_2, Q_2) \subset \mathcal{E}(\widetilde{c}, \widetilde{Q}_c)$, where $\widetilde{c} = c_1 + c_2$ and $\widetilde{Q}_c = (1 + c^{-1})Q_1 + (1 + c)Q_2$ [52]. For more details about ellipsoids, please refer to Kurzhanski and Vályi [227].

3.3 Gaussian Process and Confidence Intervals

Gaussian Process (GP) is a nonparametric Bayesian approach for supervised learning of nonlinear functions. Rather than assuming a fixed form for the function, GPs define a distribution over possible functions, allowing the model to flexibly adapt to the data. One key advantage of GPs is their ability to provide calibrated uncertainty estimates. In particular, given a noisy training dataset and a testing point, a GP outputs not only a mean prediction, but also a variance prediction reflecting the uncertainty of that prediction. This makes GPs particularly useful in scenarios where managing model uncertainty is crucial, such as active learning, robust control, and decision making under uncertainty.

In this thesis, we will use GPs to learn nonlinear functions, denoted by $f\colon \mathcal{D} \mapsto \mathbb{R}^{n_y}$, where $\mathcal{D} \subset \mathbb{R}^{n_x}$ denotes the function domain. Following Koller et al. [52], we first equivalently reformulate the multi-output function, f, using a single-output *surrogate* function, $f'\colon \mathcal{D} \times \{1, \dots, n_y\} \mapsto \mathbb{R}$. In particular, we define f' by setting $f'(x, j) \coloneqq f_j(x)$ for each dimension $j = 1, \dots, n_y$, where $f_j(x)$ denotes the j-th output of $f(x)$. This reformulate allows us to conveniently apply the standard definition of GP with a scalar output and formulate confidence intervals.

We use a GP, denoted by $GP(m, k)$, to learn f', where the prior mean function, $m\colon \mathbb{R}^{n_x} \mapsto \mathbb{R}$, is set to 0. The function $k\colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \mapsto \mathbb{R}$ denotes the covariance (or kernel) function. Common kernels include the linear, squared exponential, and Matérn kernels.

The system is given a set of n training data points, denoted by $\{(x_i \in \mathcal{D}, y_i \in \mathbb{R}^{n_y})\}_{i=1,2,\dots,n}$. Suppose that each data output is corrupted by an *i.i.d.* Gaussian noise $w \in \mathbb{R}^{n_y}$, where for each dimension $j = 1, \dots, n_y$, the noise $w_j \sim \mathcal{N}(0, \lambda_w^2)$. Formally, for each data input x_i, the corresponding observed data output $y_i = f(x_i) + w$. By conditioning the GP on the n training data points, for each dimension $j = 1, \dots, n_y$, we obtain a posterior mean function, $\mu_{n,j}\colon \mathcal{D} \mapsto \mathbb{R}$, and a posterior variance function, $\sigma_{n,j}^2\colon \mathcal{D} \mapsto \mathbb{R}$, expressed as follows:

$$\mu_{n,j}(x) = k_n(x)^\top \left(K_n + \lambda_w^2 I_n\right)^{-1} y_{n,j}$$
$$\sigma_{n,j}^2(x) = k(x, x) - k_n(x)^\top \left(K_n + \lambda_w^2 I_n\right)^{-1} k_n(x)$$

where the vector $k_n(x) \coloneqq [k(x_1, x), \dots, k(x_n, x)]^\top$, the kernel matrix $K_n \coloneqq [k(x_i, x_{i'})]_{i,i'=1,2,\dots,n}$, the data $y_{n,j} \coloneqq [y_{1,j}, \dots, y_{n,j}]^\top$, and I_n denotes the n-dimensional identity matrix.

If we assume that the true function, f', belongs to the reproducing kernel Hilbert space (RKHS) associated with the kernel, k. The smoothness of f' can be measured via its RKHS norm, denoted by $\|f'\|_k$ [52]. Then, if we further assume that f' is smooth with a small RKHS norm, then we can use the following lemma to construct confidence intervals for the outputs of f', or, equivalently, of f.

Lemma 3.3.1 (Lemma 1 in [228])**.** *Fix f in RKHS with the surrogate function f' satisfies $\|f'\|_k \leq B$. Let $\delta \in (0, 1)$. Suppose that the system observes n data points for the function f, where each dimension of each data point is corrupted by an i.i.d. noise $\sim \mathcal{N}(0, \lambda_w^2)$. Then, with probability at least $(1 - \delta)$, the following holds:*

$$\forall n = 1, 2, \dots, \forall j = 1, \dots, n_y, \forall x \in \mathcal{D}\colon \quad |\mu_{n,j}(x) - f_j(x)| \leq \beta_n \cdot \sigma_{n,j}(x),$$

where $\beta_n = B + \sqrt{2(\gamma_{n \cdot n_y} + \log(1/\delta))}$. Here, the term $\gamma_{n \cdot n_y}$ denotes the information capacity with $n \cdot n_y$ data points for f', which can be bounded if the domain $\mathcal{D}$ is compact [52, 229].

This lemma states that with high probability, jointly for each dimension j, the output of f_j is bounded by a confidence interval centered at the posterior mean prediction $\mu_{n,j}$.

For more details about GP posterior, GP prediction, RKHS, and information capacity, please refer to Koller et al. [52], Srinivas et al. [230], and Chowdhury and Gopalan [231].

3.4 Gaussian Noise Bound

We consider a finite sequence of *i.i.d.* Gaussian noises, denoted by $v_t \in \mathbb{R}^{n_v}$, for each time $t = 1, \ldots, T$, where $T \in \mathbb{N}$ denotes a finite time horizon. With high probability, jointly throughout all time steps, all the noises can be bounded by a zero-centered box in $\mathbb{R}^{n_v}$. Formally:

Lemma 3.4.1 (Gaussian noise bound)**.** *Let $T \in \mathbb{N}$ denote a fixed horizon and $\delta \in (0, 1)$. Let vectors $v_1, \ldots, v_T \in \mathbb{R}^{n_v}$, such that for each time $t = 1, \ldots, T$ and dimension $j = 1, \ldots, n_v$, the noise $v_{t,j} \sim \mathcal{N}(0, \lambda_v^2)$, with $\lambda_v \in \mathbb{R}$. Then, with probability at least $(1 - \delta)$, the following holds:*

$$\forall t = 1, \ldots, T \colon v_t \in \left[0 \pm \sqrt{2}\lambda_v \sqrt{\ln \frac{T n_v}{\delta}}\right]^{n_v} \subset \mathbb{R}^{n_v}.$$

Proof. The proof is similar to those for Lemma 5.1 in Srinivas et al. [229] and Lemma 4 in Berkenkamp [232]. For each time $t = 1, \ldots, T$ and dimension $j = 1, \ldots, n_v$, we bound $v_{t,j}$ by applying the Gaussian error function with a probability budget, $\delta/(T \cdot n_v)$; we then obtain the result via a union bound over all t and j. □

Chapter 4

Reducing Uncertainty about Preferences Using Cognitive Feedback

"Between stimulus and response there is a space.
In that space is our power to choose our response.
In our response lies our growth and our freedom."

—Steven Covey or Viktor E. Frankl

A central question of this thesis is: *How should a robot behave when it is uncertain about the human?* In interactive robotics, such as assistive dressing, shared autonomy, or autonomous driving, robots must often infer a user's preferences from limited feedback. A common approach is to present the human with a pair of options, such as two visualized robot paths [31], and observe their binary choice. This form of comparative feedback is popular because it is easy to implement and places minimal cognitive load on users [74, 233, 234].

Even when the robot presents only one option and asks the human to rate it as "good" or "bad" [235], the feedback can still be viewed as comparative: the user is implicitly comparing the current option to an internal reference. However, whether the comparison involves two presented options or one option versus an internal benchmark, binary feedback reveals only which option is preferred, not how strongly it is preferred. This limited information hinders the robot's ability to reduce uncertainty and personalize efficiently. To address this, researchers have incorporated additional *explicit human feedback*, such as ratings [236, 237], labels [233], and slider bars [41, 74], but these approaches often complicate interfaces and increase cognitive demands [234, 238].

This chapter proposes leveraging *implicit human feedback*, specifically response times, to provide additional insights into preference strength. Unlike explicit feedback, response time is unobtrusive and effortless to measure [239], offering valuable information that complements binary choices [45, 240]. For instance, consider a dressing robot that assists a user each morning and then asks, "Was that good or bad?" Some users may respond "good" most of the time, either out of politeness, low expressiveness, or because they find most executions acceptable. This consistent positive feedback makes it difficult for the robot to determine which of the many "good" trajectories the user truly prefers. Response time can help disambiguate this. Psychological research shows an inverse relationship between response time and preference strength [239]: a fast "good" response may signal strong approval, while a slow "good" may indicate hesitation or weak preference. Thus, even when choices appear similar, response time can uncover subtle differences in preference strength, helping to accelerate personalization.

We focus here on a simplified setting: the robot repeatedly asks a user to choose between static options (e.g., snacks, visualized robot paths). This abstraction captures the core challenge of learning from limited binary feedback while sidestepping the complexity of physical robot execution. This work forms the foundation for future extensions to robotics, where comparative queries may involve dynamic trajectories that unfold over time. In such settings, the user would first experience two robot behaviors (e.g., two dressing motions [32, 47, 48]) before giving feedback. This chapter, therefore, provides a computational and theoretical basis for using response times as a scalable signal in real-time, interactive systems.

Leveraging response times for preference learning presents notable challenges. Psychological research has extensively studied the relationship between human choices and response times [239, 241] using complex models like Drift-Diffusion Models [242] and Race Models [243, 244]. While these models align with both behavioral and neurobiological evidence [245], they rely on computationally intensive methods, such as hierarchical Bayesian inference [246] and maximum likelihood estimation (MLE) [247], to estimate the underlying human utility functions from both human choices and response times, making them impractical for real-time interactive systems. Although faster estimators exist [43, 248–251], they typically estimate the utility functions for a single pair of options without aggregating data across multiple pairs. This limits their ability to leverage structures like linear utility functions, which are widely adopted both in preference learning with large option spaces [31, 34, 36, 252, 253] and in cognitive models for human multi-attribute decision-making [254–256].

To address these challenges, this chapter proposes a computationally efficient method for estimating linear human utility functions from both choices and response times, grounded in the difference-based EZ diffusion model [43, 251]. Our method leverages response times to transform binary choices into richer continuous signals, framing utility estimation as a *linear regression* problem that aggregates data across multiple pairs of options. We compare our estimator to traditional *logistic regression* methods that rely solely on choices [257, 258]. For queries with strong preferences, our theoretical and empirical analyses show that response times complement choices by providing additional information about preference strength. This significantly improves utility estimation compared to using choices alone. For queries with weak preferences, response times add little value but do not degrade performance. *In summary, response times complement choices, particularly for queries with strong preferences.*

Our linear-regression-based estimator integrates seamlessly into algorithms for preference-based bandits with linear human utility functions [257, 258], enabling interactive learning systems to leverage response times for faster learning. We specifically integrated our estimator into the Generalized Successive Elimination algorithm [257] for fixed-budget best-arm identification [259, 260]. Simulations using three real-world datasets [44–46] consistently show that incorporating response times significantly reduces identification errors, compared to traditional methods that rely solely on choices. *To the best of our knowledge, this is the first work to integrate response times into bandits and reinforcement learning.*

In the broader context of this thesis, this chapter contributes to the answer of how robots should behave when uncertain about human preferences: by leveraging naturally available cognitive signals like response time, robots can more efficiently reduce uncertainty without increasing user burden.

Section 4.1 introduces the preference-based linear bandit problem and the difference-based EZ diffusion model. Section 4.2 presents our utility estimator, incorporating both choices and response times, and offers a theoretical comparison to the choice-only estimator. Section 4.3 integrates both estimators into the Generalized Successive Elimination algorithm. Section 4.4 presents empirical results for estimation and bandit learning. Section 4.5 discusses the limitations of our approach. Appendix A.1 reviews response time models, parameter estimation techniques, and their connection to preference-based RL.

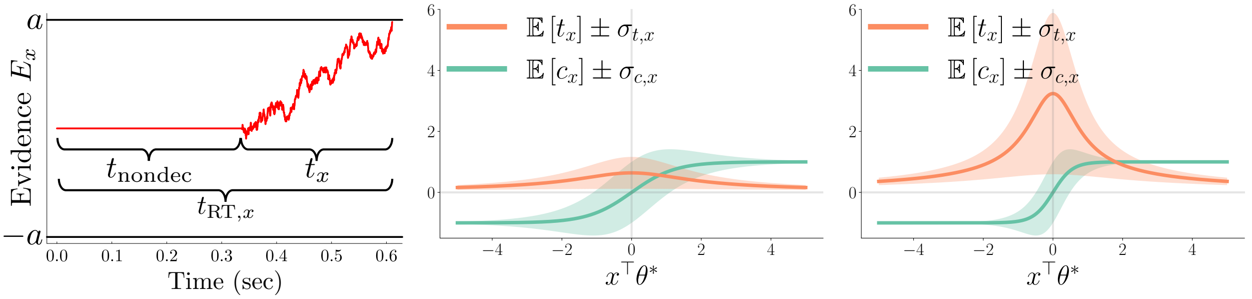
4.1 Problem setting and preliminaries

We model the task of learning human preferences from feedback as a preference-based bandit problem [258, 261]. In each round, the system (or "learner") presents a query consisting of a pair of options. The human chooses the option they prefer, and the system uses this binary choice to update its estimate of the human's underlying utility function. Over time, the goal is to efficiently learn this utility function to identify the most preferred option.

4.1.1 Preference-Based Bandits with a Linear Utility Function

The learner is given a finite set of options (or "arms"), each represented by a feature vector in $\mathcal{Z} \subset \mathbb{R}^d$, and a finite set of binary queries, where each query is the difference between two arms, denoted by $\mathcal{X} \subset \mathbb{R}^d$. For instance, if the learner can query any pair of arms, the query space is $\mathcal{X} = \{z - z' \colon z, z' \in \mathcal{Z}\}$. In the dressing robot example from the beginning of this chapter, the query space is $\mathcal{X} = \{z - z_{\text{skip}} \colon z \in \mathcal{Z}\}$, where z represents purchasing a product and z_{skip} represents skipping (often set as $\mathbf{0}$). For each arm $z \in \mathcal{Z}$, the human utility is assumed to be linear in the feature space, defined as $u_z := z^\top \theta^*$, where $\theta^* \in \mathbb{R}^d$ represents the human's preference parameters. For any query $x \in \mathcal{X}$, the utility difference is then defined as $u_x := x^\top \theta^*$.

Given a query $x := z_1 - z_2 \in \mathcal{X}$, we model human choices and response times using the difference-based EZ-Diffusion Model (dEZDM) [43, 251], integrated with our linear utility structure. (See appendix A.1.1 for a comparison with other models.) This model interprets human decision-making as a stochastic process in which evidence accumulates over time to compare two options. As shown in fig. 4.1a, after receiving a query x, the human first spends a fixed amount of non-decision time, denoted by $t_{\text{nondec}} > 0$, to perceive and encode the query. Then, evidence E_x accumulates over time following a Brownian motion with drift $x^\top \theta^*$ and two symmetric absorbing barriers, $a > 0$ and $-a$. Specifically, at time $t_{\text{nondec}} + \tau$ where $\tau \geq 0$, the evidence is $E_{x,\tau} = x^\top \theta^* \cdot \tau + B(\tau)$, where $B(\tau) \sim \mathcal{N}(0, \tau)$ is standard Brownian motion. This process continues until the evidence reaches either the upper barrier a or the lower barrier $-a$, at which point a decision is made. The random stopping time, $t_x := \min\{\tau > 0 \colon E_{x,\tau} \in \{a, -a\}\}$, represents the decision time. If $E_{x,t_x} = a$, the human chooses z_1; if $E_{x,t_x} = -a$, they choose z_2. The choice is represented by the random variable c_x, where $c_x = 1$ if z_1 is chosen, and -1 if z_2 is chosen. The total response time, $t_{\text{RT},x}$, is the sum of the non-decision time and the decision time: $t_{\text{RT},x} = t_{\text{nondec}} + t_x$. The choice probability, expected choice, choice variance, and expected decision time are given as follows



(a) (b) $a = 0.8$ (c) $a = 1.8$

Figure 4.1: (a) depicts the human decision-making process for a binary query $x \in \mathcal{X}$, where the human selects between two arms. The human first spends a fixed non-decision time t_{nondec} encoding the query. Then, the human's evidence accumulates according to a Brownian motion with drift $x^\top\theta^*$. When the evidence reaches the upper barrier a or lower barrier $-a$, the human makes a choice, denoted by $c_x = 1$ or $c_x = -1$, respectively. The random stopping time of the accumulation process is the decision time t_x, and the total response time is $t_{\text{RT},x} = t_{\text{nondec}} + t_x$. (b) and (c) plot the expected choice $\mathbb{E}[c_x]$ and the expected decision time $\mathbb{E}[t_x]$, with shaded regions representing one standard deviation, plotted as functions of the utility difference $x^\top\theta^*$ for two barrier values a.

[262, eq. (A.16) and (A.17)]:

$$\begin{aligned}\forall x \in \mathcal{X}\colon \mathbb{P}\left[c_x = 1\right] = \frac{1}{1+\exp(-2ax^\top\theta^*)},\quad \mathbb{E}\left[c_x\right] = \tanh(ax^\top\theta^*)\\ \mathbb{V}\left[c_x\right] = 1 - \tanh^2(ax^\top\theta^*),\quad \mathbb{E}\left[t_x\right] = \begin{cases}\frac{a}{x^\top\theta^*}\tanh(ax^\top\theta^*) & \text{if } x^\top\theta^* \neq 0\\ a^2 & \text{if } x^\top\theta^* = 0\end{cases}.\end{aligned} \tag{4.1}$$

This choice probability matches that of the Bradley and Terry [263] model. If the learner relies solely on choices, then our bandit problem reduces to the transductive linear logistic bandit problem [258].

Figures 4.1b and 4.1c illustrate the roles of the parameters $x^\top\theta^*$ and a. First, the absolute drift (or the absolute utility difference), $|x^\top\theta^*|$, reflects the human's preference strength for the query x. Larger values indicate stronger preferences, leading to faster decisions and more consistent choices. Smaller values suggest weaker preferences, resulting in slower decisions and less consistent choices. Second, the barrier a represents the human's conservativeness in decision-making [264]. A more conservative human (higher a) requires more evidence to decide, resulting in slower but more consistent choices. In contrast, a less conservative human (lower a) decides faster but makes less consistent choices.

We adopt the common assumption that t_{nondec} is constant across all queries for a given human [45, 256] and further assume that t_{nondec} is known to the learner. This assumption enables the learner to perfectly recover t_x from the observed $t_{\text{RT},x}$. In section 4.4.2, we empirically show that even when t_{nondec} is unknown, its impact on the performance of our method that relies on decision times is negligible.

4.1.2 Learning Objective: Best-Arm Identification with a Fixed Budget

We focus on the fixed-budget best-arm identification problem [259, 260]. The learner is provided with a total interaction time budget $B > 0$, an arm space $\mathcal{Z}$, a query space $\mathcal{X}$, and a non-decision time t_{nondec}. Both the human's preference vector θ^* and the decision barrier a are unknown. In each episode $s \in \mathbb{N}$, the learner selects a query $x_s \in \mathcal{X}$, receives human feedback $(c_{x_s,s}, t_{x_s,s})$ generated by the dEZDM, and consumes $t_{\text{RT},x_s,s}$ time. When the cumulative interaction time exceeds the budget B at some episode S, i.e., $\sum_{s=1}^{S} t_{\text{RT},x_s,s} > B$, the learner must stop and recommend an arm $\widehat{z} \in \mathcal{Z}$. The goal is to recommend the unique best arm $z^* := \arg\max_{z \in \mathcal{Z}} z^\top \theta^*$, minimizing the error probability $\mathbb{P}\left[\widehat{z} \neq z^*\right]$.

To address this problem, we adopt the Generalized Successive Elimination (GSE) algorithm [257, 265, 266]. GSE divides the total budget B into multiple phases. In each phase, it strategically samples queries until the phase's budget is exhausted, collecting both human choices and decision times. It then estimates the preference vector θ^* and eliminates arms with low estimated utilities. Decision times play a key role in the estimation step by providing complementary information about preference strength, which can enable more accurate estimation of θ^* than choices alone. Next, in section 4.2, we introduce a novel estimator that combines decision times and choices to estimate θ^*. Then, in section 4.3, we discuss how this estimator is integrated into GSE to improve preference learning.

4.2 Utility estimation

This section addresses the problem of estimating human preference θ^* from a fixed dataset, denoted by $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{x \in \mathcal{X}_{\text{sample}}, i \in [n_x]}$. Here, $\mathcal{X}_{\text{sample}}$ denotes the set of queries in the dataset, n_x denotes the number of samples for each query $x \in \mathcal{X}_{\text{sample}}$, and $s_{x,i}$ denotes the episode when x is sampled for the i-th time. Samples from the same query x are i.i.d., while samples from different queries are independent. Section 4.2.1 introduces a new estimator, the "choice-decision-time estimator," which uses both choices and decision times, in contrast to the commonly used "choice-only estimator" that only uses choices [257, 258]. Sections 4.2.2 and 4.2.3 theoretically compares these estimators, analyzing both asymptotic and non-asymptotic performance and highlighting the advantages of incorporating decision times. Section 4.4.1 presents empirical results that validate our theoretical insights.

4.2.1 Choice-decision-time estimator and choice-only estimator

The choice-decision-time estimator is based on the following relationship between human utilities, choices, and decision times, derived from eq. (4.1):

$$\forall x \in \mathcal{X}\colon x^\top \frac{\theta^*}{a} = \frac{\mathbb{E}\left[c_x\right]}{\mathbb{E}\left[t_x\right]}. \tag{4.2}$$

Intuitively, when a human provides consistent choices (i.e., large $|\mathbb{E}[c_x]|$) and makes decisions quickly (i.e., small $\mathbb{E}[t_x]$), it implies a strong preference (i.e., large $|x^\top \theta^*|$). This relationship formulates the estimation of θ^* as a *linear regression* problem. Accordingly, the choice-decision-time estimator calculates the empirical means of both choices and decision times, aggregates the ratios across all sampled queries, and applies ordinary least squares (OLS) to estimate θ^*/a. Since the ranking of arm utilities based on θ^*/a is identical to that based on θ^*, estimating θ^*/a is sufficient for identifying the best arm. Formally, this estimate of θ^*/a, denoted by $\widehat{\theta}_{\text{CH,DT}}$, is given by:

$$\widehat{\theta}_{\text{CH,DT}} \coloneqq \left(\sum_{x \in \mathcal{X}_{\text{sample}}} n_x \, x x^\top \right)^{-1} \sum_{x \in \mathcal{X}_{\text{sample}}} n_x \, x \, \frac{\sum_{i=1}^{n_x} c_{x,s_{x,i}}}{\sum_{i=1}^{n_x} t_{x,s_{x,i}}}. \tag{4.3}$$

In contrast, the choice-only estimator is based on eq. (4.1), which shows that for each query $x \in \mathcal{X}$, the random variable $(c_x + 1)/2$ follows a Bernoulli distribution with mean $1/[1 + \exp(-x^\top \cdot 2a\theta^*)]$. Similar to the choice-decision-time estimator, the parameter $2a$ does not impact the ranking of arms, so estimating $2a\theta^*$ is sufficient for best-arm identification. This estimation is formulated as a *logistic regression* problem [257, 258], with MLE providing the following estimate of $2a\theta^*$, denoted by $\widehat{\theta}_{\text{CH}}$:

$$\widehat{\theta}_{\text{CH}} \coloneqq \operatorname*{arg\,max}_{\theta \in \mathbb{R}^d} \sum_{x \in \mathcal{X}_{\text{sample}}} \sum_{i=1}^{n_x} \log \mu(c_{x,s_{x,i}} \, x^\top \theta), \tag{4.4}$$

where $\mu(y) \coloneqq 1/[1 + \exp(-y)]$ is the standard logistic function. While this MLE lacks a closed-form solution, it can be efficiently solved using optimization methods like Newton's algorithm [267, 268].

4.2.2 Asymptotic normality of the two estimators

The choice-decision-time estimator from eq. (4.3) satisfies the following asymptotic normality result:

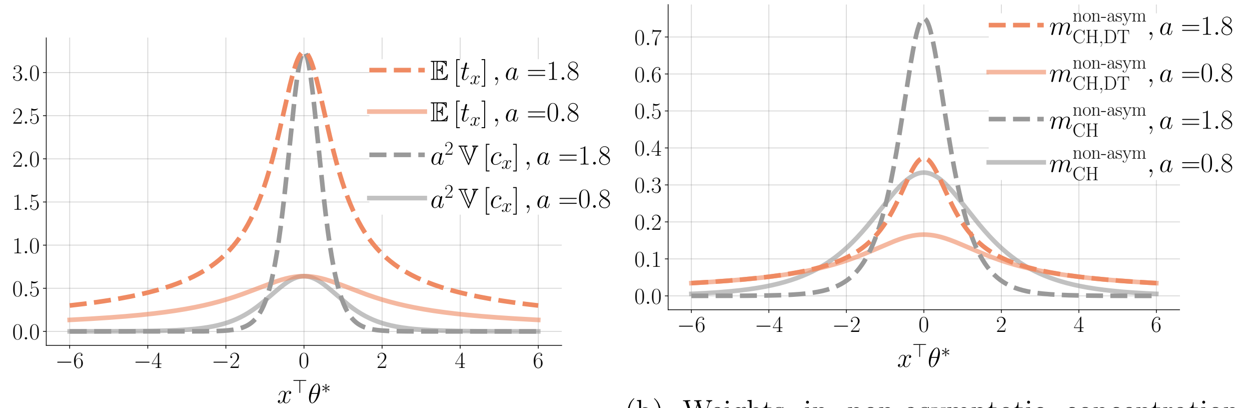
Theorem 4.2.1 (Asymptotic normality of $\widehat{\theta}_{\text{CH,DT}}$)**.** *Given a fixed i.i.d. dataset, denoted by* $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{i \in [n]}$ *for each* $x \in \mathcal{X}_{sample}$, *where* $\sum_{x \in \mathcal{X}_{sample}} x x^\top \succ 0$, *and assuming that the datasets for different* $x \in \mathcal{X}_{sample}$ *are independent, then, for any vector* $y \in \mathbb{R}^d$, *as* $n \to \infty$, *the following holds:*

$$\sqrt{n} \, y^\top \left(\widehat{\theta}_{CH,DT,n} - \theta^*/a \right) \xrightarrow{D} \mathcal{N}(0, \zeta^2/a^2).$$

Here, the asymptotic variance depends on a problem-specific constant, ζ^2, *with an upper bounded:*

$$\zeta^2 \leq \|y\|^2_{\left(\sum_{x \in \mathcal{X}_{sample}} \left[\min_{x' \in \mathcal{X}_{sample}} \mathbb{E}[t_{x'}]\right] \cdot x x^\top\right)^{-1}} .$$

The proof is provided in appendix A.2.2. The asymptotic variance upper bound shows that all sampled queries are weighted by a common factor $\min_{x' \in \mathcal{X}_{\text{sample}}} \mathbb{E}\left[t_{x'}\right]$, which is the smallest expected decision time among all the sampled queries in $\mathcal{X}_{\text{sample}}$. This weight represents the amount of information provided by each query's choices and decision times for utility estimation. A larger weight indicates that all queries in $\mathcal{X}_{\text{sample}}$ provide more information, leading to lower variance and better estimates.



(a) $\mathbb{E}[t_x]$ and $a^2\,\mathbb{V}[c_x]$ in asymptotic variances

(b) Weights in non-asymptotic concentration bounds

Figure 4.2: This figure illustrates key terms from our theoretical analyses, highlighting the different contributions of choices and decision times to utility estimation. These terms are functions of the utility difference $x^\top\theta^*$ and are plotted for two barrier values, a. (a) compares the weights $\mathbb{E}[t_x]$ and $a^2\,\mathbb{V}[c_x]$ in the asymptotic variances for the choice-decision-time estimator (orange, theorem 4.2.1) and the choice-only estimator (gray, theorem 4.2.2), respectively. This comparison shows that *decision times complement choices, particularly for queries with strong preferences.* (b) compares the weights in the non-asymptotic concentration bounds (theorems 4.2.3 and 4.2.4), showing similar trends, though these weights may not be optimal due to proof techniques.

In contrast, the choice-only estimator from eq. (4.4) has the following asymptotic normality result, as derived from Fahrmeir and Kaufmann [269, corollary 1]:

Theorem 4.2.2 (Asymptotic normality of $\widehat{\theta}_{\text{CH}}$)**.** *Given a fixed i.i.d. dataset, denoted by $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{i\in[n]}$ for each $x \in \mathcal{X}_{sample}$, where $\sum_{x\in\mathcal{X}_{sample}} xx^\top \succ 0$, and assuming that the datasets for different $x \in \mathcal{X}_{sample}$ are independent, then, for any vector $y \in \mathbb{R}^d$, as $n \to \infty$, the following holds:*

$$\sqrt{n}y^\top\left(\widehat{\theta}_{CH,n} - 2a\theta^*\right) \xrightarrow{D} \mathcal{N}\left(0, 4a^2\,\|y\|^2_{\left(\sum_{x\in\mathcal{X}_{sample}}\left[a^2\,\mathbb{V}[c_x]\right]\cdot xx^\top\right)^{-1}}\right).$$

This asymptotic variance shows that each sampled query $x \in \mathcal{X}_{\text{sample}}$ is weighted by its own factor $a^2\,\mathbb{V}[c_x]$, representing the amount of information the query's choices contribute to utility estimation. A larger weight indicates that the query contributes more information, leading to better estimates.

The weights in both theorems highlight the different contributions of choices and decision times to utility estimation. In the choice-only estimator (theorem 4.2.2), each query is weighted by $a^2\,\mathbb{V}[c_x]$, which depends on the utility difference $x^\top\theta^*$ for a fixed barrier a. As shown by the gray curves in fig. 4.2a, this weight quickly decays to zero as preferences become stronger (i.e., as $|x^\top\theta^*|$ increases). This indicates that *choices from queries with strong preferences provide little information.* Intuitively, when preferences are strong, humans consistently select the same option, making it hard to distinguish whether their preference is moderately or very strong. As a result, choices from such queries contribute minimally to utility estimation. This intuition aligns with the dressing robot example at the beginning of this chapter.

For the choice-decision-time estimator (theorem 4.2.1), queries are weighted by the minimum expected decision time over $\mathcal{X}_{\text{sample}}$, i.e., $\min_{x' \in \mathcal{X}_{\text{sample}}} \mathbb{E}\left[t_{x'}\right]$, which depends on both $\mathcal{X}_{\text{sample}}$ and $\mathbb{E}\left[t_x\right]$. To better understand this weight, we first plot $\mathbb{E}\left[t_x\right]$ without the 'min' operator as the orange curves in fig. 4.2a. Comparing the orange and gray curves shows that $\mathbb{E}\left[t_x\right]$ is generally larger than the choice-only weight, $a^2\,\mathbb{V}\left[c_x\right]$. The actual weight in the choice-decision-time estimator, which is the minimum expected decision time across sampled queries, is less than or equal to the orange curve but is likely still higher than the choice-only weight, especially for queries with strong preferences. This suggests that *when preferences are strong, decision times complement choices by capturing preference strength, leading to improved estimation.*

When queries have weak preferences, the choice-decision-time weight may be lower than the choice-only weight. However, since the choice-decision-time weight represents only an upper bound on the asymptotic variance (theorem 4.2.1), no definitive conclusions can be drawn from the theory alone. Empirically, as shown in section 4.4.1, decision times add little value but do not degrade performance.

As the barrier a increases, the choice-decision-time weight rises. In contrast, the choice-only weight increases for queries with weak preferences, but this increase is concentrated in a narrower region, with weights decreasing elsewhere. Intuitively, a higher barrier reflects greater conservativeness in human decision-making, leading to longer decision times and more consistent choices (fig. 4.1). As a result, more queries exhibit strong preferences, making choices from these queries less informative.

4.2.3 Non-asymptotic concentration of the two estimators for utility difference estimation

In this section, we focus on the simpler problem of estimating the utility difference for a single query, without aggregating data from multiple queries. Comparing the non-asymptotic concentration bounds of both estimators, in this case, provides insights similar to those discussed in section 4.2.2. Extending this non-asymptotic analysis to the full estimation of the preference vector θ^* is left for future work.

Given a query $x \in \mathcal{X}$, the task is to estimate the utility difference $u_x \coloneqq x^\top \theta^*$ using the fixed i.i.d. dataset $\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\}_{i\in[n_x]}$. Applying the choice-decision-time estimator from eq. (4.3), we get the following estimate (for details, see appendix A.2.3), which estimates u_x/a rather than u_x:

$$\widehat{u}_{x,\text{CH,DT}} \coloneqq \frac{\sum_{i=1}^{n_x} c_{x,s_{x,i}}}{\sum_{i=1}^{n_x} t_{x,s_{x,i}}}. \tag{4.5}$$

In contrast, applying the choice-only estimator from eq. (4.4), we get the following estimate (for details, see appendix A.2.3), which estimates $2au_x$ rather than u_x:

$$\widehat{u}_{x,\text{CH}} \coloneqq \mu^{-1}\left(\frac{1}{n_x}\sum_{i=1}^{n_x}\frac{c_{x,s_{x,i}}+1}{2}\right), \tag{4.6}$$

where $(c_{x,s_{x,i}}+1)/2$ is the binary choice coded as 0 or 1, and $\mu^{-1}(p) \coloneqq \log\left(p/(1-p)\right)$ is the logit function (inverse of μ introduced in eq. (4.4)).

Notably, the choice-only estimator in eq. (4.6) aligns with the EZ-diffusion model's drift estimator [43, eq. (5)]. Moreover, the estimators in Xiang Chiong et al. [270, eq. (6)] and Berlinghieri et al. [251, eq. (7)] combine elements of both estimators from eqs. (4.5) and (4.6). In section 4.4.2, we demonstrate that both estimators from Wagenmakers, Van Der Maas, and Grasman [43, eq. (5)] and Xiang Chiong et al. [270, eq. (6)] are outperformed by our proposed estimator in eq. (4.3) for the full bandit problem.

Assuming the utility difference $u_x \neq 0$, the choice-decision-time estimator in eq. (4.5) satisfies the following non-asymptotic concentration bound, proven in appendix A.2.3:

Theorem 4.2.3 (Non-asymptotic concentration of $\widehat{u}_{x,\mathrm{CH,DT}}$)**.** *For each query $x \in \mathcal{X}$ with $u_x \neq 0$, given a fixed i.i.d. dataset, denoted by $\left\{\left(c_{x,s_{x,i}}, t_{x,s_{x,i}}\right)\right\}_{i\in[n_x]}$, for any $\epsilon > 0$ satisfying $\epsilon \leq \min\left\{|u_x|/(\sqrt{2}a), \left(1+\sqrt{2}\right)a|u_x|/\mathbb{E}\left[t_x\right]\right\}$, the following holds:*

$$\mathbb{P}\left(\left|\widehat{u}_{x,CH,DT} - \frac{u_x}{a}\right| > \epsilon\right) \leq 4\exp\left(-\left[m^{\textit{non-asym}}_{CH,DT}\left(x^\top\theta^*\right)\right]^2 n_x \left[\epsilon \cdot a\right]^2\right),$$

where $m^{\textit{non-asym}}_{CH,DT}\left(x^\top\theta^*\right) \coloneqq \mathbb{E}\left[t_x\right] \,/\, \left[\left(2+2\sqrt{2}\right)\,a\right]$.

In contrast, the choice-only estimator in eq. (4.6) has the following non-asymptotic concentration result, adapted from Jun et al. [258, theorem 5][1]:

Theorem 4.2.4 (Non-asymptotic concentration of $\widehat{u}_{x,\mathrm{CH}}$)**.** *For each query $x \in \mathcal{X}$, given a fixed i.i.d. dataset, denoted by $\left\{c_{x,s_{x,i}}\right\}_{i\in[n_x]}$, for any positive $\epsilon < 0.3$, if*

$$n_x \geq \frac{1}{\dot{\mu}(2au_x)} \cdot \max\left\{\frac{3^2\log(6e)}{\epsilon^2}, \frac{64\log(3)}{1-\epsilon^2/0.3^2}\right\},$$

then the following holds:

$$\mathbb{P}\left(\left|\widehat{u}_{x,CH} - 2au_x\right| > \epsilon\right) \leq 6\exp\left(-\left[m^{\textit{non-asym}}_{CH}\left(x^\top\theta^*\right)\right]^2 n_x \left[\epsilon/(2a)\right]^2\right),$$

where $m^{\textit{non-asym}}_{CH}\left(x^\top\theta^*\right) \coloneqq a\sqrt{\mathbb{V}\left[c_x\right]}\,/\,2.4$.

The weights $m^{\text{non-asym}}_{\text{CH,DT}}(\cdot)$ and $m^{\text{non-asym}}_{\text{CH}}(\cdot)$ from theorems 4.2.3 and 4.2.4, respectively, are functions of the utility difference $x^\top\theta^*$ for a fixed barrier a. These weights determine how quickly estimation errors decay as the dataset size n_x grows, with larger weights indicating faster error reduction. While these weights may not be optimal due to proof techniques, they highlight the distinct contributions of choices and decision times, consistent with our asymptotic analysis in section 4.2.2. Figure 4.2b compares the weights for the choice-decision-time estimator (orange, $m^{\text{non-asym}}_{\text{CH,DT}}(\cdot)$) and the choice-only estimator (gray, $m^{\text{non-asym}}_{\text{CH}}(\cdot)$). For strong preferences, the choice-only weights quickly decay to zero, while the choice-decision-time weights remain relatively large. This supports our key insight that decision times complement choices and improve estimation for queries with strong preferences.

[1] In Jun et al. [258, theorem 5], we let $x_1 = \cdots = x_t = 1$ and $t_{\text{eff}} = d = 1$.

In summary, both asymptotic (section 4.2.2) and non-asymptotic (section 4.2.3) analyses demonstrate that the choice-decision-time estimator extracts more information from queries with strong preferences. This finding aligns with prior empirical work [45] and is further supported by our results in section 4.4.1.

In fixed-budget best-arm identification, our choice-decision-time estimator's ability to extract more information from queries with strong preferences is especially valuable. Bandit learners, such as GSE [257], strategically sample queries, update estimates of θ^*, and eliminate lower-utility arms. With the choice-only estimator, learners struggle to extract information from queries with strong preferences. To resolve this, one approach is to selectively sample queries with weak preferences, but this has two drawbacks. First, queries with weak preferences take longer to answer (i.e., require more resources), potentially lowering the 'bang per buck' (information per resource) [271]. Second, since θ^* is unknown in advance, learners cannot reliably target queries with weak preferences. In contrast, with our choice-decision-time estimator, learners leverage decision times to gain more information from queries with strong preferences, improving bandit learning performance. We integrate both estimators into bandit learning in section 4.3 and evaluate their performance in section 4.4.

4.3 Interactive learning algorithm

We introduce the Generalized Successive Elimination (GSE) algorithm [257, 265, 266] for fixed-budget best-arm identification in preference-based linear bandits, and outline the key options for each GSE component, which we empirically compare in section 4.4.

The pseudo-code for GSE is shown in algorithm 1. The algorithm uses a hyperparameter η to control the number of phases, the budget per phase, and the number of arms eliminated in each phase. GSE divides the total budget B evenly across phases and reserves a buffer, sized by another hyperparameter B_{buff}, to prevent overspending in any phase (line 4). In each phase, GSE computes an experimental design λ, a probability distribution over the query space, to guide query sampling. We consider two designs: the transductive design [253], λ_{trans} (line 5), and the weak-preference design [258], λ_{weak} (line 6). Both designs minimize the worst-case variance of utility differences between surviving arms. The transductive design weights all queries equally, whereas the weak-preference design prioritizes queries with weak preferences to counter the choice-only estimator's difficulty in extracting information from queries with strong preferences (section 4.2). Since θ^* is unknown, the weak-preference design identifies queries with weak preferences based on the previous phase's estimate $\widehat{\theta}_{\text{CH}}$. Then, GSE samples queries based on the design (line 7) and, after exhausting the phase's budget, estimates θ^* using either the choice-decision-time estimator $\widehat{\theta}_{\text{CH,DT}}$ (line 8) or the choice-only estimator $\widehat{\theta}_{\text{CH}}$ (line 9). It then eliminates arms with low estimated utilities (line 10). This process repeats until only one arm remains, which GSE recommends as the best arm (line 12).

The key difference between algorithm 1 and previous GSE algorithms [257, 265, 266] is that our setting involves queries with random response times, unknown to the learner. Previous work assumes fixed resource consumption per query and uses deterministic rounding methods [253, 257] to pre-allocate queries. This approach does not handle random resource usage. Instead, we adopt a random sampling procedure [272, 273] in line 7 to allocate queries based on the design. Random resource usage also requires tuning the elimination parameter η, to balance data collection and arm elimination, and the buffer size B_{buff}, to prevent overspending. In our empirical study (section 4.4.2), we manually tune both parameters. Further theoretical analysis is needed to better understand and optimize them.

Algorithm 1 Generalized Successive Elimination (GSE) [257]

```
**Input:** Arm space $\mathcal{Z}$, query space $\mathcal{X}$, non-decision time $t_{\text{nondec}}$, and total budget $B$.
**Hyperparameters:** Elimination parameter $\eta$ and buffer size $B_{\text{buff}}$.
**Initialization:** $\mathcal{Z}_1 \leftarrow \mathcal{Z}$.
**for** each phase $k = 1, \dots, K := \lceil \log_\eta |\mathcal{Z}| \rceil$ with the budget $B_k := B/K - B_{\text{buff}}$ **do**
    Design 1. $\lambda_k := \lambda_{\text{trans},k} \leftarrow \arg\min_{\lambda \in \blacktriangle^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|^2_{\left(\sum_{x \in \mathcal{X}} \lambda_x x x^\top\right)^{-1}}$.
    Design 2. $\lambda_k := \lambda_{\text{weak},k} \leftarrow \arg\min_{\lambda \in \blacktriangle^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|^2_{\left(\sum_{x \in \mathcal{X}} \dot{\mu}(x^\top \widehat{\theta}_{k-1}) \lambda_x x x^\top\right)^{-1}}$.
    Sample queries $x_j \sim \lambda_k$ and stop at $J_k$ if $\sum_{j=1}^{J_k - 1} t_{\text{RT},x_j,j} \leq B_k$ and $\sum_{j=1}^{J_k} t_{\text{RT},x_j,j} > B_k$.
    Estimate 1. $\widehat{\theta}_k := \widehat{\theta}_{\text{CH,DT},k} \leftarrow$ apply eq. (4.3) to all the $J_k$ samples.
    Estimate 2. $\widehat{\theta}_k := \widehat{\theta}_{\text{CH},k} \leftarrow$ apply eq. (4.4) to all the $J_k$ samples.
    Update $\mathcal{Z}_{k+1} \leftarrow$ Top-$\left\lceil \frac{|\mathcal{Z}_k|}{\eta} \right\rceil$ arms in $\mathcal{Z}_k$, ranked by the estimated utility $z^\top \widehat{\theta}_k$.
**end for**
**Output:** the single one $\widehat{z} \in \mathcal{Z}_{K+1}$.
```

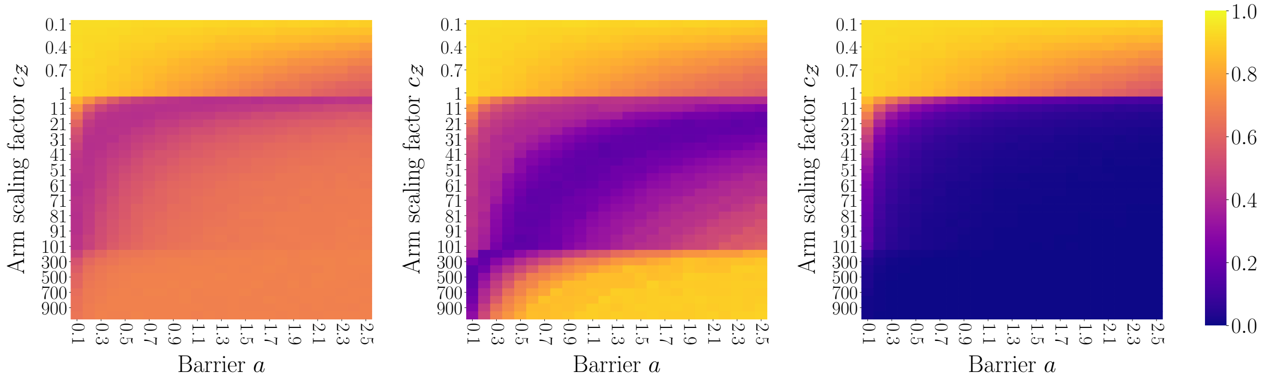
4.4 Empirical results

This section empirically compares the GSE variations introduced in section 4.3: (1) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$: Transductive design with choice-decision-time estimator. (2) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$: Transductive design with choice-only estimator. (3) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$: Weak-preference design with choice-only estimator.

4.4.1 Estimation performance on synthetic data

We evaluate the estimation performance of the GSE variations on the "sphere" synthetic problem, a standard linear bandit problem in the literature [272, 274, 275]. Details are provided in appendix A.3.1.

Estimation performance, as discussed in section 4.2, depends on the utility difference $x^\top \theta^*$ and the barrier a. We vary a over a range of values commonly used in psychology [45, 246]. To examine how preference strength impacts estimation, we scale each arm z to $c_{\mathcal{Z}} \cdot z$, effectively scaling each utility difference $x^\top \theta^*$ to $c_{\mathcal{Z}} \cdot x^\top \theta^*$. Small $c_{\mathcal{Z}}$ values correspond to problems with weak preferences, while large values correspond to strong preferences. For each $(c_{\mathcal{Z}}, a)$ pair, the system generates 100 random problem instances and runs 100 repeated simulations per instance. In each simulation, the GSE variations sample 50 queries, ignoring the response time budget, and compute $\widehat{\theta}$. Performance is evaluated by $\mathbb{P}[\arg\max_{z \in \mathcal{Z}} z^\top \widehat{\theta} \neq z^*]$, which reflects the best-arm identification goal defined in section 4.1. To isolate the effect of estimation, we



(a) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$

(b) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$

(c) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$

Figure 4.3: Three heatmaps show estimation error probabilities, $\mathbb{P}[\arg\max_{z \in \mathcal{Z}} z^\top \widehat{\theta} \neq z^*]$, for three GSE variations, shown as functions of the arm scaling factor $c_{\mathcal{Z}}$ and barrier a. Darker colors indicate better estimation. (a) The choice-only estimator $\widehat{\theta}_{\text{CH}}$ with the transductive design λ_{trans} struggles as $c_{\mathcal{Z}}$ increases (i.e., preferences become stronger), highlighting that choices from queries with strong preferences provide limited information. (b) The weak-preference design λ_{weak} improves (a) by sampling queries with weak preferences but assumes perfect knowledge of θ^* and equal resource consumption across queries. (c) The choice-decision-time estimator $\widehat{\theta}_{\text{CH,DT}}$ with λ_{trans} outperforms both choice-only methods in (a) and (b), showing that decision times complement choices and improve estimation, especially for strong preferences.

allow λ_{weak} access to the true θ^*, enabling it to perfectly compute the terms $\dot{\mu}(x^\top \theta^*)$ used in line 6 of algorithm 1.

As shown in fig. 4.3a, fixing the barrier a and examining the vertical line, as $c_{\mathcal{Z}}$ increases and preferences become stronger, the performance of the choice-only estimator with the transductive design first improves and then declines. The initial improvement arises because larger $c_{\mathcal{Z}}$ increases utility differences between the best arm and others, theoretically simplifying best-arm identification. The subsequent decline, highlighted by the dark curved band, supports our insight from section 4.2 that choices from queries with strong preferences provide limited information. Fixing $c_{\mathcal{Z}}$ and examining the horizontal line, performance first improves and then declines. This trend aligns with fig. 4.2a and section 4.2.2, where higher barriers a increase the choice-only weights for queries with weak preferences, initially improving performance. However, as a grows, fewer queries exhibit increased weights, while most queries' weights decrease, leading to a later performance drop.

In Figure 4.3b, for moderate $c_{\mathcal{Z}}$, the choice-only estimator with the weak-preference design outperforms the transductive design (fig. 4.3a), demonstrating that focusing on queries with weak preferences improves estimation. However, as $c_{\mathcal{Z}}$ becomes too large, performance declines because many $\dot{\mu}(x^\top \theta^*)$ in line 6 of algorithm 1 approach zero, preventing informative queries from being sampled. This advantage of the weak-preference design assumes perfect knowledge of θ^* and equal resource consumption across queries. In practice, where θ^* is unknown and weak-preference queries require longer response times, the transductive design performs better, as shown in section 4.4.2.

Figure 4.3c shows that the choice-decision-time estimator consistently outperforms the choice-only estimators under both the transductive and weak-preference designs, particularly for strong preferences. This suggests that for queries with strong preferences, decision times complement choices and improve estimation, confirming our theoretical insights from section 4.2, while for queries with weak preferences, decision times add little value but do not degrade performance. The performance also improves with a higher barrier a, supporting the insights conveyed by fig. 4.2a and section 4.2.2.

4.4.2 Fixed-budget best-arm identification performance on real datasets

This section compares the bandit performance of six GSE variations. The first three are as previously defined at the beginning of section 4.4: $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, and $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$.

The 4th GSE variation, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, evaluates the performance of the choice-decision-time estimator when the non-decision time t_{nondec} is unknown. The estimator, $\widehat{\theta}_{\text{CH},\mathbb{RT}}$, is identical to the original choice-decision-time estimator from eq. (4.3), but with response times used in place of decision times.

The 5th GSE variation, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, is based on Wagenmakers, Van Der Maas, and Grasman [43, eq. (5)], which states that $x^\top \cdot (2a\theta^*) = \mu^{-1}(\mathbb{P}[c_x = 1])$, where $\mu^{-1}(p) := \log\left(p/\left(1-p\right)\right)$. By incorporating our linear utility structure, we obtain the following choice-only estimator $\widehat{\theta}_{\text{CH,logit}}$:

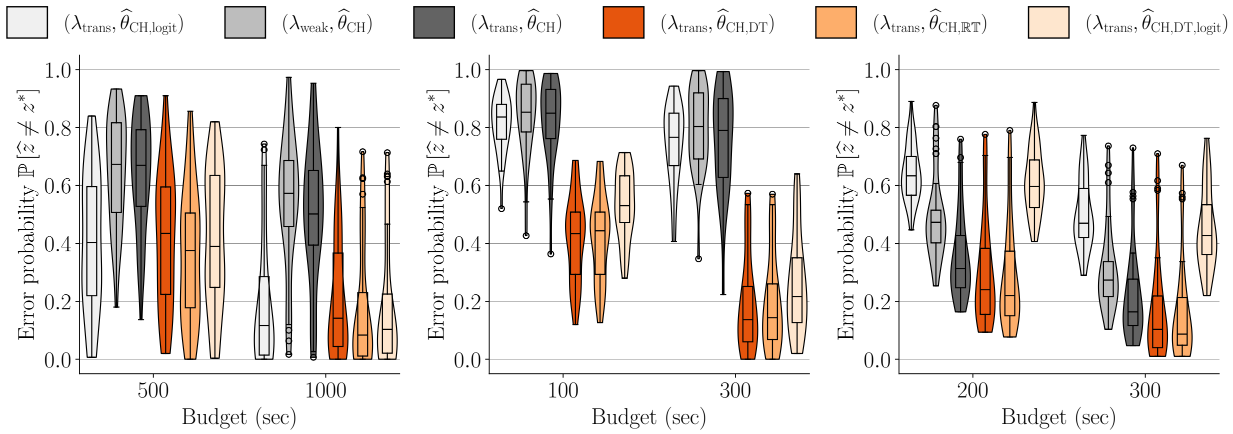
$$\widehat{\theta}_{\text{CH,logit}} := \left(\sum_{x\in\mathcal{X}_{\text{sample}}} n_x\, xx^\top\right)^{-1} \sum_{x\in\mathcal{X}_{\text{sample}}} n_x\, x\cdot\mu^{-1}\left(\widehat{\mathfrak{C}}_x\right),$$

where $\widehat{\mathfrak{C}}_x := \frac{1}{n_x}\sum_{i=1}^{n_x}\frac{1}{2}\left(c_{x,s_{x,i}}+1\right)$ is the empirical mean of the binary choices coded as 0 or 1.

The 6th GSE variation, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$, is based on Xiang Chiong et al. [270, eq. (6)], which states that $x^\top\theta^* = \operatorname{sgn}\left(c_x\right)\sqrt{\mathbb{E}\left[c_x\right]/\mathbb{E}\left[t_x\right]\cdot 0.5\,\mu^{-1}\left(\mathbb{P}\left[c_x=1\right]\right)}$. This identity forms the foundation of the estimator in Berlinghieri et al. [251, eq. (7)]. By incorporating our linear utility structure, we obtain the following choice-decision-time estimator $\widehat{\theta}_{\text{CH,DT,logit}}$:

$$\widehat{\theta}_{\text{CH,DT,logit}} := \left(\sum_{x\in\mathcal{X}_{\text{sample}}} n_x\, xx^\top\right)^{-1} \sum_{x\in\mathcal{X}_{\text{sample}}} n_x\, x\cdot\operatorname{sgn}\left(c_x\right)\sqrt{\frac{\mathbb{E}\left[c_x\right]}{\mathbb{E}\left[t_x\right]}\cdot\frac{1}{2}\,\mu^{-1}\left(\widehat{\mathfrak{C}}_x\right)}.$$

We evaluate six GSE variations on bandit instances constructed from three real-world datasets of human choices and response times. Each dataset includes multiple participants. For each participant, we estimated dEZDM parameters, built a bandit instance, and simulated the GSE variations to assess performance. Details on experimental procedures are provided in appendix A.3. Key results for the three domains are shown in fig. 4.4, with full results in appendix A.3. First, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$ consistently outperforms $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, demonstrating the benefit of incorporating decision times. Second, both of these variations



(a) Food-risk dataset [44]

(b) Snack dataset [45]

(c) Snack dataset [46]

Figure 4.4: This figure shows violin plots (with overlaid box plots) for datasets (a), (b), and (c), showing the distribution of best-arm identification error probabilities, $\mathbb{P}[\widehat{z} \neq z^*]$, for all bandit instances across six GSE variations and two budgets. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within 1.5× the interquartile range. Flier points indicate outliers beyond the whiskers.

outperform $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, as discussed in section 4.4.1. Third, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$ performs similarly to $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, suggesting that not knowing the non-decision time has minimal impact. Finally, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$ [43] and $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$ [270] do not perform as consistently well as $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, highlighting the effectiveness of our proposed choice-decision-time estimator (eq. (4.3)).

4.5 Conclusion

This chapter presented the first contribution of the thesis: a cognitive model-based algorithm that uses both human binary choices and response times to reduce uncertainty in preference learning. By integrating a well-established psychological model of decision-making, the drift diffusion model, into a linear bandit algorithm, we showed that response times provide additional information about the strength of user preferences.

This insight addresses a key challenge in human-robot interaction: how to make learning from humans more sample-efficient, particularly when user feedback is scarce or expensive. Empirical results in simulated recommender systems confirmed that incorporating response times significantly reduces misidentification of preferred options compared to choice-only methods.

In the broader context of this thesis, this work demonstrates that implicit feedback, which is naturally available and costless to the user, can be systematically leveraged to reduce the robot's uncertainty about the human preferences. This finding contributes to answering the central question of this thesis: *How should a robot behave when it is uncertain about the human?* It does so by showing that uncertainty can be reduced more efficiently through richer models of human cognition. While this chapter focuses on a simplified setting with static options as discussed at the beginning of this chapter, it lays the theoretical and computational groundwork for future extensions to robotics, where feedback will be gathered over dynamic, time-extended robot behaviors. In such settings, understanding response time will be even more crucial for efficient personalization under uncertainty (see section 7.2 for more details).

Chapter 5

Representing and Respecting Uncertainty for Robust State Estimation

"It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so."

—Mark Twain

State estimation is critical for human-robot interaction, where human physical states may be partially observed due to occlusions [276], and latent mental states can influence the interaction [130, 277]. To safely interact with humans, robots typically estimate such states using models [278], including a dynamics model that describes how the human arm moves during interaction, and an observation model that captures how that motion is perceived through the robot's sensors.

Due to the complexity of the human body and behavior, manually specifying these models is difficult. Instead, they are typically learned from data [56]. Because user-specific data is often limited, the resulting learned models can be inaccurate. This introduces "epistemic uncertainty", which is the uncertainty about the parameters or structure of the models themselves, arising from limited or imperfect training data [279]. At the same time, human motion is inherently stochastic, and sensor measurements are noisy, introducing "aleatoric uncertainty," which is the uncertainty due to intrinsic stochasticity in the system [279].

Both types of uncertainty must be accounted for to ensure reliable state estimation and safe robot behavior. If these uncertainties are ignored, the robot may become overconfident, potentially misestimating the human state and leading motion planning to generate unsafe or ineffective behavior. For example, in robot-assisted dressing (see fig. 5.1), uncertainty can cause the robot to believe the arm is in the wrong place, leading to overly aggressive actions.

To address this challenge, this chapter focuses on how robots can explicitly represent and respect both epistemic and aleatoric uncertainty in state estimation, especially when the underlying models are learned. Our approach builds on the notion of *consistency*, a key property in estimation: the estimator's belief should never be more confident than justified by the available information.

In the broader context of this thesis, this chapter contributes to the answer of how robots should estimate hidden human states when they are uncertain about human behavior: by explicitly modeling what the robot does not know

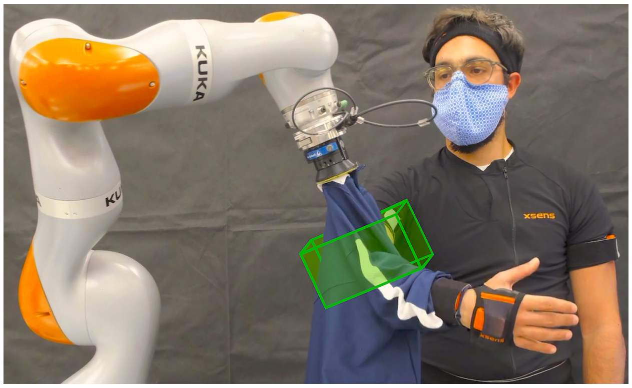


Figure 5.1: In a robot-assisted dressing scenario, we deployed our set-based estimator, GP-ZKF, to estimate the visually occluded human elbow position [276]. With human dynamics and observation models learned via Gaussian Process regression, GP-ZKF constructs zonotopic state estimates (illustrated with the green box) based on the force measurements at the robot end effector. By handling epistemic uncertainties in the learned models, GP-ZKF guarantees probabilistic consistency, i.e., the true human elbow positions are contained within the zonotopes across all time steps, with a high probability.

and estimating conservatively, the robot can maintain reliable awareness of the human's state, even under occlusions and model errors.

In the *stochastic estimation* paradigm, such as the Extended Kalman Filter (EKF), a consistent estimate is defined as an unbiased point estimate together with a covariance matching the actual estimation error [280]. However, under nonlinear or learned models, this definition can break down due to accumulated linearization and modeling errors. In SLAM, the inconsistency of EKF-based approaches such as GP-EKF [281] has been widely studied in terms of linearization error [282, 283] and state unobservability [284, 285]. Prior work has attempted to address these issues by constraining the Jacobians [282, 284] and defining local frames to handle nonlinear errors [285].

Instead, we adopt a *set-based estimation* paradigm, which constructs sets, rather than points, as state estimates. In this view, consistency means that the true state lies within the estimated set [222]. Set-based methods allow principled reasoning about both aleatoric and epistemic uncertainty. Prior literature has focused on settings where models are known, i.e., epistemic uncertainty can be ignored, and provides guarantees under aleatoric uncertainty [222, 223]. In contrast, we address the more realistic setting where both the dynamics and observation models are nonlinear and learned, and thus uncertainty arises from both sources.

We introduce the Gaussian Process-Zonotopic Kalman Filter (GP-ZKF), a set-based estimation algorithm that provides a probabilistic consistency guarantee under learned dynamics and observation models. GP-ZKF learns both models using Gaussian Process (GP) regression, and uses their confidence intervals [52] to calibrate epistemic uncertainty. This extends Combastel [286] that assumed bounded epistemic uncertainties in the linear parameter-varying enclosures of the nonlinear models.

Similar to set-based estimators [222, 223], but specifically for scenarios with learned models, our approach recursively produces set-based estimates that are represented as zonotopes (a special type of polytope). These zonotopes are designed to respect both epistemic and aleatoric uncertainties and are guaranteed to contain the true states across all time steps, with high probability, rendering GP-ZKF consistent when both nonlinear models are learned.

We also formally connect our *set-based* estimator, GP-ZKF, with the corresponding *stochastic* estimator, GP-EKF [281], and prove that GP-ZKF reduces to GP-EKF if GP-ZKF omits linearization errors and aleatoric, and simplifies epistemic uncertainties. This theoretical connection under nonlinear and learned models extends Combastel [287], which connects set-based and stochastic estimators under linear and known models.

Our contributions are:

- We propose GP-ZKF, a set-based state estimator with probabilistic consistency guarantees under both epistemic and aleatoric uncertainty, for the case where both dynamics and observation models are nonlinear and learned.
- We formally relate GP-ZKF to its stochastic counterpart, GP-EKF [281], and analyze their equivalence under simplified uncertainty assumptions.

We evaluate GP-ZKF in both a simulated pendulum environment and a real-world robot-assisted dressing scenario. Our results show that GP-ZKF provides not only more consistent, but also less conservative state estimates than the stochastic baselines (GP-EKF, GP-UKF, and GP-PF [281]). *To the best of our knowledge, this is the first method to offer probabilistic consistency guarantees for state estimation with learned nonlinear models.*

Section 5.1 defines the system setup, and section 5.2 introduces probabilistic consistency. Section 5.3 presents our algorithm, section 5.4 contains theoretical results, and section 5.5 provides empirical evaluation.

5.1 System Formulation

We model the human-robot system as a discrete-time dynamical system with finite-horizon $T \in \mathbb{N}$, nonlinear dynamics and observation functions, and additive noises. Formally, for $t = 1, \ldots, T$, the system can be described as follows:

$$\begin{aligned} x_t &= d(x_{t-1}, u_{t-1}, w_{t-1}) \\ &= f(x_{t-1}, u_{t-1}) + g(x_{t-1}, u_{t-1}) + w_{t-1}, \qquad (5.1) \\ y_t &= o(x_t, u_t, v_t) = h(x_t, u_t) + v_t. \qquad (5.2) \end{aligned}$$

Here, $x_t \in \mathbb{R}^{n_x}$ represents the hidden system state at time t, for instance, the position of the human elbow. The robot applies a control signal $u_t \in \mathfrak{U} \subset \mathbb{R}^{n_u}$ and receives a sensor measurement $y_t \in \mathbb{R}^{n_y}$. The system is affected by process noise $w_t \in \mathbb{R}^{n_x}$ and measurement noise $v_t \in \mathbb{R}^{n_y}$.

The dynamics function $d(\cdot)$ models how the hidden human state evolves over time, based on the previous state x_{t-1} and robot control u_{t-1}. This function is decomposed into two parts: (1) $f(\cdot)$ is a known nominal dynamics model (e.g., a parametric physics model). (2) $g(\cdot)$ is an unknown residual dynamics model to be learned from data.

The observation function $o(\cdot)$ describes how the current hidden state and control are perceived through the robot's sensors. It is composed of: (1) $h(\cdot)$, an unknown observation model to be learned from data, and (2) v_t, the sensor noise.

To make these Gaussian noise assumptions compatible with set-based estimation, we bound the noise terms within boxes using concentration inequalities:

Remark 5.1.1. By applying lemma 3.4.1, we construct boxes, denoted by $\mathcal{W} \subset \mathbb{R}^{n_x}$ and $\mathcal{V} \subset \mathbb{R}^{n_y}$, that bound the process noise, w, and the measurement noise, v, respectively. Formally, for a given confidence level $\delta^w \in (0, 1)$, with a probability at least $(1 - \delta^w)$, jointly for each time step $t = 1, \ldots, T$, we have that $w_{t-1} \in \mathcal{W}$. Similarly, for a given confidence level $\delta^v \in (0, 1)$, with a probability at least $(1 - \delta^v)$, jointly for each time step $t = 1, \ldots, T$, we have that $v_t \in \mathcal{V}$.

This system formulation captures both sources of uncertainty introduced in the beginning of this chapter: (1) epistemic uncertainty, arising from the unknown components $g(\cdot)$ and $h(\cdot)$ in the learned models, and (2) aleatoric uncertainty, arising from the inherent stochasticity of the system and noisy sensor readings. Our goal in the following sections is to develop a state estimator that respects both types of uncertainty and provides conservative, reliable estimates of the hidden human state.

5.2 Problem Definition

Our goal is to develop a set-based state estimator that can produce *consistent* estimates. Let $\widehat{\mathcal{X}}_t \subset \mathbb{R}^{n_x}$ denote a set-based state estimate produced by our algorithm at time t. By assuming that the controls are given, the estimation process, at time $t = 1, \ldots, T$ can be represented as a recursive function, $\widehat{\mathcal{X}}_t = E(\widehat{\mathcal{X}}_{t-1}, u_{t-1}, u_t, y_t)$.

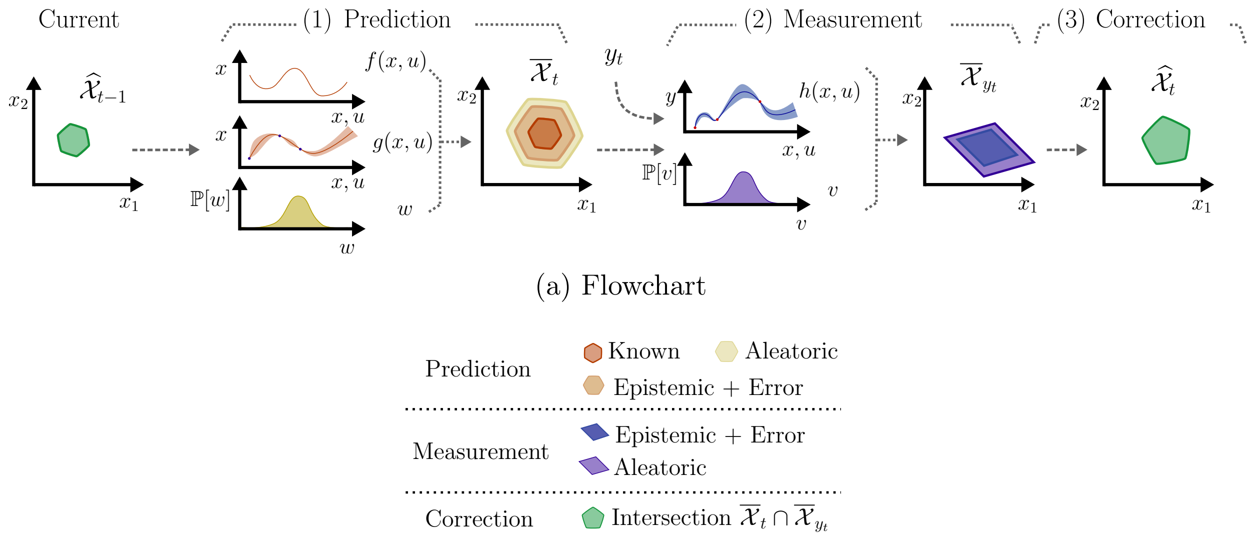
When the dynamics and observation models, $d(\cdot)$ and $o(\cdot)$ (respectively), are known, prior arts in set-based state estimation [222, 223, 288] have achieved *strict* consistency guarantees. In contrast, we focus on a scenario where both $d(\cdot)$ and $o(\cdot)$ are learned with a limited amount of data; hence, we relax the strict consistency and focus on *probabilistic* consistency, or δ-consistency, defined as follows:

Definition 5.2.1 (δ-consistency). *Given $\delta \in (0, 1)$; an initial set-based estimate, $\widehat{\mathcal{X}}_0 \subset \mathbb{R}^{n_x}$ such that $x_0 \in \widehat{\mathcal{X}}_0$; a sequence of controls, $\{u_t\}_{t=0}^T \subset \mathfrak{U}$; and a sequence of measurements, $\{y_t\}_{t=1}^T \subset \mathbb{R}^{n_y}$; Then, a state estimator is δ-consistent if the sequence of estimates, $\{\widehat{\mathcal{X}}_t\}_{t=1}^T$, computed via E, satisfies:*

$$\mathbb{P}\left[\forall t = 1 \ldots T \colon x_t \in \widehat{\mathcal{X}}_t\right] \geq 1 - \delta.$$

This definition states that δ-consistent estimators are able to guarantee that, with high probability, jointly for each time step within a finite time horizon, the set-based estimate contains the true state. Note that: (1) Definition 5.2.1 implies that the high-probability consistency guarantee holds *jointly* throughout the finite time horizon, rather than *per time-step* in the form of $\forall t = 1 \ldots T \colon \Pr[x_t \in \widehat{\mathcal{X}}_t] \geq 1 - \delta$ [289]. (2) Definition 5.2.1 indicates a filtering (rather than smoothing) problem, as future measurements are never used to estimate past or current states.

Problem Statement. *Design a set-based estimator that guarantees δ-consistency under the epistemic uncertainties in the learned models, $g(\cdot)$ and $h(\cdot)$, and the aleatoric uncertainties in the noises, w and v.*



(a) Flowchart

(b) Legend

Figure 5.2: A flowchart illustrating the three phases of the set-based estimator, GP-ZKF, at time $t = 1, \ldots, T$: (1) *Prediction*: Given the previous zonotopic estimate $\widehat{\mathcal{X}}_{t-1}$ and control u_{t-1} (omitted in the figure), GP-ZKF predicts a dynamics-consistent zonotope $\overline{\mathcal{X}}_t$ using the learned dynamics model. The dynamics contains a known function $f(\cdot)$, a learned function $g(\cdot)$, and a process noise w (eq. (5.1)). (2) *Measurement*: Given a new sensor measurement y_t, control u_t (omitted in the figure), and the predicted zonotope $\overline{\mathcal{X}}_t$, GP-ZKF computes a measurement-consistent polytope $\overline{\mathcal{X}}_{y_t}$ using the learned observation model. The observation function contains a learned function $h(\cdot)$, and a measurement noise v (eq. (5.2)). (3) *Correction*: The new state estimate $\widehat{\mathcal{X}}_t$ is formed by intersecting the prediction and measurement sets, i.e., $\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t}$.

5.3 Method

Our method, GP-ZKF, is designed to represent and respect both epistemic and aleatoric uncertainties during state estimation. To represent epistemic uncertainty, we leverage Gaussian Processes (GPs) for learning both the dynamics and observation models. GPs provide not only a predictive mean but also confidence intervals that quantify the epistemic uncertainty due to limited or noisy training data in a principled way (lemma 3.3.1).

GP-ZKF builds on the recursive structure of traditional Kalman filters and set-based estimation techniques [222, 290]. As shown in fig. 5.2, GP-ZKF performs estimation in three sequential phases, *prediction*, *measurement*, and *correction*, to maintain a consistent estimate of the hidden human state.

We represent state estimates using zonotopes, a class of convex polytopes with favorable properties for set-based estimation. Zonotopes are closed under affine transformations and Minkowski sums, which makes them particularly well-suited for recursive state estimation under nonlinear, uncertain models. For more details on zonotopes and their operations, see section 3.2.3.

At each time step $t = 1, \ldots, T$, GP-ZKF computes a zonotopic estimate $\widehat{\mathcal{X}}_t \subset \mathbb{R}^{n_x}$ that contains the true hidden state x_t with a high probability. This estimate is constructed in three stages:

1. *Prediction*: Given the previous control input $u_{t-1} \in \mathfrak{U}$ and the previous zonotopic estimate $\widehat{\mathcal{X}}_{t-1} \subset \mathbb{R}^{n_x}$ that contains the true state x_{t-1}, GP-ZKF uses the learned dynamics model $d(\cdot)$ (see eq. (5.1)) to construct a *dynamics-consistent* zonotope $\overline{\mathcal{X}}_t \subset \mathbb{R}^{n_x}$. With a high probability, this set contains all possible next states reachable from $\widehat{\mathcal{X}}_{t-1}$ under the dynamics model and process noise, capturing epistemic and aleatoric uncertainty.

2. *Measurement*: Given the new sensor measurement $y_t \in \mathbb{R}^{n_y}$, the control input $u_t \in \mathfrak{U}$, and the predicted zonotope $\overline{\mathcal{X}}_t$, GP-ZKF uses the learned observation model $o(\cdot)$ (see eq. (5.2)) to construct a *measurement-consistent* polytope $\overline{\mathcal{X}}_{y_t} \subset \mathbb{R}^{n_x}$. With a high probability, this set contains all possible states that could have generated the measurement y_t under the observation model and measurement noise. Because this set may be asymmetric, we use general polytopes rather than zonotopes.

3. *Correction*: The final estimate $\widehat{\mathcal{X}}_t$ is obtained by intersecting the prediction and measurement sets $\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t}$. This intersection yields a conservative estimate that is consistent with both the dynamics and the sensor observations, while still preserving computational tractability.

We now introduce each of the three phases, prediction, measurement, and correction, in detail in the following subsections.

5.3.1 Phase 1: Prediction

In the prediction phase, the goal is to anticipate the range of possible future human states given the previous estimate and control. Specifically, given the previous zonotopic estimate $\widehat{\mathcal{X}}_{t-1}$ and robot control input u_{t-1}, GP-ZKF constructs a new *dynamics-consistent* zonotope $\overline{\mathcal{X}}_t$ that bounds the output of the dynamics function $d(\cdot)$ with a high probability. As defined in eq. (5.1), $d(\cdot)$ comprises three components: (1) a known model $f(\cdot)$ (e.g., physics-based prior), (2) an unknown residual model $g(\cdot)$ learned from data, and (3) additive process noise w. This section describes how we conservatively bound each component and integrate them to compute $\overline{\mathcal{X}}_t$.

Bounding the Known Dynamics Function

Accurately bounding the output of an arbitrary nonlinear function is difficult in general. We assume the known function $f(\cdot)$ is smooth and structured enough to allow for principled linear approximation:

Assumption 5.3.1. *The known function $f(\cdot)$ satisfy:*

(i) $f(\cdot)$ is twice continuously differentiable, so that we can perform linearization and apply standard reachability analysis tools.

(ii) *$f(\cdot)$ is L_f-Lipschitz continuous with respect to the 2-norm. That is, small changes in the input lead to proportionally bounded changes in the output.*

(iii) *The deviation of $f(x_0, u)$ from the initial state x_0 is bounded by a constant B^f, i.e., $\|f(x_0, u) - x_0\|_2 \leq B^f$, for each initial state $x_0 \in \widehat{\mathcal{X}}_0$ and each control $u \in \mathfrak{U}$. This ensures that the state does not change too drastically in a single step.*

In this section, we use assumption 5.3.1(i) to bound the outputs of $f(\cdot)$; we will incorporate assumption 5.3.1(ii,iii) (based on [232, assumption 3]) in the following subsubsection.

Assumption 5.3.1(i) allows us to directly apply reachability analysis [224] to bound the outputs of $f(\cdot)$, for any given u_{t-1} and $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$. In particular, we first linearize $f(\cdot)$ around a reference point $\overline{x}_{t-1} \in \mathbb{R}^{n_x}$, which is chosen to be the center of $\widehat{\mathcal{X}}_{t-1}$. The linearized function $\overline{f}(x_{t-1}, u_{t-1}) = f(\overline{x}_{t-1}, u_{t-1}) + J^f_x \cdot (x_{t-1} - \overline{x}_{t-1})$, where J^f_x is the Jacobian of $f(\cdot)$ with respect to x_{t-1}, evaluated at $(\overline{x}_{t-1}, u_{t-1})$. Given control u_{t-1}, for each state $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, the linearization error can be bounded by a zero-centered box that $\subset \mathbb{R}^{n_x}$ (see Althoff [224, proposition 3.7]). Formally, for each state $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$ and $u_{t-1} \in \mathfrak{U}$, we have the following bound:

$$f(x_{t-1}, u_{t-1}) - \overline{f}(x_{t-1}, u_{t-1}) \in R^f(\widehat{\mathcal{X}}_{t-1}, u_{t-1}) \subset \mathbb{R}^{n_x}. \tag{5.3}$$

Here, $R^f(\cdot)$ denotes the function used to compute the box that bounds the error based on $\widehat{\mathcal{X}}_{t-1}$ and u_{t-1}.

Bounding the Learned Residual Dynamics

The unknown function $g(\cdot)$ is learned via GP regression (see section 3.3). In this section, we formulate a high-probability bound for the output of $g(x_{t-1}, u_{t-1})$, given u_{t-1} and $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, in the following five steps:

- (i) *Regularity assumptions*: We begin by stating regularization assumptions about the target function $g(\cdot)$, which ensure that the learned model behaves well within the domain of interest.
- (ii) *Bounding the reachable state space*: We show that, with a high probability, the state remains within a compact space over time.
- (iii) *Bounding the GP posterior mean*: Within this compact space, we derive a bound for the GP posterior mean function, $\mu^g(\cdot)$.
- (iv) *Bounding the GP posterior standard deviation*: We present a bound for the GP posterior standard deviation function, $\sigma^g(\cdot)$.
- (v) *Combining bounds*: Finally, we integrate both bounds of $\mu^g(\cdot)$ and $\sigma^g(\cdot)$ into the GP confidence intervals, as introduced in lemma 3.3.1, to bound the output of $g(\cdot)$.

We now introduce each of the five steps.

(i) *Regularity Assumptions*:
We denote the kernel for $g(\cdot)$ as k^g and the single-output surrogate function as $g'(\cdot)$ (see

section 3.3). We make the following regularity assumptions for the function $g(\cdot)$ and its GP kernel k^g:

Assumption 5.3.2. *The unknown dynamics function $g(\cdot)$ and its GP kernel k^g satisfy:*

(i) Smoothness: The kernel k^g is 2-times continuously differentiable. This ensures that the learned function is smooth enough for reliable linearization. (See Steinwart and Christmann [291, definition 4.35].)

(ii) Boundedness: The kernel k^g is bounded, meaning its values do not grow without limit $\|k^g\|_\infty < \infty$ (See Steinwart and Christmann [291, Eq. (4.15)].)

(iii) Lipschitz continuity of derivatives: The derivatives of the kernel k^g are also bounded, which ensures the GP posterior mean and variance change smoothly with the input. (Adapted from Berkenkamp [232, Assumption 4].)

(iv) Function complexity bound: The target function $g'(\cdot)$ has a bounded norm in the reproducing kernel Hilbert space (RKHS) associated with k^g: $\|g'\|_{k^g} \leq B^g$. This limits the complexity of the function learned by GP regression. (See section 3.3 for more details.)

This assumption states that $k^g(\cdot,\cdot)$ and $g(\cdot)$ are smooth and bounded; common smooth kernels, such as square exponential and rational quadratic kernels, satisfy this assumption. Assumption 5.3.2 (i,iii,iv) implies that $g(\cdot)$ is L_g-Lipschitz continuous with respect to the 2-norm by Berkenkamp [232, corollary 2].

(ii) *Bounding the Reachable State Space*:

Although the process noise w follows a Gaussian distribution and thus has infinite support, the robot's actual state during estimation does not explore all of $\mathbb{R}^{n_x}$. Instead, we show that, with high probability, the system's state remains within a compact space throughout the entire horizon. This result is formalized in the lemma below:

Lemma 5.3.3. *Suppose that assumption 5.3.1 (ii,iii) and assumption 5.3.2 hold, and the process noise w is Gaussian as described in section 5.1. Then, starting from an initial set $\widehat{\mathcal{X}}_0 \ni x_0$, there exists a compact box $\mathfrak{X} \subset \mathbb{R}^{n_x}$, such that, with a probability at least $(1-\delta^w)$, jointly for each time step $t = 0, \ldots, T$, we have that the state $x_t \in \mathfrak{X}$. The size of $\mathfrak{X}$ depends on the initial set $\widehat{\mathcal{X}}_0$, the time horizon T, the noise level λ^w, the failure tolerance δ^w, and model properties L^f, L^g, B^f, B^g, $\|k^g\|_\infty$, and n_x.*

The proof follows by applying [232, lemma 44], combined with our bounded noise region $\mathcal{W}$ from remark 5.1.1. Intuitively, our assumption 5.3.1 (ii,iii), assumption 5.3.2, and the Gaussian noise assumption for w "prevent" $f(\cdot)$, $g(\cdot)$, and w (respectively) from drifting arbitrarily far away from $\widehat{\mathcal{X}}_0$ over time.

(iii) *Bounding the GP Posterior Mean*:

We now derive a bound on the error between the GP posterior mean function $\mu^g(\cdot)$ and its linear approximation. At each time step $t = 1, \dots, T$, we linearize μ^g around the center of the current zonotopic estimate $\widehat{\mathcal{X}}_{t-1}$, denoted by $\overline{x}_{t-1}$. The linearized function is:

$$\overline{\mu}^g(x_{t-1}, u_{t-1}) = \mu^g(\overline{x}_{t-1}, u_{t-1}) + J_x^{\mu_g} \cdot (x_{t-1} - \overline{x}_{t-1}),$$

where $J_x^{\mu_g}$ is the Jacobian of μ^g with respect to x_{t-1}, evaluated at $\overline{x}_{t-1}$.

Lemma 5.3.3 implies that the domain of $\mu^g(\cdot)$ during the estimation process is compact, with high probability. Together with assumption 5.3.2 (i), we obtain that, with a high probability, for each dimension $j = 1, \dots, n_x$, the mean μ_j^g is twice continuously differentiable with $L_{\nabla\mu}^g$-Lipschitz gradient. We then follow the steps in Koller et al. [52, section V(A)2)] to derive a bound for the linearization error. Formally, with a probability at least $(1 - \delta^w)$, uniformly for each time step $t = 1, \dots, T$, dimension $j = 1, \dots, n_x$, state $x_{t-1} \in \mathfrak{X}$, and control $u_{t-1} \in \mathfrak{U}$, the following bound holds:

$$|\mu_j^g(x_{t-1}, u_{t-1}) - \overline{\mu}_j^g(x_{t-1}, u_{t-1})| \leq \frac{1}{2} L_{\nabla\mu}^g \cdot \|x_{t-1} - \overline{x}_{t-1}\|_2^2, \tag{5.4}$$

where the probability $(1 - \delta^w)$ is due to the usage of lemma 5.3.3.

(iv) *Bounding the GP Posterior Standard Deviation*:
We approximate the GP posterior standard deviation $\sigma^g(x_{t-1}, u_{t-1})$ by evaluating it at the zonotope center $\overline{x}_{t-1}$. Using assumption 5.3.2 (i, ii, iii) and results from Lederer, Umlauft, and Hirche [292, eq (21) and (22)], the error in this approximation can be bounded.

There exists a constant $L_\sigma^g \in \mathbb{R}$, determined by the GP kernel and training data, such that for each time step $t = 1, \dots, T$, dimension $j = 1, \dots, n_x$, state $x_{t-1} \in \mathbb{R}^{n_x}$, and control $u_{t-1} \in \mathfrak{U}$, the following holds:

$$|\sigma_j^g(x_{t-1}, u_{t-1}) - \sigma_j^g(\overline{x}_{t-1}, u_{t-1})| \leq L_\sigma^g \cdot \|x_{t-1} - \overline{x}_{t-1}\|_2^{1/2}. \tag{5.5}$$

This square-root dependence reflects how the standard deviation function is typically less sensitive to input changes than the mean, and gives a principled way to approximate uncertainty across the zonotope.

(v) *Combining Bounds*:
To bound the total error between the unknown function $g(\cdot)$ and the linearized GP posterior mean $\overline{\mu}^g(\cdot)$, we combine three sources of uncertainty: (1) the linearization error in the mean function, (2) the GP confidence interval capturing epistemic uncertainty, and (3) and the approximation error in the standard deviation.

Using Assumption 5.3.2 (iv), we apply lemma 3.3.1 to construct a high-probability confidence interval for $g(\cdot)$ with a failure probability $\delta^g \in (0, 1)$. For each dimension

$j = 1, \ldots, n_x$, we decompose the total error $|g_j(x_{t-1}, u_{t-1}) - \overline{\mu}^g_j(x_{t-1}, u_{t-1})|$ as follows:

$$\begin{aligned}
&|g_j(x_{t-1}, u_{t-1}) - \overline{\mu}^g_j(x_{t-1}, u_{t-1})| \\
&\leq |\mu^g_j(x_{t-1}, u_{t-1}) - \overline{\mu}^g_j(x_{t-1}, u_{t-1})| + |g_j(x_{t-1}, u_{t-1}) - \mu^g_j(x_{t-1}, u_{t-1})| && (5.6b) \\
&\leq \frac{1}{2} L^g_{\nabla\mu} \cdot \|x_{t-1} - \overline{x}_{t-1}\|_2^2 + \beta^g \cdot \sigma^g_j(x_{t-1}, u_{t-1}) && (5.6c) \\
&\leq \frac{1}{2} L^g_{\nabla\mu} \cdot \|x_{t-1} - \overline{x}_{t-1}\|_2^2 + \beta^g \cdot \sigma^g_j(\overline{x}_{t-1}, u_{t-1}) + \beta^g L^g_\sigma \cdot \|x_{t-1} - \overline{x}_{t-1}\|_2^{1/2} && (5.6d) \\
&\leq \frac{1}{2} L^g_{\nabla\mu} \cdot \|\widehat{\mathcal{X}}_{t-1} - \overline{x}_{t-1}\|_2^2 + \beta^g \cdot \sigma^g_j(\overline{x}_{t-1}, u_{t-1}) + \beta^g L^g_\sigma \cdot \|\widehat{\mathcal{X}}_{t-1} - \overline{x}_{t-1}\|_2^{1/2}. && (5.6e)
\end{aligned}$$

Here, inequality eq. (5.6b) applies the triangle inequality. Inequality eq. (5.6c) combines the GP posterior mean's linearization error bound (eq. (5.4)) and the GP confidence interval (lemma 3.3.1) via a union bound. In this way, all inequalities starting at eq. (5.6c) hold with a probability at least $(1 - \delta^g - \delta^w)$. Then, inequality eq. (5.6d) replaces the standard deviation at x_{t-1} with that at $\overline{x}_{t-1}$, introducing an approximation error term based on eq. (5.5). Finally, inequality eq. (5.6e) bounds the distance $\|x_{t-1} - \overline{x}_{t-1}\|_2$ over all $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$. In particular, we define the norm of the translated zonotope $\|\widehat{\mathcal{X}}_{t-1} - \overline{x}_{t-1}\|_2 \coloneqq \max_{x_{t-1} \in \widehat{\mathcal{X}}_{t-1}} \|x_{t-1} - \overline{x}_{t-1}\|_2$, which is the maximum deviation within the zonotope (a standard norm in set-based estimation [293]).

To make this bound explicit in terms of the size of the zonotopic estimate $\widehat{\mathcal{X}}_{t-1}$, we define $\epsilon \coloneqq \|\widehat{\mathcal{X}}_{t-1} - \overline{x}_{t-1}\|_2$. Substituting this into eq. (5.6e) gives the following:

$$|g_j(x_{t-1}, u_{t-1}) - \overline{\mu}^g_j(x_{t-1}, u_{t-1})| \leq \underbrace{\frac{1}{2} L^g_{\nabla\mu} \cdot \epsilon^2}_{\text{Linearization error of } \mu^g_j(\cdot)} + \underbrace{\beta^g \cdot \sigma^g_j(\overline{x}_{t-1}, u_{t-1}) + \beta^g \cdot L^g_\sigma \cdot \epsilon^{1/2}}_{\text{Epistemic uncertainty } \bigoplus \text{ Approx. error of } \sigma^g_j(\cdot)}. \tag{5.7}$$

This gives us a high-confidence bound on how much the true value of the unknown function $g_j(\cdot)$ may deviate from its linearized GP posterior mean $\overline{\mu}^g_j(\cdot)$. This bound accounts for both epistemic uncertainty (model error) and the geometric structure of our zonotopic state estimates.

We encapsulate this bound in eq. (5.7) across all dimensions in a zero-centered box $R^g(\widehat{\mathcal{X}}_{t-1}, u_{t-1}) \subset \mathbb{R}^{n_x}$, where the radius in each dimension $j = 1, \ldots, n_x$ is the right-hand side of eq. (5.7). Then, with a probability at least $(1 - \delta^g - \delta^w)$, jointly for each time step $t = 1, \ldots, T$, zonotopic estimate $\widehat{\mathcal{X}}_{t-1} \subset \mathfrak{X}$, control $u_{t-1} \in \mathfrak{U}$, and state $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, the following holds:

$$g(x_{t-1}, u_{t-1}) - \overline{\mu}^g(x_{t-1}, u_{t-1}) \in R^g(\widehat{\mathcal{X}}_{t-1}, u_{t-1}). \tag{5.8}$$

This bound plays a critical role in the prediction phase of GP-ZKF, as it captures epistemic uncertainty in the learned dynamics while maintaining computational tractability through a simple geometric representation.

The Prediction Phase

We now combine the bounds on the known function $f(\cdot)$, the learned function $g(\cdot)$ (modeled via a GP), and the process noise w to complete the first phase of GP-ZKF: the prediction phase.

We begin by defining a linear approximation of the full dynamics function $d(\cdot)$. This approximation, denoted $\overline{d}(\cdot)$, is constructed by summing the linearizations of the known model $f(\cdot)$ and the GP posterior mean $\mu^g(\cdot)$:

$$\overline{d}(x_{t-1}, u_{t-1}) := \overline{f}(x_{t-1}, u_{t-1}) + \overline{\mu}^g(x_{t-1}, u_{t-1}).$$

Next, we define an error bounding box $R^d(\widehat{\mathcal{X}}_{t-1}, u_{t-1}) \subset \mathbb{R}^{n_x}$ that captures all possible deviations between the true dynamics $d(\cdot)$ and the linearized approximation $\overline{d}(x_{t-1}, u_{t-1})$, with a high probability. This box is constructed by taking the Minkowski sum of the individual error bounds from (1) the known function $f(\cdot)$ (from eq. (5.3)), (2) the learned function $g(\cdot)$ (from eq. (5.8)), and (3) the process noise w (from remark 5.1.1). As a result, with a probability at least $(1 - \delta^g - \delta^w)$, jointly for each time step $t = 1, \ldots, T$, zonotopic estimate $\widehat{\mathcal{X}}_{t-1} \subset \mathfrak{X}$, control $u_{t-1} \in \mathfrak{U}$, and state $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, the following holds:

$$\begin{aligned} d(x_{t-1}, u_{t-1}, w_{t-1}) - \overline{d}(x_{t-1}, u_{t-1}) &\in R^d(\widehat{\mathcal{X}}_{t-1}, u_{t-1}) \\ &:= \underbrace{R^f(\widehat{\mathcal{X}}_{t-1}, u_{t-1})}_{\text{Linearization err. of } f(\cdot)} \bigoplus \underbrace{R^g(\widehat{\mathcal{X}}_{t-1}, u_{t-1})}_{\substack{\text{Epistemic} \bigoplus \text{Lin. err. of } \mu^g(\cdot) \\ \bigoplus \text{Approx. err. of } \sigma^g(\cdot)}} \bigoplus \underbrace{\mathcal{W}}_{\text{Aleatoric}} . \end{aligned} \tag{5.9}$$

Given that $\overline{d}$ is a linear function, u_{t-1} is known, and $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, the range of $\overline{d}(x_{t-1}, u_{t-1})$ is itself a zonotope, obtained by linearly transforming $\widehat{\mathcal{X}}_{t-1}$. Then, by Minkowski-summing this transformed zonotope and $R^d(\widehat{\mathcal{X}}_{t-1}, u_{t-1})$, we obtain the dynamics-consistent zonotope, $\overline{\mathcal{X}}_t$, that bounds all possible outputs of the true dynamics $d(x_{t-1}, u_{t-1}, w_{t-1})$ with a high probability. Let $D(\widehat{\mathcal{X}}_{t-1}, u_{t-1})$ denote the function to compute $\overline{\mathcal{X}}_t$. We summarize this construction in the following lemma:

Lemma 5.3.4. *Let δ^g and $\delta^w \in (0, 1)$. For each GP training data size $n \in \mathbb{N}$, let the GP confidence scaling factor β^g_n be chosen as in lemma 3.3.1. Given an initial set $\widehat{\mathcal{X}}_{n,0} \ni x_{n,0}$, then, with a probability at least $(1 - \delta^g - \delta^w)$, the following holds jointly for each data size $n \in \mathbb{N}$ and each time step $t = 1, \ldots, T$:*

(i) The state at time t lies within the dynamics-consistent zonotope:
$d(x_{n,t-1}, u_{n,t-1}, w_{n,t-1}) \in \overline{\mathcal{X}}_{n,t} := D(\widehat{\mathcal{X}}_{n,t-1}, u_{n,t-1})$,
for each zonotopic estimate $\widehat{\mathcal{X}}_{n,t-1} \subset \mathfrak{X}$, state $x_{n,t-1} \in \widehat{\mathcal{X}}_{n,t-1}$, and control $u_{n,t-1} \in \mathfrak{U}$, where $w_{n,t-1}$ is the process noise as assumed in section 5.1.

(ii) The state remains in the compact set $\mathfrak{X}$: $x_{n,t} \in \mathfrak{X}$, $x_{n,0} \in \mathfrak{X}$.

Proof. The noise bound $\mathcal{W}$ from remark 5.1.1 and the compactness of state space from lemma 5.3.3 hold jointly, with a probability at least $(1 - \delta^w)$. Then, by applying a union bound to combine the above result with the confidence intervals of $g(\cdot)$ from lemma 3.3.1, we arrive at the bound, $\overline{\mathcal{X}}_{n,t}$. □

Summary: Given the previous control input u_{t-1} and the previous zonotopic estimate $\widehat{\mathcal{X}}_{t-1}$ (which contains the true state x_{t-1}), the prediction phase of GP-ZKF computes a dynamics-consistent zonotope $\overline{\mathcal{X}}_t \subset \mathbb{R}^{n_x}$ that captures all possible states at time t under the system dynamics. This predicted zonotope accounts for uncertainty from model approximation and process noise, and is guaranteed to contain the true state x_t with a high probability. It forms the foundation for subsequent correction based on sensor observations.

5.3.2 Phase 2: Measurement

In the measurement phase, GP-ZKF uses the dynamics-consistent zonotope $\overline{\mathcal{X}}_t$ from the previous step and the current control input u_t to compute a new set, $\overline{\mathcal{X}}_{y_t}$, that captures all possible states consistent with the latest sensor measurement y_t, with a high probability. This set is referred to as the measurement-consistent polytope [290].

The observation model $h(\cdot)$, which maps states and controls to sensor measurements, is unknown and learned via GP regression. We denote the single-output surrogate function by $h'(\cdot)$, and the kernel by k^h. Similar to assumption 5.3.2 for $g(\cdot)$, we make the following regularity assumptions for the function $h(\cdot)$ and its GP kernel k^h:

Assumption 5.3.5. *The unknown observation function $h(\cdot)$ and its GP kernel k^h satisfy:*

(i) Smoothness: The kernel k^h is 2-times continuously differentiable. (See Steinwart and Christmann [291, definition 4.35].)

(ii) Boundedness: The kernel k^h is bounded: $\|k^h\|_\infty < \infty$ (See Steinwart and Christmann [291, Eq. (4.15)].)

(iii) Lipschitz continuity of derivatives: The derivatives of the kernel k^h are also bounded. (Adapted from Berkenkamp [232, Assumption 4].)

(iv) Function complexity bound: The target function $h'(\cdot)$ has a bounded RKHS norm: $\|h'\|_{k^h} \leq B^h$. (See section 3.3 for more details.)

We now bound the output of $h(\cdot)$ using the same approach developed for the learned dynamics component $g(\cdot)$ in section 5.3.1.

(iv) *Bounding the GP Posterior Mean*:

We linearize the GP posterior mean function $\mu^h(\cdot)$ around the center $\overline{x}_t$ of $\overline{\mathcal{X}}_t$, yielding:

$$\overline{\mu}^h(x_t, u_t) = \mu^h(\overline{x}_t, u_t) + J_x^{\mu_h} \cdot (x_t - \overline{x}_t),$$

where $J_x^{\mu_h}$ is the Jacobian of μ^h with respect to x_t, evaluated at $\overline{x}_t$. Since lemma 5.3.4(ii) ensures that x_t lies within a compact region $\mathfrak{X}$ with a high probability, this allows us to apply a bound on the linearization error, similar to eq. (5.4).

(iv) *Bounding the GP Posterior Standard Deviation*:

Under assumption 5.3.5, we can also bound the error between the actual standard deviation $\sigma^h(x_t, u_t)$ and its approximation $\sigma^h(\overline{x}_t, u_t)$, similar to eq. (5.5).

(iv) *Combining bounds*:

Given $\delta^h \in (0, 1)$, lemma 3.3.1 allows us to construct confidence intervals for $h(\cdot)$. Via a union bound, we obtain that the confidence intervals and noise bound $v_t \in \mathcal{V}$ (see remark 5.1.1) jointly hold with a probability at least $(1-\delta^h-\delta^v)$. Then, similar to R^g in eq. (5.8), we obtain a box, $R^h(\overline{\mathcal{X}}_t, u_t) \subset \mathbb{R}^{n_y}$, such that with a probability at least $(1-\delta^g-\delta^w)(1-\delta^h-\delta^v)$, jointly for each time step $t = 1, \dots, T$, dynamics-consistent zonotope $\overline{\mathcal{X}}_t \subset \mathfrak{X}$, control $u_t \in \mathfrak{U}$, and state $x_t \in \overline{\mathcal{X}}_t$, we have that

$$h(x_t, u_t) - \overline{\mu}^h(x_t, u_t) \in R^h(\overline{\mathcal{X}}_t, u_t). \tag{5.10}$$

Here, the product rule, $(1-\delta^g-\delta^w)(1-\delta^h-\delta^v)$, results from the assumption that noises w and v are assumed independent (see section 5.1).

By expanding $\overline{\mu}^h$ and then combining the noise bound $\mathcal{V}$ (remark 5.1.1) with eq. (5.10), we obtain that with probability at least $(1-\delta^g-\delta^w)(1-\delta^h-\delta^v)$, jointly for each time step $t = 1, \dots, T$, dynamics-consistent zonotope $\overline{\mathcal{X}}_t \subset \mathfrak{X}$, control $u_t \in \mathfrak{U}$, and state $x_t \in \overline{\mathcal{X}}_t$, the following holds:

$$\begin{aligned} \overline{\mu}^h(x_t, u_t) - o(x_t, u_t, v_t) &= \mu^h(\overline{x}_t, u_t) + J_x^{\mu_h} \cdot (x_t - \overline{x}_t) - h(x_t, u_t) - v_t \\ &= \overline{\mu}^h(x_t, u_t) - h(x_t, u_t) - v_t \\ &\in R^o(\overline{\mathcal{X}}_t, u_t) := \underbrace{R^h(\overline{\mathcal{X}}_t, u_t)}_{\substack{\text{Epistemic} \bigoplus \text{Lin. err. of } \mu^h(\cdot) \\ \bigoplus \text{Approx. err. of } \sigma^h(\cdot)}} \bigoplus \underbrace{\mathcal{V}}_{\text{Aleatoric}}, \end{aligned} \tag{5.11}$$

where $R^o(\overline{\mathcal{X}}_t, u_t)$ represents the total error in the learned observation model, capturing both epistemic and aleatoric sources.

Constructing the Measurement-Consistent Polytope

Given the actual sensor measurement $y_t = o(x_t, u_t, v_t) \in \mathbb{R}^{n_y}$, we invert the observation model to solve for the set of possible state x_t that are consistent with this measurement. Since these states must satisfy the bound in eq. (5.11), we equivalently represent this bound as a polytope, $\overline{\mathcal{X}}_{y_t} \subset \mathbb{R}^{n_x}$, with x_t as the variable, defined as follows:

$$\overline{\mathcal{X}}_{y_t} := \left\{ x_t \in \mathbb{R}^{n_x} : J_x^{\mu_h} \cdot x_t - [y_t - \mu^h(\overline{x}_t, u_t) + J_x^{\mu_h} \cdot \overline{x}_t] \in R^o(\overline{\mathcal{X}}_t, u_t) \right\}. \tag{5.12}$$

This measurement-consistent polytope $\overline{\mathcal{X}}_{y_t}$ contains all the states that align with the current measurement y_t and the learned observation model. We denote the function that computes this polytope as $O^{\text{inv}}(\overline{\mathcal{X}}_t, u_t, y_t)$, with the superscript inv indicating that O^{inv} is the "inverse" of our observation model, $o(\cdot)$.

Summary: The measurement phase produces a polytope $\overline{\mathcal{X}}_{y_t} \subset \mathbb{R}^{n_x}$, which contains all states at time t that are consistent with both the robot's sensor readings and the uncertainty-aware GP observation model. This polytope will be intersected with the predicted zonotope in the next phase to produce the final state estimate.

5.3.3 Phase 3: Correction

In the correction phase, the goal is to combine predictions from the dynamics model and evidence from the new sensor measurement to compute a refined estimate of the current state.

Specifically, we construct a new zonotope $\widehat{\mathcal{X}}_t$ that tightly bounds the possible true states at time t, by intersecting: (1) the dynamics-consistent zonotope $\overline{\mathcal{X}}_t$ from the dynamics model (lemma 5.3.4), (2) the measurement-consistent polytope $\overline{\mathcal{X}}_{y_t}$ derived from the sensor reading (eq. (5.12)), and (3) the known bounded region $\mathfrak{X}$ of reachable states.

Note: intersecting with $\mathfrak{X}$ is necessary because the derivation of the measurement-consistent polytope $\overline{\mathcal{X}}_{y_t}$ (see eq. (5.11)) assumes that dynamics-consistent zonotope $\overline{\mathcal{X}}_t$ lies entirely within $\mathfrak{X}$. Enforcing this condition ensures the validity of the GP bounds and consistency guarantees.

Because this intersection cannot be computed exactly, we compute a zonotope $\widehat{\mathcal{X}}_t$ that conservatively outer-approximates it. This intersection defines the most plausible region where the true state can lie, given both prior predictions and current observations. The resulting zonotope, $\widehat{\mathcal{X}}_t$, will be smaller (less conservative) than either $\overline{\mathcal{X}}_t$ or $\overline{\mathcal{X}}_{y_t}$ alone, because it incorporates both prediction and measurement information. We formalize this step in the following lemma:

Lemma 5.3.6. *Let $\delta^g, \delta^w, \delta^h, \delta^v \in (0, 1)$. For each GP training data sizes $n^g, n^h \in \mathbb{N}$, choose GP confidence scaling factors $\beta^g_{n^g}, \beta^h_{n^h}$ according to lemma 3.3.1*[1]*. Assume that the initial state is bounded in a known zonotope: $\widehat{\mathcal{X}}_0 \ni x_0$, then, with a probability at least $(1-\delta^g-\delta^w)(1-\delta^h-\delta^v)$, jointly for each data size $n^g, n^h \in \mathbb{N}$, time step $t = 1, \ldots, T$, zonotopic estimate $\widehat{\mathcal{X}}_{t-1} \subset \mathfrak{X}$, state $x_{t-1} \in \widehat{\mathcal{X}}_{t-1}$, controls $u_{t-1}, u_t \in \mathfrak{U}$, and measurement $y_t \in \mathbb{R}^{n_y}$, the following holds:*

$$d(x_{t-1}, u_{t-1}, w_{t-1}) \in \left(\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t} \cap \mathfrak{X}\right) \subset \widehat{\mathcal{X}}_t, \tag{5.13}$$

where $d(\cdot)$ is the true dynamics function including noise, $\overline{\mathcal{X}}_t := D(\widehat{\mathcal{X}}_{t-1}, u_{t-1})$ is the dynamics-consistent zonotope (lemma 5.3.4), $\overline{\mathcal{X}}_{y_t} := O^{inv}(\overline{\mathcal{X}}_t, u_t, y_t)$ is the measurement-consistent polytope (eq. (5.12)), and $\widehat{\mathcal{X}}_t$ is a zonotope that outer-approximates the intersection.

This lemma guarantees that the true human state at time t lies inside the new zonotope $\widehat{\mathcal{X}}_t$, with a high probability. Lemma 5.3.6 summarizes the derivations in section 5.3.2; it can be proved by directly combining lemma 5.3.4 and the bound in eq. (5.11).

We compute $\widehat{\mathcal{X}}_t$ to outer-approximate the intersection in eq. (5.13) in the following two steps: (1) GP-ZKF follows Le et al. [290, proposition 1] to obtain a zonotope denoted by $Z_t(\Lambda_t)$, parameterized by the matrix Λ_t, such that $Z_t(\Lambda_t) \supset \left(\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t}\right)$. The parameter Λ_t is obtained by analytically solving a convex program that minimizes the "size" of $Z_t(\Lambda_t)$ (see Alamo, Bravo, and Camacho [222, section 6.1]). (2) GP-ZKF follows the same procedures to construct $\widehat{\mathcal{X}}_t \supset \left(\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t} \cap \mathfrak{X}\right)$.

We summarize the correction step with the function: $\widehat{\mathcal{X}}_t = E(\widehat{\mathcal{X}}_{t-1}, u_{t-1}, u_t, y_t)$, which takes the previous estimate, controls, and the new sensor measurement as input, and outputs the updated zonotopic state estimate.

Summary: The correction phase fuses the robot's prediction (from dynamics) and observation (from sensors) to update its estimate and uncertainty about the hidden state. By intersecting dynamics-consistent zonotope and measurement-consistent polytope, and carefully outer-approximating the resulting set, GP-ZKF produces a reliable and conservative estimate of the hidden state, accounting for both epistemic and aleatoric uncertainty.

[1] We omit the subscripts n^g, n^h for every variable in this lemma for clarity.

5.4 Theoretical Guarantees

We present two key theoretical results about the proposed estimator, GP-ZKF. First, we prove that GP-ZKF provides a formal consistency guarantee: its zonotopic estimates contain the true system state with a high probability, even under epistemic and aleatoric uncertainty. Second, we show that under certain relaxations, GP-ZKF reduces exactly to GP-EKF [281], the standard stochastic estimator using Gaussian Processes within the Extended Kalman Filter (EKF) framework.

5.4.1 Consistency Guarantee

Theorem 5.4.1 (δ-Consistency of GP-ZKF)**.** *Given $\delta \in (0,1)$, GP-ZKF selects individual failure probabilities $\delta^g, \delta^h, \delta^w, \delta^v \in (0,1)$ such that $(1-\delta^g-\delta^w)(1-\delta^h-\delta^v) \geq (1-\delta)$. For each GP training data sizes $n^g, n^h \in \mathbb{N}$, GP-ZKF chooses GP confidence scaling factors $\beta^g_{n^g}, \beta^h_{n^h}$ according to lemma 3.3.1. Then, GP-ZKF is δ-consistent. In other words, with a probability at least $1-\delta$, jointly for each time step $t = 1, \ldots, T$, the true state x_t always lies within the zonotopic estimate $\widehat{\mathcal{X}}_t$.*

Proof. This is followed by recursively applying lemma 5.3.6 for each time step $t = 1, \ldots, T$, similar to the argument in Koller et al. [52, Corollary 7]. □

5.4.2 Connection to GP-EKF

GP-EKF [281] is a stochastic state estimator that uses GPs to learn both the dynamics and observation models. We see GP-EKF as the *stochastic* counterpart to our *set-based* GP-ZKF. At every time $t = 1, \ldots, T$, GP-EKF updates its Gaussian belief about the hidden state by first computing a Kalman gain $K_t \in \mathbb{R}^{n_x \times n_y}$, and then outputing a point estimate $\mu_t \in \mathbb{R}^{n_x}$ and a covariance matrix $\Sigma_t \in \mathbb{R}^{n_x \times n_x}$ (see Ko and Fox [281, table 2]).

To connect GP-ZKF with GP-EKF, similar to Combastel [287, theorem 7], we interpret the set-based elements of GP-ZKF analogously:

- GP-ZKF's point estimate analog is the center of the zonotope, denoted by $(\widehat{\mathcal{X}}_t)_c$.
- GP-ZKF's covariance matrix analog is defined as the outer product of its generator matrix: $(\widehat{\mathcal{X}}_t)_G \left((\widehat{\mathcal{X}}_t)_G\right)^\top$, which is called the zonotope covariation ([287, definition 4]).
- GP-ZKF's Kalman gain analog is the matrix Λ_t, which parametrizes the outer approximation of the intersection $\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t}$, as mentioned at the end of section 5.3.3.

The following theorem states that if we remove the uncertainty-handling mechanisms from GP-ZKF, it behaves identically to (the Joseph form of) GP-EKF:

Theorem 5.4.2 (Equivalence to GP-EKF)**.** *Assume that the zonotopic estimates produced by GP-ZKF always remain entirely within the known reachable state space $\mathfrak{X}$, i.e., for each time step $t = 1, \ldots, T$, the zonotope $Z_t(\Lambda_t) \subset \mathfrak{X}$, where $Z_t(\Lambda_t)$ is defined at the end of section 5.3.3. Suppose that GP-ZKF is initialized with the same point estimate and covariance as GP-EKF: $\mu_0 = (\widehat{\mathcal{X}}_0)_c$ and $\Sigma_0 = (\widehat{\mathcal{X}}_0)_G \left((\widehat{\mathcal{X}}_0)_G\right)^\top$. If GP-ZKF:*

- *Sets the confidence interval scalings $\beta^g = \beta^h = 1$.*
- *Omits all noise bounds for w and v by setting $\mathcal{W} = \emptyset$ and $\mathcal{V} = \emptyset$.*
- *Omits all linearization errors for $f(\cdot)$, $\mu^g(\cdot)$, $\sigma^g(\cdot)$, $\mu^h(\cdot)$, and $\sigma^h(\cdot)$ by setting $R^f(\cdot) = \emptyset$ (eq. (5.3)), $L^g_{\nabla\mu} = 0$ (eq. (5.4)), $L^g_\sigma = 0$ (eq. (5.5)), $L^h_{\nabla\mu} = 0$ (section 5.3.2), and $L^h_\sigma = 0$ (section 5.3.2);*

then, for each time step $t = 1, \ldots, T$, the estimates from GP-ZKF exactly match those from GP-EKF: $K_t = \Lambda_t$, $\mu_t = (\widehat{\mathcal{X}}_t)_c$, and $\Sigma_t = (\widehat{\mathcal{X}}_t)_G \left((\widehat{\mathcal{X}}_t)_G\right)^\top$.

Proof. Under the relaxations above, eq. (5.13) becomes $\left(\overline{\mathcal{X}}_t \cap \overline{\mathcal{X}}_{y_t} \cap \mathfrak{X}\right) \subset Z_t(\Lambda_t) = \widehat{\mathcal{X}}_t$. And each of the uncertainty bounds, $R^d(\cdot)$ (eq. (5.9)) and $R^o(\cdot)$ (eq. (5.11)), only contains one standard deviation. As introduced at the end of section 5.3.3, GP-ZKF obtains $\widehat{\mathcal{X}}_t(\Lambda_t)$ by optimizing Λ_t. With the analytical solution, Λ_t, we reach the final conclusion by induction (see the proof of Combastel [287, theorem 7]). □

This theorem states that with certain relaxations, GP-ZKF could produce the same Kalman gain, point estimate, and covariance as GP-EKF. The Kalman gain in GP-ZKF, Λ_t, weighs the dynamics-consistent zonotope, $\overline{\mathcal{X}}_t$, and the measurement-consistent polytope, $\overline{\mathcal{X}}_{y_t}$, when "mixing" them within the outer-approximated intersection. In contrast to GP-EKF, theorem 5.4.2 signifies the conservativeness of GP-ZKF in bounding the linearization errors and aleatoric and epistemic uncertainties during estimation. The conservativeness echoes GP-ZKF's consistency guarantee, as stated in theorem 5.4.1.

In summary, theorem 5.4.1 provides a formal guarantee that GP-ZKF's estimates are reliable. Theorem 5.4.2 shows that GP-ZKF is a strict superset of GP-EKF, matching it under idealized assumptions. These results highlight the strength of GP-ZKF: it maintains the structure of a Kalman-like estimator while offering robust guarantees under both epistemic and aleatoric uncertainties.

5.5 Experiment and Results

In the previous section, we established the theoretical guarantees of GP-ZKF: its probabilistic consistency and its connection to the standard GP-EKF estimator [281]. In this section, we provide empirical evidence demonstrating GP-ZKF's advantages in terms of consistency and robustness.

We compare GP-ZKF against three widely used stochastic state estimators that also leverage GP regression to learn both dynamics and observation models: GP-EKF (Extended Kalman Filter), GP-UKF (Unscented Kalman Filter), and GP-PF (Particle Filter), as introduced in Ko and Fox [281].

We evaluate all methods in two settings: (1) A simulated inverted pendulum task with high epistemic uncertainty (i.e., inaccurate learned models), and (2) A real-world robot-assisted dressing task, where the robot estimates the human arm's position using force sensors.

We assess each method using the following four metrics:

1. *Average RMSE (Root-Mean-Square Error) (per dimension)*: Measures the accuracy of the point estimate. For GP-ZKF, the point estimate is defined as the center of the zonotopic estimate $(\widehat{\mathcal{X}}_t)_c$ (see section 5.4).

2. *Inclusion Rate (%)*: Measures consistency by computing the percentage of time steps where the true state lies within the estimated set. For GP-ZKF, this is the zonotopic estimate. For the stochastic methods (GP-EKF, GP-UKF, GP-PF), we convert their covariance matrix into equivalent 95 % confidence ellipsoids. For GP-PF, the posterior is approximated as a Gaussian before constructing the ellipsoid.

   Recall that theorem 5.4.2 shows that GP-ZKF's zonotope is theoretically equivalent to the *unscaled* covariance-based estimate in GP-EKF. However, in practice, the covariance matrices from GP-EKF, GP-UKF, and GP-PF are scaled up to form 95 % ellipsoids, in order to match the consistency guarantee of GP-ZKF's set-based estimate. This allows a fair, apples-to-apples comparison of inclusion rates across methods.

3. *Average Radius (per dimension)*: Measures the conservativeness of the estimate by computing the radius of the smallest axis-aligned box that contains the set-based estimate.

4. *Average Computation Time (per time step)*: Evaluates the computational cost of each method.

5.5.1 Simulated Pendulum Domain

Experiment Design

We first evaluated GP-ZKF in a controlled simulated environment using a discrete-time 2D inverted pendulum. The pendulum is stabilized by an infinite-horizon linear quadratic regulator, with the goal of keeping the pendulum upright. The state is defined as $x = [\theta, \dot{\theta}]^\top$, where θ is the angle and $\dot{\theta}$ is the angular velocity. The set-point corresponds to the pendulum standing upright ($\theta = 0°$). The closed-loop dynamics of the pendulum are denoted by $d(\cdot)$, with additive process noise w (standard deviation $\lambda_w = 7.16°$). The robot perceives the state through an observation function $o(\cdot)$, which maps the state and control to the end-effector's position and velocity in $\mathbb{R}^4$, with additive observation noise v ($\lambda_v = 8.88°$). The known dynamics model $f(\cdot)$ corresponds to the linearized and discretized dynamics around the upright set-point.

To evaluate consistency under significant model errors, we introduced distribution shifts between training and test conditions. At test time, each method was run for $T = 15$ time steps, starting from four initial angles θ_0 sampled uniformly from the testing region $[\pi, 2\pi]$, with angular velocity fixed at $\dot{\theta}_0 = 0$. Each start state was repeated for 10 trials. We varied the training data for the learned models to create four different scenarios:

1. *Shift Both*: Both the dynamics model $g(\cdot)$ and the observation model $h(\cdot)$ were trained using a *default dataset* consisting of 9 rollouts with start states in the training region: $\theta_0 \in [0, \pi]$ and $\dot{\theta}_0 = 0$. Since none of the training data covers the test region, both models face significant distribution shift and thus high epistemic uncertainty.

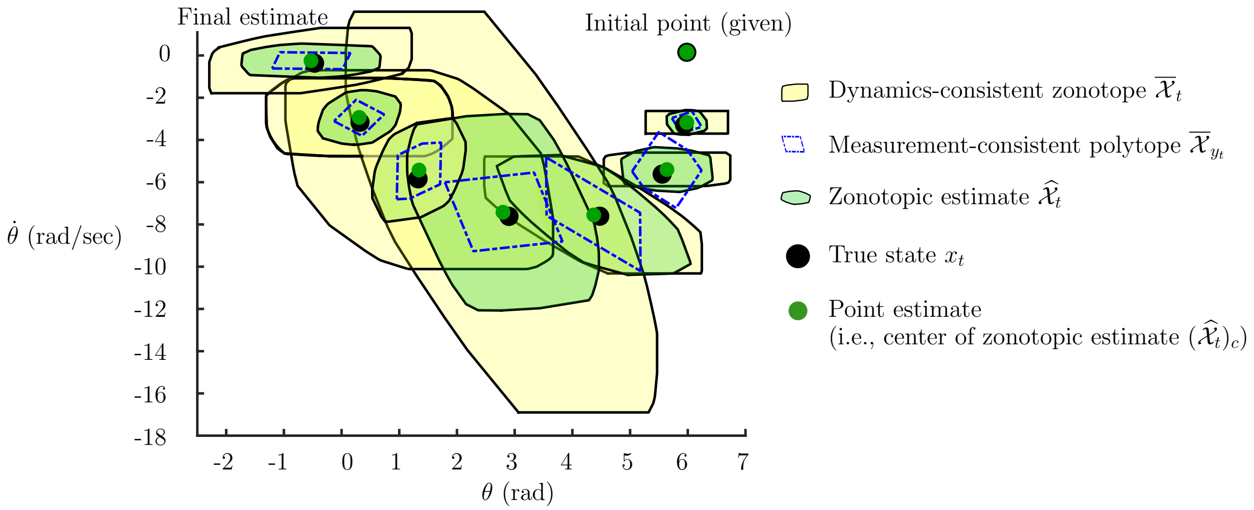


Figure 5.3: Zonotopic estimates along a trajectory produced by GP-ZKF under the *Shift None* condition (see section 5.5.1). Each zonotopic estimate, $\widehat{\mathcal{X}}_t$ (green fill), always outer-approximates the intersection of the dynamics-consistent zonotope, $\overline{\mathcal{X}}_t$ (yellow fill), and the measurement-consistent polytope, $\overline{\mathcal{X}}_{y_t}$ (blue outline). Even when the point estimate (green dot), defined as the center of $\widehat{\mathcal{X}}_t$, is inaccurate, the full zonotope still contains the true state (black dot), demonstrating the consistency of GP-ZKF. While the size of $\overline{\mathcal{X}}_t$ may increase over time due to uncertainty propagation, the correction step with informative measurements allows the estimate $\widehat{\mathcal{X}}_t$ to shrink, improving precision.

2. *Shift Dynamics Only*: The observation model $h(\cdot)$ was trained with both the default dataset and an additional supervised dataset containing 16 state-measurement pairs with states uniformly sampled from the test region $[\pi, 2\pi] \times [-3\pi, 0]$. The dynamics model $g(\cdot)$ was trained only on the default dataset. This setup reduces epistemic uncertainty for the observation model but leaves the dynamics model exposed to distribution shift.

3. *Shift Observation Only*: The dynamics model $g(\cdot)$ was trained with both the default dataset and five additional rollouts starting from the test region $\left(\theta_0 \in [\pi, 2\pi], \dot{\theta}_0 = 0\right)$. The observation model $h(\cdot)$ was trained only on the default dataset. This setup reduces epistemic uncertainty for the dynamics model but leaves the observation model exposed to a distribution shift.

4. *Shift None*: Both models were trained on the default dataset and their respective additional datasets. As both the dynamics and observation models have training data that covers the test region, neither is exposed to significant epistemic uncertainty.

In all settings, the GP confidence scaling factors β^g and β^h were manually specified and adjusted based on the amount of available data: they were scaled up when more data was available, reflecting their dependence on the GP's information capacity [52].

Table 5.1: State estimation results in the Simulated Pendulum Domain. Each row reports the performance of a different method under one of four data shift conditions: *both models*, *dynamics only*, *observation only*, and *none*. GP-ZKF consistently achieves the highest inclusion rates, indicating strong estimation consistency, while maintaining competitive accuracy and low computation time.

Data Shift	Method	Avg. RMSE $(\theta, \dot{\theta})$	Inclusion Rate	Avg. Radius $(\theta, \dot{\theta})$	Avg. Computation Time
		(°, °/sec)	(%)	(°, °/sec)	(sec)
Both models	GP-EKF	20.2, 37.5	5	1.7, 9.2	0.003
	GP-UKF	7.4, 15.6	38	10.0, 32.1	0.009
	GP-PF	90.5, 64.4	28	27.0, 54.6	1.459
	GP-ZKF	16.4, 20.4	**83**	46.7, 131.4	0.004
Dynamics only	GP-EKF	15.3, 22.1	0.17	0.2, 0.4	0.003
	GP-UKF	2.2, 10.8	27	1.6, 17.8	0.009
	GP-PF	69.6, 46.9	16	11.8, 25.5	1.500
	GP-ZKF	16.4, 21.5	**88**	44.6, 72.5	0.004
Observation only	GP-EKF	15.6, 21.7	4	1.4, 2.0	0.003
	GP-UKF	25.4, 31.2	27	35.9, 40.6	0.009
	GP-PF	222.7, 157.4	18	82.9, 100.2	1.544
	GP-ZKF	15.3, 18.6	**87**	47.9, 126.0	0.004
None	GP-EKF	15.4, 20.4	0.00	0.2, 0.4	0.003
	GP-UKF	2.2, 7.8	18	1.5, 5.1	0.009
	GP-PF	161.3, 126.7	10	47.0, 47.8	1.573
	GP-ZKF	16.6, 19.7	**92**	47.1, 72.8	0.004

Results

Figure 5.3 illustrates how the zonotopic estimates from GP-ZKF evolve over time in the *Shift None* condition. Each green zonotope, $\widehat{\mathcal{X}}_t$, reliably contains the true state, highlighting the estimator's consistency. The figure also shows how the zonotope volumes grow during open-loop dynamics propagation and shrink when informative measurements are incorporated, demonstrating the adaptive nature of the estimator.

Table 5.1 provides the quantitative results across different training and test distribution shift conditions. GP-EKF, which relies on linearization, suffered from poor performance due to the system's nonlinearity, yielding low inclusion rates and overly confident (small-radius) estimates. GP-UKF, which avoids linearization, performed better in both accuracy and consistency, but its estimates were still less reliable than GP-ZKF. GP-PF exhibited large RMSEs and only moderate inclusion, possibly due to the narrow variance of GP posteriors and particle impoverishment.

In contrast, GP-ZKF consistently achieved the highest inclusion rates across all conditions. These results empirically support its theoretical consistency guarantee (theorem 5.4.1). While GP-ZKF estimates are more conservative, resulting in larger average radii, this conservativeness appropriately reflects both epistemic and aleatoric uncertainty in the models.

Admittedly, GP-ZKF is more conservative than the others, resulting in larger radii. We argue that GP-ZKF's conservativeness actually scales appropriately with the domain, as we will next demonstrate its low conservativeness in the dressing domain.

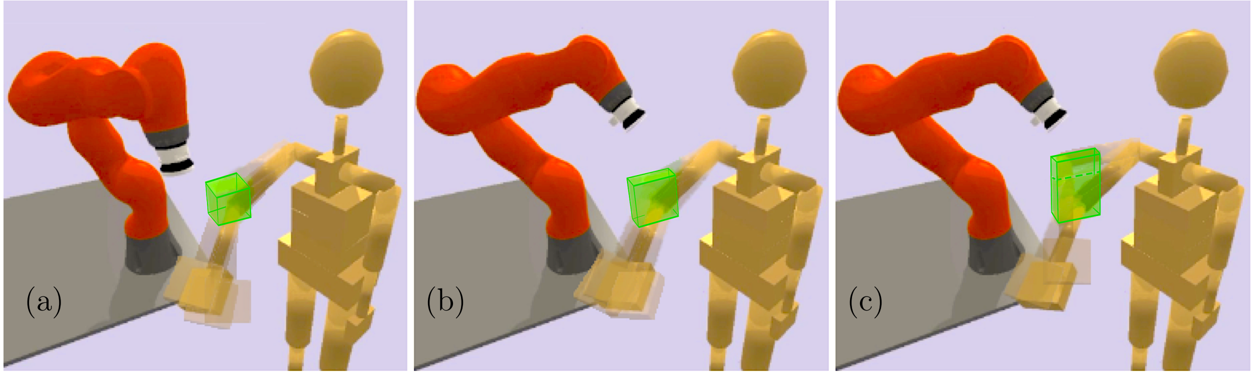


Figure 5.4: Keyframes of the robot-assisted dressing task (cloth omitted for clarity). The simulated robot and human motions are reconstructed from real-world data (fig. 5.1), as described in section 5.5.2. Each keyframe visualizes several ground-truth human arm poses, along with their corresponding zonotopic estimates of the elbow position. For visualization, each 3D zonotope produced by GP-ZKF is outer-approximated by an axis-aligned bounding box, shown as a green box.

5.5.2 Robot-Assisted Dressing Domain

Experiment Design

We evaluated GP-ZKF in a robot-assisted dressing task, where a robot arm dresses a long-sleeved jacket onto a human arm (fig. 5.1). The goal is to estimate the human elbow position, which is visually occluded by the garment [276]. All methods were evaluated offline using data collected from real-world human-robot interactions.

During data collection, the human moved their arm naturally. An Xsens motion capture system, unaffected by visual occlusion, tracked the human arm configuration. The robot executed a predefined dressing controller to move from the human hand to the elbow, and then to the shoulder. The robot was position-controlled using KUKA's impedance mode, providing some compliance during the interaction.

The dataset includes three initial arm configurations: *bend*, *lower*, and *straight*, with 17, 11, and 12 trajectories, respectively.

The state $x \in \mathbb{R}^3$ is defined as the 3D position of the human elbow. The control input $u \in \mathbb{R}^9$ includes the 3D positions of the human hand and shoulder, as well as the robot end effector. The known dynamics model is simply $f(x, u) = x$. The noise variances for w and v are automatically identified during GP training.

In our setup, the shoulder region of the garment is rigidly attached to the robot gripper. During dressing, the circular opening at the shoulder of the garment gradually slides along the human arm toward the human shoulder. At any given moment, this shoulder opening encloses a specific segment of the human arm. The 3D center position of this opening is informative about which part of the arm is currently inside the garment, making it a valuable signal for tracking dressing progress. We define this center position of the shoulder opening as the measurement $y \in \mathbb{R}^3$.

Table 5.2: State estimation results in the Robot-Assisted Dressing Domain. Each row shows the performance of a different method under one of three initial arm poses: *bend*, *lower*, or *straight*. GP-ZKF consistently achieves high inclusion rates, relatively small set estimates (Avg. Radius), competitive accuracy (Avg. RMSE), and low runtime (Avg. Computation Time), demonstrating its effectiveness in maintaining consistency, precision, accuracy, and computational efficiency.

Arm Pose	Method	Avg. RMSE $(\theta, \dot{\theta})$	Inclusion Rate	Avg. Radius $(\theta, \dot{\theta})$	Avg. Computation Time
		(cm,cm,cm)	(%)	(cm,cm,cm)	(sec)
Bend	GP-EKF	3.9, 3.3, 3.8	83	11, 19, 23	0.25
	GP-UKF	9.8, 3.0, 6.1	76	26, 10, 15	1.68
	GP-PF	2.4, 2.6, 3.0	73	6, 6, 7	29.93
	GP-ZKF	2.8, 3.4, 4.3	**88**	8, 8, 8	0.25
Lower	GP-EKF	5.2, 3.0, 6.9	94	24, 19, 22	0.11
	GP-UKF	9.2, 12.4, 13.6	80	27, 22, 33	0.71
	GP-PF	5.0, 2.7, 4.1	70	11, 8, 11	12.74
	GP-ZKF	4.6, 2.7, 4.7	**97**	12, 10, 12	0.11
Straight	GP-EKF	2.0, 3.0, 3.9	88	14, 15, 14	0.12
	GP-UKF	3.4, 3.8, 3.6	64	7, 7, 9	0.78
	GP-PF	2.5, 3.0, 2.8	62	6, 6, 6	13.95
	GP-ZKF	1.7, 2.8, 4.1	**92**	8, 8, 8	0.12

To estimate the shoulder opening's position, we use a 3D force signal [2], excluding torques, measured by an ATI force-torque sensor mounted at the robot's wrist. We apply a low-pass filter to the raw force signal and then transform it into an approximate shoulder opening position using a tether-inspired parametric model [294]. This model interprets the measured force as the tension in a virtual elastic tether connecting the shoulder opening (rigidly held by the robot) to the segment of the human arm currently enclosed by it. A large force suggests that the garment is being pulled tightly, meaning the arm segment is lagging behind the shoulder opening. Conversely, a small force implies the garment is loosely following the arm. Thus, the force gives indirect but informative clues about the location of the arm segment enclosed by the shoulder opening. We use this inferred shoulder opening center position as the measurement input to our estimator.

Results

Figure 5.4 illustrates representative keyframes from the dressing task, showing the zonotopic estimates produced by GP-ZKF. Each green box visualizes the outer-approximated 3D zonotope of the estimated elbow position at a specific time step. These boxes adapt in size depending on the underlying uncertainty, expanding when observations are ambiguous and

[2] Notably, we did not subtract the robot's commanded forces from the measured signals. This is because the robot moved slowly and the garment was lightweight, so in free motion (without human contact), the measured force was near zero. As a result, any nonzero force measured during dressing can be attributed to physical interaction between the garment and the human arm. In contrast, if the robot moved rapidly or the garment was heavier, inertial forces would need to be accounted for. In our setting, however, such correction was unnecessary.

shrinking when measurements provide clear information. Across the sequence, the green boxes consistently contain the ground-truth elbow positions, demonstrating GP-ZKF's ability to maintain accurate and consistent estimates throughout the dressing motion.

Quantitative results across all dressing conditions are presented in table 5.2. GP-ZKF consistently achieved the highest inclusion rates across all three conditions, indicating strong estimation consistency. Notably, its average set size (radius) was significantly smaller than GP-EKF and GP-UKF, demonstrating that GP-ZKF is not overly conservative despite maintaining consistency guarantees.

While GP-EKF and GP-UKF also achieved relatively high inclusion rates, they did so with much larger radius estimates. GP-PF produced smaller radii but at the cost of lower inclusion and significantly longer computation times. GP-ZKF offers a strong balance: it maintains reliable state estimates (consistency) while avoiding excessive conservativeness and remaining computationally efficient.

5.6 Conclusion

This chapter addressed the challenge of estimating a human's physical state when the robot is uncertain about its own learned models. We proposed GP-ZKF, a set-based estimation method that explicitly represents and respects epistemic uncertainty (due to scarce and noisy data) and aleatoric uncertainty (due to process and sensor noise). GP-ZKF combines Gaussian process learning with zonotope-based filtering to produce consistent state estimates, i.e., the true state remains inside the estimated set with high probability.

Our theoretical results showed that GP-ZKF is provably consistent and that, under relaxed assumptions, it reduces to a stochastic estimator (GP-EKF). Empirically, we validated our approach in two domains: a simulated inverted pendulum and a real-world robot-assisted dressing task. GP-ZKF outperformed standard GP-EKF, GP-UKF, and GP-PF methods in terms of consistency and robustness, without excessive conservativeness.

In the broader scope of this thesis, GP-ZKF contributes to answering the central question: *how should a robot behave when it is uncertain about the human?* Our answer in this chapter is: the robot should behave cautiously by estimating the human's physical state using uncertainty-aware methods that avoid overconfidence. This strategy forms a foundation for reliable human-robot interaction, especially when downstream robot actions depend critically on the estimated human state.

Chapter 6

Acting Safely under Uncertainty by Allowing Contact

"A ship in harbor is safe, but that is not what ships are built for."

—John A. Shedd

When robots physically assist humans, safety is a top priority, but overly cautious behavior can severely limit their usefulness. A key challenge is enabling robots to act to complete tasks efficiently while still ensuring human safety, especially when the robot is uncertain about how the human may move [177, 295]. This chapter addresses the central thesis question, *how should a robot behave when it is uncertain about the human?*, in the context of real-time physical interaction. I propose a new safety framework that allows the robot to act under uncertainty by relaxing traditional collision-avoidance constraints to permit low-impact contact.

In the broader context of this thesis, this chapter contributes to the answer of how robots should behave when uncertain about human behavior: by relaxing strict safety constraints to permit low-impact contact, robots can continue acting decisively without compromising safety.

In physical human-robot interaction, many existing safe motion generation approaches fall into two categories: *predictive* methods and *reactive* methods [296]. Predictive approaches allow a robot to anticipate human behavior while simultaneously planning collision-free motion [50]. Through anticipation, such approaches enable the robot to safely and effectively collaborate with humans [56, 130, 297, 298], but such safety and effectiveness heavily rely on high-quality predictive models of human behavior. By contrast, reactive approaches forgo modeling human behavior, but instead enable a robot to detect collisions in real time [50, 299, 300] and react compliantly to human behavior by ensuring reduced contact forces [301, 302].

Many prior works in safe physical human-robot interaction have integrated these two approaches *sequentially* [50, 296], with a robot first employing motion planners to find paths and then using compliant controllers for execution. However, each approach separately optimizes behavior for its own particular goal (collision avoidance for planners and contact force reduction for compliant controllers) rather than a goal *jointly held* by both approaches, ultimately exposing the weaknesses of each. First, most planners don't incorporate the fact that a compliant controller is employed to reduce contact forces in the event of a collision. As a result, planners tend to be very conservative, attempting to avoid collisions entirely.

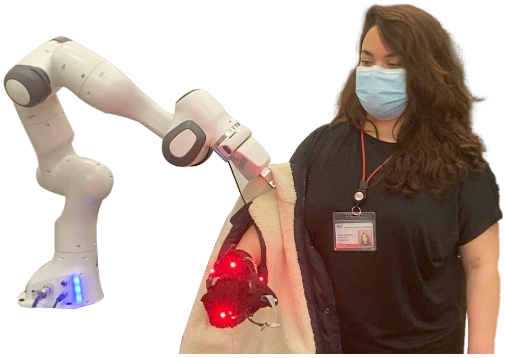


Figure 6.1: During robot-assisted dressing, the robot must remain physically close to the human arm to ensure human comfort due to the limited size of the armhole. Robot motion planners optimizing for human safety, defined as collision avoidance, might cause the robot to freeze under uncertainty, stalling progress. This chapter redefines safety to allow either collision avoidance or low-impact contact, enabling the robot to complete the task efficiently without compromising safety.

Conservative behavior can ensure safety, but could also worsen task performance or even unnecessarily freeze the robot in place [49]. Robot respecting its uncertainties about the future human behavior by avoiding regions potentially reachable by the human is necessary to ensure safety in many cases [155, 161, 297], but would exacerbate this issue of over-conservativeness. Consider the case of robot-assisted dressing shown in fig. 6.1: avoiding the (uncertain) human arm during task execution is nearly impossible, preventing a safe planner from making progress. Second, compliant controllers are usually unaware of the robot's high-level plans, making it challenging to adapt stiffness profiles in order to properly balance safety and task performance.

This chapter proposes a safe planner for integrating predictive and reactive approaches *jointly* within a framework, in order to reduce system conservativeness while maintaining safety. Human physical safety in predictive and reactive approaches were previously defined as collision avoidance [303–305] and contact force reduction [301, 302], respectively. By incorporating both definitions, I redefine safety in the context of human-aware motion planning. My new definition is two-pronged: *collision avoidance or safe impact in the event of a collision* [51]. This two-pronged definition captures the strengths of both predictive and reactive methods and enables the robot to act even when uncertainty is high. I formalize this concept within a learning-based Model Predictive Control (MPC) framework [52], which learns a model of human motion online, maintains uncertainty estimates, and plans safe actions accordingly. This integration inherits the theoretical safety guarantees of the underlying MPC algorithm, ensuring that the robot satisfies the two-pronged safety definition with high probability, even under uncertainty about future human behavior.

This safety-aware planner is evaluated both in simulation and in a real-world dressing task. Empirical results show that compared to standard MPC with collision-avoidance-only safety, the two-pronged safety constraint achieves significantly faster task completion, while maintaining safety guarantees.

In this chapter, I make two contributions:

- I redefine human physical safety for interactive manipulation as a two-pronged constraint: either avoid contact or ensure low-impact contact.
- I integrate this definition into a learning-based MPC algorithm that guarantees safety under uncertainty with high probability.

Section 6.1 formulates the motion planning problem under uncertainty. Section 6.2 presents the new safety definition. Section 6.3 introduces the learning-based MPC framework [52]. Section 6.4 describes how safety is encoded as a constraint in the MPC. Sections 6.5 and 6.6 report empirical results in simulation and physical robot experiments.

6.1 Problem Definition

In this chapter, we ignore the human and robot kinematics by making the following assumption:

Assumption 6.1.1. *The robot and human are represented as point masses in Cartesian space.*

A broad range of tasks can be modeled by this representation, including handover and space-sharing tasks [55, 306], where only the robot end-effector and human hand are modeled. As depicted in section 6.5.2, this representation enables a robot to dress a human arm (under certain task simplifications).

To reason about collision avoidance and safe impact, the human-robot system's state is designed to contain both the human and robot positions and velocities. Formally, let $p^R \in \mathbb{R}^3$, $v^R \subset \mathbb{R}^3$, $u \in \mathbb{R}^{d_u}$, and $m^R \in \mathbb{R}$ denote the robot's position, velocity, control, and mass, respectively. Similarly, let $p^H \in \mathbb{R}^3$, $v^H \in \mathbb{R}^3$, and $m^H \in \mathbb{R}$ denote the human position, velocity, and mass, respectively. A human-robot (joint) state is defined as the tuple (p^H, v^H, p^R, v^R).

6.1.1 Human Behavior

Before interacting with the human, the robot does not know how the human would move during interaction. Instead, the robot will interact with the human, collect data about the human movement, train a human dynamics model, and use that model to plan a safe robot trajectory. This chapter takes a "black-box approach" [13] to model the human movement as a first-order, deterministic [1], and discrete-time dynamical system, formulated as follows:

$$p_{t+1}^H = f^H(p_t^H, p_t^R) =: p_t^H + g(p_t^H, p_t^R), \tag{6.1}$$

$$v_{t+1}^H = \frac{1}{h}(p_{t+1}^H - p_t^H) = \frac{1}{h}\, g(p_t^H, p_t^R), \tag{6.2}$$

[1] Note that the system assumes that the human behavior is deterministic, which seems unrealistic. In fact, it is not hard to extend this chapter to stochastic human dynamics if the robot is assumed to still have access to a safe recovery controller, as introduced later in section 6.1.3 (This point is also mentioned by Koller et al. [52, Footnote 2]).

where the potentially nonlinear functions f and g are unknown. The human velocity, v^H, in eq. (6.2), is approximated as the rate of change of position, where h is a specified hyperparameter indicating the duration of a time-step. The initial condition, v_1^H, is assumed to be given by measurements, as it cannot be computed by eq. (6.2). This human dynamical system formulated above can be used to approximate low-level human movements in human-robot interactive tasks, such as reaching toward a goal [155, 307], which are usually embedded in complicated human behaviors as primitives.

6.1.2 Robot Dynamics

The robot dynamics is modeled as a known, deterministic, discrete-time system:

$$(p_{t+1}^R, v_{t+1}^R) = f^R(p_t^R, v_t^R, u_{t+1}). \tag{6.3}$$

A robot controller π is defined as a function: $u_{t+1} = \pi(p_t^H, v_t^H, p_t^R, v_t^R)$. Let f_π be a function describing the closed-loop human-robot system induced by the controller π. Then, combining the robot dynamics f^R, the human dynamics f^H and g (defined in eqs. (6.1) and (6.2)), and the robot controller π yields the closed-loop system $(p_{t+1}^H, v_{t+1}^H, p_{t+1}^R, v_{t+1}^R) = f_\pi(p_t^H, v_t^H, p_t^R, v_t^R)$.

6.1.3 Safe Recovery Controller

To ensure human safety, it is important for the robot to leverage its control authority to keep the human-robot state within the *safe set*. We denote the safe set by $\mathcal{S} := \left\{(p^H, v^H, p^R, v^R)\colon \text{human is safe}\right\}$, which will be formally defined later in eq. (6.7). The key challenge is that the robot does not know the human movement in advance, and has to rely on the human model learned from data. The noise and scarcity in the data would lead to errors in such learned models, which would potentially lead to the robot's unsafe actions. To mitigate this, we grant the robot a *safe recovery controller*, denoted by π_{red}, that can always keep the human safe. There are various ways of implementing such a controller. In this chapter, I implement π_{red} as a robot emergency stop. Intuitively, if the robot chooses to run it, then the human is assumed to always remain safe. However, note that this controller will not contribute to making progress in any interaction tasks, but only keep the human safe. So it is the robot's algorithm's job to decide when to trigger this controller, as described later in section 6.3.

Given a recovery controller π_{red}, we define the corresponding *recovery set* as the set of the human-robot states where executing π_{red} will keep the human safe. Given my implementation of π_{red} as an emergency stop, the recovery set is then defined as

$$\mathcal{S}_{\text{red}} := \left\{(p^H, v^H, p^R, v^R) \in \mathcal{S}\colon v^R = 0\right\} = \mathcal{S} \cap \left\{(p^H, v^H, p^R, v^R)\colon v^R = 0\right\}.$$

Accordingly, the assumption regarding π_{red} can be state formally as follows:

Assumption 6.1.2. *The system is given π_{red} with $\mathcal{S}_{red}$, such that:*

$$\forall \tau = 1, 2, \dots : \forall (p_\tau^H, v_\tau^H, p_\tau^R, v_\tau^R) \in \mathcal{S}_{red} \implies \forall t \geq \tau\colon (p_t^H, v_t^H, p_t^R, v_t^R) \in \mathcal{S},$$

where $\forall t \geq \tau\colon (p_{t+1}^H, v_{t+1}^H, p_{t+1}^R, v_{t+1}^R) = f_{\pi_{red}}(p_t^H, v_t^H, p_t^R, v_t^R)$ and $f_{\pi_{red}}$ denote the corresponding closed-loop system under π_{red}.

This assumption states that if the human-robot state is initially within $\mathcal{S}_{\text{red}}$, then if the robot starts to execute π_{red}, the human-robot state will always remain within the safe set $\mathcal{S}$. Intuitively, when $v^R = 0$ and the human is safe, if the robot activates the emergency stop π_{red}, then immediately from now on, the human will remain safe and never injure themselves.

6.1.4 Problem Statement

This chapter's goal is to design a robot controller to complete some interaction task, specified by a given objective function, while ensuring human physical safety throughout the operation time. However, since the robot's model of the human, learned from data, is almost never perfect, it is generally impossible for the robot controller to guarantee that the human-robot state always stays within the safe set [52]. Instead, we slightly relax this goal to *safety with a high probability* (or δ-safety), which is formally defined as follows:

Definition 6.1.3 (δ-safety (definition 3 in Koller et al. [52]))**.** *Let π be a robot controller with the closed-loop system denoted by f_π. Given $\delta \in (0, 1)$ and that the initial state $(p_1^H, v_1^H, p_1^R, v_1^R) \in \mathcal{S}_{red}$, the system is δ-safe under π iff*

$$\mathbb{P}\left[\forall t = 0, 1, \ldots : (p_t^H, v_t^H, p_t^R, v_t^R) \in \mathcal{S}\right] \geq 1 - \delta,$$

where $\forall t \in \mathbb{N}\colon (p_t^H, v_t^H, p_t^R, v_t^R) = f_\pi(p_{t-1}^H, v_{t-1}^H, p_{t-1}^R, v_{t-1}^R)$.

Intuitively, δ-safety indicates that with a high probability, the human will remain safe, given that the human-robot system is initially within the recovery set $\mathcal{S}_{\text{red}}$. Note that δ-safety is defined *jointly* throughout the operating time, rather than *per time step* [52]. Additionally, the definition does not involve any terminal time, implying that δ-safety is *independent* of the duration of operation [52].

Now we state the problem for this chapter:

Problem Statement. *Design a controller, π, to complete a given interaction task while guaranteeing δ-safety.*

6.2 Human Safety as Collision Avoidance or Safe Impact

In this section, we will expand our definition of safe set $\mathcal{S}$ introduced in section 6.1.3 to encode the concept of collision avoidance or safe impact.

Lasota, Fong, Shah, et al. [51] defined safety as (1) collision avoidance whenever possible, and (2) safe impact when collisions are required or unavoidable. We see collision avoidance and safe impact as two approaches to ensuring safety; integrating these approaches allows robot planners greater freedom to find less conservative and more efficient solutions without sacrificing safety. Hence, we define safety as collision avoidance or safe impact.

Next, we will first formulate collision avoidance and then safe impact. Finally, we will combine both formulations to formally define the safe set $\mathcal{S}$.

6.2.1 Collision Avoidance

Given a fixed human position p^H, we define the *Collision Avoidance (CA) set* as $\mathcal{S}_{\text{CA}}(p^H) \coloneqq \{(p^R, v^R)\colon \|p^R - p^H\|_2 > \epsilon\}$, with $\|p^R - p^H\|_2$ referring to the Euclidean distance between p^R and p^H and $\epsilon > 0$ denotes a specified distance threshold. Intuitively, ensuring that the robot state stays within the Collision Avoidance set would encourage the robot position to be far from the human to avoid potential collisions.

6.2.2 Safe Impact

Heinzmann and Zelinsky [308] formulated a constraint on a robot's position and velocity to ensure safe human-robot impact during collisions, while assuming that the human remains static. In this section, we adapt that constraint to the case in which the human is moving. Given the point-mass assumption (assumption 6.1.1) for both human and robot, the safe impact constraint would depend on the human's and robot's velocities.

We consider the case in which two general bodies collide. By assuming that the collision occurs within an infinitesimally small period of time $\Delta t \to 0$, we can treat both bodies during impact as rigid bodies (according to Wittenburg [309, CH 6.1]). With a slight abuse of notation, let p^H, v^H, p^R, and v^R denote the human's position, velocity, robot's position, and velocity immediately before a collision, respectively. Let Δv^H and Δv^R denote the changes to v^H and v^R immediately following the collision, respectively. As $\Delta t \to 0$, during Δt, both p^H and p^R don't change, while v^H and v^R remain finite. By Wittenburg [309, CH 6.1], the impulse, defined by $\hat{F} = \lim_{\Delta t \to 0} \int_t^{t+\Delta t} F(s)ds$, remains finite, where F is the impulsive force that tends toward infinity as $\Delta t \to 0$.

According to Walker [310, eq. (8)] or Wittenburg [309, eq. (6.9)], we have the following kinematic relationship:

$$[(v^H + \Delta v^H) - (v^R + \Delta v^R)]^\top \mathbf{n} = -e(v^H - v^R)^\top \mathbf{n} \tag{6.4}$$

where $\mathbf{n}$ is the unit normal vector to the common tangent plane at the point of collision [309]. The parameter $e \in [0, 1]$ denotes the *coefficient of restitution*. The value of e is 0 for purely plastic collisions and 1 for purely elastic collisions [310].

Under our representation where both bodies are point masses (assumption 6.1.1), by following Walker [310, sec. II(B)], we obtain that $\Delta v^H = -\hat{F}/m^H$ and $\Delta v^R = \hat{F}/m^R$. Here, we follow the convention that F is the force exerted by the human and applied to the robot. By plugging both expressions into eq. (6.4), we get the following:

$$\hat{F}^{mag} = \frac{-(e+1)((v^R)^\top - (v^H)^\top) \cdot \mathbf{n}}{\frac{1}{m^R} + \frac{1}{m^H}} \tag{6.5}$$

where $\hat{F}^{mag}$ denotes the magnitude of $\hat{F}$, i.e., $\hat{F} = \hat{F}^{mag}\mathbf{n}$. In fact, eq. (6.5) is equivalent to Wittenburg [309, eq. (6.16)] when applied to point masses.

We further adopt the assumption made by Heinzmann and Zelinsky [308] that at the moment of contact, there is sufficient friction to align $\hat{F}$ and $(v^R - v^H)$. Given this assumption, we have $\mathbf{n} = -(v^R - v^H) \,/\, \|v^R - v^H\|_2$. By plugging this expression into eq. (6.5), we arrive at the following:

$$\hat{F}_e^{mag} = \frac{(e+1)\|v^R - v^H\|_2}{\frac{1}{m^R} + \frac{1}{m^H}} =: \Omega(v^H, v^R) \tag{6.6}$$

where $\hat{F}_e^{mag}$ is the *effective impact force* magnitude [308] and we refer to this function using a new notation $\Omega(v^H, v^R)$.

Now we define *impact potential* as the maximum impact force that a robot can create in a collision with a human [308]. It is a scalar value that provides an upper limit for any impact between the robot and human. According to Heinzmann and Zelinsky [308, eq. (11)], impact potential is computed as the maximum effective impact force among all points of collision on the surfaces of the robot and human. Given our point-mass assumption, there is only one possible point of collision. Hence, our impact potential is equivalent to $\hat{F}_e^{mag}$.

Given a fixed human velocity v^H, we define the *Safe Impact (SI) set* as $\mathcal{S}_{\text{SI}}(v^H) := \left\{(p^R, v^R) \colon \Omega(v^H, v^R) \le \Omega_{\max}\right\}$, where Ω computes the impact potential between the human and the robot, as defined in eq. (6.6), and $\Omega_{\max}$ denotes the specified maximum impact potential considered to be safe. Intuitively, ensuring that the robot state stays within the safe impact set would encourage the robot velocity to be similar to the human velocity, so even when they collide, the impact will be low.

6.2.3 Safe Set Defined as Collision Safe or Safe Impact

So far, given a human position p^H, we have defined the collision avoidance set, $\mathcal{S}_{\text{CA}}(p^H)$. Given a human velocity v^H, we have defined the safe impact set, $\mathcal{S}_{\text{SI}}(v^H)$. According to my safety definition, the human is safe if the system either remains collision-free or ensures safe impact during collisions. Accordingly, we define "safety" as follows:

$$(p^R, v^R) \;\in\; \mathcal{S}_{\text{CA}}(p^H) \bigcup \left[(\mathcal{S}_{\text{CA}}(p^H))^{\mathsf{c}} \bigcap \mathcal{S}_{\text{SI}}(v^H)\right] \;=\; \mathcal{S}_{\text{CA}}(p^H) \bigcup \mathcal{S}_{\text{SI}}(v^H),$$

where the superscript c denotes the set complement operator. The $\cap$ emphasizes that safe impact is used only in the event of a collision. This above equality implies that "CA or (SI during collision)" is equivalent to "CA or SI." Hence, we formally define the *safe set* $\mathcal{S}$ as follows:

$$\mathcal{S} := \left\{(p^H, v^H, p^R, v^R) \colon (p^R, v^R) \in \mathcal{S}_{\text{CA}}(p^H) \cup \mathcal{S}_{\text{SI}}(v^H)\right\}. \tag{6.7}$$

6.3 Model Predictive Control Algorithm

We design the robot controller by adopting the learning based Model Predictive Control (MPC) algorithm from Koller et al. [52, Algorithm 1]. This MPC algorithm iteratively solves a trajectory optimization in a receding horizon fashion, and uses the recovery controller, π_{red} (defined in section 6.1.3) in case the trajectory optimization cannot find feasible solutions. The pseudo-code is presented in algorithm 2. In particular, at each time step $t = 1, 2, \ldots$, if the trajectory optimization finds a feasible finite-length trajectory, denoted by $\Pi_t := \{u_{t,1}, \ldots, u_{t,T}\}$, then the robot will execute the first control, $u_{t,1}$. In the next time step, $t+1$, if no feasible solutions can be found, the algorithm will reuse the trajectory from the previous time step t, denoted by Π_t, by shifting Π_t in a receding horizon manner and appending π_{red} to the end. In this way, the algorithm obtains a new trajectory, $\Pi_{t+1} = \{u_{t,2}, \ldots, u_{t,T}, \pi_{\text{red}}\}$, and will execute the first control, $u_{t,2}$. The algorithm will then repeat this process.

Algorithm 2 Learning-based Model Predictive Control (MPC) algorithm ([52, Algorithm 1])

```
 1: Input: Recovery controller $\pi_{\text{red}}$, human dynamics model learned via Gaussian Process,
    default trajectory $\Pi_0 \coloneqq \{u_{0,1} = \pi_{\text{red}}, u_{0,2} = \pi_{\text{red}}, \ldots, u_{0,T} = \pi_{\text{red}}\}$.
 2: for each time step $t = 1, 2, \ldots$ do
 3:     $\Pi_t \coloneqq \{u_{t,1}, \ldots, u_{t,T}\} \leftarrow$ solve Trajectory Optimization (eq. (6.8))
 4:     if infeasible then
 5:         $\Pi_t \leftarrow \{u_{t-1,2}, \ldots, u_{t-1,T}, \pi_{\text{red}}\}$
 6:     end if
 7:     Execute the first control in $\Pi_t$ and observe $p_t^R$, $v_t^R$, $p_t^H$, and $p_{t+1}^H$
 8:     Compute $v_t^H \leftarrow (p_{t+1}^H - p_t^H) \ / \ h$, based on eq. (6.2)
 9:     Update the human dynamics model using the new observation $(p_t^R, p_t^H, p_{t+1}^H)$
10: end for
```

The key step in algorithm 2 is the trajectory optimization in line 3, which will be formalized next in section 6.3.1. Intuitively, the trajectory optimization searches for a finite-length trajectory that optimizes the task objective while ensuring human safety along the trajectory. In addition, it also ensures that at the end of the trajectory, the robot is parked within its recovery set $\mathcal{S}_{\text{red}}$ (introduced in section 6.1.3). In other words, at the end of the trajectory, the robot velocity is 0. In this way, within the MPC, if a feasible trajectory is found, then this trajectory ensures safety. If no feasible trajectories are found, then by executing the recovery controller, π_{red}, the robot can still ensure human safety by assumption 6.1.2. As a result, Koller et al. [52] proves that the MPC is δ-safe (see Koller et al. [52, Theorem 8]).

6.3.1 Trajectory Optimization

The key step in the MPC formulated in algorithm 2 is the trajectory optimization in line 3. We adapt the nonlinear trajectory optimization presented in Koller et al. [52, Eq. (43)] to our human-robot scenario as follows:

$$\text{Maximize} \quad J \tag{6.8}$$

$$\text{Subject to}$$

$$\forall t = 1, 2, \ldots, T-1\colon (p_{t+1}^R, v_{t+1}^R) = f^R\left(p_t^R, v_t^R, u_t\right) \tag{6.8a}$$

$$v_T^R = 0 \tag{6.8b}$$

$$\forall t = 1, 2, \ldots, T-1\colon \mathcal{E}_{t+1}^{p^H} = \text{RobustPredict}\left(\mathcal{E}_t^{p^H}, p_t^R\right) \tag{6.8c}$$

$$\mathcal{E}_{t+1}^{v^H} = \text{RobustPredict}\left(\mathcal{E}_t^{p^H}, p_t^R\right)$$

$$\forall t = 1, 2, \ldots, T, \ \forall p^H \in \mathcal{E}_t^{p^H}, \ \forall v^H \in \mathcal{E}_t^{v^H}\colon \left(p^H, v^H, p_t^R, v_t^R\right) \in \mathcal{S} \tag{6.8d}$$

This trajectory optimization contains an objective function, denoted by J, and a set of constraints. The objective function captures the interactive task, such as moving to the human shoulder position in robot-assisted dressing used in section 6.5.2. The constraint in eq. (6.8a) specifies the robot dynamics, as defined in section 6.1.2. The constraint eq. (6.8b) ensures that at the end of the trajectory, the robot velocity is 0, or the robot has arrived within the recovery set $\mathcal{S}_{\text{red}}$. As discussed earlier, this constraint allows the MPC to leverage the recovery controller, π_{red}, to guarantee safety even if the trajectory optimization finds no feasible trajectories.

The constraint in eq. (6.8c) implements the robot's prediction about the future human behavior. As introduced next in section 6.3.2, the robot will use Gaussian Processes to learn the human dynamics, defined in eqs. (6.1) and (6.2). To mitigate potential errors in the learned models due to data noise and scarcity, the robot has to quantify its uncertainty about the true human dynamics and plan motions to ensure human safety under uncertainty. The trajectory optimization employs geometric sets, such as ellipsoids, to bound the robot's uncertainty about future human states. In particular, $\mathcal{E}_t^{p^H}$ and $\mathcal{E}_t^{v^H}$ bound the uncertain possible human positions and velocities, respectively, at time t. The constraint in eq. (6.8c) implements how such ellipsoids propagate forward in time, given the robot state, by a function called "RobustPrediction", which will be discussed next in section 6.3.2. Koller et al. [52, Corollary 7] shows that under certain assumptions, these ellipsoids are well-calibrated in the sense that they contain the true human state with a high probability.

The constraint in eq. (6.8d) specifies that for all the possible human states within the ellipsoids, the robot trajectory will ensure human safety. One key technical challenge is to apply the safe set, which is defined for a single human state, as formulated in section 6.2, to all the human states bounded by ellipsoids. Another technical challenge is to formulate the disjunction "or" in the definition of "collision avoidance or safe impact" in the nonlinear optimization. Both these challenges will be addressed later in section 6.4.

6.3.2 Robust Human Motion Prediction

The robot learns a human dynamics model from past observations of human movement. Let $\{\hat{g}_i\}_{i=1}^n$ denote a set of n past measurements of the unknown function g, formulated in eq. (6.1), at the input locations $\{(p_i^H, p_i^R)\}_{i=1}^n$. We assume that the measurements, along all dimensions, are corrupted by an *i.i.d.* Gaussian noise. Formally, for each data point $i = 1, 2, \ldots, n$, for each dimension $j = 1, 2, 3$, we have that $[\hat{g}_i]_j = [g(p_i^H, p_i^R)]_j + w$ with $w \sim \mathcal{N}(0, \lambda_w^2)$, where $[z]_j$ denotes the j-th dimension of the vector z.

To quantify the robot's uncertainty about the learned function, by following Koller et al. [52], the robot learns the multi-output function g via Gaussian Processes (GP). We denote a Gaussian process by $GP(m, k)$, where m is the prior mean function and k is the covariance (or kernel) function. For details about Gaussian Process, please refer to section 3.3.

It is challenging to ensure safety for arbitrary human behavior, because the speed of human motion can be an order of magnitude faster than that of robots [50], and a moving human can inflict arbitrary collision forces upon robots [308]. In this work, we assume that the human dynamics captured by the function g is smooth. Formally,

Assumption 6.3.1 (Assumption 1 in Koller et al. [311]). *The function g' has a bounded reproducing kernel Hilbert space norm induced by a continuously differentiable kernel k, i.e., $\|g'\|_k \leq B$.*

Here, g' denotes the surrogate single-output function, formulated in section 3.3. The reproducing kernel Hilbert space norm is also briefly introduced in section 3.3. This assumption implies that the human behavior, represented by g, remains smooth not only in collision-free cases, but also in the cases where collisions occur. Despite being strong, this above assumption is valid in practice when the robot is operating at a low speed and with a compliant controller. Compliant controllers [300, 312, 313] can use control authority to render a robot to behave like a virtual mass-spring-damper system that reacts to contact forces in a compliant manner.

With assumption 6.3.1, by lemma 3.3.1, the robot can use its GP predictions to build reliable confidence intervals for g at human states that are unseen in the past, which allows the system to robustly predict the future human states. In particular, at each time step t, given the current p^R_t and p^H_t, the system can construct an ellipsoid, denoted by $\mathcal{E}^{p^H}_{t+1}$ to bound all the reachable human position, p^H_{t+1}, at time $t+1$. Similarly, the system can also construct an ellipsoid $\mathcal{E}^{v^H}_{t+1}$ to bound all the reachable v^H_{t+1}. Furthermore, given the current p^R_{t+1} and the uncertain human position within $\mathcal{E}^{p^H}_{t+1}$, the system can leverage the desired geometric properties of ellipsoids to construct new sets of ellipsoids, $\mathcal{E}^{p^H}_{t+2}$ and $\mathcal{E}^{v^H}_{t+2}$ to bound all the reachable states at time $t+2$. This forward-in-time propagation of ellipsoids is formally defined by eq. (6.8c). The derivation of the expressions for the ellipsoids and the propagation function follows Koller et al. [52, sec. V.A.2)]. With the propagation function, the system predicts the human future positions and velocities as two sequences of ellipsoids. With my assumption 6.3.1, I obtain the theoretical result stated in Koller et al. [52, Corollary 7] that these ellipsoids are guaranteed to contain the true human state with a high probability.

6.4 Safety Constraints for Collision Avoidance Or Safe Impact

In section 6.2, we formally define the safe set $\mathcal{S}$ as the set of human-robot states where either there is no collision or the impact is safe even when the human and robot collide. To ensure safety, we need to integrate the constraint that the human-robot state always stays within $\mathcal{S}$. However, this is nontrivial for two reasons. First, recall that the safety constraint formulated in section 6.2 specifies the set of safe robot states given a single human state (p^H, v^H). However, due to the robot's uncertainty about the true human dynamics, the trajectory optimization formulated in eq. (6.8c) has to predict future human motion as sets of states, represented by ellipsoids. As a result, the safety constraint formulated in section 6.2 cannot be directly integrated into the trajectory optimization eq. (6.8). Second, the safety constraint contains a disjunction "or". Directly integrating it into trajectory optimization would result in a mixed-integer program, potentially leading to large computation overhead. Next, sections 6.4.1 and 6.4.2 extends the collision avoidance and safe impact constraints, respectively, to be compatible with human predictions in the forms of ellipsoids. Then, section 6.4.3 reformulates the disjunction to avoid the need for solving mixed-integer programs. Finally, section 6.4.4 integrates the reformulated safety constraint back into the trajectory optimization that is initially formulated in eq. (6.8).

6.4.1 Collision Avoidance with Ellipsoidal Human Predictions

To ensure collision avoidance, as defined in section 6.2.1, the robot position, p^R, has to stay away from the human position, p^H. By assumption 6.1.1 that the robot is represented as a point mass, collision avoidance can be enforced by ensuring that the robot always stays outside the ellipsoids for human positions. Formally, given that p^R is bounded by the ellipsoid $\mathcal{E}^{p^H}$ parameterized by a center vector c^{p^H} and a shape matrix Q^{p^H}, the collision avoidance (CA) constraint can be formulated as follows:

$$\mathrm{Cst}_{\mathrm{CA}}(p^R, \mathcal{E}^{p^H}) = (p^R - c^{p^H})^\top (Q^{p^H})^{-1} (p^R - c^{p^H}) - 1 > 0 \tag{6.9}$$

This constraint can be extended to the cases where the robot is modeled as a sphere, rather than a point mass, by enlarging $\mathcal{E}^{p^H}$ and constraining the center of the robot to stay outside the enlarged ellipsoid [314].

6.4.2 Safe Impact with Ellipsoidal Human Predictions

To ensure safe impact, as formulated in section 6.2.2, the robot velocity, v^R, has to satisfy this following constraint: $\Omega(v^H, v^R) \leq \Omega_{\max}$, given the human velocity, v^H, where the impact potential Ω is defined in eq. (6.6). Given that the human velocity is bounded by the ellipsoid $\mathcal{E}^{v^H}$, the constraint can be rewrite as follows:

$$\forall v^H \in \mathcal{E}^{v^H} : \Omega(v^H, v^R) = \frac{(e+1)\left\|v^R - v^H\right\|_2}{\frac{1}{m^R} + \frac{1}{m^H}} \leq \Omega_{\max},$$

which can be compactly represented by the following constraint with a constant ρ:

$$\forall v^H \in \mathcal{E}^{v^H} : \ \left\|v^R - v^H\right\|_2 \leq \rho. \tag{6.10}$$

By the relation between ℓ^2-norm and ℓ^∞-norm, eq. (6.10) can be conservatively relaxed to the following constraint:

$$\forall v^H \in \mathcal{E}^{v^H}, \forall j \in \{1,2,3\} : \ \left|[v^R]_j - [v^H]_j\right| \leq \rho/\sqrt{3}, \tag{6.11}$$

where $[z]_j$ denotes the j-th coordinate of the vector z. Then, by stacking the constraint for each dimension, we equivalently rewrite eq. (6.11) as the following polytopic constraint:

$$\forall v^H \in \mathcal{E}^{v^H} : L\, v^H \leq L\, v^R + \left[\rho/\sqrt{3}\right]_{6\times 1} =: l. \tag{6.12}$$

Here, $L \coloneqq [-I_3, I_3]^\top \in \mathbb{R}^{6\times 3}$ where I_3 denotes the identity matrix of size 3. The notation $\left[\rho/\sqrt{3}\right]_{6\times 1}$ denote the vector $\left[\rho/\sqrt{3}, \rho/\sqrt{3}, \ldots, \rho/\sqrt{3}\right]^\top \in \mathbb{R}^6$. We also denote the right hand side of eq. (6.12) by $l \in \mathbb{R}^6$.

Following Koller et al. [52, eq. (41)], we enforce the Polytopic constraint in eq. (6.12) analytically to the ellipsoid $\mathcal{E}^{v^H}$ parameterized by a center vector c^{v^H} and a shape matrix Q^{v^H}. In particular, eq. (6.12) is equivalently reformulated as the following 6 individual 1-dimensional constraints:

$$\forall i \in \{1,\ldots,6\} : [L]_{i,\cdot} \cdot c^{v^H} + \sqrt{[L]_{i,\cdot}\, Q^{v^H}\, [L]_{i,\cdot}^\top} \ \leq \ [l]_i, \tag{6.13}$$

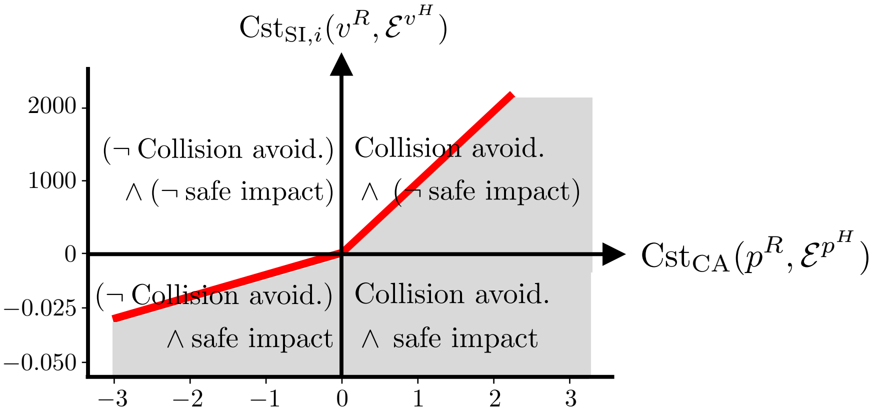


Figure 6.2: Given two ellipsoids for human position and velocity, denoted by $\mathcal{E}^{p^H}$ and $\mathcal{E}^{v^H}$, respectively, the safety constraint, defined as collision avoidance or safe impact is formulated in eq. (6.16). Given a fixed dimension $i \in \{1, \dots, 6\}$, this plot shows the constraint's feasible region by treating the function values, $\mathrm{Cst}_{\mathrm{CA}}(p^R, \mathcal{E}^{p^H})$ and $\mathrm{Cst}_{\mathrm{SI},i}(v^R, \mathcal{E}^{v^H})$, as X- and Y- axis, respectively. All pairs of $\left(\mathrm{Cst}_{\mathrm{CA}}(p^R, \mathcal{E}^{p^H}), \mathrm{Cst}_{\mathrm{SI},i}(v^R, \mathcal{E}^{v^H})\right)$ in the first quadrant indicate "collision avoidance and not safe impact". Here, note that "not safe impact" refers to the situation of "unsafe impacts during (hypothetical) collisions," which is not in conflict with "collision avoidance." All pairs in the union of the first, third, and fourth quadrants indicate "collision avoidance or safe impact," which implies safety according to our definition. All and only unsafe pairs are located within the second quadrant. We conservatively approximate the feasible region using the surrogate constraint: $\mathrm{Cst}_{\mathrm{SI},i}(v^R, \mathcal{E}^{v^H}) \leq \max\left(0.01\,\mathrm{Cst}_{\mathrm{CA}}(p^R, \mathcal{E}^{p^H}), 1000\,\mathrm{Cst}_{\mathrm{CA}}(p^R, \mathcal{E}^{p^H})\right)$, whose feasible region is plotted in gray. The corresponding equality is represented by the red line segments. Integrating this surrogate constraint, for each dimension $i \in \{1, \dots, 6\}$, into the trajectory optimization eq. (6.8) implies that the found trajectory ensures human safety, defined as collision avoidance or safe impact.

where, $[l]_i$ denotes the i-th dimension of the vector l and $[L]_{i,\cdot}$ denotes the i-th row of the matrix L. For notational shorthand, we rewrite the set of constraints, presented in eq. (6.13) as follows:

$$\forall i \in \{1, \dots, 6\} \colon \mathrm{Cst}_{\mathrm{SI},i}(v^R, \mathcal{E}^{v^H}) \leq 0, \tag{6.14}$$

where $\mathrm{Cst}_{\mathrm{SI},i}(v^R, \mathcal{E}^{v^H}) \coloneqq [L]_{i,\cdot} \cdot c^{v^H} + \sqrt{[L]_{i,\cdot}\; Q^{v^H}\; [L]_{i,\cdot}^{\top}} - [l]_i \in \mathbb{R}$.

6.4.3 Disjunction between Collision Avoidance and Safe Impact

In order to implement safety as "collision avoidance or safe impact," we must disjunctively combine the constraints for collision avoidance with those for safe impact. We combine the collision avoidance constraint in eq. (6.9) and the safe impact constraint in eq. (6.14) into the following disjunctive normal form:

$$\text{Cst}_{\text{CA}}(p^R, \mathcal{E}^{p^H}) > 0 \bigvee \left[\bigwedge_{i \in \{1,2,\ldots,6\}} \text{Cst}_{\text{SI},i}(v^R, \mathcal{E}^{v^H}) \leq 0 \right], \tag{6.15}$$

with function arguments omitted for notational convenience. Here, $\bigvee$ and $\bigwedge$ denote the operators of logical disjunction and conjunction, respectively. Equation (6.15) is then equivalently reformulated as the following conjunctive normal form:

$$\bigwedge_{i \in \{1,2,\ldots,6\}} \left[\text{Cst}_{\text{CA}}(p^R, \mathcal{E}^{p^H}) > 0 \bigvee \text{Cst}_{\text{SI},i}(v^R, \mathcal{E}^{v^H}) \leq 0 \right]. \tag{6.16}$$

Unfortunately, directly integrating the constraint eq. (6.16) into the trajectory optimization eq. (6.8) would make the optimization a mixed-integer program, leading to large computation overhead. Instead, we will conservatively approximate eq. (6.16) with a surrogate constraint, which eases the computation overhead for trajectory optimization.

To motivate the surrogate constraint, let's visualize the constraint eq. (6.16), given a particular dimension $i \in \{1, 2, \ldots, 6\}$, in a 2-dimension plane with X- and Y- axes representing Cst_{CA} and $\text{Cst}_{\text{SI},i}$, respectively. As shown in fig. 6.2, the safe region of this one-dimensional constraint is the union of the first, third, and fourth quadrants.

Inspired by the Rectified Linear Unit (ReLU) function, we conservatively approximate the constraint in eq. (6.16) for each dimension $i \in \{1, 2, \ldots, 6\}$ using the following *surrogate constraint*:

$$\bigwedge_{i \in \{1,2,\ldots,6\}} \left[\text{Cst}_{\text{SI},i}(v^R, \mathcal{E}^{v^H}) \;\leq\; \max\Big(\theta_1 \, \text{Cst}_{\text{CA}}(p^R, \mathcal{E}^{p^H}), \;\; \theta_2 \, \text{Cst}_{\text{CA}}(p^R, \mathcal{E}^{p^H}) \Big) \right], \tag{6.17}$$

with hyper-parameters $\theta_1 \geq 0$ and $\theta_2 \geq 0$. The feasible region of the surrogate constraint, in the case where $\theta_1 = 0.01$ and $\theta_2 = 1000$, is the gray region in fig. 6.2. Since the gray region is a strict subset of the union of the first, third, and fourth quadrants, satisfying the surrogate constraint is a sufficient but not necessary condition for satisfying the constraint eq. (6.16).

6.4.4 Integrating Safe Constraints into Trajectory Optimization

Next, we integrate the surrogate constraint eq. (6.17) into the eq. (6.8d) in the trajectory optimization. Recall that at the end of section 6.3.2, we have mentioned that the ellipsoids are guaranteed to contain the true future human state with a high probability. Now that the surrogate safety constraint eq. (6.17) implies that the robot state (p^R, v^R) ensures human safety given that the human state lies within the ellipsoids. Hence, the trajectory optimization will generate trajectories that ensure human safety, defined as collision avoidance or safe impact, with a high probability. By integrating this trajectory optimization within the MPC algorithm (algorithm 2), the MPC algorithm ensures δ-safety (see Koller et al. [52, Theorem 8]).

6.5 Empirical Experiment

Our goal is to empirically evaluate the improvements to task efficiency resulting from the definition of safety as collision avoidance or safe impact. We benchmarked our MPC with safety defined as collision avoidance or safe impact (henceforth referred to as *CASI*), along with its variation with safety defined only as collision avoidance (henceforth referred to as *CA*). We evaluated these algorithm variations using two tasks: a simulated 2D goal-reaching task and a real-world robot-assisted dressing task. Our hypothesis was that CASI would result in a smaller number of iterations necessary for the completion of both tasks.

We considered a realistic scenario in which a system designer could first collect non-interactive data of the human performing a task without the robot, then allow the robot to operate around the human in order to collect more interactive data. Accordingly, we collected a small, non-interactive dataset by fixing the robot at a predefined location, p^R_1, and letting the human (simulated or real) roll out their policy. Then, we ran the MPC with the trained GP until the robot reached the goal, where only the GP's posterior (and not its hyper-parameters) was updated between iterations. During our experiments, we expected our robot to act safely and efficiently under the uncertainty caused by the distribution shift from the initial training scenarios, where $p^R = p^R_1$, to the testing scenarios, where p^R moved.

For both tasks, we implemented π_{rec} as a safety stop, as described in section 6.1.3. In keeping with previous work [52, 315], since the theoretical confidence intervals for the GP model are conservative, we chose to set lemma 3.3.1's scaling parameter $\beta = 2$ to enable efficient task completion under uncertainty.

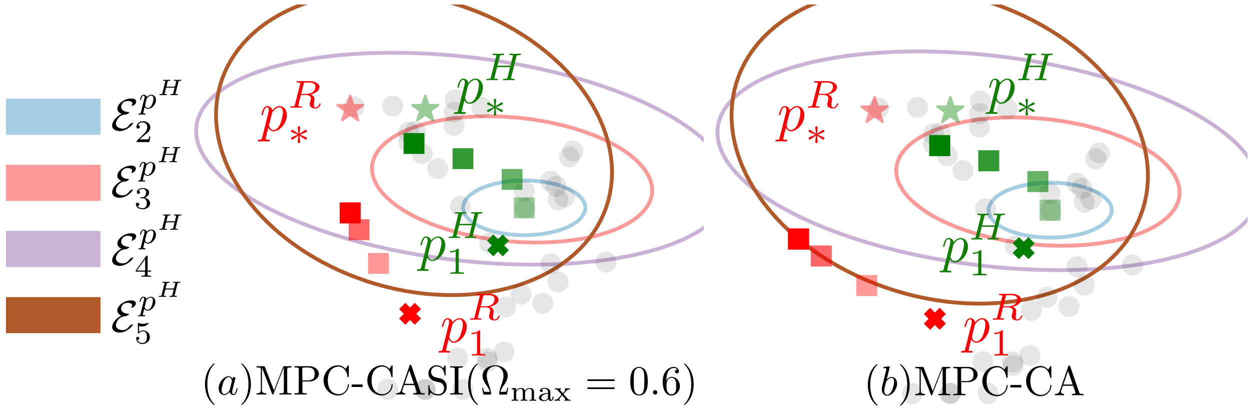
6.5.1 2D Goal-Reaching

We simulated human behavior within a 2D environment by online optimizing a continuous trajectory to track a precomputed discrete MDP policy. This policy aimed to reach a predefined goal, denoted as p^H_*, while avoiding obstacles in the environment. We generated five environments with randomly located obstacles denoted as *Env1*, ..., *Env5*; these obstacles were used only for the human, not the robot, to focus our evaluation on the robot's capability in ensuring human safety. The proposed method can handle static obstacles by adding extra constraints in the trajectory optimization section 6.3.1.

We designed three objective functions for the human's trajectory optimization. The first, *H-Indep-R*, was designed to track the MDP policy. The other two, *H-To-R* and *H-Away-R*, were minimizing and maximizing (respectively) the distance to the current robot position, besides policy tracking. We anticipated that our human model g in eq. (6.1) could capture the dependency of the human's behavior upon the robot states.

The robot's task was to efficiently reach the goal while ensuring the simulated human's safety, with a predefined starting location of p^R_1 and goal location of p^R_*. Our initial dataset contained 45 input-output pairs collected from 3 rollouts of the simulated human. Figure 6.3 shows a snapshot of the robot and human in the environment along with the human prediction.

We define an iteration as a time step where the system runs safe MPC to find a plan, and execute the first control signal, as described in section 6.3. An experiment is defined as the system running the safe MPC iteratively until p^R_* is reached or the maximal number of



(a)MPC-CASI$(\Omega_{\max} = 0.6)$ (b)MPC-CA

Figure 6.3: A comparison, in the form of robot trajectories and human ellipsoidal predictions, of the feasible solutions found by CASI and CA in the 2D goal-reaching domain. The red ■ represents the robot positions for $t \in \{2, 3, 4, 5\}$ along the trajectory. Here, the robot positions for $t = 4$ and 5 overlap due to the trajectory optimization's constraint eq. (6.8b). For $t \in \{2, 3, 4, 5\}$, the human position ellipsoids, $\mathcal{E}_t^{p^H}$, are plotted in different colors, whose centers are the green ■. The grey dots indicate the input human positions within the initial dataset for GP training. CASI produced a more efficient path by allowing the path to enter the human position ellipsoids, which CA does not allow. Thus, defining safety as collision avoidance or safe impact provided more flexibility than collision avoidance alone, allowing the planner to be less conservative while still guaranteeing safety.

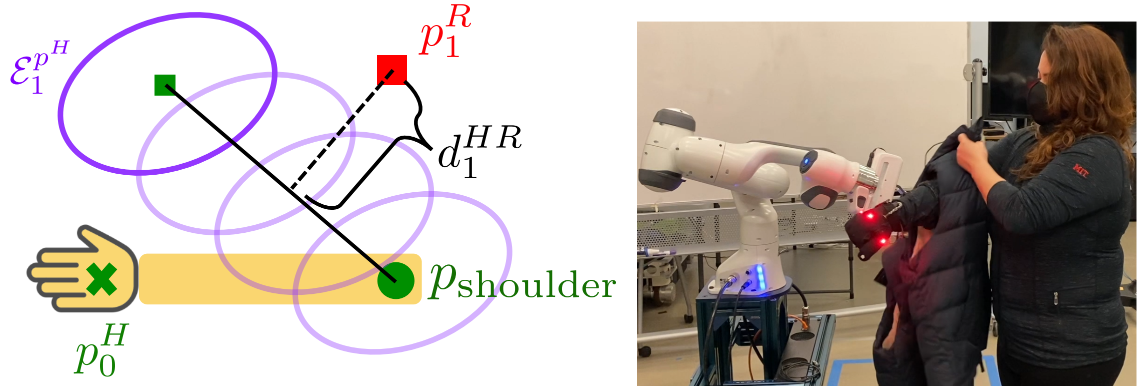
iterations, set to 50, is reached.

We benchmarked two variations of CASI, CASI ($\Omega_{\max} = 0.6$) and CASI ($\Omega_{\max} = 0.3$), against CA. Here, $\Omega_{\max}$ denotes the maximum allowable impact potential as defined in section 6.2.2 and we set its values according to Heinzmann and Zelinsky [308]. We ran 30 trials for each condition and evaluated performance based on the following measurements: (1) *#Iteration*: the number of iterations taken for the robot to reach p_*^R (per trial); (2) *#SafeCollision*: the number of collisions involving safe impact (per trial); (3) *%SafeCollision*: the percentage calculated by #SafeCollision divided by #Iteration; (4) *#UnsafeCollision*: the number of collisions involving unsafe impact (per trial); and (5) *PlanTime* (s): the amount of time taken to solve the trajectory optimization (per iteration).

6.5.2 Robot-Assisted Dressing

We deployed our algorithm to perform a real-world robot-assisted dressing task, wherein the robot must dress a sleeveless jacket onto a human arm (the human's fist is already inside the armhole upon task initiation). In this task, the robot runs the MPC, integrated with a framework that interleaves planning and execution [16], to find paths for its end-effector to reach the goal, p_*^R, near the human shoulder position, denoted by p_{shoulder}. Both p_*^R and p_{shoulder} are assumed to be known and fixed.

We assumed that the human elbow never bends during the task, and accordingly modeled the human hand as a point mass and approximated the arm configuration by linear interpolation between the hand and p_{shoulder}. Our system also duplicated the ellipsoidal predictions



(a) Formulation of the dressing task

(b) Initial data collection

Figure 6.4: (a) The robot models the human hand and robot end effector as point masses. By assuming that the human shoulder position, p_{shoulder}, is known to the robot and static during the dressing process, the robot then approximates the human arm by linearly interpolating between the hand position, p^H, and p_{shoulder}. At $t = 0$, the hand is observed directly by a sensor with the position p_0^R (green cross), and the human arm is illustrated as the orange rectangle. At time 1, the robot first predicts the human hand position as an ellipsoid $\mathcal{E}_1^{p^H}$ (in dark purple), which is constructed to bound the true hand position, p_1^H. The robot then duplicates this ellipsoid multiple times (in light purple) along the line between the center of $\mathcal{E}_1^{p^H}$ (green square) and the human shoulder position (green disc). The duplicated ellipsoids are constructed to bound the true human arm, assuming that the human arm does not bend during the dressing process. To enable dressing, the robot adds an "armhole" constraint to the trajectory optimization section 6.3.1, to ensure that the robot always stays close to the human arm during the dressing process. For example, at time $t = 1$, this constraint enforces that the distance d_1^{HR} (black curly brackets) between the robot position p_1^R and the interpolated line connecting the centers of the duplicated ellipsoid (black solid line) must be $\leq d_{\max}^{HR}$. (b) Initial non-interactive human dynamics data was collected by allowing the human to dress themselves.

for the hand positions to bound the arm configuration, as described in fig. 6.4a. Accordingly, the trajectory optimization used all ellipsoids for hand position and arm configuration, as well as ellipsoids for hand velocity, when ensuring safety constraints. We have considered lifting this assumption in future work by drawing insights from prior art [10, 276, 316–318].

As shown in fig. 6.1, the robot must remain physically close to the human arm due to the limited size of the armhole, which makes this task appropriate for evaluating the benefits of optimizing for safety, defined as collision avoidance or safe impact, rather than just collision avoidance. We encoded the armhole constraint approximately by adding an additional constraint, $d_t^{HR} \leq d_{\max}^{HR}$ for all $t = 1, 2, \ldots, T$, to the trajectory optimization section 6.3.1. Here, d_t^{HR} denotes the distance between the robot position p_t^R and the line between p_{shoulder} and the center of the ellipsoid $\mathcal{E}_t^{p^H}$, as illustrated in fig. 6.4a. The threshold $d_{\max}^{HR}$ encodes the size of the armhole.

Our initial dataset contained 16 input-output pairs collected from two human rollouts

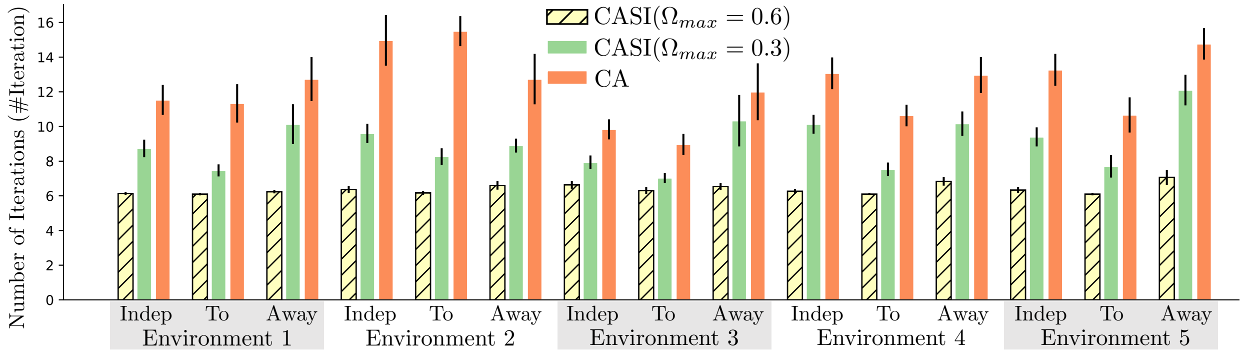


Figure 6.5: The mean and standard error of #Iteration in the 2D goal-reaching domain. The benchmark included running CASI ($\Omega_{\max} = 0.6$), CASI ($\Omega_{\max} = 0.3$), and CA, with 15 different simulated human behaviors (five environments and three objective functions). CASI ($\Omega_{\max} = 0.6$) achieved the highest efficiency, while CA yielded the lowest.

during which the human used her left arm to dress her right arm while the robot remained static, as shown in fig. 6.4b. We conducted a case study to compare CASI ($\Omega_{\max} = 1$) against CA under three conditions wherein $d^{HR}_{\max}$ was set to 0.08m, 0.085m, and 0.09m. We measured *#Iteration* and *TotalTime* (s), which is the total amount of time taken by the robot to reach p^R_* (different from #PlanTime).

6.6 Results

6.6.1 2D Goal-Reaching

We benchmarked the three algorithm variations, CASI ($\Omega_{\max} = 0.6$), CASI ($\Omega_{\max} = 0.3$), and CA, using 15 different simulated human behaviors (five environments and three objective functions), for 30 trials, and measured #Iteration, #SafeCollision, %SafeCollision, #UnsafeCollision, and PlanTime in each trial. In all conditions, #UnsafeCollision was always 0, verifying the safety guarantee provided by all three algorithm variations. The results of #Iteration, #SafeCollision, %SafeCollision, and PlanTime under all conditions are presented in table 6.1. For each measurement, we conducted a Wilcoxon signed-rank test to perform pairwise comparisons among the three algorithm variations.

The results of #Iteration are presented in table 6.1 and fig. 6.5. Both CASI ($\Omega_{\max} = 0.6$) and CASI ($\Omega_{\max} = 0.3$) had a significantly lower #Iteration than CA (both $p < 0.001$ resulting from the Wilcoxon signed-rank test); hence, CASI produced significantly more efficient plans than CA. The definition of safety as collision avoidance or safe impact, rather than just collision avoidance, allowed the planner to be less conservative while still guaranteeing safety. This is illustrated in fig. 6.3, where even though the ellipsoids in CASI and in CA had roughly similar areas, CASI produced a more efficient path by allowing the path to enter the ellipsoids, which CA does not allow. In addition, CASI ($\Omega_{\max} = 0.6$) had a significantly lower #Iteration than CASI ($\Omega_{\max} = 0.3$) ($p < 0.001$), which implies that the benefit to efficiency increases as $\Omega_{\max}$ increases.

The results of #SafeCollision and %SafeCollision are presented in table 6.1. Regarding

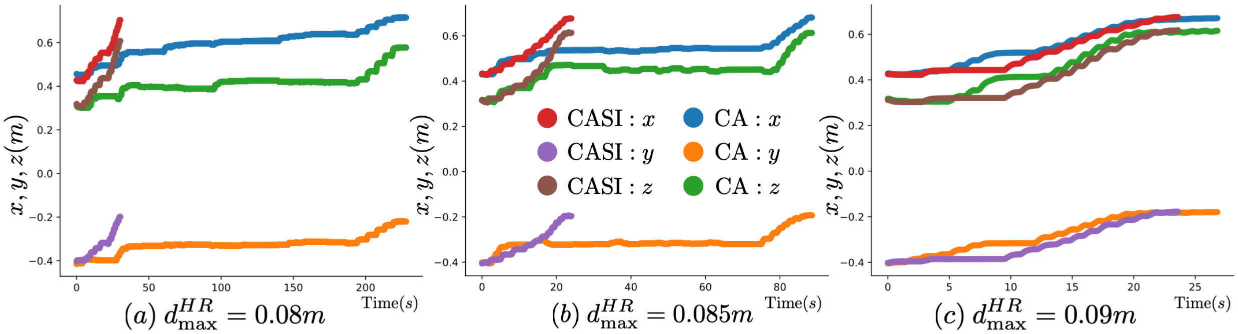
	Environment 1			Environment 2			Environment 3			Environment 4			Environment 5		
	H-Indep-R	H-To-R	H-Away-R	H-Indep-R	H-To-R	H-Away-R	H-Indep-R	H-To-R	H-Away-R	H-Indep-R	H-To-R	H-Away-R	H-Indep-R	H-To-R	H-Away-R
CASI ($\Omega_{\max} = 0.6$) #Iteration	**6.133**	**6.100**	**6.233**	**6.367**	**6.167**	**6.600**	**6.633**	**6.300**	**6.533**	**6.267**	**6.100**	**6.833**	**6.333**	**6.100**	**7.067**
CASI ($\Omega_{\max} = 0.3$) #Iteration	8.733	7.467	10.133	9.600	8.267	8.900	7.933	7.033	10.333	10.133	7.533	10.167	9.400	7.700	12.100
CA #Iteration	11.533	11.333	12.733	14.967	15.500	12.733	9.833	8.967	12.000	13.067	10.633	12.967	13.267	10.667	14.767
CASI ($\Omega_{\max} = 0.6$) #SafeCollision	0.767	1.267	**0.500**	0.467	**0.600**	0.600	0.333	0.600	0.333	1.300	1.133	0.833	**0.367**	**0.867**	0.400
CASI ($\Omega_{\max} = 0.3$) #SafeCollision	**0.633**	**0.833**	0.567	0.533	0.767	0.433	**0.167**	0.633	0.333	**1.133**	1.133	0.967	0.533	0.933	**0.367**
CA #SafeCollision	1.033	1.000	0.633	**0.333**	0.667	**0.267**	0.600	**0.500**	0.400	1.300	1.200	**0.733**	0.933	1.000	0.433
CASI ($\Omega_{\max} = 0.6$) %SafeCollision	0.107	0.178	0.070	0.067	0.081	0.075	0.044	0.085	0.046	0.180	0.160	0.105	**0.047**	0.123	0.052
CASI ($\Omega_{\max} = 0.3$) %SafeCollision	**0.065**	**0.102**	0.054	0.062	0.090	0.044	**0.021**	0.079	**0.034**	**0.106**	0.139	0.091	0.051	0.114	**0.034**
CA %SafeCollision	0.095	0.104	**0.043**	**0.023**	**0.036**	**0.022**	0.055	**0.055**	0.036	0.107	**0.112**	**0.056**	0.060	**0.105**	0.037
CASI ($\Omega_{\max} = 0.6$) PlanTime (s)	0.299	0.469	0.307	0.335	0.464	0.332	0.355	0.731	0.369	0.298	0.494	0.337	0.312	0.590	0.397
CASI ($\Omega_{\max} = 0.3$) PlanTime (s)	0.432	0.636	0.510	0.571	0.847	0.614	0.412	0.718	0.494	0.454	0.661	0.496	0.567	0.662	0.522
CA PlanTime (s)	**0.145**	**0.283**	**0.132**	**0.146**	**0.258**	**0.141**	**0.135**	**0.312**	**0.127**	**0.147**	**0.315**	**0.153**	**0.154**	**0.232**	**0.167**

Table 6.1: The means of #Iteration, #SafeCollision, %SafeCollision, and PlanTime in the 2D goal-reaching domain. The benchmark included running CASI ($\Omega_{\max} = 0.6$), CASI ($\Omega_{\max} = 0.3$), and CA, with 15 different simulated human behaviors (five environments and three objective functions), and 30 trials for each condition. The algorithm that achieved the lowest value among the three algorithm variations for each measurement is highlighted in bold.

#SafeCollision, no significant pairwise differences were found. The p-value from the Wilcoxon signed-rank test for the pair, CASI ($\Omega_{\max} = 0.6$) and CA, was 0.583. The p-value for CASI ($\Omega_{\max} = 0.3$) and CA was 0.202. The p-value for CASI ($\Omega_{\max} = 0.6$) and CASI ($\Omega_{\max} = 0.3$) was 0.552. Regarding %SafeCollision, we have found that CASI ($\Omega_{\max} = 0.6$) achieves a significantly higher %SafeCollision than both CASI ($\Omega_{\max} = 0.3$) ($p < 0.001$) and CA ($p < 0.001$), while no significance found between CASI ($\Omega_{\max} = 0.3$) and CA ($p = 0.092$).

This result implies that the extra flexibility in ($\Omega_{\max} = 0.6$) resulted in an increase percentage of collisions during the task. This does not invalidate CSAI's safety for two reasons. First, even though CASI ($\Omega_{\max} = 0.6$) has a higher percentage of collision, its #UnsafeCollision is 0, which means that it is still safe based on our definition of safety. Second, CASI ($\Omega_{\max} = 0.6$) completed the task in less number of iterations, so even though its percentage of collision is higher, its number of collisions is not significantly higher than the other methods. In summary, CASI allows collisions to occur, as long as the impacts are safe, while CA does not allow collisions at all. However, CASI does not always result in higher numbers of safe collisions than CA.

The results of PlanTime are presented in table 6.1. Both CASI ($\Omega_{\max} = 0.6$) and CASI ($\Omega_{\max} = 0.3$) had a significantly longer PlanTime than CA (both $p < 0.001$). Thus, the trajectory optimization in CASI took significantly longer to solve than the optimization in CA. One potential reason is that CASI needs to construct ellipsoids for both human positions and velocities, while CA only needs to do so for human positions. Another potential reason is the numerical challenge in handling the potentially stiff surrogate constraint, as formulated in eq. (6.17), could make the optimization in CASI harder to solve than the optimization in CA. In addition, CASI ($\Omega_{\max} = 0.3$) had a significantly longer PlanTime than CASI ($\Omega_{\max} = 0.6$) ($p < 0.001$), which implies that solving the trajectory optimization becomes harder as $\Omega_{\max}$ decreases. In real-time control scenarios where robots need to replan at 50Hz [319], the extra computation in our MPC could counteract the benefits in achieving a lower #Iteration. However, in many HRI scenarios, where the robot is equipped with compliant controllers, replanning at a much lower frequency could suffice. In such cases, our MPC's strength in finding safe and efficient solutions could potentially outweigh its weakness in longer planning



(a) $d_{\max}^{HR} = 0.08m$ (b) $d_{\max}^{HR} = 0.085m$ (c) $d_{\max}^{HR} = 0.09m$

Figure 6.6: The trajectories of the robot end-effector positions, recorded during one execution of the assisted dressing task. Each subplot depicts trajectories along three dimensions (x, y, z) produced by two algorithm variations, CASI ($\Omega_{\max} = 1$) and CA; hence, each subplot includes six trajectories. (a) presents the case of $d_{\max}^{HR} = 0.08$m, which means the robot end-effector had to remain within 0.08m of the human arm. (b) and (c) present the case of $d_{\max}^{HR} = 0.085$m and 0.09m, respectively. In both (a) and (b), CASI achieved much higher efficiency than CA. In (c), CASI and CA achieved a similar efficiency, though CASI was slightly more efficient.

time, resulting in a shorter total time in task completion, which is demonstrated in the robot-assisted dressing domain in section 6.6.2.

6.6.2 Robot-Assisted Dressing

We ran our system with one subject as a case study, leaving a full user study for future work. Figure 6.6 depicts the trajectory of robot positions recorded during a single task execution along three dimensions (x, y, and z) in Cartesian space. Figure 6.6(a) represents the execution in the case of $d_{\max}^{HR} = 0.08$m, meaning that the robot end-effector had to stay within 0.08m of the human arm. Our planner, CASI ($\Omega_{\max} = 1$), was able to reach the goal in 11 iterations and 29.30s, while CA reached the goal in 49 iterations and 223.52s. One explanation for the inefficiency observed with CA is that it is challenging to satisfy both the collision avoidance constraint and the armhole constraints, $d_t^{HR} \leq d_{\max}^{HR}$ for all $t = 1, 2, \ldots, T$ as formulated in section 6.5.2, while under large uncertainty about the future human motion. As depicted in fig. 6.4a, if the uncertainty represented by the ellipsoids is huge, then collision avoidance motivates p_R^1 to stay very far from the human arm (solid black line). However, the armhole constraint requires p_1^R to stay close to the human arm; hence, CA needed to collect much more data in order to significantly reduce the uncertainty (size of the ellipsoids) before finding feasible plans. In contrast, CASI allowed greater flexibility by ensuring safety as CA or SI. Thus, even when uncertainty about future human motion is relatively high, the robot can still find feasible plans by ensuring safe impact, resulting in more efficient task completion.

As we slightly relaxed the armhole constraint, the robot gained more freedom for collision avoidance. When $d_{\max}^{HR} = 0.085$m, CA accomplished the task in 40 iterations and 87.02s, which is still much longer than CASI ($\Omega_{\max} = 1$), which completed the task in 13 iterations and 23.22s. When $d_{\max}^{HR} = 0.09$m, both algorithm variations yielded similar performance, with CASI ($\Omega_{\max} = 1$) still a bit more efficient than CA (12 iterations and 22.57s compared with

32 iterations and 25.78s, respectively). In conclusion, CASI was able to accomplish the task and ensure safety, even when close contact between the human and robot was unavoidable.

6.7 Conclusion

This chapter presented a planning framework for safe and efficient human-robot interaction under uncertainty. Traditional approaches define safety as strict collision avoidance, which often forces the robot to freeze when uncertainty about future human motion is high. This is particularly limiting in physically assistive tasks like dressing, where contact is sometimes necessary for progress.

To address this, I introduced a two-pronged safety constraint: the robot is safe if it either avoids collisions or ensures any contact remains low-impact. I formally integrated this constraint into a learning-based model predictive control algorithm that models and updates uncertainty about human motion from data.

Empirical results in both simulation and physical dressing experiments showed that this approach significantly improved task efficiency while maintaining safety guarantees, compared to planners that enforce collision avoidance alone.

These results underscore a central insight of this thesis: that when uncertainty about human behavior cannot be eliminated, robots must reason conservatively, but not rigidly. By formally redefining safety to include both collision avoidance and safe contact, this work shows that conservative planning can still allow meaningful progress in assistive tasks. Rather than waiting passively for uncertainty to diminish, the robot acts cautiously but decisively, ensuring safety while maintaining efficiency. This approach exemplifies how robots can adapt their decision-making to hard-to-predict human behavior, not by ignoring uncertainty, but by reshaping constraints to accommodate it safely.

Chapter 7

Conclusion and Future Directions

"Knowing is not enough; we must apply. Willing is not enough; we must do."

—Johann Wolfgang von Goethe

7.1 Conclusion

Uncertainty is inevitable in human-robot interaction. Whether a robot is learning what a person prefers, estimating their hidden state, or planning how to assist them physically, it must operate with incomplete and imperfect information. This thesis has addressed the core question: *How should a robot behave when it is uncertain about the human?*

Rather than treating uncertainty as a nuisance, I propose embracing it as a guiding design principle. Across three contributions, this thesis demonstrates how robots can *reduce*, *represent and respect*, and *act under* uncertainty to provide safe and effective assistance.

7.1.1 Reducing Uncertainty about Preferences using Cognitive Feedback

In scenarios where the robot learns from human preferences, I showed that subtle behavioral cues, such as response time, can provide valuable information about preference strength. By modeling the human cognitive process behind decision-making, the robot can learn preferences more efficiently than relying on choices alone. This cognitive feedback enables faster personalization without increasing user burden.

7.1.2 Representing and Respecting Uncertainty during State Estimation

In physical interaction settings such as robot-assisted dressing, human states, such as the elbow position, may be occluded and cannot be directly observed. I introduced a set-based estimator, GP-ZKF, that explicitly represents epistemic uncertainty from learned dynamics and observation models, as well as aleatoric uncertainty from noise. By constructing conservative geometric estimates that are guaranteed to contain the true human state with high probability, GP-ZKF enables the robot to avoid overconfidence and remain consistent, even when the test-time scenarios differ from the training data.

7.1.3 Acting under uncertainty with relaxed safety constraints

Finally, I addressed how a robot should plan physical assistance in the presence of uncertainty about human motion. Traditional planners enforce strict collision avoidance, which can lead the robot to freeze in uncertain environments. I proposed a new safety constraint that allows either collision avoidance or safe low-impact contact. This redefinition of safety enables the robot to assist more effectively without compromising physical safety.

Together, these contributions form a unified framework for uncertainty-aware personalization in human-robot interaction. The proposed methods span different stages of the robot pipeline, learning, estimation, and planning, but share a common theme: explicitly modeling and responding to uncertainty to improve interaction.

7.2 Future Directions

This thesis explored how robots can act under uncertainty about humans, whether in their preferences, physical states, or movements. By developing principled approaches to reduce, respect, and act under uncertainty, this thesis lays a foundation for personalizing robot assistance in a reliable and human-centered way. Yet, many challenges remain in scaling these methods to real-world robot-assistive scenarios involving actual users. Below, I outline several future directions that build on this thesis.

7.2.1 Contact-Rich Dexterity for Real-World Assistance

The third contribution of this thesis (chapter 6) addressed uncertainty in robot-assisted dressing, enabling robots to plan trajectories that adapt to uncertain human arm movements. However, real-world assistance involves richer physical contact. Human caregivers do not just slide garments. They gently adjust and flatten clothing, regrasp clothing as needed, and apply force with care and precision, often without clear visual feedback of the user's body. These skills require fine control of both motion and force to manage the contact, which remains challenging for classical model-based methods.

Recent advances in imitation learning provide a promising alternative. For instance, Hou et al. [320] use diffusion policies to learn both movement and compliance from human demonstrations, balancing precision and responsiveness. Black et al. [321] develop a vision-language-action model to imitate humans in contact-rich and long-horizon tasks such as folding shirts. Adapting such methods to assistive settings could unlock new levels of dexterity and responsiveness in physical interaction.

However, these approaches are often black-box and difficult to interpret or control. This raises important safety challenges. A future direction is to build deployment pipelines that incorporate uncertainty estimation into policy execution, stress-test behaviors through adversarial "red-teaming," and provide supervisory interfaces so caregivers can monitor and intervene. The goal is to combine the expressiveness of learned policies with the transparency and safety needed in assistive settings.

7.2.2 Coordination through Communication

The second and third contributions of this thesis (chapters 5 and 6) focus on robot assistance where the human passively responds to the robot's actions. However, many assistive tasks require active coordination between the human and the robot. For instance, after the robot assists a user in dressing one arm, guiding the second arm into the sleeve often requires the user to move intentionally and in sync with the robot. This type of interaction demands mutual understanding: the robot must infer the human's intent, and the human must anticipate the robot's actions. In such scenarios, communication, both physical and verbal, is essential for effective collaboration.

My earlier work on human-robot coordination [55, 56] explored how robots can model uncertainty over human intent and decide both when and what to communicate. For example, the robot could use language to request a specific action from the human or inform the human about its own intended behavior. Crucially, the robot would only communicate when necessary, striking a balance between informativeness and the need to avoid over-communication.

Future work could bring these ideas to assistive robotics. With large language models (LLMs), robots now have powerful tools for communication. A promising direction is to build datasets of real caregiver-user interactions and fine-tune LLMs to communicate appropriately. The central challenge is enabling the robot to reason about uncertainty in human intent and decide when and how to communicate, whether to request, inform, or ask a question, in order to coordinate efficiently and safely.

7.2.3 Learning from Multimodal and Evolving Human Feedback

The first contribution of this thesis (chapter 4) showed that response time provides valuable information about human preference strength, augmenting binary comparisons. But response time is only one lens into the human cognitive process, and it can be unreliable, especially in real-world crowdsourcing settings where human attention fluctuates [322].

A natural next step is to integrate multiple modalities of implicit feedback, such as gaze patterns, facial expressions, and hesitation. For example, users may replay trajectories or hesitate when uncertain, a signal similar to long response times. Algorithms that interpret these behaviors as implicit feedback could facilitate faster and less burdensome preference learning. Combining implicit signals with explicit feedback and task-level metrics will lead to more grounded, human-aware personalization.

Beyond a single interaction, human preferences, trust, and goals evolve, particularly in assistive settings that involve learning or rehabilitation. For example, users may adapt as they gain familiarity or confidence with the robot. A person who initially prefers slow, conservative motions may later desire faster, more autonomous assistance. Modeling this human-robot co-adaptation is key to long-term personalization. One promising direction is to view the problem as a two-player game with asymmetric information [323], where both the robot and human adapt over time, each with partial knowledge of the other.

7.2.4 From Passive Feedback to Active Human Input

The preference learning framework developed in this thesis's first contribution (chapter 4) treats humans as passive agents who respond to robot-generated queries. But people have their own sense of autonomy [12], and may want to guide the robot proactively, for example, by saying "go slower" or "try something else." Future work should explore how to integrate such human-initiated feedback with robot-initiated querying.

A key challenge is designing a personalization system that balances both modes of interaction: allowing the robot to ask meaningful questions when uncertain, while also recognizing when users want to take the lead. Studying how different users experience autonomy, agency, and cognitive load will be crucial for creating systems that adapt not just to what users prefer, but how they prefer to interact.

Toward Lifelong Personalized Assistance

Ultimately, I envision a future in which robots do more than execute preprogrammed tasks. They continuously learn from human behavior, both explicit and implicit, and adapt their assistance accordingly. They act safely even in uncertain situations, and evolve alongside users as their needs change.

This thesis takes a step in that direction by placing uncertainty at the center of robot personalization. Future work will continue this trajectory: enabling robots to reason not just about what a person did or said, but what they need, what they are learning, and how their preferences, capabilities, and trust evolve over time.

Appendix A

Appendix for Chapter 4: Reducing Uncertainty about Preferences Using Cognitive Feedback

This appendix provides theoretical background, technical derivations, and experimental details that support the results in Chapter 4. That chapter addresses how robots can reduce uncertainty about human preferences by leveraging cognitive feedback, specifically, human response times, in addition to binary choices. To enable this integration, we build on cognitive models of human decision-making from psychology and neuroscience, especially bounded accumulation models. These models provide a principled way to interpret response times as implicit feedback that reveals preference strength.

The appendix is organized into three sections. Appendix A.1 reviews relevant literature on decision-making models and estimation methods. Appendix A.2 presents formal proofs supporting our estimator's properties, including asymptotic normality and non-asymptotic concentration bounds. Appendix A.3 describes experimental setups and data processing pipelines for the empirical results presented in Chapter 4.

A.1 Literature review

This section provides background on the cognitive modeling techniques used in Chapter 4, including bounded accumulation models and estimation methods.

A.1.1 Bounded accumulation models for choices and response times

Bounded Accumulation Models (BAMs) describe human decision-making using an accumulator (or sampling rule) and a stopping rule [245]. In binary choice tasks, such as two-alternative forced choice tasks, a widely used BAM is the drift-diffusion model (DDM) [242], which models decisions as Brownian motion with fixed boundaries. To capture differences in human response times for correct and incorrect answers, Ratcliff and McKoon [242] allows drift, starting point, and non-decision time to vary across trials. Wagenmakers, Van Der Maas, and Grasman [43] later introduced the EZ-diffusion model (EZDM), a simplified version of

DDM with closed-form solutions for choice and response time moments, making parameter estimation easier and more robust. EZDM assumes a deterministic drift, a starting point, and a non-decision time, fixed across trials, with the starting point equidistant from the boundaries. Berlinghieri et al. [251] specialized EZDM to the difference-based EZDM (dEZDM), where the drift represents the utility difference between two options. For binary queries with arms z_1 and z_2, the drift is modeled as $u_{z_1} - u_{z_2}$, where u_{z_1} and u_{z_2} are the utilities of z_1 and z_2.

As discussed in section 4.1, we impose a linear utility structure on the dEZDM, where each arm's utility is given by $u_z = z^\top \theta^*$, with θ^* denotes the human preference vector. This approach is supported by both bandit and psychology literature. In bandits, linear utility models scale efficiently with a large number of arms [252, 324]. In psychology, linear combinations of attributes are commonly used in multi-attribute decision-making models [254–256]. The standard dEZDM in [251, Definition 1] is a special case of our dEZDM with a linear utility structure, where arms correspond to the standard basis vectors in Euclidean space $\mathbb{R}^d$. This mirrors the relationship between multi-armed bandits and linear bandits.

Similarly to our approach, Shvartsman et al. [325] parameterize the human utility function as a Gaussian process and propose a moment-matching Bayesian inference method that uses both choices and response times to estimate latent utilities. Unlike our work, their focus is solely on estimation and does not address bandit optimization. Integrating their estimation techniques into bandit optimization presents an interesting avenue for future research.

Another widely used BAM is the race model [243, 326], which naturally extends to queries with more than two options. In race models, each option has its own accumulator, and the decision ends when any accumulator reaches its barrier. BAMs can also model human attention during decision-making. For example, the attentional-DDM [46, 256, 327] jointly models choices, response times, and eye movements across different options or attributes. Similarly, Thomas et al. [328] introduce the gaze-weighted linear accumulator model to study gaze bias at the trial level. To incorporate learning effects, Pedersen, Frank, and Biele [329] combines reinforcement learning (RL) with DDM, where the human adjusts the drift through RL. In contrast, our work uses RL for AI decision-making when interacting with humans. BAMs also connect to Bayesian RL models of human cognition. For example, Fudenberg, Strack, and Strzalecki [330] propose a model where humans balance decision accuracy and time cost, showing it is equivalent to a DDM with time-decaying boundaries. Neurophysiological evidence supports BAMs. For instance, EEG recordings demonstrate that neurons exhibit accumulation processes and decision thresholds [245]. Additionally, diffusion processes have been used to model neural firing rates [331].

A.1.2 Parameter estimation for bounded accumulation models

BAMs often lack closed-form density functions, so hierarchical Bayesian inference is commonly used for parameter estimation [246]. While flexible, these methods are computationally intensive, making them impractical for real-time applications in online learning systems. Faster estimators [43, 251, 270] usually estimate parameters for individual option pairs without leveraging data across pairs. To address this, we propose a computationally efficient method for estimating linear human utility functions, which we integrate into bandit learning. In section 4.4.2, we empirically show that our estimator outperforms those from prior work [43, 270].

In practice, using response time data requires pre-processing and model fitting, as outlined by Myers, Interian, and Moustafa [322]. Additionally, Alós-Ferrer, Fehr, and Netzer [240], Fudenberg et al. [250], and Baldassi et al. [332] propose statistical tests to assess the suitability of various DDM extensions for a given dataset.

A.1.3 Uses of response times

Response times serve multiple purposes, as highlighted by Clithero [239]. A primary use is improving choice prediction. For instance, Clithero [45] showed that DDM predicts choice probabilities more accurately than the logit model, with parameters estimated through Bayesian Markov chain Monte Carlo. Similarly, Alós-Ferrer, Fehr, and Netzer [240] demonstrated that response times enhance the identifiability of human preferences compared to using choices alone.

Response times also shed light on human decision-making processes. Castro et al. [333] applied DDM analysis to explore how cognitive workload, induced by secondary tasks, influences decision-making. Analyzing response times has been a long-standing method in cognitive testing to assess mental capabilities [241]. Additionally, Zhang, Kemp, and Lipovetzky [334, 335] introduced a framework that uses human planning time to infer their intended goals.

Response times can also enhance AI decision-making. In dueling bandits and preference-based RL [261], human choice models are commonly used for preference elicitation. One such model, the random utility model, can be derived from certain BAMs [240]. For example, as discussed after eq. (4.1), both the Bradley-Terry model [**BradleyTerry1952**] and dEZDM [43, 251] yield logistic choice probabilities in the form $\mathbb{P}[z_1 \succ z_2] = \sigma_{logistic}(u_{z_1} - u_{z_2}) = 1/\left(1 + \exp\left(-c \cdot (u_{z_1} - u_{z_2})\right)\right)$, where u_{z_1} and u_{z_2} denote the utilities of z_1 and z_2 and c is some constant [261, section 3.2]. Our work leverages this connection between random utility models and choice-response-time models to estimate human utilities using both choices and response times.

We hypothesize that our key insight, that response times provide complementary information, especially for queries with strong preferences, extends beyond the dEZDM and the specific logistic link function $\sigma_{logistic}$. Many psychological models capture both choices and response times but lack closed-form choice distributions. In such cases, the choice probability is often expressed as $\mathbb{P}[z_1 \succ z_2] = \sigma^{\dagger}(u_{z_1}, u_{z_2})$, where $\sigma^{\dagger}$ is a function of u_{z_1} and u_{z_2} without a closed form. Fixing u_{z_2} and varying u_{z_1} defines the psychometric function $\sigma^{\dagger}(\cdot, u_{z_2})$, which typically exhibits an "S" shape [336, fig. 1.1]. As preferences become stronger, $\sigma^{\dagger}$ flattens, similar to figs. 4.1b and 4.1c, suggesting that choices carry less information. We conjecture that response times remain a valuable complementary signal in such cases.

If we further assume the choice probability depends only on the utility difference, $u_{z_1} - u_{z_2}$, then $\mathbb{P}[z_1 \succ z_2] = \sigma^{\ddagger}(u_{z_1} - u_{z_2})$, where the link function $\sigma^{\ddagger}$ is typically assumed to be strictly monotonic and bounded within $[0, 1]$ [261, section 3.2]. These properties naturally produce an "S"-shaped curve that flattens as preferences become stronger, again suggesting that choices provide less information. In such cases, we conjecture that response times can complement choices to enhance learning.

In summary, BAMs, like DDMs and race models, offer a strong theoretical framework for understanding human decision-making, supported by both behavioral and neurophysiological evidence. These models have been widely applied to choice prediction and the study of human cognitive processes. Our work connects BAMs with bandit algorithms by introducing a computationally efficient estimator for online preference learning. Future research could explore other BAM variants to further examine the benefits of incorporating response times.

A.2 Proofs

A.2.1 Parameters of the difference-based EZ-Diffusion Model (dEZDM) [43, 251]

Given a human preference vector θ^*, for each query $x \in \mathcal{X}$, the utility difference is defined as $u_x \coloneqq x^\top \theta^*$. In the dEZDM model (introduced in section 4.1), with barrier a, according to Wagenmakers, Van Der Maas, and Grasman [43, eq. (4), (6), and (9)], the human choice c_x has the following properties:

$$\mathbb{P}\left(c_x = 1\right) = \frac{1}{1 + \exp\left(-2au_x\right)}, \quad \mathbb{P}\left(c_x = -1\right) = \frac{\exp\left(-2au_x\right)}{1 + \exp\left(-2au_x\right)}.$$

Thus, the expected choice is $\mathbb{E}\left[c_x\right] = \tanh(au_x)$, and the choice variance is $\mathbb{V}\left[c_x\right] = 1 - \tanh(au_x)^2$ (restating eq. (4.1)).

The human decision time t_x has the following properties:

$$\mathbb{E}\left[t_x\right] = \begin{cases} \frac{a}{u_x}\tanh(au_x) & \text{if } u_x \neq 0 \\ a^2 & \text{if } u_x = 0 \end{cases} \qquad \text{(restating eq. (4.1))},$$

$$\mathbb{V}\left[t_x\right] = \begin{cases} \frac{a}{{u_x}^3}\frac{\exp(4au_x) - 1 - 4au_x\exp(2au_x)}{\left(\exp(2au_x)+1\right)^2} & \text{if } u_x \neq 0 \\ 2a^4/3 & \text{if } u_x = 0 \end{cases}.$$

From this, we obtain the following key relationship:

$$\frac{\mathbb{E}\left[c_x\right]}{\mathbb{E}\left[t_x\right]} = \frac{u_x}{a} = x^\top\left(\frac{1}{a}\theta^*\right) \qquad \text{(restating eq. (4.2))}.$$

All these parameters depend solely on the utility difference $u_x \coloneqq x^\top\theta^*$ and the barrier a.

A.2.2 Asymptotic normality of the choice-decision-time estimator for estimating the human preference vector $\boldsymbol{\theta}^*$

We now present the proof of the asymptotic normality result for the choice-decision-time estimator, $\widehat{\theta}_{\text{CH,DT}}$, as stated in theorem 4.2.1, which is restated as follows:

Theorem 4.2.1 (Asymptotic normality of $\widehat{\theta}_{\text{CH,DT}}$)**.** *Given a fixed i.i.d. dataset, denoted by* $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{i\in[n]}$ *for each* $x \in \mathcal{X}_{sample}$*, where* $\sum_{x\in\mathcal{X}_{sample}} xx^\top \succ 0$*, and assuming that the datasets for different* $x \in \mathcal{X}_{sample}$ *are independent, then, for any vector* $y \in \mathbb{R}^d$*, as* $n \to \infty$*, the following holds:*

$$\sqrt{n}\, y^\top\left(\widehat{\theta}_{CH,DT,n} - \theta^*/a\right) \xrightarrow{D} \mathcal{N}(0, \zeta^2/a^2).$$

Here, the asymptotic variance depends on a problem-specific constant, ζ^2*, with an upper bounded:*

$$\zeta^2 \leq \|y\|^2_{\left(\sum_{x\in\mathcal{X}_{sample}}\left[\min_{x'\in\mathcal{X}_{sample}}\mathbb{E}[t_{x'}]\right]\cdot xx^\top\right)^{-1}}.$$

Proof. To simplify notation, we define:

$$\widehat{\mathcal{C}}_x = \frac{1}{n}\sum_{i=1}^{n} c_{x,s_{x,i}}, \quad \mathcal{C}_x = \mathbb{E}\left[c_x\right], \quad \widehat{\mathcal{T}}_x = \frac{1}{n}\sum_{i=1}^{n} t_{x,s_{x,i}}, \quad \mathcal{T}_x = \mathbb{E}\left[t_x\right]. \tag{A.1}$$

For brevity, we abbreviate $\mathcal{X}_{\text{sample}}$ as $\mathcal{X}$ and $\widehat{\theta}_{\text{CH,DT},n}$ as $\widehat{\theta}$. The estimator $\widehat{\theta}$ can be expressed as:

$$\widehat{\theta} = \left(\sum_{x'\in\mathcal{X}} nx'x'^{\top}\right)^{-1} \sum_{x\in\mathcal{X}} nx\, \frac{\widehat{\mathcal{C}}_x}{\widehat{\mathcal{T}}_x} \qquad \text{(restating eq. (4.3))}.$$

We rewrite θ^*/a as:

$$\begin{aligned} \theta^*/a &= \left(\sum_{x'\in\mathcal{X}} nx'x'^{\top}\right)^{-1} \sum_{x\in\mathcal{X}} nxx^{\top}\, \frac{\theta^*}{a} \\ &= \left(\sum_{x'\in\mathcal{X}} nx'x'^{\top}\right)^{-1} \sum_{x\in\mathcal{X}} nx\, \frac{\mathcal{C}_x}{\mathcal{T}_x}. \end{aligned} \tag{A.2}$$

Therefore, for any vector $y \in \mathbb{R}^d$, we have:

$$y^{\top}\left(\widehat{\theta} - \frac{\theta^*}{a}\right) = y^{\top}\left(\sum_{x'\in\mathcal{X}} nx'x'^{\top}\right)^{-1} \sum_{x\in\mathcal{X}} nx \left(\frac{\widehat{\mathcal{C}}_x}{\widehat{\mathcal{T}}_x} - \frac{\mathcal{C}_x}{\mathcal{T}_x}\right) =: \sum_{x\in\mathcal{X}} \xi_x \left(\frac{\widehat{\mathcal{C}}_x}{\widehat{\mathcal{T}}_x} - \frac{\mathcal{C}_x}{\mathcal{T}_x}\right), \tag{A.3}$$

where ξ_x is defined as $\xi_x \coloneqq y^{\top}\left(\sum_{x'\in\mathcal{X}} nx'x'^{\top}\right)^{-1} nx$. In eq. (A.3), the only random variables are $\widehat{\mathcal{C}}_x$ and $\widehat{\mathcal{T}}_x$. For simplicity, for any $x_i \in \mathcal{X} \coloneqq \{x_1, \cdots, x_{|\mathcal{X}|}\}$, we slighly abuse the notation and use ξ_i, c_i, t_i, $\mathcal{C}_i$, $\mathcal{T}_i$, $\widehat{\mathcal{C}}_i$ and $\widehat{\mathcal{T}}_i$ denote ξ_{x_i}, c_{x_i}, t_{x_i}, $\mathcal{C}_{x_i}$, $\mathcal{T}_{x_i}$, $\widehat{\mathcal{C}}_{x_i}$, and $\widehat{\mathcal{T}}_{x_i}$, respectively. By applying the multidimensional central limit theorem, we have:

$$\begin{aligned} \sqrt{n}\begin{bmatrix} \widehat{\mathcal{C}}_1 - \mathcal{C}_1 \\ \widehat{\mathcal{T}}_1 - \mathcal{C}_1 \\ \vdots \\ \widehat{\mathcal{C}}_{|\mathcal{X}|} - \mathcal{C}_{|\mathcal{X}|} \\ \widehat{\mathcal{T}}_{|\mathcal{X}|} - \mathcal{C}_{|\mathcal{X}|} \end{bmatrix} &\xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} \mathbb{V}\left[c_1\right] & \operatorname{cov}\left[c_1, t_1\right] & & & \\ \operatorname{cov}\left[t_1, c_1\right] & \mathbb{V}\left[t_1\right] & & & \\ & & \ddots & & \\ & & & \mathbb{V}\left[c_{|\mathcal{X}|}\right] & \operatorname{cov}\left[c_{|\mathcal{X}|}, t_{|\mathcal{X}|}\right] \\ & & & \operatorname{cov}\left[t_{|\mathcal{X}|}, c_{|\mathcal{X}|}\right] & \mathbb{V}\left[t_{|\mathcal{X}|}\right] \end{bmatrix}\right) \\ &= \mathcal{N}\left(0, \operatorname{diag}\left[\mathbb{V}\left[c_1\right], \mathbb{V}\left[t_1\right], \cdots, \mathbb{V}\left[c_{|\mathcal{X}|}\right], \mathbb{V}\left[t_{|\mathcal{X}|}\right]\right]\right). \end{aligned} \tag{A.4}$$

In the first line of eq. (A.4), the block-diagonal structure of the covariance matrix emerges because $(\widehat{\mathcal{C}}_i, \widehat{\mathcal{T}}_i)_{i\in[|\mathcal{X}|]}$ are independent of each other. For any fixed x_i, to derive the second line of eq. (A.4), we use the fact that:

$$\begin{aligned} \mathbb{E}\left[t_i c_i\right] &= \mathbb{P}\left(c_i = 1\right)\mathbb{E}\left[1 \cdot t_i | c_i = 1\right] + \mathbb{P}\left(c_i = -1\right)\mathbb{E}\left[-1 \cdot t_i | c_i = -1\right] \\ &\overset{(i)}{=} \left(\mathbb{P}\left(c_i = 1\right) - \mathbb{P}\left(c_i = -1\right)\right)\mathbb{E}\left[t_i | c_i = 1\right] \\ &= \mathbb{E}\left[c_i\right]\mathbb{E}\left[t_i\right], \end{aligned} \tag{A.5}$$

where (i) is because $\mathbb{E}[t_i|c_i=1] = \mathbb{E}[t_i|c_i=-1]$ [262, eq. (A.7) and (A.9)]. Therefore, eq. (A.5) implies that $\operatorname{cov}(c_i, t_i) = 0$ [1], which justifies the second line of eq. (A.4).

Now, let us define the function $g(c_1, t_1, \cdots, c_{|\mathcal{X}|}, t_{|\mathcal{X}|}) \coloneqq \sum_{i\in[|\mathcal{X}|]} \xi_i\ c_i/t_i$. The gradient of g is:

$$\nabla g|_{\left(c_1,t_1,\cdots,c_{|\mathcal{X}|},t_{|\mathcal{X}|}\right)} = \begin{bmatrix}\xi_1/t_1 & -\xi_1 c_1/t_1^2 & \cdots & \xi_{|\mathcal{X}|}/t_{|\mathcal{X}|} & -\xi_{|\mathcal{X}|}c_{|\mathcal{X}|}/t_{|\mathcal{X}|}^2\end{bmatrix}^\top. \tag{A.6}$$

Using the multivariate delta method, we obtain:

$$\begin{aligned}
&\sqrt{n}\sum_{i\in[|\mathcal{X}|]}\xi_i\left(\frac{\widehat{\mathcal{C}}_i}{\widehat{\mathcal{T}}_i}-\frac{\mathcal{C}_i}{\mathcal{T}_i}\right)\\
&=\sqrt{n}\left(g\left(\widehat{\mathcal{C}}_1,\widehat{\mathcal{T}}_1,\cdots,\widehat{\mathcal{C}}_{|\mathcal{X}|},\widehat{\mathcal{T}}_{|\mathcal{X}|}\right)-g\left(\mathcal{C}_1,\mathcal{T}_1,\cdots,\mathcal{C}_{|\mathcal{X}|},\mathcal{T}_{|\mathcal{X}|}\right)\right)\\
&\xrightarrow{D}\mathcal{N}\left(0,\nabla g^\top|_{\left(\mathcal{C}_1,\mathcal{T}_1,\cdots,\mathcal{C}_{|\mathcal{X}|},\mathcal{T}_{|\mathcal{X}|}\right)}\begin{bmatrix}\mathbb{V}[c_1] & & & & \\ & \mathbb{V}[t_1] & & & \\ & & \ddots & & \\ & & & \mathbb{V}\left[c_{|\mathcal{X}|}\right] & \\ & & & & \mathbb{V}\left[t_{|\mathcal{X}|}\right]\end{bmatrix}\nabla g|_{\left(\mathcal{C}_1,\mathcal{T}_1,\cdots,\mathcal{C}_{|\mathcal{X}|},\mathcal{T}_{|\mathcal{X}|}\right)}\right)\\
&=\mathcal{N}\left(0,\sum_{i\in[|\mathcal{X}|]}\xi_i^2\left(\frac{1}{\mathcal{T}_i^2}\mathbb{V}(c_i)+\frac{\mathcal{C}_i^2}{\mathcal{T}_i^4}\mathbb{V}(t_i)\right)\right)\\
&=\mathcal{N}\left(0,\frac{1}{a^2}\sum_{i\in[|\mathcal{X}|]}\xi_i^2\left(\frac{a^2}{\mathcal{T}_i^2}\mathbb{V}(c_i)+\frac{a^2\mathcal{C}_i^2}{\mathcal{T}_i^4}\mathbb{V}(t_i)\right)\right)
\end{aligned} \tag{A.7}$$

By applying the identities outlined in appendix A.2.1, we can establish the following identity:

$$\forall i\in[|\mathcal{X}|]\colon \frac{a^2}{\mathcal{T}_i^2}\mathbb{V}(c_i)+\frac{a^2\mathcal{C}_i^2}{\mathcal{T}_i^4}\mathbb{V}(t_i)=\frac{1}{\mathcal{T}_i}. \tag{A.8}$$

Substituting this identity into eq. (A.7), we obtain:

$$\sqrt{n}\sum_{i\in[|\mathcal{X}|]}\xi_i\left(\frac{\widehat{\mathcal{C}}_i}{\widehat{\mathcal{T}}_i}-\frac{\mathcal{C}_i}{\mathcal{T}_i}\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{1}{a^2}\sum_{i\in[|\mathcal{X}|]}\xi_i^2\frac{1}{\mathcal{T}_i}\right). \tag{A.9}$$

[1] Equation (A.5) implies that for any query x_i, the human choice c_i and decision time t_i are uncorrelated. Moreover, they are independent, as discussed by Drugowitsch [337, the discussion above eq. (7)] and Baldassi et al. [332, proposition 3].

Finally, the asymptotic variance can be upper bounded as follows:

$$\begin{aligned}
&\frac{1}{a^2} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \frac{1}{\mathcal{T}_i} \\
&\leq \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \\
&= \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \cdot \left(\sum_{x \in \mathcal{X}} y^\top \left(\sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} n^2 x x^\top \left(\sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} y \right) \\
&= \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \cdot y^\top \left(\sum_{x' \in \mathcal{X}} x' x'^\top \right)^{-1} y \\
&= \frac{1}{a^2} y^\top \left(\sum_{x' \in \mathcal{X}} \left[\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i \right] x' x'^\top \right)^{-1} y \\
&\equiv \frac{1}{a^2} \|y\|^2_{\left(\sum_{x' \in \mathcal{X}} \left[\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i \right] x' x'^\top \right)^{-1}} .
\end{aligned} \tag{A.10}$$

□

A.2.3 Non-asymptotic concentration of the two estimators for estimating the utility difference $\boldsymbol{u_x}$ given a query $\boldsymbol{x}$

The choice-decision-time estimator

Section 4.2.3 focuses on the problem of estimating the utility difference for a single query. Given a query $x \in \mathcal{X}$, the objective is to estimate the utility difference $u_x \coloneqq x^\top \theta^*$ using an i.i.d. dataset, denoted by $\left\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\right\}_{i\in[n_x]}$.

We begin by applying the choice-decision-time estimator from eq. (4.3), which is derived by solving the following least squares problem:

$$\widehat{\theta}_{\mathrm{CH,DT}} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{x\in\mathcal{X}_{\mathrm{sample}}} n_x \left(x^\top\theta - \frac{\sum_{i\in[n_x]} c_{x,s_{x,i}}}{\sum_{i\in[n_x]} t_{x,s_{x,i}}} \right)^2 .$$

Similarly, the utility difference for a single query is estimated as the solution to the following least squares problem, yielding the estimate:

$$\widehat{u}_{x,\mathrm{CH,DT}} = \operatorname*{arg\,min}_{u\in\mathbb{R}} \left(u - \frac{\sum_{i\in[n_x]} c_{x,s_{x,i}}}{\sum_{i\in[n_x]} t_{x,s_{x,i}}} \right)^2 = \frac{\sum_{i\in[n_x]} c_{x,s_{x,i}}}{\sum_{i\in[n_x]} t_{x,s_{x,i}}} \quad \text{(restating eq. (4.5))}.$$

The resulting estimate, $\widehat{u}_{x,\mathrm{CH,DT}}$, approximates u_x/a rather than u_x. However, since the ranking of arm utilities is preserved between u_x/a and u_x, estimating u_x/a is sufficient for the purpose of best-arm identification.

For the case where the utility difference $u_x \neq 0$, the non-asymptotic concentration inequality for this estimator is presented in theorem 4.2.3. To prove this, we first introduce lemma A.2.1, which demonstrates that for any given query x, the decision time is a sub-exponential random variable.

To simplify notation, we define:

$$\widehat{\mathcal{C}}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} c_{x,s_{x,i}}, \quad \mathcal{C}_x = \mathbb{E}\left[c_x\right], \quad \widehat{\mathcal{T}}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} t_{x,s_{x,i}}, \quad \mathcal{T}_x = \mathbb{E}\left[t_x\right], \quad \widehat{u}_{x,\mathrm{CH,DT}} = \frac{\widehat{\mathcal{C}}_x}{\widehat{\mathcal{T}}_x}. \tag{A.11}$$

Lemma A.2.1. *If $u_x \neq 0$, then $(t_x - \mathcal{T}_x)$ is sub-exponential $SE\left(\nu_x^2, \alpha_x\right)$, where $\nu_x = \sqrt{2}a/|u_x|$ and $\alpha_x = 2/u_x^2$.*

Proof. For simplicity, we will omit the subscript x throughout the proof and assume, without loss of generality, that $u > 0$.

Our objective is to establish the following inequality, which holds for all $s \in (-u^2/2, u^2/2)$:

$$\mathbb{E}\left(\exp\left(s\left(t - \mathcal{T}\right)\right)\right) \leq \exp\left(\frac{2a^2/u^2}{2} s^2\right). \tag{A.12}$$

This implies that $(t - \mathcal{T})$ is sub-exponential $\mathrm{SE}\left(\nu^2, \alpha\right)$, as defined by Wainwright [338, Definition 2.7].

Step 1: Transform eq. (A.12) into a more manageable inequality (eq. (A.18)).

Using Cox [339, eq. (128)], with $\Delta \coloneqq u^2 - 2s$, $\theta_1 \coloneqq -u - \sqrt{\Delta}$ and $\theta_2 \coloneqq -u + \sqrt{\Delta}$, we have[2]:

$$\begin{aligned}
\mathbb{E}\left(\exp\left(st\right)\right) &= \frac{\exp\left(a\theta_1\right) - \exp\left(2a\theta_2 + a\theta_1\right)}{\exp\left(2a\theta_1\right) - \exp\left(2a\theta_2\right)} - \frac{\exp\left(a\theta_2\right) - \exp\left(2a\theta_1 + a\theta_2\right)}{\exp\left(2a\theta_1\right) - \exp\left(2a\theta_2\right)} \\
&= \frac{\exp\left(a\theta_1\right)\left[1 + \exp\left(a\theta_1 + a\theta_2\right)\right]}{\exp\left(2a\theta_1\right) - \exp\left(2a\theta_2\right)} - \frac{\exp\left(a\theta_2\right)\left[1 + \exp\left(a\theta_2 + a\theta_1\right)\right]}{\exp\left(2a\theta_1\right) - \exp\left(2a\theta_2\right)} \\
&= \frac{\left[\exp\left(a\theta_1\right) - \exp\left(a\theta_2\right)\right]\left[1 + \exp\left(a\theta_2 + a\theta_1\right)\right]}{\exp\left(2a\theta_1\right) - \exp\left(2a\theta_2\right)} \\
&= \frac{1 + \exp\left(a\theta_2 + a\theta_1\right)}{\exp\left(a\theta_1\right) + \exp\left(a\theta_2\right)} \\
&= \frac{\exp\left(-au\right) + \exp\left(au\right)}{\exp\left(-a\sqrt{\Delta}\right) + \exp\left(a\sqrt{\Delta}\right)} \\
&=: \frac{N}{D(s)}.
\end{aligned} \tag{A.13}$$

In the last line, we define $N = 2\cosh(au)$ and $D(s) = 2\cosh(a\sqrt{\Delta})$. Thus, we arrive at:

$$\mathbb{E}\left(\exp\left(s \cdot \left(t - \mathcal{T}\right)\right)\right) = \frac{N}{D(s)} \cdot \frac{1}{\exp\left(s \cdot \mathcal{T}\right)} = \frac{N}{\exp\left(sa\tanh(au)/u\right) D(s)}. \tag{A.14}$$

To prove the original inequality in eq. (A.12), it is now sufficient to show:

$$D(s) \cdot \exp\left(\frac{a}{u}\tanh(au)s + \frac{a^2}{u^2}s^2\right) \geq N. \tag{A.15}$$

For $s = 0$, the inequality holds trivially, as:

$$D(0) \cdot 1 = 2\cosh(au) = N. \tag{A.16}$$

For $s \neq 0$, taking the derivative of the left-hand side of eq. (A.15) yields:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}s}\left(D(s) \cdot \exp\left(\frac{a}{u}\tanh(au)s + \frac{a^2}{u^2}s^2\right)\right) \\
&= \exp\left(\frac{a}{u}\tanh(au)s + \frac{a^2}{u^2}s^2\right) \cdot \left(-\frac{2a}{\sqrt{\Delta}}\sinh\left(a\sqrt{\Delta}\right) + 2\cosh\left(a\sqrt{\Delta}\right) \cdot \left(\frac{a}{u}\tanh(au) + 2\frac{a^2}{u^2}s\right)\right) \\
&= 2\exp\left(\frac{a}{u}\tanh(au)s + \frac{a^2}{u^2}s^2\right)\cosh\left(a\sqrt{\Delta}\right) \cdot \left(-\frac{a}{\sqrt{\Delta}}\tanh\left(a\sqrt{\Delta}\right) + \frac{a}{u}\tanh(au) + 2\frac{a^2}{u^2}s\right).
\end{aligned} \tag{A.17}$$

In step 2, we will prove the following inequality:

$$-\frac{a}{\sqrt{\Delta}}\tanh\left(a\sqrt{\Delta}\right) + \frac{a}{u}\tanh(au) + 2\frac{a^2}{u^2}s \begin{cases} \geq 0, & \forall s \geq 0, \\ < 0, & \forall s < 0, \end{cases} \tag{A.18}$$

[2] In Cox [339, eq. (128)], setting $a = 2a$ and $x_0 = a$ leads to the desired result.

Equation (A.18) implies that $D(s) \cdot \exp\left(\frac{a}{u}\tanh(au)s + \frac{a^2}{u^2}s^2\right) \geq N$, which finishes the proof.

Step 2. Prove eq. (A.18).

For $s \geq 0$, the following holds:

$$\begin{aligned}
&-\frac{a}{\sqrt{\Delta}}\tanh\left(a\sqrt{\Delta}\right) + \frac{a}{u}\tanh(au) + 2\frac{a^2}{u^2}s \\
\overset{(i)}{\geq}\ & a\tanh\left(a\sqrt{\Delta}\right)\left(\frac{1}{u} - \frac{1}{\sqrt{\Delta}}\right) + 2\frac{a^2}{u^2}s \\
=\ & a\tanh\left(a\sqrt{\Delta}\right)\frac{-2s}{u\sqrt{\Delta}\left(\sqrt{\Delta}+u\right)} + 2\frac{a^2}{u^2}s \\
=\ & -2s\cdot\frac{a^2}{u\left(\sqrt{\Delta}+u\right)}\cdot\frac{\tanh\left(a\sqrt{\Delta}\right)}{a\sqrt{\Delta}} + 2\frac{a^2}{u^2}s \\
\overset{(ii)}{\geq}\ & -2s\frac{a^2}{u^2}\cdot 1 + 2\frac{a^2}{u^2}s \\
=\ & 0.
\end{aligned} \tag{A.19}$$

Here, (i) follows from $\tanh(au) \geq \tanh(a\sqrt{\Delta}) = \tanh(a\sqrt{u^2-2s})$ and (ii) follows from $\tanh(x)/x \leq 1$.

For $s < 0$, the following holds:

$$\begin{aligned}
&-\frac{a}{\sqrt{\Delta}}\tanh\left(a\sqrt{\Delta}\right) + \frac{a}{u}\tanh(au) + 2\frac{a^2}{u^2}s \\
\overset{(i)}{\leq}\ & a\tanh\left(a\sqrt{\Delta}\right)\left(\frac{1}{u} - \frac{1}{\sqrt{\Delta}}\right) + 2\frac{a^2}{u^2}s \\
=\ & -2s\cdot\frac{a^2}{u\left(\sqrt{\Delta}+u\right)}\cdot\frac{\tanh\left(a\sqrt{\Delta}\right)}{a\sqrt{\Delta}} + 2\frac{a^2}{u^2}s \\
\overset{(ii)}{\leq}\ & -2s\frac{a^2}{u^2}\cdot 1 + 2\frac{a^2}{u^2}s \\
=\ & 0.
\end{aligned} \tag{A.20}$$

Here, (i) follows from $\tanh(au) \leq \tanh(a\sqrt{\Delta}) = \tanh(a\sqrt{u^2-2s})$ and (ii) follows from $\tanh(x)/x \leq 1$.

By combining both cases, we conclude that the inequality in eq. (A.18) holds, which completes Step 2 and proves the desired result. □

Next, we prove theorem 4.2.3, which provides the non-asymptotic concentration inequality for the estimator from eq. (4.5), restated as follows:

Theorem 4.2.3 (Non-asymptotic concentration of $\widehat{u}_{x,\mathrm{CH,DT}}$)**.** *For each query $x \in \mathcal{X}$ with $u_x \neq 0$, given a fixed i.i.d. dataset, denoted by $\left\{\left(c_{x,s_{x,i}}, t_{x,s_{x,i}}\right)\right\}_{i\in[n_x]}$, for any $\epsilon > 0$ satisfying*

$\epsilon \leq \min\left\{|u_x|/(\sqrt{2}a), \left(1+\sqrt{2}\right) a|u_x|/\mathbb{E}\left[t_x\right]\right\}$, *the following holds:*

$$\mathbb{P}\left(\left|\widehat{u}_{x,CH,DT} - \frac{u_x}{a}\right| > \epsilon\right) \leq 4\exp\left(-\left[m^{\textit{non-asym}}_{CH,DT}\left(x^\top\theta^*\right)\right]^2 n_x \left[\epsilon\cdot a\right]^2\right),$$

where $m^{\textit{non-asym}}_{CH,DT}\left(x^\top\theta^*\right) \coloneqq \mathbb{E}\left[t_x\right] \,/\, \left[(2+2\sqrt{2})\; a\right]$.

Proof. For clarity, we will omit the subscripts x throughout this proof. Based on lemma A.2.1, we define the constants $\nu \coloneqq \sqrt{2}a/|u|$ and $\alpha \coloneqq 2/u^2$.

We begin by introducing $\epsilon_{\mathcal{C}} \coloneqq \mathcal{T}/\left(\sqrt{2}+\sqrt{2}\nu|\mathcal{C}|/\mathcal{T}\right)\cdot\epsilon$ and $\epsilon_{\mathcal{T}} \coloneqq \nu\epsilon_{\mathcal{C}}$. From the identities provided in appendix A.2.1, we know that $\nu|\mathcal{C}|/\mathcal{T} = \sqrt{2}a/|u|\cdot|u|/a = \sqrt{2}$. This allows us to simplify the constants $\epsilon_{\mathcal{C}}$ and $\epsilon_{\mathcal{T}}$ as:

$$\epsilon_{\mathcal{C}} = \frac{\mathcal{T}}{\sqrt{2}\left(\sqrt{2}+1\right)}\epsilon \quad\text{and}\quad \epsilon_{\mathcal{T}} = \frac{\nu\mathcal{T}}{\sqrt{2}\left(\sqrt{2}+1\right)}\epsilon. \tag{A.21}$$

For any ϵ satisfying the following condition:

$$\epsilon \leq \min\left\{\frac{1}{\nu}, \frac{\sqrt{2}(1+\sqrt{2})\nu}{\alpha\mathcal{T}}\right\}, \tag{A.22}$$

we observe that $\epsilon_{\mathcal{T}} < \min\left\{\mathcal{T}(1-1/\sqrt{2}), \nu^2/\alpha\right\}$. We can now apply lemma A.2.2 to derive the following:

$$\mathbb{P}\left(\left|\widehat{\mathcal{T}} - \mathcal{T}\right| > \epsilon_{\mathcal{T}}\right) \leq 2\exp\left(-\frac{n\epsilon_{\mathcal{T}}^2}{2\nu^2}\right). \tag{A.23}$$

Thus, by combining the results, we conclude:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| > \sqrt{2}\frac{\epsilon_{\mathcal{C}} + \epsilon_{\mathcal{T}}\cdot|\mathcal{C}|/\mathcal{T}}{\mathcal{T}}\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(\left|\widehat{\mathcal{C}} - \mathcal{C}\right| > \epsilon_{\mathcal{C}}\right) + \mathbb{P}\left(\left|\widehat{\mathcal{T}} - \mathcal{T}\right| > \epsilon_{\mathcal{T}}\right) \\
&\overset{(ii)}{\leq} 2\exp\left(-\frac{n\epsilon_{\mathcal{C}}^2}{2}\right) + 2\exp\left(-\frac{n\epsilon_{\mathcal{T}}^2}{2\nu^2}\right) \\
&\overset{(iii)}{=} 4\exp\left(-\frac{n\epsilon_{\mathcal{C}}^2}{2}\right) \\
&= 4\exp\left(-\frac{\mathcal{T}^2}{4\left(1+\sqrt{2}\right)^2}\cdot n\epsilon^2\right).
\end{aligned} \tag{A.24}$$

Here, (i) follows from lemma A.2.3, (ii) uses lemma A.2.2 and eq. (A.23), and (iii) follows from eq. (A.21). □

Supporting Details

Lemma A.2.2. *For each query x with $u_x \neq 0$, and constants $\epsilon_{\mathcal{C}} > 0$ and $\epsilon_{\mathcal{T}} \in (0, \nu_x^2/\alpha_x]$, the following inequalities hold:*

$$\mathbb{P}\left(\left|\widehat{\mathcal{C}}_x - \mathcal{C}_x\right| \geq \epsilon_{\mathcal{C}}\right) \leq 2\exp\left(-\frac{n\epsilon_{\mathcal{C}}^2}{2}\right), \quad \mathbb{P}\left(\left|\widehat{\mathcal{T}}_x - \mathcal{T}_x\right| \geq \epsilon_{\mathcal{T}}\right) \leq 2\exp\left(-\frac{n\epsilon_{\mathcal{T}}^2}{2\nu_x^2}\right). \tag{A.25}$$

Here, the constants are $\nu_x \coloneqq \sqrt{2}a/|u_x|$ and $\alpha_x \coloneqq 2/u_x^2$.

Proof. Since $c_x \in \{-1, 1\}$, by applying Hoeffding's inequality [338, proposition 2.5], we obtain:

$$\mathbb{P}\left(\left|\widehat{\mathcal{C}}_x - \mathcal{C}_x\right| \geq \epsilon_{\mathcal{C}}\right) \leq 2\exp\left(-\frac{n\epsilon_{\mathcal{C}}^2}{2}\right). \tag{A.26}$$

From lemma A.2.1, we know that t_x is sub-exponential $SE(\nu_x^2, \alpha_x)$. By applying Wainwright [338, proposition 2.9 and eq. (2.18)], we obtain:

$$\mathbb{P}\left(\left|\widehat{\mathcal{T}}_x - \mathcal{T}_x\right| \geq \epsilon_{\mathcal{T}}\right) \leq 2\exp\left(-\frac{n\epsilon_{\mathcal{T}}^2}{2\nu_x^2}\right), \quad \forall \epsilon_{\mathcal{T}} \in (0, \nu_x^2/\alpha_x]. \tag{A.27}$$

□

Lemma A.2.3. *Consider constants $\mathcal{C} \in \mathbb{R}$, $\mathcal{T} > 0$, $\epsilon_{\mathcal{C}} > 0$, and $\epsilon_{\mathcal{T}} \in \left(0, (1 - 1/\sqrt{2})\mathcal{T}\right)$. For any $\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_{\mathcal{C}}, \mathcal{C} + \epsilon_{\mathcal{C}}]$ and $\widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]$, the following inequality holds*

$$\left|\frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| \leq \sqrt{2}\frac{\epsilon_{\mathcal{C}} + \epsilon_{\mathcal{T}} \cdot |\mathcal{C}|/\mathcal{T}}{\mathcal{T}}. \tag{A.28}$$

Proof. The maximum value of $\left|\widehat{\mathcal{C}}/\widehat{\mathcal{T}} - \mathcal{C}/\mathcal{T}\right|$ is attained at the extremum of $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$. Since $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$ is linear in $\widehat{\mathcal{C}}$, the extremum of $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$ is attained at $C^* \in \{\mathcal{C} - \epsilon_{\mathcal{C}}, \mathcal{C} + \epsilon_{\mathcal{C}}\}$ for any $\widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}] > 0$. Given that $\widehat{\mathcal{T}} > 0$, the extremum of $C^*/\widehat{\mathcal{T}}$ is attained at $T^* \in \{\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}\}$. Therefore, the extremum of $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$ lies in the set:

$$\max_{\substack{\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_{\mathcal{C}}, \mathcal{C} + \epsilon_{\mathcal{C}}] \\ \widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]}} \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} \in \left\{\frac{\mathcal{C} - \epsilon_{\mathcal{C}}}{\mathcal{T} - \epsilon_{\mathcal{T}}}, \quad \frac{\mathcal{C} - \epsilon_{\mathcal{C}}}{\mathcal{T} + \epsilon_{\mathcal{T}}}, \quad \frac{\mathcal{C} + \epsilon_{\mathcal{C}}}{\mathcal{T} - \epsilon_{\mathcal{T}}}, \quad \frac{\mathcal{C} + \epsilon_{\mathcal{C}}}{\mathcal{T} + \epsilon_{\mathcal{T}}}\right\}. \tag{A.29}$$

For any combination $(s_{\mathcal{C}}, s_{\mathcal{T}}) \in \{\pm 1\} \times \{\pm 1\}$, and using the function $\epsilon_{\mathcal{T}} \leq (1 - 1/\sqrt{2})\mathcal{T}$, we have:

$$\left|\frac{\mathcal{C} + s_{\mathcal{C}}\epsilon_{\mathcal{C}}}{\mathcal{T} + s_{\mathcal{T}}\epsilon_{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| = \left|\frac{s_{\mathcal{C}}\epsilon_{\mathcal{C}}\mathcal{T} - s_{\mathcal{T}}\epsilon_{\mathcal{T}}\mathcal{C}}{\mathcal{T}\left(\mathcal{T} + s_{\mathcal{T}}\epsilon_{\mathcal{T}}\right)}\right| \leq \frac{\epsilon_{\mathcal{C}}\mathcal{T} + \epsilon_{\mathcal{T}}|\mathcal{C}|}{\mathcal{T}\left(\mathcal{T} - \epsilon_{\mathcal{T}}\right)} \leq \sqrt{2}\frac{\epsilon_{\mathcal{C}}\mathcal{T} + \epsilon_{\mathcal{T}}|\mathcal{C}|}{\mathcal{T}^2}. \tag{A.30}$$

By combining these results, we conclude that:

$$\max_{\substack{\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_{\mathcal{C}}, \mathcal{C} + \epsilon_{\mathcal{C}}] \\ \widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]}} \left|\frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| = \max_{(s_{\mathcal{C}}, s_{\mathcal{T}}) \in \{\pm 1\} \times \{\pm 1\}} \left|\frac{\mathcal{C} + s_{\mathcal{C}}\epsilon_{\mathcal{C}}}{\mathcal{T} + s_{\mathcal{T}}\epsilon_{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}}\right| \leq \sqrt{2}\frac{\epsilon_{\mathcal{C}} + \epsilon_{\mathcal{T}}|\mathcal{C}|/\mathcal{T}}{\mathcal{T}}.$$

□

The choice-only estimator

We now apply the logistic-regression-based choice-only estimator from eq. (4.4) to estimate the utility difference for a single query. Recall that for each query $x \in \mathcal{X}$, the human choice $c_x \in \{-1, 1\}$. We define the binary-encoded choice as $e_x \coloneqq (c_x + 1)/2 \in \{0, 1\}$. We reformulate the MLE in eq. (4.4) into a utility difference estimation problem for a single query, leading to the following optimization problem:

$$\begin{aligned}\widehat{u}_{x,\mathrm{CH}} &= \arg\max_{u \in \mathbb{R}} \sum_{i \in [n_x]} \log \mu(c_{x,s_{x,i}}\, u) \\ &= \arg\max_{u \in \mathbb{R}} \sum_{i \in [n_x]} \log \left[(\mu(u))^{e_{x,s_{x,i}}} \cdot (\mu(-u))^{1-e_{x,s_{x,i}}} \right].\end{aligned}$$

The first-order optimality condition provides the optimal solution:

$$\widehat{u}_{x,\mathrm{CH}} = \mu^{-1} \left(\frac{1}{n_x} \sum_{i \in [n_x]} e_{x,s_{x,i}} \right) \qquad \text{(restating eq. (4.6))},$$

where $\mu^{-1}(p) \coloneqq \log(p/(1-p))$ is the logit function (also known as the log-odds), defined as the inverse of the function $\mu(\cdot)$ introduced in eq. (4.4).

The resulting estimate, $\widehat{u}_{x,\mathrm{CH}}$, from eq. (4.6) gives an estimate of $2au_x$, not u_x. However, since the ranking of arm utilities based on $2au_x$ is the same as that based on the true u_x, estimating $2au_x$ suffices for identifying the best arm.

The non-asymptotic concentration inequality for this estimator is stated in theorem 4.2.4. This result is directly adapted from Jun et al. [258, theorem 5], by letting $x_1 = \cdots = x_t = 1$ and $t_{\text{eff}} = d = 1$.

A.3 Experiment details

Our empirical experiments (section 4.4) were conducted on a MacBook Pro (M3 Pro, Nov 2023) with 36 GB of memory.

Our implementation is available via https://shenlirobot.github.io/pages/NeurIPS24.html. The code is written in Julia and builds on the implementation by Tirinzoni and Degenne [340], where the transductive and weak-preference designs are solved using the Frank–Wolfe algorithm [253]. Their code is accessible at https://github.com/AndreaTirinzoni/bandit-elimination. Simulations and Bayesian inference for the DDM are implemented using the Julia package `SequentialSamplingModels.jl`, available at https://itsdfish.github.io/SequentialSamplingModels.jl/dev/#SequentialSamplingModels.jl.

For a query $x \in \mathcal{X}$, the estimators from Wagenmakers, Van Der Maas, and Grasman [43] and Xiang Chiong et al. [270], analyzed in section 4.2.3 and benchmarked in section 4.4.2, require calculating $\mu^{-1}(p) \coloneqq \log\left(p/\left(1-p\right)\right)$, where $\mu^{-1}(\cdot)$ is the logit function and $p \coloneqq 1/n_x \cdot \sum_{i=1}^{n_x} \left(c_{x,s_{x,i}} + 1\right)/2$ represents the empirical mean of the human binary choices coded as 0 or 1. Since $p = 0$ or $p = 1$ makes this calculation undefined, we follow Wagenmakers, Van Der Maas, and Grasman [43, the discussion below fig. 6] and approximate p as $1 - 1/(2n_x)$ when $p = 1$ and $1/(2n_x)$ when $p = 0$.

A.3.1 The "Sphere" Synthetic Problem for Evaluating Estimation Performance in section 4.4.1

We evaluate estimation performance using the "sphere" synthetic problem, a standard benchmark in the linear bandit literature [272, 274, 275]. In this problem, the arm space $\mathcal{Z} \subset \{z \in \mathbb{R}^5 \colon \|z\|_2 = 1\}$ contains 10 randomly generated arms. To define the true preference vector θ^*, we select the two arms z and z' that are closest in direction, i.e., $(z, z') \in \arg\max_{z,z' \in \mathcal{Z}} z^\top z'$, and set $\theta^* = z + 0.01(z' - z)$. In this way, z is the best arm. The query space is $\mathcal{X} \coloneqq \{z - z' \colon z \in \mathcal{Z}\}$.

A.3.2 Processing the food-risk dataset with choices (-1 or 1) [44]

We accessed the food-risk dataset with choices (-1 or 1) [44] through Yang and Krajbich [256]'s repository (https://osf.io/d7s6c/). This dataset includes the choices and response times of 42 participants, each responding to between 60 and 200 queries. Each query compares two arms, with each arm containing two food items. By selecting an arm, participants had an equal chance of receiving either food item, hence the name "food risk" (or "food-gamble") task. Additionally, participants' eye movements were tracked during the experiment. Yang and Krajbich [256] modeled each participant's choices, response times, and eye movements using the attentional DDM [46], where the drift for each query is a linear combination of the participant's ratings of the four food items in the query, with the weights adjusting based on their eye movements. The ratings, $\in \{-10, -9, \ldots, 0, \ldots, 9, 10\}$, were collected before the participants interacted with the binary queries.

In our work, for each participant, we define each arm's feature vector as the participant's ratings of the two corresponding food items, augmented with second-order polynomials. We fit each participant's data to a difference-based EZ-diffusion model [43, 251] with a linear utility structure, as introduced in section 4.1. For each participant, using Bayesian inference with non-informative priors [45], we estimated the preference vector $\theta^* \in \mathbb{R}^5$, non-decision time t_{nondec}, and barrier a. Across participants, the barrier a ranged from 0.715 to 2.467, with a mean of 1.437, and t_{nondec} ranged from 0.206 to 1.917 seconds, with a mean of 0.746 seconds. This procedure generated one bandit instance per participant, with a preference vector $\theta^* \in \mathbb{R}^5$, an arm space $\mathcal{Z} \subset \mathbb{R}^5$ where $|\mathcal{Z}| \in [31, 95]$, and a query space $\mathcal{X} \coloneqq \{z - z' \colon z \in \mathcal{Z}\}$. Then, we used the fitted models to simulate human feedback for bandit experiments.

For each bandit instance, we benchmarked the following six GSE variations (introduced in section 4.4.2): $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, and $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$. For each GSE variation, we ran 300 repeated simulations under different random seeds, with human choices and response times sampled from the dEZDM with the identified parameters. Since each bandit instance contains a different number of arms, rather than tuning the elimination parameter η in algorithm 1 for each instance, we set $\eta = 2$, following the convention in previous bandit research, e.g., Azizi, Kveton, and Ghavamzadeh [257, section 3]. We manually tuned the buffer size B_{buff} in algorithm 1 to 20, 30, or 50 seconds based on empirical performance, ensuring the budget was not exceeded in each phase. The full results are shown in fig. A.1, with selected results highlighted in fig. 4.4a.

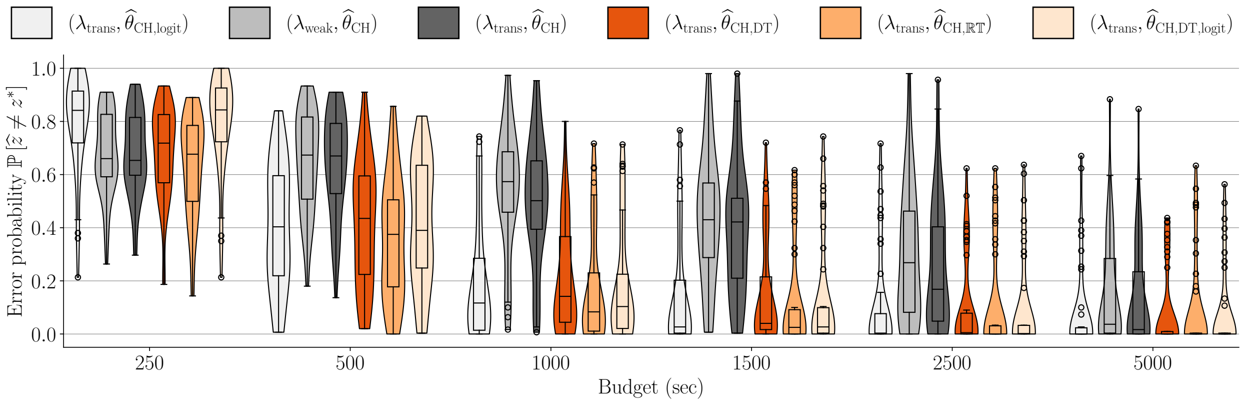


Figure A.1: A violin plot overlaid with a box plot showing the best-arm identification error probability, $\mathbb{P}[\widehat{z} \neq z^*]$, as a function of budget for each GSE variation, simulated using the food-risk dataset with choices (-1 or 1) [44], as described in appendix A.3.2. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within 1.5× the interquartile range. Flier points indicate outliers beyond the whiskers.

A.3.3 Processing the snack dataset with choices (yes or no) [45]

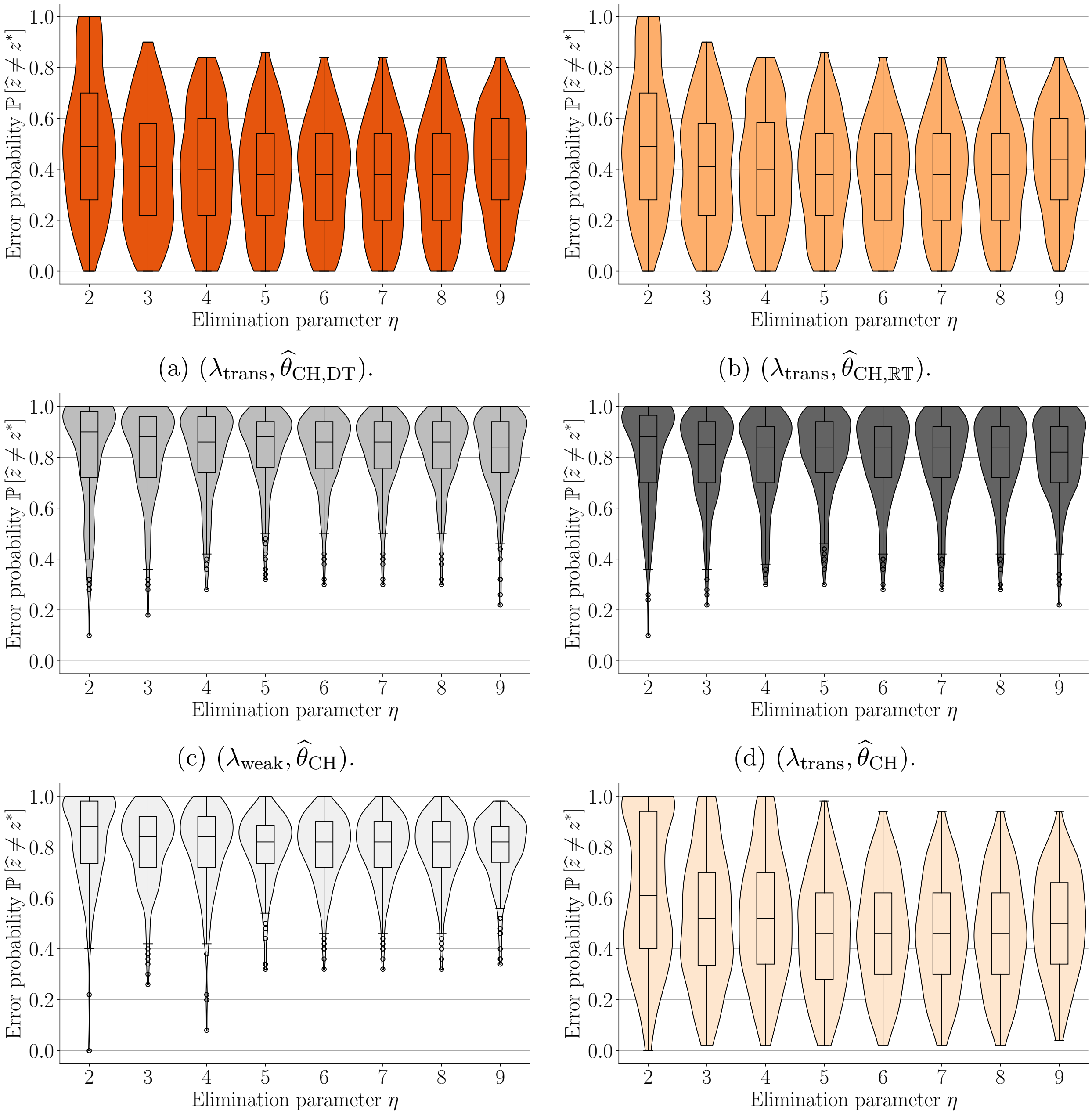
We accessed the snack dataset with choices (yes or no) [45] through the supplementary material provided by Alós-Ferrer, Fehr, and Netzer [240] at https://www.journals.uchicago.edu/doi/abs/10.1086/713732. This dataset consists of training and testing data. The training data was collected from a "YN" task, where 31 participants provided binary feedback ("Yes" or "No") and response times for queries comparing each of the 17 snack items to a fixed reference snack, with each query repeated 10 times. The reference snack, assigned a utility of 0, remained fixed throughout the experiment. The testing data was collected using a two-alternative forced-choice task, where participants provided binary choices and response times for queries comparing two snack items, with each query repeated once. Clithero [45] fit a difference-based EZ-diffusion model [43, 251] to the training data using Bayesian inference with non-informative priors, without imposing a linear utility structure, and tested the model using the testing data.

In our work, we fit each participant's training data to a difference-based EZ-diffusion model with a linear utility structure, as described in section 4.1, and used the fitted model to simulate human feedback for bandit experiments. We pre-processed the data by removing outliers, following Clithero [45, footnote 22], excluding trials with response times below 200 ms or greater than five standard deviations above the mean. After cleaning, the number of trials per participant ranged from 167 to 170. Since the dataset does not provide feature vectors for the 17 non-reference snack items, we used one-hot encoding to represent each snack item as a feature vector in $\mathbb{R}^{17}$. This allowed us to construct a bandit instance for each participant with a preference vector $\theta^* \in \mathbb{R}^{17}$, an arm space $\mathcal{Z} \subset \mathbb{R}^{17}$ with $|\mathcal{Z}| = 17$, and a query space $\mathcal{X} := \{z - \mathbf{0} \colon z \in \mathcal{Z}\}$ to represent comparisons with the reference snack. We applied Bayesian inference with non-informative priors [45] to estimate each participant's preference vector θ^*, non-decision time t_{nondec}, and barrier a. Across participants, the barrier a ranged from 0.759 to 1.399, with a mean of 1.1, and t_{nondec} ranged from 0.139 to 0.485 seconds, with a mean of 0.367 seconds.

For each of the following six GSE variations (introduced in section 4.4.2): $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, and $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$, we tuned the elimination parameter η in algorithm 1 using the following procedure: We considered $\eta \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, resulting in the number of phases $:= \lceil \log_\eta |\mathcal{Z}| \rceil = \lceil \log_\eta (17) \rceil$ (line 4 of algorithm 1) being $\{5, 3, 3, 2, 2, 2, 2, 2\}$, respectively. We excluded $\eta > \lceil 17/2 \rceil = 9$, as those cases also result in 2 phases, the same as $\eta \in \{5, 6, 7, 8, 9\}$. Then, for each η, for each of the 31 bandit instances, and for each time budget $\in \{50, 75, 100, 125, 150, 200, 250, 300\}$ seconds, we ran 50 repeated simulations per GSE variation under different random seeds, sampling human feedback from the fitted dEZDM. We then aggregated the results into a single best-arm identification error probability for each GSE variation, η, bandit instance, and budget. These error probabilities were compiled into violin and box plots, as shown in fig. A.2.

For each GSE variation, we selected the η that minimized the median error probability, as shown in the box plots in fig. A.2. If multiple η values yielded the same median, we used the third quartile, and if necessary, the first quartile, to break ties. Based on this approach, we selected: $\eta = 6$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $\eta = 6$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, $\eta = 9$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, $\eta = 9$ for $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, $\eta = 9$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, and $\eta = 5$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$.

After tuning η, we manually set the buffer size B_{buff} in algorithm 1 to 10 seconds based on empirical results, ensuring the budget was not exceeded in any phase. We then benchmarked each GSE variation on all 31 bandit instances using its own manually tuned η and B_{buff}. Each variation was evaluated over 300 repeated simulations with different random seeds, where human choices and response times were sampled from the dEZDM with the identified parameters. The full results are shown in fig. A.3, with selected results presented in fig. 4.4b.



(a) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$.

(b) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$.

(c) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$.

(d) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$.

(e) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$.

(f) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$.

Figure A.2: Violin plots overlaid with box plots, used for tuning the elimination parameter η in algorithm 1 for each GSE variation, simulated based on the snack dataset with choices (yes or no) [45], as discussed in appendix A.3.3. Each plot shows the best-arm identification error probability, $\mathbb{P}[\widehat{z} \neq z^*]$, as a function of η. The box plots follow the convention of the `matplotlib` Python package. The horizontal line in each box represents the median of the error probabilities across all bandit instances and budgets. Each error probability is averaged over 50 repeated simulations under different random seeds. The top and bottom borders of the box represent the third and first quartiles, respectively, while the whiskers extend to the farthest points within 1.5× the interquartile range. Flier points are the outliers past the end of the whiskers.

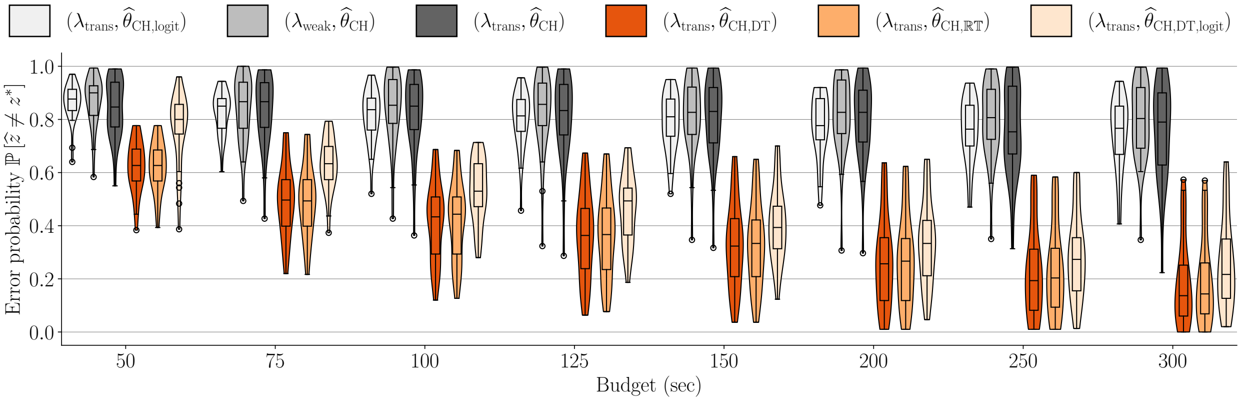


Figure A.3: A violin plot overlaid with a box plot showing the best-arm identification error probability, $\mathbb{P}\left[\widehat{z} \neq z^*\right]$, as a function of budget for each GSE variation, simulated using the snack dataset with choices (yes or no) [45], as described in appendix A.3.3. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within 1.5× the interquartile range. Flier points indicate outliers beyond the whiskers.

A.3.4 Processing the snack dataset with choices (-1 or 1) [46]

We accessed the snack dataset with choices (-1 or 1) [46] via Fudenberg, Strack, and Strzalecki [330]'s replication package at https://www.aeaweb.org/articles?id=10.1257/aer.20150742. This dataset contains choices and response times from 39 participants, each responding to between 49 and 100 queries comparing two snack items. Participants' eye movements were tracked during the experiment. Krajbich, Armel, and Rangel [46] modeled each participant's choices, response times, and eye movements using the attentional DDM, where the drift for each query is a linear combination of the participant's ratings of both snack items in the query, with the weights influenced by their eye movements. The ratings, $\in \{-10,-9,\dots,0,\dots,9,10\}$, were collected before participants interacted with the binary queries.

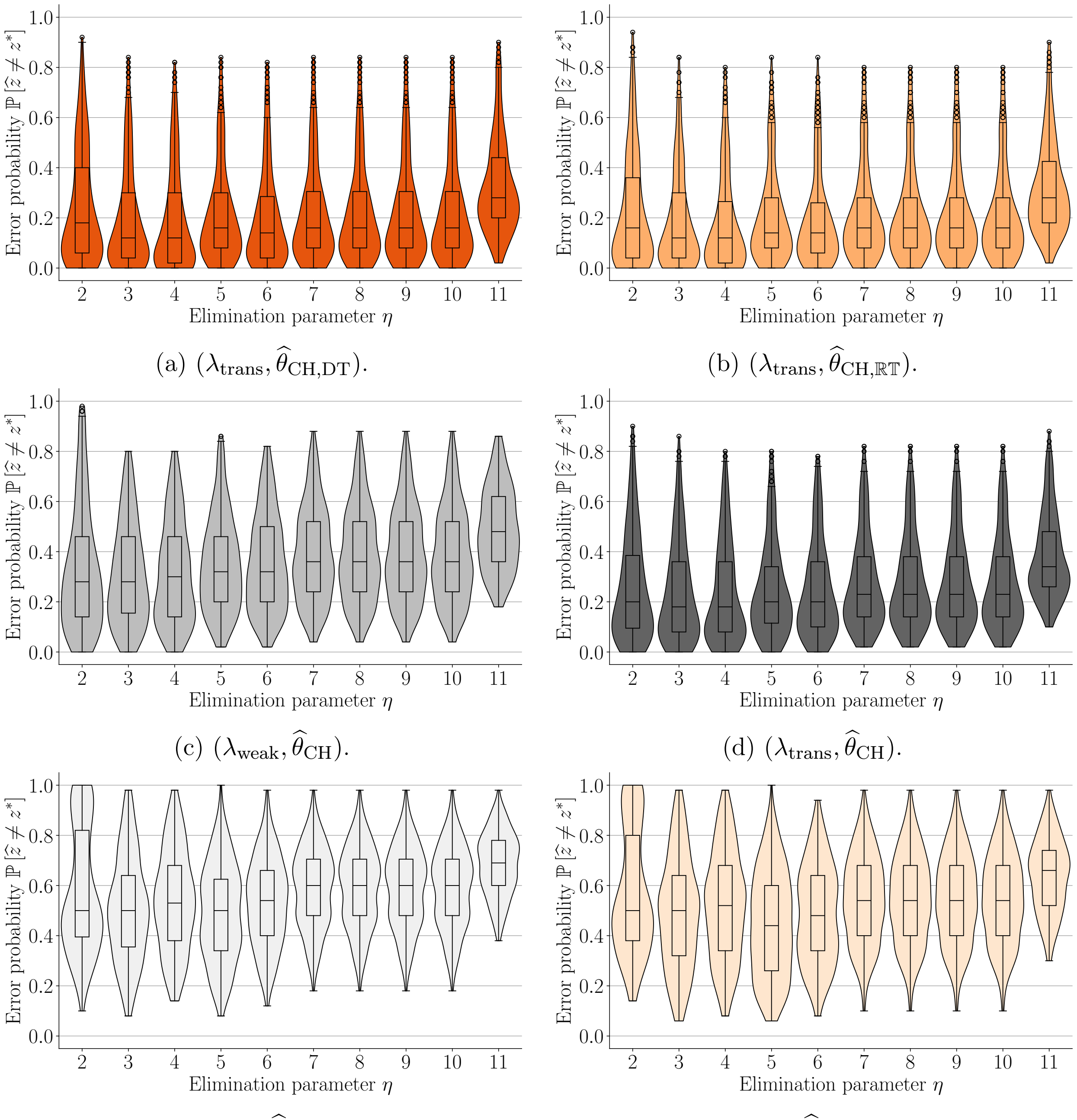
In our work, to avoid creating trivial bandit problems by encoding snack items as 1-dimensional vectors (as done in appendix A.3.2), we defined the feature vector for each snack item with a participant rating $r_z \in \{-10,-9,\dots,0,\dots,9,10\}$ as a one-hot vector in $\mathbb{R}^{21}$, where the (r_z+11)-th element is 1 and the rest are 0. The preference vector θ^* is structured as $\beta^* \cdot [-10,-9,\dots,0,\dots,9,10]^\top \in \mathbb{R}^{21}$, where β^* is participant-specific and unknown to the learner. This ensures that, for each arm z, the participant's utility is $u_z \coloneqq z^\top\theta^* = r_z\beta^*$. In this way, each participant's data generated a bandit instance with a preference vector $\theta^* \in \mathbb{R}^{21}$, a set of arms $\mathcal{Z} \subset \mathbb{R}^{21}$ with $|\mathcal{Z}| = 21$, and a query space $\mathcal{X} \coloneqq \{z - z' \colon z \in \mathcal{Z}\}$.

We fit each participant's data to a difference-based EZ-diffusion model [43, 251] using the linear utility structure described above. For each participant, using Bayesian inference with non-informative priors [45], we estimated the preference vector θ^* (or equivalently, the parameter β^*), non-decision time t_{nondec}, and barrier a. Across participants, the barrier a ranged from 0.75 to 2.192 with a mean of 1.335, and t_{nondec} ranged from 0.387 to 1.22 seconds with a mean of 0.641 seconds. We then used these fitted models to simulate human feedback for bandit experiments, assuming the learner did not know the underlying structure $\theta^* = \beta^* \cdot [-10,-9,\dots,0,\dots,9,10]^\top$.

For each of the following GSE variations (introduced in section 4.4.2): $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, and $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$, we tuned the elimination parameter η in algorithm 1 using the following procedure: We considered $\eta \in \{2,3,4,5,6,7,8,9,10,11\}$, which resulted in the number of phases $\coloneqq \lceil \log_\eta |\mathcal{Z}| \rceil = \lceil \log_\eta(17) \rceil$ (line 4 of algorithm 1) being $\{5,3,3,2,2,2,2,2,2,2\}$, respectively. We excluded cases where $\eta > \lceil 21/2 \rceil = 11$, as these result in 2 phases, identical to when $\eta \in \{5,6,7,8,9,10,11\}$. Then, for each η, for each of the 39 bandit instances, and also for each time budget $\in \{150,200,250,300,350,400,450,500\}$ seconds, we ran 50 repeated simulations per GSE variation under different random seeds, sampling human feedback from the fitted dEZDM. We then aggregated the results into a single best-arm identification error probability for each GSE variation, η, bandit instance, and budget. These error probabilities were compiled into violin and box plots, as shown in fig. A.4.

For each GSE variation, we selected the η that minimized the median error probability, as shown in the box plots in fig. A.4. If multiple η values yielded the same median, we used the third quartile, and if necessary, the first quartile, to break ties. Based on this approach, we selected: $\eta = 4$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $\eta = 4$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, $\eta = 4$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, $\eta = 2$ for $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, $\eta = 5$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, and $\eta = 5$ for $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$.

After tuning η, we manually set the buffer size B_{buff} in algorithm 1 to 20 seconds based on empirical results, ensuring the budget was not exceeded in any phase. We then benchmarked each GSE variation on all 39 bandit instances using its own manually tuned η. Each variation was evaluated over 300 repeated simulations with different random seeds, where human choices and response times were sampled from the dEZDM with the identified parameters. The full results are shown in fig. A.5, with selected results presented in fig. 4.4c.



(a) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$.

(b) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$.

(c) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$.

(d) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$.

(e) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$.

(f) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$.

Figure A.4: Violin plots overlaid with box plots, used for tuning the elimination parameter η in algorithm 1 for each GSE variation, simulated based on the snack dataset with choices (-1 or 1) [46], as discussed in appendix A.3.4. Each plot shows the best-arm identification error probability, $\mathbb{P}[\widehat{z} \neq z^*]$, as a function of η. The box plots follow the convention of the `matplotlib` Python package. The horizontal line in each box represents the median of the error probabilities across all bandit instances and budgets. Each error probability is averaged over 50 repeated simulations under different random seeds. The top and bottom borders of the box represent the third and first quartiles, respectively, while the whiskers extend to the farthest points within 1.5× the interquartile range. Flier points are the outliers past the end of the whiskers.

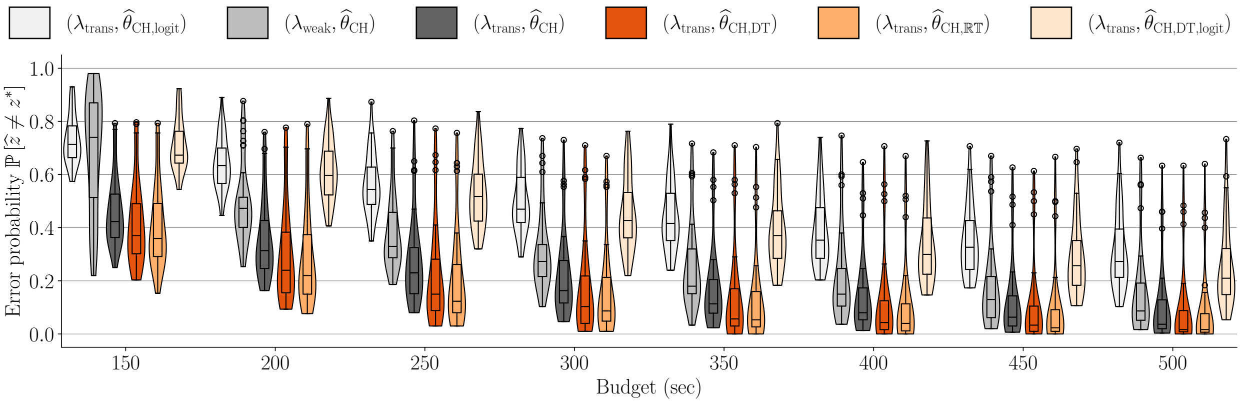


Figure A.5: A violin plot overlaid with a box plot showing the best-arm identification error probability, $\mathbb{P}[\widehat{z} \neq z^*]$, as a function of budget for each GSE variation, simulated using the snack dataset with choices (-1 or 1) [46], as described in appendix A.3.4. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within 1.5× the interquartile range. Flier points indicate outliers beyond the whiskers.

References

[1] R. Picard. “What every engineer and computer scientist should know: the biggest contributor to happiness”. In: *Communications of the ACM* 64.12 (2021), pp. 40–42.

[2] U. N. P. Division. *World population prospects: The 2024 revision.* 2024. URL: https://population.un.org/wpp/.

[3] World Health Organization. *10 Facts on Disability.* Accessed: 2025-05-08. 2023. URL: https://www.who.int/news-room/facts-in-pictures/detail/disabilities.

[4] W. H. Organization et al. “Global strategy on human resources for health: workforce 2030”. In: *Global strategy on human resources for health: workforce 2030.* 2016.

[5] E. Hertog, S. Fukuda, R. Matsukura, N. Nagase, and V. Lehdonvirta. “The future of unpaid work: Estimating the effects of automation on time spent on housework and care work in Japan and the UK”. In: *Technological Forecasting and Social Change* 191 (2023), p. 122443.

[6] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers, and A. Sandygulova. “Perceived safety in physical human–robot interaction–A survey”. In: *Robotics and Autonomous Systems* 151 (2022), p. 104047.

[7] N. Akalin, A. Kiselev, A. Kristoffersson, and A. Loutfi. “A taxonomy of factors influencing perceived safety in human–robot interaction”. In: *International Journal of Social Robotics* 15.12 (2023), pp. 1993–2004.

[8] D. Gopinath, S. Jain, and B. D. Argall. “Human-in-the-loop optimization of shared autonomy in assistive robotics”. In: *IEEE Robotics and Automation Letters* 2.1 (2016), pp. 247–254.

[9] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell. “Shared autonomy via hindsight optimization for teleoperation and teaming”. In: *The International Journal of Robotics Research* 37.7 (2018), pp. 717–742.

[10] A. Kapusta, Z. Erickson, H. M. Clever, W. Yu, C. K. Liu, G. Turk, and C. C. Kemp. “Personalized collaborative plans for robot-assisted dressing via optimization and simulation”. In: *Autonomous Robots* 43 (2019), pp. 2183–2207.

[11] E. Pignat and S. Calinon. “Learning adaptive dressing assistance from human demonstration”. In: *Robotics and Autonomous Systems* 93 (2017), pp. 61–75.

[12] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa. “Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding”. In: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 2020, pp. 181–190.

[13] G. Swamy, J. Schulz, R. Choudhury, D. Hadfield-Menell, and A. Dragan. “On the Utility of Model Learning in HRI”. In: *arXiv preprint arXiv:1901.01291* (2019). URL: https://arxiv.org/abs/1901.01291.

[14] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal. “A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 687–694.

[15] B. Hayes and B. Scassellati. “Autonomously constructing hierarchical task networks for planning and human-robot collaboration”. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 5469–5476.

[16] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah. “Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time”. In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 2394–2401.

[17] T. Iqbal, S. Li, C. Fourie, B. Hayes, and J. A. Shah. “Fast Online Segmentation of Activities from Partial Trajectories”. In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*. 2019.

[18] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus. “Social behavior for autonomous vehicles”. In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[19] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić. “Object handovers: a review for robotics”. In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 1855–1873.

[20] S. Reddy, A. D. Dragan, and S. Levine. “Shared autonomy via deep reinforcement learning”. In: *arXiv preprint arXiv:1802.01744* (2018).

[21] J. Zhang, P. Fiers, K. A. Witte, R. W. Jackson, K. L. Poggensee, C. G. Atkeson, and S. H. Collins. “Human-in-the-loop optimization of exoskeleton assistance during walking”. In: *Science* 356.6344 (2017), pp. 1280–1284.

[22] F. Khadivar, V. Mendez, C. Correia, I. Batzianoulis, A. Billard, and S. Micera. “EMG-driven shared human-robot compliant control for in-hand object manipulation in hand prostheses”. In: *Journal of Neural Engineering* 19.6 (2022), p. 066024.

[23] F. Zhang and Y. Demiris. “Learning garment manipulation policies toward robot-assisted dressing”. In: *Science robotics* 7.65 (2022), eabm6010.

[24] D. Park, Y. Hoshi, H. P. Mahajan, H. K. Kim, Z. Erickson, W. A. Rogers, and C. C. Kemp. “Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned”. In: *Robotics and Autonomous Systems* 124 (2020), p. 103344.

[25] S. Li, N. Figueroa, A. Shah, and J. Shah. "Provably Safe and Efficient Motion Planning with Uncertain Human Dynamics". In: *Proceedings of the Robotics: Science and Systems (RSS)*. 2021.

[26] S. Li*, T. Stouraitis*, M. Gienger, S. Vijayakumar, and J. A. Shah. "Set-based state estimation with probabilistic consistency guarantee under epistemic uncertainty". In: *IEEE Robotics and Automation Letters* (2022).

[27] Y. Gao, H. J. Chang, and Y. Demiris. "Iterative path optimisation for personalised dressing assistance using vision and force information". In: *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2016, pp. 4398–4403.

[28] G. Canal, E. Pignat, G. Alenyà, S. Calinon, and C. Torras. "Joining high-level symbolic planning with low-level motion primitives in adaptive HRI: application to dressing assistance". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 3273–3278.

[29] N. Koganti, T. Shibata, T. Tamei, and K. Ikeda. "Data-efficient learning of robotic clothing assistance using Bayesian Gaussian process latent variable model". In: *Advanced Robotics* 33.15-16 (2019), pp. 800–814.

[30] S. Li*, Y. Zhang*, Z. Ren, C. Liang, N. Li, and J. A. Shah. "Enhancing Preference-based Linear Bandits via Human Response Time". In: *Advances in Neural Information Processing Systems* (2024).

[31] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia. "Active Preference-Based Learning of Reward Functions". In: *Proceedings of Robotics: Science and Systems*. Cambridge, Massachusetts, July 2017. DOI: 10.15607/RSS.2017.XIII.053.

[32] M. Tucker, E. Novoseller, C. Kann, Y. Sui, Y. Yue, J. W. Burdick, and A. D. Ames. "Preference-Based Learning for Exoskeleton Gait Optimization". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2020, pp. 2351–2357. DOI: 10.1109/ICRA40945.2020.9196661.

[33] M. Karimi, D. Jannach, and M. Jugovac. "News recommender systems – Survey and roads ahead". In: *Information Processing & Management* 54.6 (2018), pp. 1203–1227. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2018.04.008. URL: https://www.sciencedirect.com/science/article/pii/S030645731730153X.

[34] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha. "Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions". In: *Expert Systems with Applications* 197 (2022), p. 116669. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116669. URL: https://www.sciencedirect.com/science/article/pii/S0957417422001543.

[35] V. Bogina, T. Kuflik, D. Jannach, M. Bielikova, M. Kompan, and C. Trattner. "Considering temporal aspects in recommender systems: a survey". In: *User Modeling and User-Adapted Interaction* 33.1 (2023), pp. 81–119. DOI: 10.1007/s11257-022-09335-w. URL: https://doi.org/10.1007/s11257-022-09335-w.

[36] Y. Deldjoo, M. Schedl, and P. Knees. "Content-driven music recommendation: Evolution, state of the art, and challenges". In: *Computer Science Review* 51 (2024), p. 100618. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2024.100618. URL: https://www.sciencedirect.com/science/article/pii/S1574013724000029.

[37] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 3008–3021. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[38] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, et al. "Teaching language models to support answers with verified quotes". In: *arXiv preprint arXiv:2203.11147* (2022).

[39] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. "Webgpt: Browser-assisted question-answering with human feedback". In: *arXiv preprint arXiv:2112.09332* (2021).

[40] L. Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[41] Y. Bai et al. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. 2022. arXiv: 2204.05862 [cs.CL]. URL: https://arxiv.org/abs/2204.05862.

[42] R. Ratcliff. "A theory of memory retrieval." In: *Psychological review* (1978). [PDF].

[43] E.-J. Wagenmakers, H. L. Van Der Maas, and R. P. Grasman. "An EZ-diffusion model for response time and accuracy". In: *Psychonomic bulletin & review* (2007). [PDF].

[44] S. M. Smith and I. Krajbich. "Attention and choice across domains." In: *Journal of Experimental Psychology: General* 147.12 (2018), p. 1810.

[45] J. A. Clithero. "Improving out-of-sample predictions using response times and a model of the decision process". In: *Journal of Economic Behavior & Organization* 148 (2018), pp. 344–375. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2018.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0167268118300398.

[46] I. Krajbich, C. Armel, and A. Rangel. "Visual fixations and the computation and comparison of value in simple choice". In: *Nature Neuroscience* 13.10 (2010), pp. 1292–1298. DOI: 10.1038/nn.2635. URL: https://doi.org/10.1038/nn.2635.

[47] M. Tucker, M. Cheng, E. Novoseller, R. Cheng, Y. Yue, J. W. Burdick, and A. D. Ames. "Human preference-based learning for high-dimensional optimization of exoskeleton walking gaits". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 3423–3430.

[48] M. Tucker, N. Csomay-Shanklin, W.-L. Ma, and A. D. Ames. "Preference-based learning for user-guided hzd gait generation on bipedal walking robots". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 2804–2810.

[49] P. Trautman and A. Krause. "Unfreezing the robot: Navigation in dense, interacting crowds". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 797–803.

[50] S. Haddadin, A. De Luca, and A. Albu-Schäffer. "Robot collisions: A survey on detection, isolation, and identification". In: *IEEE Transactions on Robotics* (2017).

[51] P. A. Lasota, T. Fong, J. A. Shah, et al. "A survey of methods for safe human-robot interaction". In: *Foundations and Trends® in Robotics* 5.4 (2017), pp. 261–349.

[52] T. Koller, F. Berkenkamp, M. Turchetta, J. Boedecker, and A. Krause. *Learning-based Model Predictive Control for Safe Exploration and Reinforcement Learning*. 2019. URL: https://arxiv.org/abs/1906.12189.

[53] W. B. Knox and P. Stone. "Interactively shaping agents via human reinforcement: The TAMER framework". In: *Proceedings of the fifth international conference on Knowledge capture*. 2009, pp. 9–16.

[54] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard. "Recent advances in robot learning from demonstration". In: *Annual review of control, robotics, and autonomous systems* 3 (2020), pp. 297–330.

[55] V. V. Unhelkar*, S. Li*, and J. A. Shah. "Semi-Supervised Learning of Decision-Making Models for Human-Robot Collaboration". In: *Proceedings of the Conference on Robot Learning (CoRL)*. 2020.

[56] V. V. Unhelkar*, S. Li*, and J. A. Shah. "Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2020.

[57] R. M. Aronson, N. Almutlak, and H. Admoni. "Inferring goals with gaze during teleoperated manipulation". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 7307–7314.

[58] R. Zhang, F. Torabi, L. Guan, D. H. Ballard, and P. Stone. "Leveraging human guidance for deep reinforcement learning tasks". In: *arXiv preprint arXiv:1909.09906* (2019).

[59] W. B. Knox and P. Stone. "Combining manual feedback with subsequent MDP reward signals for reinforcement learning." In: *AAMAS*. 2010, pp. 5–12.

[60] W. B. Knox and P. Stone. "Reinforcement learning from simultaneous human and MDP reward." In: *AAMAS*. Vol. 1004. Valencia. 2012, pp. 475–482.

[61] W. B. Knox and P. Stone. "Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance". In: *Artificial Intelligence* 225 (2015), pp. 24–50.

[62] M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil. "Teaching with rewards and punishments: Reinforcement or communication?" In: *CogSci*. 2015.

[63] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone. "Deep tamer: Interactive agent shaping in high-dimensional state spaces". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[64] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman. "Interactive learning from policy-dependent human feedback". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2285–2294.

[65] Y. Efroni, N. Merlis, and S. Mannor. "Reinforcement learning with trajectory feedback". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 7288–7295.

[66] N. Chatterji, A. Pacchiano, P. Bartlett, and M. Jordan. "On the theory of reinforcement learning with once-per-episode feedback". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3401–3412.

[67] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[68] E. Biyik and D. Sadigh. "Batch active preference-based learning of reward functions". In: *Conference on robot learning*. PMLR. 2018, pp. 519–528.

[69] E. Bıyık, D. A. Lazar, D. Sadigh, and R. Pedarsani. "The green choice: Learning and influencing human decisions on shared roads". In: *2019 IEEE 58th conference on decision and control (CDC)*. IEEE. 2019, pp. 347–354.

[70] E. Bıyık, M. Palan, N. C. Landolfi, D. P. Losey, and D. Sadigh. "Asking easy questions: A user-friendly approach to active reward learning". In: *arXiv preprint arXiv:1910.04365* (2019).

[71] C. Basu, E. Bıyık, Z. He, M. Singhal, and D. Sadigh. "Active learning of reward dynamics from hierarchical queries". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 120–127.

[72] E. Bıyık, N. Huynh, M. J. Kochenderfer, and D. Sadigh. "Active preference-based gaussian process regression for reward learning". In: *arXiv preprint arXiv:2005.02575* (2020).

[73] V. Myers, E. Biyik, N. Anari, and D. Sadigh. "Learning multimodal rewards from rankings". In: *Conference on Robot Learning*. PMLR. 2022, pp. 342–352.

[74] N. Wilde, E. Biyik, D. Sadigh, and S. L. Smith. "Learning Reward Functions from Scale Feedback". In: *Proceedings of the 5th Conference on Robot Learning*. Ed. by A. Faust, D. Hsu, and G. Neumann. Vol. 164. Proceedings of Machine Learning Research. PMLR, Aug. 2022, pp. 353–362. URL: https://proceedings.mlr.press/v164/wilde22a.html.

[75] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh. “Learning reward functions by integrating human demonstrations and preferences”. In: *arXiv preprint arXiv:1906.08928* (2019).

[76] C. Basu, M. Singhal, and A. D. Dragan. “Learning from richer human guidance: Augmenting comparison-based learning with feature queries”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, pp. 132–140.

[77] K. Lee, L. Smith, and P. Abbeel. “Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training”. In: *arXiv preprint arXiv:2106.05091* (2021).

[78] J. Hejna and D. Sadigh. “Few-Shot Preference Learning for Human-in-the-Loop RL”. In: *arXiv preprint arXiv:2212.03363* (2022).

[79] D. Brown, W. Goo, P. Nagarajan, and S. Niekum. “Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations”. In: *International conference on machine learning*. PMLR. 2019, pp. 783–792.

[80] D. S. Brown, W. Goo, and S. Niekum. “Better-than-demonstrator imitation learning via automatically-ranked demonstrations”. In: *Conference on robot learning*. PMLR. 2020, pp. 330–359.

[81] L. Chen, R. Paleja, and M. Gombolay. “Learning from suboptimal demonstration via self-supervised reward regression”. In: *Conference on robot learning*. PMLR. 2021, pp. 1262–1277.

[82] D. Brown, R. Coleman, R. Srinivasan, and S. Niekum. “Safe imitation learning via fast bayesian reward inference from preferences”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1165–1177.

[83] D. Shin, D. S. Brown, and A. D. Dragan. “Offline preference-based apprenticeship learning”. In: *arXiv preprint arXiv:2107.09251* (2021).

[84] E. Novoseller, Y. Wei, Y. Sui, Y. Yue, and J. Burdick. “Dueling posterior sampling for preference-based reinforcement learning”. In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1029–1038.

[85] Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski. “Preference-based reinforcement learning with finite-time guarantees”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18784–18794.

[86] A. Pacchiano, A. Saha, and J. Lee. “Dueling rl: reinforcement learning with trajectory preferences”. In: *arXiv preprint arXiv:2111.04850* (2021).

[87] X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. “Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3773–3793.

[88] S. Akbarzadeh, E. Lobarinas, and N. Kehtarnavaz. “Online personalization of compression in hearing aids via maximum likelihood inverse reinforcement learning”. In: *IEEE Access* 10 (2022), pp. 58537–58546.

[89] K. Li, M. Tucker, E. Bıyık, E. Novoseller, J. W. Burdick, Y. Sui, D. Sadigh, Y. Yue, and A. D. Ames. “Roial: Region of interest active learning for characterizing exoskeleton gait preference landscapes”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 3212–3218.

[90] R. Cosner, M. Tucker, A. Taylor, K. Li, T. Molnar, W. Ubelacker, A. Alan, G. Orosz, Y. Yue, and A. Ames. “Safety-Aware Preference-Based Learning for Safety-Critical Control”. In: *Learning for Dynamics and Control Conference*. PMLR. 2022, pp. 1020–1033.

[91] N. Csomay-Shanklin, M. Tucker, M. Dai, J. Reher, and A. D. Ames. “Learning controller gains on bipedal walking robots via user preferences”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 10405–10411.

[92] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. “Dynamical movement primitives: learning attractor models for motor behaviors”. In: *Neural computation* 25.2 (2013), pp. 328–373.

[93] A. Billard, S. Mirrazavi, and N. Figueroa. *Learning for Adaptive and Reactive Robot Control: A Dynamical Systems Approach*. Mit Press, 2022.

[94] J. Ho and S. Ermon. “Generative adversarial imitation learning”. In: *Advances in neural information processing systems* 29 (2016).

[95] S. Ross, G. Gordon, and D. Bagnell. “A reduction of imitation learning and structured prediction to no-regret online learning”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 627–635.

[96] D. P. Losey, A. Bajcsy, M. K. O’Malley, and A. D. Dragan. “Physical interaction as communication: Learning robot objectives online from human corrections”. In: *The International Journal of Robotics Research* 41.1 (2022), pp. 20–44.

[97] A. Bajcsy, D. P. Losey, M. K. O’Malley, and A. D. Dragan. “Learning from physical human corrections, one feature at a time”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, pp. 141–149.

[98] D. P. Losey and M. K. O’Malley. “Including uncertainty when learning from human corrections”. In: *Conference on Robot Learning*. PMLR. 2018, pp. 123–132.

[99] A. Bobu, A. Bajcsy, J. F. Fisac, and A. D. Dragan. “Learning under misspecified objective spaces”. In: *Conference on Robot Learning*. PMLR. 2018, pp. 796–805.

[100] Y. Cui and S. Niekum. “Active reward learning from critiques”. In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 6907–6914.

[101] D. P. Losey and M. K. O’Malley. “Learning the correct robot trajectory in real-time from physical human interactions”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.1 (2019), pp. 1–19.

[102] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. “Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections”. In: *IEEE Transactions on Robotics* 36.3 (2020), pp. 835–854.

[103] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa. “Learning from interventions: Human-robot interaction as both explicit and implicit feedback”. In: *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals. 2020.

[104] M. Li, A. Canberk, D. P. Losey, and D. Sadigh. “Learning human objectives from sequences of physical corrections”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 2877–2883.

[105] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan. “Inducing structure in reward learning by learning features”. In: *The International Journal of Robotics Research* 41.5 (2022), pp. 497–518.

[106] Q. Li, Z. Peng, and B. Zhou. “Efficient learning of safe driving policy via human-ai copilot optimization”. In: *arXiv preprint arXiv:2202.10341* (2022).

[107] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu. “Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment”. In: *arXiv preprint arXiv:2211.08416* (2022).

[108] Y. Gao, H. J. Chang, and Y. Demiris. “User modelling using multimodal information for personalised dressing assistance”. In: *IEEE Access* 8 (2020), pp. 45700–45714.

[109] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler. “Human preferences for robot-human hand-over configurations”. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2011, pp. 1986–1993.

[110] I. Batzianoulis, F. Iwane, S. Wei, C. G. P. R. Correia, R. Chavarriaga, J. d. R. Millán, and A. Billard. “Customizing skills for assistive robotic manipulators, an inverse reinforcement learning approach with error-related potentials”. In: *Communications biology* 4.1 (2021), p. 1406.

[111] P. Slade, M. J. Kochenderfer, S. L. Delp, and S. H. Collins. “Personalizing exoskeleton assistance while walking in the real world”. In: *Nature* 610.7931 (2022), pp. 277–282.

[112] J. Gao, S. Reddy, G. Berseth, N. Hardy, N. Natraj, K. Ganguly, A. D. Dragan, and S. Levine. “X2T: Training an x-to-text typing interface with online learning from user feedback”. In: *arXiv preprint arXiv:2203.02072* (2022).

[113] I. Hochberg, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and E. Yom-Tov. “A reinforcement learning system to encourage physical activity in diabetes patients”. In: *arXiv preprint arXiv:1605.04070* (2016).

[114] W. Felt, J. C. Selinger, J. M. Donelan, and C. D. Remy. “" Body-In-The-Loop": Optimizing device parameters using measures of instantaneous energetic cost”. In: *PloS one* 10.8 (2015), e0135342.

[115] M. Kim, Y. Ding, P. Malcolm, J. Speeckaert, C. J. Siviy, C. J. Walsh, and S. Kuindersma. “Human-in-the-loop Bayesian optimization of wearable device parameters”. In: *PloS one* 12.9 (2017), e0184054.

[116] T.-C. Wen, M. Jacobson, X. Zhou, H.-J. Chung, and M. Kim. "The personalization of stiffness for an ankle-foot prosthesis emulator using Human-in-the-loop optimization". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE. 2020, pp. 3431–3436.

[117] T. L. Wu, A. Murphy, C. Chen, and D. Kulić. "Human-in-the-loop auditory cueing strategy for gait modification". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 3521–3528.

[118] B. Clement, D. Roy, P.-Y. Oudeyer, and M. Lopes. "Multi-armed bandits for intelligent tutoring systems". In: *arXiv preprint arXiv:1310.3174* (2013).

[119] P.-A. Andersen, C. Kråkevik, M. Goodwin, and A. Yazidi. "Adaptive task assignment in online learning environments". In: *Proceedings of the 6th international conference on web intelligence, mining and semantics.* 2016, pp. 1–10.

[120] T. Mu, K. Goel, and E. Brunskill. "Program2Tutor: combining automatic curriculum generation with multi-armed bandits for intelligent tutoring systems". In: *Conference on Neural Information Processing Systems.* 2017.

[121] T. Mu, S. Wang, E. Andersen, and E. Brunskill. "Combining adaptivity with progression ordering for intelligent tutoring systems". In: *Proceedings of the Fifth Annual ACM Conference on Learning at Scale.* 2018, pp. 1–4.

[122] A. Segal, Y. Ben David, J. J. Williams, K. Gal, and Y. Shalom. "Combining difficulty ranking with multi-armed bandits to sequence educational content". In: *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19.* Springer. 2018, pp. 317–321.

[123] T. Mu, S. Wang, E. Andersen, and E. Brunskill. "Automatic adaptive sequencing in a webgame". In: *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17.* Springer. 2021, pp. 430–438.

[124] S. Spaulding, J. Shen, H. W. Park, and C. Breazeal. "Lifelong personalization via Gaussian process modeling for long-term HRI". In: *Frontiers in Robotics and AI* 8 (2021), p. 683066.

[125] A. S. Lan and R. G. Baraniuk. "A Contextual Bandits Framework for Personalized Learning Action Selection." In: *EDM.* 2016, pp. 424–429.

[126] T. Schodde, K. Bergmann, and S. Kopp. "Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* 2017, pp. 128–136.

[127] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. "Faster teaching via pomdp planning". In: *Cognitive science* 40.6 (2016), pp. 1290–1332.

[128] F. Ai, Y. Chen, Y. Guo, Y. Zhao, Z. Wang, G. Fu, and G. Wang. "Concept-Aware Deep Knowledge Tracing and Exercise Recommendation in an Online Learning System." In: *International Educational Data Mining Society* (2019).

[129] S. Nikolaidis, P. Lasota, R. Ramakrishnan, and J. Shah. "Improved human–robot team performance through cross-training, an approach inspired by human team training practices". In: *The International Journal of Robotics Research* 34.14 (2015), pp. 1711–1730.

[130] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa. "Trust-aware decision making for human-robot collaboration: Model learning and planning". In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.2 (2020), pp. 1–23.

[131] H. Modares, I. Ranatunga, F. L. Lewis, and D. O. Popa. "Optimized assistive human–robot interaction using reinforcement learning". In: *IEEE transactions on cybernetics* 46.3 (2015), pp. 655–667.

[132] A. Ghadirzadeh, J. Bütepage, A. Maki, D. Kragic, and M. Björkman. "A sensorimotor reinforcement learning framework for physical human-robot interaction". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 2682–2688.

[133] C. Moro, G. Nejat, and A. Mihailidis. "Learning and personalizing socially assistive robot behaviors to aid with activities of daily living". In: *ACM Transactions on Human-Robot Interaction (THRI)* 7.2 (2018), pp. 1–25.

[134] Y. Wen, J. Si, A. Brandt, X. Gao, and H. H. Huang. "Online reinforcement learning control for the personalization of a robotic knee prosthesis". In: *IEEE transactions on cybernetics* 50.6 (2019), pp. 2346–2356.

[135] A. Shafti, J. Tjomsland, W. Dudley, and A. A. Faisal. "Real-world human-robot collaborative reinforcement learning". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 11161–11166.

[136] Y. Wen, M. Li, J. Si, and H. Huang. "Wearer-prosthesis interaction for symmetrical gait: A study enabled by reinforcement learning prosthesis control". In: *IEEE transactions on neural systems and rehabilitation engineering* 28.4 (2020), pp. 904–913.

[137] X. Tu, M. Li, M. Liu, J. Si, and H. H. Huang. "A data-driven reinforcement learning solution framework for optimal and adaptive personalization of a hip exoskeleton". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 10610–10616.

[138] M. Li, Y. Wen, X. Gao, J. Si, and H. Huang. "Toward expedited impedance tuning of a robotic prosthesis for personalized gait assistance by reinforcement learning control". In: *IEEE Transactions on Robotics* 38.1 (2021), pp. 407–420.

[139] R. Wu, M. Li, Z. Yao, J. Si, et al. "Reinforcement learning enabled automatic impedance control of a robotic knee prosthesis to mimic the intact knee motion in a co-adapting environment". In: *arXiv preprint arXiv:2101.03487* (2021).

[140] W. Liu, R. Wu, J. Si, and H. Huang. "A New Robotic Knee Impedance Control Parameter Optimization Method Facilitated by Inverse Reinforcement Learning". In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 10882–10889.

[141] M. Li, X. Gao, Y. Wen, J. Si, and H. H. Huang. "Offline policy iteration based reinforcement learning controller for online robotic knee prosthesis parameter tuning". In: *2019 International conference on robotics and automation (ICRA)*. IEEE. 2019, pp. 2831–2837.

[142] S. Tesfazgi, A. Lederer, J. F. Kunz, A. J. Ordóñez-Conejo, and S. Hirche. "Personalized Rehabilitation Robotics based on Online Learning Control". In: *arXiv preprint arXiv:2110.00481* (2021).

[143] Y. Li and S. S. Ge. "Human–robot collaboration based on motion intention estimation". In: *IEEE/ASME Transactions on Mechatronics* 19.3 (2013), pp. 1007–1014.

[144] Y. Li and S. S. Ge. "Force tracking control for motion synchronization in human-robot collaboration". In: *Robotica* 34.6 (2016), pp. 1260–1281.

[145] S. Cremer, S. K. Das, I. B. Wijayasinghe, D. O. Popa, and F. L. Lewis. "Model-free online neuroadaptive controller with intent estimation for physical human–robot interaction". In: *IEEE Transactions on Robotics* 36.1 (2019), pp. 240–253.

[146] X. Yu, W. He, Y. Li, C. Xue, J. Li, J. Zou, and C. Yang. "Bayesian estimation of human impedance and motion intention for human–robot collaboration". In: *IEEE transactions on cybernetics* 51.4 (2019), pp. 1822–1834.

[147] A. Takagi, Y. Li, and E. Burdet. "Flexible Assimilation of Human's Target for Versatile Human-Robot Physical Interaction". In: *IEEE Transactions on Haptics* 14.2 (2020), pp. 421–431.

[148] Y. Li, L. Yang, D. Huang, C. Yang, and J. Xia. "A proactive controller for human-driven robots based on force/motion observer mechanisms". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.10 (2022), pp. 6211–6221.

[149] Z. Li, X. Li, Q. Li, H. Su, Z. Kan, and W. He. "Human-in-the-loop control of soft exosuits using impedance learning on different terrains". In: *IEEE Transactions on Robotics* 38.5 (2022), pp. 2979–2993.

[150] X. Gao, J. Si, Y. Wen, M. Li, and H. Huang. "Reinforcement learning control of robotic knee with human-in-the-loop by flexible policy iteration". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.10 (2021), pp. 5873–5887.

[151] R. Wu, Z. Yao, J. Si, and H. H. Huang. "Robotic knee tracking control to mimic the intact human knee profile based on actor-critic reinforcement learning". In: *IEEE/CAA Journal of Automatica Sinica* 9.1 (2021), pp. 19–30.

[152] S. Nikolaidis, D. Hsu, and S. Srinivasa. "Human-robot mutual adaptation in collaborative tasks: Models and experiments". In: *The International Journal of Robotics Research* 36.5-7 (2017), pp. 618–634.

[153] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa. "Human-robot mutual adaptation in shared autonomy". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 2017, pp. 294–302.

[154] S. Nikolaidis, M. Kwon, J. Forlizzi, and S. Srinivasa. "Planning with verbal communication for human-robot collaboration". In: *ACM Transactions on Human-Robot Interaction (THRI)* 7.3 (2018), pp. 1–21.

[155] J. S. Park, C. Park, and D. Manocha. "I-planner: Intention-aware motion planning using learning-based human motion prediction". In: *The International Journal of Robotics Research* 38.1 (2019), pp. 23–39.

[156] S. Jain and B. Argall. "Probabilistic human intent recognition for shared autonomy in assistive robotics". In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.1 (2019), pp. 1–23.

[157] A. Ghosh, S. Tschiatschek, H. Mahdavi, and A. Singla. "Towards deployment of robust cooperative ai agents: An algorithmic framework for learning adaptive policies". In: (2020).

[158] C. Z. Qiao, M. Sakr, K. Muelling, and H. Admoni. "Learning from demonstration for real-time user goal prediction and shared assistive control". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 3270–3275.

[159] A. Ghadirzadeh, X. Chen, W. Yin, Z. Yi, M. Björkman, and D. Kragic. "Human-centered collaborative robots with deep reinforcement learning". In: *IEEE Robotics and Automation Letters* 6.2 (2020), pp. 566–571.

[160] K. Backman, D. Kulić, and H. Chung. "Learning to assist drone landings". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 3192–3199.

[161] V. V. Unhelkar and J. A. Shah. "Learning Models of Sequential Decision-Making with Partial Specification of Agent Behavior". In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2019). [PDF].

[162] D. Gopinath, M. N. Javaremi, and B. Argall. "Customized handling of unintended interface operation in assistive robots". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 10406–10412.

[163] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh. "Learning latent representations to influence multi-agent interaction". In: *Conference on robot learning*. PMLR. 2021, pp. 575–588.

[164] W. Z. Wang, A. Shih, A. Xie, and D. Sadigh. "Influencing towards stable multi-agent interactions". In: *Conference on robot learning*. PMLR. 2022, pp. 1132–1143.

[165] S. Parekh and D. P. Losey. "Learning Latent Representations to Co-Adapt to Humans". In: *arXiv preprint arXiv:2212.09586* (2022).

[166] H. B. Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters. "Interaction primitives for human-robot cooperation tasks". In: *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2014, pp. 2831–2837.

[167] G. Maeda, M. Ewerton, R. Lioutikov, H. B. Amor, J. Peters, and G. Neumann. "Learning interaction for collaborative tasks with probabilistic movement primitives". In: *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE. 2014, pp. 527–534.

[168] J. Campbell and H. B. Amor. "Bayesian interaction primitives: A slam approach to human-robot interaction". In: *Conference on Robot Learning*. PMLR. 2017, pp. 379–387.

[169] J. Campbell, S. Stepputtis, and H. B. Amor. “Probabilistic multimodal modeling for human-robot interaction tasks”. In: *arXiv preprint arXiv:1908.04955* (2019).

[170] H. S. Koppula and A. Saxena. “Anticipating human activities using object affordances for reactive robotic response”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2015), pp. 14–29.

[171] J. Mainprice and D. Berenson. “Human-robot collaborative manipulation planning using early prediction of human motion”. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 299–306.

[172] C. Pérez-D’Arpino and J. A. Shah. “Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification”. In: *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2015, pp. 6175–6182.

[173] P. A. Lasota and J. A. Shah. “A multiple-predictor approach to human motion prediction”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 2300–2307.

[174] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone. “Multimodal probabilistic model-based planning for human-robot interaction”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 3399–3406.

[175] B. Ivanovic and M. Pavone. “The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2375–2384.

[176] B. Ivanovic, A. Elhafsi, G. Rosman, A. Gaidon, and M. Pavone. “Mats: An interpretable trajectory forecasting representation for planning and control”. In: *arXiv preprint arXiv:2009.07517* (2020).

[177] K. Leung, E. Schmerling, M. Zhang, M. Chen, J. Talbot, J. C. Gerdes, and M. Pavone. “On infusing reachability-based safety assurance within planning frameworks for human–robot vehicle interactions”. In: *The International Journal of Robotics Research* 39.10-11 (2020), pp. 1326–1345.

[178] A. D. Dragan and S. S. Srinivasa. “A policy-blending formalism for shared control”. In: *The International Journal of Robotics Research* 32.7 (2013), pp. 790–805.

[179] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan. “Planning for autonomous cars that leverage effects on human actions.” In: *Robotics: Science and systems*. Vol. 2. Ann Arbor, MI, USA. 2016, pp. 1–9.

[180] Y. Li, K. P. Tee, W. L. Chan, R. Yan, Y. Chua, and D. K. Limbu. “Continuous role adaptation for human–robot shared control”. In: *IEEE Transactions on Robotics* 31.3 (2015), pp. 672–681.

[181] Y. Li, K. P. Tee, R. Yan, W. L. Chan, and Y. Wu. “A framework of human–robot coordination based on game theory and policy iteration”. In: *IEEE Transactions on Robotics* 32.6 (2016), pp. 1408–1418.

[182] D. Fridovich-Keil, A. Bajcsy, J. F. Fisac, S. L. Herbert, S. Wang, A. D. Dragan, and C. J. Tomlin. “Confidence-aware motion prediction for real-time collision avoidance”. In: *The International Journal of Robotics Research* 39.2-3 (2020), pp. 250–265.

[183] R. Tian, M. Tomizuka, A. Dragan, and A. Bajcsy. “Towards Modeling and Influencing the Dynamics of Human Learning”. In: *arXiv preprint arXiv:2301.00901* (2023).

[184] H. Hu, K. Nakamura, and J. F. Fisac. “SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 5591–5598.

[185] H. Hu and J. F. Fisac. “Active uncertainty learning for human-robot interaction: An implicit dual control approach”. In: *arXiv preprint arXiv:2202.07720* (2022).

[186] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah. “Efficient model learning from joint-action demonstrations for human-robot collaborative tasks”. In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, pp. 189–196.

[187] N. Buckman, W. Schwarting, S. Karaman, and D. Rus. “Semi-Cooperative Control for Autonomous Emergency Vehicles”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 7052–7059.

[188] A. T. Le, P. Kratzer, S. Hagenmayer, M. Toussaint, and J. Mainprice. “Hierarchical Human-Motion Prediction and Logic-Geometric Programming for Minimal Interference Human-Robot Tasks”. In: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 7–14.

[189] L. Wang, Q. Li, J. Lam, Z. Wang, and Z. Zhang. “Intent inference in shared-control teleoperation system in consideration of user behavior”. In: *Complex & Intelligent Systems* (2021), pp. 1–11.

[190] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan. “Safety assurances for human-robot interaction via confidence-aware game-theoretic human models”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 11229–11235.

[191] J. R. Wright and K. Leyton-Brown. “Predicting human behavior in unrepeated, simultaneous-move games”. In: *Games and Economic Behavior* 106 (2017), pp. 16–37.

[192] N. Li, I. Kolmanovsky, A. Girard, and Y. Yildiz. “Game theoretic modeling of vehicle interactions at unsignalized intersections and application to autonomous vehicle control”. In: *2018 Annual American Control Conference (ACC)*. IEEE. 2018, pp. 3215–3220.

[193] R. Tian, S. Li, N. Li, I. Kolmanovsky, A. Girard, and Y. Yildiz. “Adaptive game-theoretic decision making for autonomous vehicle control at roundabouts”. In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 321–326.

[194] S. Li, N. Li, A. Girard, and I. Kolmanovsky. “Decision making in dynamic and interactive environments based on cognitive hierarchy theory, Bayesian inference, and predictive control”. In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 2181–2187.

[195] R. Tian, L. Sun, M. Tomizuka, and D. Isele. "Anytime game-theoretic planning with active reasoning about humans' latent states for human-centered robots". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 4509–4515.

[196] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa. "Game-theoretic modeling of human adaptation in human-robot collaboration". In: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 2017, pp. 323–331.

[197] D. P. Losey and D. Sadigh. "Robots that take advantage of human trust". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 7001–7008.

[198] J. Bragg and E. Brunskill. "Fake it till you make it: Learning-compatible performance support". In: *Uncertainty in artificial intelligence*. PMLR. 2020, pp. 915–924.

[199] M. Srivastava, E. Biyik, S. Mirchandani, N. Goodman, and D. Sadigh. "Assistive Teaching of Motor Control Tasks to Humans". In: *arXiv preprint arXiv:2211.14003* (2022).

[200] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. "Cooperative inverse reinforcement learning". In: *Advances in neural information processing systems* 29 (2016).

[201] R. Shah, P. Freire, N. Alex, R. Freedman, D. Krasheninnikov, L. Chan, M. D. Dennis, P. Abbeel, A. Dragan, and S. Russell. "Benefits of assistance over reward learning". In: (2020).

[202] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan. "The assistive multi-armed bandit". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 354–363.

[203] B. Yang, G. Habibi, P. Lancaster, B. Boots, and J. Smith. "Motivating Physical Activity via Competitive Human-Robot Interaction". In: *Conference on Robot Learning*. PMLR. 2022, pp. 839–849.

[204] J. F. Fisac, E. Bronstein, E. Stefansson, D. Sadigh, S. S. Sastry, and A. D. Dragan. "Hierarchical game-theoretic planning for autonomous vehicles". In: *2019 International conference on robotics and automation (ICRA)*. IEEE. 2019, pp. 9590–9596.

[205] S. Aoki, C.-W. Lin, and R. Rajkumar. "Human-robot cooperation for autonomous vehicles and human drivers: Challenges and solutions". In: *IEEE communications magazine* 59.8 (2021), pp. 35–41.

[206] E. OhnBar, K. Kitani, and C. Asakawa. "Personalized dynamics models for adaptive assistive navigation systems". In: *Conference on Robot Learning*. PMLR. 2018, pp. 16–39.

[207] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard. "Personalized machine learning for robot perception of affect and engagement in autism therapy". In: *Science Robotics* 3.19 (2018), eaao6760.

[208] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard. “Personalized estimation of engagement from videos using active learning with deep reinforcement learning”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2019, pp. 217–226.

[209] J. Z.-Y. He, A. Raghunathan, D. S. Brown, Z. Erickson, and A. D. Dragan. “Learning Representations that Enable Generalization in Assistive Tasks”. In: *arXiv preprint arXiv:2212.03175* (2022).

[210] M. L. Schrum, E. Sumner, M. C. Gombolay, and A. Best. “MAVERIC: A Data-Driven Approach to Personalized Autonomous Driving”. In: *arXiv preprint arXiv:2301.08595* (2023).

[211] C. Huang, W. Luo, and R. Liu. “Meta preference learning for fast user adaptation in human-supervisory multi-robot deployments”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 5851–5856.

[212] R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht. “A Survey of Ad Hoc Teamwork Research”. In: *Multi-Agent Systems: 19th European Conference, EUMAS 2022, Düsseldorf, Germany, September 14–16, 2022, Proceedings*. Springer. 2022, pp. 275–293.

[213] S. V. Albrecht and S. Ramamoorthy. “A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems”. In: *arXiv preprint arXiv:1506.01170* (2015).

[214] S. V. Albrecht and S. Ramamoorthy. “On convergence and optimality of best-response learning with policy types in multiagent systems”. In: *arXiv preprint arXiv:1907.06995* (2019).

[215] P. J. Gmytrasiewicz and P. Doshi. “A framework for sequential planning in multi-agent settings”. In: *Journal of Artificial Intelligence Research* 24 (2005), pp. 49–79.

[216] S. Barrett and P. Stone. “Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[217] H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. ““other-play” for zero-shot coordination”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4399–4410.

[218] D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. “Collaborating with humans without human data”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14502–14515.

[219] B. Cui, H. Hu, L. Pineda, and J. Foerster. “K-level reasoning for zero-shot coordination in hanabi”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8215–8228.

[220] J. Treutlein, M. Dennis, C. Oesterheld, and J. Foerster. “A new formalism, method and open issues for zero-shot coordination”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10413–10423.

[221] A. Shih, A. Sawhney, J. Kondic, S. Ermon, and D. Sadigh. “On the critical role of conventions in adaptive human-AI collaboration”. In: *arXiv preprint arXiv:2104.02871* (2021).

[222] T. Alamo, J. M. Bravo, and E. F. Camacho. “Guaranteed state estimation by zonotopes”. In: *Automatica* (2005).

[223] B. S. Rego, G. V. Raffo, J. K. Scott, and D. M. Raimondo. “Guaranteed methods based on constrained zonotopes for set-valued state estimation of nonlinear discrete-time systems”. In: *Automatica* (2020).

[224] M. Althoff. “Reachability analysis and its application to the safety assessment of autonomous cars”. PhD thesis. Technische Universität München, 2010.

[225] J. K. Scott, D. M. Raimondo, G. R. Marseglia, and R. D. Braatz. “Constrained zonotopes: A new tool for set-based estimation and fault detection”. In: *Automatica* (2016).

[226] Wikipedia contributors. *Ellipsoid — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Ellipsoid. Accessed: 2025-04-30. 2025.

[227] A. Kurzhanski and I. Vályi. *Ellipsoidal calculus for estimation and control*. Springer, 1997.

[228] I. Bogunovic, A. Krause, and J. Scarlett. “Corruption-tolerant Gaussian process bandit optimization”. In: *International Conference on Artificial Intelligence and Statistics*. 2020.

[229] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. “Information-theoretic regret bounds for gaussian process optimization in the bandit setting”. In: *Transactions on Information Theory* (2012).

[230] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. “Gaussian Process Bandits without Regret: An Experimental Design Approach”. In: *International Conference on Machine Learning*. 2010. URL: http://arxiv.org/abs/0912.3995.

[231] S. R. Chowdhury and A. Gopalan. “On kernelized multi-armed bandits”. In: *International Conference on Machine Learning*. 2017.

[232] F. Berkenkamp. “Safe exploration in reinforcement learning: Theory and applications in robotics”. PhD thesis. ETH Zurich, 2019.

[233] Y. Xu, H. Zhang, K. Miller, A. Singh, and A. Dubrawski. “Noise-Tolerant Interactive Learning Using Pairwise Comparisons”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2017/file/e11943a6031a0e6114ae69c257617980-Paper.pdf.

[234] P. Koppol, H. Admoni, and R. Simmons. “Interaction Considerations in Learning from Humans”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Z.-H. Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 283–291. DOI: 10.24963/ijcai.2021/40. URL: https://doi.org/10.24963/ijcai.2021/40.

[235] H. Yu, R. M. Aronson, K. H. Allen, and E. S. Short. "From "Thumbs Up" to "10 out of 10": Reconsidering Scalar Feedback in Interactive Reinforcement Learning". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2023, pp. 4121–4128. DOI: 10.1109/IROS55552.2023.10342458.

[236] T. Somers, N. R. Lawrance, and G. A. Hollinger. "Efficient learning of trajectory preferences using combined ratings and rankings". In: *Robotics: Science and Systems Conference Workshop on Mathematical Models, Algorithms, and Human-Robot Interaction*. 2017.

[237] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk. "From Pairwise Comparisons and Rating to a Unified Quality Scale". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 1139–1151. DOI: 10.1109/TIP.2019.2936103.

[238] P. Koppol, H. Admoni, and R. Simmons. "Iterative interactive reward learning". In: *Participatory Approaches to Machine Learning, International Conference on Machine Learning Workshop*. 2020.

[239] J. A. Clithero. "Response times in economics: Looking through the lens of sequential sampling models". In: *Journal of Economic Psychology* 69 (2018), pp. 61–86. ISSN: 0167-4870. DOI: https://doi.org/10.1016/j.joep.2018.09.008. URL: https://www.sciencedirect.com/science/article/pii/S0167487016306444.

[240] C. Alós-Ferrer, E. Fehr, and N. Netzer. "Time Will Tell: Recovering Preferences When Choices Are Noisy". In: *Journal of Political Economy* 129.6 (2021), pp. 1828–1877. DOI: 10.1086/713732. eprint: https://doi.org/10.1086/713732. URL: https://doi.org/10.1086/713732.

[241] P. De Boeck and M. Jeon. "An Overview of Models for Response Times and Processes in Cognitive Tests". In: *Frontiers in Psychology* 10 (2019). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00102. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00102.

[242] R. Ratcliff and G. McKoon. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks". In: *Neural Computation* 20.4 (Apr. 2008), pp. 873–922. ISSN: 0899-7667. DOI: 10.1162/neco.2008.12-06-420. eprint: https://direct.mit.edu/neco/article-pdf/20/4/873/817277/neco.2008.12-06-420.pdf. URL: https://doi.org/10.1162/neco.2008.12-06-420.

[243] M. Usher and J. L. McClelland. "The time course of perceptual choice: the leaky, competing accumulator model." In: *Psychological review* 108.3 (2001), p. 550.

[244] S. D. Brown and A. Heathcote. "The simplest complete model of choice response time: Linear ballistic accumulation". In: *Cognitive Psychology* 57.3 (2008), pp. 153–178. ISSN: 0010-0285. DOI: https://doi.org/10.1016/j.cogpsych.2007.12.002. URL: https://www.sciencedirect.com/science/article/pii/S0010028507000722.

[245] R. Webb. "The (Neural) Dynamics of Stochastic Choice". In: *Management Science* 65.1 (2019), pp. 230–255. DOI: 10.1287/mnsc.2017.2931. eprint: https://doi.org/10.1287/mnsc.2017.2931. URL: https://doi.org/10.1287/mnsc.2017.2931.

[246] T. V. Wiecki, I. Sofer, and M. J. Frank. "HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python". In: *Frontiers in Neuroinformatics* 7 (2013). ISSN: 1662-5196. DOI: 10.3389/fninf.2013.00014. URL: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00014.

[247] R. Ratcliff and F. Tuerlinckx. "Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability". In: *Psychonomic Bulletin & Review* 9.3 (2002), pp. 438–481. DOI: 10.3758/BF03196302. URL: https://doi.org/10.3758/BF03196302.

[248] E.-J. Wagenmakers, H. L. J. van der Maas, C. V. Dolan, and R. P. P. P. Grasman. "EZ does it! Extensions of the EZ-diffusion model". In: *Psychonomic Bulletin & Review* 15.6 (2008), pp. 1229–1235. DOI: 10.3758/PBR.15.6.1229. URL: https://doi.org/10.3758/PBR.15.6.1229.

[249] R. P. Grasman, E.-J. Wagenmakers, and H. L. van der Maas. "On the mean and variance of response times under the diffusion model with an application to parameter estimation". In: *Journal of Mathematical Psychology* 53.2 (2009), pp. 55–68. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2009.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0022249609000066.

[250] D. Fudenberg, W. Newey, P. Strack, and T. Strzalecki. "Testing the drift-diffusion model". In: *Proceedings of the National Academy of Sciences* 117.52 (2020), pp. 33141–33148. DOI: 10.1073/pnas.2011446117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2011446117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2011446117.

[251] R. Berlinghieri, I. Krajbich, F. Maccheroni, M. Marinacci, and M. Pirazzini. "Measuring utility with diffusion models". In: *Science Advances* 9.34 (2023), eadf1665. DOI: 10.1126/sciadv.adf1665. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adf1665. URL: https://www.science.org/doi/abs/10.1126/sciadv.adf1665.

[252] L. Li, W. Chu, J. Langford, and R. E. Schapire. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 661–670. ISBN: 9781605587998. DOI: 10.1145/1772690.1772758. URL: https://doi.org/10.1145/1772690.1772758.

[253] T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. "Sequential Experimental Design for Transductive Linear Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2019/file/8ba6c657b03fc7c8dd4dff8e45defcd2-Paper.pdf.

[254] J. S. Trueblood, S. D. Brown, and A. Heathcote. "The multiattribute linear ballistic accumulator model of context effects in multialternative choice." In: *Psychological review* 121.2 (2014), p. 179.

[255] G. Fisher. "An attentional drift diffusion model over binary-attribute choice". In: *Cognition* 168 (2017), pp. 34–45. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2017.06.007. URL: https://www.sciencedirect.com/science/article/pii/S0010027717301695.

[256] X. Yang and I. Krajbich. "A dynamic computational model of gaze and choice in multi-attribute decisions." In: *Psychological Review* 130.1 (2023), p. 52.

[257] M. Azizi, B. Kveton, and M. Ghavamzadeh. "Fixed-Budget Best-Arm Identification in Structured Bandits". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Ed. by L. D. Raedt. Main Track. International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 2798–2804. DOI: 10.24963/ijcai.2022/388. URL: https://doi.org/10.24963/ijcai.2022/388.

[258] K.-S. Jun, L. Jain, B. Mason, and H. Nassif. "Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 5148–5157. URL: https://proceedings.mlr.press/v139/jun21a.html.

[259] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. "Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2012/file/8b0d268963dd0cfb808aac48a549829f-Paper.pdf.

[260] E. Kaufmann, O. Cappé, and A. Garivier. "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models". In: *Journal of Machine Learning Research* 17.1 (2016), pp. 1–42. URL: http://jmlr.org/papers/v17/kaufman16a.html.

[261] V. Bengs, R. Busa-Fekete, A. E. Mesaoudi-Paul, and E. Hüllermeier. "Preference-based Online Learning with Dueling Bandits: A Survey". In: *Journal of Machine Learning Research* 22.7 (2021), pp. 1–108. URL: http://jmlr.org/papers/v22/18-546.html.

[262] J. Palmer, A. C. Huk, and M. N. Shadlen. "The effect of stimulus strength on the speed and accuracy of a perceptual decision". In: *Journal of Vision* 5.5 (May 2005), pp. 1–1. ISSN: 1534-7362. DOI: 10.1167/5.5.1. eprint: https://arvojournals.org/arvo/content_public/journal/jov/933510/jov-5-5-1.pdf. URL: https://doi.org/10.1167/5.5.1.

[263] R. A. Bradley and M. E. Terry. "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* (1952). [PDF].

[264] V. Lerche, A. Voss, and M. Nagler. "How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria". In: *Behavior Research Methods* 49.2 (2017), pp. 513–537. DOI: 10.3758/s13428-016-0740-2. URL: https://doi.org/10.3758/s13428-016-0740-2.

[265] A. Alieva, A. Cutkosky, and A. Das. "Robust Pure Exploration in Linear Bandits with Limited Budget". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 187–195. URL: https://proceedings.mlr.press/v139/alieva21a.html.

[266] J. Yang and V. Tan. "Minimax Optimal Fixed-Budget Best Arm Identification in Linear Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 12253–12266. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2022/file/4f9342b74c3bb63f6e030d8263082ab6-Paper-Conference.pdf.

[267] T. P. Minka. "A comparison of numerical optimizers for logistic regression". In: *Unpublished draft* (2003). eprint: https://tminka.github.io/papers/logreg/minka-logreg.pdf. URL: https://tminka.github.io/papers/logreg/minka-logreg.pdf.

[268] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. "Parametric Bandits: The Generalized Linear Case". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

[269] L. Fahrmeir and H. Kaufmann. "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models". In: *The Annals of Statistics* 13.1 (1985), pp. 342–368. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/2241164 (visited on 10/22/2024).

[270] K. Xiang Chiong, M. Shum, R. Webb, and R. Chen. "Combining Choice and Response Time Data: A Drift-Diffusion Model of Mobile Advertisements". In: *Management Science* 70.2 (2024), pp. 1238–1257. DOI: 10.1287/mnsc.2023.4738. eprint: https://doi.org/10.1287/mnsc.2023.4738. URL: https://doi.org/10.1287/mnsc.2023.4738.

[271] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. "Bandits with knapsacks". In: *Journal of the ACM (JACM)* 65.3 (2018), pp. 1–55.

[272] C. Tao, S. Blanco, and Y. Zhou. "Best Arm Identification in Linear Bandits with Linear Dimension Dependency". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4877–4886. URL: https://proceedings.mlr.press/v80/tao18a.html.

[273] R. Camilleri, K. Jamieson, and J. Katz-Samuels. "High-dimensional Experimental Design and Kernel Bandits". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 1227–1237. URL: https://proceedings.mlr.press/v139/camilleri21a.html.

[274] Z. Li, K. Jamieson, and L. Jain. "Optimal Exploration is no harder than Thompson Sampling". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by S. Dasgupta, S. Mandt, and Y. Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, Feb. 2024, pp. 1684–1692. URL: https://proceedings.mlr.press/v238/li24h.html.

[275] R. Degenne, P. Menard, X. Shang, and M. Valko. “Gamification of Pure Exploration for Linear Bandits”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 2432–2442. URL: https://proceedings.mlr.press/v119/degenne20a.html.

[276] F. Zhang, A. Cully, and Y. Demiris. “Probabilistic real-time user posture tracking for personalized robot-assisted dressing”. In: *IEEE Transactions on Robotics* 35.4 (2019), pp. 873–888.

[277] A. Xu and G. Dudek. “Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations”. In: *International Conference on Human-Robot Interaction*. 2015.

[278] R. Choudhury, G. Swamy, D. Hadfield-Menell, and A. D. Dragan. “On the utility of model learning in HRI”. In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. [PDF]. 2019.

[279] Y. Gal. “Uncertainty in Deep Learning”. PhD thesis. University of Cambridge, 2016.

[280] J. A. Castellanos, J. Neira, and J. D. Tardós. “Limits to the consistency of EKF-based SLAM”. In: *IFAC Proceedings Volumes* (2004).

[281] J. Ko and D. Fox. “GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models”. In: *Autonomous Robots* (2009). [PDF].

[282] S. J. Julier and J. K. Uhlmann. “A counter example to the theory of simultaneous localization and map building”. In: *International Conference on Robotics and Automation*. 2001.

[283] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis. “Analysis and improvement of the consistency of extended Kalman filter based SLAM”. In: *International Conference on Robotics and Automation*. 2008.

[284] S. Huang and G. Dissanayake. “Convergence and consistency analysis for extended Kalman filter based SLAM”. In: *Transactions on Robotics* (2007).

[285] A. Barrau and S. Bonnabel. “An EKF-SLAM algorithm with consistency properties”. In: *ArXiv* (2015).

[286] C. Combastel. “An Extended Zonotopic and Gaussian Kalman Filter (EZGKF) merging set-membership and stochastic paradigms: Toward non-linear filtering and fault detection”. In: *Annual Reviews in Control* (2016).

[287] C. Combastel. “Zonotopes and Kalman observers: Gain optimality under distinct uncertainty paradigms and robust convergence”. In: *Automatica* (2015).

[288] C. Combastel. “A state bounding observer for uncertain non-linear continuous-time systems based on zonotopes”. In: *Conference on Decision and Control*. 2005.

[289] A. Shetty and G. X. Gao. “Predicting State Uncertainty Bounds Using Non-Linear Stochastic Reachability Analysis for Urban GNSS-Based UAS Navigation”. In: *Transactions on Intelligent Transportation Systems* (2020).

[290] V. T. H. Le, C. Stoica, T. Alamo, E. F. Camacho, and D. Dumur. “Zonotope-based set-membership estimation for multi-output uncertain systems”. In: *International Symposium on Intelligent Control.* 2013.

[291] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[292] A. Lederer, J. Umlauft, and S. Hirche. “Uniform error bounds for gaussian process regression with application to safe control”. In: *Advances in Neural Information Processing Systems.* 2019.

[293] M. Althoff. “CORA 2021 Manual”. In: (2021).

[294] M. Tognon, R. Alami, and B. Siciliano. “Physical human-robot interaction with a tethered aerial vehicle: Application to a force-based human guiding problem”. In: *Transactions on Robotics* (2021).

[295] K. A. Wyrobek, E. H. Berger, H. M. Van der Loos, and J. K. Salisbury. “Towards a personal robotics development platform: Rationale and design of an intrinsically safe personal robot”. In: *International Conference on Robotics and Automation.* 2008.

[296] R. Schiavi, A. Bicchi, and F. Flacco. “Integration of active and passive compliance control for safe human-robot coexistence”. In: *IEEE International Conference on Robotics and Automation.* 2009.

[297] S. Pellegrinelli, H. Admoni, S. Javdani, and S. Srinivasa. “Human-robot shared workspace collaboration via hindsight optimization”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2016.

[298] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami. “Artificial cognition for social human–robot interaction: An implementation”. In: *Artificial Intelligence* (2017).

[299] F. Flacco, T. Kröger, A. De Luca, and O. Khatib. “A depth space approach to human-robot collision avoidance”. In: *IEEE International Conference on Robotics and Automation.* 2012.

[300] E. Mariotti, E. Magrini, and A. De Luca. “Admittance control for human-robot interaction using an industrial robot equipped with a F/T sensor”. In: *International Conference on Robotics and Automation.* 2019.

[301] A. De Luca, A. Albu-Schaffer, S. Haddadin, and G. Hirzinger. “Collision detection and safe reaction with the DLR-III lightweight manipulator arm”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2006.

[302] S. Haddadin, A. Albu-Schaffer, A. De Luca, and G. Hirzinger. “Collision detection and reaction: A contribution to safe physical human-robot interaction”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2008.

[303] J. Mainprice, E. A. Sisbot, L. Jaillet, J. Cortés, R. Alami, and T. Siméon. “Planning human-aware motions using a sampling-based costmap planner”. In: *2011 IEEE International Conference on Robotics and Automation.* 2011.

[304] J. S. Park, C. Park, and D. Manocha. “Intention-Aware Motion Planning Using Learning Based Human Motion Prediction.” In: *Robotics: Science and Systems.* 2017.

[305] S. Li and J. A. Shah. "Safe and Efficient High Dimensional Motion Planning in Space-Time with Time Parameterized Prediction". In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*. 2019.

[306] G. Hoffman. "Evaluating fluency in human–robot collaboration". In: *IEEE Transactions on Human-Machine Systems* 49.3 (2019), pp. 209–218.

[307] P. A. Lasota and J. A. Shah. "Analyzing the effects of human-aware motion planning on close-proximity human–robot collaboration". In: *Human factors* (2015).

[308] J. Heinzmann and A. Zelinsky. "Quantitative safety guarantees for physical human-robot interaction". In: *The International Journal of Robotics Research* (2003).

[309] J. Wittenburg. *Dynamics of multibody systems*. Springer-Verlag Berlin Heidelberg, 2008.

[310] I. D. Walker. "Impact configurations and measures for kinematically redundant and multiple armed robot systems". In: *IEEE transactions on robotics and automation* (1994).

[311] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause. "Learning-based model predictive control for safe exploration". In: *IEEE Conference on Decision and Control*. 2018.

[312] B. Vanderborght, A. Albu-Schäffer, A. Bicchi, E. Burdet, D. G. Caldwell, R. Carloni, M. Catalano, O. Eiberger, W. Friedl, G. Ganesh, et al. "Variable impedance actuators: A review". In: *Robotics and autonomous systems* (2013).

[313] A. Q. Keemink, H. van der Kooij, and A. H. Stienen. "Admittance control for physical human–robot interaction". In: *The International Journal of Robotics Research* (2018).

[314] H. Zhu and J. Alonso-Mora. "Chance-constrained collision avoidance for mavs in dynamic environments". In: *IEEE Robotics and Automation Letters* (2019).

[315] F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause. "Safe model-based reinforcement learning with stability guarantees". In: *Advances in Neural Information Processing Systems*. 2017.

[316] Z. Erickson, M. Collier, A. Kapusta, and C. C. Kemp. "Tracking human pose during robot-assisted dressing using single-axis capacitive proximity sensing". In: *Robotics and Automation Letters* (2018).

[317] A. Clegg, Z. Erickson, P. Grady, G. Turk, C. C. Kemp, and C. K. Liu. "Learning to collaborate from simulation for robot-assisted dressing". In: *Robotics and Automation Letters* (2020).

[318] R. P. Joshi, N. Koganti, and T. Shibata. "A framework for robotic clothing assistance by imitation learning". In: *Advanced Robotics* (2019).

[319] L. Hewing, J. Kabzan, and M. N. Zeilinger. "Cautious model predictive control using Gaussian process regression". In: *IEEE Transactions on Control Systems Technology* (2019).

[320] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, S. Feng, B. Burchfiel, and S. Song. "Adaptive Compliance Policy: Learning Approximate Compliance for Diffusion Guided Control". In: *arXiv preprint arXiv:2410.09309* (2024).

[321] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. "π_0: A Vision-Language-Action Flow Model for General Robot Control". In: *arXiv preprint arXiv:2410.24164* (2024).

[322] C. E. Myers, A. Interian, and A. A. Moustafa. "A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences". In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.1039172. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1039172.

[323] K. Zhang, Z. Yang, and T. Başar. "Multi-agent reinforcement learning: A selective overview of theories and algorithms". In: *Handbook of reinforcement learning and control* (2021), pp. 321–384.

[324] W. Chu, L. Li, L. Reyzin, and R. Schapire. "Contextual Bandits with Linear Payoff Functions". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 208–214. URL: https://proceedings.mlr.press/v15/chu11a.html.

[325] M. Shvartsman, B. Letham, E. Bakshy, and S. L. Keeley. "Response Time Improves Gaussian Process Models for Perception and Preferences". In: *The 40th Conference on Uncertainty in Artificial Intelligence*. 2024.

[326] S. Brown and A. Heathcote. "A ballistic model of choice response time." In: *Psychological review* 112.1 (2005), p. 117.

[327] I. Krajbich. "Accounting for attention in sequential sampling models of decision making". In: *Current Opinion in Psychology* 29 (2019). Attention & Perception, pp. 6–11. ISSN: 2352-250X. DOI: https://doi.org/10.1016/j.copsyc.2018.10.008. URL: https://www.sciencedirect.com/science/article/pii/S2352250X18301866.

[328] A. W. Thomas, F. Molter, I. Krajbich, H. R. Heekeren, and P. N. C. Mohr. "Gaze bias differences capture individual choice behaviour". In: *Nature Human Behaviour* 3.6 (2019), pp. 625–635. DOI: 10.1038/s41562-019-0584-8. URL: https://doi.org/10.1038/s41562-019-0584-8.

[329] M. L. Pedersen, M. J. Frank, and G. Biele. "The drift diffusion model as the choice rule in reinforcement learning". In: *Psychonomic Bulletin & Review* 24.4 (2017), pp. 1234–1251. DOI: 10.3758/s13423-016-1199-y. URL: https://doi.org/10.3758/s13423-016-1199-y.

[330] D. Fudenberg, P. Strack, and T. Strzalecki. "Speed, accuracy, and the optimal timing of choices". In: *American Economic Review* (2018). [PDF].

[331] R. Ratcliff, P. L. Smith, S. D. Brown, and G. McKoon. “Diffusion Decision Model: Current Issues and History”. In: *Trends in Cognitive Sciences* 20.4 (2016), pp. 260–281. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2016.01.007. URL: https://www.sciencedirect.com/science/article/pii/S1364661316000255.

[332] C. Baldassi, S. Cerreia-Vioglio, F. Maccheroni, M. Marinacci, and M. Pirazzini. “A Behavioral Characterization of the Drift Diffusion Model and Its Multialternative Extension for Choice Under Time Pressure”. In: *Management Science* 66.11 (2020), pp. 5075–5093. DOI: 10.1287/mnsc.2019.3475. eprint: https://doi.org/10.1287/mnsc.2019.3475. URL: https://doi.org/10.1287/mnsc.2019.3475.

[333] S. C. Castro, D. L. Strayer, D. Matzke, and A. Heathcote. “Cognitive workload measurement and modeling under divided attention.” In: *Journal of experimental psychology: human perception and performance* 45.6 (2019), p. 826.

[334] C. Zhang, C. Kemp, and N. Lipovetzky. “Goal Recognition with Timing Information”. In: *Proceedings of the International Conference on Automated Planning and Scheduling* 33.1 (July 2023), pp. 443–451. DOI: 10.1609/icaps.v33i1.27224. URL: https://ojs.aaai.org/index.php/ICAPS/article/view/27224.

[335] C. Zhang, C. Kemp, and N. Lipovetzky. “Human Goal Recognition as Bayesian Inference: Investigating the Impact of Actions, Timing, and Goal Solvability”. In: *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. AAMAS ’24. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2024, pp. 2066–2074. ISBN: 9798400704864.

[336] T. Strzalecki. *Stochastic Choice Theory*. Econometric Society Monographs. Cambridge University Press, 2025. URL: https://scholar.harvard.edu/sites/scholar.harvard.edu/files/tomasz/files/manuscript%5C%5F01.pdf.

[337] J. Drugowitsch. “Fast and accurate Monte Carlo sampling of first-passage times from Wiener diffusion models”. In: *Scientific reports* 6.1 (2016), p. 20490.

[338] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[339] D. R. Cox. *The theory of stochastic processes*. Routledge, 2017.

[340] A. Tirinzoni and R. Degenne. “On Elimination Strategies for Bandit Fixed-Confidence Identification”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 18586–18598. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2022/file/760564ebba4797d0dcf1678e96e8cbcb-Paper-Conference.pdf.