RESOLVING NON-CONVERGENCE IN DISTRIBUTIONAL CONTROL: A HOMOTOPY APPROACH WITH DOUBLE DISTRIBUTIONAL VALUE ITERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributional reinforcement learning models the full probability distribution of cumulative returns rather than just their expected values. In control settings, standard algorithms select actions using a greedy policy based on the mean returns of the current distributional estimates. However, when multiple actions have the same optimal expected return but different return distributions, small estimation errors in their means cause the greedy policy to abruptly jump back and forth between these actions. Because the policy keeps switching, the target return distributions for Bellman updates oscillate continuously, preventing the estimated return distributions from converging even after value errors become negligible. To resolve this chattering issue, we provide an algorithm,at each step, an auxiliary update estimates action values and turns them into a softmax reference policy, and a main update learns the return distribution by distributional Bellman evaluation under that policy. The softmax is evaluated at a positive temperature λ_t. Cooling the temperature to zero throughout training brings the policy back to the optimal actions and, on a tie, shares its mass uniformly over them, while the objective and the return target stay unchanged. Theoretically, for tabular MDPs, we prove for the planning version of our algorithm that the softmax policy converges to the uniform distribution over optimal actions and that the return distributions converge to its law in maximal 1-Wasserstein distance. For its stochastic approximation version, we prove almost-sure convergence to the same policy and return law. We provide a practical implementation using categorical networks. Experiments in tabular environments confirm that algorithm eliminates distributional oscillation and guarantees convergence, while continuous-control benchmarks in MuJoCo demonstrate improved distributional stability during training.

1 INTRODUCTION

Reinforcement learning typically summarizes the future consequences of a state–action pair by the expected discounted return (Sutton & Barto, 2018; Watkins & Dayan, 1992). Distributional reinforcement learning instead models the law of the random return (Bellemare et al., 2017; 2023). This representation captures variability and multimodality that are invisible to an expectation-only value function, and extends algorithms such as C51 (Bellemare et al., 2017), quantile-regression DQN (Dabney et al., 2018b), and implicit quantile networks (Dabney et al., 2018a).

Q-learning and distributional methods such as C51 and quantile-regression DQN select the action in the Bellman target greedily, by maximizing the estimated action value. In the distributional case this value is the mean of the estimated return distribution, and the greedy action is the bootstrap target of the distributional Bellman update (Bellemare et al., 2017; Dabney et al., 2018b; Bellemare et al., 2023). Suppose, however, that several actions share the same optimal expected return—each attains the optimal action value $Q^*(s, \cdot)$, the largest mean return achievable from that state—while their return distributions differ. The gaps among the estimated means are then only value-estimation errors, the discrepancies between those means and Q^*. Because this greedy rule is discontinuous in the means, arbitrarily small errors make the selected action jump back and forth among these actions. The estimated return distributions therefore need not converge, even after the value errors themselves have become negligible (Bellemare et al., 2017; 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Entropy-regularized methods replace greedy selection by a regularization term, and the resulting policy is a softmax at a fixed positive temperature. Because that temperature enters the objective, the return they learn is biased away from the original problem (Haarnoja et al., 2017; 2018; Neu et al., 2017; Geist et al., 2019; Ma et al., 2025). We propose double distributional value iteration (DDVI), which learns the reference policy by a vanishing-temperature homotopy. This path removes the non-convergence of the distributional backup and introduces no such bias: the temperature indexes only the policy, while the expected-return objective stays unchanged. The auxiliary iterate $\{\widetilde{Z}_t\}$ follows a greedy distributional backup and is used only to estimate action values, $\widetilde{Q}_t(s,a) = \mathbb{E}[\widetilde{Z}_t(s,a)]$. From these means the homotopy builds the reference policy on the finite action set $\mathcal{A}$,

$$\pi_t(a \mid s) = \frac{\exp(\widetilde{Q}_t(s,a)/\lambda_t)}{\sum_{b\in\mathcal{A}} \exp(\widetilde{Q}_t(s,b)/\lambda_t)},$$

where the temperature $\lambda_t > 0$ decreases to zero. For each fixed λ_t the map from means to π_t is continuous, so a small value-estimation error shifts probabilities slightly and no longer swaps the Bellman target. Our homotopy approach is motivated by this softmax,the Gibbs measure that converges,as the temperature vanishes,to the uniform distribution on the maximizers of its argument (Hwang, 1980);a cooling schedule makes that limit algorithmic and tracks a smooth path back to greedy selection (Geman & Geman, 1984; Hajek, 1988; Rose, 1998; Allgower & Georg, 1990). The main iterate $\{Z_t\}$ is updated by distributional Bellman evaluation under this policy, $Z_{t+1} = \mathcal{T}^{\pi_t} Z_t$. Here $\mathcal{T}^{\pi}$ pushes a return-distribution estimate forward by one reward and a discounted next-state transition drawn from π. The expected-return criterion and this unregularized target stay unchanged; only the policy that selects the target is smoothed and then cooled.

For the planning version of DDVI in tabular MDPs, if $\lambda_t \to 0$ and $\gamma^t/\lambda_t \to 0$, the reference policy converges to π^*, the uniform policy over optimal actions, and the main return distributions converge to its return law in maximal 1-Wasserstein distance. The stochastic approximation version achieves the same limits almost surely under suitable cooling, exploration, and stepsize conditions.

The proof uses a common argument: value errors that vanish faster than the temperature equalize the softmax weights on tied optimal actions and suppress suboptimal actions. The resulting policy error gives a vanishing perturbation of evaluation under π^*, whose Wasserstein contraction yields distributional convergence. For the stochastic approximation version, martingale-noise bounds and the Robbins–Monro stepsizes additionally control the sampling noise and yield almost-sure convergence.

Our contributions are as follows:

1. To address non-convergence in distributional control, we propose DDVI, a homotopy algorithm that replaces greedy selection in the main Bellman target with a vanishing-temperature softmax policy. An auxiliary return-distribution estimator supplies the action-value estimates used by the softmax policy, while the main estimator learns return distributions under that policy. We also provide a practical implementation (Algorithm 2, Appendix D).
2. For tabular MDPs, we prove convergence of the planning version and almost-sure convergence of the stochastic approximation version under the stated cooling and sampling conditions. Both versions select the uniform policy over optimal actions and converge to its return law in maximal 1-Wasserstein distance, without requiring a unique optimal policy.
3. Tabular experiments verify convergence to the selected policy and return law while preserving optimal expected returns. MuJoCo experiments show improved diagnostics of distributional stability, with return gains varying across tasks.

2 Related Work

Distributional RL. Distributional reinforcement learning represents the random return by its full distribution rather than by its mean (Bellemare et al., 2017; 2023). For a policy, the distributional Bellman operator—the backup that pushes the return distribution forward by one reward and a discounted transition under that policy—is a contraction, so evaluation converges to the return distribution of the policy being evaluated (Bellemare et al., 2017; Rowland et al., 2018). Control does not inherit this guarantee. C51, quantile-regression DQN, and implicit quantile networks convert the estimated distributions into a greedy policy by comparing their means (Bellemare et al., 2017; Dabney

et al., 2018b;a). When several actions share the optimal mean but have different return distributions, that greedy choice can switch under small errors in the means. The distributional optimality operator is then not a contraction, and the iterates need not converge to the return distribution of any one optimal policy (Bellemare et al., 2017; 2023).

Vanishing-temperature Homotopy Vanishing-temperature homotopy is a path of softmax policies, indexed by a temperature $\lambda > 0$, that becomes greedy as λ goes to zero. On a finite set, the softmax with scores q puts mass on action a proportional to $\exp(q(a)/\lambda)$. As $\lambda \to 0$ it concentrates on the maximizers of q, and when several maximizers tie the limit is uniform on that set (Hwang, 1980). This is the canonical choice we need among actions that share an optimal mean. Cooling schedules turn the same limit into a procedure, lowering the temperature while the estimates are still moving (Geman & Geman, 1984; Hajek, 1988). Deterministic annealing and numerical continuation describe the same construction as tracking a smooth path until it reaches a non-smooth selection rule (Rose, 1998; Allgower & Georg, 1990). We use the path only to form the reference policy inside an otherwise unregularized iteration. The temperature rescales the estimated means; it does not enter the reward or the return distribution being learned.

Entropy-regularized control. The policy is rewarded for expected return and for its own randomness, and at a fixed positive temperature the optimal policy is a softmax. The regularized Bellman operator is then a contraction, which is why soft Q-learning and soft actor-critic are well posed (Haarnoja et al., 2017; 2018; Neu et al., 2017; Geist et al., 2019).

The closest work to ours is Jhaveri et al. (2025), who study policy and return-distribution convergence through vanishing entropy regularization and decoupled temperatures. With a uniform reference policy and discrete actions, their construction also selects the uniform policy over optimal actions. For each prescribed accuracy, their Theorem 4.7 establishes the existence of suitable fixed temperatures and separate numbers of control and evaluation steps that approximate the limiting policy's unregularized return law. This is an accuracy-dependent, two-stage approximation guarantee. Although they discuss progressive annealing, they do not analyze a single coupled recursion with a prescribed cooling schedule. DDVI provides such a recursion: it updates the auxiliary and main estimators together while cooling the policy temperature, using unregularized Bellman backups throughout. We prove convergence of this sequence as $t \to \infty$ for the planning version and almost-sure convergence for the stochastic approximation version under the stated conditions.

3 Algorithm

3.1 Preliminaries

3.1.1 Markov Decision Processes

A tabular Markov decision process (MDP) is a 5-tuple $M = \langle \mathcal{S}, \mathcal{A}, P_R, P, \gamma \rangle$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ a finite action space, $P_R : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ the reward distribution, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the transition kernel, and $\gamma \in (0,1)$ a discount factor. Here $\Delta(\cdot)$ denotes the set of probability distributions over some set.

The return is $G^\pi(s_0, a_0) := \sum_{t=0}^{\infty} \gamma^t R_t$, and $G^\pi(s,a) \in [0, 1/(1-\gamma)]$. The state-action value is $Q^\pi(s,a) := \mathbb{E}[G^\pi(s,a)]$. We write $P_{s,a} := P(\cdot \mid s,a)$, $P^R_{s,a} := P_R(\cdot \mid s,a)$, $\pi_s := \pi(\cdot \mid s)$, and $Q^\pi_s := Q^\pi(s,\cdot)$. The Bellman equation is $Q^\pi(s,a) = \mathbb{E}_{S' \sim P_{s,a},\, A' \sim \pi_{S'},\, R \sim P^R_{s,a}}\big[R + \gamma Q^\pi(S', A')\big], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$, or $Q^\pi = T^\pi Q^\pi$, where $T^\pi : [0, 1/(1-\gamma)]^{\mathcal{S}\times\mathcal{A}} \to [0, 1/(1-\gamma)]^{\mathcal{S}\times\mathcal{A}}$ is the scalar Bellman operator.

The set of optimal policies is $\Pi^* := \arg\max_\pi Q^\pi(s_0, a_0)$ for an arbitrary initial pair (s_0, a_0); the maximizer is independent of this choice. For any $\pi^* \in \Pi^*$, $Q^{\pi^*}(s,a) = \max_\pi Q^\pi(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. The optimal value $Q^*(s,a) := \max_\pi Q^\pi(s,a)$ satisfies the Bellman optimality equation $Q^*(s,a) = \mathbb{E}_{S' \sim P_{s,a},\, A' \sim \mathcal{G}(Q^*)_{S'},\, R \sim P^R_{s,a}}\big[R + \gamma Q^*(S', A')\big]$, or $Q^* = T^{\mathcal{G}} Q^*$. Here $\mathcal{G}(Q)$ is a greedy selection rule, i.e. $\mathcal{G}(Q)(a \mid s) > 0 \implies Q(s,a) = \max_{a'} Q(s,a')$. The operator $T^{\mathcal{G}} : [0, 1/(1-\gamma)]^{\mathcal{S}\times\mathcal{A}} \to [0, 1/(1-\gamma)]^{\mathcal{S}\times\mathcal{A}}$ is the Bellman optimality operator.

In distributional RL one tracks return distributions $\eta^\pi(s,a) := \mathrm{Law}(G^\pi(s,a))$. They satisfy the distributional Bellman equation $\eta^\pi(s,a) = \mathbb{E}_{S'\sim P(\cdot|s,a),\, A'\sim\pi_{S'},\, R\sim P^R_{s,a}}\big[(b_{R,\gamma})_\#(\eta^\pi(S',A'))\big]$, where $(b_{x,y})_\#$ is the push-forward of the affine map $z \mapsto x + yz$, i.e. $(b_{x,y})_\#(\mathrm{Law}(X)) = \mathrm{Law}(x + yX)$. Different optimal policies $\pi^*_1, \pi^*_2 \in \Pi^*$ may induce different return distributions. Nevertheless, for any $\pi^* \in \Pi^*$, $\eta^{\pi^*}(s,a) = \mathbb{E}_{S'\sim P_{s,a},\, A'\sim\mathcal{G}(\eta^{\pi^*})_{S'},\, R\sim P^R_{s,a}}\big[(b_{R,\gamma})_\#(\eta^{\pi^*}(S',A'))\big]$, or $\eta^{\pi^*} = \mathcal{T}^{\mathcal{G}}\eta^{\pi^*}$. We write $\mathcal{G}(\eta) := \mathcal{G}(Q_\eta)$, where $Q_\eta(s,a)$ is the mean of $\eta(s,a)$. The operator $\mathcal{T}^{\mathcal{G}} : \Delta\big([0,1/(1-\gamma)]\big)^{\mathcal{S}\times\mathcal{A}} \to \Delta\big([0,1/(1-\gamma)]\big)^{\mathcal{S}\times\mathcal{A}}$ is the distributional Bellman optimality operator. We write T^π and $T^{\mathcal{G}}$ for the scalar Bellman operators acting on action values, and $\mathcal{T}^\pi$ and $\mathcal{T}^{\mathcal{G}}$ for the distributional Bellman operators acting on return laws.

3.1.2 Value Iteration and Distributional Value Iteration

Control aims at finding some $\pi^* \in \Pi^*$. Since $\mathcal{G}(Q^*) \in \Pi^*$ for any greedy rule $\mathcal{G}$, it suffices to compute Q^*. Value iteration starts from $Q^{(0)} \in [0,1/(1-\gamma)]^{\mathcal{S}\times\mathcal{A}}$ and sets $Q^{(t+1)} = T^{\mathcal{G}}Q^{(t)}, \qquad t = 0,1,\dots$. The operator $T^{\mathcal{G}}$ is a γ-contraction in $\|\cdot\|_\infty$, so Banach's fixed-point theorem yields $Q^{(t)} \to Q^*$ geometrically: $\|Q^{(t)} - Q^*\|_\infty = O(\gamma^t)$.

Distributional value iteration uses return distributions: given $\eta^{(0)}$ and $\mathcal{G}$, $\eta^{(t+1)} = \mathcal{T}^{\mathcal{G}}\eta^{(t)}, \qquad t = 0,1,\dots$. The scheme is mean-preserving: taking expectations recovers ordinary value iteration.

One might expect $\overline{W}_1(\eta^{(t)}, \eta^{\pi^*}) \to 0$ for some $\pi^* \in \Pi^*$, where $\overline{W}_1(\eta_1,\eta_2) := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_1(\eta_1(s,a),\eta_2(s,a))$ is the supremum 1-Wasserstein distance. This need not hold without extra assumptions (e.g. uniqueness of the optimal policy). Unlike $T^{\mathcal{G}}$ on values, the distributional operator $\mathcal{T}^{\mathcal{G}}$ is not a contraction in $\overline{W}_1$; see Chapter 7 of Bellemare, Dabney, and Rowland Bellemare et al. (2023).

3.2 Double Distributional Value Iteration

We propose double distributional value iteration (DDVI) to mitigate the instability and possible non-convergence of standard distributional value iteration. The method runs two coupled return-distribution iterates. An auxiliary iterate produces a sequence of reference policies; the main iterate replaces greedy selection by these reference policies. We first present the planning version of DDVI, which uses exact model-based backups, and then its stochastic approximation version, which uses sampled transitions. We also provide a practical implementation.

Planning version. When the transition and reward kernels are known, DDVI reduces to an exact coupled backup. At step t, the auxiliary return law is first updated by the distributional optimality operator. Its mean $Q^{(t+1)}(s,a) = \mathbb{E}_{Z\sim\nu^{(t+1)}(s,a)}[Z]$ then defines a softmax reference policy at temperature $\lambda_t > 0$, and the main return law is evaluated under that policy:

$$\nu^{(t+1)} = \mathcal{T}^{\mathcal{G}}\nu^{(t)}, \tag{1}$$

$$\pi^{(t+1)}(a \mid s) = \frac{\exp\big(Q^{(t+1)}(s,a)/\lambda_t\big)}{\sum_{a'\in\mathcal{A}}\exp\big(Q^{(t+1)}(s,a')/\lambda_t\big)}, \tag{2}$$

$$\eta^{(t+1)} = \mathcal{T}^{\pi^{(t+1)}}\eta^{(t)}. \tag{3}$$

The temperature is annealed to zero, $\lambda_t \to 0$. For any fixed $\lambda_t > 0$, the map from means to $\pi^{(t+1)}$ is continuous, so a small error in $Q^{(t+1)}$ only perturbs the action probabilities and no longer replaces the Bellman target. In the limit, $\pi^{(t+1)}$ concentrates on the maximizers of $Q^{(t+1)}$ and, when several actions attain the same maximum, places equal mass on all of them. The intended limit is therefore the uniform optimal policy π^* and the return law η^{π^*} it induces.

Stochastic approximation version. When the model is unavailable, the stochastic approximation version of DDVI uses sampled Bellman targets to update a visited pair (S_t, A_t). This is the recursion analyzed in Section 4.2. Stepsizes satisfy $\alpha_t, \beta_t \in [0,1]$. Unvisited coordinates stay unchanged, and the two branches may share the transition (R_t, S'_t). Algorithm 1 records the input, output, and update of each step.

Algorithm 1 Stochastic approximation version of DDVI

Input: Stepsizes α_t, β_t, temperature $\lambda_t \to 0$, laws ν_t, η_t, and visited pair (S_t, A_t).
Output: Updated laws ν_{t+1}, η_{t+1} and reference policy $\pi^{(t+1)}$.
Sample $(R_t, S'_t) \sim P^R_{S_t,A_t} \otimes P_{S_t,A_t}$ and $g_t \sim \mathcal{G}(\nu_t)(\cdot \mid S'_t)$.
Update the auxiliary law at (S_t, A_t):

$$\nu_{t+1}(S_t, A_t) = (1 - \beta_t)\nu_t(S_t, A_t) + \beta_t(b_{R_t,\gamma})_{\#}\nu_t(S'_t, g_t). \tag{4}$$

Compute the reference policy:

$$q_{t+1}(s, a) = \mathbb{E}_{Z\sim\nu_{t+1}(s,a)}[Z], \qquad \pi^{(t+1)} = \mathcal{G}_{\lambda_t}(\nu_{t+1}). \tag{5}$$

Sample $B_t \sim \pi^{(t+1)}(\cdot \mid S'_t)$ and update the main law at (S_t, A_t):

$$\eta_{t+1}(S_t, A_t) = (1 - \alpha_t)\eta_t(S_t, A_t) + \alpha_t(b_{R_t,\gamma})_{\#}\eta_t(S'_t, B_t). \tag{6}$$

Leave all other coordinates of ν_t and η_t unchanged.

Taking means in equation 4 recovers ordinary tabular Q-learning, so the auxiliary means can approach Q^* whether or not several actions are optimal. The cooled softmax in Step 4, rather than a greedy rule on those noisy means, is what selects among optimal actions.

Practical implementation. We implement DDVI with two categorical neural networks, replay-buffer sampling, and gradient updates. The auxiliary network forms the cooled reference policy, and the main network learns return distributions under that policy. The full training procedure is given in Algorithm 2 in Appendix D; this implementation is used in the deep RL experiments.

4 Theoretical Analysis

We analyze the two tabular versions of DDVI separately. Section 4.1 establishes convergence of the planning version with exact Bellman backups. Section 4.2 establishes almost-sure convergence of the stochastic approximation version in Algorithm 1.

4.1 Analysis of the planning version

Now we formally state the convergence guarantee of our proposed algorithm.

Theorem 1. *Suppose $\gamma^t/\lambda_t \to 0$, then for any initialized return distribution $\eta^{(0)}$, we have*

$$\lim_{t\to\infty} \bar{W}_1(\eta^{(t)}, \eta^{\pi^*}) = 0. \tag{7}$$

Here $\pi^* \in \Pi^*$ is the optimal policy satisfying:

$$\pi^*(a \mid s) = \begin{cases} 1/|\arg\max_{a\in A} Q^*(s, a)| & a \in \arg\max_{a\in A} Q^*(s, a) \\ 0 & \text{else} \end{cases}, \quad \forall s \in S. \tag{8}$$

For a signed Borel measure μ on $[0, (1-\gamma)^{-1}]$ and a function class $\mathcal{F}$ on this interval, define $\|\mu\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mu f|$. Let $M_{\mathcal{F}}$ denote the vector space of such measures with finite $\|\cdot\|_{\mathcal{F}}$ norm. The product $\left(M_{\mathcal{F}_{W_1}}\right)^{S\times A}$ also constitutes a normed vector space, where the infinity norm is defined as $\|\cdot\|_{\mathcal{F}_{W_1},\infty} := \sup_{(s,a)\in S\times A}\|\cdot\|_{\mathcal{F}}$. $\mathcal{T}^\pi$ can be naturally extended to $(M_{\mathcal{F}})^{S\times A}$ without modifying its original definition.

Recall that the W_1 metric is an integral probability metric. If $\mu \in M_{\mathcal{F}}$ can be written as the difference of two probability distributions $\mu_+, \mu_- \in \Delta\left([0, \frac{1}{1-\gamma}]\right)$, we have $\|\mu\|_{\mathcal{F}_{W_1}} = W_1(\mu_+, \mu_-)$, where $\mathcal{F}_{W_1} := \{f \mid f \text{ is 1-Lipschitz}\}$. If the vector of distribution $\eta \in (M_{\mathcal{F}})^{S\times A}$ can be written as the difference of two vectors of probability distributions η_+, η_-, then we have $\|\eta(s, a)\|_{\mathcal{F}_{W_1},\infty} = \bar{W}_1(\eta_+, \eta_-)$.

4.2 Analysis of the stochastic approximation version

We analyze the stochastic approximation version of DDVI in Algorithm 1, written coordinate-wise as equation 4–equation 6. All unvisited coordinates are left unchanged. The two branches may share the transition (R_t, S'_t).

Assumption 1 (Correct sampling and adaptability)**.** *Let $\mathcal{F}_t$ denote the information available before the environment sample at time t, which contains $S_t, A_t, \nu^{(t)}, \eta^{(t)}, \lambda_t, \alpha_t, \beta_t$. All of these quantities are $\mathcal{F}_t$-measurable. Conditional on $\mathcal{F}_t$, and according to the reward–transition model of the paper, $(R_t, S_{t+1}) \sim P^R_{S_t,A_t} \otimes P_{S_t,A_t}$. The auxiliary target action satisfies*

$$\mathbb{P}\big(A^{\nu}_{t+1} = a' \mid \mathcal{F}_t, R_t, S_{t+1}\big) = G(\nu^{(t)})(a' \mid S_{t+1}), \tag{9}$$

and the main-branch target action satisfies

$$\mathbb{P}\big(A^{\eta}_{t+1} = a' \mid \mathcal{F}_t, R_t, S_{t+1}\big) = \pi^{(t+1)}(a' \mid S_{t+1}). \tag{10}$$

In particular,

$$\mathbb{E}\big[\widehat{\eta}^{(t+1)} \mid \mathcal{F}_t\big] = \big(\mathcal{T}^{\pi^{(t+1)}}\eta^{(t)}\big)(S_t, A_t). \tag{11}$$

All random measures and related maps satisfy the usual measurability conditions.

Assumption 2 (Stepsizes)**.** *Stepsizes vanish on unvisited coordinates. Almost surely, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, $0 \le \alpha_t(s,a) \le 1, \sum_{t=0}^{\infty} \alpha_t(s,a) = \infty, \sum_{t=0}^{\infty} \alpha_t(s,a)^2 < \infty, 0 \le \beta_t(s,a) \le 1, \sum_{t=0}^{\infty} \beta_t(s,a) = \infty, \sum_{t=0}^{\infty} \beta_t(s,a)^2 < \infty$.*

Assumptions 2 in particular require that every state–action pair receives infinitely many effective updates.

Taking means in equation 4 and using $m((b_{r,\gamma})_{\#}\mu) = r + \gamma m(\mu)$ yields the ordinary tabular Q-learning recursion

$$q_{t+1} = (1 - \beta_t)\, q_t + \beta_t\big(R_t + \gamma \max_b q_t(S'_t, b)\big). \tag{12}$$

Thus the auxiliary means track Q^*, while the main distributions track the return law of the reference policy. The intended limits are the uniform optimal policy π^* defined in equation 8 and its return law η^{π^*}.

We impose the following temperature condition and then state the convergence guarantee.

Assumption 3 (Error–temperature matching)**.** *Almost surely,*

$$\lambda_t > 0, \qquad \lambda_t \to 0, \qquad \frac{\big\|Q^{(t)} - Q^*\big\|_\infty}{\lambda_t} \to 0. \tag{13}$$

Theorem 2. *Under Assumptions 1, 2, and 3, for any initialization $\eta^{(0)}$,*

$$\lim_{t\to\infty} \overline{W}_1\big(\eta^{(t)}, \eta^{\pi^*}\big) = 0 \qquad \textit{almost surely}, \tag{14}$$

where $\pi^* \in \Pi^*$ is defined in equation 8.

Corollary 1 (Temperature matching under tabular Q-learning)**.** *Suppose the auxiliary means $\{q_t\}$ follow tabular Q-learning under the assumptions of Qu & Wierman (2020, Theorem 7): bounded rewards, zero initialization, and a fixed behavior policy inducing an ergodic state–action chain. The visited-coordinate stepsizes are $\beta_t = h/(t + t_0)$, with h and t_0 satisfying that theorem's requirements. Write $e_t = \|q_t - Q^*\|_\infty$. That result yields, after absorbing constants independent of t and δ,*

$$\mathbb{P}\left(e_t > C_1\sqrt{\frac{\log(C_2 t^2/\delta)}{t + t_0}} + \frac{C_3}{t + t_0}\right) \le \delta. \tag{15}$$

The bound controls only the ℓ_∞ error of the action values: it does not require uniqueness of an optimal policy, nor convergence of the greedy policy induced by q_t. Choosing $\delta_t = (t+1)^{-2}$ and applying the Borel–Cantelli lemma gives $e_t = O(\sqrt{\log t/t})$ almost surely. Thus any schedule $\lambda_t = (t+1)^{-p}$ with $0 < p < \frac{1}{2}$ satisfies $\lambda_t \to 0$ and $e_t/\lambda_t \to 0$ almost surely.

5 EXPERIMENTS

5.1 TABULAR CONVERGENCE EXPERIMENTS

We compare our algorithm with tabular Q-learning and C51 with deterministic greedy tie-breaking on finite MDPs. All methods use synchronous, model-based updates with exact transition and reward distributions for 500 iterations, isolating convergence behavior from sampling variability. C51 and our algorithm use 121 uniformly spaced categorical atoms, and our algorithm uses $\lambda_t = 0.8/t$. We measure start-state expected return, policy distance d_π to the tie-uniform optimal policy, and maximal Wasserstein-1 distance $\bar{W}_1$ to its return distribution. For Q-learning, the distributional diagnostic uses a Dirac surrogate centered at its action-value estimate and thus evaluates a mean-only representation.

Table 1 reports three additional environments, where all methods achieve the same expected returns. On `looping_reward`, C51 and our algorithm have negligible distributional errors. On `gridworld_4x4`, Q-learning and C51 retain a policy distance of 1.0, while our algorithm reduces it to 3.469×10^{-14} and 7.074×10^{-10}, respectively. C51 nevertheless has zero reported distributional error in both environments, illustrating that distinct optimal policies can share the same return distributions. Overall, our algorithm preserves expected-return performance while approaching the tie-uniform policy; policy and distributional metrics reveal distinctions that expected return alone misses.

Table 1: Final tabular results after 500 iterations. Errors are measured against the tie-uniform optimal policy and its return distribution; Q-learning uses a Dirac surrogate.

Environment	Algorithm	Return ↑	$d_\pi \downarrow$	$\bar{W}_1 \downarrow$
looping_reward	Q-learning	0.675	0	5.84×10^{-1}
	C51	0.675	0	5.23×10^{-12}
	our algorithm	0.675	1.12×10^{-18}	5.23×10^{-12}
gridworld_4x4	Q-learning	0.696	1.000	1.79×10^{-1}
	C51	0.696	1.000	0
	our algorithm	0.696	7.07×10^{-10}	2.13×10^{-11}

We further ablate the temperature on a 4×4 gridworld in which east and south are both optimal at the start but induce different return laws; the suboptimality gap is $\Delta \approx 0.045$. The backup, atom count, and horizon match the runs above, and only λ_t changes. Greedy tie-breaking already attains the optimal start return 0.9025, yet finishes at $d_\pi = 1$ and $\bar{W}_1 = 0.0428$. The main-text schedule $\lambda_t = 0.8/t$ keeps this return and reduces the two errors to 6.30×10^{-11} and 1.53×10^{-12}. Cooling as $\lambda_t = 2^{-t}$ reaches numerical zero on both. Schedules that remain large relative to Δ do not: $\lambda_t = 1$, $(t+1)^{-1/4}$, and $(t+1)^{-1/2}$ finish at start returns 0.403, 0.501, and 0.766. The environment, learning curves, and full table are in Appendix D.5.

5.2 DEEP EXPERIMENTS

5.2.1 INTRODUCTION

We further evaluate the proposed double distributional mechanism in deep continuous-control domains. The goal of this experiment is not only to compare episodic return, but also to test the main qualitative prediction of Theorem 2: when the policy used by the distributional Bellman update is produced by a convergent auxiliary process, the resulting return-distribution dynamics should be less prone to policy oscillation than a purely greedy distributional update. We instantiate this idea in MuJoCo-v5 tasks by comparing a C51-style distributional baseline with a Double DRL variant, denoted Double DRL (π from ν), in which an auxiliary distributional critic ν generates the reference policy used in the main distributional update. All curves are averaged over three random seeds and trained for 5×10^5 environment steps. For each environment, we report three diagnostics: the mean episodic return, a proxy for the policy discrepancy $\|\pi - \pi^\star\|_1$, and a proxy for the supremum Wasserstein distance $\bar{W}_1$ between return-distribution estimates.

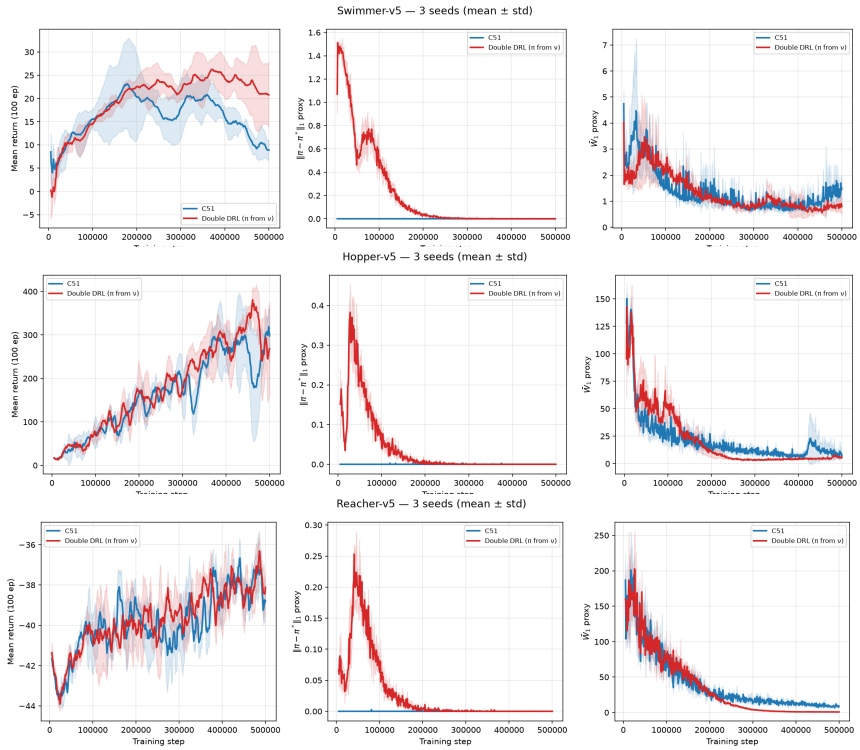


Figure 1: Deep MuJoCo-v5 results over three seeds. Each row is one environment: mean return (left), policy-discrepancy proxy (middle), and $\bar{W}_1$ proxy (right). Shaded regions denote one standard deviation across seeds.

Table 2: Final multi-seed performance summary. Values are mean ± one standard deviation over three seeds. Higher return is better.

Environment	C51	Double DRL (π from ν)
Swimmer-v5	8.91 ± 2.24	$\mathbf{20.79 \pm 7.05}$
Hopper-v5	$\mathbf{297.79 \pm 61.26}$	267.84 ± 106.83
Reacher-v5	-38.78 ± 0.71	$\mathbf{-38.11 \pm 0.60}$

5.2.2 Results and Analysis.

The results in Fig. 1 show that the proposed double construction changes the learning dynamics in the direction predicted by the theory. In all three tasks, the policy-discrepancy proxy of Double DRL exhibits a clear transient phase followed by rapid decay toward zero, typically within the first 2–3×10^5 steps. This behavior is consistent with the assumption used in Theorem 2 that the reference policy sequence converges to an optimal-policy limit. Importantly, the decay of the policy proxy is accompanied by a systematic reduction of the $\bar{W}_1$ proxy, suggesting that the main distributional iterates become increasingly stable once the auxiliary policy has stopped moving substantially.

The effect is most visible in Swimmer-v5. Standard C51 improves quickly but then loses performance after roughly 2×10^5 steps, while Double DRL continues to maintain returns around a higher plateau. The final return improves from 8.91 ± 2.24 for C51 to 20.79 ± 7.05 for Double DRL. This pattern matches the motivation of the method: when greedy distributional updates induce unstable policy switches, decoupling policy generation from the main return-distribution update can produce more persistent improvement. Reacher-v5 shows a smaller but consistent gain, with Double DRL reaching -38.11 ± 0.60 compared with -38.78 ± 0.71 for C51, while also driving the $\bar{W}_1$ proxy close to zero by the end of training. Hopper-v5 is more mixed: C51 attains a slightly higher final mean return, but Double DRL obtains a higher greedy evaluation score and its $\bar{W}_1$ proxy remains

lower in the late stage of training. The larger variance of Double DRL on Hopper indicates that the practical implementation still inherits sensitivity from continuous-control optimization, even though the distributional-dynamics diagnostics improve.

Overall, these experiments support the theoretical picture at the level of deep function approximation. The auxiliary policy sequence in Double DRL acts as a stabilizing homotopy from exploratory soft policies toward the limiting greedy policy, and the main distributional update benefits from this smoother policy path. The empirical gains are therefore not merely return improvements: the diagnostic curves provide evidence that Double DRL reduces the oscillatory behavior of greedy distributional control and promotes convergence of the learned return distribution in Wasserstein geometry.

6 Conclusion

We introduced double distributional value iteration to address non-convergence in distributional control caused by switching among actions with equal optimal expected returns. The method separates action-value estimation from return-distribution evaluation and uses a vanishing-temperature softmax policy to stabilize action selection without changing the original objective or return target. For tabular MDPs, we established convergence of both the planning and stochastic approximation versions of DDVI to the return distribution induced by the policy that is uniform over optimal actions.

Our limitations are listed as follows. Our guarantees rely on tabular assumptions and a cooling schedule matched to value-estimation error. They therefore do not directly extend to deep function approximation or continuous action spaces. In practice, temperature selection and the additional auxiliary estimator may also introduce tuning and computational costs. The theorems apply to exact distributional operators, while the reported runs use a categorical projection and synchronous model-based backups.

AI use statement

Generative AI tools were used to assist with proofs and to improve the wording and readability of the manuscript. The proof ideas and overall proof strategies were developed by the authors. The authors take full responsibility for the mathematical arguments and all other content of the paper.

References

Eugene L. Allgower and Kurt Georg. Numerical Continuation Methods: An Introduction. Springer, 1990.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 449–458. PMLR, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. Distributional Reinforcement Learning. MIT Press, Cambridge, MA, 2023. URL https://www.distributional-rl.org/.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 1096–1105. PMLR, 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, pp. 2892–2901, 2018b. doi: 10.1609/aaai.v32i1.11791. URL https://ojs.aaai.org/index.php/AAAI/article/view/11791.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 2160–2169, 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6): 721–741, 1984.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 1352–1361. PMLR, 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Bruce Hajek. Cooling schedules for optimal annealing. Mathematics of Operations Research, 13 (2):311–329, 1988.

Chii-Ruey Hwang. Laplace's method revisited: Weak convergence of probability measures. The Annals of Probability, 8(6):1177–1182, 1980.

Yash Jhaveri, Harley Wiltzer, Patrick Shafto, Marc G. Bellemare, and David Meger. Convergence theorems for entropy-regularized and distributional reinforcement learning. In Advances in Neural Information Processing Systems, 2025. URL https://arxiv.org/abs/2510.08526.

Xiaoteng Ma, Junyao Chen, Li Xia, Jun Yang, Qianchuan Zhao, and Zhengyuan Zhou. DSAC: Distributional soft actor-critic for risk-sensitive reinforcement learning. Journal of Artificial Intelligence Research, 83(4), 2025. doi: 10.1613/jair.1.17526. URL https://www.jair.org/index.php/jair/article/view/17526.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. arXiv preprint arXiv:1705.07798, 2017. URL https://arxiv.org/abs/1705.07798.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, pp. 3185–3205, 2020.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi (ed.), Optimizing Methods in Statistics, pp. 233–257. Academic Press, New York, 1971.

Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE, 86(11):2210–2239, 1998.

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pp. 29–37. PMLR, 2018. URL https://proceedings.mlr.press/v84/rowland18a.html.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT Press, 2 edition, 2018. ISBN 9780262039246. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. Machine Learning, 8(3-4):279–292, 1992.

A CONVERGENCE PROOF FOR THE PLANNING VERSION

A.1 PROOF OF THEOREM 1

Now we give some supporting lemmas that will help the proof of Theorem 1: i) the convergence of the policy sequence $\{\pi^{(t)}\}_{t=1}^{\infty}$; ii) the non-expansiveness of the operator $\mathcal{T}^{\pi}$, for any policy π; iii) the Lipschitz continuity of the operator $\mathcal{T}^{\pi}$ w.r.t. π.

Lemma 1. *Suppose $\gamma^t/\lambda_t \to 0$, then $\|\pi^{(t)} - \pi^*\|_1 \to 0$.*

Lemma 2. *For any policy π and for any $\eta \in M_{\mathcal{F}_{W_1}}^{S\times A}$,*

$$\|\mathcal{T}^{\pi}\eta\|_{\mathcal{F}_{W_1},\infty} \leq \|\eta\|_{\mathcal{F}_{W_1},\infty}. \tag{16}$$

Lemma 3. *For any policy π_1, π_2 and for any $\eta \in M_{\mathcal{F}_{W_1}}^{S\times A}$,*

$$\|(\mathcal{T}^{\pi_1} - \mathcal{T}^{\pi_2})\eta\|_{\mathcal{F}_{W_1},\infty} \leq \|\pi_1 - \pi_2\|_1 \, \|\eta\|_{\mathcal{F}_{W_1},\infty}. \tag{17}$$

Proof. Since $\|\pi^{(t)} - \pi^*\|_1 \to 0$, for any $\varepsilon > 0$, there exists $T \in \mathbb{N}$ such that for any $t > T$, $\|\pi^{(t)} - \pi^*\|_1 < \varepsilon$. Now we have

$$\eta^{(T+N)} = \mathcal{T}^{\pi^{(T+N)}} \cdots \mathcal{T}^{\pi^{(T+1)}} \eta^{(T)} \tag{18}$$

$$= \big(\mathcal{T}^{\pi^*} + (\mathcal{T}^{\pi^{(T+N)}} - \mathcal{T}^{\pi^*})\big) \cdots \big(\mathcal{T}^{\pi^*} + (\mathcal{T}^{\pi^{(T+1)}} - \mathcal{T}^{\pi^*})\big)\eta^{(T)} \tag{19}$$

$$= \prod_{i=1}^{N} \big(\mathcal{T}^{\pi^*} + (\mathcal{T}^{\pi^{(T+i)}} - \mathcal{T}^{\pi^*})\big)\eta^{(T)}. \tag{20}$$

For simplicity, we write $(\mathcal{T}^{\pi^{(T+i)}} - \mathcal{T}^{\pi^*})$ as $\Delta^{(i)}$. Expanding the product, we obtain

$$\eta^{(T+N)} = \sum_{d\in\{0,1\}^N} \prod_{i=1}^{N} \big(\mathcal{T}^{\pi^*}\big)^{d_i} \big(\Delta^{(i)}\big)^{1-d_i} \eta^{(T)}. \tag{21}$$

By Lemma 2, for any $\eta \in \big(M_{\mathcal{F}_{W_1}}\big)^{S\times A}$,

$$\|\mathcal{T}^{\pi^*}\eta\|_{\mathcal{F}_{W_1},\infty} \leq \|\eta\|_{\mathcal{F}_{W_1},\infty}. \tag{22}$$

Moreover, by Lemma 3 and the fact that $\|\pi^{(t)} - \pi^*\|_1 < \varepsilon$ for all $t > T$, it follows that for all $i \in \{1,\ldots,N\}$ and all $\eta \in \big(M_{\mathcal{F}_{W_1}}\big)^{S\times A}$,

$$\|\Delta^{(i)}\eta\|_{\mathcal{F}_{W_1},\infty} \leq \varepsilon \, \|\eta\|_{\mathcal{F}_{W_1},\infty}. \tag{23}$$

Therefore, we have

$$\left\| \prod_{i=1}^{N} \big(\mathcal{T}^{\pi^*}\big)^{d_i} \big(\Delta^{(i)}\big)^{1-d_i} \eta^{(T)} \right\|_{\mathcal{F}_{W_1},\infty} \leq \varepsilon^{N-\sum_{i=1}^{N} d_i} \, \|\eta^{(T)}\|_{\mathcal{F}_{W_1},\infty}. \tag{24}$$

and further

$$\left\|\eta^{(T+N)} - \big(\mathcal{T}^{\pi^*}\big)^N \eta^{(T)}\right\|_{\mathcal{F}_{W_1},\infty} = \left\| \sum_{d\in\{0,1\}^N,\, d\neq(0,\ldots,0)} \prod_{i=1}^{N} \big(\mathcal{T}^{\pi^*}\big)^{d_i} \big(\Delta^{(i)}\big)^{1-d_i} \eta^{(T)} \right\|_{\mathcal{F}_{W_1},\infty} \tag{25}$$

$$\leq \sum_{d\in\{0,1\}^N,\, d\neq(0,\ldots,0)} \left\| \prod_{i=1}^{N} \big(\mathcal{T}^{\pi^*}\big)^{d_i} \big(\Delta^{(i)}\big)^{1-d_i} \eta^{(T)} \right\|_{\mathcal{F}_{W_1},\infty} \tag{26}$$

$$\leq \sum_{d\in\{0,1\}^N,\, d\neq(0,\ldots,0)} \varepsilon^{N-\sum_{i=1}^{N} d_i} \, \|\eta^{(T)}\|_{\mathcal{F}_{W_1},\infty} \tag{27}$$

$$= \big((1+\varepsilon)^N - 1\big) \, \|\eta^{(T)}\|_{\mathcal{F}_{W_1},\infty} \tag{28}$$

$$\leq C_0\big((1+\varepsilon)^N - 1\big), \tag{29}$$

where C_0 is a universal constant as $\eta^{(T)}$ is a vector of bounded distributions. Let $N = \log_\gamma \varepsilon$, then as $\mathcal{T}^{\pi^*}$ is a γ-contraction w.r.t. the $\bar{W}_1$ metric,

$$\left\|\eta^{\pi^*} - \big(\mathcal{T}^{\pi^*}\big)^N \eta^{(T)}\right\|_{\mathcal{F}_{W_1},\infty} = \bar{W}_1\Big(\big(\mathcal{T}^{\pi^*}\big)^N \eta^{(T)}, \eta^{\pi^*}\Big) \tag{30}$$

$$\leq \varepsilon\, \bar{W}_1(\eta^{\pi^*}, \eta^{(T)}) \leq C_1\varepsilon. \tag{31}$$

Here C_1 is a universal constant as both η^{π^*} and $\eta^{(T)}$ have bounded supports. Putting all the pieces together, we can get

$$\bar{W}_1(\eta^{(T+N)}, \eta^{\pi^*}) = \|\eta^{(T+N)} - \eta^{\pi^*}\|_{\mathcal{F}_{W_1},\infty} \tag{32}$$

$$\leq \left\|\eta^{\pi^*} - \big(\mathcal{T}^{\pi^*}\big)^N \eta^{(T)}\right\|_{\mathcal{F}_{W_1},\infty} + \left\|\eta^{(T+N)} - \big(\mathcal{T}^{\pi^*}\big)^N \eta^{(T)}\right\|_{\mathcal{F}_{W_1},\infty} \tag{33}$$

$$\leq C_1\varepsilon + C_0\big((1+\varepsilon)^{\log_\gamma \varepsilon} - 1\big) \tag{34}$$

$$\leq C_1\varepsilon + C_0'\varepsilon \log_\gamma \varepsilon. \tag{35}$$

Note that in 35 C_0' is a universal constant and the inequality is valid for sufficiently small ε. Since ε is arbitrarily chosen, we finish the proof. □

A.2 Proofs of Lemma on Theorem 1

Proof of Lemma 1. It suffices to consider the non-distributional counterpart of the auxiliary iteration. Let $Q^{(t)}$ be the state-action value function associated with $\nu^{(t)}$, then we have $Q^{(t+1)} = T^{\mathcal{G}}Q^{(t)}$. We demonstrate the soft greedy policy $\pi^{(t)} = G_{\lambda_t}(Q^{(t)})$ will converge to π^*. Since the Bellman optimality operator $T^{\mathcal{G}}$ is a γ-contraction w.r.t. the $\|\cdot\|_\infty$, we have $\|Q^{(t)} - Q^*\|_\infty \leq C\gamma^t$. Let $Q^*_{\max}(s)$ denote $\max_{a\in A} Q^*(s,a)$, we have

$$G_{\lambda_t}(Q^{(t)})(a \mid s) = \frac{\exp(Q^{(t)}(s,a)/\lambda_t)}{\sum_{a\in A}\exp(Q^{(t)}(s,a')/\lambda_t)} \tag{36}$$

$$\leq \frac{\exp((Q^*(s,a) + C\gamma^t)/\lambda_t)}{\sum_{a\in A}\exp((Q^*(s,a') - C\gamma^t)/\lambda_t)} \tag{37}$$

$$= \frac{\exp((Q^*(s,a) - Q^*_{\max}(s))/\lambda_t + C\gamma^t/\lambda_t)}{\sum_{a\in A}\exp((Q^*(s,a') - Q^*_{\max}(s))/\lambda_t - C\gamma^t/\lambda_t)} \tag{38}$$

$$\to \pi^*(a \mid s) \quad (t\to\infty). \tag{39}$$

The last limit 39 holds due to $\lambda_t \to 0$ and $\gamma^t/\lambda_t \to 0$ as $t\to\infty$. Similarly,

$$G_{\lambda_t}(Q^{(t)})(a \mid s) = \frac{\exp(Q^{(t)}(s,a)/\lambda_t)}{\sum_{a\in A}\exp(Q^{(t)}(s,a')/\lambda_t)} \tag{40}$$

$$\geq \frac{\exp((Q^*(s,a) - C\gamma^t)/\lambda_t)}{\sum_{a\in A}\exp((Q^*(s,a') + C\gamma^t)/\lambda_t)} \tag{41}$$

$$= \frac{\exp((Q^*(s,a) - Q^*_{\max}(s))/\lambda_t - C\gamma^t/\lambda_t)}{\sum_{a\in A}\exp((Q^*(s,a') - Q^*_{\max}(s))/\lambda_t + C\gamma^t/\lambda_t)} \tag{42}$$

$$\to \pi^*(a \mid s) \quad (t\to\infty). \tag{43}$$

Therefore, we have $\pi^{(t)}(a \mid s) = G_{\lambda_t}(Q^{(t)})(a \mid s) \to \pi^*(a \mid s)$, $\forall (s,a) \in S\times A$. And the conclusion follows. □

Proof of Lemma 2. We have

$$\|[\mathcal{T}^\pi\eta](s,a)\|_{\mathcal{F}_{W_1}} = \left\|\sum_{s'\in S}\sum_{a'\in A} P(s' \mid s,a)\,\pi(a' \mid s')\int (b_{r,\gamma})_{\#}\big(\eta(s',a')\big)\, P_R(dr \mid s,a)\right\|_{\mathcal{F}_{W_1}} \tag{44}$$

$$\leq \sum_{s' \in S} \sum_{a' \in A} P(s' \mid s, a)\, \pi(a' \mid s') \int \|(b_{r,\gamma})_{\#}\big(\eta(s', a')\big)\|_{\mathcal{F}_{W_1}} P_R(dr \mid s, a) \tag{45}$$

$$\leq \max_{(s',a') \in S \times A} \int \|(b_{r,\gamma})_{\#}\big(\eta(s', a')\big)\|_{\mathcal{F}_{W_1}} P_R(dr \mid s, a) \tag{46}$$

$$\leq \|\eta\|_{\mathcal{F}_{W_1}, \infty}. \tag{47}$$

Here 45 is by the triangular inequality of Bochner integral, and 47 is by the non-expansiveness of the push-forward operator on $M_{\mathcal{F}_{W_1}}$. Specifically, for any $\mu \in M_{\mathcal{F}_{W_1}}$,

$$\|(b_{r,\gamma})_{\#}\mu\|_{\mathcal{F}_{W_1}} = \sup_{f \in \mathcal{F}_{W_1}} \mathbb{E}_{X \sim \mu} f(r + \gamma X) \leq \sup_{f \in \mathcal{F}_{W_1}} \mathbb{E}_{X \sim \mu} f(X) = \|\mu\|_{\mathcal{F}_{W_1}}. \tag{48}$$

The inequality 48 holds since for any $f \in \mathcal{F}_{W_1}$, $f(r + \gamma \cdot) \in \mathcal{F}_{W_1}$. The final conclusion follows by taking supremum over $(s, a) \in S \times A$. □

Proof of Lemma 3. We have $\forall (s, a) \in S \times A$,

$$\|[(\mathcal{T}^{\pi_1} - \mathcal{T}^{\pi_2})\eta](s, a)\|_{\mathcal{F}_{W_1}} = \|[\mathcal{T}^{\pi_1}\eta](s, a) - [\mathcal{T}^{\pi_2}\eta](s, a)\|_{\mathcal{F}_{W_1}} \tag{49}$$

$$= \left\| \sum_{s' \in S} \sum_{a' \in A} P(s' \mid s, a)\, (\pi_1(a' \mid s') - \pi_2(a' \mid s')) \int (b_{r,\gamma})_{\#}\big(\eta(s', a')\big)\, P_R(dr \mid s, a) \right\|_{\mathcal{F}_{W_1}} \tag{50}$$

$$\leq \sum_{s' \in S} P(s' \mid s, a) \sum_{a' \in A} \big|\pi_1(a' \mid s') - \pi_2(a' \mid s')\big| \int \|(b_{r,\gamma})_{\#}\big(\eta(s', a')\big)\|_{\mathcal{F}_{W_1}} P_R(dr \mid s, a) \tag{51}$$

$$\leq \|\pi_1 - \pi_2\|_1 \max_{(s',a') \in S \times A} \int \|(b_{r,\gamma})_{\#}\big(\eta(s', a')\big)\|_{\mathcal{F}_{W_1}} P_R(dr \mid s, a) \tag{52}$$

$$\leq \|\pi_1 - \pi_2\|_1\, \|\eta\|_{\mathcal{F}_{W_1}, \infty}. \tag{53}$$

Here 51 is by the triangular inequality of Bochner integral, and 53 is by the non-expansiveness of the push-forward operator on $M_{\mathcal{F}_{W_1}}$. Specifically, for any $\mu \in M_{\mathcal{F}_{W_1}}$,

$$\|(b_{r,\gamma})_{\#}\mu\|_{\mathcal{F}_{W_1}} = \sup_{f \in \mathcal{F}_{W_1}} \mathbb{E}_{X \sim \mu} f(r + \gamma X) \leq \sup_{f \in \mathcal{F}_{W_1}} \mathbb{E}_{X \sim \mu} f(X) = \|\mu\|_{\mathcal{F}_{W_1}}. \tag{54}$$

The inequality 54 holds since for any $f \in \mathcal{F}_{W_1}$, $f(r + \gamma \cdot) \in \mathcal{F}_{W_1}$. The final conclusion follows by taking supremum over $(s, a) \in S \times A$. □

B Convergence proof for the stochastic approximation version

B.1 Proof of Theorem

We write $H := (1 - \gamma)^{-1}$ and work with the dual formulation of W_1 from the preliminaries: if μ is the difference of two probability measures on $[0, H]$, then

$$\|\mu\|_{\mathcal{F}_{W_1}} = W_1(\mu_+, \mu_-), \qquad \|\eta\|_{\mathcal{F}_{W_1}, \infty} = \overline{W}_1(\eta_+, \eta_-).$$

The operator $\mathcal{T}^{\pi}$ is extended linearly to $(\mathcal{M}_{\mathcal{F}_{W_1}})^{\mathcal{S} \times \mathcal{A}}$ without changing its definition.

Now we give some lemmas that will help proof of Theorem 2:

Lemma 4 (Auxiliary means and reference policy)**.** *Under Assumptions 2,*

$$\left\|Q^{(t)} - Q^*\right\|_{\infty} \to 0 \quad \textit{a.s.} \tag{55}$$

If in addition Assumption 3 holds, then

$$\left\|\pi^{(t)} - \pi^*\right\|_1 \to 0 \quad \textit{a.s.}, \tag{56}$$

where $\pi^* \in \Pi^*$ is given by

$$\pi^*(a \mid s) = \begin{cases} 1 \big/ \left|\arg\max_{a' \in \mathcal{A}} Q^*(s, a')\right| & \textit{if } a \in \arg\max_{a' \in \mathcal{A}} Q^*(s, a'), \\ 0 & \textit{otherwise}, \end{cases} \qquad \forall s \in \mathcal{S}.$$

Proof. Taking means in the auxiliary update yields

$$Q^{(t+1)}(S_t, A_t) = \big(1 - \beta_t(S_t, A_t)\big)Q^{(t)}(S_t, A_t) + \beta_t(S_t, A_t)\Big[R_t + \gamma Q^{(t)}\big(S_{t+1}, A^{\nu}_{t+1}\big)\Big].$$

Since the auxiliary target action is selected by the greedy rule $\mathcal{G}(\nu^{(t)})$,

$$Q^{(t)}\big(S_{t+1}, A^{\nu}_{t+1}\big) = \max_{a'} Q^{(t)}(S_{t+1}, a').$$

Hence

$$Q^{(t+1)}(S_t, A_t) = Q^{(t)}(S_t, A_t) + \beta_t(S_t, A_t)\Big[R_t + \gamma \max_{a'} Q^{(t)}(S_{t+1}, a') - Q^{(t)}(S_t, A_t)\Big]. \tag{57}$$

This is the standard tabular Q-learning recursion. By the correct sampling assumption, the boundedness of rewards, and the coordinate-wise stochastic approximation stepsizes, the classical Q-learning convergence theorem gives

$$\big\|Q^{(t)} - Q^*\big\|_\infty \to 0 \quad \text{a.s.}$$

Then we proof the convergence of the reference policy. Fix a probability-one event on which equation 55 and Assumption 3 hold simultaneously. Write

$$e^{(t)}(s, a) := Q^{(t)}(s, a) - Q^*(s, a),$$
$$Q^*_{\max}(s) := \max_{a'} Q^*(s, a'),$$
$$\Delta(s, a) := Q^*_{\max}(s) - Q^*(s, a).$$

We have:

$$\max_{s,a}\left|\frac{e^{(t)}(s, a)}{\lambda_t}\right| \to 0. \tag{58}$$

From the softmax definition of the reference policy,

$$\pi^{(t+1)}(a \mid s) = \frac{\exp\big(-\Delta(s, a)/\lambda_t + e^{(t)}(s, a)/\lambda_t\big)}{\sum_{a'} \exp\big(-\Delta(s, a')/\lambda_t + e^{(t)}(s, a')/\lambda_t\big)}. \tag{59}$$

If $a \in \arg\max_{a'} Q^*(s, a')$, then $\Delta(s, a) = 0$, so the corresponding exponential weight tends to 1 by equation 58. If $a \notin \arg\max_{a'} Q^*(s, a')$, then $\Delta(s, a) > 0$; combined with $\lambda_t \to 0$ and equation 58, the corresponding exponential weight tends to 0. Therefore

$$\pi^{(t+1)}(a \mid s) \to \pi^*(a \mid s) \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Since $\mathcal{S}$ and $\mathcal{A}$ are finite,

$$\big\|\pi^{(t+1)} - \pi^*\big\|_1 \to 0.$$

□

By definition of π^*,

$$\sum_{a'} \pi^*(a' \mid s)\, Q^*(s, a') = \max_{a'} Q^*(s, a').$$

Hence $T^{\pi^*} Q^* = T^{\mathcal{G}} Q^* = Q^*$. The operator T^{π^*} is a γ-contraction under $\|\cdot\|_\infty$, so its fixed point is unique. Thus $Q^{\pi^*} = Q^*$, i.e. $\pi^* \in \Pi^*$.

Lemma 5 (Contraction of the distributional Bellman operator)**.** *For any policy π and any $\mu \in (\mathcal{M}_{\mathcal{F}_{W_1}})^{\mathcal{S}\times\mathcal{A}}$,*

$$\big\|\mathcal{T}^\pi \mu\big\|_{\mathcal{F}_{W_1},\infty} \le \gamma \big\|\mu\big\|_{\mathcal{F}_{W_1},\infty}. \tag{60}$$

In particular, for any vectors of probability distributions η_1, η_2,

$$\overline{W}_1\big(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2\big) \le \gamma\, \overline{W}_1(\eta_1, \eta_2). \tag{61}$$

Proof. For $\mu \in \mathcal{M}_{\mathcal{F}_{W_1}}$ with $\mu([0,H]) = 0$ and $f \in \mathcal{F}_{W_1}$. Then

$$\int f(r+\gamma z)\,\mu(dz) = \int \big[f(r+\gamma z) - f(r)\big]\,\mu(dz).$$

The map

$$z \mapsto \frac{f(r+\gamma z) - f(r)}{\gamma}$$

is 1-Lipschitz, hence belongs to $\mathcal{F}_{W_1}$. Therefore

$$\big\|(b_{r,\gamma})_{\#}\mu\big\|_{\mathcal{F}_{W_1}} \le \gamma \|\mu\|_{\mathcal{F}_{W_1}}. \tag{62}$$

Mixing over the reward, the next state and the next action, and using the triangle inequality,

$$\begin{aligned}\big\|(\mathcal{T}^{\pi}\mu)(s,a)\big\|_{\mathcal{F}_{W_1}} &\le \mathbb{E}_{S'\sim P_{s,a},\, A'\sim\pi_{S'},\, R\sim P^R_{s,a}} \big\|(b_{R,\gamma})_{\#}\mu(S',A')\big\|_{\mathcal{F}_{W_1}} \\ &\le \gamma \max_{s',a'} \big\|\mu(s',a')\big\|_{\mathcal{F}_{W_1}}.\end{aligned}$$

Taking the supremum over (s,a) yields equation 60. Setting $\mu = \eta_1 - \eta_2$ and using linearity together with the identification $\|\eta_1 - \eta_2\|_{\mathcal{F}_{W_1},\infty} = \overline{W}_1(\eta_1,\eta_2)$ gives equation 61. □

Lemma 6 (Continuity of $\mathcal{T}^{\pi}$ in the policy)**.** *For any policies π_1, π_2 and any $\eta \in \Delta([0,H])^{\mathcal{S}\times\mathcal{A}}$,*

$$\big\|(\mathcal{T}^{\pi_1} - \mathcal{T}^{\pi_2})\eta\big\|_{\mathcal{F}_{W_1},\infty} \le \frac{\gamma}{2(1-\gamma)}\,\|\pi_1 - \pi_2\|_1. \tag{63}$$

Proof. For the next state s' and we define

$$\rho_1 := \sum_{a'} \pi_1(a' \mid s')\,\eta(s',a'), \qquad \rho_2 := \sum_{a'} \pi_2(a' \mid s')\,\eta(s',a').$$

Pairing the common mass of the two action probabilities, the unmatched mass equals

$$\frac{1}{2}\sum_{a'} \big|\pi_1(a' \mid s') - \pi_2(a' \mid s')\big|.$$

Since every distribution is supported on $[0,H]$, the unit transportation cost of this mass is at most H. Thus

$$W_1(\rho_1,\rho_2) \le \frac{H}{2}\sum_{a'} \big|\pi_1(a' \mid s') - \pi_2(a' \mid s')\big|. \tag{64}$$

After the common affine map $z \mapsto r + \gamma z$, the W_1 distance is multiplied by at most γ. Mixing over the reward and the next state then gives

$$\big\|((\mathcal{T}^{\pi_1} - \mathcal{T}^{\pi_2})\eta)(s,a)\big\|_{\mathcal{F}_{W_1}} \le \frac{\gamma H}{2}\,\|\pi_1 - \pi_2\|_1.$$

Taking the supremum over (s,a) and substituting $H = (1-\gamma)^{-1}$ yields equation 63. □

Lemma 7 (Averaged sampling noise vanishes)**.** *Let $M^{(t+1)} \in (\mathcal{M}_{\mathcal{F}_{W_1}})^{\mathcal{S}\times\mathcal{A}}$ satisfy, for all (s,a) and all $z \in [0,H]$,by Assumption 1 we have*

$$\mathbb{E}\big[M^{(t+1)}(s,a)([0,z]) \mid \mathcal{F}_t\big] = 0 \tag{65}$$

and

$$\big|M^{(t+1)}(s,a)([0,z])\big| \le 1. \tag{66}$$

Define $Z^{(0)} = 0$ and

$$Z^{(t+1)}(s,a) = \big(1 - \alpha_t(s,a)\big)Z^{(t)}(s,a) + \alpha_t(s,a)\,M^{(t+1)}(s,a). \tag{67}$$

Under Assumption 2,

$$\big\|Z^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \to 0 \quad \textit{a.s.} \tag{68}$$

Proof. On a coordinate (s, a) and write

$$Y^{(t)}(z) := Z^{(t)}(s,a)([0,z]), \qquad \xi^{(t+1)}(z) := M^{(t+1)}(s,a)([0,z]).$$

Then

$$Y^{(t+1)}(z) = \big(1 - \alpha_t(s,a)\big)Y^{(t)}(z) + \alpha_t(s,a)\,\xi^{(t+1)}(z). \tag{69}$$

By the initialization and equation 66, $\big|Y^{(t)}(z)\big| \le 1$. Define

$$V^{(t)}(s,a) := \int_0^H \big|Y^{(t)}(z)\big|^2\, dz, \tag{70}$$

so that $0 \le V^{(t)}(s,a) \le H$. Squaring equation 69, taking the conditional expectation, cancelling the cross term via equation 65, and integrating in z gives

$$\begin{aligned}\mathbb{E}\big[V^{(t+1)}(s,a) \mid \mathcal{F}_t\big] &\le \big(1 - \alpha_t(s,a)\big)^2 V^{(t)}(s,a) + H\alpha_t(s,a)^2 \\ &\le V^{(t)}(s,a) - \alpha_t(s,a)\, V^{(t)}(s,a) + H\alpha_t(s,a)^2,\end{aligned} \tag{71}$$

where the last step uses $\big(1 - \alpha_t(s,a)\big)^2 \le 1 - \alpha_t(s,a)$. By the Robbins–Siegmund almost supermartingale theorem Robbins & Siegmund (1971), $V^{(t)}(s,a)$ converges almost surely and

$$\sum_t \alpha_t(s,a)\, V^{(t)}(s,a) < \infty \quad \text{a.s.} \tag{72}$$

Combined with $\sum_t \alpha_t(s,a) = \infty$, the limit of $V^{(t)}(s,a)$ must be zero. The dual representation of W_1 and Cauchy–Schwarz yield

$$\big\|Z^{(t)}(s,a)\big\|_{\mathcal{F}_{W_1}} = \int_0^H \big|Y^{(t)}(z)\big|\, dz \le \sqrt{H\, V^{(t)}(s,a)} \to 0 \quad \text{a.s.}$$

Since there are finitely many coordinates, equation 68 follows. □

Robbins–Siegmund almost supermartingale theorem. Let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration, and let X_t, A_t, B_t, C_t be nonnegative and $\mathcal{F}_t$-measurable. Suppose that almost surely

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] \le (1 + A_t)X_t + B_t - C_t, \qquad \sum_{t=0}^{\infty} A_t < \infty, \qquad \sum_{t=0}^{\infty} B_t < \infty.$$

Then X_t converges almost surely to a finite (random) limit, and $\sum_{t=0}^{\infty} C_t < \infty$ almost surely Robbins & Siegmund (1971).

Lemma 8 (Asynchronous contraction with vanishing perturbation)**.** *Let $\{d^{(t)}(s,a)\}$ be a nonnegative array satisfying*

$$D^{(t)} := \max_{s,a} d^{(t)}(s,a), \qquad \sup_t D^{(t)} < \infty, \tag{73}$$

and

$$d^{(t+1)}(s,a) \le \big(1 - \alpha_t(s,a)\big)d^{(t)}(s,a) + \alpha_t(s,a)\big[\gamma D^{(t)} + r_t\big], \tag{74}$$

where $0 \le \alpha_t(s,a) \le 1$, $\sum_t \alpha_t(s,a) = \infty$, and $r_t \ge 0$, $r_t \to 0$. Then

$$D^{(t)} \to 0. \tag{75}$$

Proof. Let $L := \limsup_{t\to\infty} D^{(t)} < \infty$. For any $\varepsilon > 0$ there exists T such that $D^{(t)} \le L + \varepsilon$ and $r_t \le \varepsilon$ for all $t \ge T$.

Let $c_\varepsilon := \gamma(L + \varepsilon) + \varepsilon$. Then equation 74 becomes

$$d^{(t+1)}(s,a) \le \big(1 - \alpha_t(s,a)\big)d^{(t)}(s,a) + \alpha_t(s,a)\, c_\varepsilon.$$

Hence

$$\big(d^{(t)}(s,a) - c_\varepsilon\big)_+ \le \big(d^{(T)}(s,a) - c_\varepsilon\big)_+ \prod_{k=T}^{t-1} \big(1 - \alpha_k(s,a)\big).$$

Since $\sum_k \alpha_k(s,a) = \infty$, the product tends to zero, and $\limsup_t d^{(t)}(s,a) \le c_\varepsilon$.

Finiteness of $\mathcal{S} \times \mathcal{A}$ yields $L \le \gamma(L + \varepsilon) + \varepsilon$. Letting $\varepsilon \downarrow 0$ gives $L \le \gamma L$. As $0 < \gamma < 1$, one has $L = 0$. □

Then we can proof our main Theorem 2:

Proof. By Lemma 4,It remains to prove convergence of the main distributions. Fix the target η^{π^*}, which satisfies

$$\mathcal{T}^{\pi^*}\eta^{\pi^*} = \eta^{\pi^*}. \tag{76}$$

Define

$$\delta^{(t)} := \big(\mathcal{T}^{\pi^{(t+1)}} - \mathcal{T}^{\pi^*}\big)\eta^{(t)}. \tag{77}$$

By Lemma 6,

$$\big\|\delta^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \le \frac{\gamma}{2(1-\gamma)}\,\big\|\pi^{(t+1)} - \pi^*\big\|_1.$$

Combined with Lemma 4,

$$\big\|\delta^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \to 0 \quad \text{a.s.} \tag{78}$$

On the updated coordinate set

$$M^{(t+1)}(S_t, A_t) := \widehat{\eta}^{(t+1)} - \big(\mathcal{T}^{\pi^{(t+1)}}\eta^{(t)}\big)(S_t, A_t), \tag{79}$$

and set $M^{(t+1)}(s,a) = 0$ elsewhere. We have:

$$\mathbb{E}\big[M^{(t+1)}(s,a)([0,z]) \mid \mathcal{F}_t\big] = 0.$$

Moreover, $M^{(t+1)}(s,a)$ is either zero or a difference of two probability measures, so $M^{(t+1)}(s,a)([0,H]) = 0$ and

$$\big|M^{(t+1)}(s,a)([0,z])\big| \le 1.$$

Let $Z^{(0)} = 0$ and

$$Z^{(t+1)}(s,a) = \big(1-\alpha_t(s,a)\big)Z^{(t)}(s,a) + \alpha_t(s,a)\,M^{(t+1)}(s,a). \tag{80}$$

Lemma 7 yields

$$\big\|Z^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \to 0 \quad \text{a.s.} \tag{81}$$

Then define the total error

$$E^{(t)} := \eta^{(t)} - \eta^{\pi^*}. \tag{82}$$

From the main-process update, equation 76–equation 79, and linearity of $\mathcal{T}^{\pi^*}$,

$$\begin{aligned} E^{(t+1)}(s,a) = \big(1-\alpha_t(s,a)\big)E^{(t)}(s,a) \\ + \alpha_t(s,a)\Big[\big(\mathcal{T}^{\pi^*}E^{(t)}\big)(s,a) + \delta^{(t)}(s,a) + M^{(t+1)}(s,a)\Big]. \end{aligned} \tag{83}$$

Define the denoised error $U^{(t)} := E^{(t)} - Z^{(t)}$. Subtracting equation 80 from equation 83 gives

$$\begin{aligned} U^{(t+1)}(s,a) = \big(1-\alpha_t(s,a)\big)U^{(t)}(s,a) \\ + \alpha_t(s,a)\Big[\big(\mathcal{T}^{\pi^*}U^{(t)}\big)(s,a) + \mathcal{T}^{\pi^*}Z^{(t)}(s,a) + \delta^{(t)}(s,a)\Big], \end{aligned} \tag{84}$$

By Lemma 5,

$$\big\|\mathcal{T}^{\pi^*}Z^{(t)} + \delta^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \le \gamma\big\|Z^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} + \big\|\delta^{(t)}\big\|_{\mathcal{F}_{W_1},\infty}.$$

Using equation 78 and equation 81,

$$r_t := \big\|\mathcal{T}^{\pi^*}Z^{(t)} + \delta^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} \to 0 \quad \text{a.s.} \tag{85}$$

Set

$$d^{(t)}(s,a) := \big\|U^{(t)}(s,a)\big\|_{\mathcal{F}_{W_1}}, \qquad D^{(t)} := \big\|U^{(t)}\big\|_{\mathcal{F}_{W_1},\infty} = \max_{s,a} d^{(t)}(s,a).$$

From equation 84 and Lemma 5,

$$d^{(t+1)}(s,a) \le \big(1-\alpha_t(s,a)\big)d^{(t)}(s,a) + \alpha_t(s,a)\big[\gamma D^{(t)} + r_t\big]. \tag{86}$$

It remains to check that $D^{(t)}$ is bounded. Since $\eta^{(t)}(s,a)$ and $\eta^{\pi^*}(s,a)$ are supported on $[0,H]$,

$$\left\|E^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} = \overline{W}_1\big(\eta^{(t)}, \eta^{\pi^*}\big) \le H. \tag{87}$$

The convex-combination structure of equation 80 gives $\left|Z^{(t)}(s,a)([0,z])\right| \le 1$, hence

$$\left\|Z^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} \le H. \tag{88}$$

Therefore

$$D^{(t)} \le \left\|E^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} + \left\|Z^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} \le 2H. \tag{89}$$

Restricting to a probability-one event on which the convergences above and the stepsize conditions hold simultaneously, Lemma 8 yields $D^{(t)} \to 0$.

Since $E^{(t)} = U^{(t)} + Z^{(t)}$,

$$\overline{W}_1\big(\eta^{(t)}, \eta^{\pi^*}\big) = \left\|E^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} \le D^{(t)} + \left\|Z^{(t)}\right\|_{\mathcal{F}_{W_1},\infty} \to 0 \quad \text{a.s.}$$

We finish the proof. □

C TECHNICAL LEMMAS

We collect several technical lemmas and their proofs in this section.

Lemma 9. *Consider the softmax function $\sigma(x) : \mathbb{R}^n \to \mathbb{R}^n$. Specifically,*

$$\sigma(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)}. \tag{90}$$

Then we have $\|\sigma(x_1) - \sigma(x_2)\|_\infty \le \frac{1}{2}\|x_1 - x_2\|_\infty$.

Proof. It suffices to analyze its Jacobian matrix and show that its induced norm $\|\cdot\|_\infty$ is bounded. We have for $i, j \in \{1, 2, \ldots, n\}$,

$$\frac{\partial \sigma(x)_j}{\partial x_i} = \begin{cases} -\sigma(x)_i \sigma(x)_j & i \ne j \\ \sigma(x)_i(1 - \sigma(x)_i) & i = j \end{cases}. \tag{91}$$

Let $J(x)$ denote the Jacobian matrix at x, we have

$$\|J(x)\|_\infty = \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^n \left| \frac{\partial \sigma(x)_i}{\partial x_j} \right| \tag{92}$$

$$= \max_{i \in \{1,\ldots,n\}} \left[\sigma(x)_i(1 - \sigma(x)_i) + \sum_{j=1,\, j \ne i}^n \sigma(x)_i \sigma(x)_j \right] \tag{93}$$

$$= \max_{i \in \{1,\ldots,n\}} 2\sigma(x)_i(1 - \sigma(x)_i) \le \frac{1}{2}. \tag{94}$$

The third equality 94 holds by $\sum_{i=1}^n \sigma(x)_i = 1$ and the last inequality holds by the AM-GM inequality. □

Lemma 10. *For $x \in \mathbb{R}^n$, $\lambda > 0$, consider the function*

$$g(x, \lambda) = \frac{\sum_{i=1}^n x_i \exp(x_i/\lambda)}{\sum_{i=1}^n \exp(x_i/\lambda)}. \tag{95}$$

Then we have for any $x \in \mathbb{R}^n$, $|g(x, \lambda_1) - g(x, \lambda_2)| \le n(n-1)|\lambda_1 - \lambda_2|/4$.

Proof. It suffices to bound $\sup_x |\partial g(x,\lambda)/\partial\lambda|$. Define $p_i(x,\lambda) = \exp(x_i/\lambda)/\sum_{j=1}^n \exp(x_j/\lambda)$. Then we have

$$\frac{\partial g(x,\lambda)}{\partial\lambda} = -\frac{1}{\lambda^2}\left[\left(\sum_{i=1}^n x_i^2 p_i(x,\lambda)\right) - \left(\sum_{i=1}^n x_i p_i(x,\lambda)\right)^2\right] \tag{96}$$

$$= -\frac{1}{2\lambda^2}\sum_{i=1}^n\sum_{j=1}^n (x_i - x_j)^2 p_i(x,\lambda)p_j(x,\lambda) \tag{97}$$

$$= -\frac{1}{\lambda^2}\sum_{1\le i<j\le n} (x_i - x_j)^2 p_i(x,\lambda)p_j(x,\lambda). \tag{98}$$

For any $i \neq j$,

$$\frac{(x_i - x_j)^2 p_i(x,\lambda)p_j(x,\lambda)}{\lambda^2} = \frac{(x_i - x_j)^2\exp(x_i/\lambda)\exp(x_j/\lambda)}{\lambda^2\big(\sum_{i=1}^n \exp(x_i/\lambda)\big)^2} \tag{99}$$

$$\le \frac{(x_i - x_j)^2\exp(x_i/\lambda)\exp(x_j/\lambda)}{\lambda^2\big(\exp(x_i/\lambda) + \exp(x_j/\lambda)\big)^2} \tag{100}$$

$$= \frac{(x_i - x_j)^2\exp((x_i - x_j)/\lambda)}{\lambda^2\big(\exp((x_i - x_j)/\lambda) + 1\big)^2}. \tag{101}$$

WLOG we can assume $x_i > x_j$. Let $z = (x_i - x_j)/\lambda \in (0,\infty)$, define the function $h(z) = \dfrac{z^2\exp(z)}{(1+\exp(z))^2}$. The function $h(z)$ is zero at $z = 0$ and tends to zero as $z \to \infty$. It is continuous and positive for $z > 0$, so it must have a finite maximum. By solving $h'(z) = 0$, one finds that the maximum occurs at $z_0 \approx 2.3994$, and $h(z_0) \approx 0.449$. Therefore, we have

$$\sup_x \big|\partial g(x,\lambda)/\partial\lambda\big| \le \frac{0.439\, n(n-1)}{2} < \frac{n(n-1)}{4}. \tag{102}$$

The conclusion follows. □

Lemma 11 (Banach's fixed point theorem for a sequence of convergent contractions.)**.** *Let (M,d) be a complete metric space. We call $f : M \to M$ a α-contraction in M if and only if there exists $\alpha \in (0,1)$ such that for any $x, y \in M$, $d(f(x), f(y)) < \alpha d(x,y)$. Suppose $\{f_n\}$ is a sequence of α-contractions, f_∞ is a α-contraction in M. Assume that*

1. *$d(f_n(x), f_{n+1}(x)) < C_n$, $n = 1, 2, \ldots$, $\forall x \in M$, such that $\sum_{n=1}^\infty \sum_{k=1}^n C_k \alpha^{n-k} < \infty$.*
2. *$f_n(x) \Rightarrow f_\infty(x)$, $\forall x \in M$.*

Let $x_0 \in M$ be an arbitrary initialization and $x_n = f_n(x_{n-1})$ for $n = 1, 2, \ldots$, then we have $x_n \to x_\infty$, where x_∞ denotes the fixed point of f_∞. Specifically, $d(x_n, x_\infty) = O\big(\sum_{j=n}^\infty \sum_{k=1}^j C_k\alpha^{j-k} + \alpha^n\big)$.

Proof. We have

$$d(x_n, x_{n+1}) = d(f_n(x_{n-1}), f_{n+1}(x_n)) \tag{103}$$

$$\le d(f_n(x_n), f_{n+1}(x_n)) + d(f_n(x_{n-1}), f_n(x_n)) \tag{104}$$

$$\le C_n + \alpha d(x_{n-1}, x_n) \tag{105}$$

$$\le C_n + \alpha\big(C_{n-1} + \alpha d(x_{n-2}, x_{n-1})\big) \tag{106}$$

$$\le \cdots \tag{107}$$

$$\le (C_n + \alpha C_{n-1} + \alpha^2 C_{n-2} + \cdots + \alpha^{n-1} C_1) + \alpha^n d(x_0, x_1) \tag{108}$$

$$= \sum_{k=1}^n C_k \alpha^{n-k} + \alpha^n d(x_0, x_1). \tag{109}$$

Therefore, we have

$$d(x_n, x_{n+N}) \leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \dots d(x_{n+N-1}, x_{n+N}) \tag{110}$$

$$= \sum_{i=1}^{N} d(x_{n+i-1}, x_{n+i}) \tag{111}$$

$$\leq \sum_{i=1}^{N} \sum_{k=1}^{n+i-1} C_k \alpha^{n+i-1-k} + \sum_{i=1}^{N} \alpha^{n+i-1} d(x_0, x_1) \tag{112}$$

$$\leq \sum_{j=n}^{\infty} \sum_{k=1}^{j} C_k \alpha^{j-k} + \sum_{j=n}^{\infty} \alpha^j d(x_0, x_1). \tag{113}$$

Since $\sum_{j=1}^{\infty} \sum_{k=1}^{j} C_k \alpha^{j-k} < \infty$, $\sum_{j=n}^{\infty} \sum_{k=1}^{j} C_k \alpha^{j-k} \to 0$, and $\{x_n\}$ are Cauchy. Therefore, $x_n \to x' \in M$. Now we show $x' = x_\infty$. We have

$$d(x', f_\infty(x')) \leq d(x', x_{n+1}) + d(x_{n+1}, f_\infty(x_n)) + d(f_\infty(x_n), f_\infty(x')). \tag{114}$$

Since $x_n \to x'$, $d(x', x_{n+1}) \to 0$ and $d(f_\infty x_n, f_\infty x_{n+1}) \leq \alpha d(x', x_{n+1}) \to 0$. We also have $d(x_{n+1}, f_\infty x_n) = d(f_n x_n, f_\infty x_n) \to 0$ as $f_n(x) \Rightarrow f_\infty(x)$. Since n is arbitrarily chosen, $x' = f_\infty(x')$ and thus $x' = x_\infty$. □

D Additional Experimental Details

D.1 Practical implementation

The practical implementation of DDVI used in Section 5.2 trains two categorical networks. The auxiliary network ν is updated by a greedy distributional target and supplies both the ε-greedy behavior and the cooled reference policy. The main network η is updated by distributional Bellman evaluation under that policy, so the softmax enters only the Bellman target. Algorithm 2 summarizes the full training procedure.

D.2 Common deep-learning configuration

The default implementation uses the following settings:the reference policy is generated from the auxiliary network ν; $\lambda_{\text{init}} = 1$ unless otherwise specified; $\alpha = 0.5$; $\lambda_{\min} = 10^{-6}$; the main soft policy is used only in the Bellman target; behavior actions are selected by an ϵ-greedy policy based on greedy auxiliary values; the auxiliary learner uses a target network; Atari observations are processed with a Nature-style CNN; vector observations are processed with a two-layer 128–128 MLP; Atari rewards are clipped; and MuJoCo action dimensions are discretized into five bins each.

D.3 Advanced ablations

The following ablations were completed after the main deep comparison and are reported here for completeness rather than as primary paper claims:

A1: reference policy from η. Hopper-v5, three seeds, 500k steps. Double DRL mean greedy evaluation: 370.8 ± 97.2.

A2–A3: temperature exponents. Hopper-v5 with $\alpha = 0.3$ (two seeds) and $\alpha = 0.7$ (two seeds). Mean greedy evaluation: 238.3 ± 152.5 and 291.3 ± 228.9, respectively.

A4: seven action bins. Hopper-v5, two seeds. C51 mean greedy evaluation 210.2 ± 69.1; Double 156.0 ± 170.5.

A5: longer training. Hopper-v5 for one million steps, two seeds. C51 mean greedy evaluation 263.3 ± 5.6; Double 308.3 ± 114.6.

A6: Atari canary. Freeway-v5, three seeds, one million steps. Both methods have near-zero `ret100`; mean greedy evaluation is 7.3 ± 12.6 for C51 and 7.2 ± 12.5 for Double.

Algorithm 2 Practical implementation of DDVI

Input: Categorical networks ν and η; temperature $\lambda_t = \max\{\lambda_{\min}, \lambda_{\text{init}}(t+1)^{-\alpha}\}$ with $\lambda_{\text{init}} = 1$, $\alpha = 0.5$, $\lambda_{\min} = 10^{-6}$; exploration ε_t; replay buffer $\mathcal{D}$; target period K.
Output: Trained networks ν and η.

Initialize the auxiliary target $\nu^- \leftarrow \nu$.
Let Π denote projection onto the fixed categorical support.
for environment step $t = 0, 1, 2, \ldots$ **do**
Observe S_t and read the auxiliary means $Q_\nu(S_t, a) = \mathbb{E}_{Z\sim\nu(S_t,a)}[Z]$.
Select A_t by ε-greedy on these means: with probability $1 - \varepsilon_t$, $A_t \in \arg\max_a Q_\nu(S_t, a)$; otherwise draw A_t uniformly from $\mathcal{A}$.
Step the environment, observe (R_t, S_{t+1}), and store (S_t, A_t, R_t, S_{t+1}) in $\mathcal{D}$.
Sample a minibatch from $\mathcal{D}$.
On each transition, form the reference policy from the current auxiliary means:

$$Q_\nu(s, a) = \mathbb{E}_{Z\sim\nu(s,a)}[Z], \qquad \pi^{(t)}(a \mid s) = \frac{\exp\big(Q_\nu(s, a)/\lambda_t\big)}{\sum_{a'\in\mathcal{A}} \exp\big(Q_\nu(s, a')/\lambda_t\big)}. \tag{115}$$

Auxiliary target: take a greedy action of the target network, $A^\star \in \arg\max_{a'} Q_{\nu^-}(S', a')$, and set

$$\widehat{\nu} = \Pi\big((b_{R,\gamma})_\# \nu^-(S', A^\star)\big). \tag{116}$$

Main target: draw $A' \sim \pi^{(t)}(\cdot \mid S')$ and set

$$\widehat{\eta} = \Pi\big((b_{R,\gamma})_\# \eta(S', A')\big). \tag{117}$$

The softmax $\pi^{(t)}$ is used only in this target.
Descend on the categorical cross-entropy of $\nu(S, A)$ against $\widehat{\nu}$ and of $\eta(S, A)$ against $\widehat{\eta}$.
If t is a multiple of K, copy $\nu^- \leftarrow \nu$.
end for

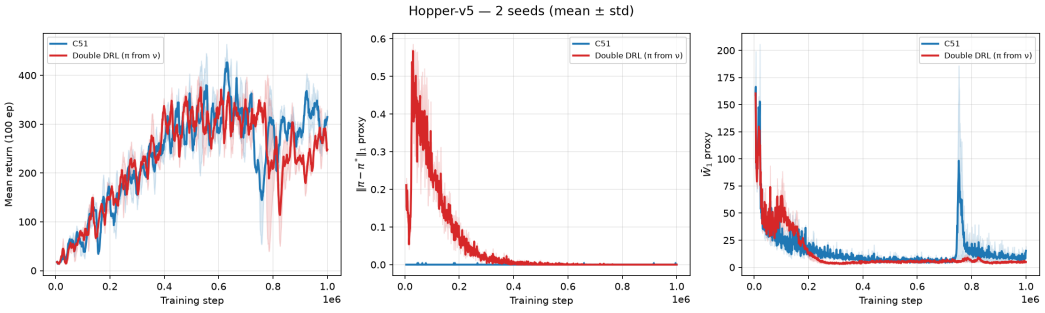


Figure 2: A5 ablation: Hopper-v5 trained for one million steps (two seeds). Double DRL retains a higher mean greedy evaluation than C51 (308.3 vs. 263.3), and its training return is much more stable than in the 500k-step runs (247.9 ± 3.5 vs. 267.8 ± 130.8), although the two-seed comparison remains high variance.

These runs suggest that, over longer Hopper training (A5, Figure 2), Double DRL keeps a higher mean greedy evaluation than C51 with considerably more stable training returns, while finer action discretization (A4) remains difficult under the current categorical setup. On the Freeway canary (Figure 3), neither method learns a strong policy within one million steps, and the occasional nonzero greedy evaluations are driven by single seeds; we treat this environment as a smoke test rather than evidence of a performance difference.

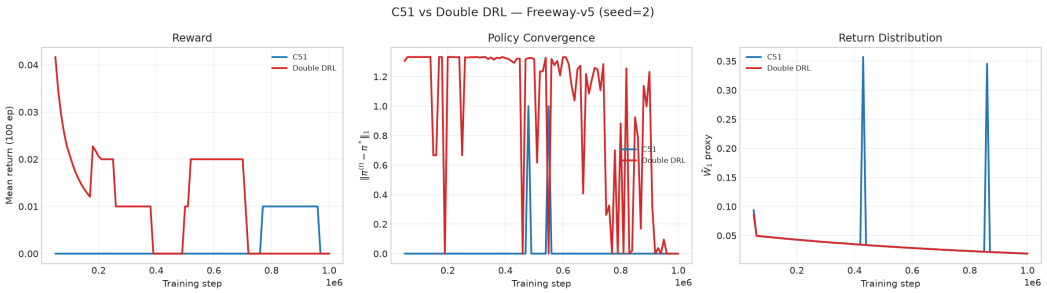


Figure 3: A6 Atari canary on Freeway-v5 (seed 0). Training returns remain near zero for both methods within one million steps, but the greedy evaluation of the best Double DRL seed reaches 21.6 whereas the matched C51 seed stays at 0.

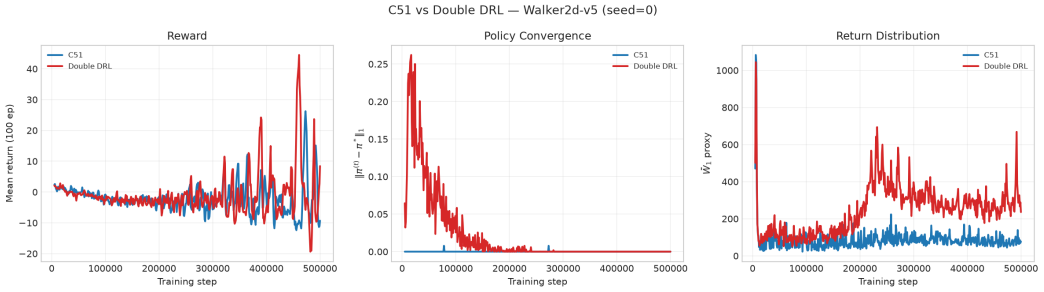


Figure 4: Walker2d-v5 comparison on the completed Double seed. Neither method learns a strong locomotion policy within 500k discretized steps, but the Double run attains a positive greedy evaluation (46.5) while the matched C51 runs stay near zero.

D.4 Supplementary MuJoCo runs

Walker2d-v5 was also evaluated with five-bin action discretization over 500,000 steps. C51 remains near zero on two seeds ($\texttt{ret100} = -10.0 \pm 1.0$, greedy evaluation -8.0 ± 0.3). On the single completed Double seed, $\texttt{ret100}$ is 8.4 and greedy evaluation is 46.5. These results are exploratory and are not included in the main comparison because the matched seed count is incomplete.

We also ran a single 2M-step C51 run on HalfCheetah-v5 ($\texttt{ret100} = -206.8$, greedy evaluation -234.6); the matched Double run has not been completed, so HalfCheetah is omitted from the comparison.

D.5 Temperature Ablation on a 4×4 Gridworld

Section 5.1 fixes the temperature at $\lambda_t = 0.8/t$. The runs below keep that model-based backup and change only the schedule. The grid is built so that an optimal tie changes the return law as well as the policy, which makes both d_π and $\bar{W}_1$ respond to λ_t.

Environment. Figure 5 shows the grid. The discount is $\gamma = 0.95$. The episode starts in the blue cell. The green cell is an absorbing goal: entering it yields reward 1, and every action there stays with reward 0. The hatched cells are obstacles. A move into an obstacle or off the grid stays in place and yields reward 0.

East from the start follows the corridor and reaches the goal in three steps, for a deterministic return γ^2. South enters the orange cell and receives a bonus, equal to $2(\gamma^2 - \gamma^4)$ with probability $1/2$ and equal to 0 otherwise. The mean of this bonus is $\gamma^2 - \gamma^4$, so the optimal action values tie,

$$Q^\star(s_0, \text{east}) = Q^\star(s_0, \text{south}) = \gamma^2 = 0.9025,$$

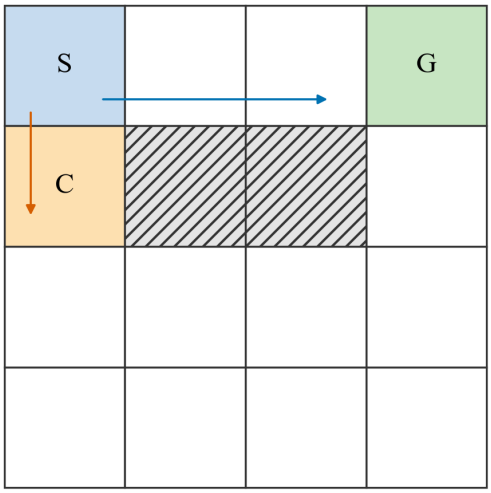


Figure 5: 4×4 gridworld for the temperature ablation. The blue cell is the start and the green cell is the goal. Hatched cells are obstacles. The blue arrow is a deterministic three-step corridor of return γ^2. The orange arrow marks a stochastic bonus whose mean is matched to that corridor, so east and south are both optimal at the start.

while north and west, which bump the boundary, have value γ^3. The gap between optimal and suboptimal actions at the start is $\Delta = \gamma^2(1-\gamma) \approx 0.0451$. The tie-uniform policy puts mass $1/2$ on east and on south. These two actions do not share a return law: east is a point mass at γ^2, and south mixes the bonus into the later return. Cells in the lower part of the grid add further two-action ties along equal-length shortest paths. Those ties are symmetric, so any positive temperature splits them evenly once their means agree. Only the tie at the start changes the return law, so it is the tie that appears in $\bar{W}_1$.

Protocol. Updates follow the synchronous model-based scheme of Section 5.1. Auxiliary means apply the Bellman optimality operator exactly. The reference policy is the softmax of those means at temperature λ_t. The greedy ablation instead places mass one on the lowest-index maximizer, which selects east whenever several actions tie. The main estimator then applies one projected distributional backup under that reference policy. The categorical support is 121 equally spaced atoms on $[0, 2\gamma^2]$, the same atom count as in the main tabular runs. The upper endpoint is the supremum return from collecting the high bonus on every visit to the start. Action values start at zero, and each categorical distribution starts as a point mass at zero. Every schedule runs for 500 iterations.

The reported return is the exact start-state value of the current reference policy, from the linear Bellman equation. Policy distance is the maximum ℓ_1 distance to the tie-uniform optimal policy,

$$d_\pi = \max_s \|\pi(\cdot \mid s) - \pi^\star(\cdot \mid s)\|_1, \tag{118}$$

where the maximum is over nonterminal, non-obstacle cells. Distributional error is the maximum categorical 1-Wasserstein distance between the main estimator and the categorical fixed point of $\pi^\star$, over those cells and every action.

Results. Figure 6 and Table 3 show three regimes.

Greedy tie-breaking reaches the optimal start return 0.9025 by iteration 10 and stays there, while d_π remains 1 and $\bar{W}_1$ levels off at 0.0428. The action it selects is optimal, so expected return does not register the mistake. Table 1 shows the same split for C51 on `gridworld_4x4`, where a deterministic tie-break leaves $d_\pi = 1$. On that environment the tied actions share a return law, and the distributional error is reported as zero. Here the tied actions do not, so the unresolved tie is visible in $\bar{W}_1$ as well.

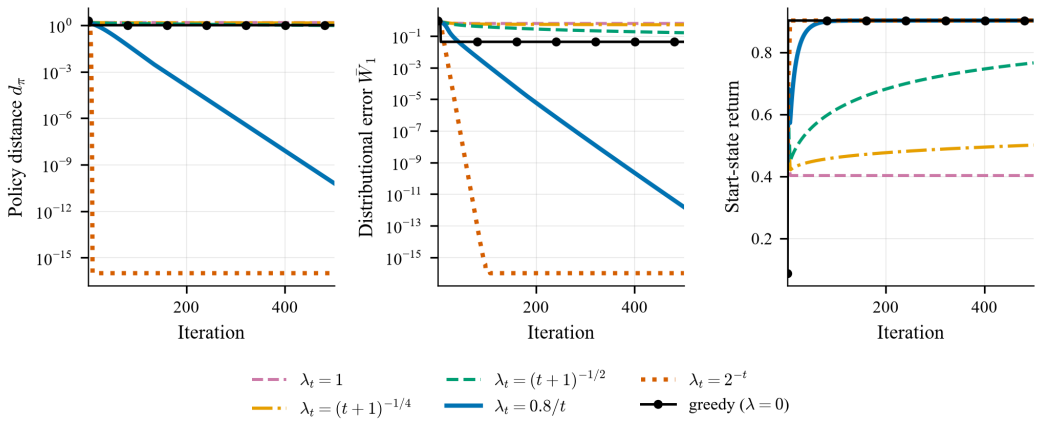


Figure 6: Temperature schedules on the gridworld of Figure 5. All runs use synchronous model-based algorithm for 500 iterations with 121 atoms. Left: policy distance to the tie-uniform optimum. Middle: categorical $\bar{W}_1$ to the return law of that policy. Right: exact start-state return of the reference policy. The dotted line is $V^\star(s_0) = \gamma^2$. Greedy tie-breaking, $\lambda_t = 2^{-t}$, and $\lambda_t = 0.8/t$ all reach this return; only the positive temperatures also drive d_π and $\bar{W}_1$ to zero. Schedules that are still large relative to the suboptimality gap $\Delta \approx 0.045$ remain suboptimal at iteration 500.

Table 3: Gridworld temperature ablation after 500 iterations. Return is the exact start-state value of the reference policy; the optimum is $\gamma^2 = 0.9025$. The schedule $\lambda_t = 0.8/t$ is the one used in Section 5.1.

Schedule	λ_{500}	Return ↑	d_π ↓	$\bar{W}_1$ ↓
greedy ($\lambda = 0$)	0	0.9025	1	0.0428
$\lambda_t = 1$	1	0.4028	1.479	0.618
$\lambda_t = (t+1)^{-1/4}$	0.211	0.5009	1.395	0.513
$\lambda_t = (t+1)^{-1/2}$	0.0447	0.7659	0.958	0.160
$\lambda_t = 0.8/t$	0.0016	0.9025	6.30×10^{-11}	1.53×10^{-12}
$\lambda_t = 2^{-t}$	2^{-500}	0.9025	0	0

Positive temperatures that become small relative to Δ remove the error inside the same horizon. Under $\lambda_t = 0.8/t$, policy distance falls from 0.250 at iteration 50 to 1.87×10^{-2} at iteration 100, and the run ends at $d_\pi = 6.30\times10^{-11}$ and $\bar{W}_1 = 1.53\times10^{-12}$, with return 0.9025. Geometric cooling $\lambda_t = 2^{-t}$ concentrates sooner. By iteration 10 the two optimal means agree, so the softmax splits them evenly at every positive temperature, and the smaller temperature only suppresses north and west earlier. Policy distance is numerically zero by iteration 10. The distributional error lags the policy: $\bar{W}_1$ is 2.03×10^{-8} at iteration 50 and reaches zero before iteration 200. The schedule $\lambda_t = 2^{-t}$ decays faster than γ^t, so it lies outside the sufficient condition $\gamma^t/\lambda_t \to 0$ of Theorem 1. The tie residual on this grid is already zero while the temperature is still positive, and equal means receive equal softmax mass, so the run still reaches $\pi^\star$.

Schedules that stay large relative to Δ are unfinished at iteration 500. Constant $\lambda_t = 1$ ends at return 0.4028, policy distance 1.479, and $\bar{W}_1 = 0.618$. The corollary schedule $\lambda_t = (t+1)^{-1/4}$ is still 0.211 at iteration 500, several times Δ, and ends at return 0.5009, policy distance 1.395, and $\bar{W}_1 = 0.513$. The intermediate choice $\lambda_t = (t+1)^{-1/2}$ reaches $\lambda_{500} = 0.0447$, comparable to Δ, with return 0.7659, policy distance 0.958, and $\bar{W}_1 = 0.160$. Corollary 1 chooses the exponent $1/4$ so that a $\sqrt{(\log t)/t}$ error from sampled Q-learning stays negligible relative to λ_t. On these exact backups that rate has not yet concentrated the softmax on the optimal actions.

The gridworld runs separate the two jobs of the temperature in the planning version of DDVI. A positive temperature equalizes actions whose estimated means already tie; greedy tie-breaking skips

that step and leaves both $d_\pi = 1$ and a nonzero Wasserstein gap. Sending λ_t to zero is what removes actions that are only near-optimal. Within five hundred exact backups, $\lambda_t = 0.8/t$ and $\lambda_t = 2^{-t}$ complete both steps. Holding the temperature fixed, or cooling it only as $(t+1)^{-1/4}$ or $(t+1)^{-1/2}$, does not.