Conditions on Preference Relations that Guarantee the Existence of Optimal Policies

Jonathan Colaço Carr
McGill University
Mila - Quebec AI Institute

Prakash Panangaden
McGill University
Mila - Quebec AI Institute

Doina Precup
McGill University
Mila - Quebec AI Institute
Google DeepMind

Abstract

Learning from Preferential Feedback (LfPF) plays an essential role in training Large Language Models, as well as certain types of interactive learning agents. However, a substantial gap exists between the theory and application of LfPF algorithms. Current results guaranteeing the existence of optimal policies in LfPF problems assume that both the preferences and transition dynamics are determined by a Markov Decision Process. We introduce the Direct Preference Process, a new framework for analyzing LfPF problems in partially-observable, non-Markovian environments. Within this framework, we establish conditions that guarantee the existence of optimal policies by considering the ordinal structure of the preferences. We show that a decision-making process can have optimal policies, that are characterized by recursive optimality equations, even when no reward function can express the learning goal. Our findings narrow the gap between the empirical success and theoretical understanding of LfPF algorithms and provide future practitioners with the tools necessary for a more principled design of LfPF agents.

1 INTRODUCTION

Learning from Preferential Feedback (LfPF) is an important part of many real-world applications of artificial intelligence (AI). At a high level, it describes an interactive learning problem in which an agent's objectives are determined by a collection of relative preferences over outcomes. LfPF has been effectively employed in a diverse range of applications, from robotics tasks (Christiano et al., 2017; Lee et al., 2021) to the fine-tuning of Large Language Models (Bai et al., 2022; Lee et al., 2023; OpenAI, 2023; Rafailov et al., 2023; Stiennon et al., 2020).

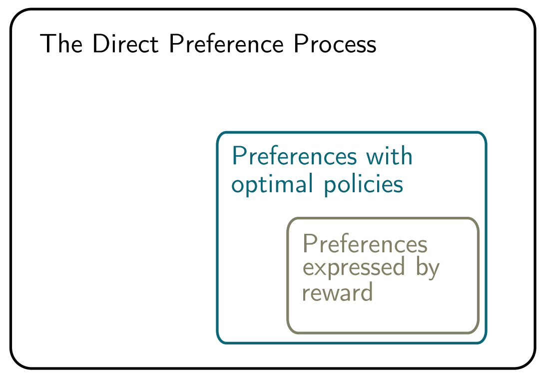

Figure 1: The Direct Preference Process (DPP) is a new framework for sequential decision-making from preferences. In Section 5 we present conditions guaranteeing the existence of optimal policies within a DPP which are strictly weaker than the conditions necessary for preferences to be expressed by expected reward. Therefore, it is possible for decision-making problems to have well-defined optimal policies even when no reward function can express the learning goal.

However, the current theory of LfPF lags far behind the success it has demonstrated in applications. There are no performance guarantees for LfPF problems beyond those defined through fully observable, Markovian environments. Moreover, current results (Chatterji et al., 2021; Kong and Yang, 2022; Saha et al., 2023; Xu et al., 2020; Zhu et al., 2023) assume that the preferences in a LfPF problem are generated by an underlying reward function. This assumption is known

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD. Copyright 2024 by the author(s).

to be both unrealistic (Bobu et al., 2020; Pandey et al., 2022; Tversky and Kahneman, 1974) and hard to verify (Casper et al., 2023). As a consequence, there are no theoretical guarantees for LfPF methods that are used in real-world scenarios.

In this paper, we define the Direct Preference Process, a model of preference-based learning in partially-observable, non-Markovian environments. A key feature of the Direct Preference Process is that abstracts away the details of *how* feedback is given to a learning agent, instead working "directly" with the ordinal structure inferred from the feedback. This abstraction is particularly well suited for LfPF problems, where a variety of feedback mechanisms are used during training, including offline reward modelling (Ziegler et al., 2020; Bai et al., 2022), once-per-episode trajectory feedback (Chatterji et al., 2021) and online feedback between trajectory segments (Christiano et al., 2017).

Our Contributions

- We define the Direct Preference Process, a model of preference-based learning in partially-observable, non-Markovian environments (Section 4). We provide necessary and sufficient conditions that determine when a Direct Preference Process can be cast as an instance of reinforcement learning (RL).
- We show that optimal policies exist in a Direct Preference Process even when the preferences cannot be expressed by reward, and generalize the Bellman Optimality Equations to a larger class of order relations (Section 5). In doing so, we highlight the properties of reward-based objectives that are *not necessary* in order for optimal policies to exist.
- We derive conditions that determine when a computationally-constrained agent is able to behave optimally (Section 6).

We focus on conditions that guarantee the existence of optimal policies in order to determine when a LfPF problem has well-defined solutions. Our work opens up interesting areas of future research, both for the theory and practice of preference-based learning.

2 RELATED WORK

In this section we review relevant sub-fields of LfPF.

Preference-based RL. Preference-based Reinforcement Learning (PbRL) (Abdelkareem et al., 2022; Wirth et al., 2017) describes a collection of RL techniques used to solve sequential decision-making problems whose objectives are determined by a set of relative preferences. This sub-field of LfPF includes RL from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020) and RL from AI Feedback (RLAIF) (Bai et al., 2022; Lee et al., 2023), both of which are popular methods of fine-tuning Large Language Models. To the best of our knowledge, the current results that guarantee the existence of optimal policies in PbRL (Chatterji et al., 2021; Kong and Yang, 2022; Saha et al., 2023; Xu et al., 2020; Zhu et al., 2023) rely on the assumption that there is an underlying controlled Markov process and reward function which describe the transition dynamics and preferences of the PbRL problem. We will relax both of these assumptions in this paper, and instead analyze preference-based learning problems in terms of the ordinal structure of the preferences.

Ordinal Dynamic Programs. Our work is reminiscent of ordinal dynamic programs (Mitten, 1974; Sobel, 1975; Carraway and Morin, 1988; Weng, 2011). Our model is most similar to Mitten's Preference Order Dynamic Program (Mitten, 1974), which searched for conditions on the ordinal structure of the objectives that could guarantee the existence of an optimal policy. While Mitten assumed access to a set of preferences between "intermediate policies" for each state, we assume that the objectives are given by a *single* set of preferences between distributions over trajectories, which seems like a more reasonable assumption given that feedback is typically collected over trajectories or trajectory segments.

Our analysis significantly extends that of Mitten and other prior ordinal dynamic programs. First, we abandon the Markov assumption and consider problems that occur in partially observable, non-Markovian environments. Second, we highlight two structural properties (convexity and interpolation) as well as two concrete examples (Examples 12 and 17) of goals that lead to the existence of optimal policies in the absence of expected reward. These properties and examples are essential to our theory, since they determine when it is impossible for ordinal decision problems to be reconsidered as instances of reinforcement learning. To the best of our knowledge, the connections between preference relations and expected reward in RL have only recently started to be considered (Bowling et al., 2023; Shakerinava and Ravanbakhsh, 2022; Pitis et al., 2022). Lastly, we provide additional conditions which determine when it is possible for a computationally constrained agent to behave optimally–an essential result for practical applications which has not been studied in prior work.

3 BACKGROUND

Given a finite set X, let $\mathrm{Dist}(X)$ be the set of probability distributions over X. For distributions A and B over X and non-negative number α less than or equal to one, the distribution $\alpha A+(1-\alpha)B$ assigns the probability of an element $x \in X$ as $\alpha A(x) + (1-\alpha)B(x)$.

We interpret a binary relation $\preceq$ on $\text{Dist}(X)$ as a set of relative preferences, so that for any two distributions A and B over X, the statement "$A \preceq B$" means that B is at least as desirable as A. Outcome B is "strictly preferred" to A, written $A \prec B$, if $A \preceq B$ and $\neg(B \preceq A)$. Outcomes A and B are "$\preceq$-equivalent", written $A \sim B$, if both $A \preceq B$ and $B \preceq A$.

Definition 1. A binary relation $\preceq$ on $\text{Dist}(X)$ is a **preorder** if it satisfies both of the following properties:

- *(reflexivity)* for any distribution A over X, $A \preceq A$.
- *(transitivity)* for any three distributions A,B and C over X, if $A \preceq B$ and $B \preceq C$ then $A \preceq C$.

A preorder is *total* if for any two distributions A, B over X, either $A \preceq B$ or $B \preceq A$.

3.1 Agents and Environments

To describe interactions between a learning agent and its environment, we consider a finite version the agent-environment interface (Abel et al., 2023). This framework is draws from related models of partially-observable, non-Markovian learning problems (Dong et al., 2022; Lu et al., 2023; Majeed, 2022; Hutter, 2016).

Definition 2. An **agent-environment interface** $(\mathcal{O}, \mathcal{A}, T)$ consists of a finite set of observations $\mathcal{O}$, a finite set of actions $\mathcal{A}$ and a time horizon $T \in \mathbb{N}$.

The finite horizon assumption is motivated by the fact that in practice, human labelers rank trajectories, so only a finite number of time steps is available to train. However, we impose no restrictions on the transition dynamics, so there may be an arbitrary (but finite) number of training "episodes". To avoid trivialities we assume that both the action and observation sets are non-empty. Extensions to infinite action and observation sets is left as an important area for future work.

For an agent-environment interface $(\mathcal{O}, \mathcal{A}, T)$, a set of *t-histories* is defined for each non-negative integer t less than or equal to T as follows: $\mathcal{H}_0 = \mathcal{O}$ and $\mathcal{H}_{t+1} = \mathcal{H}_t \times (\mathcal{A} \times \mathcal{O})$. We define $\mathcal{H}$ as the set of all histories,

$$\mathcal{H} = \bigcup_{t=0}^{T} \mathcal{H}_t. \tag{1}$$

We will refer to histories of length T as *trajectories* and we will denote their set as Ω instead of $\mathcal{H}_T$. For each non-negative integer t less than or equal to T, the projection $\xi_{0:t} : \Omega \to \mathcal{H}_t$ maps each trajectory to its first sub-history of length t. The environment determines which histories are attainable in a given learning problem.

Definition 3. An **environment** with respect to the interface $(\mathcal{O}, \mathcal{A}, T)$ is a tuple $e = (\rho_0, \rho)$ consisting of an initial distribution over observations $\rho_0 \in \text{Dist}(\mathcal{O})$ and a transition probability function $\rho : (\bigcup_{t=0}^{T-1} \mathcal{H}_t) \times \mathcal{A} \to \text{Dist}(\mathcal{O})$.

Notice that the transition dynamics in a learning environment may depend on the entire history, making it a realistic model for practical applications.

Example 4 (Generative Language Models 1)**.** An agent-environment interface $(\mathcal{O}, \mathcal{A}, T)$ can describe the interactions between a language model and a user, where the set of observations consists of the possible messages the user sends to the language model and the set of actions consists of the possible messages the model is able to send to the user. The environment e models the user's question patterns and prompts, which may depend on the full conversation history.

The behaviour of an agent is defined by its policy.

Definition 5. A **policy** π with respect to interface $(\mathcal{O}, \mathcal{A}, T)$ is a function $\pi : \mathcal{H} \to \text{Dist}(\mathcal{A})$.

It is important to allow policies to depend on the full history in Section 5 because we seek optimality conditions that do not depend on the agent's "state", or memory constraints. We address agents with memory constraints in Section 6, where the decision problem becomes partially-observable and non-Markovian. This is typical when function approximation is used.

Important Distributions. Agents will be evaluated according to the distributions their policies induce over Ω. For each policy π and history h_t, we define $D^\pi(h_t)$ as the distribution over Ω induced by starting from history h_t and following π in environment e thereafter. More precisely, for each trajectory h_T, $D^\pi(h_T)$ is equal to the Dirac distribution concentrated at h_T and for each history h_t of length less than T,

$$D^\pi(h_t) = \sum_{a\in\mathcal{A}} \pi(a|h_t) \sum_{o\in\mathcal{O}} \rho(o|h_t, a) D^\pi(h_t \cdot (a, o)), \tag{2}$$

where $h_t \cdot (a, o)$ is the history of length $t + 1$ obtained by appending the action-observation pair (a, o) to h_t. Similarly, we define $D^\pi(h_t \cdot a)$ as the distribution over Ω induced by starting from history h_t, selecting action a and following π thereafter. Note that we are overloading notation here; D^π may take either a history or a history appended with an action as its argument. The distributions $D^\pi(h_t)$, $D^\pi(h_t \cdot a)$ and $D^\pi(h_t \cdot (a, o))$ are related as follows:

$$D^\pi(h_t \cdot a) = \sum_{o\in\mathcal{O}} \rho(o|h_t, a) D^\pi(h_t \cdot (a, o)) \tag{3a}$$

$$D^\pi(h_t) = \sum_{a\in\mathcal{A}} \pi(a|h_t) D^\pi(h_t \cdot a). \tag{3b}$$

Attainable Histories. As in Abel et al. (2023), we will only consider the performance of policies in histories that occur can with non-zero probability in a given environment e under some policy. For each non-negative integer t less than or equal to T, the *set of attainable t-histories in e*, denoted by $\mathcal{H}_t^e$, is defined recursively as follows: $\mathcal{H}_0^e$ is equal to the support of ρ_0 and

$$\mathcal{H}_{t+1}^e = \{h_t \cdot (a, o) \in \mathcal{H}_{t+1} : \ h_t \in \mathcal{H}_t^e \text{ and } \rho(o|h_t, a) > 0\}. \quad (4)$$

We define $\mathcal{H}^e = \bigcup_{t=0}^{T} \mathcal{H}_t^e$ as the *set of attainable histories in e* and $\Omega^e = \mathcal{H}_T^e$ as the *set of attainable trajectories in e*.

4 THE DIRECT PREFERENCE PROCESS

An agent-environment interface, an environment and a binary relation on the set of distributions over trajectories define a Direct Preference Process.

Definition 6. A **Direct Preference Process** $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ consists of an agent-environment interface $(\mathcal{O}, \mathcal{A}, T)$, an environment e and a binary relation $\preceq$ on the set of distributions over Ω.

The distinctive feature, and fundamental premise of the Direct Preference Process is that the preference relation $\preceq$ defines the goals of a learning problem. Notably, we do not assume that these objectives have any quantifiable metric structure. However, when a numerical objective function does convey the goals of a decision problem, there is an implicit Direct Preference Process.

Example 7 (Generative Language Models 2)**.** Given the interface $(\mathcal{O}, \mathcal{A}, T)$ and environment e from Example 4, the goal of the language model may be to maximize a performance function $\varphi : \text{Dist}(\Omega) \to \mathbb{R}$. This induces a preference relation $\preceq_\varphi$ on $\text{Dist}(\Omega)$, defined for each pair of distributions A and B over Ω as:

$$A \preceq_\varphi B \iff \varphi(A) \leq \varphi(B).$$

The Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq_\varphi)$ underlies this decision problem.

A policy π is optimal in a Direct Preference Process if it achieves the most desirable outcome in every attainable start history.

Definition 8. Given a Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$, a policy π is **$\preceq$-optimal** (or simply **optimal**) if for every attainable history h_t and policy π', $D^{\pi'}(h_t) \preceq D^\pi(h_t)$.

In Example 7, the language model's policy is optimal for a given user if it achieves the best performance in every attainable conversation history.

4.1 Reward-Based Objectives

As noted in Section 2, the current analyses of LfPF problems assume that preferences are derived from an underlying reward function. While ordinal dynamic programs do not make this assumption outright, it is unclear whether or not an underlying reward is implied by the ordinal structure imposed. In contrast to both of these models, the Direct Preference Process comes with necessary and sufficient conditions that determine whether goals can be expressed by the expected cumulative sum of numerical rewards.

Definition 9. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process. We say that $\preceq$ is expressed by the **expected reward criterion** if there is a function $r : \mathcal{H} \to \mathbb{R}$ such that for any two distributions A and B over Ω,

$$A \preceq B \iff \mathbb{E}_A \left[\sum_{t=0}^{T} r(H_t) \right] \leq \mathbb{E}_B \left[\sum_{t=0}^{T} r(H_t) \right]. \quad (5)$$

We say that r **expresses** $\preceq$ if (5) holds for any two distributions A and B over Ω.

As an important sanity check, we now establish, in Theorem 10, that when goals are expressed by the expected reward criterion, the above definition of an optimal policy can be re-stated in terms the value function criterion found in the RL literature (Sutton and Barto, 2018; Puterman, 1994). For a reward function $r : \mathcal{H} \to \mathbb{R}$, we define the *r-value* of a policy π in history h_t as

$$V_\pi(h_t; r) = \mathbb{E}_\pi \left[\sum_{s=t}^{T} r(H_s) | H_t = h_t \right], \quad (6)$$

where the conditional expectation is taken with respect to $D^\pi(h_t)$.

Theorem 10. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process and suppose that a reward function $r : \mathcal{H} \to \mathbb{R}$ expresses $\preceq$. A policy π is $\preceq$-optimal if and only if for each attainable history h_t,

$$V_\pi(h_t; r) = \sup_{\pi'} V_{\pi'}(h_t; r).$$

As a consequence of Theorem 10, the standard RL problem can be seen as a Direct Preference Process, where the performance of a policy is only considered on histories that are attainable in an environment. A natural next question is what kinds of Direct Preference Processes can be cast as RL problems? Stated in Theorem 13, the von Neumann-Morgenstern (vNM) Expected Utility Theorem (von Neumann and Morgenstern, 1947) provides a decisive answer to this question. Their result depends on the following properties.

Definition 11. Let X be a finite set. A total preorder $\preceq$ on the set of distributions over X is said to satisfy:

i. **consistency** (or is **consistent**) if for every $\alpha \in (0,1)$ and any distributions A,B and C over X, $A \preceq B$ implies

$$\alpha A + (1-\alpha)C \preceq \alpha B + (1-\alpha)C.$$

ii. **convexity** (or is **convex**) if for every $\alpha \in (0,1)$ and any distributions A,B and C over X, $A \preceq B$ if and only if

$$\alpha A + (1-\alpha)C \preceq \alpha B + (1-\alpha)C.$$

iii. **interpolation** if for any distributions A,B and C over X, if $A \preceq B$ and $B \preceq C$ then there exists $\alpha \in [0,1]$ such that

$$\alpha A + (1-\alpha)C \sim B.$$

The following example clarifies the difference between consistency and convexity, drawing from a scenario with unacceptable risk (Jensen, 2012).

Example 12. Let E be a proper non-empty subset of Ω, interpreted as an event of "unacceptable risk". Given a real-valued function u on Ω and a real number β such that such that u is strictly greater than β on Ω, define the performance $\varphi : \mathrm{Dist}(\Omega) \to \mathbb{R}$ as:

$$\varphi(A) = \begin{cases} \sum_{\omega\in\Omega} u(\omega)A(\omega) & A(E) = 0 \\ \beta e^{A(E)} & A(E) > 0, \end{cases}$$

where $A(E)$ is the probability of event E under A. Assuming that u is non-constant on the complement of E, the relation $\preceq_\varphi$ on $\mathrm{Dist}(\Omega)$ defined by

$$A \preceq_\varphi B \iff \varphi(A) \leq \varphi(B),$$

is a total consistent preorder that is not convex. Moreover, $\preceq_\varphi$ does not satisfy interpolation.

Totality, transitivity, convexity and interpolation are the axioms of von Neumann and Morgenstern's seminal result. We state their result in Theorem 13 using our notation. In the general case Ω may be replaced with any finite set.

Theorem 13 (von Neumann-Morgenstern)**.** A binary relation $\preceq$ on the set of distributions over Ω is a total convex preorder satisfying interpolation if and only if there is a reward function $r : \mathcal{H} \to \mathbb{R}$ that expresses $\preceq$. Furthermore, the function $u_r : \Omega \to \mathbb{R}$ given by $u_r(\omega) = \sum_{t=0}^{T} r(\xi_{0:t}(\omega))$ is unique up to positive affine transformations.

Theorem 13 is a crucial feature of the Direct Preference Process. On one hand, it highlights the backdrop assumptions that are made in the PbRL literature (Chatterji et al., 2021; Kong and Yang, 2022; Saha et al., 2023; Xu et al., 2020; Zhu et al., 2023) which assume that preferences are derived from an underlying reward function. On the other hand, it will allow us to illustrate structural properties and concrete examples of decision problems that have optimal policies *in the absence of expected reward.*

5 CONDITIONS FOR OPTIMAL POLICIES

Without any assumptions on the goals of a Direct Preference Process, optimal policies may not exist, making it impossible to proceed with a meaningful theory of preference-based learning. Therefore, in this section we address the following question:

Q1: *Given a Direct Preference Process* $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$*, what conditions on* $\preceq$ *are sufficient to guarantee the existence of a* $\preceq$*-optimal policy?*

Our main result of this section, Theorem 15, concludes that (Q1) is satisfied whenever the restriction of $\preceq$ onto the set of distributions over attainable trajectories is a total, consistent preorder. One might have hoped that "rational" preferences, given by total preorders, would have been sufficient to guarantee that optimal policies exist. The next proposition shows that this is not the case.

Proposition 14. There is an agent-environment interface $(\mathcal{O}, \mathcal{A}, T)$, environment e and a total preorder $\preceq$ on the set of distributions over Ω such that the Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ has no optimal policy.

Proof. Let $\mathcal{O} = \{o^0, o^1\}$, $\mathcal{A} = \{a^0, a^1\}$, and $T = 2$. To keep notation light, define $h_1^0 = (o^0, a^0, o^0)$. Suppose that an environment e starts in o^0 and that for each history h_t, action a and observation o, $\rho(o|h_t, a) = 1/2$. For each trajectory ω, let $u(\omega; a^1)$ be the number of times that action a^1 occurs in ω. We define the performance $\varphi : \mathrm{Dist}(\Omega) \to \mathbb{R}$ as:

$$\varphi(A) = \begin{cases} \sum_{\omega\in\Omega} A(\omega)u(\omega; a^1) & \mathrm{supp}(A) \subseteq \mathrm{Cyl}(h_1^0) \\ -\sum_{\omega\in\Omega} A(\omega)u(\omega; a^1) & \text{else}, \end{cases} \tag{7}$$

where $\mathrm{Cyl}(h_1^0)$ is the subset of Ω consisting of every trajectory that begins with h_1^0. The performance φ measures the expected number of times that action a^1 occurs in a given distribution. It may be "bad" or "good" for a^1 to occur under a distribution A, depending on whether or not A is supported by $\mathrm{Cyl}(h_1^0)$. The performance induces a total preorder $\preceq_\varphi$ on $\mathrm{Dist}(\Omega)$

defined for any two distributions A and B over Ω as:

$$A \preceq_\varphi B \iff \varphi(A) \leq \varphi(B).$$

For contradiction, assume there is an optimal policy $\pi^\star$ for the Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq_\varphi)$. The distribution $D^{\pi^\star}((o^0))$ minimizes the expected number of times that a^1 occurs since for any policy π, $D^\pi((o^0))$ is not supported by $\mathrm{Cyl}(h_1^0)$. In particular, $\pi^\star$ must select action a^0 in history h_1^0. Let π' be a policy that selects action a^1 in history h_1^0. Then $\varphi(D^{\pi^\star}(h_1^0)) = 0$ and $\varphi(D^{\pi'}(h_1^0)) = 1$. So $D^{\pi^\star}(h_1^0) \prec_\varphi D^{\pi'}(h_1^0)$, contradicting the assumption that $\pi^\star$ is optimal. □

Notice that the relation $\preceq_\varphi$ defined in the proof of Proposition 14 does not satisfy consistency. To see this, consider $\omega_1 = (o^0, a^0, o^0, a^1, o^0)$, $\omega_2 = (o^0, a^0, o^0, a^0, o^0)$, $\omega_3 = (o^1, a^0, o^0, a^0, o^0)$. For each $i \in \{1, 2, 3\}$, let $\delta(\omega_i)$ be the Dirac distribution concentrated at ω_i. Then $\delta(\omega_1) \succ_\varphi \delta(\omega_2)$ but for any positive number α less than one,

$$\alpha\delta(\omega_1) + (1-\alpha)\delta(\omega_3) \prec_\varphi \alpha\delta(\omega_2) + (1-\alpha)\delta(\omega_3).$$

If, however, the goals of a Direct Preference Process also satisfy consistency, then we have the following result extending far beyond (Q1) that characterizes optimal policies with a series of recursive relations.

Theorem 15. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process. Whenever the restriction of $\preceq$ onto $\mathrm{Dist}(\Omega^e)$ is a total, consistent preorder:

i. There is a deterministic $\preceq$-optimal policy.

ii. If a policy π satisfies the following relation for each attainable history h_t of length less than T and action a,

$$D^\pi(h_t) \succeq D^\pi(h_t \cdot a), \tag{8}$$

then π is a $\preceq$-optimal policy.

Paired with vNM's Expected Utility Theorem, Theorem 15 has two profound implications. First, it is possible for agents to solve preference-based learning problems even when the objectives cannot be expressed by the expected reward criterion.

Corollary 16. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process. If the restriction of $\preceq$ onto $\mathrm{Dist}(\Omega^e)$ is a total consistent preorder that is either not convex or does not satisfy interpolation, then an optimal policy exists but $\preceq$ cannot be expressed by the expected reward criterion.

This is the case in Example 12 as well as our next example.

Example 17 (Tie-breaking Criterion)**.** Let u_1 and u_2 be two real-valued functions on Ω. For each $i \in \{1, 2\}$ and distribution A over Ω, let $u_i(A)$ denote the expected value of u_i under A. Define the relation $\preceq$ on $\mathrm{Dist}(\Omega)$ according to the following two rules:

R1: For any two distributions A and B over Ω, if $u_1(A) < u_1(B)$ then $A \prec B$.

R2: For any two distributions A and B over Ω, if $u_1(A) = u_1(B)$ then $(A \preceq B \iff u_2(A) \leq u_2(B))$.

Under these rules, u_2 acts as a "tie-breaking criterion" when two distributions achieve the same performance on u_1. Assuming that u_1 is non-constant and there are distributions A and B such that $u_1(A) = u_1(B)$ and $u_2(A) \neq u_2(B)$, the relation $\preceq$ defined by (R1) and (R2) is a total, convex preorder that does not satisfy interpolation.

A second implication of Theorem 15 is that the Bellman Optimality Equations (Bellman, 1957) that characterize optimal policies in RL are a consequence of a more general result that holds for total consistent preorders. We obtain Bellman's equations as a consequence of the second part of Theorem 15.

Corollary 18. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process. Whenever $\preceq$ is expressed by a reward function $r : \mathcal{H} \to \mathbb{R}$, a policy π is optimal if and only if it satisfies the following equation for each attainable history h_t of length less than T:

$$V_\pi(h_t; r) = \max_{a \in \mathcal{A}} \left(r(h_t) + \sum_{o \in \mathcal{O}} \rho(o | h_t, a) V_\pi(h_t \cdot (a, o); r) \right).$$

In light of Proposition 14 and Theorem 15, a minimal and robust assumption to further develop a theory of LfPF is the following.

Assumption 1. The restriction of $\preceq$ onto the set of distributions over attainable trajectories is a total consistent preorder.

5.1 Optimal Action Sets

A third consequence of Theorem 15 is that all optimal policies in a Direct Preference Process satisfying Assumption 1 are characterized by a set of "optimal actions" for each attainable history. This gives rise to a useful characterization of optimal policies which we will use in the next section.

Definition 19. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process. For each policy π and history h_t of length less than T, define $\mathcal{A}^\star_\pi(h_t)$ as the set of actions a for which $D^\pi(h_t \cdot a)$ is a least upper bound for the set $\{D^\pi(h_t \cdot a') : a' \in \mathcal{A}\}$. More precisely, $\mathcal{A}^\star_\pi(h_t)$ consists of every action a for which the following holds:

$$\forall a' \in \mathcal{A}, \quad D^\pi(h_t \cdot a) \succeq D^\pi(h_t \cdot a'). \tag{9}$$

Lemma 20. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process that satisfies Assumption 1. For any two optimal policies π and π' and attainable history h_t of length less than T, $\mathcal{A}^\star_\pi(h_t) =$ is equal to $\mathcal{A}^\star_{\pi'}(h_t)$.

In view of this lemma, we drop the dependence of $\mathcal{A}^\star_\pi(h_t)$ on π when π is an optimal policy. We call $\mathcal{A}^\star(h_t)$ the *optimal action set for h_t*.

Corollary 21. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process that satisfies Assumption 1. A policy π is optimal if and only if for each attainable history h_t of length less than T, $\pi(\cdot|h_t)$ is supported by $\mathcal{A}^\star(h_t)$.

6 OPTIMAL FEATURE-BASED POLICIES

In real-world applications, agents face computational constraints and make decisions based on a limited set of relevant information, known as "features", derived from their history. Hence, the concept of an optimal "feature-based" policy is crucial for a theory of preference-based learning. This is especially important for decision problems whose objectives lack a quantitative metric structure, since we cannot evaluate feature-based policies in terms of "near-optimal" behaviour.

The main question of this section (Q2) concerns a computationally-constrained agent which can only access a finite set of features, denoted as $\mathcal{X}$. A *feature map* $\phi : \mathcal{H} \to \mathcal{X}$ determines the feature retained from each history.

Example 22. Given a positive integer $k < T$, the feature of each history can be the sub-string of the most recent k observations and $k-1$ actions. In this case, $\mathcal{X} = \bigcup_{l=0}^{k-1} \mathcal{H}_l$ and the feature map ϕ is defined for each history $h_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$ as:

$$\phi(h_t) := \begin{cases} h_t & t < k \\ (o_{t-k+1}, a_{t-k+1}, \ldots, a_{t-1}, o_t) & t \geq k. \end{cases}$$

We define a feature-based policy as one whose action selection in each history h_t depends only on $\phi(h_t)$.

Definition 23. Given a feature map ϕ, π is a **feature-based policy** if for each pair of t-histories h_t, h'_t of length less than T,

$$(\phi(h_t) = \phi(h'_t)) \implies (\pi(\cdot|h_t) = \pi(\cdot|h'_t)). \tag{10}$$

We define Π^ϕ as the set of feature-based policies.

In Example 22, Π^ϕ is the set of policies whose action selection in each history depends only on the history through its final k observations and $k-1$ actions. The core objective of this section is to address (Q2):

Q2: *Given a Direct Preference Process that satisfies Assumption 1, what conditions does a feature map ϕ need to satisfy in order to guarantee that Π^ϕ contains an optimal policy?*

The optimal action sets described in Section 5.1 provide a necessary and sufficient condition to address (Q2).

Proposition 24. If a Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ satisfies Assumption 1 then for any feature map ϕ, Π^ϕ contains an optimal policy if and only if for each attainable history h_t of length less than T,

$$\bigcap_{h'_t \in \phi^{-1}(\phi(h_t)) \cap \mathcal{H}^e_t} \mathcal{A}^\star(h'_t) \neq \emptyset. \tag{11}$$

Roughly speaking, Proposition 24 shows that when the goals of a Direct Preference Process satisfy Assumption 1, an optimal feature-based policy exists if, and only if, for each attainable t-history h_t, there is an action that is simultaneously optimal for every attainable t-history in the preimage of $\phi(h_t)$.

6.1 Embedded Preferences

Although Proposition 24 gives both a necessary and sufficient condition that answers (Q2), the condition is rather generic and it is hard to check whether or not a system satisfies it. In this section we present Theorem 31, which provides verifiable conditions to answer (Q2). While not necessary, these conditions offer practical ways to ensure that optimal feature-based policies exist. They rely on the following notion of weighted averages.

Definition 25 ((ϕ, γ)-Frequency)**.** Let $\phi : \mathcal{H} \to \mathcal{X}$ be a feature map and $(\gamma_t)_{t=1}^{T-1}$ be a sequence of non-negative numbers that are not all zero. Given two non-negative integers t_1 and t_2 such that $t_1 \leq t_2 \leq T$ and for which $\sum_{t=t_1}^{t_2-1} \gamma_t$ is non-zero, define the function $f^{(\phi,\gamma)}_{t_1:t_2} : \mathcal{X} \times \mathcal{A} \times \mathrm{Dist}(\Omega) \to [0, 1]$ as

$$f^{(\phi,\gamma)}_{t_1:t_2}(x, a|D) := \frac{1}{\sum_{t=t_1}^{t_2-1} \gamma_t} \sum_{t=t_1}^{t_2-1} \gamma_t \mathbb{P}_D((X_t, A_t) = (x, a)), \tag{12}$$

where $\mathbb{P}_D((X_t, A_t) = (x, a))$ is the probability that the feature-action pair (x, a) is visited at time t under distribution D. We say that $f^{(\phi,\gamma)}_{t_1:t_2}(x, a|D)$ is the **(ϕ, γ)-frequency of (x, a) in distribution D in between t_1 and t_2.** When $\sum_{t=t_1}^{t_2-1} \gamma_t = 0$ we define $f^{(\phi,\gamma)}_{t_1:t_2}(x, a|D) = 0$. We abbreviate $f^{(\phi,\gamma)}_{0:T}(x, a|D)$ to $f^{(\phi,\gamma)}(x, a|D)$.

Interpretation of γ. The (ϕ, γ)-frequency is a weighted measure of how often each feature-action pair is visited in a given distribution over Ω. The weights $(\gamma_t)_{t=1}^{T-1}$ measure the importance of the time at which feature-action pairs are visited. For instance, if γ_t is equal to one for each time t, the distribution $f^{(\phi,\gamma)}(\cdot|D)$ measures the relative frequency of feature-action pairs visited under D. If $\gamma_t = \alpha^t$ for some positive number α less than one, $f^{(\phi,\gamma)}(\cdot|D)$ measures the α-discounted frequency of feature-action pairs visited under D.

Lemma 26. When $\sum_{t=t_1}^{t_2-1} \gamma_t$ is non-zero the function $(x, a) \mapsto f_{t_1:t_2}^{(\phi,\gamma)}(x, a|D)$ defines a probability distribution over the set of feature-action pairs, which we denote by $f_{t_1:t_2}^{(\phi,\gamma)}(\cdot|D)$.

Using the (ϕ, γ)-frequency map, we are now able to describe goals that "only depend" on the weighted frequency of feature action pairs.

Definition 27. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process and let $\preceq_\circ$ be a binary relation on the set of distributions over $\mathcal{X} \times \mathcal{A}$. We say that $\preceq$ **preserves and reflects** $\preceq_\circ$ **via** (ϕ, γ)**-frequency** if for any two distributions A and B over Ω,

$$A \preceq B \iff f^{(\phi,\gamma)}(\cdot|A) \preceq_\circ f^{(\phi,\gamma)}(\cdot|B). \tag{13}$$

We say that $\preceq$ **embeds into** $\preceq_\circ$ **via** (ϕ, γ)**-frequency** whenever $\preceq$ preserves and reflects $\preceq_\circ$ via (ϕ, γ)-frequency, despite the fact that the map $D \mapsto f^{(\phi,\gamma)}(\cdot|D)$ is neither injective nor surjective, and thus not an order embedding.

The next two examples show how the (ϕ, γ)-frequency embedding is useful when preferences are given between observation-action pairs (Stiennon et al., 2020) or trajectory segments (Christiano et al., 2017; Kim et al., 2022). In these situations, we can use the (ϕ, γ)-frequency map to define a preference relation on $\text{Dist}(\Omega)$ from the preference data.

Example 28 (Preferences over Observation-Action Pairs)**.** Consider $v_1, v_2 : \mathcal{O} \times \mathcal{A} \to [0, 1]$ and define the relation $\preceq_\circ$ on $\text{Dist}(\mathcal{O} \times \mathcal{A})$ according to the Tie-breaking Criterion from Example 16. If ϕ maps each history to its most recent observation, then $\mathcal{X} = \mathcal{O}$ and $\preceq_\circ$ is an ordering on $\text{Dist}(\mathcal{X} \times \mathcal{A})$. For a sequence of positive weights $(\gamma_t)_{t=0}^{T-1}$, we can define a relation $\preceq$ on $\text{Dist}(\Omega)$ via Equation 13. In this case, the preferences given by $\preceq$ depend only on the weighted frequency of observation-action pairs. In particular, distributions A and B over Ω are $\preceq$-equivalent if they visit all observation action pairs with the same weighted frequency.

Example 29 (Preferences over Trajectory Segments)**.** Let k be a positive integer less than T. Suppose that preference data is available for histories of length up to k, giving rise to a binary relation $\preceq_\circ$ on the set of distributions over $\bigcup_{l=0}^{k}(\mathcal{H}_l \times \mathcal{A})$. With the feature map defined in Example 22 and a sequence of non-negative numbers $\gamma = (\gamma_t)_{t=0}^{T-1}$, the goals of a Direct Preference Process can be defined using $\preceq_\circ$ and Equation 13.

Although Definition 27 ensures that the learning objectives are fully described by distributions over feature-action pairs, we require additional assumptions on the transition dynamics of the environment to ensure that optimal feature-based policies exist.

Definition 30. A Direct Preference Process $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ and feature map ϕ satisfy the **Markov Feature Assumption** when the following two statements hold for each pair of attainable t-histories h_t, h'_t of length less than T:

- If $\phi(h_t) = \phi(h'_t)$ then for each action a, $\rho(\cdot|h_t, a) = \rho(\cdot|h'_t, a)$.
- If $\phi(h_t) = \phi(h'_t)$ then for each action a and observation o, $\phi(h_t \cdot (a, o)) = \phi(h'_t \cdot (a, o))$.

The latter conveys the notion that if a feature accounts for all the retained information in each history, then the feature in each history depends on its sub-histories only through previous features. This is satisfied by the agents considered in many popular agent designs (Lu et al., 2023; Mnih et al., 2015; Osband et al., 2016). Combined with Definition 27, the Markov Feature Assumption guarantees that optimal feature-based policies exist.

Theorem 31. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be a Direct Preference Process and $\preceq_\circ$ a total consistent preorder on $\text{Dist}(\mathcal{X} \times \mathcal{A})$ such that $\preceq$ embeds into $\preceq_\circ$ via (ϕ, γ)-frequency.

i. Every policy π that satisfies the following recursive relation for each attainable history h_t of length less than T and action a is a $\preceq$-optimal policy:

$$f_{t:T}^{(\phi,\gamma)}(\cdot|D^\pi(h_t)) \succeq_\circ f_{t:T}^{(\phi,\gamma)}(\cdot|D^\pi(h_t \cdot a)). \tag{14}$$

ii. If the Markov Feature Assumption is satisfied then Π^ϕ contains a $\preceq$-optimal policy.

The first part of Theorem 31 shows that if the goals of a Direct Preference Process are preserved and reflected by a total consistent preorder on the set of distributions over feature-action pairs, then a policy is optimal whenever it achieves the most desirable distribution over future feature-action pairs in every starting history. However, without any assumptions on the environment's transition dynamics, a feature-based policy may not satisfy the conditions in part (i). The Markov

Feature Assumption is a strong assumption and relaxing these conditions is an important area for future work. When the Markov Feature Assumption does not hold, Proposition 24 provides an alternative means to guarantee the existence of feature-based policies.

6.2 Connection to Markov Rewards

We introduced the embedding of preferences via (ϕ, γ)-frequency as an abstract property that might underpin the goals of a Direct Preference Process. Example 22 demonstrates the practical utility of this property when preference data is collected on trajectory segments rather than full-length trajectories. Nevertheless, some readers may be hesitant about its justification. To address these doubts, our final result shows that the (ϕ, γ)-frequency embedding is implied by any objective defined by Markov rewards.

Theorem 32. Let $(\mathcal{O}, \mathcal{A}, T, e, \preceq)$ be Direct Preference Process, $\phi : \mathcal{H} \to \mathcal{X}$ be a feature map and $(\gamma_t)_{t=1}^{T-1}$ be a sequence of non-negative numbers that are not all zero. The following two statements are equivalent:

1. $\preceq$ embeds into a total convex preorder $\preceq_\circ$ satisfying interpolation via (ϕ, γ)-frequency.
2. There is a reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that for any two distributions D and D' over Ω, $D \preceq D'$ if and only if

$$\mathbb{E}_D\left[\sum_{t=1}^{T-1} \gamma_t r(X_t, A_t)\right] \leq \mathbb{E}_{D'}\left[\sum_{t=1}^{T-1} \gamma_t r(X_t, A_t)\right].$$

It is interesting to compare this result to vNM's Expected Utility Theorem. If the goals are expressed by a feature-action reward function, as opposed to a history-based reward, then $\preceq$ embeds into an underlying feature-action preference via (ϕ, γ)-frequency. However, the feature-action reward function which expressed $\preceq$ may not be unique, as multiple feature-action preferences could preserve and reflect $\preceq$. This result complements previous work on Markov reward expressiveness in finite MDPs (Abel et al., 2021; Skalse and Abate, 2023).

7 CONCLUSION

We introduced the Direct Preference Process, a model of preference-based learning in partially-observable, non-Markovian environments. Unlike previous work, we did not assume that preferences were generated by an underlying reward function. Instead we used conditions on the ordinal structure of the preferences to guarantee the existence of optimal policies. We showed that it is possible for an agent to behave optimally with respect to a given set of preferences even when there is no corresponding reward function that captures the same learning goal. Lastly, we provided two results to determine when it is possible for a computationally-constrained agent to behave optimally, as well as a characterization of goals expressed by Markov rewards.

The Direct Preference Process opens up many interesting avenues for future work. An extension of this framework for infinite observation and action sets is important for preference-based robotics tasks. For practitioners, it is interesting to study whether agents can perform well without learning a reward model. Recent findings (Rafailov et al., 2023; An et al., 2023; Kang et al., 2023) have shown that this may be the case. Moreover, the notion of a (ϕ, γ)-frequency embedding could be used to *derive* relevant features for a preference-based decision problem. Finally, it would be very useful to study the hardness of learning feature maps. We suspect that this may highlight differences between reward-based vs purely preference-based agents.

Acknowledgements

We thank the reviewers for their helpful feedback. JCC is grateful for the support of the NSERC CGS-M scholarship. PP is grateful for the support of an NSERC research grant. DP acknowledgements...

References

Abdelkareem, Y., Shehata, S., and Karray, F. (2022). Advances in preference-based reinforcement learning: A review. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2527–2532.

Abel, D., Barreto, A., Van Roy, B., Precup, D., van Hasselt, H., and Singh, S. (2023). A definition of continual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems.*

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., and Singh, S. (2021). On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812.

An, G., Lee, J., Zuo, X., Kosaka, N., Kim, K.-M., and Song, H. O. (2023). Direct preference-based policy optimization without reward modeling. In *Thirty-seventh Conference on Neural Information Processing Systems.*

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition.

Bobu, A., Bajcsy, A., Fisac, J. F., Deglurkar, S., and Dragan, A. D. (2020). Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, 36(3):835–854.

Bowling, M., Martin, J. D., Abel, D., and Dabney, W. (2023). Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR.

Carraway, R. and Morin, T. (1988). Theory and applications of generalized dynamic programming: An overview. *Computers & Mathematics with Applications*, 16(10):779–788.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback.

Chatterji, N., Pacchiano, A., Bartlett, P., and Jordan, M. (2021). On the theory of reinforcement learning with once-per-episode feedback. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3401–3412. Curran Associates, Inc.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dong, S., Roy, B. V., and Zhou, Z. (2022). Simple agent, complex environment: Efficient reinforcement learning with agent states. *J. Mach. Learn. Res.*, 23:255:1–255:54.

Hutter, M. (2016). Extreme state aggregation beyond markov decision processes. *Theoretical Computer Science*, 650:73–91. Algorithmic Learning Theory.

Jensen, K. K. (2012). Unacceptable risks and the continuity axiom. *Economics & Philosophy*, 28(1):31–42.

Kang, Y., Shi, D., Liu, J., He, L., and Wang, D. (2023). Beyond reward: Offline preference-guided policy optimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15753–15768. PMLR.

Kim, C., Park, J., Shin, J., Lee, H., Abbeel, P., and Lee, K. (2022). Preference transformer: Modeling human preferences using transformers for rl. In *The Eleventh International Conference on Learning Representations*.

Kong, D. and Yang, L. (2022). Provably feedback-efficient reinforcement learning via active reward learning. *Advances in Neural Information Processing Systems*, 35:11063–11078.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Lee, K., Smith, L. M., and Abbeel, P. (2021). Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*.

Lu, X., Roy, B. V., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. (2023). Reinforcement learning, bit by bit. *Foundations and Trends® in Machine Learning*, 16(6):733–865.

Majeed, S. (2022). *Abstractions of General Reinforcement Learning*. PhD thesis.

Mitten, L. G. (1974). Preference order dynamic programming. *Management Science*, 21(1):43–46.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

OpenAI (2023). Gpt-4 technical report.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped dqn. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Pandey, R., Purohit, H., Castillo, C., and Shalin, V. L. (2022). Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

Pitis, S., Bailey, D., and Ba, J. (2022). Rational multi-objective agents must admit non-markov reward representations. In *NeurIPS ML Safety Workshop*.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Saha, A., Pacchiano, A., and Lee, J. (2023). Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 6263–6289. PMLR.

Shakerinava, M. and Ravanbakhsh, S. (2022). Utility theory for sequential decision making. In *International Conference on Machine Learning*, pages 19616–19625. PMLR.

Skalse, J. and Abate, A. (2023). On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1974–1984. PMLR.

Sobel, M. J. (1975). Ordinal dynamic programming. *Management Science*, 21(9):967–975.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Weng, P. (2011). Markov decision processes with ordinal rewards: Reference point-based preferences. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 21, pages 282–289.

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. (2017). A survey of preference-based reinforcement learning methods. *J. Mach. Learn. Res.*, 18(1):4945–4990.

Xu, Y., Wang, R., Yang, L. F., Singh, A., and Dubrawski, A. (2020). Preference-based reinforcement learning with finite-time guarantees. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Zhu, B., Jordan, M., and Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2020). Fine-tuning language models from human preferences.