RL without Rewards: A Social Choice Approach to Intermediate Preferences in Context-Ordinal MDPs

Anonymous authors
Paper under double-blind review

Keywords: Preference-based RL, social choice, dynamic programming, Markov decision processes, game theory

Summary

Preference-based RL (PbRL) traditionally relies on sparse, trajectory-level feedback, which bottlenecks credit assignment and restricts temporal-difference (TD) bootstrapping to propagating delayed terminal signals. We propose the Context-Ordinal MDP, a framework that natively incorporates dense, intermediate ordinal preferences without collapsing them into scalar proxy models. Drawing from social choice theory, we formulate a dynamic programming approach that aggregates preferences over actions across time using temporal discounting.

We prove this ordinal Q-learning framework is mathematically well-posed, converges under regularized voting rules, and strictly generalizes classical cardinal RL. However, we also expose a fundamental limitation: given non-transitive preference cycles, dynamic programming acts as a dictatorial voting agenda. This "Sequential Tournament Paradox" forces the sample-efficient local solution to diverge from the unexploitable global solution, potentially falling into the "2nd-Worst" trap. Ultimately, we establish a strict theoretical trade-off in PbRL between the computational efficiency of local dynamic programming and the guaranteed unexploitability of global trajectory evaluation.

Contribution(s)

1. Formulating a preference-based RL framework that aggregates intermediate ordinal preferences across time via discounted voting.
   Context: Extends the Context-Ordinal Games framework (Gemp et al., 2026) from single-state settings to sequential Markov Decision Processes.
2. Proving ordinal Q-learning converges to a guaranteed fixed point and acts as a strict contraction under KL-regularized voting rules.
   Context: Unlike the non-expansive $\max$ operator in traditional RL, discontinuous ordinal voting rules fail to induce contractions, necessitating regularization.
3. Demonstrating this framework strictly generalizes classical RL, recovering standard Bellman equations when applying Score Voting.
   Context: This validates temporal discounting as a mathematically sound weighting mechanism for sequential voter populations.
4. Identifying the Sequential Tournament Paradox, proving local ordinal dynamic programming structurally diverges from global optimality given preference cycles.
   Context: Establishes that Bellman's Principle of Optimality breaks down in non-transitive spaces, forcing a trade-off between tractability and unexploitability.
5. Showing that state-action voting statistics eliminate the need for synchronous counterfactuals, enabling dense, intermediate off-policy TD bootstrapping.
   Context: Standard preference-based RL relies on sparse trajectory-level feedback, which sacrifices dense credit assignment and sample efficiency.

RL without Rewards: A Social Choice Approach to Intermediate Preferences in Context-Ordinal MDPs

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning traditionally relies on cardinal reward functions to optimize behavior. However, human preferences—often the true objective in alignment tasks—are frequently elicited as ordinal comparisons. In this work, we present a preference-based reinforcement learning framework for Markov Decision Processes (MDPs) that replaces cardinal rewards entirely with dense, intermediate ordinal feedback. Drawing on recent approaches that apply social choice to preference-based feedback, we formulate a dynamic programming approach that aggregates preferences over actions across time using temporal discounting. We prove that this ordinal Q-learning framework is mathematically well-posed, converges to a guaranteed fixed point under regularized voting rules, and strictly generalizes classical cardinal RL. However, we expose a fundamental limitation in non-transitive environments: the sequential nature of dynamic programming acts as a dictatorial voting agenda. This "Sequential Tournament Paradox" forces local, time-consistent solutions to structurally diverge from global optimality, establishing a formal theoretical trade-off between the sample efficiency of local bootstrapping and the guaranteed unexploitability of global trajectory evaluation.

1 Introduction

Preference-based reinforcement learning (PbRL) has emerged as a cornerstone for aligning agents with human intent (Wirth et al., 2017). Beyond alignment (Christiano et al., 2017), framing objectives through ordinal preferences offers a powerful paradigm for general RL: it provides inherent robustness against arbitrary reward shaping, miscalibrated sensors, and hyperparameter sensitivity by remaining strictly scale-invariant. However, a stark structural gap remains between traditional cardinal RL and modern PbRL. In classical RL, environments routinely provide dense, intermediate rewards at every time step (Ng et al., 1999), allowing agents to efficiently evaluate local transitions and utilize temporal-difference (TD) bootstrapping. By contrast, pure PbRL frameworks typically elicit ordinal preferences exclusively over full, completed trajectories or terminal states (Munos et al., 2024; Shani et al., 2024).

This restriction is conceptually artificial. If an environment or human designer is capable of specifying a cardinal reward function that evaluates intermediate states, they are certainly capable of expressing ordinal preferences over those same intermediate states. The barrier to utilizing intermediate preferences is not elicitation, but mathematics: unlike scalar rewards, step-by-step ordinal rankings (e.g., $A \succ B$ at $t = 1$, and $B \succ A$ at $t = 2$) cannot be algebraically summed across time. To utilize intermediate feedback, standard PbRL pipelines are forced to compress these rankings into scalar proxy rewards (e.g., via Bradley-Terry models), abandoning the pure ordinal space. This scalarization destroys the inherent scale-invariance of ordinal metrics, reintroducing well-documented vulnerabilities to magnitude-based reward hacking and brittle multi-objective tuning that plague traditional RL.

In this work, we ask: *Can we natively integrate intermediate preference feedback into the transition dynamics of an MDP without collapsing it into scalar rewards?*

We propose the Context-Ordinal MDP (CO-MDP), a framework that treats sequential evaluation as a temporal election. Drawing on approaches to preference-based feedback problems that use social choice (Gemp et al., 2026), we show that by strict time-synchronization and temporal discounting, ordinal statistics can be linearly accumulated. This organically restores Bellman's equations, unlocking off-policy dynamic programming over purely ordinal distributions.

By mapping the theoretical boundaries of this framework, we uncover a rich set of guarantees and fundamental limitations. On the positive side, we prove that local ordinal Q-learning acts as a strict contraction mapping under KL-regularization, recovers the global optimum if a Condorcet winner exists, and is mathematically immune to selecting a Condorcet loser. On the negative side, we expose the *Sequential Tournament Paradox*: in the presence of non-transitive preference cycles, the sequential nature of dynamic programming acts as a dictatorial voting agenda. This forces the local, time-consistent solution to structurally diverge from the unexploitable global solution. Ultimately, we establish that while native intermediate feedback enables ordinal dynamic programming, practitioners must carefully navigate the fundamental trade-off between local time-consistency and global unexploitability.

2 Preliminaries

To establish our preference-based framework, we first briefly review standard cardinal reinforcement learning and the mathematics of preference aggregation.

2.1 Markov Decision Processes

A Markov Decision Process (MDP) (Puterman, 2014) is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ defines the transition dynamics, $r(s) \in \mathbb{R}$ is a cardinal reward function (we assume state-dependent rewards for simplicity of early exposition, explicitly deferring the fully general state-action formulation to Section 5.2), and $\gamma \in [0, 1)$ is the discount factor.

The agent acts according to a policy $\pi(a \mid s)$. The traditional objective is to find a policy that maximizes the expected discounted reward. The state-action value function, $Q^\pi(s, a)$, represents the expected reward of taking action a in state s and following π thereafter, satisfying the Bellman expectation equation:

$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[V^\pi(s')] \tag{1}$$

where the state value is $V^\pi(s) = \mathbb{E}_{a \sim \pi}[Q^\pi(s, a)]$. The optimal policy acts greedily with respect to the optimal values, satisfying the Bellman optimality equation: $\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$.

2.2 Social Choice Theory

Social choice theory studies the aggregation of individual preferences into a collective decision. Let $\mathcal{C}$ be a finite set of candidates (in our setting, discrete actions) and $\mathcal{V}$ be a population of voters.

Voters express preferences via ranked ballots (e.g., $c_1 \succ c_2 \succ c_3$). The aggregated set of these ballots forms a *preference profile*, which we define formally as a probability distribution over all possible ranked ballots. A *voting rule* (or *social choice function*) ν maps a preference profile to a winning set of candidates or a probability distribution (lottery) over them.[1] Philosophically, a social choice function can be interpreted as a decision rule for determining a "reasonable" action given potentially divergent ranked preferences over all available actions.

[1] Typically preference profiles are defined as a finite tuple of rankings, but nearly all natural voting rules can be extended to apply to distributions.

For example, in a 2-candidate election, *Majority Rule* selects the candidate preferred by more than 50% of the voters; it is arguably the only "reasonable" rule in this setting (May, 1952). For elections with three or more candidates, impossibility theorems preclude a single canonical rule (Arrow, 1950). Instead, numerous sophisticated voting rules have been proposed, each exhibiting different axiomatic properties (Brandt et al., 2016).

Maximal Lotteries. Because of its central role in this work, we specifically highlight the *Maximal Lotteries* voting rule (Fishburn, 1984). For any two candidates $c, c' \in \mathcal{C}$, let $P(c \succ c')$ denote the fraction of the electorate that ranks c above c'. The pairwise margin $M(c, c') = P(c \succ c') - P(c' \succ c)$ defines a skew-symmetric matrix $M \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$. Maximal Lotteries treats M as a symmetric zero-sum game, identifying the winning lottery as the Nash Equilibrium mixed strategy $p^* \in \Delta(\mathcal{C})$ satisfying $(p^*)^\top M q \geq 0$ for all $q \in \Delta(\mathcal{C})$. Crucially, unlike deterministic voting rules, Maximal Lotteries natively resolves non-transitive Condorcet cycles (e.g., $A \succ B \succ C \succ A$) by returning a mixed stochastic strategy (e.g., $p^* = [1/3, 1/3, 1/3]$), mathematically guaranteeing an unexploitable aggregate outcome.

3 Context-Ordinal MDPs

We now introduce our base framework, extending the principles of Context-Ordinal Games (Gemp et al., 2026) to sequential decision-making.

3.1 Motivating Intuition: The Marshmallow Election

Consider the famous Stanford marshmallow experiment (Mischel et al., 1972; Shoda et al., 1990). A child at state s_0 faces two actions: a_0 (eat the single marshmallow now) or a_0' (wait 15 minutes to receive two marshmallows).

Imagine two parallel timelines unrolling from this decision. At time $t = 1$ (one minute later), the "timeline a_0" self is happily chewing a marshmallow, while the "timeline a_0'" self is staring at the table, impatient and salivating. If we ask this $t = 1$ self to rank the two outcomes, they submit the vote: $a_0 \succ a_0'$.

However, at time $t = 15$, the "timeline a_0" self has nothing left, whereas the "timeline a_0'" self is happily eating two marshmallows. The $t = 15$ self submits the opposite vote: $a_0' \succ a_0$.

In typical RL, the optimal action is the one that maximizes expected discounted reward. With strictly ordinal data, this decision rule cannot be implemented. Instead, we treat the decision as an election over actions. Every future time step t constitutes a vote that ranks the original actions, and we downweight the voting power of the t-th time step by γ^t. The agent's optimal action is then selected by applying a voting rule to this temporally discounted preference profile.

3.2 Formal Problem Definition

We formalize this intuition by defining a **Context-Ordinal MDP (CO-MDP)** as a tuple $\langle \mathcal{S}, \mathcal{A}, P, \rho, \gamma \rangle$. We denote tuples of states by bold letters $\mathbf{s}$.

The defining differences from a traditional MDP lie in the replacement of scalar rewards with ordinal preferences, the explicit definition of evaluation contexts, and the corresponding redefinition of the agent's objective.

1. **No Numeric Rewards:** The cardinal reward function is entirely removed. Following our state-based formulation, we replace it with a state-based preference function $\rho(\mathbf{s})$, which takes a tuple of states and returns a strict or weak ordinal ranking over them. (We extend this to fully general state-action preferences $\rho(\mathbf{x})$ where $\mathbf{x} = \big((s_1, a_1), \ldots, (s_{|\mathcal{A}|}, a_{|\mathcal{A}|})\big)$ in Section 5.2).
2. **Temporal Contexts:** To evaluate actions, we must define the context in which preferences can be expressed. Because the environment's transition dynamics P and the agent's policy π may be

stochastic, timelines unfolding from the initial actions are probabilistic. A *context* $\mathbf{s}_t$ at time t is thus a specific realization (a sampled tuple of states) resulting from taking each available action in the root state and following π thereafter.[2]

Applying the preference function to this context, $\rho(\mathbf{s}_t)$, yields a ranking over the states. Crucially, this state ranking implicitly expresses a preference ranking over the initial actions $a \in \mathcal{A}$ that generated those respective timelines.

To build the overall preference profile over actions, we have each context submit a vote. We weight each context by two factors: probability and time. First, to account for environment stochasticity, a context's weight is proportional to the probability of it actually occurring. Second, to implement temporal discounting ($\gamma < 1$), we reduce the voting power of populations further in the future. Thus, the final weight of a ballot generated at time t is proportional to its occurrence probability multiplied by γ^t (see Figure 1).

3.3 Ordinal Dynamic Programming

In traditional reinforcement learning, value functions evaluate individual states in isolation because scalar rewards possess independent, absolute magnitude. In contrast, ordinal preferences are fundamentally comparative—they possess no meaning outside the context of a specific set of alternatives. Consequently, to generalize Bellman's equations from scalar to preference feedback, our value functions must be upgraded from single-state evaluators to tuple-operators that simultaneously evaluate distinct states across parallel trajectories.

Let $\mathcal{R}$ denote the set of all possible ordinal rankings over the root action space $\mathcal{A}$. The space of all possible preference profiles is the probability simplex over $\mathcal{R}$, denoted $\Delta(\mathcal{R})$. A voting rule ν maps this simplex to a distribution over actions, $\nu : \Delta(\mathcal{R}) \Rightarrow \Delta(\mathcal{A})$. We let $v(\mathbf{s})$ denote the one-hot "vote" vector corresponding to the deterministic ranking $\rho(\mathbf{s})$, e.g., a profile with 100% of the voting mass on $\rho(\mathbf{s})$.

To ensure the aggregated votes remain a valid probability distribution, we normalize the temporal weights by $(1-\gamma)$. Because $\sum_{t=0}^{\infty}(1-\gamma)\gamma^t = 1$, this forms a strict convex combination, allowing the ordinal state-value $V^\pi(\mathbf{s})$ to track the expected ballot distribution across time:

$$V^\pi(\mathbf{s}) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}(1-\gamma)\gamma^t v(\mathbf{s}_t) \;\middle|\; \mathbf{s}_0 = \mathbf{s}\right] = (1-\gamma)v(\mathbf{s}) + \gamma\mathbb{E}_{\mathbf{s}'\sim P(\cdot|\mathbf{s},\pi(\mathbf{s}))}[V^\pi(\mathbf{s}')]$$

Here, the transition notation $P(\cdot \mid \mathbf{s}, \pi(\mathbf{s}))$ compactly represents the synchronous, parallel unrolling of independent trajectories. Because the context $\mathbf{s} = (s^{(1)}, s^{(2)}, \dots)$ is a tuple of states, this operator specifies sampling an independent action for each individual state according to the agent's policy π, and then transitioning each state independently according to the environment dynamics P to form the subsequent context tuple $\mathbf{s}'$.

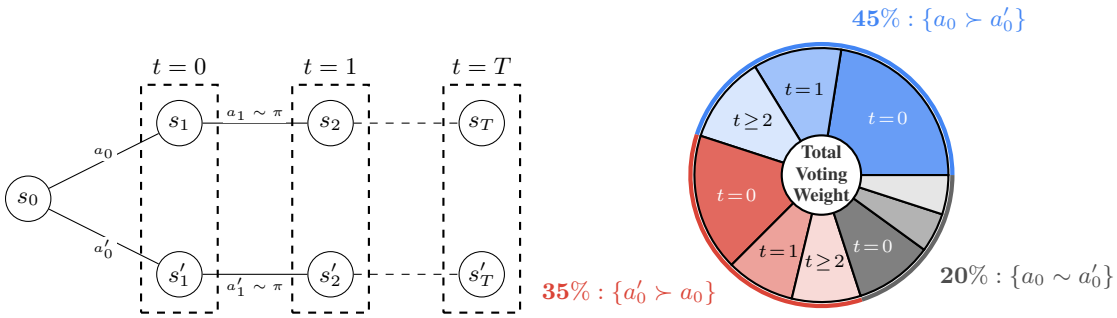
A Q-like function can be recovered for a single root state s by evaluating the expected preference profile for the immediate context of next states resulting from all available actions $\mathcal{A}$:

$$Q^\pi(s) = \mathbb{E}_{\mathbf{s}'\sim P(\cdot|s,\mathcal{A})}[V^\pi(\mathbf{s}')] \tag{2}$$

The optimal policy is then computed by applying a voting rule ν to this final aggregated distribution over rankings: $\pi^*(s) \in \nu(Q^\pi(s))$. This forms the foundational mathematical model of Context-Ordinal RL.

The Objective of a CO-MDP. In traditional RL, the explicit objective is to find a policy π^* that maximizes the expected discounted return from every state. Crucially, standard MDP theory establishes that achieving this maximum is mathematically equivalent to finding the unique fixed point

[2] While contexts can be defined abstractly to accommodate highly general evaluation settings (e.g., cross-temporal comparisons that intrinsically model human time-preference; see Appendix D), we focus here on synchronous temporal contexts.



(a) Parallel Rollouts and Context Generation

(b) Aggregated Temporal Votes

Figure 1: **Preference-Based RL as an Ordinal Election.** **(a)** To evaluate the initial actions a_0 and a'_0 at state s_0, we perform parallel rollouts using the agent's policy π. At each time step t, the resulting states (dashed boxes) are compared using a preference relation ρ to generate a ranked ballot. **(b)** Across all parallel rollouts, the resulting ballots form a temporal voting population. The total voting pool is categorized by ballot type, where each category is subdivided by the time step t of the observation. Crucially, the voting power (slice size and opacity) of an observation at time t is weighted by its discounted state visitation frequency ($\propto \gamma^t$). For $\gamma = 0.5$, $t = 0$ contributes exactly 50% of the total voting weight. A voting rule (e.g., Majority Rule) aggregates this distribution to determine the optimal policy at the root.

of the Bellman optimality equation. In a CO-MDP, because there is no scalar return to maximize, we cannot define optimality through a global maximization objective. Instead, we must rely entirely on this self-consistent fixed-point perspective: the objective is to find a policy whose actions are consistently "elected" by the temporally discounted voter population it generates.

Let $Q^\pi(s)$ denote the preference profile over possible actions $\mathcal{A}$, accumulated by starting from the root state s following policy π. Let ν be the chosen voting rule. An optimal policy π^* is defined as a fixed point of the voting rule applied to its own induced preference profile: $\pi^*(s) \in \nu\big(Q^{\pi^*}(s)\big)$ for all $s \in \mathcal{S}$.

Theorem 1 (Existence of an Ordinal Fixed Point)**.** *Assume a discount factor $\gamma \in [0, 1)$, and finite state and action spaces. If the voting rule ν is an upper hemicontinuous correspondence mapping preference profiles to non-empty, convex, and compact sets of distributions over actions, and the environment dynamics are continuous with respect to the policy, then there exists at least one optimal policy π^* satisfying $\pi^*(s) \in \nu\big(Q^{\pi^*}(s)\big)$ for all $s \in \mathcal{S}$.*

This existence is mathematically guaranteed via Kakutani's Fixed Point Theorem (Kakutani, 1941) (see Appendix A.1 for the formal proof). These topological conditions are natively satisfied by randomized voting rules that resolve cycles via lotteries, such as Maximal Lotteries. A similar argument was utilized by Gemp et al. (2026) to prove the existence of equilibria in single-state Context-Ordinal Games.

In a finite-horizon setting, this optimal policy can also be constructed via backward induction, structurally identical to solving for a subgame perfect equilibrium in a perfect-information extensive-form game. In infinite-horizon settings, finding the optimal policy requires finding this self-consistent fixed point of the ordinal dynamic programming equations.

3.4 Convergence via Contraction Mappings

For dynamic programming algorithms (such as Value Iteration or Q-learning) to reliably converge, the underlying Bellman operator must be a *contraction mapping*. In traditional reinforcement learning, the standard Bellman optimality operator guarantees convergence to a unique fixed point because the $\max$ operator is non-expansive (i.e., it has a Lipschitz constant of 1).

In a Context-Ordinal MDP, the Bellman-like operator evaluates the time-discounted preferences through the chosen voting rule ν. For ordinal dynamic programming to converge, the mapping $Q \mapsto \nu(Q)$ must preserve this non-expansive property. This introduces a profound theoretical constraint: the requirement of a contraction mapping explicitly filters which social choice functions are admissible for sequential dynamic programming.

If a voting rule is discontinuous, it cannot form a contraction. For example, a deterministic Majority Rule acts as a step function (e.g., shifting 100% of the probability mass to action A the moment its margin over B crosses 0). An infinitesimally small perturbation in the accumulated preference profiles Q can cause the target policy $\nu(Q)$ to violently flip, yielding an infinite Lipschitz constant. Consequently, deterministic ordinal Q-learning may oscillate endlessly.

To guarantee convergence, we can employ a randomized, regularized voting rule that smoothly interpolates probabilities, such as KL-regularized Maximal Lotteries.[3] Crucially, the regularization temperature τ must be large enough to dampen downstream errors, explicitly linking the voting rule to the environment's discount factor γ.

Theorem 2 (Contraction Mapping and Temperature Bound)**.** *Let the voting rule ν_τ be a KL-regularized Maximal Lottery with temperature $\tau > 0$. Then, the ordinal dynamic programming operator is guaranteed to form a strict contraction mapping in policy space provided $\tau > \frac{2\gamma}{1-\gamma}$. Under this condition, ordinal Q-learning unconditionally converges to a unique local fixed point.*

This theorem provides a strict, environment-independent stability threshold for ordinal RL. Unlike cardinal RL, where theoretical bounds are often loose due to unbounded reward magnitudes ($\frac{R_{max}}{1-\gamma}$), the explicitly normalized, proportional nature of ordinal voting margins enables this precise algebraic condition. We provide the full derivation and proof via Banach's Fixed Point Theorem in Appendix A.2.

3.5 Reduction to Cardinal RL

Before proceeding further, we note that formulating the objective as an election where contexts provide temporally-discounted "votes" over actions does not restrict the framework's expressiveness; rather, it serves as a strict generalization of classical reinforcement learning. If we allow contexts to express cardinal preferences instead of strictly ordinal rankings, we recover standard expected-return maximization.

This connection becomes clear under *score voting* (also known as range voting), a rule where each voter assigns a scalar score to every candidate, and the winner is determined by maximizing the aggregate score across the electorate. If we set our voting rule ν to be score voting, and have each context $\mathbf{s}_t$ submit a ballot where the initial action $a \in \mathcal{A}$ receives a score equal to its realized cardinal reward $r(s_t^a)$, the framework collapses back to classical RL. Under these conditions, the aggregated dynamic programming equations evaluate exactly to the standard cardinal Bellman equations (up to an inconsequential $1-\gamma$ scaling factor), and the voting-based policy update $\pi^*(s) \in \nu(Q^\pi(\mathbf{s}))$ perfectly matches the standard greedy policy improvement step.

4 The Agenda Paradox: Local vs. Global Optimality

In traditional reinforcement learning, Bellman's Principle of Optimality guarantees that local, state-by-state optimization perfectly recovers the globally optimal policy. When we transition from cardinal scalar rewards to ordinal social choice functions, we must rigorously verify if this *Optimal Substructure* holds. In this section, we expose the fundamental mathematical tension between local dynamic programming and global evaluation in preference-based RL.

[3] Gemp et al. (2026) introduces *trembling hand* / convolution techniques to render nearly all voting rules continuous.

4.1 Global Policy Elections

To evaluate the optimality of our Context-Ordinal MDP, we first define the theoretical "global" benchmark. One might intuitively define optimality by staging a single, massive election where the candidates are all possible deterministic policies $\pi \in \Pi$. Two policies can be directly compared by the same temporally-discounted comparison of their independent roll outs.

If we apply the Maximal Lotteries voting rule to this policy space, we construct a symmetric zero-sum normal form game of size $|\Pi| \times |\Pi|$, where the payoffs are the expected discounted margins between full trajectories. The optimal global policy is the Nash Equilibrium of this massive matrix. This global evaluation is precisely the objective optimized by recent trajectory-level PbRL algorithms, such as Nash-MD (Munos et al., 2024; Shani et al., 2024).

Does the computationally efficient local fixed point of our CO-MDP match the winner of this intractable global election? In classical game theory, the equivalence between local sequential optimization and global normal-form equilibria relies on the *One-Shot Deviation Principle* (Osborne & Rubinstein, 1994). As we show below, the success or failure of this principle depends entirely on whether the ordinal environment contains Condorcet cycles.

4.2 Condorcet Winners and Losers

A policy is a Condorcet winner (loser) if it wins (loses) against every other policy in a heads-up matchup. Ideally, Q-learning selects winners (if they exist) and avoids losers. A rule that selects Condorcet winners when they exist is called Condorcet consistent. Proofs are in Appendix A.3.

We first establish Condorcet consistency of Q-learning. If the environment's preferences are strictly transitive, or if a ground-truth optimal path clearly exists, the local dynamic programming loop perfectly and safely recovers the global optimum.

Theorem 3 (Condorcet Consistency)**.** *If the environment permits a policy π^* that is globally preferred to all other policies (a global Condorcet Winner), then ordinal dynamic programming under a Condorcet-consistent voting rule (e.g., Maximal Lotteries) will deterministically recover π^*, perfectly matching the global optimum.*

Equally importantly, the framework guarantees a strict lower bound on performance, protecting the agent from the absolute worst-case global policies, i.e., avoiding Condorcet losers.

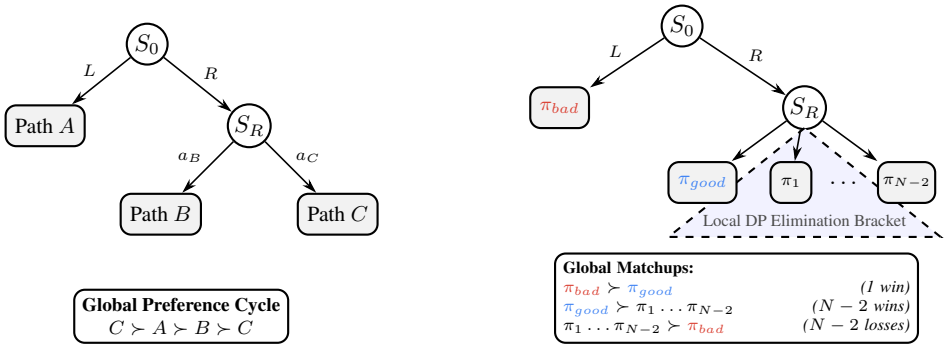
Theorem 4 (The Condorcet Loser Criterion)**.** *If there exists a policy π_{loser} that is globally stochastically dominated by all other policies, ordinal dynamic programming will never converge to it. The local framework mathematically guarantees protection against global Condorcet Losers.*

4.3 The Sequential Tournament Paradox

When preferences are non-transitive, Optimal Substructure fundamentally collapses. Due to phenomena such as *Simpson's Paradox* (Simpson, 1951; Blyth, 1972), non-transitive Condorcet cycles over global policies can emerge organically from temporal discounting, even if the preferences over the underlying state space are completely transitive (see Appendix B.1 for the "Temporal Intransitivity" proof).

In the presence of a global cycle (e.g., $C \succ A \succ B \succ C$), the global Maximal Lottery outputs a mixed stochastic strategy to prevent exploitation. However, because the local dynamic programming loop evaluates the MDP from the bottom up, it functionally acts as a sequential voting agenda (a single-elimination tournament). This structural rigidity results in our first fundamental impossibility theorem.

Theorem 5 (Divergence of Local and Global Lotteries)**.** *There exist Context-Ordinal MDPs where the time-consistent local fixed point computed via ordinal dynamic programming strictly diverges from the global Maximal Lottery computed over the full policy space.*



(a) The Sequential Tournament Paradox

(b) The 2nd-Worst Trap (Agenda Control)

Figure 2: **Divergence of Local Ordinal DP from Global Optimality. (a)** The MDP acts as a single-elimination tournament. At S_R, path C is locally pruned by B. At S_0, A defeats the surviving path B. The local DP converges to A, entirely missing that it is globally exploited by the pruned counterfactual C. **(b)** The agenda can be manipulated to elect the 2nd-worst global policy. Because π_{good} is the Condorcet winner of the right sub-tree, it survives the local DP backups. At the root, π_{bad} faces π_{good} and secures its single, isolated victory, forcing the agent to select a policy that loses to 99.9% of the global policy space.

Proof. We prove this constructively via the counterexample in Figure 2a. At the local state S_R, the agent evaluates B versus C. Because $B \succ C$ locally, the agent deterministically prunes path C. When the root state subsequently evaluates A against the surviving path B, it selects A (since $A \succ B$). The local framework confidently converges to a deterministic policy (A) that is globally exploited by the pruned counterfactual (C), structurally diverging from the unexploitable global mixed strategy. □

Because local dynamic programming is inherently myopic—evaluating one-shot deviations against a fixed future baseline—it systematically breaks macroscopic preference cycles. Crucially, this divergence does not require a deep time horizon. If the horizon is $T = 1$ (a contextual bandit), local and global elections are identical. However, the moment $T \geq 2$, the sequential agenda emerges, forcing a strict structural divergence between local time-consistency and global unexploitability.

4.4 The 2nd-Worst Trap

Establishing that local DP diverges from global evaluation naturally raises the question of severity: how suboptimal can the local fixed point be? Because the transition graph of the MDP acts as an inflexible, dictatorial agenda setter, the divergence can be maximally catastrophic.

Theorem 6 (Arbitrarily Loose Suboptimality)**.** *For any $N \geq 3$, there exists a Context-Ordinal MDP with N global deterministic policies where the local dynamic programming loop deterministically converges to a policy that is globally defeated by $N - 1$ alternative policies in pairwise matchups.*

Proof. We prove this constructively using the "2nd-Worst" trap depicted in Figure 2b. This construction adapts McKelvey's theorems on agenda control (McKelvey, 1986) to the sequential MDP setting. Suppose policy π_{bad} loses to every global policy except for exactly one: the otherwise globally dominant policy π_{good}.

By isolating π_{bad} on the left branch and placing all $N - 1$ other policies on the right branch, π_{good} acts as the strict Condorcet Winner of the right sub-tree. Consequently, local dynamic programming mathematically guarantees π_{good} will systematically defeat all competitors and propagate to S_R. At the root, π_{bad} faces π_{good} and secures its single, isolated pairwise victory. Thus, the local DP loop selects a policy that loses to an $\frac{N-2}{N}$ fraction of the global policy space. □

Natively within the ordinal space, there is no known non-trivial, environment-agnostic bound on the ordinal exploitability of the local policy. Theorem 6 formally cements the fundamental trade-off of Preference-Based RL: one must strictly choose between the computational efficiency of local dynamic programming and the guaranteed unexploitability of global trajectory evaluation.

5 Scaling to Practical Algorithms

The base Context-Ordinal MDP provides a rigorous theoretical foundation, but applying it directly poses several engineering challenges. In this section, we discuss the algorithmic optimizations required to scale this framework to large state-action spaces and deep RL.

5.1 Sufficient Voting Statistics

As defined above, $V^\pi(\mathbf{s})$ tracks a full probability distribution over all possible strict rankings of the available actions. This requires maintaining a histogram of size $\mathcal{O}(|\mathcal{A}|!)$ at each state, which is mathematically general but computationally intractable for large action spaces.

However, we can drastically compress this representation if we know which voting rule ν we intend to use. The choice of voting rule dictates the *sufficient statistics* of the election required to determine the winner. As long as these are linear functions of the histogram, we can seamlessly replace the histogram with these statistics.

For example, if ν is a positional scoring rule like the Borda Count, the agent only needs to track the expected Borda score for each action. By redefining $\mathcal{V}$ as a vector of scores, the dimensionality drops massively from $\mathcal{O}(|\mathcal{A}|!)$ to just $\mathcal{O}(|\mathcal{A}|)$, requiring no more memory than standard Q-learning. Alternatively, Condorcet methods like Maximal Lotteries are based entirely on pairwise comparisons. Thus, it suffices to track an $|\mathcal{A}| \times |\mathcal{A}|$ margin matrix of pairwise win-rates. By tailoring the tracked statistics to the chosen voting rule, the social choice framework becomes highly efficient and practical to implement using standard RL representations.

5.2 State-Action Preferences and Function Approximation

While our foundational formulation defined preferences purely over subsequent states—analogous to an $r(s')$ transition reward model for simplicity of exposition—this introduces a slight temporal offset between a decision and its evaluation. Furthermore, generating a valid state context $\mathbf{s}_t$ rigidly requires unrolling *synchronous, parallel, counterfactual* trajectories from the exact same state, which is often impossible in standard environment interactions.

We can resolve the temporal ambiguity and the data collection constraint by extending the preference relation to evaluate state-action tuples: $\rho(\mathbf{x})$. Let $\mathbf{x}_s = \big((s, a_1), \ldots, (s, a_{|\mathcal{A}|})\big)$ denote the context of all available actions at state s. Ordinal Q-learning acts greedily over this context by defining the current policy as $\pi_k(\cdot \mid s) \in \nu\big(Q_k(\mathbf{x}_s)\big)$.

By directly tracking the voting statistics over state-action pairs, we can bypass the state-value function entirely and bootstrap using standard temporal-difference logic:

$$Q_{k+1}(\mathbf{x}) \leftarrow Q_k(\mathbf{x}) + \alpha\Big((1-\gamma)v(\mathbf{x}) + \gamma\mathbb{E}_{\mathbf{a}'\sim\pi_k(\cdot|\mathbf{s}'),\,\mathbf{s}'\sim P(\cdot|\mathbf{x})}\big[Q_k(\mathbf{s}', \mathbf{a}')\big] - Q_k(\mathbf{x})\Big) \quad (3)$$

This single update elegantly extends standard off-policy Q-learning to the ordinal regime.

Eliminating Counterfactuals via Function Approximation. In tabular settings, generating valid evaluation contexts requires unrolling synchronous, counterfactual trajectories from the exact same root state (Figure 1). By formulating voting statistics over state-action pairs, $Q^\pi(X)$, alongside function approximation, we completely decouple this requirement. The agent can execute independent trajectories and store them in a standard off-policy replay buffer. During training, a neural network parameterized over (s, a) learns to predict the discounted voting statistics from arbitrary

batches of offline data. At decision time, to evaluate the specific context of all available actions $\boldsymbol{x}_s = \{(s, a_1), \ldots, (s, a_{|\mathcal{A}(s)|})\}$, the network generalizes across the rich state-action feature space, effectively inferring the relative voting statistics of these unobserved counterfactuals without ever requiring synchronous multi-branch data collection.

6 Related Work

Our formulation of recursive voting statistics generalizes the standard Bellman expectation and optimality operators, situating our work alongside the rich literature on abstract dynamic programming (Bertsekas, 2022). Early foundational works explored dynamic programming directly under preference relations (Mitten, 1974; Sobel, 1975), and recent advances have formalized dynamic programming operations over partially ordered sets (Sargent & Stachurski, 2025) and ordered vector spaces (Peng & Stachurski, 2026). While these works establish the mathematical conditions (e.g., monotonicity and discounting) required for convergence in abstract spaces, they rely heavily on the assumption that the underlying preferences form a strict, transitive partial order. Our Context-Ordinal framework fundamentally departs from this classical assumption. By aggregating temporal evaluations via social choice voting rules, our framework naturally permits non-transitive preference cycles (the Condorcet paradox), requiring DP operators that solve for game-theoretic equilibria (e.g., Maximal Lotteries) rather than standard order-theoretic maximal elements.

7 Conclusion

By treating reinforcement learning through the lens of social choice, we introduce a fundamentally new way to evaluate trajectories without explicit numeric rewards. By analyzing the theoretical boundaries of Context-Ordinal MDPs, we uncover two profound implications for the modern RL from Human Feedback (RLHF) pipeline.

The Condorcet Winner: Safe, Reward-Free RL. If a policy stochastically dominates all alternatives (acting as a strict global Condorcet Winner), practitioners can bypass delicate, error-prone scalar reward models and apply Ordinal Q-learning directly. Global dominance natively restores Optimal Substructure, ensuring local dynamic programming safely and efficiently recovers the global optimum. Furthermore, operating exclusively on ordinal margins renders this policy structurally immune to the heavy-tailed proxy misspecifications and magnitude-based reward hacking that plague traditional alignment methods.

The Emergence of Cycles: Why RLHF is Hard. In practice, Condorcet winners are exceptionally rare. Modern RLHF relies on crowdsourced feedback, which inevitably yields cyclic preferences (Arrow, 1950). Even with perfectly rational evaluators, temporal discounting mathematically manufactures global cycles (Appendix B.1). Under cyclic preferences, the decision problem fundamentally loses its Markovian Optimal Substructure, explaining why training in these environments is notoriously unstable. Using local temporal-difference learning turns the MDP's sequential structure into a dictatorial voting agenda. Local Bellman updates aggressively prune branches based on myopic skirmishes, causing the learned policy to catastrophically diverge from the global optimum. Ultimately, this establishes a fundamental theoretical limit: when aligning agents in cyclic environments, computational tractability and global unexploitability are mutually exclusive.

Finally, we note that while our framework structurally isolates global Condorcet winners, the theoretical requirement of KL-regularization for algorithmic convergence inherently smooths the resulting policies. Formally bridging the gap between this regularized contraction mapping and the exact, deterministic global optimum remains an exciting open challenge for preference-based RL.

A Proofs of Main Results

A.1 Proof of the Existence of a Fixed-Point Policy (Theorem 1)

Let Π denote the space of all stationary Markov policies. Because $\mathcal{S}$ and $\mathcal{A}$ are finite, $\Pi = [\Delta(\mathcal{A})]^{\mathcal{S}}$ is a non-empty, compact, and convex subset of a finite-dimensional Euclidean space.

First, we establish the continuity of the voting statistics with respect to the policy. Under any policy π, the discounted state visitation frequencies are determined by the resolvent matrix $(I - \gamma P^{\pi})^{-1}$. Because the transition matrix P^{π} is continuous in π, and the matrix inversion operation is continuous for strictly diagonally dominant matrices (guaranteed by $\gamma < 1$), the mapping $\pi \mapsto (I - \gamma P^{\pi})^{-1}$ is continuous. Consequently, the mapping from the policy to the resulting expected temporal voting statistics at any state s, $\pi \mapsto Q^{\pi}(s)$, is a continuous function.

Second, we define the state-conditional policy update correspondence $\Phi_s : \Pi \rightrightarrows \Delta(\mathcal{A})$ as $\Phi_s(\pi) = \nu(Q^{\pi}(s))$. Because $Q^{\pi}(s)$ is a continuous function and ν is an upper hemicontinuous correspondence, their composition Φ_s is also an upper hemicontinuous correspondence. By the premise of the theorem, $\Phi_s(\pi)$ yields non-empty, compact, and convex sets.

Finally, we define the global policy update correspondence $\Phi : \Pi \rightrightarrows \Pi$ by concatenating the operators for each state: $\Phi(\pi) = \prod_{s \in \mathcal{S}} \Phi_s(\pi)$. Because the Cartesian product of upper hemicontinuous correspondences with compact values is itself upper hemicontinuous, Φ is u.h.c. Furthermore, the Cartesian product of non-empty, compact, and convex sets is non-empty, compact, and convex.

Therefore, Φ is an upper hemicontinuous correspondence mapping a non-empty, compact, convex set Π into non-empty, compact, convex subsets of itself. By Kakutani's Fixed Point Theorem (Kakutani, 1941), there exists at least one fixed point $\pi^* \in \Pi$ such that $\pi^* \in \Phi(\pi^*)$. This implies $\pi^*(s) \in \nu\big(Q^{\pi^*}(s)\big)$ for all $s \in \mathcal{S}$, concluding the proof.

Remark: The assumption that ν returns non-empty, compact, and convex sets is a standard property of randomized voting rules that resolve cycles via lotteries. For instance, the Maximal Lotteries rule computes the Nash equilibrium of a zero-sum preference game, and the set of Nash equilibria in finite zero-sum games is well-known to be non-empty, compact, and convex. A similar topological argument was utilized in Gemp et al. (2026) to prove the existence of equilibria in context-ordinal games.

A.2 Proof of Contraction Mapping (Theorem 2)

Here, we prove that KL-regularized Maximal Lotteries induces a strict contraction mapping in policy space, provided the temperature parameter τ satisfies $\tau > 2\gamma/(1-\gamma)$.

1. Lipschitz Continuity of the Voting Rule. The KL-regularized Maximal Lottery is defined as the Nash Equilibrium of the regularized zero-sum preference game. Evaluating this rule over an action-value matrix Q yields:

$$\nu_{\tau}(Q) = \arg\max_{p \in \Delta} \min_{q \in \Delta} p^{\top} Q q - \tau \mathrm{KL}(p \parallel p_{\text{ref}}) + \tau \mathrm{KL}(q \parallel q_{\text{ref}}) \tag{4}$$

By Pinsker's Inequality, the unnormalized negative entropy function is strictly 1-strongly convex with respect to the L_1-norm on the probability simplex (Beck & Teboulle, 2003). This regularization formulates the game as a strongly monotone variational inequality with modulus τ (Mertikopoulos & Sandholm, 2016). Through the sensitivity analysis of strongly convex saddle-point problems (Facchinei & Pang, 2003), the Jacobian of the equilibrium policy with respect to the payoffs is strictly bounded by the inverse of the strong convexity modulus (Ling et al., 2018). Therefore, the policy update is Lipschitz continuous, such that for any two action-value matrices Q_1, Q_2:

$$\|\nu_{\tau}(Q_1) - \nu_{\tau}(Q_2)\|_1 \leq \frac{1}{\tau} \|Q_1 - Q_2\|_{\infty} \tag{5}$$

2. Sensitivity of the Backed-Up Margin. We must now bound how much the expected backed-up margin matrix shifts between any two arbitrary policies π_1 and π_2. Consider a root state s, and let $Q^\pi(s)$ denote the $|\mathcal{A}| \times |\mathcal{A}|$ action-value margin matrix. For any specific pair of actions a and b, the transition yields the subsequent state context $\mathbf{s}' = (s^a, s^b)$. The expected backed-up margin for this scalar entry is:

$$Q^\pi(s)[a,b] = (1-\gamma)\rho(a,b) + \gamma \mathbb{E}_{\mathbf{s}'}\left[V^\pi(\mathbf{s}')\right] \tag{6}$$

where $V^\pi(\mathbf{s}') = \pi(s^a)^\top Q^\pi(\mathbf{s}')\pi(s^b)$.

The exact difference in the expected future state-value under π_1 versus π_2 requires accounting for both the policy shift and the recursive shift in the future action-value matrix:

$$\Delta V(\mathbf{s}') = V^{\pi_1}(\mathbf{s}') - V^{\pi_2}(\mathbf{s}') = \pi_1(s^a)^\top Q^{\pi_1}(\mathbf{s}')\pi_1(s^b) - \pi_2(s^a)^\top Q^{\pi_2}(\mathbf{s}')\pi_2(s^b)$$

To bound this, we add and subtract mixed terms $\pi_2(s^a)^\top Q^{\pi_1}(\mathbf{s}')\pi_1(s^b)$ and $\pi_2(s^a)^\top Q^{\pi_1}(\mathbf{s}')\pi_2(s^b)$:

$$\begin{aligned}\Delta V(\mathbf{s}') = &(\pi_1(s^a) - \pi_2(s^a))^\top Q^{\pi_1}(\mathbf{s}')\pi_1(s^b) \\ &+ \pi_2(s^a)^\top Q^{\pi_1}(\mathbf{s}')(\pi_1(s^b) - \pi_2(s^b)) \\ &+ \pi_2(s^a)^\top \big(Q^{\pi_1}(\mathbf{s}') - Q^{\pi_2}(\mathbf{s}')\big)\pi_2(s^b)\end{aligned} \tag{7}$$

Because temporal votes are normalized, $\|Q^\pi\|_\infty \leq 1$. Let $\Delta\pi = \pi_1 - \pi_2$ and $\Delta Q = Q^{\pi_1} - Q^{\pi_2}$. Applying Hölder's Inequality ($|x^\top Q y| \leq \|x\|_1 \|Q\|_\infty \|y\|_1$), we bound the magnitude of the scalar entry shift $\Delta Q(s)[a,b] = Q^{\pi_1}(s)[a,b] - Q^{\pi_2}(s)[a,b]$:

$$\begin{aligned}|\Delta Q(s)[a,b]| &\leq \gamma \mathbb{E}_{\mathbf{s}'}\Big[\|\Delta\pi(s^a)\|_1(1)\|\pi_1(s^b)\|_1 + \|\pi_2(s^a)\|_1(1)\|\Delta\pi(s^b)\|_1 \\ &\quad + \|\pi_2(s^a)\|_1 \|\Delta Q(\mathbf{s}')\|_\infty \|\pi_2(s^b)\|_1\Big] \\ &\leq \gamma\left(2\max_{s'}\|\Delta\pi(s')\|_1 + \|\Delta Q\|_\infty\right)\end{aligned} \tag{8}$$

Because this upper bound holds identically for all scalar action pairs $[a,b]$, the L_∞-norm of the entire matrix shift is bounded by the exact same quantity:

$$\|\Delta Q\|_\infty \leq 2\gamma\|\Delta\pi\|_1 + \gamma\|\Delta Q\|_\infty \tag{9}$$

Rearranging terms to solve for the recursive infinite-horizon accumulation yields:

$$\|\Delta Q\|_\infty \leq \frac{2\gamma}{1-\gamma}\|\Delta\pi\|_1 \tag{10}$$

3. The Contraction Mapping. Let $\Phi(\pi) = \nu_\tau(Q^\pi)$ denote the ordinal dynamic programming operator. Chaining the bounds from equation 5 and equation 10:

$$\|\Phi(\pi_1) - \Phi(\pi_2)\|_1 \leq \frac{1}{\tau}\|\Delta Q\|_\infty \leq \left(\frac{2\gamma}{\tau(1-\gamma)}\right)\|\pi_1 - \pi_2\|_1 \tag{11}$$

For Φ to be a strict contraction mapping, its Lipschitz constant must be strictly less than 1, requiring $\tau > \frac{2\gamma}{1-\gamma}$.

If this strict temperature bound holds, Banach's Fixed Point Theorem guarantees that the policy space operator is a strict contraction, ensuring that iterative updates (ordinal Q-learning) will unconditionally converge to a unique, stable local fixed point.

A.3 Proofs of Condorcet Winners and Losers

Theorem 3. *If the environment permits a policy π^* that is globally preferred to all other policies (a global Condorcet Winner), then ordinal dynamic programming under a Condorcet-consistent voting rule (e.g., Maximal Lotteries) will deterministically recover π^*, perfectly matching the global optimum.*

Proof. By definition, if π^* is globally preferred to all other policies over the temporal context distribution, it acts as the strict global Condorcet Winner. Because Maximal Lotteries is Condorcet-consistent, the global Nash Equilibrium collapses to a pure, deterministic strategy assigning 100% probability to π^*. In the local dynamic programming framework, the policy update at any state s evaluates actions assuming the agent follows the baseline policy thereafter. Consider evaluating π^* at state s against any unilateral action deviation. Because π^* strictly defeats all alternative global policies, it must strictly defeat any policy that deviates for a single step. Therefore, the action $\pi^*(s)$ will systematically win the local pairwise election at every state. Consequently, π^* is mathematically guaranteed to be a stable local fixed point of the ordinal Bellman equations. □

Equally importantly, the framework guarantees a strict lower bound on performance, protecting the agent from the absolute worst-case global policies, i.e., avoiding Condorcet losers.

Theorem 4. *If there exists a policy π_{loser} that is globally stochastically dominated by all other policies, ordinal dynamic programming will never converge to it. The local framework mathematically guarantees protection against global Condorcet Losers.*

Proof. Suppose, for the sake of contradiction, that the local ordinal dynamic programming loop converges to the globally dominated policy π_{loser}. For π_{loser} to be the converged solution, it must be a local fixed point, meaning no unilateral action deviation at any state s can strictly defeat it in the local election. However, consider a one-shot deviation policy π' that diverges from π_{loser} for a single action at state s, but perfectly mimics π_{loser} thereafter. Because π_{loser} is globally dominated by all policies, it is globally defeated by π'. Furthermore, because π' and π_{loser} share the exact same future trajectory distribution from state s onward, their global margin is mathematically identical to their local margin. Consequently, the local election at state s will strictly prefer the deviation, breaking the fixed point. Thus, it is mathematically impossible for a global Condorcet loser to be a local fixed point. □

B Additional Illustrative Examples & Counterexamples

B.1 The Temporal Intransitivity Counterexample

One might assume that the Sequential Tournament Paradox only occurs if the human evaluator possesses inherently cyclic, irrational preferences over the state space (see Figure 2a). Remarkably, this is false. Due to a phenomenon analogous to non-transitive dice (Savage Jr, 1994), the synchronous temporal discounting of rollouts can manufacture strict Condorcet cycles over global policies, even when the underlying state space is perfectly transitive.

Consider an MDP where all states can be cleanly evaluated into three strictly transitive tiers: High (3), Medium (2), and Low (1), such that $3 \succ 2 \succ 1$.

Let the discount factor be $\gamma = 0.9$. The pairwise margin between any two time-synchronized states is $+1$ if the state is strictly preferred, -1 if dispreferred, and 0 if tied.

Global Evaluation: If we evaluate the global election, the expected margins $M(\tau, \tau') = \sum \gamma^t m_t$ are:

- $M(A, B) = (+1) + 0.9(-1) + 0.81(+1) = +\mathbf{0.91} \implies A \succ B$
- $M(B, C) = (0) + 0.9(+1) + 0.81(-1) = +\mathbf{0.09} \implies B \succ C$
- $M(C, A) = (-1) + 0.9(+1) + 0.81(+1) = +\mathbf{0.71} \implies C \succ A$

Despite the state space lacking any cyclic preferences, the global policy space exhibits a massive Condorcet cycle. The global Maximal Lottery correctly outputs a mixed strategy to prevent exploitation.

Local DP Evaluation: The local dynamic programming loop evaluates the tree from the bottom up. At state s_R ($t = 1$), the local election compares the sub-paths of B and C. The local margin

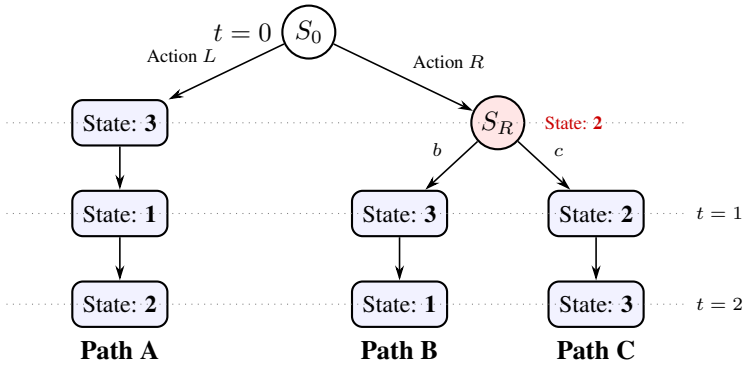


Figure 3: **Temporal Intransitivity.** The state preferences are strictly ordered ($3 \succ 2 \succ 1$). However, the temporal alignment of the rewards causes the global trajectories to form a Condorcet cycle ($A \succ B \succ C \succ A$).

is $M_{local}(B, C) = (+1) + 0.9(-1) = +\mathbf{0.1}$. Because $B \succ C$ locally, the agent deterministically prunes path C and commits to B. At the root s_0, the agent compares A against B. Because $A \succ B$ (margin $+0.91$), the root deterministically selects A.

Conclusion: The local CO-MDP converges to a 100% deterministic policy (Path A), which is globally exploited by Path C. This proves that the Agenda Paradox is not an artifact of irrational human evaluators. Time-synchronization inherently manufactures cyclic preferences over trajectories, guaranteeing that local ordinal dynamic programming will systematically diverge from global optimality in strictly transitive state spaces.

B.2 The Binary Tree Suboptimality Trap

This example demonstrates the potentially slow learning speed of the voting framework in sequential environments.

Consider a deterministic binary tree MDP of depth T with two actions: u (up/optimal) and d (down/sub-optimal). The initial policy is uniformly random. All intermediate transition rewards are 0; non-zero rewards are only distributed at the terminal leaves. If the agent deviates from the optimal path by taking action d at depth t, it enters a sub-tree where *every* leaf yields a terminal reward of r_t. We constrain the rewards such that $2^{-T} > r_0 > r_1 > \cdots > r_{T-1} > 0$.

Observation 7 (Slower Backpropagation)**.** *Under a uniform random policy, cardinal Q-learning identifies the optimal path at every node in $O(1)$ steps. The voting approach requires $O(T)$ steps for the optimal policy to reach the root.*

Cardinal Evaluation: At any state s_t on the optimal path, $Q^\pi(s_t, d) = r_t$. Conversely, $Q^\pi(s_t, u) > 2^{-(T-t-1)} \cdot 1 \geq 2^{-(T-1)} > 2^{-T} > r_0 \geq r_t$. Cardinal Q-learning correctly identifies u as optimal at *every* node instantly.

Voting Evaluation: Because all intermediate rewards are exactly 0, any context comparisons prior to the terminal step result in strict ties ($u \sim d$). Under majority rule (or any margin-based symmetric voting rule), these ties contribute zero net margin to the election, meaning the overall winner is strictly determined by the votes at the terminal leaves. At the terminal step, a rollout for d always yields exactly r_t. A rollout for u yields $+1$ with probability $p = 2^{-(T-t-1)}$, but otherwise yields some r_k (where $k > t$). Because $r_t > r_k$, the u action strictly loses the terminal vote whenever it fails to hit $+1$. At shallow depths, $p \ll 0.5$, so d decisively wins the overall election. The framework relies on the sub-trees "slowly fixing" their policies round-by-round from the leaves upward, requiring $O(T)$ steps for the optimal policy to reach the root.

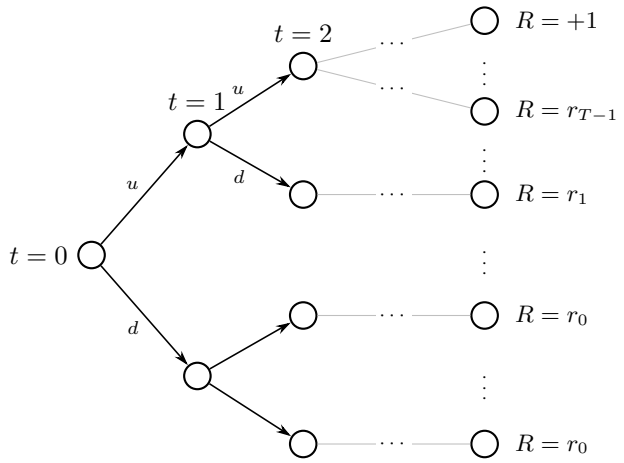


Figure 4: The binary tree structure causing $O(T)$ backpropagation delay in voting RL.

B.3 Bounding the Suboptimality Gap

Because strict voting rules aggregate ordinal rankings — evaluating *which* trajectory is better rather than by *how much* — they fundamentally act as *median-seeking* estimators rather than mean-seeking estimators. In a 2-action setting, this rank-based evaluation mathematically reduces to comparing pairwise win-rates. While this bounds the influence of extreme outliers, it renders the voting framework vulnerable to highly skewed distributions. Specifically, the agent becomes structurally blind to tail risks, favoring policies that yield frequent marginal gains while systematically ignoring low-probability, catastrophic losses.

Consider a 1-step bandit MDP where the cardinal approach correctly identifies action a_c as optimal, but the voting approach erroneously prefers a_v because it wins the majority of rollouts ($P(R_v > R_c) > 0.5$). Let the difference in rewards be $Z = R_v - R_c$. By applying Cantelli's inequality (the one-sided Chebyshev bound), we can strictly bound the expected suboptimality of the voting policy.

Theorem 8 (Cantelli Upper Bound)**.** *The maximum expected suboptimality gap $\Delta = \mathbb{E}[R_c] - \mathbb{E}[R_v]$ incurred by a voting-based policy is strictly bounded by the standard deviation of the difference in rewards: $\Delta \leq \sigma_Z \leq \sigma_v + \sigma_c$.*

Proof. Let $Z = R_v - R_c$. Let $\Delta = \mathbb{E}[R_c] - \mathbb{E}[R_v] \geq 0$. By definition, $\mathbb{E}[Z] = -\Delta$. Let $\sigma_Z^2 = \text{Var}(Z)$. Because the voting approach erroneously selected a_v, we know $P(Z > 0) \geq 0.5$.

Cantelli's inequality states that for any random variable Z and $k > 0$:

$$P(Z - \mathbb{E}[Z] \geq k) \leq \frac{\sigma_Z^2}{\sigma_Z^2 + k^2}$$

Setting $k = \Delta$, we get $P(Z \geq 0) \leq \frac{\sigma_Z^2}{\sigma_Z^2 + \Delta^2}$. Because $P(Z \geq 0) \geq 0.5$, we have:

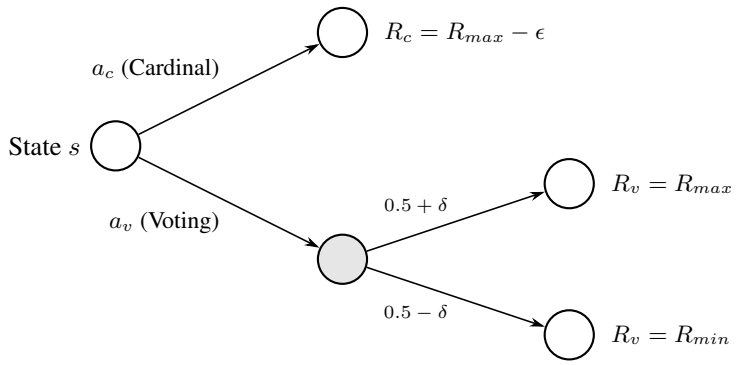
$$0.5 \leq \frac{\sigma_Z^2}{\sigma_Z^2 + \Delta^2} \implies \Delta^2 \leq \sigma_Z^2 \implies \Delta \leq \sigma_Z$$

Assuming worst-case negative correlation ($\rho = -1$), $\text{Var}(Z) \leq (\sigma_v + \sigma_c)^2$, yielding $\Delta \leq \sigma_v + \sigma_c$. □

While Theorem 8 establishes a general upper bound, this bound can be remarkably loose in highly skewed environments. If numeric rewards represent the absolute ground truth, the voting framework can be seduced by a slight majority win-rate, blinding it to massive catastrophic downsides. We formalize this vulnerability by establishing a strict lower bound on the worst-case loss.

Theorem 9 (Bounded Reward Lower Bound)**.** *In environments where all rewards fall within a finite bounded range $[R_{min}, R_{max}]$, the worst-case expected loss of the voting approach is bounded below by exactly half the reward span: $\frac{1}{2}(R_{max} - R_{min})$.*

Proof. Consider an adversarial MDP where a_c yields $R_c = R_{max} - \epsilon$ deterministically. Action a_v yields R_{max} with probability $p = 0.5 + \delta$, and R_{min} with $p = 0.5 - \delta$.



The voting framework compares the actions: $P(R_v > R_c) = 0.5 + \delta > 0.5$. The sub-optimal action a_v strictly wins the election. However, evaluating the cardinal expectations yields $\Delta = (0.5 - \delta)(R_{max} - R_{min}) - \epsilon$. As $\delta, \epsilon \to 0$, the gap approaches $\frac{1}{2}(R_{max} - R_{min})$. □

B.4 The Condorcet Paradox: Emergence of Stochastic Policies

Because cardinal reinforcement learning compresses reward distributions into well-ordered scalar expectations ($\mathbb{E}[A] > \mathbb{E}[B] > \mathbb{E}[C]$), it is mathematically guaranteed that a deterministic optimal policy exists. Voting frameworks, however, evaluate pairwise probabilities, which are not strictly ordered. This appendix provides a concrete 1-step MDP demonstrating how ordinal evaluation natively gives rise to non-transitive Condorcet cycles, forcing the emergence of stochastic policies.

Consider a 1-step MDP (a 3-armed bandit) where actions yield independent, stochastic reward distributions that mimic non-transitive dice:

- **Action a:** Yields a deterministic reward $R_a = 3$.
- **Action b:** Yields $R_b = 4$ with probability 0.6, and $R_b = 0$ with probability 0.4.
- **Action c:** Yields $R_c = 5$ with probability 0.4, and $R_c = 1$ with probability 0.6.

Pairwise Voting Preferences

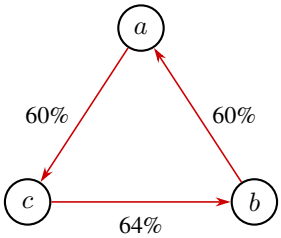


Figure 5: The non-transitive Condorcet cycle formed by the ordinal evaluation of independent reward distributions.

Cardinal Evaluation: The cardinal approach seeks the policy that maximizes the scalar expected return:

$$\mathbb{E}[R_a] = 3.0$$
$$\mathbb{E}[R_b] = (0.6)(4) + (0.4)(0) = 2.4$$
$$\mathbb{E}[R_c] = (0.4)(5) + (0.6)(1) = 2.6$$

Because expected values are scalars, no cycles can exist. Since $\mathbb{E}[R_a] > \mathbb{E}[R_c] > \mathbb{E}[R_b]$, the cardinal approach effortlessly returns a deterministic optimal policy: $\pi(a) = 1$.

Voting Evaluation: The voting approach compares actions via pairwise rollouts. We evaluate the win-rates for each pair (note that exact ties have 0% probability in this specific continuous-like formulation):

- b **vs.** a**:** Action b gets 4 (beating a's 3) with probability 0.6. Thus, $P(R_b > R_a) = \mathbf{0.60}$.
- c **vs.** b**:** Action c wins if it gets 5 (prob 0.4). If c gets 1 (prob 0.6), it still wins if b gets 0 (prob 0.4). Thus, $P(R_c > R_b) = 0.4 + (0.6 \times 0.4) = \mathbf{0.64}$.
- a **vs.** c**:** Action a gets 3, which beats c only when c gets 1 (prob 0.6). Thus, $P(R_a > R_c) = \mathbf{0.60}$.

This creates a strict Condorcet cycle: $b \succ a \succ c \succ b$. In the ordinal voting framework, *no deterministic action is optimal*. If the policy deterministically selects a, it is strictly exploited by b in a majority of rollouts; if b, it is exploited by c; if c, it is exploited by a.

To resolve this, the Maximal Lotteries framework treats the preferences as a symmetric zero-sum game with payoff matrix $M(x, y) = P(R_x > R_y) - P(R_y > R_x)$:

$$M = \begin{pmatrix} 0 & -0.20 & +0.20 \\ +0.20 & 0 & -0.28 \\ -0.20 & +0.28 & 0 \end{pmatrix}$$

The solution to the Maximal Lottery is the Nash Equilibrium of this preference game. Solving for the mixed strategy π_{ML} yields:

$$\pi_{ML} = \left[\frac{7}{17}, \frac{5}{17}, \frac{5}{17}\right] \approx [41.2\%,\ 29.4\%,\ 29.4\%]$$

Conclusion: Even in a fully observable, 1-step environment, ordinal evaluation natively breaks transitivity. This fundamentally forces the voting approach to converge on a *stochastic (mixed) policy* to avoid exploitation, contrasting sharply with the deterministic pure strategies yielded by cardinal Q-learning.

B.5 Robustness to Reward Misspecification

The primary vulnerability of the voting approach (ignoring magnitudes) becomes its greatest strength when the environment's numeric rewards are untrustworthy. In practical alignment tasks, the observed reward $\tilde{R}$ is often a noisy proxy for true human utility R^*.

Consider an "Outlier Trap" MDP with two actions: a reliable action ($R^* = 1$, observed flawlessly as $\tilde{R} = 1$) and a risky action. The risky action yields a true utility of $R^* = 0$ in 99% of cases and $R^* = 50$ in the remaining 1%. However, due to a sensor glitch, the rare 50 reward event is misreported as a massive hallucinated reward of $\tilde{R} = 500$.

Observation 10 (The Outlier Trap)**.** *Under heavy-tailed monotonic reward misspecification, cardinal RL suffers catastrophic failure by maximizing the hallucinated outlier. The voting-based approach maintains a suboptimality gap of zero on the true objective.*

Implication: Because the cardinal approach maximizes the algebraic expected value $\mathbb{E}[\tilde{R}]$, its influence function is unbounded; a single massive outlier completely hijacks the policy. The voting

approach limits the impact of any single rollout to exactly 1 vote. It observes that the reliable action beats the risky action 99% of the time and decisively ignores the magnitude of the 500 proxy glitch. This mathematical invariance to extreme magnitudes provides a rigorous justification for using PbRL to prevent reward hacking.

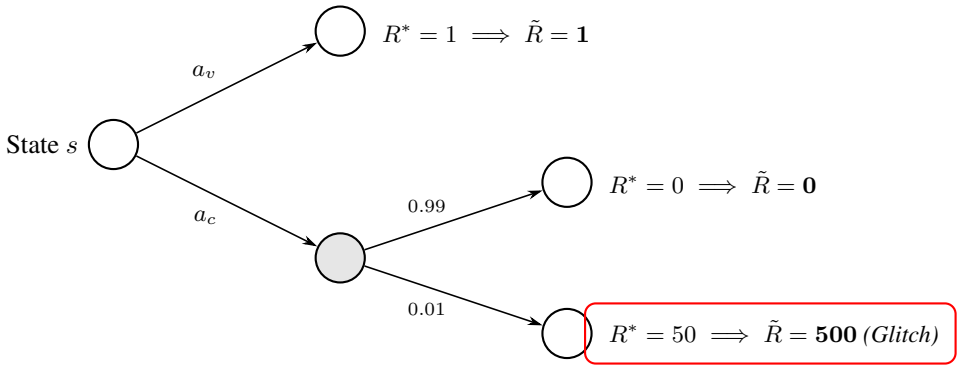


Figure 6: A 2-action MDP with one stochastic terminal reward.

Evaluating the true objective, $\mathbb{E}[R^*(a_v)] = 1 > 0.5 = \mathbb{E}[R^*(a_c)]$. The cardinal agent operates on $\tilde{R}$, calculates $\mathbb{E}[\tilde{R}(a_c)] = 5 > 1$, and is completely hijacked by the outlier. The voting framework evaluates ordinal preferences: $P(\tilde{R}(a_v) > \tilde{R}(a_c)) = P(\tilde{R}(a_c) = 0) = 0.99$. It decisively selects the robust action a_v, perfectly immune to the extreme magnitude of the proxy glitch.

C Extended Discussion on Global Trajectory Evaluation and Nash-MD

Prior works on preference-based RL approaches have specifically considered Maximal Lotteries as the voting rule (Munos et al., 2024; Shani et al., 2024), although the social choice perspective was not recognized until later, and only for the single-turn setting (Maura-Rivero et al., 2025). Here, we provide a detailed walkthrough of how our proposed framework conceptually connects to—and structurally diverges from—the approach proposed for the multi-turn setting by Shani et al. (2024).

C.1 Derivation of Global Mirror Descent

Maximal lotteries are computed by solving for the Nash Equilibrium (NE) of an appropriately constructed 2-player, zero-sum game. Let $P(a, a' \mid s)$ denote the frequency (probability) with which action a was preferred to action a' in terms of their resulting trajectories beginning at state s. The margin matrix is defined as $M(s) = P(s) - P(s)^T$, where $P(s)$ is the $|\mathcal{A}| \times |\mathcal{A}|$ preference matrix.

If no votes of indifference ($a \sim a'$) exist, then $P(a', a \mid s) = 1 - P(a, a' \mid s)$ and the margin evaluates to $M(a, a' \mid s) = 2P(a, a' \mid s) - 1$. A positive affine transformation of $P(s)$ does not change the equilibrium computed, so we can focus on maximizing the expected margin:

$$\max_{\pi(s)} \min_{\pi'(s)} \pi(s)^\top M(s) \pi'(s) \tag{12}$$

Critically, M is skew-symmetric, meaning the game is symmetric across players. Consequently, we can safely restrict our computation to symmetric Nash Equilibria.

In the multi-turn generalization of NashMD (Shani et al., 2024), the goal is to construct an appropriate state-action value function Q such that Q-learning converges to this global equilibrium. The authors recognize a bridge between this process and a typical gradient ascent process used to compute the NE of a maximin game. Specifically, gradient ascent on a KL-regularized game:

$$\max_{\pi(s)} \min_{\pi'(s)} \pi(s)^\top M(s) \pi'(s) - \alpha \, \mathrm{KL}(\pi \parallel \pi_{\text{ref}}) + \alpha \, \mathrm{KL}(\pi' \parallel \pi'_{\text{ref}}) \tag{13}$$

converges to the KL-regularized equilibrium. Convergence is guaranteed because the KL divergence induces a strongly monotone game. A generalization of gradient ascent, mirror descent, is typically

written in a proximal form as:

$$\pi_{t+1} = \text{argmax}_\pi \left[\pi(s)^\top M(s)\pi'(s) - \alpha \, \text{KL}(\pi \parallel \pi_{\text{ref}}) - \frac{1}{\eta_t} \text{KL}(\pi \parallel \pi_t) \right] \tag{14}$$

where the Bregman divergence term effectively controls the step size via η_t. Because the solution is invariant to scale, it can also be rewritten as it appears in their paper:

$$\pi_{t+1} = \text{argmax}_\pi \left[\eta_t \pi(s)^\top M(s)\pi'(s) - \alpha\eta_t \, \text{KL}(\pi \parallel \pi_{\text{ref}}) - (1 - \alpha\eta_t) \, \text{KL}(\pi \parallel \pi_t) \right] \tag{15}$$

Crucially, they recognize that if they define the Q-function using a self-play assumption where the opponent policy π' is set to the current policy π, $Q(s) = M(s)\pi$, then this mirror descent update takes the exact form of entropy-regularized Q-learning. This forms the theoretical basis of their approach: taking the entropy-regularized argmax of this learned Q-function effectively performs mirror descent towards a KL-regularized Maximal Lottery.

C.2 The Role of Reach Probabilities in Global Policy Gradients

As discussed in Section 4.1, global trajectory evaluation algorithms (such as MTPO and Nash-MD) face strict structural limitations regarding the Markov property. Here, we formally derive the global policy gradient to illuminate exactly where reach probabilities must be explicitly accounted for, and when they can be implicitly handled by Monte Carlo sampling.

Let π_θ be the agent's policy and μ be an independent opponent policy. Let $\tau = (s_0, a_0, \ldots, s_T)$ be the agent's trajectory, and τ' be the opponent's trajectory. The global objective is to maximize the expected trajectory-level margin: $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta,\tau'\sim\mu}[M(\tau, \tau')]$. Using the standard score-function estimator, the gradient is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^{\pi,\mu}(h_t, a_t) \right] \tag{16}$$

where $h_t = (s_0, a_0, \ldots, s_t)$ is the specific history of the trajectory up to time t. The action-value function is formally defined over this history:

$$Q^{\pi,\mu}(h_t, a_t) = \mathbb{E}_{\tau_{>t}\sim\pi,\tau'\sim\mu} \Big[M(\tau, \tau') \;\Big|\; h_t, a_t \Big] \tag{17}$$

If we attempt to compress this history-dependent value function into a standard Markovian state-action value function $Q^{\pi,\mu}(s_t, a_t)$, we must marginalize over the *posterior reach probabilities*—the probability that the agent experienced history h_t given that it is currently in state s_t under policy π:

$$Q^{\pi,\mu}(s_t, a_t) = \sum_{h_t} P(h_t \mid s_t, \pi_\theta) Q^{\pi,\mu}(h_t, a_t) \tag{18}$$

The tractability of Equation 18 dictates the boundary conditions of global evaluation.

Case 1: Trajectory Preferences in Append-Only States (Trees). In Large Language Model (LLM) alignment, the environment is a deterministic tree where the state s_t consists of the prompt plus all generated tokens. Therefore, the state *is* the history ($s_t \equiv h_t$). Because it is physically impossible to reach s_t via any other sequence of actions, the posterior reach probability collapses deterministically: $P(h_t \mid s_t, \pi) = 1$ for the unique matching history, and 0 otherwise. Consequently, $Q^{\pi,\mu}(s_t, a_t) = Q^{\pi,\mu}(h_t, a_t)$. The Q-value is perfectly well-defined, explaining why global trajectory evaluation functions flawlessly in append-only environments.

Case 2: Terminal Preferences in Loopy MDPs. Suppose the environment is a standard, loopy MDP where state s_t can be reached via thousands of different histories. If the preference model evaluates the *entire* trajectory, $P(h_t \mid s_t, \pi)$ is highly complex and shifts constantly as π_θ updates, making $Q^{\pi,\mu}(s_t, a_t)$ generally intractable.

However, if human preferences evaluate *only the terminal states*—meaning $M(\tau, \tau') = M(s_T, s'_T)$—we can mathematically demonstrate that the dependence on the past history h_t vanishes:

$$
\begin{aligned}
Q^{\pi,\mu}(h_t, a_t) &= \mathbb{E}_{\tau_{>t}\sim\pi,\tau'\sim\mu}\Big[M(\tau,\tau') \;\Big|\; h_t, a_t\Big] && \text{(Definition)}\\
&= \mathbb{E}_{\tau_{>t}\sim\pi,\tau'\sim\mu}\Big[M(s_T,s'_T) \;\Big|\; h_t, a_t\Big] && \text{(Terminal preference assumption)}\\
&= \mathbb{E}_{s_T\sim P(\cdot|h_t,a_t,\pi)}\left[\mathbb{E}_{s'_T\sim P(\cdot|s_0,\mu)}\Big[M(s_T,s'_T)\Big]\right] && \text{(Independence of opponent rollout)}\\
&= \mathbb{E}_{s_T\sim P(\cdot|s_t,a_t,\pi)}\left[\mathbb{E}_{s'_T\sim P(\cdot|s_0,\mu)}\Big[M(s_T,s'_T)\Big]\right] && \text{(Markov property of } P \text{ and } \pi)\\
&\equiv Q^{\pi,\mu}(s_t, a_t) && \text{(History independence)}
\end{aligned}
$$

In this specific regime, the posterior reach probabilities are no longer required inside the Q-function. Instead, the forward reach probabilities are implicitly handled by the outer expectation of the policy gradient in Equation 16, analytically expressed via the discounted state visitation frequency $d^\pi(s)$. Because on-policy Monte Carlo rollouts naturally sample states according to $d^\pi(s)$, the global evaluation algorithm succeeds without requiring tree-structured states.

Case 3: Synchronous Intermediate Contexts in Loopy MDPs (Our Framework). To contrast with the global evaluation constraints, we formally derive how the Context-Ordinal MDP preserves the Markov property even when providing dense feedback in loopy environments. While global frameworks evaluate the full historical trajectory, our framework assumes the preference function evaluates intermediate contexts: $\rho(\mathbf{x}_t)$.

Consider an agent arriving at state s_t via some arbitrary history h_{t-1}. From this state, we evaluate the synchronous local context $\mathbf{x}_t = \big((s_t, a_1), \ldots, (s_t, a_{|\mathcal{A}|})\big)$. We calculate the history-dependent expected voting statistic, $Q^\pi(h_{t-1}, \mathbf{x}_t)$, which aggregates the intermediate preference evaluations $\rho(\mathbf{x}_{t+k})$ over all future time steps $k \geq 0$:

$$
\begin{aligned}
Q^\pi(h_{t-1}, \mathbf{x}_t) &= \mathbb{E}_{\mathbf{x}_{>t}\sim P,\pi}\left[\sum_{k=0}^{\infty}(1-\gamma)\gamma^k \rho(\mathbf{x}_{t+k}) \;\middle|\; h_{t-1}, \mathbf{x}_t\right] && \text{(Definition)}\\
&= \sum_{k=0}^{\infty}(1-\gamma)\gamma^k\, \mathbb{E}_{\mathbf{x}_{t+k}}\Big[\rho(\mathbf{x}_{t+k}) \;\Big|\; h_{t-1}, \mathbf{x}_t\Big] && \text{(Linearity of expectation)}\\
&= \sum_{k=0}^{\infty}(1-\gamma)\gamma^k\, \mathbb{E}_{\mathbf{x}_{t+k}}\Big[\rho(\mathbf{x}_{t+k}) \;\Big|\; \mathbf{x}_t\Big] && \text{(Markov property of } P \text{ and } \pi) \qquad (19)\\
&\equiv Q^\pi(\mathbf{x}_t) && \text{(History independence)}
\end{aligned}
$$

The Mathematical Mechanism: The critical step occurs in (19). Because the preference model $\rho(\mathbf{x}_{t+k})$ only evaluates the intermediate state-action tuples, it requires no knowledge of the past. Furthermore, because the branches are perfectly synchronous, their future trajectories evolve via independent Markovian transitions. Consequently, the distribution of the future joint context $\mathbf{x}_{t+k}$ is entirely conditionally independent of the arrival history h_{t-1} given the current context $\mathbf{x}_t$. The unobservable posterior reach probabilities $P(h_{t-1} \mid s_t, \pi)$ that plague global evaluation mathematically factor out, allowing the Context-Ordinal framework to natively execute dense, intermediate temporal-difference bootstrapping in highly loopy environments.

D Generalizing the Context Distribution: Cross-Temporal Comparisons

In the main text, we defined a context $\mathbf{s}_t$ by pairing states that occur at the exact same time step t across parallel rollouts. Implicitly, we assumed that contexts are drawn from a perfectly correlated joint distribution over time: we only ever evaluate the tuple $(s_{t_1}^{a_1}, s_{t_2}^{a_2}, \dots)$ when $t_1 = t_2 = \cdots = t$. This synchronous temporal structure is what enables the expected voting statistics to decompose additively across time, unlocking efficient temporal-difference (TD) bootstrapping.

However, the Context-Ordinal framework can be generalized by relaxing this correlated joint distribution into a more uniform distribution over all possible time step pairings.

Implicit Discounting. We can bypass explicit mathematical discounting (γ) entirely if the preference model is capable of cross-temporal comparisons. Instead of strictly pairing states at the same time step t, we can construct context tuples by drawing states from disparate time steps, querying rankings over combinations like $(s_{t_1}^{a_1}, s_{t_2}^{a_2})$ for all $t_1, t_2 \geq 0$.

This natively handles scenarios where the human evaluator explicitly discounts the future intrinsically. For example, a human evaluator might strictly prefer a slightly worse outcome today over a substantially better outcome a decade from now. Under cross-temporal evaluation, the preference function ρ internalizes the time preference, rendering an external MDP discount factor γ redundant.

The Trade-off: Efficiency vs. Generality. If we sample contexts from the fully uniform "all-pairs" joint distribution, we gain maximum generality. The voting rule evaluates whether an action's entire resultant timeline is holistically preferred to another's, explicitly weighing near-term states in one timeline directly against long-term states in another.

However, this comes at a steep computational cost. By breaking the synchronous temporal structure, the aggregated voting statistics no longer cleanly factorize into the recursive Bellman expectation equation $Q \approx v + \gamma \mathbb{E}[Q]$. The value of an action at $t = 0$ becomes dependent on the combinatorial cross-evaluation of all future states against all other future states. In this generalized setting, we trade the efficiency of Bellman bootstrapping for ultimate preference generality. We leave the conceptualization of alternative, efficient dynamic programming avenues for cross-temporal ordinal RL to future work.

References

Kenneth J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58 (4):328–346, 1950. DOI: 10.1086/256963.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Dimitri Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, 2022.

Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

Peter C Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

Ian Gemp, Crystal Qian, Marc Lanctot, and Kate Larson. Nash without numbers: A social choice approach to mixed equilibria in context-ordinal games, 2026. URL https://arxiv.org/abs/2605.07996.

Shizuo Kakutani. A generalization of Brouwers fixed point theorem. *Duke Mathematical Journal*, 8(3):457, 1941.

Chun Kai Ling, Fei Fang, and J Zico Kolter. What game are we playing? end-to-end learning in normal and extensive form games. *arXiv preprint arXiv:1805.02777*, 2018.

Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! Alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

Kenneth O May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica: Journal of the Econometric Society*, pp. 680–684, 1952.

Richard D McKelvey. Covering, dominance, and institution-free properties of social choice. *American Journal of Political Science*, pp. 283–314, 1986.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Walter Mischel, Ebbe B Ebbesen, and Antonette Raskoff Zeiss. Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, 21(2):204, 1972.

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

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Chengyuan Peng and John Stachurski. Dynamic programming in ordered vector space. *Operations Research*, 2026.

Martin L Puterman. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Thomas J Sargent and John Stachurski. Dynamic programs on partially ordered sets. *SIAM Journal on Control and Optimization*, 63(2):778–795, 2025.

Richard P Savage Jr. The paradox of nontransitive dice. *The American Mathematical Monthly*, 101 (5):429–436, 1994.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2024.

Yuichi Shoda, Walter Mischel, and Philip K Peake. Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental psychology*, 26(6):978, 1990.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

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

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.