EFFICIENT WASSERSTEIN AND SINKHORN POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Trust-region methods based on Kullback-Leibler divergence are pervasively used to stabilize policy optimization in reinforcement learning. In this paper, we examine two natural extensions of policy optimziation with Wasserstein and Sinkhorn trust regions, namely *Wasserstein policy optimization (WPO)* and *Sinkhorn policy optimization (SPO)*. Instead of restricting the policy to a parametric distribution class, we directly optimize the policy distribution and derive their close-form policy updates based on the Lagrangian duality. Theoretically, we show that WPO guarantees a monotonic performance improvement, and SPO provably converges to WPO as the entropic regularizer diminishes. Experiments across tabular domains and robotic locomotion tasks further demonstrate the performance improvement of both approaches, more robustness of WPO to sample insufficiency, and faster convergence of SPO, over state-of-art policy gradient methods.

1 INTRODUCTION

Policy-based reinforcement learning (RL) approaches have received remarkable success in many domains, including video games (Mnih et al., 2013; Mnih et al., 2015), board games (Silver et al., 2016; Heinrich & Silver, 2016), robotics (Grudic et al., 2003; Gu et al., 2017), and continuous control tasks (Duan et al., 2016; Schulman et al., 2016). One prominent example is policy gradient method (Grudic et al., 2003; Peters & Schaal, 2006; Lillicrap et al., 2016; Sutton et al., 1999; Williams, 1992; Mnih et al., 2016; Silver et al., 2014). The core idea is to represent the policy with a probability distribution $\pi_\theta(a|s) = P[a|s;\theta]$, such that the action a in state s is chosen stochastically following the policy π_θ controlled by parameter θ. Determining the right step size to update the policy is crucial for maintaining the stability of policy gradient methods: too conservative choice of stepsizes result in slow convergence, while too large stepsizes may lead to catastrophically bad updates.

To control the size of policy updates, Kullback-Leibler (KL) divergence is commonly adopted to measure the difference between two policies. For example, the seminal work on trust region policy optimization (TRPO) by Schulman et al. (2015) introduced KL divergence based constraints (trust region constraints) to restrict the size of the policy update; see also Peng et al. (2019); Abdolmaleki et al. (2018). Kakade (2001) and Schulman et al. (2017) introduced a KL-based penalty term to the objective to prevent excessive policy shift.

Though KL-based policy optimization has achieved promising results, it remains interesting whether using other metrics to gauge the similarity between policies could bring additional benefits. Recently, few work (Richemond & Maginnis, 2017; Zhang et al., 2018; Moskovitz et al., 2020; Pacchiano et al., 2020) has explored the Wasserstein metric to restrict the deviation between consecutive policies. Compared with KL divergence, Wasserstein metric has several desirable properties. Firstly, it is a true symmetric distance measure. Secondly, it allows flexible user-defined costs between actions and is less sensitive to ill-posed likelihood ratios. Thirdly but most importantly, Wasserstein metric has a weaker topology (Arjovsky et al., 2017), thus possibly leading to a more robust policy and more stable performance.

However, the challenge of applying Wasserstein metric for policy optimization is also evident: evaluating Wasserstein distance requires solving an optimal transport problem, which could be computationally expensive. To avoid this computation hurdle, existing work resorts to different techniques to *approximate the policy update* under Wasserstein regularization. For example, Richemond & Maginnis (2017) solved the resulting RL problem using Fokker-Planck equations; Zhang et al.

(2018) introduced particle approximation method to estimate the Wasserstein gradient flow. Recently, Moskovitz et al. (2020) instead considered the second-order Taylor expansion of Wasserstein distance based on Wasserstein information matrix to characterize the local behavioral structure of policies. Pacchiano et al. (2020) tackled behavior-guided policy optimization with smooth Wasserstein regularization by solving an approximate dual reformulation defined on reproducing kernel Hilbert spaces. Aside from such approximation, some of these work also limits the policy representation to a particular parametric distribution class, As indicated in Tessler et al. (2019), since parametric distributions are not convex in the distribution space, optimizing over such distributions results in local movements in the action space and thus leads to convergence to a sub-optimal solution. Henceforth, the theoretical performance of policy optimization under Wasserstein metric remains elusive in light of these approximation errors.

In this paper, we study policy optimization with trust regions based on Wasserstein distance and Sinkhorn divergence. The latter is a smooth variant of Waserstein distance by imposing an entropic regularization to the optimal transport problem (Cuturi, 2013). We call them, *Wasserstein Policy Optimization (WPO)* and *Sinkhorn Policy Optimization (SPO)*, respectively. Instead of confining the distribution of policy to a particular distribution class, we work on the space of policy distribution directly, and consider all admissible policies that are within the trust regions with the goal of avoiding approximation errors. Unlike existing work, we focus on *exact characterization* of the policy updates. We highlight our contributions as follows:

1. **Algorithms:** We develop close-form expressions of the policy updates for both WPO and SPO based on the corresponding optimal Lagrangian multipliers of the trust region constraints. In particular, the optimal Lagrangian multiplier of SPO admits a simple form and can be computed efficiently. A practical on-policy actor-critic algorithm is proposed based on the derived expressions of policy updates and advantage value function estimation.
2. **Theory:** We theoretically show that WPO guarantees a *monotonic performance improvement* through the iterations, *even with non-optimal Lagrangian multipliers*, yielding better and more robust guarantee compared to that using KL divergence. Moreover, we prove that SPO converges to WPO as the entropic regularizer diminishes.
3. **Experiments:** A comprehensive evaluation with several types of testing environments including tabular domains and robotic locomotion tasks demonstrates the efficiency and effectiveness of WPO and SPO. Compared to state-of-art policy gradients approaches using KL divergence such as TRPO and PPO, our methods achieve better sample efficiency, faster convergence, and improved final performance. Our numerical study indicates that by properly choosing the weight of the entropic regularizer, SPO achieves a better trade-off between convergence and final performance than WPO.

Related work: Wasserstein-like metrics have only been recently studied in the context of reinforcement learning. The most related work to ours are Richemond & Maginnis (2017); Zhang et al. (2018); Moskovitz et al. (2020); Pacchiano et al. (2020). Albeit for different purposes, these work apply Wasserstein distance as an explicit (and sometimes fixed) penalty function instead of trust-region constraints. Moreover, they use different strategies to approximate the Wasserstein distance. The only work that exploited Sinkhorn divergence in RL, to our best knowledge, is Pacchiano et al. (2020). In addition, few recent work has exploited Wasserstein distance for imitation learning (Xiao et al., 2019; Dadashi et al., 2021).

2 Background and Notations

Markov Decision Process (MDP): We consider an infinite-horizon discounted MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}$ is the distribution of the initial state s_0, and γ is the discount factor. We define the return of timestep t as the accumulated discounted reward from t, $R_t = \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k})$, and the performance of a stochastic policy π as $J(\pi) = \mathbb{E}_{s_0,a_0,s_1\ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$ where $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$. As shown in Kakade & Langford (2002), the expected return of a new policy π' can be expressed in terms of the advantage over the old policy π: $J(\pi') = J(\pi) + \mathbb{E}_{s\sim\rho_v^{\pi'}, a\sim\pi'}[A^\pi(s, a)]$, where $A^\pi(s, a) = \mathbb{E}[R_t|s_t = s, a_t = a; \pi] - \mathbb{E}[R_t|s_t = s; \pi]$ represents the advantage function and ρ_v^π represents

the unnormalized discounted visitation frequencies with initial state distribution υ, i.e., $\rho_\upsilon^\pi(s) = \mathbb{E}_{s_0\sim\upsilon}[\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0)]$.

Trust Region Policy Optimization (TRPO): In TRPO (Schulman et al., 2015), the policy π is parameterized as π_θ with parameter vector θ. For notation brevity, we use θ to represent the policy π_θ. Then, the new policy θ' is found in each iteration to maximize the expected improvement $J(\pi') - J(\pi)$, or equivalently, the expected value of the advantage function:

$$\begin{aligned} &\max_{\theta'} \ \ \mathbb{E}_{s\sim\rho_\upsilon^\theta, a\sim\theta'}[A^\theta(s,a)] \\ &\text{s.t. } \mathbb{E}_{s\sim\rho_\upsilon^\theta}[d_{KL}(\theta',\theta)] \le \delta, \end{aligned} \tag{1}$$

where d_{KL} represents the KL divergence and δ is the threshold of the distance between the new and the old policies. Note that here the expected value of the advantage function is an estimation as the visitation frequency ρ_υ^θ is used rather than $\rho_\upsilon^{\theta'}$, which means the changes in state visitation frequencies caused by the changes in policy are ignored.

Wasserstein Distance: Given two probability distributions of policies π and π' on the discrete action space $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$, the Wasserstein distance between the policies is defined as:

$$d_W(\pi', \pi) = \inf_{Q\in\Pi(\pi',\pi)} \langle Q, M\rangle, \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product. The infimum is taken over all joint distributions Q with marginals π' and π, and M is the cost matrix with elements $M_{ij} = d(a_i, a_j)$, where $d(a_i, a_j)$ is defined as the distance between actions a_i and a_j.

Sinkhorn Divergence: Sinkhorn divergence (Cuturi, 2013) provides a smooth approximation of the Wasserstein distance by adding an entropic regularizer. The Sinkhorn divergence is defined as follows:

$$d_S(\pi', \pi|\lambda) = \inf_{Q\in\Pi(\pi',\pi)} \{\langle Q, M\rangle - \frac{1}{\lambda}h(Q)\}, \tag{3}$$

where $h(Q) = -\sum_{i=1}^N\sum_{j=1}^N Q_{ij}\log Q_{ij}$ represents the entropy term, and $\lambda > 0$ is a regularization parameter. Similarly, we use Q^s to denote the joint distribution of $\pi(\cdot|s)$ and $\pi'(\cdot|s)$ with $\sum_{i=1}^N Q_{ij}^s = \pi(a_j|s)$ and $\sum_{j=1}^N Q_{ij}^s = \pi'(a_i|s)$. The intuition of adding the entropic regularization is: since most elements of the optimal joint distribution Q will be 0 with a high probability, by trading the sparsity with entropy, a smoother and denser coupling between distributions can be achieved (Courty et al., 2014; 2016). Therefore, when the weight of the entropic regularization decreases (i.e., λ increases), the sparsity of the divergence increases, and the Sinkhorn divergence converges to the Wasserstein metric, i.e., $\lim_{\lambda\to\infty} d_S(\pi', \pi|\lambda) = d_W(\pi', \pi)$. More critically, Sinkhorn divergence is useful to mitigate the computational burden of computing Wasserstein distance. In fact, the efficiency improvement that Sinkhorn divergence and the related algorithms brought paves the way to utilize Wassersterin-like metrics in many machine learning domains, including online learning (Cesa-Bianchi & Lugosi, 2006), model selection (Juditsky et al., 2008; Rigollet & Tsybakov, 2011), generative modeling (Genevay et al., 2018; Petzka et al., 2017; Patrini et al., 2019), dimensionality reduction (Huang et al., 2021; Lin et al., 2020; Wang et al., 2021), etc.

3 Wasserstein Policy Optimization

Motivated by TRPO, here we consider a trust region based on Wassersteim metric. Moreover, we lift the restrictive assumption that a policy has to follow a parametric distribution class by allowing all admissible policies. Then, the new policy π' is found in each iteration to maximize the estimated expected value of the advantage function. Therefore, the *Wasserstein Policy Optimization* (WPO) framework is shown as follows:

$$\begin{aligned} &\max_{\pi'\in\mathcal{D}} \quad \mathbb{E}_{s\sim\rho_\upsilon^\pi, a\sim\pi'(\cdot|s)}[A^\pi(s,a)] \\ &\text{where} \quad \mathcal{D} = \{\pi' | \mathbb{E}_{s\sim\rho_\upsilon^\pi}[d_W(\pi'(\cdot|s), \pi(\cdot|s))] \le \delta\}, \end{aligned} \tag{4}$$

where the Wasserstein distance $d_W(\cdot,\cdot)$ is defined in (2).

Before describing our main result, we adopt a mild and conventional assumption that generally holds true in most practical cases:

Assumption 1. *Assume $A^\pi(s,a)$ is bounded, i.e.,* $\sup_{a\in\mathcal{A},s\in\mathcal{S}}|A^\pi(s,a)| \le A^{max}$ *for some* $A^{max} > 0$.

With Wasserstein metric based trust region constraint, we are able to derive the closed-form of the policy update shown in Theorem 1. The main idea is to form the Lagrangian of the constrained optimization problem presented above, and the detailed proof can be found in Appendix A.

Theorem 1. ***(Closed-form policy update)*** *Let $k_s^\pi(\beta,j) = argmax_{k=1\dots N}\{A^\pi(s,a_k) - \beta M_{kj}\}$. If Assumption 1 holds, then an optimal solution to the WPO problem in (4) is given by:*

$$\pi^*(a_i|s) = \sum_{j=1}^{N} \pi(a_j|s) f_s^*(i,j), \tag{5}$$

where $f_s^*(i,j) = 1$ if $i = k_s^\pi(\beta^*,j)$ and $f_s^*(i,j) = 0$ otherwise, and β^* is an optimal Lagrangian multipler corresponds to the following dual formulation:

$$\min_{\beta\ge 0} F(\beta) = \min_{\beta\ge 0}\left\{\beta\delta + \mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\right\}. \tag{6}$$

Moreover, we have $\beta^* \le \bar{\beta} := \max_{s\in\mathcal{S},k,j=1\dots N,k\ne j}(M_{kj})^{-1}(A^\pi(s,a_k) - A^\pi(s,a_j))$.

The exact policy update for WPO in (5) requires computing the optimal Lagrangian multiplier β^* by solving the one-dimensional subproblem (6). A closed form of β^* is not easy to obtain in general, except for special cases of the distance $d(x,y)$ or cost matrix M. In Appendix G, we provide the closed form of β^* for the case when $d(x,y) = 0$ if $x = y$ and 1 otherwise.

WPO Policy Update: Based on Theorem 1, we introduce the following WPO policy updating rule:

$$\pi_{t+1}(a_i|s) = \mathbb{F}^{\text{WPO}}(\pi_t) := \sum_{j=1}^{N}\pi_t(a_j|s) f_s^t(i,j), \tag{WPO}$$

where we choose an arbitrary $k_s^{\pi_t}(\beta_t,j) \in \text{argmax}_{k=1,\dots,N}\{A^{\pi_t}(s,a_k) - \beta_t M_{kj}\}$ and set $f_s^t(k_s^{\pi_t}(\beta_t,j),j) = 1$ and other entries to be 0.

Note that different from (5), we allow β_t to be chosen arbitrarily and time dependently. We show that such policy update always leads to a monotonic improvement of the performance even when β_t is not the optimal Lagrangian multiplier. In particular, we propose two efficient strategies to update the multiplier β_t:

(i) Time-dependent β_t: To improve the computational efficiency, we can simply treat β_t as a time-dependent control parameter, e.g., we can choose β_t to be a diminishing sequence.

(ii) Approximation of optimal β_t: To improve the convergence, we can approximately solve the optimal Lagrangian multiplier based on Sinkhorn divergence. We will discuss this in more detail in Section 4.

Next, we provide theoretical justification that WPO policy update is always guaranteed to improve the true performance J monotonically if we have access to the true advantage function. If the advantage function can only be evaluated inexactly with limited samples, then an extra estimation error will incur. The detailed proof can be found in Appendix B.

Theorem 2. ***(Performance improvement)*** *For any initial state distribution μ and any $\beta_t \ge 0$, if $||\hat{A}^\pi - A^\pi||_\infty \le \epsilon$ for some $\epsilon > 0$, the WPO policy update with the inaccurate advantage function $\hat{A}^\pi$, guarantees the following performance improvement bound,*

$$J(\pi_{t+1}) \ge J(\pi_t) + \beta_t \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}}\sum\nolimits_{j=1}^{N}\pi_t(a_j|s) M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} - \frac{2\epsilon}{1-\gamma}. \tag{7}$$

Note that when the estimation error $\epsilon = 0$, we have a monotonic performance improvement $J(\pi_{t+1}) \ge J(\pi_t)$ for any $\beta_t \ge 0$. If the second term in (7) is non-zero, then we have a strict monotonic improvement. Compared to the performance bound when using KL-based trust region (see, e.g., Schulman et al. (2015); Cen et al. (2020)), using the Wasserstein metric yields a tighter performance improvement bound and is more robust to the choice of parameters β_t.

4 SINKHORN POLICY OPTIMIZATION

In this section, we introduce Sinkhorn policy optimization (SPO) by constructing trust region with Sinkhorn divergence. Invoking the definition of Sinkhorn divergence in (3), the trust region constrained problem with Sinkhorn divergence can be reformulated as:

$$\max_{Q} \quad \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}[\sum\nolimits_{i=1}^{N} A^{\pi}(s,a_i)\sum\nolimits_{j=1}^{N} Q_{ij}^{s}] \tag{8a}$$

$$s.t. \quad \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}[\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N} M_{ij}Q_{ij}^{s} + \frac{1}{\lambda}Q_{ij}^{s}\log Q_{ij}^{s}] \leq \delta \tag{8b}$$

$$\sum\nolimits_{i=1}^{N} Q_{ij}^{s} = \pi(a_j|s), \qquad \forall j = 1,\ldots,N, s\in\mathcal{S}. \tag{8c}$$

In the following theorem, we derive the optimal policy update in each step when using Sinkhorn divergence based trust region. The proof follows a similar procedure as the Wasserstein policy optimization framework by forming the Lagrangian of the above constrained optimization problem (8). Details are provided in Appendix C.

Theorem 3. *If Assumption 1 holds, then the optimal solution to the trust region constrained problem (4) with Sinkhorn divergence is:*

$$\pi_{\lambda}^{*}(a_i|s) = \sum_{j=1}^{N} \frac{\exp\left(\frac{\lambda}{\beta_{\lambda}^{*}}A^{\pi}(s,a_i) - \lambda M_{ij}\right)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda}{\beta_{\lambda}^{*}}A^{\pi}(s,a_k) - \lambda M_{kj}\right)}\pi(a_j|s), \tag{9}$$

where β_{λ}^{*} is an optimal solution to the following dual formulation:

$$\min_{\beta\geq 0}\ F_{\lambda}(\beta) = \min_{\beta\geq 0}\Big\{\beta\delta - \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\sum\nolimits_{j=1}^{N}\pi(a_j|s)(\tfrac{\beta}{\lambda} + \tfrac{\beta}{\lambda}\ln(\pi(a_j|s)) -$$

$$\tfrac{\beta}{\lambda}\ln[\sum\nolimits_{i=1}^{N}\exp\left(\tfrac{\lambda}{\beta}A^{\pi}(s,a_i) - \lambda M_{ij}\right)]) + \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\tfrac{\beta}{\lambda}\tfrac{\exp\left(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij}\right)\cdot\pi(a_j|s)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj}\right)}\Big\}. \tag{10}$$

Moreover, we have $\beta_{\lambda}^{*} \leq \frac{2A^{max}}{\delta}$.

In contrast to the Wasserstein dual formulation (6), the objective in the Sinkhorn dual formulation (10) is differentiable in β and admits closed-form gradients (shown in Appendix E). With this gradient information, we can use gradient-based global optimization algorithms (Wales & Doye, 1998; Zhan et al., 2006; Leary, 2000) to find a global optimal solution β_{λ}^{*} to (10).

Next, we show that if the entropic regularization parameter λ is large enough, then the optimal solution β_{λ}^{*} is a close approximation to the optimal solution β^{*} to the Wasserstein dual formulation. The detailed proof is provided in Appendix F.

Theorem 4. *Define $\beta_{UB} = \max\{\frac{2A^{max}}{\delta}, \bar{\beta}\}$. The following holds:*

1. *$F_{\lambda}(\beta)$ converges to $F(\beta)$ uniformly on $[0, \beta_{UB}]$,*
2. $\lim\limits_{\lambda\to\infty}\ \arg\min\limits_{0\leq\beta\leq\beta_{UB}} F_{\lambda}(\beta) \subseteq \arg\min\limits_{0\leq\beta\leq\beta_{UB}} F(\beta)$.

Although it is difficult to obtain the exact value of the optimal solution β^{*} to the Wasserstein dual formulation (6), the above theorem suggests that we can approximate β^{*} via β_{λ}^{*} by setting up a relative large λ. In practice, we can also adopt a smooth homotopy approach by setting an increasing sequence λ_t for each iteration and letting $\lambda_t \to \infty$.

SPO Policy Update: Based on Theorem 3, we introduce the following SPO policy updating rule:

$$\pi_{t+1}(a_i|s) = \mathbb{F}^{\text{SPO}}(\pi_t) := \sum_{j=1}^{N} \frac{\exp\left(\frac{\lambda_t}{\beta_t}A^{\pi_t}(s,a_i) - \lambda_t M_{ij}\right)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda_t}{\beta_t}A^{\pi_t}(s,a_k) - \lambda_t M_{kj}\right)}\pi_t(a_j|s), \tag{SPO}$$

Here $\lambda_t \geq 0$ and $\beta_t \geq 0$ are some control parameters. The parameter β_t can be either computed via solving the one-dimensional subproblem (10) or simply set as a diminishing sequence.

5 A PRACTICAL ALGORITHM

In practice, the advantage value functions are often estimated from sampled trajectories. In this section, we provide a practical on-policy actor-critic algorithm, described in Algorithm 1, that combines WPO/SPO with advantage function estimation.

Algorithm 1: Practical on-policy WPO/SPO algorithm

Input: number of iterations K, learning rate α
Initialize policy π_0 and value network V_{ψ_0} with random parameter ψ_0
for $k = 0, 1, 2 \ldots K$ **do**
  Collect a set of trajectories $\mathcal{D}_k$ on policy π_k
  For each timestep t in each trajectory, compute total returns G_t and estimate advantages $\hat{A}_t^{\pi_k}$
  Update value:
$$\psi_{k+1} \leftarrow \psi_k - \alpha \nabla_{\psi_k} \sum (G_t - V_{\psi_k}(s_t))^2$$
  Update policy:
$$\pi_{k+1} \leftarrow \mathbb{F}(\pi_k) \text{ according to WPO or SPO with } \hat{A}^{\pi_k}$$
end

In each iteration of Algorithm 1, the first step is to collect trajectories, which can be either complete or partial. The difference is whether the return is considered thoroughly to the end of a planning horizon. If the trajectory is complete, the total return can be directly expressed as the accumulated discounted rewards $R_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k}$. If the trajectory is partial, it can be estimated by applying the multi-step temporal difference (TD) methods (De Asis et al., 2017): $\hat{R}_{t:t+n} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n V(s_{t+n})$. Then for the advantage estimation, we can use Monte Carlo advantage estimation, i.e., $\hat{A}_t^{\pi_k} = R_t - V_{\psi_k}(s_t)$ or Generalized Advantage Estimation (GAE) (Schulman et al., 2016), which provides a more explicit control over the bias-variance trade-off. In the value update step, we use a neural net to represent the value function, where ψ is the parameter that specifies the value net $s \rightarrow V(s)$. Then, we can update ψ by using gradient descent, which significantly reduces the computational burden of computing advantage directly.

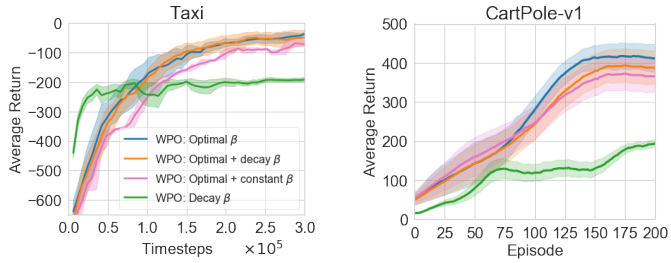
6 EXPERIMENTS

In this section, we evaluate the proposed WPO and SPO approaches on tabular domains and robotic locomotion tasks as presented in Algorithm 1. We compare the performance of our proposed methods with several benchmarks, including TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), and A2C (Mnih et al., 2016)[1]. We compare with A2C because it is similar to our framework in the sense that both of them are simple on-policy actor-critic methods that utilize the advantage information to perform policy updates. For environments with a discrete state space (e.g., tabular domains), policy updates are performed for all states at each iteration. For environments with a continuous state space, a random subset of states is sampled at each iteration to perform policy updates.

6.1 ABLATION STUDY

In this experiment, we first examine the sensitivity of WPO in terms of different strategies of β_t. We test four settings of β value for WPO policy update: (1) Setting 1: computing optimal β value for all policy update; (2) Setting 2: computing optimal β value for first 20% of policy updates and decaying β for the remaining; (3) Setting 3: computing optimal β value for first 20% of policy updates and fix β as its last updated value for the remaining; (4) Setting 4: decaying β for all policy updates (e.g., $\beta_t = \frac{1}{t^2}$). In particular, Setting 3 is rooted in the observation that β^* does not change significantly throughout all the policy updates, especially in the later stage in the experiments carried out in the paper. Small perturbations are added to the approximate values to avoid any stagnation in updating. Taxi task (Dietterich, 1998) from tabular domain and CartPole task (Barto et al., 1983) from locomotion domain are selected for this experiment. The performance comparisons and average run times are shown in Figure 1.

[1]We use the implementations of TRPO, PPO and A2C from OpenAI Baselines (Dhariwal et al., 2017) for MuJuCo tasks and Stable Baselines (Hill et al., 2018) for other tasks.



Run time	Taxi (min)	CartPole
Setting 1	20.4	130s
Setting 2	10.8	63s
Setting 3	10.5	67s
Setting 4	8.7	44s

Figure 1: Episode rewards and run time during the training process for the Taxi and the CartPole tasks, averaged across 3 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

Figure 1 clearly indicates a tradeoff between computation efficiency and accuracy in terms of different choices of β value. Setting 2 is the most effective way to balance the tradeoff between the performance and run time. In the rest of the experiments, we adopt this setting for both WPO and SPO.

We compare WPO with SPO under different λ values on the Chain (Dearden et al., 1998) and Taxi tasks. As shown in Figure 2, SPO converges faster than WPO. With more weight on the entropic regularization of Sinkhorn divergence (i.e., smaller λ), SPO can speed up its convergence more; while as λ increases, the convergence becomes slower but the final performance of SPO improves and becomes closer to the final performance of WPO, which verifies the convergence property of Sinkhorn to Wasserstein distance shown in Theorem 4. Therefore, the choice of λ can effectively adjust the trade-off between convergence and final performance. With a proper λ choice, SPO is able to attain a faster convergence speed with an optimum that is only slightly lower than WPO.

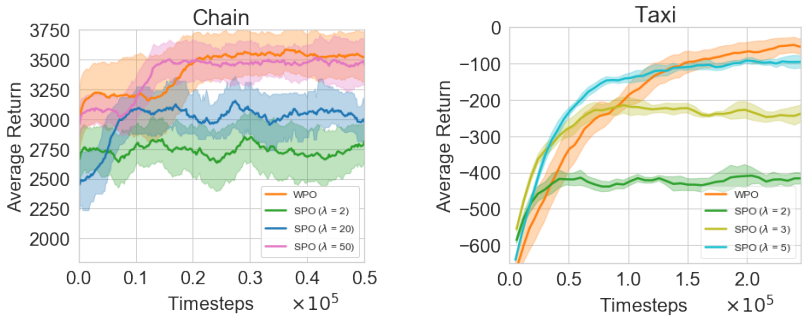


Figure 2: Episode rewards during the training process for the tabular domain tasks, averaged across 3 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

6.2 Tabular Domains

We evaluate WPO and SPO on a set of tasks including Taxi, Chain, and Cliff Walking (Sutton & Barto, 2018), which are intentionally designed to test the exploration ability of the algorithms. The tabular domain has a special environment structure with a discrete state space and a discrete action space. Thus, we use an array of size $|\mathcal{S}| \times |\mathcal{A}|$ to represent the policy $\pi(a|s)$. For the value function, we use a neural net to smoothly update the values. The performance of WPO and SPO are compared to the performance of TRPO, PPO and A2C under the same neural net structure. Each algorithm is evaluated 5 times with a random initialization. Results are reported in Table 1 and Figure 3. The setting of hyperparamaters and network sizes of our algorithms and additional results are provided in Appendix H.

As shown in Figure 3, the performances of WPO, SPO and TRPO are manifestly better than A2C and PPO. Between the trust region based methods, WPO and SPO outperform TRPO in most tasks, except in Chain, where the performances of these three methods are not significantly different. In Taxi and Cliff Walking, SPO converges to the optimum the fastest, while in Taxi, WPO converges to the best optimum, among all methods. We further analyze the performance of the trained agent for

each algorithm on the Taxi environment. As shown in Table 1, WPO has a higher successful drop-off rate and a lower task completion time while the original TRPO reaches the time limit with a drop-off rate 0. Therefore, the results in Taxi show that WPO finds a better policy than the original TRPO.

Table 1: Trained agents performance on Taxi (averaged over 1000 episodes)

	Successful Dropoff (+20)	Illegal Pickup/ Dropoff (-10)	Timesteps (-1)	Average Return
WPO	0.753	0.232	70.891	-58.151
TRPO	0	0	200	-200

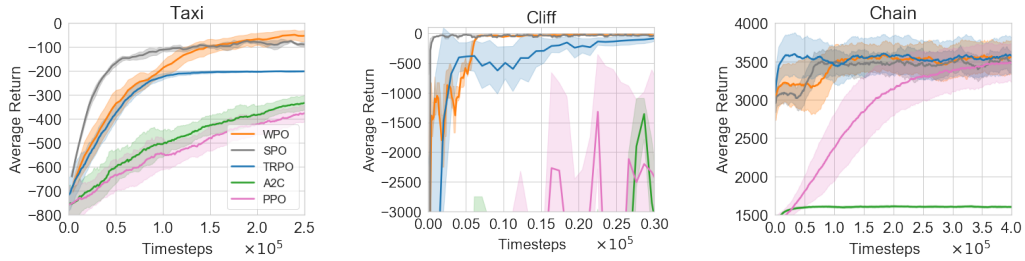
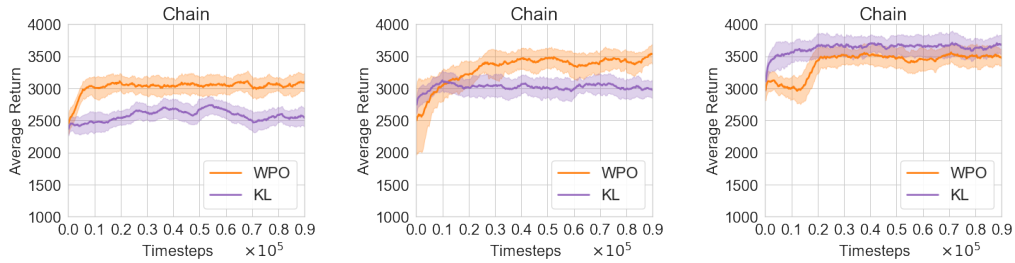


Figure 3: Episode rewards during the training process for the tabular domain tasks, averaged across 5 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

We also show that compared with the KL divergence, which is used in traditional TRPO and PPO approaches, the utilization of Wassersteim metric can cope with the inaccurate advantage function estimations caused by the lack of samples. We compare WPO with KL (Algorithm 1 framework with KL based policy update derived in Peng et al. (2019)) on the Chain task. We evaluate the performance of these two algorithms under different N_A, which denotes the number of samples used to estimate the advantage function at each iteration. As shown in Figure 4, when N_A is 1000, KL performs slightly better than WPO. However, when N_A decreases to 100 or 250, WPO outperforms KL. These results indicate that WPO is more robust than KL under inaccurate advantage values.



(a) $N_A = 100$ (b) $N_A = 250$ (c) $N_A = 1000$

Figure 4: Episode rewards during the training process for the Chain task, averaged across 3 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

6.3 Robotic Locomotion Tasks

We then integrate deep neural network architecture into MPO and SPO and evaluate their performance on several discrete locomotion tasks, including CartPole and Acrobot (Geramifard et al., 2015). We use two separate neural nets to represent the policy and the value. The policy neural net receives state s as an input and outputs the categorical distribution of $\pi(a|s)$. The performance of WPO and SPO are compared to that of TRPO, PPO and A2C under the same neural net structure. We run each algorithm 5 times with a random initialization.

Final Performance: Figure 5 shows the episode rewards during training process for WPO, SPO and baseline algorithms. As seen in Figure 5, WPO and SPO outperform TRPO, PPO and A2C in most tasks in terms of final performance, except in Acrobot where PPO performs the best. In most cases, SPO converges faster but WPO has a better final performance.

Training Time: To train 10^5 timesteps in the discrete locomotion tasks, the training wall-clock time is around 63s for WPO, 65s for SPO, 59s for TRPO and 70s for PPO. Therefore, WPO has a similar computational efficiency as TRPO and PPO.

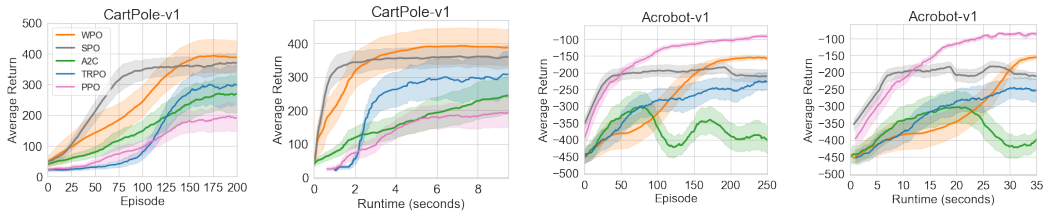
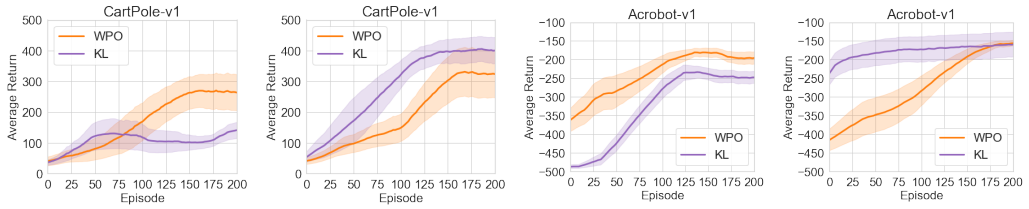


Figure 5: Episode rewards during the training process for the locomotion tasks, averaged across 5 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

The performances of WPO and KL are also compared for the discrete locomotion tasks under different N_A. As shown in Figure 6, when N_A is 500, KL performs better than WPO. However, when N_A decreases to 100, WPO significantly outperforms KL. These results indicate that for discrete locomotion tasks, WPO is more robust than KL when advantage values are inaccurate.



(a) $N_A = 100$ (b) $N_A = 500$ (c) $N_A = 100$ (d) $N_A = 500$

Figure 6: Episode rewards during the training process for the locomotion tasks, averaged across 3 runs with a random initialization. The shaded area depicts the mean $\pm$ the standard deviation.

7 Discussion and Future Work

In this paper, we present two policy optimization frameworks, WPO and SPO, which can exactly characterize the policy updates instead of confining their distributions to particular distribution class or requiring any approximation. Our policy optimization methods outperform TRPO and PPO with better sample efficiency, faster convergence, and improved final performance. Our numerical results show that Wassersteim metric is more robust to the ambiguity of advantage functions, as compare with the KL divergence that is used in most literature. Our strategy for adjusting β value for WPO can reduce the computational time and boost the convergence without noticeable performance degradation. SPO can further improve the convergence speed of WPO by properly choosing the weight of the entropic regularizer.

There are many interesting directions we plan to explore in the future: First, the idea of exact characterization of policies can also be applied to PPO and natural policy gradient, which penalize the policy update instead of imposing trust region constraint. Second, both WPO and SPO are theoretically grounded methods, yet at the same time, on-policy actor-critic algorithms. One focus for the future work will be on integrating WPO and SPO with other well-known off-policy frameworks, such as Soft Actor Critic (Haarnoja et al., 2018) and Twin Delayed DDPG (Fujimoto et al., 2018). Last, extending to continuous control space and exploring the convergence rate, are directions that will also be considered in future research.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *ArXiv Preprint*, pp. arXiv:1806.06920, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *ArXiv Preprint*, pp. arXiv:1701.07875, 2017.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *ArXiv Preprint*, pp. arXiv:2007.06558, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–289. Springer, 2014.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. In *Proceedings of 9th International Conference on Learning Representations*, 2021.

Kristopher De Asis, J. Fernando Hernandez-Garcia, G. Zacharias Holland, and Richard S. Sutton. Multi-step reinforcement learning: A unifying algorithm. *ArXiv Preprint*, pp. arXiv:1703.01327, 2017.

Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference*, pp. 761–768, 1998.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. OpenAI baselines. `https://github.com/openai/baselines`, 2017.

Thomas G. Dietterich. The MAXQ method for hierarchical reinforcement learning. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 118–126, 1998.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1329–1338, 2016.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1582–1591, 2018.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Alborz Geramifard, Christoph Dann, Robert H. Klein, William Dabney, and Jonathan P. How. RLPy: A value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research*, 16(46):1573–1578, 2015.

Gregory Z. Grudic, Vijay Kumar, and Lyle H. Ungar. Using policy gradient reinforcement learning on autonomous robot controllers. In *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 406–411, 2003.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation*, pp. 3389–3396, 2017.

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*, pp. 1861–1870, 2018.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *ArXiv Preprint*, pp. arXiv:1603.01121, 2016.

Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Stable baselines. `https://github.com/hill-a/stable-baselines`, 2018.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A Riemannian block coordinate descent method for computing the projection robust Wasserstein distance. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4446–4455, 2021.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Robert Leary. Global optimization on funneling landscapes. *Journal of Global Optimization*, 18: 367–383, 2000.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael I Jordan. Projection robust Wasserstein distance and Riemannian optimization. *ArXiv Preprint*, pp. arXiv:2006.07458, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *ArXiv Preprint*, pp. arXiv:1312.5602, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1928–1937, 2016.

Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient Wasserstein natural gradients for reinforcement learning. *ArXiv Preprint*, pp. arXiv:2010.05380, 2020.

Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Krzysztof Choromanski, Anna Choromanska, and Michael Jordan. Learning to score behaviors for guided policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7445–7454, 2020.

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pp. 733–743, 2019.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *ArXiv Preprint*, pp. arXiv:1910.00177, 2019.

Jan Peters and Stefan Schaal. Policy gradient methods for robotics. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2219–2225, 2006.

Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of Wasserstein GANs. *ArXiv Preprint*, pp. arXiv:1709.08894, 2017.

Pierre H. Richemond and Brendan Maginnis. On Wasserstein reinforcement learning and the Fokker-Planck equation. *ArXiv Preprint*, pp. arXiv:1712.07185, 2017.

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Springer, 1998.

Johannes O. Royset. Approximations and solution estimates in optimization. *Mathematical Programming*, 170:479–506, 2018.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv Preprint*, pp. arXiv:1707.06347, 2017.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395, 2014.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 1999.

Chen Tessler, Guy Tennenholtz, and Shie Mannor. Distributional policy optimization: An alternative approach for continuous control. In *Advances in Neural Information Processing Systems*, pp. 1350–1360, 2019.

David Wales and Jonathan Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1998.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test using projected Wasserstein distance. In *Proceedings of IEEE International Symposium on Information Theory*, volume 21, 2021.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *ArXiv Preprint*, pp. arXiv:1906.08113, 2019.

Lixin Zhan, Jeff Chen, and Wing-Ki Liu. Monte Carlo basin paving: An improved global optimization method. *Physical Review E*, 73:015701, 2006.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as Wasserstein gradient flows. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5737–5746, 2018.

A PROOF OF THEOREM 1

Theorem 1. ***(Closed-form policy update)*** *Let $k_s^\pi(\beta, j) = argmax_{k=1\dots N}\{A^\pi(s, a_k) - \beta M_{kj}\}$. If Assumption 1 holds, then an optimal solution to the WPO problem in (4) is given by:*

$$\pi^*(a_i|s) = \sum_{j=1}^{N} \pi(a_j|s) f_s^*(i, j), \tag{5}$$

where $f_s^*(i, j) = 1$ if $i = k_s^\pi(\beta^*, j)$ and $f_s^*(i, j) = 0$ otherwise, and β^* is an optimal Lagrangian multipler corresponds to the following dual formulation:

$$\min_{\beta \geq 0} \; F(\beta) = \min_{\beta \geq 0} \left\{ \beta\delta + \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s, a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}] \right\}. \tag{6}$$

Moreover, we have $\beta^* \leq \bar{\beta} := \max_{s\in\mathcal{S},k,j=1\dots N,k\neq j} (M_{kj})^{-1}(A^\pi(s, a_k) - A^\pi(s, a_j))$.

Proof of Theorem 1. First, we denote Q^s as the joint distribution of $\pi(\cdot|s)$ and $\pi'(\cdot|s)$ with $\sum_{i=1}^N Q_{ij}^s = \pi(a_j|s)$ and $\sum_{j=1}^N Q_{ij}^s = \pi'(a_i|s)$. Also, let $f_s(i, j)$ represent the conditional distribution of $\pi'(a_i|s)$ under $\pi(a_j|s)$. Then $Q_{ij}^s = \pi(a_j|s) f_s(i, j)$, $\pi'(a_i|s) = \sum_{j=1}^N Q_{ij}^s = \sum_{j=1}^N \pi(a_j|s) f_s(i, j)$. In addition:

$$\begin{aligned} d_W(\pi'(\cdot|s), \pi(\cdot|s)) \quad &= \min_{Q_{ij}^s} \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij} Q_{ij}^s = \min_{f_s(i,j)} \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}\pi(a_j|s) f_s(i, j), \text{ and} \\ \mathbb{E}_{a\sim\pi'(\cdot|s)}[A^\pi(s, a)] \quad &= \sum_{i=1}^{N} A^\pi(s, a_i)\pi'(a_i|s) = \sum_{i=1}^{N}\sum_{j=1}^{N} A^\pi(s, a_i)\pi(a_j|s) f_s(i, j). \end{aligned}$$

Thus, the WPO problem in (4) can be reformulated as:

$$\max_{f_s(i,j)\geq 0} \quad \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^{N}\sum_{j=1}^{N} A^\pi(s, a_i)\pi(a_j|s) f_s(i, j) \tag{11a}$$

$$s.t. \quad \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}\pi(a_j|s) f_s(i, j) \leq \delta, \tag{11b}$$

$$\sum_{i=1}^{N} f_s(i, j) = 1, \qquad \forall s \in \mathcal{S}, j = 1\dots N. \tag{11c}$$

Note here that (11b) is equivalent to $\mathbb{E}_{s\sim\rho_\upsilon^\pi} \min_{f_s(i,j)} \sum_{i=1}^N \sum_{j=1}^N M_{ij}\pi(a_j|s) f_s(i, j) \leq \delta$ because if we have a feasible $f_s(i, j)$ to make (11b) hold, we must have $\mathbb{E}_{s\sim\rho_\upsilon^\pi} \min_{f_s(i,j)} \sum_{i=1}^N \sum_{j=1}^N M_{ij}\pi(a_j|s) f_s(i, j) \leq \delta$.

Since both the objective function and the constraint are linear in $f_s(i, j)$, (11) is a convex optimization problem. Also, Slater's condition holds for (11) as the feasible region has an interior point, which is $f_s(i, i) = 1\, \forall i$, and $f_s(i, j) = 0\, \forall i \neq j$. Meanwhile, since $A^\pi(s, a)$ is bounded based on Assumption 1, the objective is bounded above. Therefore, strong duality holds for (11). At this point we can derive the dual problem of (11) as its equivalent reformulation:

$$\begin{aligned} \min_{\beta\geq 0, \zeta_j^s} \quad & \beta\delta + \int_{s\in\mathcal{S}} \sum_{j=1}^{N} \zeta_j^s ds \\ s.t. \quad & A^\pi(s, a_i)\pi(a_j|s) - \beta M_{ij}\pi(a_j|s) - \frac{\zeta_j^s}{\rho_\upsilon^\pi(s)} \leq 0, \qquad \forall s \in \mathcal{S}, i, j = 1\dots N. \end{aligned} \tag{12}$$

We observe that with a fixed β, the optimal ζ_j^s will be achieved at:

$$\zeta_j^{s*}(\beta) = \max_{i=1\dots N} \rho_\upsilon^\pi(s)\pi(a_j|s)(A^\pi(s, a_i) - \beta M_{ij}). \tag{13}$$

Denote β^* as an optimal solution to (12) and $f_s^*(i,j)$ as an optimal solution to (11). Due to the complimentary slackness, the following equations hold:

$$(A^\pi(s,a_i)\pi(a_j|s) - \beta^* M_{ij}\pi(a_j|s) - \frac{\zeta_j^{s*}(\beta^*)}{\rho_\upsilon^\pi(s)})f_s^*(i,j) = 0, \qquad \forall s,i,j.$$

In this case, $f_s^*(i,j)$ can have non-zero values only when $A^\pi(s,a_i)\pi(a_j|s) - \beta^* M_{ij}\pi(a_j|s) - \frac{\zeta_j^{s*}(\beta^*)}{\rho_\upsilon^\pi(s)} = 0$, which means $\zeta_j^{s*}(\beta^*) = \rho_\upsilon^\pi(s)\pi(a_j|s)(A^\pi(s,a_i) - \beta^* M_{ij})$. Given the expression of the optimal ζ_j^{s*} in (13), $f_s^*(i,j)$ can have non-zero values only when $i \in \mathcal{K}_s^\pi(\beta^*,j)$, where $\mathcal{K}_s^\pi(\beta,j) = \text{argmax}_{k=1\dots N} A^\pi(s,a_k) - \beta M_{kj}$. Since $\sum_{i=1}^N f_s^*(i,j) = 1$ as indicated in (11c), we can choose an arbitrary optimizer $k_s^\pi(\beta^*,j) \in \mathcal{K}_s^\pi(\beta^*,j)$ to derive an optimal solution $f_s^*(i,j)$:

$$f_s^*(i,j) = \begin{cases} 1 & \text{if } i = k_s^\pi(\beta^*,j) \\ 0 & \text{otherwise,} \end{cases}$$

And then the corresponding optimal solution is, $\pi^*(a_i|s) = \sum_{j=1}^N \pi(a_j|s) f_s^*(i,j)$.

Last, by substituting $\zeta_j^{s*}(\beta) = \rho_\upsilon^\pi(s)\pi(a_j|s)(A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j})$ into the dual problem (12), we can reformulate (12) into:

$$\min_{\beta\geq 0}\{\beta\delta + \int_{s\in\mathcal{S}} \sum_{j=1}^N \zeta_j^{s*}(\beta)ds\} = \min_{\beta\geq 0}\{\beta\delta + \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^N \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\}. \tag{14}$$

The optimal β can then be obtained by solving (14).

We will further show that $\beta^* \leq \bar{\beta} := \max_{s\in\mathcal{S},k,j=1\dots N,k\neq j} (M_{kj})^{-1}(A^\pi(s,a_k) - A^\pi(s,a_j))$.

In the general case, i.e., $\beta \geq 0$, (11a) is non-negative because:

$$\begin{aligned}
&\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^N \sum_{j=1}^N A^\pi(s,a_i)\pi(a_j|s) f_s^*(i,j) && \text{(15a)}\\
&= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^N \pi(a_j|s) \sum_{i=1}^N A^\pi(s,a_i) f_s^*(i,j) && \text{(15b)}\\
&= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^N \pi(a_j|s) A^\pi(s,a_{k_s^\pi(\beta^*,j)}) && \text{(15c)}\\
&\geq \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^N \pi(a_j|s)\{A^\pi(s,a_j) + \beta^* M_{k_s^\pi(\beta^*,j)j}\} && \text{(15d)}\\
&= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^N \pi(a_j|s)\beta^* M_{k_s^\pi(\beta^*,j)j} && \text{(15e)}\\
&\geq 0, && \text{(15f)}
\end{aligned}$$

where (15d) holds since $A^\pi(s,a_{k_s^\pi(\beta^*,j)}) - \beta^* M_{k_s^\pi(\beta^*,j)j} \geq A^\pi(s,a_j) - \beta^* M_{jj} = A^\pi(s,a_j)$.

When $\beta^* > \max_{s\in\mathcal{S},k,j=1\dots N,k\neq j}\{\frac{A^\pi(s,a_k)-A^\pi(s,a_j)}{M_{kj}}\}$, we have that for all $s \in \mathcal{S}$, $k_s^\pi(\beta^*,j) = j$. Thus, $f_s^*(i,i) = 1$, $\forall i$ and $f_s^*(i,j) = 0$, $\forall i \neq j$. The objective value (11a) will be 0 because $\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^N \sum_{j=1}^N A^\pi(s,a_i)\pi(a_j|s) f_s^*(i,j) = \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^N A^\pi(s,a_i)\pi(a_i|s) = 0$. The left hand side of (11b) equals to $\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^N \sum_{j=1}^N M_{ij}\pi(a_j|s) f_s^*(i,j) = \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^N M_{ii}\pi(a_i|s) = 0$. Thus, for any $\delta > 0$, (11b) is always satisfied.

Since the objective of the primal Wasserstein trust-region constrained problem in (6) constantly evaluates to 0 when $\beta^* > \max_{s\in\mathcal{S},k,j=1\dots N,k\neq j}\{\frac{A^\pi(s,a_k)-A^\pi(s,a_j)}{M_{kj}}\}$, and is non-negative when $\beta^* \leq$

$\max_{s\in\mathcal{S},k,j=1\dots N,k\neq j}\{\frac{A^\pi(s,a_k)-A^\pi(s,a_j)}{M_{kj}}\}$, we can use $\max_{s\in\mathcal{S},k,j=1\dots N,k\neq j}\{\frac{A^\pi(s,a_k)-A^\pi(s,a_j)}{M_{kj}}\}$ as an upper bound for the optimal dual variable β^*. □

B PROOF OF THEOREM 2

Theorem 2. ***(Performance improvement)*** *For any initial state distribution μ and any $\beta_t \geq 0$, if $||\hat{A}^\pi - A^\pi||_\infty \leq \epsilon$ for some $\epsilon > 0$, the WPO policy update with the inaccurate advantage function $\hat{A}^\pi$, guarantees the following performance improvement bound,*

$$J(\pi_{t+1}) \geq J(\pi_t) + \beta_t \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum\nolimits_{j=1}^{N} \pi_t(a_j|s) M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} - \frac{2\epsilon}{1-\gamma}. \tag{7}$$

Proof of Theorem 2.

$$\begin{aligned}
J(\pi_{t+1}) - J(\pi_t) &= \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \mathbb{E}_{a\sim\pi_{t+1}}[A^{\pi_t}(s,a)] && \text{(16a)}\\
&= \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{i=1}^{N} \pi_{t+1}(a_i|s) A^{\pi_t}(s,a_i) && \text{(16b)}\\
&= \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_t(a_j|s) \hat{f}_s^t(i,j) A^{\pi_t}(s,a_i) && \text{(16c)}\\
&= \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{j=1}^{N} \pi_t(a_j|s) \sum_{i=1}^{N} \hat{f}_s^t(i,j) A^{\pi_t}(s,a_i) && \text{(16d)}\\
&= \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{j=1}^{N} \pi_t(a_j|s) A^{\pi_t}(s,a_{\hat{k}_s^{\pi_t}(\beta_t,j)}) && \text{(16e)}\\
&\geq \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{j=1}^{N} \pi_t(a_j|s)[A^{\pi_t}(s,a_j) + \beta_t M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} - 2\epsilon] && \text{(16f)}\\
&= \beta_t \mathbb{E}_{s\sim\rho_\mu^{\pi_{t+1}}} \sum_{j=1}^{N} \pi_t(a_j|s) M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} - \frac{2\epsilon}{1-\gamma}, && \text{(16g)}
\end{aligned}$$

where (16a) holds due to the performance difference lemma in Kakade & Langford (2002); (16f) follows from the definition of $\hat{k}_s^{\pi_t}(\beta_t,j)$ and the fact that $||\hat{A}^{\pi_t} - A^{\pi_t}||_\infty \leq \epsilon$, therefore $[A^{\pi_t}(s,a_{\hat{k}_s^{\pi_t}(\beta_t,j)}) + \epsilon] - \beta_t M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} \geq \hat{A}^{\pi_t}(s,a_{\hat{k}_s^{\pi_t}(\beta_t,j)}) - \beta_t M_{\hat{k}_s^{\pi_t}(\beta_t,j)j} \geq \hat{A}^{\pi_t}(s,a_j) - \beta_t M_{jj} = \hat{A}^{\pi_t}(s,a_j) \geq A^{\pi_t}(s,a_j) - \epsilon$; (16g) holds since $\mathbb{E}_{a\sim\pi}[A^\pi(s,a)] = 0$. □

C PROOF OF THEOREM 3

Theorem 3. *If Assumption 1 holds, then the optimal solution to the trust region constrained problem (4) with Sinkhorn divergence is:*

$$\pi_\lambda^*(a_i|s) = \sum_{j=1}^{N} \frac{\exp\left(\frac{\lambda}{\beta_\lambda^*} A^\pi(s,a_i) - \lambda M_{ij}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\lambda}{\beta_\lambda^*} A^\pi(s,a_k) - \lambda M_{kj}\right)} \pi(a_j|s), \tag{9}$$

where β_λ^* is an optimal solution to the following dual formulation:

$$\min_{\beta\geq 0}\ F_\lambda(\beta) = \min_{\beta\geq 0} \Big\{\beta\delta - \mathbb{E}_{s\sim\rho_v^\pi} \sum\nolimits_{j=1}^{N} \pi(a_j|s)\big(\tfrac{\beta}{\lambda} + \tfrac{\beta}{\lambda}\ln(\pi(a_j|s)) - \tfrac{\beta}{\lambda}\ln[\sum\nolimits_{i=1}^{N}\exp\left(\tfrac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)]\big) + \mathbb{E}_{s\sim\rho_v^\pi}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\tfrac{\beta}{\lambda}\tfrac{\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i)-\lambda M_{ij}\right)\cdot\pi(a_j|s)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_k)-\lambda M_{kj}\right)}\Big\}. \tag{10}$$

Moreover, we have $\beta_\lambda^* \leq \frac{2A^{max}}{\delta}$.

Proof of Theorem 3. Let β and ω represent the dual variables of constraints (8b) and (8c) respectively, then the Lagrangian duality of (8) can be derived as:

$$\max_{Q}\min_{\beta\geq 0,\omega} L(Q,\beta,\omega) = \max_{Q}\min_{\beta\geq 0,\omega} \mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N} A^\pi(s,a_i)\sum_{j=1}^{N} Q_{ij}^s]$$
$$+\int_{s\in\mathcal{S}}\sum_{j=1}^{N}\omega_j^s(\sum_{i=1}^{N} Q_{ij}^s - \pi(a_j|s))ds + \beta(\delta - \mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}Q_{ij}^s + \frac{1}{\lambda}Q_{ij}^s\log Q_{ij}^s]) \tag{17a}$$

$$= \max_{Q}\min_{\beta\geq 0,\omega} \mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N} A^\pi(s,a_i)\sum_{j=1}^{N} Q_{ij}^s] + \int_{s\in\mathcal{S}}\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{\omega_j^s}{\rho_\upsilon^\pi(s)}Q_{ij}^s\rho_\upsilon^\pi(s)ds$$
$$-\int_{s\in\mathcal{S}}\sum_{j=1}^{N}\omega_j^s\pi(a_j|s)ds + \beta\delta - \beta\mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}Q_{ij}^s + \frac{1}{\lambda}Q_{ij}^s\log Q_{ij}^s]) \tag{17b}$$

$$= \max_{Q}\min_{\beta\geq 0,\omega} \beta\delta - \int_{s\in\mathcal{S}}\sum_{j=1}^{N}\omega_j^s\pi(a_j|s)ds$$
$$+\mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N}\sum_{j=1}^{N}(A^\pi(s,a_i) - \beta M_{ij} + \frac{\omega_j^s}{\rho_\upsilon^\pi(s)})Q_{ij}^s - \frac{\beta}{\lambda}Q_{ij}^s\log Q_{ij}^s] \tag{17c}$$

$$= \min_{\beta\geq 0,\omega}\max_{Q} \beta\delta - \int_{s\in\mathcal{S}}\sum_{j=1}^{N}\omega_j^s\pi(a_j|s)ds$$
$$+\mathbb{E}_{s\sim\rho_\upsilon^\pi}[\sum_{i=1}^{N}\sum_{j=1}^{N}(A^\pi(s,a_i) - \beta M_{ij} + \frac{\omega_j^s}{\rho_\upsilon^\pi(s)})Q_{ij}^s - \frac{\beta}{\lambda}Q_{ij}^s\log Q_{ij}^s], \tag{17d}$$

where (17d) holds since the Lagrangian function $L(Q,\beta,\omega)$ is concave in Q and linear in β and ω, and we can exchange the $\max$ and the $\min$ following the Minimax theorem (Sion, 1958).

Note that the inner max problem of (17d) is an unconstrained concave problem, and we can obtain the optimal Q by taking the derivatives and setting them to 0. That is,

$$\frac{\partial L}{\partial Q_{ij}^s} = A^\pi(s,a_i) - \beta M_{ij} + \frac{\omega_j^s}{\rho_\upsilon^\pi(s)} - \frac{\beta}{\lambda}(\log Q_{ij}^s + 1) = 0,\ \ \forall i,j = 1,\cdots,N, s\in\mathcal{S}. \tag{18}$$

Therefore, we have the optimal Q_{ij}^{s*} as:

$$Q_{ij}^{s*} = \exp(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij})\exp(\frac{\lambda\omega_j^s}{\beta\rho_\upsilon^\pi(s)} - 1),\ \ \forall i,j = 1,\cdots,N, s\in\mathcal{S}. \tag{19}$$

In addition, since $\sum_{i=1}^{N} Q_{ij}^{s*} = \pi(a_j|s)$, we have the following hold:

$$\exp(\frac{\lambda\omega_j^s}{\beta\rho_\upsilon^\pi(s)} - 1) = \frac{\pi(a_j|s)}{\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij})}. \tag{20}$$

By substituting the left hand side of (20) into (19), we can further reformulate the optimal Q_{ij}^{s*} as:

$$Q_{ij}^{s*} = \frac{\exp(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_k) - \lambda M_{kj})}\pi(a_j|s),\ \ \forall i,j = 1,\cdots,N, s\in\mathcal{S}. \tag{21}$$

To obtain the optimal dual variables, based on (20), we have the optimal ω^* as:

$$\omega_j^{s*} = \rho_\upsilon^\pi(s)\{\frac{\beta}{\lambda} + \frac{\beta}{\lambda}\ln(\pi(a_j|s)) - \frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij})]\},\ \ \forall j = 1,\cdots,N, s\in\mathcal{S} \tag{22}$$

By substituting (21) and (22) into (17d), we can obtain the optimal β^* via:

$$\min_{\beta\geq 0} \quad \beta\delta - \mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\{\frac{\beta}{\lambda}+\frac{\beta}{\lambda}\ln(\pi(a_j|s))-\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_i)-\lambda M_{ij})]\}$$
$$+\mathbb{E}_{s\sim\rho_v^\pi}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\beta}{\lambda}\frac{\exp(\frac{\lambda}{\beta}A^\pi(s,a_i)-\lambda M_{ij})\cdot\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_k)-\lambda M_{kj})}.$$

The proof for the upper bound of sinkhorn optimal β can be found in Appendix D. □

D UPPER BOUND OF SINKHORN OPTIMAL BETA

In this section, we will derive the upper bound of Sinkhorn optimal β. First, for a given β, the optimal $Q_{ij}^{s*}(\beta)$ to the Lagrangian dual $L(Q,\beta,\omega)$ can be expressed in (21). With this, we will present the following two lemmas:

Lemma 1. *The objective function (8a) with respect to $Q_{ij}^{s*}(\beta)$ decreases as the dual variable β increases.*

Lemma 2. *If Assumption 1 holds, then for every $\delta>0$, $Q_{ij}^{s*}(\frac{2A^{max}}{\delta})$ is feasible to (8b) for any λ.*

We provide proofs for Lemma 1 and Lemma 2 in Appendix D.1 and Appendix D.2 respectively. Given the above two lemmas, we are able to prove the following proposition on the upper bound of Sinkhorn optimal β:

Proposition 1. *If β_λ^* is the optimal dual solution to the Sinkhorn dual formulation (10), then $\beta_\lambda^*\leq\frac{2A^{max}}{\delta}$ for any λ.*

Proof of Proposition 1. We will prove it by contradiction. According to Lemma 2, $Q_{ij}^{s*}(\frac{2A^{\max}}{\delta})$ is feasible to (8b). Since β_λ^* is the optimal dual solution, $Q_{ij}^{s*}(\beta_\lambda^*)$ is optimal to (8). If $\beta_\lambda^*>\frac{2A^{\max}}{\delta}$, according to Lemma 1, the objective value in (8a) with respect to $\frac{2A^{\max}}{\delta}$ is smaller than the objective value in (8a) with respect to β_λ^*, which contradicts the fact that $Q_{ij}^{s*}(\beta_\lambda^*)$ is the optimal solution to (8). □

D.1 PROOF OF LEMMA 1

Lemma 1. *The objective function (8a) with respect to $Q_{ij}^{s*}(\beta)$ decreases as the dual variable β increases.*

Proof of Lemma 1. Let $G_\lambda(\beta)$ represent the objective function (8a). By substituting the optimal Q_{ij}^{s*} in (21) into (8a), we have:

$$G_\lambda(\beta)=\mathbb{E}_{s\sim\rho_v^\pi}[\sum_{i=1}^{N}A^\pi(s,a_i)\sum_{j=1}^{N}\frac{\exp(\frac{\lambda}{\beta}A^\pi(s,a_i)-\lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_k)-\lambda M_{kj})}\pi(a_j|s)] \tag{23a}$$
$$=\mathbb{E}_{s\sim\rho_v^\pi}[\sum_{j=1}^{N}\pi(a_j|s)\sum_{i=1}^{N}A^\pi(s,a_i)\frac{\exp(\frac{\lambda}{\beta}A^\pi(s,a_i)-\lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_k)-\lambda M_{kj})}]. \tag{23b}$$

For any $\beta_2>\beta_1>0$, we have:

$$G_\lambda(\beta_1)-G_\lambda(\beta_2)$$
$$=\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\sum_{i=1}^{N}A^\pi(s,a_i)\{\frac{\exp(\frac{\lambda}{\beta_1}A^\pi(s,a_i)-\lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta_1}A^\pi(s,a_k)-\lambda M_{kj})}$$
$$-\frac{\exp(\frac{\lambda}{\beta_2}A^\pi(s,a_i)-\lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta_2}A^\pi(s,a_k)-\lambda M_{kj})}\} \tag{24a}$$

$$= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \sum_{i=1}^{N} A^\pi(s, a_{[i]}) \{ \frac{\exp(\frac{\lambda}{\beta_1} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp(\frac{\lambda}{\beta_1} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})}$$

$$- \frac{\exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})} \}, \tag{24b}$$

where $[i]$ denotes sorted indices that satisfy $A^\pi(s, a_{[1]}) \geq A^\pi(s, a_{[2]}) \geq \cdots \geq A^\pi(s, a_{[N]})$. Let

$$f_s(i) = \frac{\exp(\frac{\lambda}{\beta_1} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp(\frac{\lambda}{\beta_1} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})} - \frac{\exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})} \tag{25a}$$

$$= \frac{\exp((\frac{\lambda}{\beta_1} - \frac{\lambda}{\beta_2}) A^\pi(s, a_{[i]})) \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp((\frac{\lambda}{\beta_1} - \frac{\lambda}{\beta_2}) A^\pi(s, a_{[k]})) \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})}$$

$$- \frac{\exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[i]}) - \lambda M_{[i]j})}{\sum_{k=1}^{N} \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[k]}) - \lambda M_{[k]j})}. \tag{25b}$$

For notation brevity, we let $m_s(i) = \exp((\frac{\lambda}{\beta_1} - \frac{\lambda}{\beta_2}) A^\pi(s, a_{[i]})) > 0$, $w_s(i) = \exp(\frac{\lambda}{\beta_2} A^\pi(s, a_{[i]}) - \lambda M_{[i]j}) > 0$ and $q_s(i) = \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(i) w_s(k)}$. Then we have

$$(25b) = \frac{m_s(i) w_s(i)}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{w_s(i)}{\sum_{k=1}^{N} w_s(k)} \tag{26a}$$

$$= m_s(i) w_s(i) (\frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(i) w_s(k)}) \tag{26b}$$

$$= m_s(i) w_s(i) q_s(i). \tag{26c}$$

Since $\frac{\lambda}{\beta_1} - \frac{\lambda}{\beta_2} > 0$, $m_s(i)$ decreases as i increases. Thus, $q_s(i)$ decreases as i increases. Since $m_s(1) \geq m_s(k)$ and $m_s(N) \leq m_s(k)$ for all $k = 1, \ldots, N$, we have $q_s(1) = \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(1) w_s(k)} \geq \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} = 0$, and $q_s(N) = \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(N) w_s(k)} \leq \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} - \frac{1}{\sum_{k=1}^{N} m_s(k) w_s(k)} = 0$. Since $q_s(1) \geq 0$, $q_s(N) \leq 0$ and $q_s(i)$ decreases as i increases, there exists an index $1 \leq k_s \leq N$ such that $q_s(i) \geq 0$ for $i \leq k_s$ and $q_s(i) < 0$ for $i > k_s$. Since $m_s(i), w_s(i) > 0$, we have $f_s(i) \geq 0$ for $i \leq k_s$ and $f_s(i) < 0$ for $i > k_s$. In addition, we have $\sum_{i=1}^{N} f_s(i) = 0$ directly follows from the definition. Thus, $\sum_{i=1}^{N} f_s(i) = \sum_{i=1}^{k_s} |f_s(i)| - \sum_{i=k_s+1}^{N} |f_s(i)| = 0$. Therefore,

$$G_\lambda(\beta_1) - G_\lambda(\beta_2) = \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \sum_{i=1}^{N} A^\pi(s, a_{[i]}) f_s(i) \tag{27a}$$

$$= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \{ \sum_{i=1}^{k_s} A^\pi(s, a_{[i]}) |f_s(i)| - \sum_{i=k_s+1}^{N} A^\pi(s, a_{[i]}) |f_s(i)| \} \tag{27b}$$

$$\geq \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \{ \sum_{i=1}^{k_s} A^\pi(s, a_{[k_s]}) |f_s(i)| - \sum_{i=k_s+1}^{N} A^\pi(s, a_{[k_s+1]}) |f_s(i)| \} \tag{27c}$$

$$= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \{ A^\pi(s, a_{[k_s]}) \sum_{i=1}^{k_s} |f_s(i)| - A^\pi(s, a_{[k_s+1]}) \sum_{i=k_s+1}^{N} |f_s(i)| \} \tag{27d}$$

$$= \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s) \{ A^\pi(s, a_{[k_s]}) \sum_{i=1}^{k_s} |f_s(i)| - A^\pi(s, a_{[k_s+1]}) \sum_{i=1}^{k_s} |f_s(i)| \} \tag{27e}$$

$$= \mathbb{E}_{s\sim\rho_v^\pi} \sum_{j=1}^{N} \pi(a_j|s)(A^\pi(s, a_{[k_s]}) - A^\pi(s, a_{[k_s+1]})) \sum_{i=1}^{k_s} |f_s(i)| \tag{27f}$$

$$\geq 0. \tag{27g}$$

where (27c) and (27g) hold since $A^\pi(s, a_{[i]})$ is non-increasing as i increases. Furthermore, at least one inequality of (27c) and (27g) will not hold at equality since $\sum_{i=1}^{N} \pi(a_i|s)A^\pi(s, a_i) = 0, \forall s \in \mathcal{S}$, and for non-trivial cases, $Pr\{A^\pi(s, a) = 0, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}\} < 1$, which means $Pr\{\exists s_1, s_2 \in \mathcal{S}, a_1, a_2 \in \mathcal{A},\ s.t.\ A^\pi(s_1, a_1) \neq A^\pi(s_2, a_2)\} > 0$. Therefore, we have $G_\lambda(\beta_1) - G_\lambda(\beta_2) > 0$. □

D.2 Proof of Lemma 2

Lemma 2. *If Assumption 1 holds, then for every $\delta > 0$, $Q_{ij}^{s*}(\frac{2A^{max}}{\delta})$ is feasible to (8b) for any λ.*

Proof of Lemma 2. By substituting the optimal Q_{ij}^{s*} in (21) into (8b), we can reformulate the left hand side of (8b) as follows:

$$\mathbb{E}_{s\sim\rho_v^\pi}[\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}Q_{ij}^{s*} + \frac{1}{\lambda}Q_{ij}^{s*}\log Q_{ij}^{s*}] \tag{28a}$$

$$= \mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}Q_{ij}^{s*} + \frac{1}{\lambda}Q_{ij}^{s*}[\frac{\lambda}{\beta}A^\pi(s, a_i) - \lambda M_{ij} + \log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s, a_k) - \lambda M_{kj})}]\} \tag{28b}$$

$$= \mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\beta}Q_{ij}^{s*}A^\pi(s, a_i) + \frac{1}{\lambda}Q_{ij}^{s*}\log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s, a_k) - \lambda M_{kj})}\}. \tag{28c}$$

Now we prove that when $\beta = \frac{2A^{\max}}{\delta}$, $\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\beta}Q_{ij}^{s*}(\beta)A^\pi(s, a_i)\} \leq \frac{\delta}{2}$ and $\mathbb{E}_{s\sim\rho_v^\pi}\{\frac{1}{\lambda}Q_{ij}^{s*}(\beta)\log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s,a_k)-\lambda M_{kj})}\} \leq \frac{\delta}{2}$ hold. For the first part, we have:

$$\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\beta}Q_{ij}^{s*}A^\pi(s, a_i)\} \tag{29a}$$

$$= \frac{1}{\beta}\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}[\sum_{j=1}^{N}Q_{ij}^{s*}]A^\pi(s, a_i)\} \tag{29b}$$

$$= \frac{1}{\beta}\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\pi'(a_i|s)A^\pi(s, a_i)\} \tag{29c}$$

$$\leq \frac{1}{\beta}\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\pi'(a_i|s)|A^\pi(s, a_i)|\} \tag{29d}$$

$$\leq \frac{A^{max}}{\beta} = \frac{\delta}{2}. \tag{29e}$$

For the second part, the followings hold:

$$\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\lambda}Q_{ij}^{s*}\log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s, a_k) - \lambda M_{kj})}\} \tag{30a}$$

$$= \mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{j=1}^{N}\frac{1}{\lambda}(\sum_{i=1}^{N}Q_{ij}^{s*})\log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s, a_k) - \lambda M_{kj})}\} \tag{30b}$$

$$= \frac{1}{\lambda}\mathbb{E}_{s\sim\rho_v^\pi}\{\sum_{j=1}^{N}\pi(a_j|s)\log\frac{\pi(a_j|s)}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^\pi(s, a_k) - \lambda M_{kj})}\} \tag{30c}$$

$$\leq \frac{1}{\lambda}\mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\{\sum_{j=1}^{N}\pi(a_j|s)\log\frac{\pi(a_j|s)}{\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_j))}\} \tag{30d}$$

$$\leq \frac{1}{\lambda}\mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\{\sum_{j=1}^{N}\pi(a_j|s)\log\frac{1}{\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_j))}\} \tag{30e}$$

$$= \frac{1}{\lambda}\mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\{\sum_{j=1}^{N}\pi(a_j|s)(-\frac{\lambda}{\beta}A^{\pi}(s,a_j))\} \tag{30f}$$

$$\leq \frac{1}{\beta}\mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\{\sum_{j=1}^{N}\pi(a_j|s)|A^{\pi}(s,a_j)|\} \tag{30g}$$

$$\leq \frac{A^{max}}{\beta} = \frac{\delta}{2}. \tag{30h}$$

Therefore, $Q_{ij}^{s*}(\frac{2A^{\max}}{\delta})$ is feasible to (8b) for any λ. □

E Gradient of the Objective in the Sinkhorn Dual Formulation

The closed-form gradient of the objective in the Sinkhorn dual formulation (10) is as follows:

$$\begin{aligned}
&\delta - \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\sum_{j=1}^{N}\pi(a_j|s)\Big\{\frac{1}{\lambda}+\frac{1}{\lambda}\ln(\pi(a_j|s))-\frac{1}{\lambda}\ln[\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})]\\
&-\frac{\beta}{\lambda}\cdot\frac{1}{\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})}\times\sum_{i=1}^{N}[\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})\times-\lambda A^{\pi}(s,a_i)\beta^{-2}]\Big\}\\
&+\mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big\{\frac{\pi(a_j|s)}{\lambda}\frac{\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})}{\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj})}\\
&+\frac{\beta\pi(a_j|s)}{\lambda}\cdot\frac{\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})\times-\lambda A^{\pi}(s,a_i)\beta^{-2}\times\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj})}{(\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj}))^2}\\
&-\frac{\beta\pi(a_j|s)}{\lambda}\cdot\frac{\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})\times\sum_{k=1}^{N}[\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj})\times-\lambda A^{\pi}(s,a_k)\beta^{-2}]}{(\sum_{k=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_k)-\lambda M_{kj}))^2}\Big\}.
\end{aligned}$$

F Proof of Theorem 4

Given the upper bound of Wassertein optimal β in Theorem 1 and the upper bound of Sinkhorn optimal β in Proposition 1, we are able to derive the following theorem:

Theorem 4. *Define* $\beta_{UB} = \max\{\frac{2A^{max}}{\delta}, \bar{\beta}\}$. *The following holds:*

1. $F_{\lambda}(\beta)$ *converges to* $F(\beta)$ *uniformly on* $[0, \beta_{UB}]$,
2. $\lim_{\lambda\to\infty}\arg\min_{0\leq\beta\leq\beta_{UB}} F_{\lambda}(\beta) \subseteq \arg\min_{0\leq\beta\leq\beta_{UB}} F(\beta)$.

Proof of Theorem 4. To show that $F_{\lambda}(\beta)$ converges to $F(\beta)$ uniformly on $[0, \beta_{\text{UB}}]$, it is equivalent to show that $\lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\Big|F_{\lambda}(\beta) - F(\beta)\Big| = 0$. Let $\epsilon_s^{\pi}(\beta,i,j) = A^{\pi}(s,a_{k_s^{\pi}(\beta,j)}) - \beta M_{k_s^{\pi}(\beta,j)j} - [A^{\pi}(s,a_i) - \beta M_{ij}]$ where $k_s^{\pi}(\beta,j) \in \mathcal{K}_s^{\pi}(\beta,j) = \text{argmax}_{k=1\ldots N} A^{\pi}(s,a_k) - \beta M_{kj}$ is an arbitrary optimizer, and $\epsilon_s^{\pi}(\beta,i,j) \geq 0$. First, we have

$$\begin{aligned}
&\Big|F_{\lambda}(\beta) - F(\beta)\Big|\\
&= \Big|\beta\delta - \mathbb{E}_{s\sim\rho_{\upsilon}^{\pi}}\sum_{j=1}^{N}\pi(a_j|s)\{\frac{\beta}{\lambda}+\frac{\beta}{\lambda}\ln(\pi(a_j|s))-\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp(\frac{\lambda}{\beta}A^{\pi}(s,a_i)-\lambda M_{ij})]\}
\end{aligned}$$

$$+ \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\beta}{\lambda} \frac{\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)\cdot \pi(a_j|s)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_k) - \lambda M_{kj}\right)} - \beta\delta$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big| \tag{31a}$$

$$\leq \Big|\frac{\beta}{\lambda}\mathbb{E}_{s\sim\rho_\upsilon^\pi}\sum_{j=1}^{N}\pi(a_j|s)\Big| + \Big|\frac{\beta}{\lambda}\mathbb{E}_{s\sim\rho_\upsilon^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln(\pi(a_j|s))\Big|$$

$$+ \Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\beta}{\lambda} \frac{\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)\cdot \pi(a_j|s)}{\sum_{k=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_k) - \lambda M_{kj}\right)}\Big|$$

$$+ \Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)]$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big| \tag{31b}$$

$$\leq 2\Big|\frac{\beta}{\lambda}\mathbb{E}_{s\sim\rho_\upsilon^\pi}\sum_{j=1}^{N}\pi(a_j|s)\Big| + \Big|\frac{\beta}{\lambda}\mathbb{E}_{s\sim\rho_\upsilon^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln(\pi(a_j|s))\Big|$$

$$+ \Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)]$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big|. \tag{31c}$$

In addition,

$$\Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_i) - \lambda M_{ij}\right)]$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big| \tag{32a}$$

$$=\Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\ln[\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_{k_s^\pi(\beta,j)}) - \lambda M_{k_s^\pi(\beta,j)j}\right)\sum_{i=1}^{N}\exp\left(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j)\right)]$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big| \tag{32b}$$

$$=\Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\{\ln[\exp\left(\frac{\lambda}{\beta}A^\pi(s,a_{k_s^\pi(\beta,j)}) - \lambda M_{k_s^\pi(\beta,j)j}\right)] + \ln[\sum_{i=1}^{N}\exp\left(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j)\right)]\}$$

$$- \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)[A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j}]\Big| \tag{32c}$$

$$=\Big|\mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)\frac{\beta}{\lambda}\ln[\sum_{i=1}^{N}\exp\left(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j)\right)]\Big|. \tag{32d}$$

Therefore,

$$\lim_{\lambda\longrightarrow\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\Big|F_\lambda(\beta) - F(\beta)\Big| \tag{33a}$$

$$\leq \lim_{\lambda\to\infty} \frac{2\beta_{\text{UB}}}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\Big| + \lim_{\lambda\to\infty}\frac{\beta_{\text{UB}}}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln(\pi(a_j|s))\Big|$$

$$+ \lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\frac{\beta}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j))]\Big| \tag{33b}$$

$$= \lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\frac{\beta}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j))]\Big|. \tag{33c}$$

In addition, $\forall\beta\in[0,\beta_{\text{UB}}]$ and $\forall\lambda$, $\epsilon_s^\pi(\beta,i,j)$ is bounded since

$$\begin{aligned}\Big|\epsilon_s^\pi(\beta,i,j)\Big| &= \Big|A^\pi(s,a_{k_s^\pi(\beta,j)}) - \beta M_{k_s^\pi(\beta,j)j} - [A^\pi(s,a_i)-\beta M_{ij}]\Big| \qquad (34)\\ &= \Big|(A^\pi(s,a_{k_s^\pi(\beta,j)}) - A^\pi(s,a_i)) - (\beta M_{k_s^\pi(\beta,j)j} - \beta M_{ij})\Big|\\ &\leq \Big|A^\pi(s,a_{k_s^\pi(\beta,j)}) - A^\pi(s,a_i)\Big| + \Big|\beta M_{k_s^\pi(\beta,j)j} - \beta M_{ij}\Big|\\ &\leq 2\max_{s,a}A^\pi(s,a) + \beta_{\text{UB}}\max_{i,j}M_{ij}\\ &\leq 2A^{\max} + \beta_{\text{UB}}\max_{i,j}M_{ij} < \infty. \qquad (35)\end{aligned}$$

Then, $\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j))]\Big|$ is bounded. Therefore in (33c), the optimal β can be achieved. Let $\beta^\lambda = \arg\max_{0\leq\beta\leq\beta_{\text{UB}}}\frac{\beta}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j))]\Big|$, and then we have:

$$\lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\frac{\beta}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta}\epsilon_s^\pi(\beta,i,j))]\Big| \tag{36a}$$

$$= \lim_{\lambda\to\infty}\frac{\beta^\lambda}{\lambda}\Big|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^{N}\pi(a_j|s)\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))]\Big|. \tag{36b}$$

Define $\sigma_s(j) = \min_{0\leq\beta\leq\beta_{\text{UB}}}\min_{i=1\ldots N, i\notin\mathcal{K}_s^\pi(\beta,j)}\epsilon_s^\pi(\beta,i,j)$. Then since $\epsilon_s^\pi(\beta,i,j) > 0$ for $i\notin\mathcal{K}_s^\pi(\beta,j)$ based on its definition, we have $\sigma_s(j) > 0$. On one hand, we have

$$\lim_{\lambda\to\infty}\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))] \tag{37a}$$

$$= \lim_{\lambda\to\infty}\ln[\sum_{i=1|i\notin\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j)) + \sum_{i=1|i\in\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))] \tag{37b}$$

$$\leq \lim_{\lambda\to\infty}\ln[\sum_{i=1|i\notin\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta_{\text{UB}}}\sigma_s(j)) + \sum_{i=1|i\in\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(0)] \tag{37c}$$

$$= \lim_{\lambda\to\infty}\ln[\sum_{i=1|i\notin\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta_{\text{UB}}}\sigma_s(j)) + |\mathcal{K}_s^\pi(\beta_\lambda,j)|] \tag{37d}$$

$$= \lim_{\lambda\to\infty}\ln[|\mathcal{K}_s^\pi(\beta_\lambda,j)|]. \tag{37e}$$

On the other hand, we have

$$\lim_{\lambda\to\infty}\ln[\sum_{i=1}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))] \tag{38a}$$

$$= \lim_{\lambda\to\infty}\ln[\sum_{i=1|i\notin\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j)) + \sum_{i=1|i\in\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N}\exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))] \tag{38b}$$

$$\geq \lim_{\lambda\to\infty} \ln[\sum_{i=1|i\in\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N} \exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))] \tag{38c}$$

$$= \lim_{\lambda\to\infty} \ln[\sum_{i=1|i\in\mathcal{K}_s^\pi(\beta_\lambda,j)}^{N} \exp(0)] \tag{38d}$$

$$= \lim_{\lambda\to\infty} \ln[|\mathcal{K}_s^\pi(\beta_\lambda,j)|]. \tag{38e}$$

Therefore, $\lim_{\lambda\to\infty}\left|\ln[\sum_{i=1}^N \exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))]\right| = \lim_{\lambda\to\infty}\ln[|\mathcal{K}_s^\pi(\beta_\lambda,j)|]$. Based on that, we have

$$\lim_{\lambda\to\infty}\frac{\beta^\lambda}{\lambda}\left|\mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^N \pi(a_j|s)\ln[\sum_{i=1}^N \exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))]\right| \tag{39a}$$

$$\leq \lim_{\lambda\to\infty}\frac{\beta^\lambda}{\lambda}\left|\sum_{j=1}^N \ln[\sum_{i=1}^N \exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))]\right| \tag{39b}$$

$$\leq \lim_{\lambda\to\infty}\frac{\beta^\lambda}{\lambda}\sum_{j=1}^N\left|\ln[\sum_{i=1}^N \exp(-\frac{\lambda}{\beta^\lambda}\epsilon_s^\pi(\beta^\lambda,i,j))]\right| \tag{39c}$$

$$= \lim_{\lambda\to\infty}\frac{\beta^\lambda}{\lambda}\sum_{j=1}^N \ln[|\mathcal{K}_s^\pi(\beta_\lambda,j)|] \tag{39d}$$

$$\leq \lim_{\lambda\to\infty}\frac{\beta_{\text{UB}}}{\lambda}N\ln N = 0, \tag{39e}$$

which means $\lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}\left|F_\lambda(\beta) - F(\beta)\right| \leq 0$. Furthermore, since $\lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}|F_\lambda(\beta)-F(\beta)| \geq 0$ holds naturally, we have $\lim_{\lambda\to\infty}\sup_{0\leq\beta\leq\beta_{\text{UB}}}|F_\lambda(\beta) - F(\beta)| = 0$. Therefore, $F_\lambda(\beta)$ converges to $F(\beta)$ uniformly on $[0,\beta_{\text{UB}}]$, which also indicates $F_\lambda(\beta)$ epi-converges to $F(\beta)$ on $[0,\beta_{\text{UB}}]$ (Royset, 2018; Rockafellar & Wets, 1998). By properties of epi-convergence, we have that $\lim_{\lambda\to\infty}\arg\min_{0\leq\beta\leq\beta_{\text{UB}}} F_\lambda(\beta) \subseteq \arg\min_{0\leq\beta\leq\beta_{\text{UB}}} F(\beta)$ (Rockafellar & Wets, 1998). □

G Optimal Beta for a Special Distance

Proposition 2. *(1). If the initial point β_0 is in $[\max_{s,j}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\},+\infty)$, the local optimal β solution is $\max_{s,j}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\}$.*
(2). If the initial point β_0 is in $[0,\min_{s,j\neq k_s}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\}]$: if $\delta - \int_{s\in\mathcal{S}}\rho^\pi(s)(1-\pi(a_{k_s}|s))ds < 0$, the local optimal β is $\min_{s,j\neq k_s}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\}$; otherwise, the local optimal β solution is 0.

(3). If the initial point β_0 is in $(\min_{s,j\neq k_s}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\},\max_{s,j}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\})$, we construct sets I_s^1 and I_s^2 as:

for $s\in\mathcal{S}, j\in\{1,2\ldots N\}$: **if** $\beta_0 \geq A^\pi(s,a_{k_s}) - A^\pi(s,a_j)$ **then** Add j to I_s^1 **else** Add j to I_s^2.
Then, if $\delta - \mathbb{E}_{s\sim\rho^\pi}\sum_{j\in I_s^2}\pi(a_j|s) < 0$, the local optimal β is $\min_{s\in\mathcal{S},j\in I_s^2}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\}$; otherwise, the local optimal β is $\max_{s\in\mathcal{S},j\in I_s^1}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\}$.

Proof of Proposition 2. (1). When $\beta \in [\max_{s,j}\{A^\pi(s,a_{k_s}) - A^\pi(s,a_j)\},+\infty)$, we have $A^\pi(s,a_j) \geq A^\pi(s,a_{k_s}) - \beta$ for all $s\in\mathcal{S}, j = 1\ldots N$. Since $A^\pi(s,a_{k_s}) - \beta \geq A^\pi(s,a_k) - \beta$ for all $k = 1\ldots N$, we have $A^\pi(s,a_j) \geq A^\pi(s,a_k) - \beta$ for all $s\in\mathcal{S}, j = 1\ldots N, k = 1\ldots N$. Thus, $j\in\mathcal{K}_s^\pi(\beta^*,j)$, for all $s\in\mathcal{S}, j = 1\ldots N$. Therefore, (6) can be reformulated as:

$$\min_{\beta\geq 0}\{\beta\delta + \mathbb{E}_{s\sim\rho_v^\pi}\sum_{j=1}^N \pi(a_j|s)A^\pi(s,a_j)\}.$$

Since $\delta \geq 0$, we have the local optimal $\beta = \max_{s,j}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}$.

(2). When $\beta \in [0, \min_{s,j\neq k_s}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}]$, we have $A^\pi(s, a_j) \leq A^\pi(s, a_{k_s}) - \beta$ for all $s \in \mathcal{S}$, $j = 1 \dots N$, $j \neq k_s$. Thus $k_s \in \mathcal{K}_s^\pi(\beta^*, j)$ for all $s \in \mathcal{S}$, $j = 1 \dots N$. The inner part of (6) then is:

$$\beta\delta + \mathbb{E}_{s\sim\rho_\upsilon^\pi}\{\sum_{j=1,j\neq k_s}^{N} \pi(a_j|s)(A^\pi(s, a_{k_s}) - \beta) + \pi(a_{k_s}|s)A^\pi(s, a_{k_s})\}$$
$$= \beta(\delta - \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1,j\neq k_s}^{N} \pi(a_j|s)) + \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)A^\pi(s, a_{k_s})$$
$$= \beta(\delta - \int_{s\in\mathcal{S}} \rho_\upsilon^\pi(s)(1 - \pi(a_{k_s}|s))ds) + \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j=1}^{N} \pi(a_j|s)A^\pi(s, a_{k_s}).$$

If $\delta - \int_{s\in\mathcal{S}} \rho_\upsilon^\pi(s)(1 - \pi(a_{k_s}|s))ds < 0$, we have the local optimal $\beta = \min_{s,j\neq k_s}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}$. If $\delta - \int_{s\in\mathcal{S}} \rho_\upsilon^\pi(s)(1 - \pi(a_{k_s}|s))ds \geq 0$, we have the local optimal $\beta = 0$.

(3). For an initial point β_0 in $(\min_{s,j\neq k_s}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}, \max_{s,j}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\})$, we construct partitions I_s^1 and I_s^2 of the set $\{1, 2 \dots N\}$ in the way described in Proposition 2 for all $s \in \mathcal{S}$. Consider β in the neightborhood of β_0, i.e., $\beta \geq A^\pi(s, a_{k_s}) - A^\pi(s, a_j)$ for $s \in \mathcal{S}, j \in I_s^1$ and $\beta \leq A^\pi(s, a_{k_s}) - A^\pi(s, a_j)$ for $s \in \mathcal{S}, j \in I_s^2$. Then the inner part of (6) can be reformulated as:

$$\beta\delta + \mathbb{E}_{s\sim\rho_\upsilon^\pi}\{\sum_{j\in I_s^1} \pi(a_j|s)A^\pi(s, a_j) + \sum_{j\in I_s^2} \pi(a_j|s)(A^\pi(s, a_{k_s}) - \beta)\}$$
$$= \beta(\delta - \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j\in I_s^2} \pi(a_j|s)) + \mathbb{E}_{s\sim\rho_\upsilon^\pi}\{\sum_{j\in I_s^1} \pi(a_j|s)A^\pi(s, a_j) + \sum_{j\in I_s^2} \pi(a_j|s)A^\pi(s, a_{k_s})\}.$$

If $\delta - \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j\in I_s^2} \pi(a_j|s) < 0$, we have the local optimal $\beta = \min_{s\in\mathcal{S},j\in I_s^2}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}$. If $\delta - \mathbb{E}_{s\sim\rho_\upsilon^\pi} \sum_{j\in I_s^2} \pi(a_j|s) \geq 0$, we have the local optimal $\beta = \max_{s\in\mathcal{S},j\in I_s^1}\{A^\pi(s, a_{k_s}) - A^\pi(s, a_j)\}$. □

H Implementation Details

Visitation Frequencies Estimation: The unnormalized discounted visitation frequencies are needed to compute the global optimal β^*. At the k-th iteration, the visitation frequencies ρ_k^π are estimated using samples of the trajectory set $\mathcal{D}_k$. Specifically, we first initialize $\rho_k^\pi(s) = 0,\ \forall s \in S$. Then for each timestep t in each trajectory from $\mathcal{D}_k$, we update ρ_k^π as $\rho_k^\pi(s_t) \leftarrow \rho_k^\pi(s_t) + \gamma^t/|\mathcal{D}_k|$.

Policy Representation: The general approach depicted in Algorithm 1 allows various policy representations including arrays and neural networks. Let $\mathcal{S}_k \subseteq \mathcal{S}$ represent a subset of states to perform the policy update at the k-th iteration. When an array is used, the policy update step is simply $\pi_{k+1}(\cdot|s) = \mathbb{F}(\pi_k)(\cdot|s),\ \forall s \in \mathcal{S}_k$. When a neural network is employed, the policy update step can be achieved by obtaining the gradient descent, i.e., $\nabla \sum_{s\in\mathcal{S}_k} (\mathbb{F}(\pi_k)(\cdot|s) - \pi_k(\cdot|s))^2$.

Policy Updating Strategy: State space: For environments with a discrete state space (e.g., tabular domains), the policy update is performed for all states at each iteration. For the environments with a continuous state space, a random subset of states is sampled at each iteration to perform the policy update.

Hyperparamaters and Additional Results:

Our main experimental results are reported in Section 6. In addition, we provide the setting of hyperparamaters and network sizes of our WPO/SPO algorithms in Table 2. And we present the numerical results of the final performance comparison among our algorithms and the baseline methods (i.e., TRPO, PPO, A2C) in Table 3.

Table 2: Hyperparamaters and network sizes

	Taxi-v3, NChain-v0 CliffWalking-v0	CartPole-v1	Acrobot-v1
γ	0.9	0.95	0.95
lr_π	\	10^{-2}	5×10^{-3}
lr_{value}	10^{-2}	10^{-2}	5×10^{-3}
$\lvert\mathcal{D}_k\rvert$	60 (Taxi); 1 (Chain); 3 (CliffWalking)	2	3
π size	2D array	$[64, 64]$	$[64, 64]$
Q/v size	$[10, 7, 5]$	$[64, 64]$	$[64, 64]$
λ	$5, 50, 10$	10	10

Table 3: Averaged rewards over last 10% episodes during the training process

Environment	WPO	SPO	TRPO	PPO	A2C
Taxi-v3	-45 ± 27	-87 ± 11	-202 ± 3	-381 ± 34	-338 ± 30
NChain-v0	3549 ± 197	3432 ± 131	3522 ± 258	3506 ± 237	1606 ± 10
CliffWalking-v0	-35 ± 15	-25 ± 1	-159 ± 94	-3290 ± 2106	-5587 ± 1942
CartPole-v1	388 ± 54	370 ± 30	297 ± 65	193 ± 45	267 ± 61
Acrobot-v1	-162 ± 8	-185 ± 15	-248 ± 33	-103 ± 5	-379 ± 39