Towards an Understanding of Default Policies in Multitask Policy Optimization

Ted Moskovitz, Michael Arbel, Jack Parker-Holder, Aldo Pacchiano

Towards an Understanding of Default Policies in Multitask Policy Optimization

Ted Moskovitz
Gatsby Unit, UCL
ted@gatsby.ucl.ac.uk

Michael Arbel
Université Grenoble Alpes, Inria, CNRS
michael.n.arbel@gmail.com

Jack Parker-Holder
University of Oxford
jackph@robots.ox.ac.uk

Aldo Pacchiano
Microsoft Research
apacchiano@microsoft.com

Abstract

Much of the recent success of deep reinforcement learning has been driven by regularized policy optimization (RPO) algorithms, with strong performance across multiple domains. In this family of methods, agents are trained to maximize cumulative reward while penalizing deviation in behavior from some reference, or *default* policy. In addition to empirical success, there is a strong theoretical foundation for understanding RPO methods applied to single tasks, with connections to natural gradient, trust region, and variational approaches. However, there is limited formal understanding of desirable properties for default policies in the *multitask* setting, an increasingly important domain as the field shifts towards training more generally capable agents. Here, we take a first step towards filling this gap by formally linking the quality of the default policy to its effect on optimization. Using these results, we then derive a principled RPO algorithm for multitask learning with strong performance guarantees.

1 Introduction

Appropriate regularization has been a key factor in the widespread success of policy-based deep reinforcement learning (RL) (Levine, 2018; Furuta et al., 2021). The key idea underlying such *regularized policy optimization* (RPO) methods is to train an agent to maximize reward while minimizing some cost which penalizes deviations from useful behavior, typically encoded as a *default policy*. In addition to being easily scalable and compatible with function approximation, these methods have been shown to ameliorate the high sample complexity of deep RL methods, making them an attractive choice for high-dimensional problems (Berner et al., 2019; Espeholt et al., 2018).

A natural question underlying this success is *why* these methods are so effective. Fortunatley, there is a strong foundation for the formal understanding of regularizers in the single-task setting. These methods can be seen as approximating a form of natural gradient ascent (Kakade, 2002; Pacchiano et al., 2020; Moskovitz et al., 2021), trust region or proximal point optimization (Schulman et al., 2015, 2017), or variational inference (Levine, 2018; Haarnoja et al., 2018; Marino et al., 2020; Abdolmaleki et al., 2018), and thus are well-understood by theory (Agarwal et al., 2020).

However, as interest has grown in training general agents capable of providing real world utility, there has been a shift in emphasis towards *multitask* learning. Accordingly, there are a number of approaches to learning or constructing default policies for regularized policy optimization in multitask settings (Galashov et al., 2019; Teh et al., 2017; Goyal et al., 2019, 2020). The basic idea is to obtain a default policy which is generally useful for some family of tasks, thus offering a form of supervision to the learning process. However, there is little theoretical understanding of how the choice of default policy affects optimization. Our goal in this paper is to take a first step towards bridging this gap, asking:

What properties does a default policy need to have in order to improve optimization on new tasks?

This is a nuanced question. The choice of penalty, structural commonalities among the tasks encountered by the agent, and even the distribution space in which the regularization is applied have dramatic effects on the resulting algorithm and the agent's performance characteristics.

In this work, we focus on methods using the Kullback-Leibler (KL) divergence with respect to the default policy, as they are the most common in the literature. We first consider this form of regularized policy optimization applied to a single task, with the goal of understanding how the relationship between the default and optimal policies for a given problem affect optimization. We then generalize these results to the multitask setting, where we not only quantify the advantages of this family of approaches, but also identify its limitations, both fundamental and algorithm-specific.

In the process of garnering new understanding of these algorithms, our results also imply a new framework through which to understand families of tasks. Because different algorithms are sensitive to different forms of structure, this leads to another guiding question, closely tied to the first:

What properties does a group of tasks need to share for a given algorithm to provide a measurable benefit?

It's clear that in order to be effective, any multitask learning algorithm must be applied to a task distribution with some form of structure identifiable by that algorithm: if tasks have nothing in common, no understanding gained from one task will be useful for accelerating learning on another. Algorithms may be designed to accommodate—or learn—a broader array of structures, but at increased computational costs. In high-dimensional problems, function approximation mandates new compromises. In this paper, which we view as a first step towards understanding these trade-offs, we make the following contributions:

- We show the error bound and iteration complexity for optimization using an α-optimal default policy, where sub-optimality is measured via distance from the optimal policy for a given task.
- From these results, we derive a principled RPO algorithm for multitask learning, which we term *total variation policy optimization* (TVPO). We show that popular multitask KL-based algorithms can be seen as approximations of TVPO and demonstrate the strong performance of TVPO on simple tasks.
- We offer novel insights on the optimization characteristics—both limitations and advantages—of common multitask RPO frameworks in the literature.

2 Regularized Policy Optimization

Reinforcement learning In reinforcement learning (RL), an agent learns how to act within its environment in order to maximize its performance on a given task or tasks. We model a task as a finite *Markov decision process* (MDP; (Puterman, 2010)) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where $\mathcal{S}, \mathcal{A}$ are finite state and action spaces, respectively, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition distribution, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, $\gamma \in [0, 1)$ is the discount factor, and $\rho \in \Delta(\mathcal{S})$ is the starting state distribution. $\Delta(\cdot)$ is used to denote the simplex over a given space. We also assume access to a restart distribution for training $\mu \in \Delta(\mathcal{S})$ such that $\mu(s) > 0 \ \forall s \in \mathcal{S}$, as is common in the literature (Kakade and Langford, 2002; Agarwal et al., 2020). The agent takes actions using a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which, in conjunction with the transition dynamics, induces a distribution over trajectories $\tau = (s_t, a_t)_{t=0}^{\infty}$.

The value function $V^\pi : \mathcal{S} \to \mathbb{R}^+$ measures the expected discounted cumuluative reward obtained by following π from state s, $V^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_t = s\right]$, where the expectation is with respect to the distribution over trajectories induced by π in M. We overload notation and define $V^\pi(\rho) \coloneqq \mathbb{E}_{s_0 \sim \rho} \left[V^\pi(s_0)\right]$ as the expected value for initial state distribution ρ. The action-value and advantage functions are given by

$$Q^\pi(s, a) \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_t = s, a_t = a\right], \qquad A^\pi(s, a) \coloneqq Q^\pi(s, a) - V^\pi(s).$$

By $d^{\pi}_{s_0}$, we denote the discounted state visitation distribution of π with starting state distribution μ, so that

$$d^{\pi}_{s_0}(s) = \mathbb{E}_{s_0 \sim \mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}^{\pi}(s_t = s|s_0)\right], \tag{2.1}$$

where $d^{\pi}_{\mu} \coloneqq \mathbb{E}_{s_0\sim\mu}\left[d^{\pi}_{s_0}(s)\right]$. The goal of the agent is to adapt its policy so as to maximize its value, i.e., optimize $\max_{\pi} V^{\pi}(\rho)$. We use $\pi^{\star} \in \mathrm{argmax}_{\pi} V^{\pi}(\rho)$ to denote the optimal policy and $V^{\star}$ and $Q^{\star}$ as shorthand for $V^{\pi^{\star}}$ and $Q^{\pi^{\star}}$, respectively.

Policy parameterizations In practice, this problem typically takes the form $\max_{\theta\in\Theta} V^{\pi_\theta}$, where $\{\pi_\theta|\theta\in\Theta\}$ is a class of parametric policies. In this work, we primarily consider the softmax policy class, which may be tabular or *complete* (able to represent any stochastic policy), as in the case of the tabular softmax

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}}\exp(\theta_{s,a'})}, \tag{2.2}$$

where $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, or *restricted*, where $\pi_\theta(a|s) \propto \exp(f_\theta(s,a))$, with $f_\theta : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ some parametric function class (e.g., a neural network).

The general form of the *regularized policy optimization* (RPO) objective function is given by

$$\mathcal{J}_\lambda(\theta) \coloneqq V^{\pi_\theta}(\mu) - \lambda\Omega(\theta), \tag{2.3}$$

where Ω is some convex regularization functional. Gradient ascent updates proceed according to

$$\theta^{(t+1)} = \theta^{(t)} + \eta\nabla_\theta \mathcal{J}_\lambda(\theta^{(t)}). \tag{2.4}$$

For simplicity of notation, from this point forward, for iterative algorithms which obtain successive estimates of parameters $\theta^{(t)}$, we denote the associated policy and value functions as $\pi^{(t)}$ and $V^{(t)}$, respectively. The choice of Ω plays a signficant role in algorithm design and practice, as we discuss below. It's also important to note that the error bounds and convergence rates we derive are based on the basic policy gradient framework in Appendix Algorithm 2, in which update Eq. (2.4) is applied every after a fixed number B of trajectories $\{\tau_b\}_{b=1}^{B}$ is sampled from the environment. Therefore, the iteration complexities below are proportional to the associated sample complexity.

3 Related Work

Single-task learning The majority of the theoretical (Agarwal et al., 2020; Grill et al., 2020) and empirical (Schulman et al., 2015, 2017; Abdolmaleki et al., 2018; Pacchiano et al., 2020) literature has focused on the use of RPO in a single-task setting, i.e., applied to a single MDP M. The majority of these methods place a soft or hard constraint on the Kullback-Leibler (KL) divergence between the updated policy at each time step and the current policy, maximizing an objective of the form

$$\mathcal{J}_\lambda(\pi_p, \pi_q) = \sum_{t=0}^{\infty}\mathbb{E}_{s_t\sim d^{\pi_q}_{\mu}}\mathbb{E}_{a_t\sim\pi_q(\cdot|s_t)}\left[\mathcal{G}(s_t,a_t) - \lambda\mathrm{KL}(\pi_p(\cdot|s_t),\pi_q(\cdot|s_t))\right], \tag{3.1}$$

where $\mathcal{G}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is typically the Q- or advantage function and $\pi_q, \pi_p \in \{\pi_\theta, \pi_0\}$ (Furuta et al., 2021). At each update, then, the idea is to maximize reward while minimizing the regularization cost. From a theoretical perspective, such methods can often be framed as a form of approximate variational inference, with either learned (Abdolmaleki et al., 2018; Song et al., 2019; Peng et al., 2021; Nair et al., 2021; Peters et al., 2010) or fixed (Todorov, 2007; Toussaint and Storkey, 2006; Rawlik et al., 2013; Fox et al., 2016) π_0. When $\pi_0 \approx \pi_\theta$, we can also understand such approaches as

approximating the natural policy gradient (Kakade, 2002), which is known to accelerate convergence (Agarwal et al., 2020). Similarly, regularizing the objective using the Wasserstein distance (Pacchiano et al., 2020) rather than the KL divergence produces updates which approximate those of the Wasserstein natural policy gradient (Moskovitz et al., 2021). Other approaches can be understood as trust region or proximal point methods (Schulman et al., 2015, 2017; Touati et al., 2020), or even model-based approaches (Grill et al., 2020). It's also important to note the special case of entropy regularization, where $\Omega(\theta) = -\mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\mathsf{H}[\pi_\theta(\cdot|s)] = \mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\mathrm{KL}(\pi_\theta(\cdot|s), \mathrm{U}_\mathcal{A})$ (where $\mathrm{U}_\mathcal{X}$ denotes the uniform distribution over a space $\mathcal{X}$) which is perhaps the most common form of RPO (Haarnoja et al., 2018; Levine, 2018; Mnih et al., 2016; Williams and Peng, 1991; Schulman et al., 2018) and has been shown to aid optimization by encouraging exploration and smoothing the objective landscape (Ahmed et al., 2019).

Multitask learning Less common in the literature are policy regularizers designed explicitly for multitask settings. In many multitask RL algorithm which apply RPO, shared task structure is leveraged in other forms (e.g., importance weighting), and the regularizer itself doesn't reflect shared information (Espeholt et al., 2018; Riedmiller et al., 2018). However, in cases where the penalty is designed for multitask learning, the policy is penalized for deviating from a more general *task-agnostic* default policy meant to encode behavior which is generally useful for the *family* of tasks at hand. The use of such a behavioral default is intuitive: by distilling the common structure of the tasks the agent encounters into behaviors which have shown themselves to be useful, optimization on new tasks can be improved through with the help of prior knowledge. For example, some approaches Goyal et al. (2019, 2020) construct a default policy by marginalizing over goals g for a set of goal-conditioned policies $\pi_0(a|s) = \sum_g P(g)\pi_\theta(a|s, g)$. Such partitioning of the input into goal-dependent and goal-agnostic features can be used to create structured internal representations via an information bottleck (Tishby et al., 2000), shown empirically to improve generalization. In other multitask RPO algorithms, the default policies are derived from a Bayesian framework which views π_0 as a prior (Wilson et al., 2007; O'Donoghue et al., 2020). Still other methods learn π_0 online through distillation (Hinton et al., 2015) by minimizing $\mathrm{KL}(\pi_0, \pi)$ with respect to π_0 (Galashov et al., 2019; Teh et al., 2017). When π_0 is preserved across tasks but π_θ is re-initialized, π_0 learns the average behavior across task-specific policies. However, to our knowledge, there has been no investigation of the formal optimization properties of explicitly multitask approaches, and basic questions remain unanswered.

4 A Basic Theory for Default Policies

At an intuitive level, the question we'd like to explore is: *What properties does a default policy need in order to improve optimization?* By "improve" we refer to a reduction in the error at convergence with respect to the optimal value function or a reduction in the number of updates required to reach a given error threshold. To begin, we consider perhaps the simplest default: the uniform policy. The proofs for this section are provided in Appendix C.

4.1 Log-barrier regularization

For now, we'll restrict ourselves to the direct softmax parameterization (Eq. (2.2)) with access to exact gradients. Our default is a uniform policy over actions, i.e.: $\pi_0(a|s) = \mathrm{U}_\mathcal{A}$, resulting in the objective

$$\begin{aligned}\mathcal{J}_\lambda(\theta) &\coloneqq V^{\pi_\theta}(\mu) - \lambda\mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\left[\mathrm{KL}(\mathrm{U}_\mathcal{A}, \pi_\theta(\cdot|s))\right] \\ &\equiv V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}||\mathcal{A}|}\sum_{s,a}\log\pi_\theta(a|s),\end{aligned} \tag{4.1}$$

where we have dropped terms that are constant with respect to θ. Importantly, it's known that even this default policy has beneficial effects on optimization by erecting a log-barrier against low

values of $\pi_\theta(a|s)$. This barrier prevents gradients from quickly dropping to zero due to exponential scaling, leading to a polynomial convergence rate[1]. We now briefly restate convergence error and iteration complexity results for this case, due to Agarwal et al. (2020):

Lemma 4.1 (Error bound for log-barrier regularization)**.** *Suppose θ is such that $\|\nabla_\theta \mathcal{J}_\lambda(\theta)\|_2 \leq \epsilon_{\text{opt}}$, with $\epsilon_{\text{opt}} \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$. Then we have for all starting state distributions ρ,*

$$V^{\pi_\theta}(\rho) \geq V^\star(\rho) - \frac{2\lambda}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty .$$

We briefly comment on the term $\left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty$ (in which the division refers to component-wise division), known as the *distribution mismatch coefficient*, which roughly quantifies the difficulty of the exploration problem faced by the optimization algorithm. We note that μ is the starting distribution used for training/optimization, while the ultimate goal is to perform well on the target starting state distribution ρ. The iteration complexity is given below.

Lemma 4.2 (Iteration complexity for log-barrier regularization)**.** *Let $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. Starting from any initial $\theta^{(0)}$, consider the updates Eq. (2.4) with $\lambda = \frac{\epsilon(1-\gamma)}{2\left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty}$ and $\eta = 1/\beta_\lambda$. Then for all starting state distribution ρ, we have*

$$\min_{t \leq T} \{V^\star(\rho) - V^{(t)}(\rho)\} \leq \epsilon \quad \textit{whenever} \quad T \geq \frac{320|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty .$$

These results will act as useful reference points for the following investigation. At a minimum, we'd like a default policy to provide guarantees that are at least as good as those of log-barrier regularization.

4.2 Regularization with an α-optimal policy

To understand what properties are required of the default policy, we place an upper-bound on the suboptimality of π_0 via the TV distance. For each $s \in \mathcal{S}$, we have

$$d_{\text{TV}}(\pi^\star(\cdot|s), \pi_0(\cdot|s)) \leq \alpha(s) \tag{4.2}$$

Our regularized objective is

$$\begin{aligned} \mathcal{J}_\lambda^\alpha(\theta) &= V^{\pi_\theta}(\mu) - \lambda \mathbb{E}_{s \sim \text{U}_\mathcal{S}} \left[\text{KL}(\pi_0(\cdot|s), \pi(\cdot|s)) \right] \\ &\equiv V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}|} \sum_{s,a} \pi_0(a|s) \log \pi_\theta(a|s) \end{aligned} \tag{4.3}$$

for starting state distribution $\mu \in \Delta(\mathcal{S})$. We then have

$$\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}(\pi_0(a|s) - \pi_\theta(a|s)). \tag{4.4}$$

Our first result presents the error bound for first-order stationary points of the π_0-regularized objective.

[1] It remains an open question whether entropy regularization, which is gentler in penalizing low probabilities, produces a polynomial convergence rate.

Lemma 4.3 (Error bound for $\alpha(s)$-optimal π_0)**.** *Suppose θ is such that $\|\nabla \mathcal{J}_\lambda^\alpha(\theta)\|_\infty \leq \epsilon_{\text{opt}}$. Then we have that for all states $s \in \mathcal{S}$ and starting distributions ρ:*

$$V^{\pi_\theta}(\rho) \geq V^*(\rho) - \min\Bigg\{ \frac{1}{1-\gamma}\mathbb{E}_{s\sim \text{Unif}_\mathcal{S}}\left[\frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\max\left\{1-\alpha(s)-\frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}, 0\right\}} + \lambda\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty, \\ \frac{|\mathcal{A}|-1}{(1-\gamma)^2}\left(\mathbb{E}_{s\sim\mu}\left[\alpha(s)\right]\left\|\frac{d_\rho^{\pi_\theta}}{\mu}\right\|_\infty + \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}\right)\Bigg\}$$

The $\min\{\cdot\}$ operation above reflects the fact that the value of λ effectively determines whether reward-maximization or the regularization dominates the optimization of Eq. (4.3). Note that a similar effect also applies to log-barrier regularization, but the "high" λ setting is excluded in that instance because as $\lambda \to \infty$, $\pi_\theta(a|s) \to \text{U}_\mathcal{A}$. In this case, however, as $\alpha \to 0$, a high value of λ might be preferable, as it would amount to doing supervised learning with respect to a (nearly) optimal policy. When the reward-maximization dominates, we can see that the error bound becomes vacuous as $\alpha(s)$ approaches $\alpha^- \coloneqq 1 - (\epsilon_{\text{opt}}|\mathcal{S}|/\lambda)$ from below. In other words, as α approaches this point from below, the error can grow arbitrarily high.

In the KL-minimizing case, we can see that as the policy error $\alpha \to 0$, the value gap is given by $\frac{\epsilon_{\text{opt}}|\mathcal{S}|(|\mathcal{A}|-1)}{\lambda(1-\gamma)^2}$. Intuitively, then, as the default policy moves closer to $\pi^\star$, we can drive the value error to zero as $\lambda \to \infty$. Interestingly, we can also see that as the distribution mismatch $\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty \to 0$, the influence of the policy distance α diminishes and the error can again be driven to zero by increasing λ. We leave a more detailed discussion of the impact of the distribution mismatch coefficient to future work. Note that in most practical cases, neither α nor $\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty$ will be low enough to achieve a lower error via KL minimization alone. We will therefore focus on the reward-maximizing case ($\lambda < 1$) for the majority of our further analysis.

Before considering iteration complexity however, it's also helpful to note that Lemma 4.3 generalizes Lemma 4.1 given the same upper-bound on ϵ_{opt} as Agarwal et al. (2020).

Corollary 4.1. *Suppose θ is such that $\|\nabla \mathcal{J}_\lambda^\alpha(\theta)\|_\infty \leq \epsilon_{\text{opt}}$, with $\epsilon_{\text{opt}} \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$ and $\lambda < 1$. Then we have that for all states $s \in \mathcal{S}$,*

$$V^{\pi_\theta}(\rho) \geq V^\star(\rho) - \frac{\mathbb{E}_{s\sim \text{U}_\mathcal{S}}\left[\kappa_\mathcal{A}^\alpha(s)\right]\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty$$

where $\kappa_\mathcal{A}^\alpha(s) = \frac{2|\mathcal{A}|(1-\alpha(s))}{2|\mathcal{A}|(1-\alpha(s))-1}$.

We can see that in this case, the coefficient $\kappa_\mathcal{A}^\alpha(s)$ takes on key importance. In particular, we can see that the error-bound becomes vacuous as $\alpha(s)$ approaches $\alpha^- = 1 - 1/(2|\mathcal{A}|)$ from below. The error bound is improved with respect to log-barrier regularization when the coefficient $\kappa_\mathcal{A}^\alpha(s) < 2$, which occurs for $\alpha(s) < 1 - 1/|\mathcal{A}|$. These relationships are visualized in Fig. 4.1. We can see that the range of values over which α-optimal regularization will result in lower error than log-barrier regularization grows as the size of the action space increases. This may have implications for the use of a uniform default policy in continuous action spaces, which we leave to future work.

We can then combine this result with standard results for the convergence of gradient ascent to first order stationary points to get the iteration complexity for convergence. First, however, we require an upper bound on the smoothness of $\mathcal{J}_\lambda^\alpha$ defined in Eq. (4.3).

Lemma 4.4 (Smoothness of $\mathcal{J}_\lambda^\alpha$)**.** *For the softmax parameterization, we have that*

$$||\nabla_\theta \mathcal{J}_\lambda^\alpha(\theta) - \nabla_\theta \mathcal{J}_\lambda^\alpha(\theta')||_2 \leq \beta_\lambda ||\theta - \theta'||_2$$

where $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$.

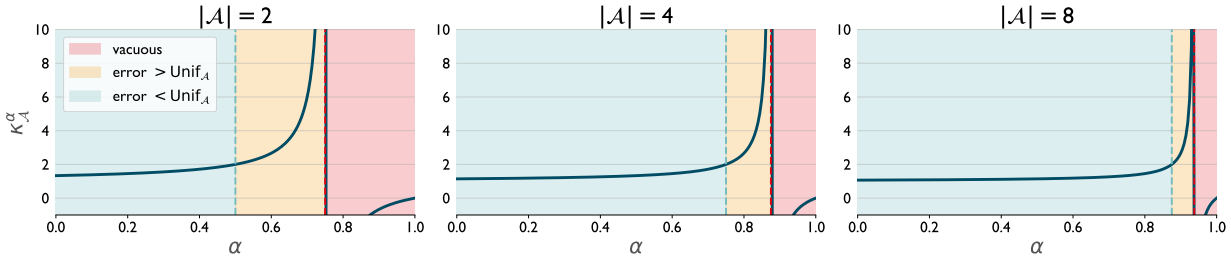

Figure 4.1: As $|\mathcal{A}|$ grows, regularizing using π_0 with larger $d_{\mathrm{TV}}(\pi^\star(\cdot|s), \pi_0(\cdot|s))$ will converge to a lower error than log-barrier regularization.

We can now bound the iteration complexity.

Lemma 4.5 (Iteration complexity for $\mathcal{J}^{\alpha}_{\lambda}$)**.** *Let ρ be a starting state distribution. Following Lemma 4.4, let $\beta_\lambda = \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. From any initial $\theta^{(0)}$ and following Eq. (2.4) with $\eta = 1/\beta_\lambda$*

$$\lambda = \frac{\epsilon(1-\gamma)}{\mathbb{E}_{s\sim \mathrm{U}_{\mathcal{S}}}\left[\kappa^{\alpha}_{\mathcal{A}}(s)\right] \left\| \frac{d^{\pi^\star}_{\rho}}{\mu} \right\|_{\infty}},$$

we have

$$\min_{t\leq T}\{V^\star(\rho) - V^{(t)}(\rho)\} \leq \epsilon \quad \text{whenever} \quad T \geq \frac{80\mathbb{E}_{s\sim \mathrm{U}_{\mathcal{S}}}\left[\kappa^{\alpha}_{\mathcal{A}}(s)\right]^2 |\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2} \left\| \frac{d^{\pi^\star}_{\rho}}{\mu} \right\|^2_{\infty}.$$

It is also natural consider the case in which π_0 is used as an initialization for π_θ.

Corollary 4.2. *Given the same assumptions as Lemma 4.5, if the initial policy is chosen to be π_0, i.e., $\pi_{\theta^{(0)}} = \pi_0$ where $\pi_0(\cdot|s)$ is $\alpha(s)$-optimal with respect to $\pi^\star(\cdot|s)$ $\forall s$, then*

$$\min_{t\leq T}\{V^\star(\rho) - V^{(t)}(\rho)\} \leq \epsilon \quad \text{whenever} \quad T \geq \frac{320|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7} \left\| \frac{d^{\pi^\star}_{\rho}}{\mu} \right\|^2_{\infty} \left\| \frac{1}{\mu} \right\|_{\infty} \mathbb{E}_{s\sim\mu}\left[\alpha(s)\right].$$

In the case of random initialization, note that when $\alpha(s) = \alpha = 1 - 1/|\mathcal{A}|$, $\mathbb{E}\kappa^{\alpha}_{\mathcal{A}}(s) = 2$, recovering the iteration complexity for log-barrier reuglarization, as expected. We also see that as the error α moves higher or lower than $1 - 1/|\mathcal{A}|$, the iteration complexity grows or shrinks quadratically. Therefore, a default policy within this range will not only linearly reduce the error at convergence, but will also quadratically increase the rate at which that error is reached. When the initial policy is π_0, the iteration complexity depends on the factor $\mathbb{E}_{s\sim \mathrm{U}_{\mathcal{S}}}[\alpha(s)]$. Hence, for good initialization, α is small, resulting in fewer iterations. The natural question, then, is how to find such a default policy, with high probability, for some family of tasks.

5 Extension to Multitask Learning

These results provide guidance on the construction of default policies in the multitask setting. The key insight is that if the optimal policies for the tasks drawn from a given task distribution have commonalities, the agent can use the optimal policies it learns from previous tasks to construct a useful π_0. More precisely, consider a distribution $\mathcal{P}_{\mathcal{M}}$ over a family of tasks $\mathcal{M} \coloneqq \{M_k\}$ with shared state and action spaces $\mathcal{S}$ and $\mathcal{A}$ and optimal deterministic policies $\{\pi^\star_k\}$. We assume that the other task components (reward function, transition distribution, etc.) are independent. Then by Corollary 4.1 and Lemma 4.5, if the TV barycenter of these policies at a given state s, given by

$$\pi_0(\cdot|s) = \operatorname*{argmin}_{\pi} \mathbb{E}_{M_k\sim\mathcal{P}_{\mathcal{M}}}\left[d_{\mathrm{TV}}(\pi^\star_k(\cdot|s), \pi(\cdot|s))\right] \tag{5.1}$$

is such that $\mathbb{E}[d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \pi_0(\cdot|s))] < 1 - 1/|\mathcal{A}|$, then regularizing with π_0 will, in expectation, result in faster convergence and lower error than using a uniform distribution. Crucially, when there is a lack of shared structure, which in this particular approach is manifested as a lack of agreement among optimal policies, $\pi_0(\cdot|s)$ collapses to $\mathrm{U}_{\mathcal{A}}$. Therefore, in the worst case, regularizing with $\pi_0(\cdot|s)$ can do no worse than log-barrier regularization, which already enjoys polynomial iteration complexity.

When the optimal policies $\{\pi_k^\star\}$ are deterministic, the following result gives a convenient expression for the TV-barycenter policy:

Lemma 5.1 (TV barycenter)**.** *Let $\mathcal{P}_{\mathcal{M}}$ be a distribution over tasks $\mathcal{M} = \{M_k\}$, each with a deterministic policy $\pi_k^\star : \mathcal{S} \to \mathcal{A}$. Define the average optimal action as*

$$\xi(s, a) \coloneqq \mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}} \left[\mathbb{1}(\pi_k^\star(s) = a) \right]. \tag{5.2}$$

Then, the TV barycenter $\pi_0(\cdot|s)$ defined in Eq. (5.1) *is given by a greedy policy over ξ, i.e., $\pi_0(a|s) = \delta(a \in \operatorname{argmax}_{a' \in \mathcal{A}} \xi(s, a'))$, where $\delta(\cdot)$ is the Dirac delta distribution.*

The proof, along with the rest of the proofs for this section, is provided in Appendix D. Interestingly, this result also holds for the KL barycenter, which we show in Appendix Lemma D.1. Because the average optimal action ξ is closely related to a recently-proposed computational model of *habit formation* in cognitive psychology (Miller et al., 2016), from now on we refer to it as the *habit function* for task family $\mathcal{M}$. When the agent has observed K tasks sampled from $\mathcal{P}_{\mathcal{M}}$, ξ is approximated by the sample average $\hat{\xi}(s, a) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(\pi_k^\star(s) = a)$ provided that the optimal policies $\pi_k^\star$ are available. In practice, however, the agent only has access to an approximation $\tilde{\pi}_k$ of $\pi_k^\star$, for instance, through the use of a learning algorithm A, such as Appendix Algorithm 3. Hence, $\hat{\xi}(s, a)$ is instead, given by $\hat{\xi}(s, a) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(\tilde{\pi}_k(s) = a)$ which induces an approximate barycenter $\hat{\pi}_0$ by taking the greedy policy over $\hat{\xi}$. The following result provides the iteration complexity for the multitask setting when using $\hat{\pi}_0$ as the default policy.

Lemma 5.2 (Multitask iteration complexity)**.** *Let $M_k \sim \mathcal{P}_{\mathcal{M}}$ and denote by $\pi_k^\star : \mathcal{S} \to \mathcal{A}$ its optimal policy. Denote by T_k the number of iterations to reach ϵ-error for M_k in the sense that:*

$$\min_{t \leq T_k} \{V^*(\rho) - V^{(t)}(\rho)\} \leq \epsilon.$$

Set λ, β_λ, and η as in Lemma 4.5. From any initial $\theta^{(0)}$, and following Eq. (2.4)*, $\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}}[T_k]$ satisfies:*

$$\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}}[T_k] \geq \frac{80|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^6} \mathbb{E}_{\substack{M_k \sim \mathcal{P}_{\mathcal{M}} \\ s \sim \mathrm{U}_{\mathcal{S}}}} \left[\kappa_{\mathcal{A}}^{\alpha_k}(s) \left\| \frac{d_\rho^{\pi_k^*}}{\mu} \right\|_\infty^2 \right],$$

where $\alpha_k(s) \coloneqq d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \hat{\pi}_0(\cdot|s))$. If $\hat{\pi}_0$ is also used for initialization, then $\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}}[T_k]$ satisfies:

$$\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}}[T_k] \geq \frac{320|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7} \left\| \frac{1}{\mu} \right\|_\infty^3 \mathbb{E}_{\substack{M_k \sim \mathcal{P}_{\mathcal{M}} \\ s \sim \mu}} \left[\alpha_k(s) \right],$$

Lemma 5.2 characterizes the average iteration complexity over tasks when using $\hat{\pi}_0$ as a default policy. In particular, when the learning algorithm is also initialized with $\hat{\pi}_0$, we obtain that the average iteration to reach ϵ accuracy is proportional to the expected TV-distance of $\hat{\pi}_0$ to optimal policies $\pi_k^\star$ for tasks M_k. We expect this distance to approach $\mathbb{E}\left[d_{\mathrm{TV}}(\pi_0(\cdot|s), \pi_k^\star(\cdot|s))\right]$ as the number of tasks increases and $\tilde{\pi}_k$ become more accurate. Note that even in this case, the regularization is *still* required to assure polynomial convergence. To provide a precise quantification, we let $\tilde{\pi}_k(\cdot|s)$ be, on average, $\zeta(s)$-optimal in state s across tasks M_k, i.e. $\mathbb{E}_{M_k \sim \mathcal{M}}\left[d_{\mathrm{TV}}(\tilde{\pi}_k(\cdot|s), \pi_k^\star(\cdot|s))\right] \leq \zeta(s)$ for some $\zeta(s) \in [0, 1]$. The following lemma quantifies how close $\hat{\pi}_0$ grows to the TV barycenter of $\{\pi_k^\star\}_{k=1}^K$ as $K \to \infty$:

Algorithm 1: TV Policy Optimization (TVPO)

1: **Input** Task set $\mathcal{M}$, policy class Θ, fixed-π_0 RPO algorithm $A(M, \Theta, \pi_0, \lambda)$, as in Appendix Algorithm 3
2: initialize $\pi_0(\cdot|s) = \xi^{(0)}(s, .) = \mathrm{U}_{\mathcal{A}}\ \forall s \in \mathcal{S}$
3: **for** iteration $k = 1, 2, ...$ **do**
4: Sample a task $M^{(k)} \sim \mathcal{P}_{\mathcal{M}}$
5: Solve the task: $\tilde{\theta}^{(k)} = A(M_k, \Theta, \pi_0^{(k-1)}, \lambda)$
6: Set $\tilde{\pi}_k \leftarrow \pi_{\tilde{\theta}^{(k)}}$.
7: Update habit moving average $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\xi^{(k)}(s, a) \leftarrow \frac{k-1}{k}\xi^{(k-1)}(s, a) + \frac{1}{k}\mathbb{1}\left(a = \operatorname*{argmax}_{a'} \tilde{\pi}_k(a'|s)\right)$$

8: Update default policy $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\pi_0^{(k)}(a|s) \propto \exp(\xi^{(k)}(s, a)/\beta(k))$$

9: **end for**

Lemma 5.3 (Barycenter concentration)**.** *Let δ be $0 < \delta < 1$. Then with probability higher than $1 - \delta$, for all $s \in \mathcal{S}$, it holds that:*

$$\begin{aligned}|\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}}[d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \hat{\pi}_0(\cdot|s)) - d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \pi_0(\cdot|s))]| \\ \leq 2\zeta(s) + \sqrt{\frac{2\log(\frac{2}{\delta})}{K}} + 2C\sqrt{\frac{|\mathcal{A}|}{K}},\end{aligned}$$

for some constant C that depends only on $|\mathcal{A}|$.

In other words, in order to produce a default policy which improves over log-barrier regularization as $K \to \infty$, the margin of error for the trained policies is half that which is required for the default policy.

In practice, due to the epistemic uncertainty about the task family early in training, regularizing using $\hat{\pi}_0$ risks misleading π_θ by placing all of the default policy's mass on a sub-optimal action. We can therefore define $\hat{\pi}_0$ using a softmax $\hat{\pi}_0(a|s) \propto \exp(\hat{\xi}(s, a)/\beta(K))$ with some temperature parameter $\beta(K)$ tending to zero as the number of observed tasks K approaches infinity so that π_0 converges to the optimal default policy in the limit. This suggests the simple approach to multitask RPO presented in Algorithm 1, which we call *total variation policy optimization* (TVPO). Note that if $\mathcal{P}_{\mathcal{M}}$ is non-stationary, the moving average in Line 8 can be changed to an exponentially weighted moving average to place more emphasis on recent tasks.

6 Understanding the Literature

As stated previously, many approaches to multitask RPO in the literature learn a default policy $\pi_0(a|s; \phi)$ parameterized by ϕ via gradient descent on the KL divergence (Galashov et al., 2019; Teh et al., 2017), e.g., via

$$\phi = \operatorname*{argmin}_{\phi'} \mathbb{E}_{s \sim \mathrm{U}_{\mathcal{S}}}\left[\mathrm{KL}(\pi_\theta(\cdot|s), \pi_0(\cdot|s; \phi))\right]. \tag{6.1}$$

The idea is that by updating ϕ across multiple tasks, π_0 will acquire the average behaviors of the goal-directed policies π_θ. This objective can be seen as an approximation of Eq. (5.1) in which we

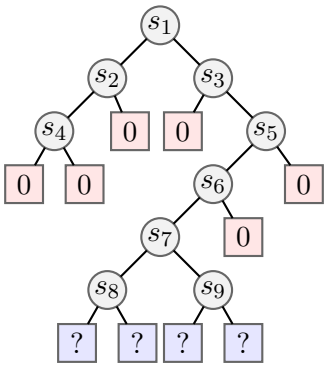

Figure 7.1: A tree environment. Each task within the family randomly distributes rewards among the leaves marked with a '?', while all other states result in zero reward.

can view the use of the KL as a relaxation of the TV distance:

$$\begin{aligned}\pi_0(\cdot|s) &= \operatorname*{argmin}_{\pi} \mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}} \left[d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \pi(\cdot|s)) \right] \\ &\leq \operatorname*{argmin}_{\pi} \mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}} \left[d_{\mathrm{TV}}(\pi_k^\star(\cdot|s), \pi(\cdot|s))^2 \right] \\ &\leq \operatorname*{argmin}_{\pi} \mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}} \left[\mathrm{KL}(\pi_k^\star(\cdot|s), \pi(\cdot|s)) \right],\end{aligned}$$

where the first inequality is due to Jensen's inequality and the second is due to Pollard (2000) and where $\pi_\theta(\cdot|s) \approx \pi^\star(\cdot|s)$. The use of the KL is natural due to its easy computation and differentiability, however the last approximation is crucial. By distilling π_0 from π_θ via Eq. (6.1) from the outset of each task, there is an implicit assumption that $\pi_\theta \approx \pi^\star$ even early in training. This is a source of suboptimality, as we discuss in Section 7.

7 Experiments

We now study the implications of these ideas in a simple empirical setting: a family of tasks whose state space follows the tree structure shown in Fig. 7.1. In these tasks, the agent starts at the root s_1 and at each timestep chooses whether to proceed down its left subtree or right subtree ($|\mathcal{A}| = 2$). The episode ends when the agent reaches a leaf node. In this setup, there is zero reward in all states other than the leaf nodes marked with a '?', for which one or more are randomly assigned a reward of one for each draw from the task distribution, with the number of rewards drawn from a geometric distribution with success parameter $p = 0.5$ to encourage sparsity. One training run consisted of five rounds of randomly sampling a task and solving it. Despite the simplicity of this environment, we found that it could prove surprisingly difficult for many algorithms to solve consistently. As can be seen in Fig. 7.1, the key structural consistency in this task is that every optimal policy makes the same choices in states $\{s_1, s_3, s_5, s_6\}$, with any exploration limited to the lower subtree rooted at s_7.

For comparison, we selected RPO approaches with both *fixed* default policies (LOG-BARRIER, ENTROPY, and NONE) and *learned* default policies: DISTRAL ($-\mathrm{KL}(\pi_\theta, \pi_0) + \mathsf{H}[\pi_\theta]$; (Teh et al., 2017)), FORWARD KL ($-\mathrm{KL}(\pi_0, \pi_\theta)$), and REVERSE KL ($-\mathrm{KL}(\pi_\theta, \pi_0)$). To make the problem more challenging for the learned default policies, the reward distribution was made sparser by setting $p = 0.7$. Each approach was applied over 20 random seeds, with results plotted in Fig. 7.2 (fixed π_0) and Fig. 7.3 (learned π_0), where we see that TVPO most consistently solves the tasks. Hyperparameters were kept constant across methods (further experimental details can be found in Appendix E). We can see that TVPO matches or outperforms all other algorithms. This is not surprising, as $\mathbb{E}_{M_k \sim \mathcal{P}_{\mathcal{M}}} [\alpha_k(s)] = 0$ for all states en route to the rewarded leaves until s_7. Thus, $\hat{\pi}_0(\cdot|s) \to \pi_k^\star(s)$ quickly for these states as the number of tasks grows. This dramatically reduces the size of the exploration problem for TVPO, confining it to the subtree rooted at s_7.

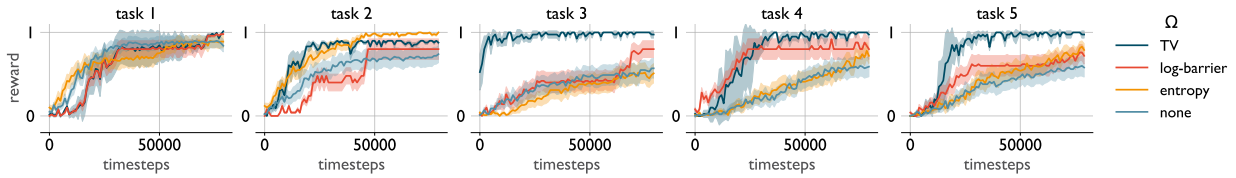

Figure 7.2: Fixed π_0 baselines. Results are averaged over 20 seeds, with the shaded region denoting one standard deviation.

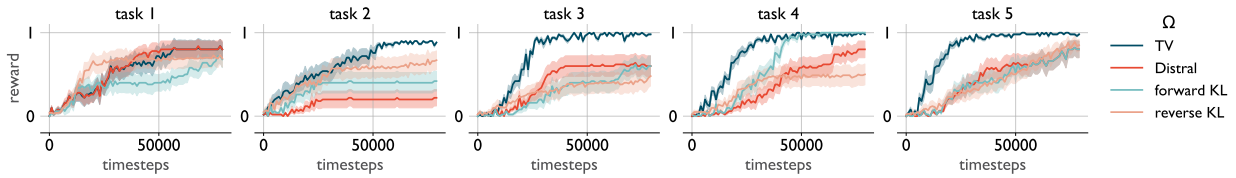

Figure 7.3: Learned π_0 baselines. Results are averaged over 20 seeds, with the shaded region denoting one standard deviation.

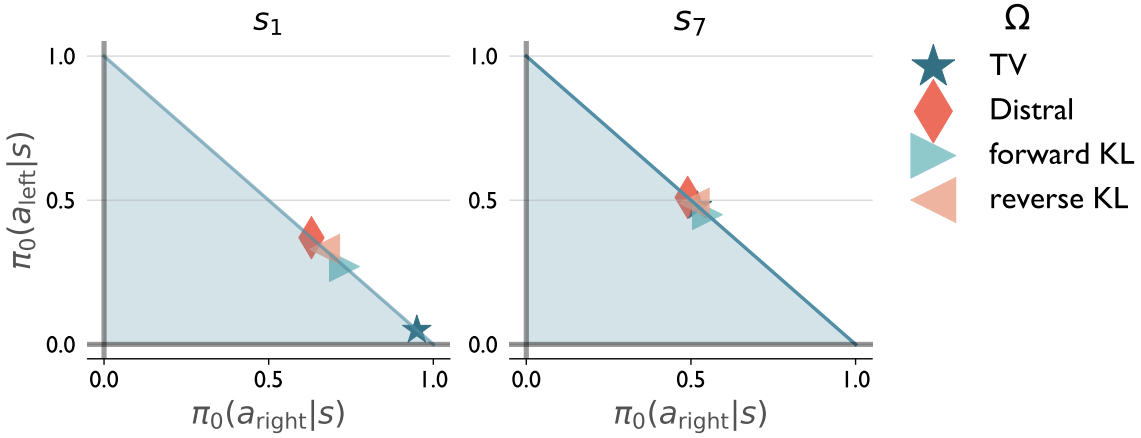

Figure 7.4: Learned default policies in states s_1 and s_7 after five tasks. In the simplex for s_7, the marker for TVPO is behind the markers for the other methods.

To gain a better understanding of the results and the learned default policies, we plotted the average default policies for each method on the 2-simplex for states s_1 and s_7 in Fig. 7.4. For all tasks in the family, the optimal policy goes right in s_1, while, on average, reward could be located in either subtree rooted at s_7. This is reflected in the default policies, which prefer right in s_1 and are close to uniform in s_7. There is a notable difference, however, in that the KL-, gradient-based methods are much less deterministic in s_1. The critical difference is that the KL-based methods are trained online via distillation from suboptimal $\pi_\theta \not\approx \pi^\star$. Early in training, π_θ is inconsistent across tasks and runs, resulting in a more uniform target for π_0. This delays its convergence across tasks to the shared TV/KL barycenter. To test this effect empirically, we repeated the same experiment with REVERSE KL but started training π_0 progressively later within each task.

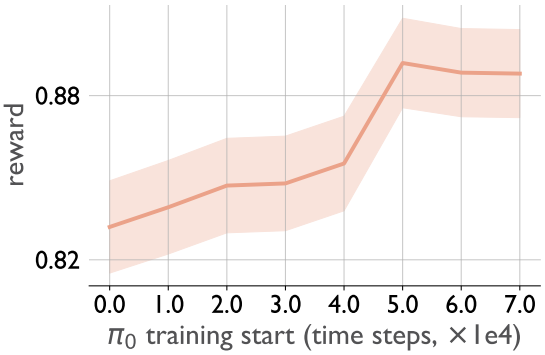

Figure 7.5: Delayed training of π_0 improves performance.

Fig. 7.5 depicts the average final reward across tasks for different time steps at which the default policy began training. Note, however, that π_0 is still used to regularize π_θ, it just isn't updated based on π_θ until π_θ is a reasonable approximation of $\pi^\star$. We can see that, as predicted, delaying training within each task improves performance. There is a slight drop performance in performance if π_0 does not have a sufficient number of updates at the end of training.

8 Discussion

In this work, we introduce novel, more general bounds on the error and iteration complexity of KL-regularized policy optimization. We then show how these bounds apply to the multitask setting, showing the first formal results for a popular class of algorithms and deriving a novel multitask RPO algorithm with formal guarantees. We then demonstrate the implications of these results in a simple experimental setting.

There are several important implications for future work. First, these results imply an algorithm-dependent definition of task families, such that a group of tasks can be considered a *family* for a given algorithm if that algorithm can leverage their shared properties to improve optimization. For RPO algorithms, then, the choice of divergence measure, default policy, and distribution space implicitly determines task groupings. As an example, the particular class of algorithm we investigate here is sensitive to state-dependent similarities in the space of optimal policies for a group of tasks. There are a multitude of other forms of shared structure which alternative approaches can leverge, however, such consistent transition dynamics (Barreto et al., 2020) or even structure in an abstract *behavioral* space (Pacchiano et al., 2020; Moskovitz et al., 2021; Agarwal et al., 2021). Conducting an effective taxonomy of algorithms and associated task families will be crucial for the development of practical real-world agents.

We also believe this work provides a formal framework for settings where forward transfer is possible during lifelong learning scenario with multiple interrelated tasks (Lopez-Paz and Ranzato, 2017). While we test these ideas in a toy setting, the underlying theory has implications for state-of-the-art deep RL methods. When state and action spaces grow large, however, π_0 is necessarily represented by a restricted policy class. Both TVPO and the learned π_0 baseline methods can be scaled to this domain, with TVPO's π_0 being trained online to predict the next action taken by π_θ. One useful lesson which equally applies to KL-based methods, however, is that it's preferable from an optimization standpoint to distill π_0 from π_θ only late in training when $\pi_\theta \approx \pi^\star$. Given the promise of this general class of methods, we hope that the insight garnered by these results will help propel the field towards more robust and general algorithms.

References

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018.

Hiroki Furuta, Tadashi Kozuno, Tatsuya Matsushima, Yutaka Matsuo, and Shixiang Shane Gu. Co-adaptation of algorithmic and implementational innovations in inference-based deep reinforcement learning, 2021.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *CoRR*, abs/1912.06680, 2019.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures, 2018.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Anna Choromanska, Krzysztof Choromanski, and Michael I Jordan. Learning to score behaviors for guided policy optimization. In *The International Conference on Machine Learning*. 2020.

Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. In *International Conference on Learning Representations*. ICLR, 2021.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015. URL `http://arxiv.org/abs/1502.05477`.

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

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

Joseph Marino, Alexandre Piché, Alessandro Davide Ialongo, and Yisong Yue. Iterative amortized policy optimization, 2020.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation, 2018.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 2020. URL `https://proceedings.mlr.press/v125/agarwal20a.html`.

Alexandre Galashov, Siddhant M. Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M. Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized RL. *CoRR*, abs/1905.01240, 2019. URL `http://arxiv.org/abs/1905.01240`.

Yee Whye Teh, Victor Bapst, Wojciech Marian Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4499–4509, 2017.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Yoshua Bengio, and Sergey Levine. Infobot: Transfer and exploration via the information bottleneck, 2019.

Anirudh Goyal, Yoshua Bengio, Matthew Botvinick, and Sergey Levine. The variational bandwidth bottleneck: Stochastic evaluation on an information budget, 2020.

Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley and Sons, 2010.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 267–274, 2002.

Jean-Bastien Grill, Florent Altche, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Remi Munos. Monte-carlo tree search as regularized policy optimization, 2020.

H. Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W. Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Dan Belov, Martin Riedmiller, and Matthew M. Botvinick. V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control, 2019.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2021. URL `https://openreview.net/forum?id=ToWi1RjuEr8`.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. {AWAC}: Accelerating online reinforcement learning with offline datasets, 2021. URL `https://openreview.net/forum?id=OJiM1R3jAtZ`.

Jan Peters, Katharina Mülling, and Yasemin Altün. Relative entropy policy search. *AAAI Conference on Artificial Intelligence*, 2010.

Emanuel Todorov. Linearly-solvable markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. URL `https://proceedings.neurips.cc/paper/2006/file/d806ca13ca3449af72a1ea5aedbed26a-Paper.pdf`.

Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. volume 2006, pages 945–952, 01 2006. doi: 10.1145/1143844.1143963.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference (extended abstract). In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 3052–3056. AAAI Press, 2013.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 202–211, Arlington, Virginia, USA, 2016. AUAI Press.

Ahmed Touati, Amy Zhang, Joelle Pineau, and Pascal Vincent. Stable policy optimization via off-policy divergence regularization. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of

Proceedings of Machine Learning Research, pages 1328–1337. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/touati20a.html.

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 Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/mniha16.html.

Ronald Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3:241–, 09 1991. doi: 10.1080/09540099108946587.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning, 2018.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ahmed19a.html.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Van de Wiele, Volodymyr Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing - solving sparse reward tasks from scratch, 2018.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv e-prints*, April 2000.

Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: A hierarchical bayesian approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 1015–1022, New York, NY, USA, 2007. Association for Computing Machinery. doi: 10.1145/1273496.1273624. URL https://doi.org/10.1145/1273496.1273624.

Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xitgHtvS.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Kevin J. Miller, Amitai Shenhav, and Elliot A. Ludvig. Habits without values. *bioRxiv*, 2016. doi: 10.1101/067603. URL https://www.biorxiv.org/content/early/2016/08/03/067603.

David Pollard. *Chapter 3*. 2000. URL http://www.stat.yale.edu/~pollard/Books/Asymptopia.

Andre Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117 (48):30079–30087, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907370117. URL https://www.pnas.org/content/117/48/30079.

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning, 2021.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. 2018.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A Appendix Overview

Appendix B contains policy gradients pseudocode, Appendix C contains proofs for the single-task results (including additional results for state-dependent λ and ϵ), Appendix D contains proofs for the multitask RPO results, and Appendix E contains experimental details.

B Generic Policy Optimization Algorithms

Algorithm 2: Generic policy gradient algorithm

Input MDP M, policy class Θ
initialize $\theta^{(0)} \in \Theta$
for iteration $k = 0, 1, 2, \dots$ **do**
  sample a trajectory:

$$\tau = (s_0, a_0, s_1, \dots) \sim \Pr_\mu^{\pi_{\theta^{(k)}}}(\cdot) = \mu(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \pi_{\theta^{(k)}}(a_t|s_t)$$

  update parameters:

$$\theta^{(k+1)} = \theta^{(k)} + \eta \widehat{\nabla V^{\pi_\theta}}(\mu)$$

  where

$$\widehat{\nabla V^{\pi_\theta}}(\mu) = \sum_{t=0}^{\infty} \gamma^t \widehat{Q^{\pi_\theta}}(s_t, a_t) \nabla \log \pi_\theta(a_t|s_t), \qquad \widehat{Q^{\pi_\theta}}(s_t, a_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

end for

Algorithm 3: Regularized policy gradient algorithm

Input MDP M, policy class Θ, regularization strength λ, default policy π_0
initialize $\theta^{(0)} \in \Theta$
for iteration $k = 0, 1, 2, \dots, K$ **do**
  sample a trajectory:

$$\tau = (s_0, a_0, s_1, \dots) \sim \Pr_\mu^{\pi_{\theta^{(k)}}}(\cdot) = \mu(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \pi_{\theta^{(k)}}(a_t|s_t)$$

  update parameters:

$$\theta^{(k+1)} = \theta^{(k)} + \eta \widehat{\nabla_{\theta^{(k)}} \mathcal{J}_\lambda}(\theta^{(k)})$$

  where

$$\widehat{\nabla_\theta \mathcal{J}_\lambda}(\theta) = \widehat{\nabla_\theta V^{\pi_\theta}}(\mu) - \lambda \nabla_\theta \Omega(\pi_0, \pi_\theta)$$

  and $\widehat{\nabla_\theta V^{\pi_\theta}}(\mu)$ is as in Algorithm 2.
end for
return $\theta^{(K)}$

C Single-task results

We now consider the error bound for π_0 such that $d_{\mathrm{TV}}(\pi^*(\cdot|s), \pi_0(\cdot|s)) \leq \alpha(s)\ \forall s \in \mathcal{S}$.

Lemma 4.3 (Error bound for $\alpha(s)$-optimal π_0)**.** *Suppose θ is such that $\|\nabla \mathcal{J}_\lambda^\alpha(\theta)\|_\infty \leq \epsilon_{\mathrm{opt}}$. Then we have that for all states $s \in \mathcal{S}$ and starting distributions ρ:*

$$V^{\pi_\theta}(\rho) \geq V^*(\rho) - \min\Bigg\{ \frac{1}{1-\gamma}\mathbb{E}_{s\sim \mathrm{Unif}_{\mathcal{S}}}\left[\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\max\left\{1-\alpha(s)-\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}, 0\right\}} + \lambda\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty, \\ \frac{|\mathcal{A}|-1}{(1-\gamma)^2}\left(\mathbb{E}_{s\sim\mu}\left[\alpha(s)\right]\left\|\frac{d_\rho^{\pi_\theta}}{\mu}\right\|_\infty + \frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}\right)\Bigg\}$$

Proof. Let's assume that π^* is a deterministic policy. By Puterman (2010) such an optimal policy always exists for an MDP. We'll use the notation $a^*(s)$ to denote the optimal action at state s. This, combined with the assumption that $d_{\mathrm{TV}}(\pi^*(\cdot|s), \pi_0(\cdot|s)) \leq \alpha(s)$ for all $s \in \mathcal{S}$, tells us that $\pi_0(a^*(s)|s) \geq \pi^*(a^*(s)|s) - \alpha(s) = 1 - \alpha(s)$. Similarly, for $a \neq a^*(s)$, $\pi_0(a|s) \leq \alpha(s)$. Using this, we can start by showing that whenever $A^{\pi_\theta}(s, a^*(s)) \geq 0$ we can lower bound $\pi_\theta(a^*(s)|s)$ for all s.

The gradient norm assumption $\|\nabla \mathcal{J}_\lambda^\alpha(\theta)\|_\infty \leq \epsilon_{\mathrm{opt}}$ implies that for all s, a:

$$\epsilon_{\mathrm{opt}} \geq \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}(\pi_0(a|s) - \pi_\theta(a|s))$$

In particular for all s,

$$\begin{aligned}\epsilon_{\mathrm{opt}} \geq \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a^*(s)}} &\overset{(i)}{\geq} \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a^*(s)|s)A^{\pi_\theta}(s,a^*(s)) + \frac{\lambda}{|\mathcal{S}|}(\pi^*(a^*(s)|s) - \alpha(s) - \pi_\theta(a^*(s)|s)) \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a^*(s)|s)A^{\pi_\theta}(s,a^*(s)) + \frac{\lambda}{|\mathcal{S}|}(1 - \alpha(s) - \pi_\theta(a^*(s)|s))\end{aligned} \tag{C.1}$$

And therefore if $A^{\pi_\theta}(s, a^*(s)) \geq 0$,

$$\epsilon_{\mathrm{opt}} \geq \frac{\lambda}{|\mathcal{S}|}(1 - \alpha(s) - \pi_\theta(a^*(s)|s))$$

Thus,

$$\pi_\theta(a^*(s)|s) \geq 1 - \alpha(s) - \frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}. \tag{C.2}$$

We then have

$$\begin{aligned}A^{\pi_\theta}(s, a^*(s)) &\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a^*(s)|s)}\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} - \frac{\lambda}{|\mathcal{S}|}\frac{1}{\pi_\theta(a^*(s)|s)}\left(1 - \alpha(s) - \pi_\theta(a^*(s)|s)\right)\right) \\ &= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a^*(s)|s)}\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{|\mathcal{S}|}\left(1 - \frac{1-\alpha(s)}{\pi_\theta(a^*(s)|s)}\right)\right) \\ &\overset{(i)}{\leq} \frac{1}{\mu(s)}\left(\frac{1}{\max\left\{1-\alpha(s)-\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}, 0\right\}}\cdot\epsilon_{\mathrm{opt}} + \frac{\lambda}{|\mathcal{S}|}\left(1 - (1-\alpha(s))\right)\right) \\ &\leq \frac{1}{\mu(s)}\left(\frac{1}{\max\left\{(1-\alpha(s))-\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}, 0\right\}}\cdot\epsilon_{\mathrm{opt}} + \frac{\lambda}{|\mathcal{S}|}\alpha(s)\right)\end{aligned}$$

where (i) follows because $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$, $\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} \leq \epsilon_{\mathrm{opt}}$ and $\max(1 - \alpha(s) - \frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|}{\lambda}, 0) \leq \pi_\theta(a^*(s)|s) \leq 1$. Then applying the performance difference lemma (Kakade and Langford, 2002)

gives

$$
\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^*}(s) \pi^*(a|s) A^{\pi_\theta}(s,a) \\
&= \frac{1}{1-\gamma} \sum_{s} d_\rho^{\pi^*}(s) A^{\pi_\theta}(s, a^*(s)) \\
&\leq \frac{1}{1-\gamma} \sum_{s} d_\rho^{\pi^*}(s) A^{\pi_\theta}(s, a^*(s)) \mathbb{1}(A^{\pi_\theta}(s, a^*(s)) \geq 0) \\
&\leq \frac{1}{1-\gamma} \sum_{s} \frac{d_\rho^{\pi^*}(s)}{\mu(s)} \left(\frac{1}{\left\{1 - \alpha(s) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}, 0\right\}} \cdot \epsilon_{\text{opt}} + \frac{\lambda}{|\mathcal{S}|} \alpha(s) \right) \mathbb{1}(A^{\pi_\theta}(s, a^*(s)) \geq 0) \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim \text{Unif}_{\mathcal{S}}} \left[\frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\left\{1 - \alpha(s) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}, 0\right\}} + \lambda \alpha(s) \right] \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty .
\end{aligned}
$$

Now let's relate the values of π^* and π_θ. We will again apply the performance difference lemma, this time in the other direction:

$$
\begin{aligned}
V^{\pi_\theta}(\rho) - V^*(\rho) &= \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi^*}(s,a) \\
&\overset{(1)}{=} \frac{1}{1-\gamma} \sum_{s} \left(\sum_{a \neq a^*(s)} d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi^*}(s,a) \right) \\
&\overset{(2)}{=} \frac{-1}{1-\gamma} \sum_{s} d_\rho^{\pi_\theta}(s) \left(\alpha(s) + \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \right) \frac{|\mathcal{A}| - 1}{1-\gamma} \\
&= -\frac{|\mathcal{A}| - 1}{1-\gamma} \left(\sum_{s} \frac{d_\rho^{\pi_\theta}(s)}{1-\gamma} \alpha(s) + \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \sum_{s} \frac{d_\rho^{\pi_\theta}(s)}{1-\gamma} \right) \quad (\text{b/c } \sum_s d_\rho^{\pi_\theta}(s) = 1) \\
&= -\frac{|\mathcal{A}| - 1}{1-\gamma} \left(\sum_{s} \frac{d_\rho^{\pi_\theta}(s)}{1-\gamma} \alpha(s) \right) - \frac{(|\mathcal{A}| - 1)\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda(1-\gamma)^2} \\
&= -\frac{|\mathcal{A}| - 1}{(1-\gamma)^2} \left(\sum_{s} d_\rho^{\pi_\theta}(s) \alpha(s) \right) - \frac{(|\mathcal{A}| - 1)\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda(1-\gamma)^2} \\
&= -\frac{|\mathcal{A}| - 1}{(1-\gamma)^2} \left(\sum_{s} \frac{d_\rho^{\pi_\theta}(s)}{\mu(s)} \mu(s) \alpha(s) \right) - \frac{(|\mathcal{A}| - 1)\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda(1-\gamma)^2} \\
&\geq -\frac{|\mathcal{A}| - 1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi_\theta}}{\mu} \right\|_\infty \left(\sum_{s} \mu(s) \alpha(s) \right) - \frac{(|\mathcal{A}| - 1)\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda(1-\gamma)^2}
\end{aligned}
$$

where (1) is due to the fact that $A^{\pi^*}(s, a^*(s)) = 0$, and (2) is due to the fact that $A^{\pi^*}(s,a)$ for $a \neq a^*$ is lower-bounded by $-1/(1-\gamma)$ and Eq. (C.21). Therefore,

$$
V^{\pi_\theta}(\rho) + \frac{|\mathcal{A}| - 1}{(1-\gamma)^2} \left(\mathbb{E}_{s \sim \mu} [\alpha(s)] \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty + \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \right) \geq V^*(\rho).
$$

This completes the proof. □

We now present a comparatively looser bound which applies the same upper bound on the norm of the gradient used by Agarwal et al. (2020).

Corollary 4.1. *Suppose θ is such that $\|\nabla \mathcal{J}_\lambda^\alpha(\theta)\|_\infty \leq \epsilon_{\text{opt}}$, with $\epsilon_{\text{opt}} \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}$ and $\lambda < 1$. Then we have that for all states $s \in \mathcal{S}$,*

$$V^{\pi_\theta}(\rho) \geq V^\star(\rho) - \frac{\mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\left[\kappa_\mathcal{A}^\alpha(s)\right]\lambda}{1-\gamma}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty$$

where $\kappa_\mathcal{A}^\alpha(s) = \frac{2|\mathcal{A}|(1-\alpha(s))}{2|\mathcal{A}|(1-\alpha(s))-1}$.

Proof. The proof proceeds as in Lemma 4.3, except that we use the upper bound on ϵ_{opt} in Eq. (C.21) to get

$$\pi_\theta(a^*(s)|s) \geq 1 - \alpha(s) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda} \geq 1 - \alpha(s) - \frac{1}{2|\mathcal{A}|} = \frac{2|\mathcal{A}|(1-\alpha(s))-1}{2|\mathcal{A}|} \tag{C.3}$$

In this case we can upper bound $A^{\pi_\theta}(s, a^*(s))$. From Eq. (C.1) inequality (i), we have

$$\begin{aligned}
A^{\pi_\theta}(s,a^*(s)) &\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a^*(s)|s)}\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} - \frac{\lambda}{|\mathcal{S}|}\frac{1}{\pi_\theta(a^*(s)|s)}\left(1-\alpha(s)-\pi_\theta(a^*(s)|s)\right)\right)\\
&= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_\theta(a^*(s)|s)}\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{|\mathcal{S}|}\left(1 - \frac{1-\alpha(s)}{\pi_\theta(a^*(s)|s)}\right)\right)\\
&\overset{(i)}{\leq} \frac{1}{\mu(s)}\left(\frac{1}{(1-\alpha(s)) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}}\cdot\epsilon_{\text{opt}} + \frac{\lambda}{|\mathcal{S}|}\left(1-(1-\alpha(s))\right)\right)\\
&\leq \frac{1}{\mu(s)}\left(\frac{1}{(1-\alpha(s)) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}}\cdot\epsilon_{\text{opt}} + \frac{\lambda}{|\mathcal{S}|}\alpha(s)\right)\\
&\overset{(ii)}{\leq} \frac{1}{\mu(s)}\left(\frac{2|\mathcal{A}|}{(2|\mathcal{A}|(1-\alpha(s))-1)}\frac{\lambda}{2|\mathcal{S}||\mathcal{A}|} + \frac{\lambda}{|\mathcal{S}|}\alpha(s)\right)\\
&= \frac{\lambda}{|\mathcal{S}|\mu(s)}\left(\frac{1}{2|\mathcal{A}|(1-\alpha(s))-1} + \underbrace{\alpha(s)}_{\leq 1}\right)\\
&\leq \frac{\lambda}{|\mathcal{S}|\mu(s)}\left(\underbrace{\frac{2|\mathcal{A}|(1-\alpha(s))}{2|\mathcal{A}|(1-\alpha(s))-1}}_{:=\kappa_\mathcal{A}^\alpha(s)}\right)
\end{aligned}$$

Where (i) follows because $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$, $\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} \leq \epsilon_{\text{opt}}$ and $\max(1-\alpha(s) - \frac{\epsilon_{\text{opt}}|\mathcal{S}|}{\lambda}, 0) \leq \pi_\theta(a^*(s)|s) \leq 1$. (ii) is obtained by plugging in the upper bound on ϵ_{opt}.

We now make use of the performance difference lemma:

$$\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma}\sum_{s,a} d_\rho^{\pi^*}(s)\pi^*(a|s)A^{\pi_\theta}(s,a) &\text{(C.4)}\\
&= \frac{1}{1-\gamma}\sum_{s} d_\rho^{\pi^*}(s)A^{\pi_\theta}(s,a^*(s)) &\text{(C.5)}\\
&\leq \frac{1}{1-\gamma}\sum_{s} d_\rho^{\pi^*}(s)A^{\pi_\theta}(s,a^*(s))\mathbb{1}(A^{\pi_\theta}(s,a^*(s)) \geq 0) &\text{(C.6)}\\
&\leq \frac{\lambda}{(1-\gamma)|\mathcal{S}|}\sum_{s} \kappa_\mathcal{A}^\alpha(s)\frac{d_\rho^{\pi^*}(s)}{\mu(s)}\mathbb{1}(A^{\pi_\theta}(s,a^*(s)) \geq 0) &\text{(C.7)}\\
&\leq \frac{\lambda}{(1-\gamma)}\mathbb{E}_{s\sim\mathrm{Unif}_\mathcal{S}}\left[\kappa_\mathcal{A}^\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty &\text{(C.8)}
\end{aligned}$$

This completes the proof.

□

We can bound the smoothness of the objective as follows.

Lemma 4.4 (Smoothness of $\mathcal{J}_\lambda^\alpha$)**.** *For the softmax parameterization, we have that*

$$||\nabla_\theta \mathcal{J}_\lambda^\alpha(\theta) - \nabla_\theta \mathcal{J}_\lambda^\alpha(\theta')||_2 \leq \beta_\lambda ||\theta - \theta'||_2$$

where $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$.

Proof. We can first bound the smoothness of $V^{\pi_\theta}(\mu)$ using Lemma D.4 from Agarwal et al. (2020). We get

$$||\nabla_\theta V^{\pi_\theta}(\mu) - \nabla_\theta V^{\pi_{\theta'}}(\mu)||_2 \leq \beta ||\theta - \theta'||_2$$

for

$$\beta = \frac{8}{(1-\gamma)^3}.$$

We now need to bound the smoothness of the regularizer $\frac{\lambda}{|\mathcal{S}|}\Omega(\theta)$ where

$$\Omega(\theta) = \sum_{s,a} \pi_0(a|s) \log \pi_\theta(a|s).$$

Using that $\frac{\partial}{\partial \theta_{s',a'}} \log \pi_\theta(a|s) = \mathbb{1}(s = s')[\mathbb{1}(a = a') - \pi_\theta(a'|s)]$ for the softmax parameterization, we get

$$\begin{aligned}
\nabla_{\theta_s} \Omega(\theta) &= \pi_0(\cdot|s) - \pi_\theta(\cdot|s), \\
\nabla^2_{\theta_s} \Omega(\theta) &= -\text{diag}(\pi_\theta(\cdot|s)) + \pi_\theta(\cdot)\pi_\theta(\cdot|s)^\mathsf{T}.
\end{aligned}$$

The remainder of the proof follows directly from that of Lemma D.4 in Agarwal et al. (2020), as the second-order gradients are identical. We then have that $\Omega(\theta)$ is 2-smooth and therefore $\frac{\lambda}{|\mathcal{S}|}\Omega(\theta)$ is $\frac{2\lambda}{|\mathcal{S}|}$-smooth, completing the proof. □

Note that the second value of λ will nearly always be greater than 1 for most values of $\epsilon, \epsilon_{\text{opt}}, |\mathcal{S}|, |\mathcal{A}|$, as that's the case when $\mathbb{E}_\mu[\alpha(s)] > \frac{(1-\gamma)^2\epsilon}{|\mathcal{A}|-1} - \epsilon_{\text{opt}}|\mathcal{S}|$, which is usually negative, thus trivially satisfying the inequality for $\alpha(s) \in [0, 1]\ \forall s \in \mathcal{S}$.

Lemma 4.5 (Iteration complexity for $\mathcal{J}_\lambda^\alpha$)**.** *Let ρ be a starting state distribution. Following Lemma 4.4, let $\beta_\lambda = \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. From any initial $\theta^{(0)}$ and following Eq.* (2.4) *with $\eta = 1/\beta_\lambda$*

$$\lambda = \frac{\epsilon(1-\gamma)}{\mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\left[\kappa_\mathcal{A}^\alpha(s)\right] \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty},$$

we have

$$\min_{t\leq T}\{V^\star(\rho) - V^{(t)}(\rho)\} \leq \epsilon \quad \text{whenever} \quad T \geq \frac{80\mathbb{E}_{s\sim \mathrm{U}_\mathcal{S}}\left[\kappa_\mathcal{A}^\alpha(s)\right]^2 |\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2.$$

Proof. The proof rests on bounding the iteration complexity of making the gradient sufficiently small. Because the optimization process is deterministic and unconstrained, we can use the standard result that after T updates with stepsize $1/\beta_\lambda$, we have

$$\min_{t\leq T} ||\nabla_\theta \mathcal{J}_\lambda^\star(\theta^{(t)})||_2^2 \leq \frac{2\beta_\lambda(\mathcal{J}_\lambda^\star(\theta^*) - \mathcal{J}_\lambda^\star(\theta^{(0)}))}{T} = \frac{2\beta_\lambda}{(1-\gamma)T}, \tag{C.9}$$

where β_λ upper-bounds the smoothness of $\mathcal{J}_\lambda^\star(\theta)$. Using the above and Corollary 4.1, we want

$$\epsilon_{\text{opt}} \leq \sqrt{\frac{2\beta_\lambda}{(1-\gamma)T}} \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}.$$

Solving the above inequality for T gives $T \geq \frac{8|\mathcal{S}|^2|\mathcal{A}|^2\beta_\lambda}{\lambda^2(1-\gamma)}$. From Lemma 4.4, we can set $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. Plugging this in gives

$$T \geq \frac{8|\mathcal{S}|^2|\mathcal{A}|^2\beta_\lambda}{(1-\gamma)\lambda^2} = \left(\frac{64|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda^2} + \frac{16|\mathcal{S}||\mathcal{A}|^2}{(1-\gamma)\lambda}\right).$$

Corollary 4.1 gives us the possible values for λ for value error margin ϵ. Then if

$$\lambda = \frac{\epsilon(1-\gamma)}{\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty} < 1,$$

we can write

$$\begin{aligned}\frac{64|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda^2} + \frac{16|\mathcal{S}||\mathcal{A}|^2}{(1-\gamma)\lambda} &\leq \frac{80|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda^2} \\ &= \frac{80\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]^2|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^2(1-\gamma)^6}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2.\end{aligned}$$

□

Corollary 4.2. *Given the same assumptions as Lemma 4.5, if the initial policy is chosen to be π_0, i.e., $\pi_{\theta^{(0)}} = \pi_0$ where $\pi_0(\cdot|s)$ is $\alpha(s)$-optimal with respect to $\pi^\star(\cdot|s)$ $\forall s$, then*

$$\min_{t\leq T}\{V^\star(\rho) - V^{(t)}(\rho)\} \leq \epsilon \quad \text{whenever} \quad T \geq \frac{320|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2\left\|\frac{1}{\mu}\right\|_\infty\mathbb{E}_{s\sim\mu}\left[\alpha(s)\right].$$

Proof. The proof rests on bounding the iteration complexity of making the gradient sufficiently small. Because the optimization process is deterministic and unconstrained, we can use the standard result that after T updates with stepsize $1/\beta_\lambda$, we have

$$\min_{t\leq T}||\nabla_\theta\mathcal{J}_\lambda^\star(\theta^{(t)})||_2^2 \leq \frac{2\beta_\lambda(\mathcal{J}_\lambda^\star(\theta^*) - \mathcal{J}_\lambda^\star(\theta^{(0)}))}{T} = \frac{2\beta_\lambda}{T}\Delta, \tag{C.10}$$

where β_λ upper-bounds the smoothness of $\mathcal{J}_\lambda^\star(\theta)$ and we. define $\Delta := \mathcal{J}_\lambda^\star(\theta^*) - \mathcal{J}_\lambda^\star(\theta^{(0)}))$ for conciseness. Using the above and Corollary 4.1, we want

$$\epsilon_{\text{opt}} \leq \sqrt{\frac{2\beta_\lambda\Delta}{T}} \leq \frac{\lambda}{2|\mathcal{S}||\mathcal{A}|}.$$

Solving the above inequality for T gives $T \geq \Delta\frac{8|\mathcal{S}|^2|\mathcal{A}|^2\beta_\lambda}{\lambda^2}$. From Lemma 4.4, we can set $\beta_\lambda = \frac{8}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. Plugging this in gives

$$T \geq \Delta\frac{8|\mathcal{S}|^2|\mathcal{A}|^2\beta_\lambda}{\lambda^2} = \Delta\left(\frac{64|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^3\lambda^2} + \frac{16|\mathcal{S}||\mathcal{A}|^2}{\lambda}\right).$$

Corollary 4.1 ensures that $\min_{t\leq T} V^\star(\rho) - V^{(t)}(\rho) \leq \epsilon$ provided that λ is of the form:

$$\lambda = \frac{\epsilon(1-\gamma)}{\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty} < 1,$$

we can therefore write:

$$\frac{64|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^3\lambda^2}+\frac{16|\mathcal{S}||\mathcal{A}|^2}{\lambda}\leq\frac{80|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^3\lambda^2}$$
$$=\frac{80\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]^2|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^2(1-\gamma)^5}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2.$$

This implies the following condition on T:

$$T\geq\frac{80\Delta|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^5}\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]^2\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2 \tag{C.11}$$

It remains to control the error Δ due to initialization with policy π_0. Denote by $\pi_\lambda^\star$ the optimal policy maximizing $\mathcal{J}_\lambda^\star$. We have the following:

$$\begin{aligned}\Delta :=&V^{\pi_\lambda^\star}(\rho)-V^{\pi_0}(\rho)-\lambda\mathrm{KL}(\pi_0,\pi_\lambda^\star) &\text{(C.12)}\\ \leq&V^{\pi_\lambda^\star}(\rho)-V^\star(\rho)+V^\star(\rho)-V^{\pi_0}(\rho) &\text{(C.13)}\\ \leq&V^\star(\rho)-V^{\pi_0}(\rho) &\text{(C.14)}\\ \leq&\frac{1}{(1-\gamma)^2}\left\|\frac{d_\rho^{\pi_0}}{\mu'}\right\|_\infty\mathbb{E}_{s\sim\mu'}\left[\alpha(s)\right] &\text{(C.15)}\end{aligned}$$

where the first line is by definition of Δ, the second line uses that the KL term is non-positive. The third line uses that $V^{\pi_\lambda^\star}(\rho)-V^\star(\rho)\leq 0$ and the last line follows from Lemma C.2. Hence, it suffice to choose T satisfying:

$$T\geq\frac{80|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2\left\|\frac{d_\rho^{\pi_0}}{\mu'}\right\|_\infty\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]^2\mathbb{E}_{s\sim\mu'}\left[\alpha(s)\right] \tag{C.16}$$

As a final step, we simply observe that $\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]\leq 2$.

□

Lemma C.1. *Following Lemma 4.4, let* $\beta_\lambda=\frac{8\gamma}{(1-\gamma)^3}+\frac{2\lambda}{|\mathcal{S}|}$. *From any initial* $\theta^{(0)}$ *and following Eq.* (2.4) *with* $\eta=1/\beta_\lambda$ *and*

$$\lambda=\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|(|\mathcal{A}|-1)}{(1-\gamma)^2\epsilon-(|\mathcal{A}|-1)\mathbb{E}_\mu\left[\alpha(s)\right]}, \tag{C.17}$$

for all starting state distributions ρ, *we have,*

$$\min_{t<T}\{V^*(\rho)-V^{(t)}(\rho)\}\leq\epsilon$$
$$\textit{whenever}\quad T\geq\min\left\{\frac{80\mathbb{E}_\mu\left[\kappa_{\mathcal{A}}^\alpha(s)\right]^2|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\epsilon^2}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2,\ 80|\mathcal{S}||\mathcal{A}|^2\left(\frac{\epsilon}{\epsilon_{\mathrm{opt}}(1-\gamma)^2(|\mathcal{A}|-1)}-\frac{\mathbb{E}_\mu\left[\alpha(s)\right]}{\epsilon_{\mathrm{opt}}(1-\gamma)^4}\right)\right\}. \tag{C.18}$$

Proof. The proof proceeds identically as above, except we set

$$\lambda=\frac{\epsilon_{\mathrm{opt}}|\mathcal{S}|(|\mathcal{A}|-1)}{(1-\gamma)^2\epsilon-(|\mathcal{A}|-1)\mathbb{E}_\mu\left[\alpha(s)\right]}>1,$$

we have

$$\begin{aligned}\frac{64|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda^2}+\frac{16|\mathcal{S}||\mathcal{A}|^2}{(1-\gamma)\lambda}\leq&\frac{80|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\lambda}\\ =&\frac{80|\mathcal{S}||\mathcal{A}|^2\left((1-\gamma)^2\epsilon-(|\mathcal{A}|-1)\mathbb{E}_\mu\left[\alpha(s)\right]\right)}{(1-\gamma)^4\epsilon_{\mathrm{opt}}(|\mathcal{A}-1)}\\ =&80|\mathcal{S}||\mathcal{A}|^2\left(\frac{\epsilon}{\epsilon_{\mathrm{opt}}(1-\gamma)^2(|\mathcal{A}|-1)}-\frac{\mathbb{E}_\mu\left[\alpha(s)\right]}{\epsilon_{\mathrm{opt}}(1-\gamma)^4}\right)\end{aligned}$$

completing the proof. □

Note that this value of λ will nearly always be greater than 1 for most values of $\epsilon, \epsilon_{\text{opt}}, |\mathcal{S}|, |\mathcal{A}|$, as that's the case when $\mathbb{E}_\mu[\alpha(s)] > \frac{(1-\gamma)^2\epsilon}{|\mathcal{A}|-1} - \epsilon_{\text{opt}}|\mathcal{S}|$, which is usually negative, thus trivially satisfying the inequality for $\alpha(s) \in [0,1]\ \forall s \in \mathcal{S}$.

Lemma C.2. *Assume that π is such that $\pi(a^\star(s)|s) \geq 1 - \beta(s)$ for some state dependent error $s \mapsto \beta(s)$ and that $\rho(s) > 0$ for all states s. Then the following inequality holds:*

$$V^\pi(\rho) - V^*(\rho) \geq -\frac{1}{(1-\gamma)^2} \left\| \frac{d_\rho^\pi}{\rho} \right\|_\infty \mathbb{E}_\rho[\beta(s)] \tag{C.19}$$

Proof.

$$\begin{aligned}
V^\pi(\rho) - V^*(\rho) &= \frac{1}{1-\gamma} \sum_{s,a} d_\rho^\pi(s)\pi(a|s)A^{\pi^*}(s,a) \\
&= \frac{1}{1-\gamma} \sum_s \sum_{a \neq a^\star(s)} \left(d_\rho^\pi(s)\pi(a|s)A^{\pi^*}(s,a) \right) \\
&\geq -\frac{1}{(1-\gamma)^2} \sum_s d_\rho^\pi(s) \sum_{a \neq a^\star(s)} \pi(a|s) \\
&\geq -\frac{1}{(1-\gamma)^2} \sum_s \left(d_\rho^\pi(s)\beta(s) \right) \\
&\geq -\frac{1}{(1-\gamma)^2} \left\| \frac{d_\rho^\pi}{\mu} \right\|_\infty \mathbb{E}_{s\sim\mu}[\beta(s)]
\end{aligned}$$

where the first line follows by application of the performance different lemma (Agarwal et al., 2020, Lemma 3.2), the second line is due to the fact that $A^{\pi^*}(s, a^*(s)) = 0$, the third line from $A^{\pi^*}(s,a) \geq -1/(1-\gamma)$ for $a \neq a^*$. The fourth line uses that $\sum_{a \neq a^\star(s)} \pi(a|s) = 1 - \pi(a^\star(s)|s) \leq \beta(s)$ for $a \neq a^\star(s)$ since by assumption $\pi(a^\star(s)|s) \geq 1 - \beta(s)$. Finally, the last line uses that d_ρ^π is a probability distribution over states s satisfying $\sum_{s\in\mathcal{S}} d_\rho^\pi(s) = 1$.

□

C.1 State dependent λ and ϵ

We can further generalize these results by allowing the error ϵ and regularization weight λ to be state-dependent. The gradient with state dependent regularized λ equals

$$\mathcal{J}^{\pi_0}(\theta) = V^{\pi_\theta}(\mu) + \sum_{s,a} \frac{\lambda(s)}{|\mathcal{S}|} \pi_0(a|s) \log \pi_\theta(a|s)$$

Lemma C.3. *Suppose θ is such that $\left(\nabla \mathcal{J}_\lambda^\alpha(\theta)\right)_{s,a} \leq \epsilon_{\text{opt}}(s,a)$. Then we have that for all states $s \in \mathcal{S}$,*

$$\begin{aligned}
V^{\pi_\theta}(\rho) \geq V^*(\rho) - \min\Bigg\{ &\frac{1}{1-\gamma} \mathbb{E}_{s\sim\text{Unif}_\mathcal{S}} \left[\frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\max\left((1-\alpha(s)) - \frac{\epsilon_{\text{opt}}(s,a^*(s))|\mathcal{S}|}{\lambda(s)}, 0 \right)} + \lambda(s)\alpha(s) \right] \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty, \\
&\frac{|\mathcal{A}|}{(1-\gamma)^2} \mathbb{E}_{s\sim\mu}[\alpha(s)] \left\| \frac{d_\rho^{\pi_\theta}}{\mu} \right\|_\infty + \frac{|\mathcal{S}|}{(1-\gamma)^2} \left\| \frac{\sum_a \epsilon_{\text{opt}}(s,a)}{\lambda(s)} \right\|_\infty \Bigg\}
\end{aligned}$$

Proof. Let's assume that π^* is a deterministic policy. By Puterman (2010) such an optimal policy always exists for an MDP. We'll use the notation $a^*(s)$ to denote the optimal action at state s. This, combined with the assumption that $d_{\text{TV}}(\pi^*(\cdot|s), \pi_0(\cdot|s)) \leq \alpha(s)$ for all $s \in \mathcal{S}$, tells us that

$\pi_0(a^*(s)|s) \geq \pi^*(a^*(s)|s) - \alpha(s) = 1 - \alpha(s)$. Similarly, for $a \neq a^*(s)$, $\pi_0(a|s) \leq \alpha(s)$. Using this, we can start by showing that whenever $A^{\pi_\theta}(s, a^*(s)) \geq 0$ we can lower bound $\pi_\theta(a^*(s)|s)$ for all s.

The gradient norm assumption $\left(\nabla \mathcal{J}_\lambda^\alpha(\theta)\right)_{s,a} \leq \epsilon_{\text{opt}}(s, a)$ implies that for all s, a:

$$\epsilon_{\text{opt}}(s, a) \geq \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s, a) + \frac{\lambda(s)}{|\mathcal{S}|}(\pi_0(a|s) - \pi_\theta(a|s))$$

In particular for all s,

$$\begin{aligned}
\epsilon_{\text{opt}}(s, a^*(s)) \geq \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a^*(s)}} &\overset{(i)}{\geq} \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a^*(s)|s) A^{\pi_\theta}(s, a^*(s)) + \frac{\lambda(s)}{|\mathcal{S}|}(\pi^*(a^*(s)|s) - \alpha(s) - \pi_\theta(a^*(s)|s)) \\
&= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta(a^*(s)|s) A^{\pi_\theta}(s, a^*(s)) + \frac{\lambda(s)}{|\mathcal{S}|}(1 - \alpha(s) - \pi_\theta(a^*(s)|s))
\end{aligned} \tag{C.20}$$

And therefore if $A^{\pi_\theta}(s, a^*(s)) \geq 0$,

$$\epsilon_{\text{opt}}(s, a) \geq \frac{\lambda(s)}{|\mathcal{S}|}(1 - \alpha(s) - \pi_\theta(a^*(s)|s))$$

Thus,

$$\pi_\theta(a^*(s)|s) \geq \max\left(1 - \alpha(s) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda(s)}, 0\right) \geq 1 - \alpha(s) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda(s)}. \tag{C.21}$$

In this case we can upper bound $A^{\pi_\theta}(s, a^*(s))$. From Eq. (C.1) inequality (i), we have

$$\begin{aligned}
A^{\pi_\theta}(s, a^*(s)) &\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a^*(s)|s)} \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a^*(s)}} - \frac{\lambda(s)}{|\mathcal{S}|} \frac{1}{\pi_\theta(a^*(s)|s)} (1 - \alpha(s) - \pi_\theta(a^*(s)|s)) \right) \\
&= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \left(\frac{1}{\pi_\theta(a^*(s)|s)} \frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a^*(s)}} + \frac{\lambda(s)}{|\mathcal{S}|} \left(1 - \frac{1 - \alpha(s)}{\pi_\theta(a^*(s)|s)}\right) \right) \\
&\overset{(i)}{\leq} \frac{1}{\mu(s)} \left(\frac{1}{\max\left((1 - \alpha(s)) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda}, 0\right)} \cdot \epsilon_{\text{opt}}(s, a^*(s)) + \frac{\lambda(s)}{|\mathcal{S}|} (1 - (1 - \alpha(s))) \right) \\
&\leq \frac{1}{\mu(s)} \left(\frac{1}{\max\left((1 - \alpha(s)) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda}, 0\right)} \cdot \epsilon_{\text{opt}}(s, a^*(s)) + \frac{\lambda(s)}{|\mathcal{S}|} \alpha(s) \right)
\end{aligned}$$

Where (i) follows because $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\mu(s)$, $\frac{\partial \mathcal{J}_\lambda^\alpha(\theta)}{\partial \theta_{s,a}} \leq \epsilon_{\text{opt}}$ and $\max\left(1 - \alpha(s) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda}, 0\right) \leq \pi_\theta(a^*(s)|s) \leq 1$.

We now make use of the performance difference lemma:

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^*}(s) \pi^*(a|s) A^{\pi_\theta}(s, a) \tag{C.22}$$

$$= \frac{1}{1-\gamma} \sum_{s} d_\rho^{\pi^*}(s) A^{\pi_\theta}(s, a^*(s)) \tag{C.23}$$

$$\leq \frac{1}{1-\gamma} \sum_{s} d_\rho^{\pi^*}(s) A^{\pi_\theta}(s, a^*(s)) \mathbb{1}(A^{\pi_\theta}(s, a^*(s)) \geq 0) \tag{C.24}$$

$$\leq \frac{1}{1-\gamma} \sum_{s} \frac{d_\rho^{\pi^*}(s)}{\mu(s)} \left(\frac{1}{\max\left((1 - \alpha(s)) - \frac{\epsilon_{\text{opt}}(s, a)|\mathcal{S}|}{\lambda}, 0\right)} \cdot \epsilon_{\text{opt}}(s, a^*(s)) + \frac{\lambda(s)}{|\mathcal{S}|} \alpha(s) \right) \mathbb{1}(A^{\pi_\theta}(s, a^*(s) \tag{C.25}$$

$$\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim \text{Unif}_\mathcal{S}} \left[\frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\max\left((1 - \alpha(s)) - \frac{\epsilon_{\text{opt}}(s, a^*(s))|\mathcal{S}|}{\lambda(s)}, 0\right)} + \lambda(s)\alpha(s) \right] \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty \tag{C.26}$$

Now let's relate the values of π^* and π_θ. We will again apply the performance difference lemma, this time in the other direction:

$$\begin{aligned}
V^{\pi_\theta}(\rho) - V^*(\rho) &= \frac{1}{1-\gamma}\sum_{s,a} d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi^*}(s,a) \\
&\overset{(1)}{=} \frac{1}{1-\gamma}\sum_s \left(\sum_{a\neq a^*(s)} d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi^*}(s,a)\right) \\
&\overset{(2)}{\geq} \frac{-1}{1-\gamma}\sum_s\sum_a d_\rho^{\pi_\theta}(s)\left(\alpha(s)(|\mathcal{A}|-1) + \frac{\sum_{a\neq a^*(s)}\epsilon_{\text{opt}}(s,a)|\mathcal{S}|}{\lambda(s)}\right)\frac{1}{1-\gamma} \\
&= -\frac{1}{1-\gamma}\left(\sum_s \frac{d_\rho^{\pi_\theta}(s)}{1-\gamma}\alpha(s)(|\mathcal{A}|-1) + \sum_s |\mathcal{S}| d_\rho^{\pi_\theta}(s)\sum_{a\neq a^*(s)}\frac{\epsilon_{\text{opt}}(s,a)}{(1-\gamma)\lambda(s)}\right) \\
&\overset{(3)}{\geq} -\sum_s \frac{d_\rho^{\pi_\theta}(s)}{(1-\gamma)^2}\alpha(s)(|\mathcal{A}|-1) - \frac{|\mathcal{S}|}{(1-\gamma)^2}\left\|\frac{\sum_{a\neq a^*(s)}\epsilon_{\text{opt}}(s,a)}{\lambda(s)}\right\|_\infty \\
&\geq -\frac{|\mathcal{A}|}{(1-\gamma)^2}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}[\alpha(s)] - \frac{|\mathcal{S}|}{(1-\gamma)^2}\left\|\frac{\sum_a \epsilon_{\text{opt}}(s,a)}{\lambda(s)}\right\|_\infty \\
&\geq -\frac{|\mathcal{A}|}{(1-\gamma)^2}\mathbb{E}_{s\sim\mu}[\alpha(s)]\left\|\frac{d_\rho^{\pi_\theta}}{\mu}\right\|_\infty - \frac{|\mathcal{S}|}{(1-\gamma)^2}\left\|\frac{\sum_a \epsilon_{\text{opt}}(s,a)}{\lambda(s)}\right\|_\infty
\end{aligned}$$

where (1) is due to the fact that $A^{\pi^*}(s,a^*(s)) = 0$, and (2) is due to the fact that $A^{\pi^*}(s,a)$ for $a \neq a^*$ is lower-bounded by $-1/(1-\gamma)$ and Eq. (C.21). and (3) holds because of Holder and $\sum_s d_\rho^{\pi_\theta}(s) = 1$. Therefore,

$$V^{\pi_\theta}(\rho) + \frac{|\mathcal{A}|}{(1-\gamma)^2}\mathbb{E}_{s\sim\mu}[\alpha(s)]\left\|\frac{d_\rho^{\pi_\theta}}{\mu}\right\|_\infty + \frac{|\mathcal{S}|}{(1-\gamma)^2}\left\|\frac{\sum_a \epsilon_{\text{opt}}(s,a)}{\lambda(s)}\right\|_\infty \geq V^*(\rho).$$

□

A simple corollary of Lemma C.3 is,

Corollary C.1. *If* $\epsilon_{\text{opt}}(s,a) \leq \frac{(1-\alpha(s))\lambda(s)}{|\mathcal{S}|}$ *then*

$$\begin{aligned}
V^{\pi_\theta}(\rho) \geq V^*(\rho) - \min\Bigg\{ &\frac{1}{1-\gamma}\mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\left[\frac{2\epsilon_{\text{opt}}(s,a^*(s))|\mathcal{S}|}{1-\alpha(s)} + \lambda(s)\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty, \\
&\frac{|\mathcal{A}|}{(1-\gamma)^2}\mathbb{E}_{s\sim\mu}[\alpha(s)]\left\|\frac{d_\rho^{\pi_\theta}}{\mu}\right\|_\infty + \frac{|\mathcal{S}|}{(1-\gamma)^2}\left\|\frac{\sum_a \epsilon_{\text{opt}}(s,a)}{\lambda(s)}\right\|_\infty\Bigg\}
\end{aligned}$$

Proof. If $\epsilon(s,a) \leq \frac{(1-\alpha(s))\lambda(s)}{|\mathcal{S}|}$ then $\max\left((1-\alpha(s)) - \frac{\epsilon_{\text{opt}}(s,a)|\mathcal{S}|}{\lambda(s)}, 0\right) \geq \frac{1-\alpha(s)}{2}$. The result follows. □

Corollary C.1 recovers the results of Agarwal et al. (2020) by noting the TV distance between the optimal policy and the uniform one equals $1 - \frac{1}{|\mathcal{A}|}$ and therefore $1 - \alpha(s) = \frac{1}{|\mathcal{A}|}$.

We now concern ourselves with the problem of finding a true $\epsilon > 0$ optimal policy. This will require us to set the values of $\lambda(s)$ appropriately. We restrict ourselves to the following version of the results of Corollary C.1. If $\epsilon(s,a) \leq \frac{(1-\alpha(s))\lambda(s)}{|\mathcal{S}|}$ then

$$V^{\pi_\theta}(\rho) \geq V^*(\rho) - \frac{1}{1-\gamma}\mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\left[\frac{2\epsilon_{\text{opt}}(s,a^*(s))|\mathcal{S}|}{1-\alpha(s)} + \lambda(s)\alpha(s)\right]\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty$$

By setting $\lambda(s) = \frac{\epsilon(1-\gamma)}{2\alpha(s)\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty}$ and $\epsilon_{\text{opt}}(s,a) = \min\left(\frac{(1-\alpha(s))\epsilon(1-\gamma)}{4|\mathcal{S}|\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty}, \frac{(1-\alpha(s))\lambda(s)}{|\mathcal{S}|}\right)$ we get

$$V^{\pi_\theta}(\rho) \geq V^*(\rho) - \epsilon.$$

Observe that the level of regularization depends on the state's error. If the error is very low, the regularizer $\lambda(s)$ should be set to a larger value.

D Multitask learning

Assume we are given K i.i.d. tasks M_k sampled from $\mathcal{P}_\mathcal{M}$, denote by $\pi_k^\star(\cdot|s)$ their corresponding optimal policies and let $\tilde{\pi}_k(\cdot|s)$ be $\alpha(s)$ policies, i.e. $d_{TV}(\tilde{\pi}_k(\cdot|s), \pi_k^\star(\cdot|s)) \leq \alpha(s)$ for some $\alpha(s) \leq 1$. To simplify notation, we may also refer to $\mathcal{P}$ directly as the distribution over these optimal policies. Let $\hat{\pi}_0$ be the total variation barycenter of the policies $\tilde{\pi}_k$, i.e.: $\hat{\pi}_0 = \arg\min_\pi \frac{1}{K}\sum_{k=1}^K d_{TV}(\pi, \tilde{\pi}_i)$, while $\pi_0 = \arg\min_\pi \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[d_{TV}(\pi, \pi_i^\star)]$.

Lemma 5.1 (TV barycenter)**.** *Let $\mathcal{P}_\mathcal{M}$ be a distribution over tasks $\mathcal{M} = \{M_k\}$, each with a deterministic policy $\pi_k^\star : \mathcal{S} \to \mathcal{A}$. Define the average optimal action as*

$$\xi(s,a) \coloneqq \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\left[\mathbb{1}(\pi_k^\star(s) = a)\right]. \tag{5.2}$$

Then, the TV barycenter $\pi_0(\cdot|s)$ defined in Eq. (5.1) is given by a greedy policy over ξ, i.e., $\pi_0(a|s) = \delta(a \in \text{argmax}_{a'\in\mathcal{A}}\,\xi(s,a'))$, where $\delta(\cdot)$ is the Dirac delta distribution.

Proof. Let's first express the barycenter loss in a more convenient form:

$$\begin{aligned}
\mathbb{E}_{\pi'\sim P}\left[d_{TV}(\pi(.|s), \pi'(.|s))\right] =& \mathbb{E}_{\pi'\sim P}\left[\frac{1}{2}\sum_{a\neq a_{\pi'}(s)}\pi(a) + \frac{1}{2}(1-\pi(a_{\pi'}(s), s))\right] && \text{(D.1)}\\
=& \mathbb{E}_{\pi'\sim P}\left[(1-\pi(a_{\pi'}(s), s))\right] && \text{(D.2)}\\
=& 1 - \sum_a \mathbb{P}(\pi'(a|s) = 1)\pi(a|s) && \text{(D.3)}\\
=& 1 - \sum_a \pi_{soft}(a|s)\pi(a|s). && \text{(D.4)}
\end{aligned}$$

Therefore, the barycenter loss is minimized when $\pi(a|s)$ puts all its mass on the maximum value of $\pi_{soft}(a|s)$ over actions $a \in \mathcal{A}$.

□

The KL barycenter can be described as follows.

Lemma D.1 (KL barycenter)**.** *Let $\mathcal{P}_\mathcal{M}$ be a distribution over tasks such that for every $M_k \in \mathcal{M}$, there exists a unique optimal action $a_k^\star(s)$ for each state s such that $\pi_k^\star(s) = a^\star$. Then the KL barycenter for state s is:*

$$\operatorname*{argmin}_\pi \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\text{KL}(\pi_k^\star(\cdot|s), \pi(\cdot|s)) = \delta(a = \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\pi_k^\star(s)) \tag{D.5}$$

where $\delta(\cdot)$ is the Dirac delta distribution. This holds for both directions of the KL.

Proof. We have

$$\begin{aligned}\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\mathrm{KL}(\pi_k^\star(\cdot|s),\pi(\cdot|s)) &= \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\sum_a \pi_k^\star(a|s)\log\frac{\pi_k^\star(a|s)}{\pi(a|s)}\\ &= \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}\left[-\log\pi(a_k^\star(s)|s) + \underbrace{\sum_{a\neq a_k^\star(s)} 0\cdot\log\frac{0}{\pi(a|s)}}_{=0}\right]\\ &= \mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[-\log\pi(a_k^\star(s)|s)]\end{aligned}$$

Therefore, the barycenter loss is minimized when $\pi(a|s)$ puts all its mass on the expected $a_k^\star(s)$. Note that we consider the underbrace term zero because $\lim_{x\to 0} x\log x = 0$. It is easy to verify that this result holds for the reverse KL. □

Lemma 5.2 (Multitask iteration complexity)**.** *Let $M_k\sim\mathcal{P}_\mathcal{M}$ and denote by $\pi_k^\star:\mathcal{S}\to\mathcal{A}$ its optimal policy. Denote by T_k the number of iterations to reach ϵ-error for M_k in the sense that:*

$$\min_{t\le T_k}\{V^*(\rho) - V^{(t)}(\rho)\}\le\epsilon.$$

Set λ,β_λ, and η as in Lemma 4.5. From any initial $\theta^{(0)}$, and following Eq. (2.4), $\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[T_k]$ satisfies:

$$\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[T_k]\ge\frac{80|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^6}\mathbb{E}_{\substack{M_k\sim\mathcal{P}_\mathcal{M}\\ s\sim\mathrm{U}_\mathcal{S}}}\left[\kappa_\mathcal{A}^{\alpha_k}(s)\left\|\frac{d_\rho^{\pi_k^*}}{\mu}\right\|_\infty^2\right],$$

where $\alpha_k(s)\coloneqq d_{\mathrm{TV}}(\pi_k^\star(\cdot|s),\hat\pi_0(\cdot|s))$. If $\hat\pi_0$ is also used for initialization, then $\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[T_k]$ satisfies:

$$\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[T_k]\ge\frac{320|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7}\left\|\frac{1}{\mu}\right\|_\infty^3\mathbb{E}_{\substack{M_k\sim\mathcal{P}_\mathcal{M}\\ s\sim\mu}}[\alpha_k(s)],$$

Proof. Let M_i be a random task sampled according to $\mathcal{M}$ and denote by $\pi_i^\star$ its corresponding optimal policy. Set $\alpha(s) = d_{TV}(\hat\pi_0,\pi_i^\star)$ and choose $\lambda = \frac{\epsilon(1-\gamma)}{2\|\frac{d_\rho^{\pi^\star}}{\mu}\|}$. By **??**, we have that:

$$\begin{aligned}&\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\}\le\epsilon\\ &\textit{whenever}\quad T\ge\frac{160|\mathcal{A}|^2|\mathcal{S}|^2}{\epsilon^2(1-\gamma)^7}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2\left\|\frac{d_\rho^{\hat\pi_0}}{\mu'}\right\|_\infty\mathbb{E}_{s\sim\mu'}[\alpha(s)].\end{aligned}\tag{D.6}$$

By choosing μ' to be uniform and recalling that $d_\rho^{\hat\pi_0}\le 1$, it suffice to have:

$$T\ge\frac{160|\mathcal{A}|^2|\mathcal{S}|^3}{\epsilon^2(1-\gamma)^7}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2\mathbb{E}_{s\sim\mu'}[d_{TV}(\hat\pi_0,\pi_i^\star)].\tag{D.7}$$

Taking the expectation over the tasks and treating T as a random variable depending on the task, we get that:

$$\mathbb{E}[T]\ge\frac{160|\mathcal{A}|^2|\mathcal{S}|^3}{\epsilon^2(1-\gamma)^7}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2\mathbb{E}_{s\sim\mu'\pi'\sim\mathbb{P}}\left[d_{TV}(\hat\pi_0,\pi')\right].\tag{D.8}$$

□

The following lemma quantifies how $\hat{\pi}_0$ is close to be the TV barycenter of $\{\pi_k^\star\}_{1\leq k\leq K}$ when K grows to infinity. We let $\tilde{\pi}_k(\cdot|s)$ be, on average, $\zeta(s)$-optimal in state s across tasks M_k, i.e. $\mathbb{E}_{M_k\sim\mathcal{M}}\left[d_{\mathrm{TV}}(\tilde{\pi}_k(\cdot|s),\pi_k^\star(\cdot|s))\right]\leq\zeta(s)$ for some $\zeta(s)\in[0,1]$. For concision, we shorten $\pi(\cdot|s)$ as π.

Lemma 5.3 (Barycenter concentration)**.** *Let δ be $0<\delta<1$. Then with probability higher than $1-\delta$, for all $s\in\mathcal{S}$, it holds that:*

$$\begin{aligned}&\left|\mathbb{E}_{M_k\sim\mathcal{P}_\mathcal{M}}[d_{\mathrm{TV}}(\pi_k^\star(\cdot|s),\hat{\pi}_0(\cdot|s))-d_{\mathrm{TV}}(\pi_k^\star(\cdot|s),\pi_0(\cdot|s))]\right|\\&\qquad\leq 2\zeta(s)+\sqrt{\frac{2\log(\frac{2}{\delta})}{K}}+2C\sqrt{\frac{|\mathcal{A}|}{K}},\end{aligned}$$

for some constant C that depends only on $|\mathcal{A}|$.

Proof. To simplify the proof, we fix a state s and omit the dependence in s. We further introduce the following notations:

$$f(\pi)=\mathbb{E}_{M_i\sim\mathcal{P}_\mathcal{M}}\left[d_{TV}(\pi,\pi_i^\star)\right] \tag{D.9}$$

$$\tilde{f}(\pi)=\frac{1}{K}\sum_{i=1}^K d_{TV}(\pi,\tilde{\pi}_i) \tag{D.10}$$

$$\hat{f}(\pi)=\frac{1}{K}\sum_{i=1}^K d_{TV}(\pi,\pi_i^\star) \tag{D.11}$$

Let $\pi_0=\arg\min_\pi f(\pi)$ and $\hat{\pi}_0=\arg\min_\pi\tilde{f}(\pi)$. It is easy to see that:

$$\begin{aligned}f(\hat{\pi}_0)\leq&\tilde{f}(\hat{\pi}_0)+|\hat{f}(\hat{\pi}_0)-f(\hat{\pi}_0)|+|\tilde{f}(\hat{\pi}_0)-\hat{f}(\hat{\pi}_0)|\\\leq&\hat{f}(\pi_0)+|\hat{f}(\hat{\pi}_0)-f(\hat{\pi}_0)|+|\hat{f}(\hat{\pi}_0)-\hat{f}(\hat{\pi}_0)|\\\leq&f(\pi_0)+|\hat{f}(\hat{\pi}_0)-f(\hat{\pi}_0)|+|\tilde{f}(\hat{\pi}_0)-\hat{f}(\hat{\pi}_0)|\\&+|\hat{f}(\pi_0)-f(\pi_0)|+|\tilde{f}(\pi_0)-\hat{f}(\pi_0)|\\\leq&f(\pi_0)+2\sup_\pi|\hat{f}(\pi)-f(\pi)|+2\sup_\pi|\hat{f}(\pi)-\tilde{f}(\pi)|.\end{aligned}$$

where the first line follows by a triangular inequality, the second line uses that $\hat{f}(\hat{\pi}_0)\leq\hat{f}(\pi_0)$ since $\hat{\pi}_0$ minimizes $\hat{f}$. The third line uses a triangular inequality again while the last line follows by definition of the supremum. Moreover, recall that $f(\pi_0)\leq f(\hat{\pi}_0)$ as π_0 minimizes f and that $|\hat{f}(\pi)-\tilde{f}(\pi)|\leq\zeta$ since, by assumption, we have that $d_{TV}(\pi_i^\star,\tilde{\pi}_i)\leq\zeta$. Therefore, it follows that:

$$|f(\hat{\pi}_0)-f(\pi_0)|\leq 2\zeta+2\sup_\pi|\hat{f}(\pi)-f(\pi)|. \tag{D.12}$$

By application of the bounded difference inequality (McDiarmid's inequality) (Sen, 2018, Theorem 13.8), we know that for any $t>0$:

$$\mathbb{P}\left[|\sup_\pi|\hat{f}(\pi)-f(\pi)|-\mathbb{E}\left[\sup_\pi|\hat{f}(\pi)-f(\pi)|\right]|>t\right]\leq 2e^{-2t^2K} \tag{D.13}$$

This implies that for any $0<\eta<1$, we have with probability higher than $1-\eta$ that:

$$\sup_\pi|\hat{f}(\pi)-f(\pi)|\leq\sqrt{\frac{\log(\frac{2}{\delta})}{2K}}+\mathbb{E}\left[\sup_\pi|\hat{f}(\pi)-f(\pi)|\right] \tag{D.14}$$

Combining Eq. (D.12) with Eq. (D.14) and using Lemma D.2 to control $\mathbb{E}\left[\sup_\pi|\hat{f}(\pi)-f(\pi)|\right]$, we have that for any $0<\delta<1$, with probability higher than $1-\delta$, it holds that:

$$|f(\hat{\pi}_0)-f(\pi_0)|\leq 2\zeta+\sqrt{\frac{2\log(\frac{2}{\delta})}{K}}+2C\sqrt{\frac{|\mathcal{A}|}{K}}, \tag{D.15}$$

for some constant C that depends only on $|\mathcal{A}|$.

□

Lemma D.2.

$$\mathbb{E}\left[\sup_{\pi} |\hat{f}(\pi) - f(\pi)|\right] \leq C\sqrt{\frac{|\mathcal{A}|}{N}}, \tag{D.16}$$

where C is a constant that depends only on $|\mathcal{A}|$.

Proof. To control the quantity $\mathbb{E}\left[\sup_{\pi} |\hat{f}(\pi) - f(\pi)|\right]$, we will use a classical result from empirical process theory (Van der Vaart, 2000, Corollary 19.35). We begin by introducing some useful notions to state the result. Denote by $\mathcal{F}$ the set of functions $\pi' \mapsto d_{TV}(\pi, \pi')$ that are indexed by a fixed π. Given a random task $M_i \sim \mathcal{M}$, we call $\pi_i^\star$ its optimal policy and denote by P the probability distribution of $\pi_i^\star$ when the task M_i is drawn from $\mathcal{M}$. Note that we can express $f(\pi)$ as an expectation w.r.t. P: $f(\pi) = \mathbb{E}_{\pi' \sim P}\left[d_{TV}(\pi, \pi')\right]$. Moreover, $\hat{f}(\pi)$ is an empirical average over i.i.d. samples $\pi_i^\star$ drawn from P.

The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_2(P))$ is the smallest number of functions f_j and g_j such that for any π, there exists j such that $f_j(\pi') \leq d_{TV}(\pi, \pi') \leq g_j(\pi')$ and $\|f_j - g_j\|_{L_2(P)} \leq \epsilon$. The following result is a direct application of (Van der Vaart, 2000, Corollary 19.35) and provides a control on $\mathbb{E}\left[\sup_{\pi} |\hat{f}(\pi) - f(\pi)|\right]$ in terms of the bracketing number $N_{[]}$:

$$\sqrt{N}\mathbb{E}\left[\sup_{\pi} |\hat{f}(\pi) - f(\pi)|\right] \leq \int_0^R \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}. \tag{D.17}$$

where $R^2 = \mathbb{E}_{\pi' \sim P}\left[\sup_{\pi} d_{TV}(\pi, \pi')^2\right] \leq 1$. It remains to control the bracketing number $N_{[]}$. To achieve this, note that the functions in $\mathcal{F}$ are all 1-Lipschitz, meaning that:

$$|d_{TV}(\pi, \pi) - d_{TV}(\pi', \pi)| \leq d_{TV}(\pi, \pi') \leq 1. \tag{D.18}$$

Moreover, the family $\mathcal{F}$ admits the constant function $F(\pi') = 1$ as an envelope, which means, in other words, that the following upper-bound holds:

$$\sup_{\pi} d_{TV}(\pi, \pi') \leq 1. \tag{D.19}$$

Therefore, we can apply (Van der Vaart, 2000, Example 19.7) to the family $\mathcal{F}$, which directly implies the following upper-bound on $N_{[]}$:

$$N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq K\left(\frac{1}{\epsilon}\right)^{|\mathcal{A}|} \tag{D.20}$$

where K is a constant that depends only on $|\mathcal{A}|$. Combining $D.18$ and D.20 and recalling that $R \leq 1$, it follows that:

$$\mathbb{E}\left[\sup_{\pi} |\hat{f}(\pi) - f(\pi)|\right] \leq C\sqrt{\frac{|\mathcal{A}|}{N}}. \tag{D.21}$$

where C is a constant that depends only on $|\mathcal{A}|$. □

E Experimental details

The policy model for all algorithms was given by the tabular softmax with single parameter vector $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}.$$

All agents were trained for 80,000 time steps per task using standard stochastic gradient ascent with learning rate $\eta = 0.02$. For methods with learned regularizers, the learning for the regularizer

was halved, with $\eta_{\text{reg}} = 0.01$. Each episode terminated when the agent reached a leaf node. For those using regularization, the regularization weight was $\lambda = 0.2$. For DISTRAL, this weight was applied equally to both the KL term and the entropy term. Each task was randomly sampled with $r(s) = 0$ for all nodes other than the leaf nodes of the subtree rooted at s_7 (Fig. 7.1). For those nodes, $r(s) \sim \text{Geom}(p)$ with $p = 0.5$ for experiments with fixed default policies and $p = 0.7$ for those with learned default policies. The sparsity of the reward distribution made learning challenging, and so limiting the size of the effective search space (via an effective default policy) was crucial to consistent success. A single run consisted of 5 draws from the task distribution, with each method trained for 20 runs with different random seeds. For TVPO, the softmax temperature decayed as $\beta(k) = \exp(-k/10)$, with k being the number of tasks. The plotted default policies in Fig. 7.4 were the average default policy probabilities in the selected states across these runs.