Regularization Guarantees Generalization in Bayesian Reinforcement Learning through Algorithmic Stability

Aviv Tamar, Daniel Soudry, Ev Zisselman

Technion – Israel Institute of Technology

Abstract

In the Bayesian reinforcement learning (RL) setting, a prior distribution over the unknown problem parameters – the rewards and transitions – is assumed, and a policy that optimizes the (posterior) expected return is sought. A common approximation, which has been recently popularized as *meta-RL*, is to train the agent on a *sample* of N problem instances from the prior, with the hope that for large enough N, good generalization behavior to an unseen test instance will be obtained. In this work, we study generalization in Bayesian RL under the probably approximately correct (PAC) framework, using the method of algorithmic stability. Our main contribution is showing that by adding regularization, the optimal policy becomes stable in an appropriate sense. Most stability results in the literature build on strong convexity of the regularized loss – an approach that is not suitable for RL as Markov decision processes (MDPs) are not convex. Instead, building on recent results of fast convergence rates for mirror descent in regularized MDPs, we show that regularized MDPs satisfy a certain *quadratic growth* criterion, which is sufficient to establish stability. This result, which may be of independent interest, allows us to study the effect of regularization on generalization in the Bayesian RL setting.

1 Introduction

How can an agent learn to quickly perform well in an unknown task? This is the basic question in reinforcement learning (RL). The most popular RL algorithms are designed in a *minimax* approach – seeking a procedure that will eventually learn to perform well in any task (Strehl et al. 2006; Jaksch, Ortner, and Auer 2010; Jin et al. 2018). Lacking prior information about the task, such methods must invest considerable efforts in *uninformed exploration*, typically requiring many samples to reach adequate performance. In contrast, when a *prior distribution* over possible tasks is known in advance, an agent can direct its exploration much more effectively. This is the Bayesian RL (BRL, Ghavamzadeh et al. 2016) setting. A *Bayes-optimal policy* – the optimal policy in BRL – can be orders of magnitude more sample efficient than a minimax approach, and indeed, recent studies demonstrated a quick solution of novel tasks, sometimes in just a handful of trials (Duan et al. 2016; Zintgraf et al. 2020; Dorfman, Shenfeld, and Tamar 2020).

For high-dimensional problems, and when the prior does not obey a very simple structure such as a Dirichlet prior (Ghavamzadeh et al. 2016), solving BRL is intractable, and one must resort to approximations. A common approximation, which has been popularized under the term meta-RL (Duan et al. 2016; Finn, Abbeel, and Levine 2017), is to replace the distribution over tasks with an empirical sample of tasks, and seek an optimal policy with respect to the sample, henceforth termed the empirical risk minimization policy (ERM policy). In this paper, we investigate the performance of the ERM policy on a novel task from the task distribution, that is – we ask *how well the policy generalizes*.

We focus on the Probably Approximately Correct (PAC) framework, which is popular in supervised learning, and adapt it to the BRL setting. Since the space of deterministic history-dependent policies is finite (in a finite horizon setting), a trivial generalization bound for a finite hypothesis space can be formulated. However, the size of the policy space, which such a naive bound depends on, leads us to seek alternative methods for *controlling* generalization. In particular, regularization is a well-established method in supervised learning that can be used to trade-off training error and test error. In RL, regularized MDPs are popular in practice (Schulman et al. 2017), and have also received interest lately due to their favorable optimization properties (Shani, Efroni, and Mannor 2020; Neu, Jonsson, and Gómez 2017).

The main contribution of this work is making the connection between regularized MDPs and PAC generalization, as described above. We build on the classical analysis of Bousquet and Elisseeff (2002), which bounds generalization through algorithmic stability. Establishing algorithmic stability results for regularized MDPs, however, is not trivial, as the loss function in MDPs is not convex in the policy. Our key insight is that while not convex, regularized MDPs satisfy a certain *quadratic growth* criterion, which is sufficient to establish stability. To show this, we build on the recently discovered fast convergence rates for mirror descent in regularized MDPs (Shani, Efroni, and Mannor 2020). Our result, which may be of independent interest, allows us to derive generalization bounds that can be controlled by the regularization magnitude. Furthermore, we show that when the MDP prior obeys certain structure, our results significantly improve the trivial finite hypothesis space bound.

To our knowledge, this is the first work to formally study generalization in the BRL setting. While not explicitly mentioned as such, the BRL setting has been widely used by many empirical studies on generalization in RL (Tamar et al. 2016; Cobbe et al. 2020). In fact, whenever the MDPs come from a distribution, BRL is the relevant formulation. Our results therefore also establish a formal basis for studying generalization in RL.

This paper is structured as follows. After surveying related work in Section 2, we begin with background on MDPs, BRL, and algorithmic stability in Section 3, and then present our problem formulation and straightforward upper and lower bounds for the ERM policy in Section 4. In Section 5 we discuss fundamental properties of regularized MDPs. In Section 6 we describe a general connection between a certain rate result for mirror descent and a quadratic growth condition, and in Section 7 we apply this connection to regularized MDPs, and derive corresponding generalization bounds. We discuss our results and future directions in Section 8.

2 Related Work

Generalization to novel tasks in RL has been studied extensively, often without making the explicit connection to Bayesian RL. Empirical studies can largely be classified into three paradigms. The first *increases the number of training tasks*, either using procedurally generated domains (Cobbe et al. 2020), or by means such as image augmentation (Kostrikov, Yarats, and Fergus 2020) and task interpolation (Yao, Zhang, and Finn 2021). The second paradigm *adds inductive bias* to the neural network, such as a differentiable planning or learning computation (Tamar et al. 2016; Boutilier et al. 2020), or graph neural networks (Rivlin, Hazan, and Karpas 2020). The third is *meta-RL*, where an agent is explicitely trained to generalize, either using a Bayesian RL objective (Duan et al. 2016; Zintgraf et al. 2020), or through gradient based meta learning (Finn, Abbeel, and Levine 2017). We are not aware of theoretical studies of generalization in Bayesian RL.

The Bayesian RL algorithms of Guez, Silver, and Dayan (2012) and Grover, Basu, and Dimitrakakis (2020) perform online planning in the belief space by sampling MDPs from the posterior at each time step, and optimizing over this sample. The performance bounds for these algorithms require the *correct posterior* at each step, implicitly assuming a correct prior, while we focus on *learning the prior* from data. The lower bound in our Proposition 2 demonstrates how errors in the prior can severely impact performance.

Our stability-based approach to PAC learning is based on the seminal work of Bousquet and Elisseeff (2002). More recent works investigated stability of generalized learning (Shalev-Shwartz et al. 2010), multi-task learning (Zhang 2015), and stochastic gradient descent (Hardt, Recht, and Singer 2016). To our knowledge, we provide the first stability result for regularized MDPs, which, due to their nonconvex nature, requires a new methodology. The stability results of Charles and Papailiopoulos (2018) build on quadratic growth, a property we use as well in Proposition 4. However, all the other results in this paper, including showing that quadratic growth holds for regularized MDPs, and deriving bounds for Bayesian RL, are novel.

Finally, there is recent interest in PAC-Bayes theory for meta learning (Amit and Meir 2018; Rothfuss et al. 2021; Farid and Majumdar 2021). To our knowledge, this theory has not yet been developed for meta RL.

3 Background

We give background on BRL and algorithmic stability.

3.1 MDPs and Bayesian RL

A stationary Markov decision process (MDP, Bertsekas 2006) is defined by a tuple $M = (S, A, P_{\text{init}}, C, P, H)$, where S and A are the state and actions spaces, P_{init} is an initial state distribution, $C : S \times A \to [0, C_{\max}]$ is a bounded cost function, P is the transition kernel, and H is the episode horizon, meaning that after H steps of interaction, the state is reset to $s \sim P_{\text{init}}$. We make the additional assumption that the cost $C(s, a) \in \mathcal{C}$, and $\mathcal{C}$ is a finite set.[1]

In the Bayesian RL setting (BRL, Ghavamzadeh et al. 2016), there is a distribution over MDPs $P(M)$, defined over some space of MDPs $\mathcal{M}$. For simplicity, we assume that S, A, P_{init}, and H are fixed for all MDPs in $\mathcal{M}$, and thus the only varying factors between different MDPs are the costs and transitions, denoted C_M and P_M.

A *simulator* for an MDP M is a sequential algorithm that at time t outputs s_t, and, given input a_t, outputs $c_t = C(s_t, a_t)$, and transitions the state according to $s_{t+1} \sim P(\cdot|s_t, a_t)$. After every H steps of interaction, the state is reset to $s \sim P_{\text{init}}$. Let the history at time t be $h_t = \{s_0, a_0, c_0, s_1, a_1, c_1 \ldots, s_t\}$. A *policy* π is a stochastic mapping from the history to a probability over actions $\pi(a|h_t) = P(a_t = a|h_t)$.

A typical MDP objective is to minimize the T-horizon expected return $\mathbb{E}_{\pi;M}\left[\sum_{t=0}^{T} C_M(s_t, a_t)\right]$, where the expectation is with respect to the policy π and state transitions prescribed by M. In BRL, the objective is an average over the possible MDPs in the prior:

$$\mathcal{L}(\pi) = \mathbb{E}_{M\sim P}\mathbb{E}_{\pi;M}\left[\sum_{t=0}^{T} C_M(s_t, a_t)\right]. \qquad (1)$$

We denote by $\mathcal{H}$ the space of T-length histories. Note that by our definitions above, $\mathcal{H}$ is finite. Also note that T is not necessarily equal to H.

3.2 PAC Generalization and Algorithmic Stability

Statistical learning theory (Vapnik 2013) studies the generalization performance of a prediction algorithm trained on a finite data sample. Let $S = \{z_1, \ldots, z_N\}$ denote a sample of $N \geq 1$ i.i.d. elements from some space $\mathcal{Z}$ with distribution $P(z)$. A learning algorithm $\mathcal{A}$ takes as input S and outputs a prediction function $\mathcal{A}_S$. Let $0 \leq \ell(\mathcal{A}_S, z) \leq B$,

[1]This assumption is non-standard, and required to guarantee a finite set of possible histories in the Bayesian RL setting. In practice, the reward can be discretized to satisfy the assumption.

where $z \in \mathcal{Z}$, denote the loss of the prediction on a sample z. The population risk is $R(\mathcal{A}, S) = \mathbb{E}_z [\ell(\mathcal{A}_S, z)]$, and the empirical risk is $\hat{R}(\mathcal{A}, S) = \frac{1}{N}\sum_{i=1}^{N} [\ell(\mathcal{A}_S, z_i)]$. Typically, algorithms are trained by minimizing the empirical risk. Probably approximately correct (PAC) learning algorithms are guaranteed to produce predictions with a population risk that is close to the empirical risk with high probability, and thus generalize. We recite fundamental results due to Bousquet and Elisseeff (2002) that connect algorithmic stability and PAC bounds.

Let $S^{\setminus i}$ denote the set S with element i removed. An algorithm satisfies uniform stability β if the following holds:

$$\forall S \in \mathcal{Z}^N, \forall i \in \{1, \dots, N\}, \|\ell(\mathcal{A}_S, \cdot) - \ell(\mathcal{A}_{S^{\setminus i}}, \cdot)\|_\infty \leq \beta.$$

An algorithm is said to satisfy pointwise hypothesis stability β if the following holds:

$$\forall i \in \{1, \dots, N\}, \mathbb{E}_S [|\ell(\mathcal{A}_S, z_i) - \ell(\mathcal{A}_{S^{\setminus i}}, z_i)|] \leq \beta.$$

Theorem 1 (Theorem 11 in Bousquet and Elisseeff 2002)**.** Let $\mathcal{A}$ be an algorithm with pointwise hypothesis stability β. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of S,

$$R(\mathcal{A}, S) \leq \hat{R}(\mathcal{A}, S) + \sqrt{\frac{B^2 + 12BN\beta}{2N\delta}}.$$

Theorem 2 (Theorem 12 in Bousquet and Elisseeff 2002)**.** Let $\mathcal{A}$ be an algorithm with uniform stability β. Then, for any $\delta \in (0, 1)$, with probability at least $1-\delta$ over the random draw of S,

$$R(\mathcal{A}, S) \leq \hat{R}(\mathcal{A}, S) + 2\beta + (4N\beta + B)\sqrt{\frac{\ln(1/\delta)}{2N}}.$$

The bounds in Theorems 1 and 2 are useful if one can show that for a particular problem, β scales as $o(1/\sqrt{N})$. Indeed, Bousquet and Elisseeff 2002 showed such results for several supervised learning problems. For example, L_2 regularized kernel regression has stability $\mathcal{O}(1/\lambda N)$, where λ – the regularization weight in the loss – can be chosen to satisfy the $o(1/\sqrt{N})$ condition on β.

4 Problem Formulation

We next describe our learning problem. We are given a training set of N simulators for N independently sampled MDPs, $\{M_1, \dots, M_N\}$, where each $M_i \sim P(M)$; in the following, we will sometimes refer to this training set as the *training data*. We are allowed to interact with these simulators as we wish for an unrestricted amount of time. From this interaction, our goal is to compute a policy π that obtains a low expected T-horizon cost for a *test* simulator $M \sim P(M)$, i.e., we wish to minimize the population risk (1). It is well known (e.g., Ghavamzadeh et al. 2016) that there exists a deterministic history dependent policy that minimizes (1), also known as the *Bayes-optimal policy*, and we denote it by π_{BO}. Our performance measure is the T-horizon average regret,

$$\begin{aligned}\mathcal{R}_T(\pi) =& \mathbb{E}_{M\sim P}\left[\mathbb{E}_{\pi;M}\left[\sum_{t=0}^{T} C_M(s_t, a_t)\right]\right.\\ &\left.-\mathbb{E}_{\pi_{\mathrm{BO}};M}\left[\sum_{t=0}^{T} C_M(s_t, a_t)\right]\right] = \mathcal{L}(\pi) - \mathcal{L}(\pi_{\mathrm{BO}}).\end{aligned} \tag{2}$$

Remark 1. The BRL formulation generalizes several special cases that were explored before in the context of generalization in RL. When $T = kH$, this setting is often referred to as k-shot learning, and in particular, for $T = H$, the learned policy is evaluated on solving a test task in a single shot. Another popular setting is the contextual MDP (Hallak, Di Castro, and Mannor 2015), where, in addition to the state, each task M is identified using some task identifier id_M, which is observed. By adding id_M to the state space, and modifying the dynamics such that id_M does not change throughout the episode, this setting is a special case of our formalism. Finally, many previous studies (e.g., Tamar et al. 2016) considered the same performance objective, but limited the optimization to Markov policies (i.e., policies that depend only on the current state and not on the full history). In this work, we specifically consider history dependent policies, as it allows us to meaningfully compare the learned policy with the optimum.

4.1 Analysis of an ERM Approach

Our goal is to study the generalization properties of learning algorithms in the BRL setting. An intuitive approach, in the spirit of the empirical risk minimization (ERM, Vapnik 2013) principle, is to minimize the *empirical risk*,

$$\begin{aligned}\hat{\mathcal{L}}(\pi) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi;M_i}\left[\sum_{t=0}^{T} C_{M_i}(s_t, a_t)\right]\\ &\equiv \mathbb{E}_{M\sim \hat{P}_N}\mathbb{E}_{\pi,M}\left[\sum_{t=0}^{T} C_M(s_t, a_t)\right],\end{aligned} \tag{3}$$

where $\hat{P}_N$ is the empirical distribution of the N sampled MDPs. Let $\hat{\pi}^* \in \arg\min_{\pi\in\mathcal{H}} \hat{\mathcal{L}}(\pi)$ denote the ERM policy.

Since the hypothesis space of deterministic history dependent policies is finite, and the loss is bounded by $C_{\max}T$, a trivial generalization bound can be formulated as follows (following PAC bounds for a finite hypothesis class, e.g., Shalev-Shwartz and Ben-David 2014).

Proposition 1. Consider the ERM policy $\hat{\pi}^*$, and let $\bar{\mathcal{H}}$ denote the set of deterministic T-length history dependent policies. Then with probability at least $1 - \delta$,

$$\mathcal{R}_T(\hat{\pi}^*) \leq \sqrt{\frac{2\log(2|\bar{\mathcal{H}}|/\delta)C_{\max}^2 T^2}{N}}.$$

Note that $|\bar{\mathcal{H}}| = |A|^{|\mathcal{H}|} \approx |A|^{(|S||A||\mathcal{C}|)^T}$, so $\log|\bar{\mathcal{H}}| = \mathcal{O}((|S||A||\mathcal{C}|)^T)$. The exponential dependence on T in the bound is not very satisfying, and one may ask whether a more favourable upper bound can be established for the ERM policy. To answer this, we next give a lower bound, showing that without additional assumptions on the problem, the exponential dependence on T is necessary.

Proposition 2. For any $0 \leq \delta < 1$, there is an $\epsilon > 0$, and a problem, such that for $N = 2^T$, with probability larger than δ we have $\mathcal{R}_T(\hat{\pi}^*) > \epsilon$.

Proof. (sketch; full proof in Section E.) Let $T = H$, and consider an MDP space $\mathcal{M}$ of size 2^H, where the state space has $2H + 1$ states that we label

$s_0, s_1^0, s_1^1, \ldots, s_t^0, s_t^1, \ldots, s_H^0, s_H^1$. The initial state for all MDPs in $\mathcal{M}$ is s_0. A cost is only obtained at the last time step, and depends only on the last action. Each MDP $M \in \mathcal{M}$ corresponds to a unique binary number of size H, denoted x, and the transitions for each MDP correspond to the digits in its identifier x: there is a high probability to transition to s_t^0 from either s_{t-1}^0 or s_{t-1}^1 only if the t's digit of x is zero, and similarly, there is a high probability to transition to s_t^1 from either s_{t-1}^0 or s_{t-1}^1 only if the t's digit of x is one. Thus, with high probability, a trajectory in the MDP traces the digits of its identifier x. Given a finite data sample, there is a non-negligible set of MDPs that will not appear in the data. For any trajectory that corresponds to an x from this set, the ERM policy at time H will not be able to correctly identify the most probable MDP, and will choose an incorrect action with non-negligible probability. $\square$

The results above motivate us to seek alternatives to the ERM approach, with the hope of providing more favorable generalization bounds. In the remainder of this paper, we focus on methods that add a regularization term to the loss.

5 Regularized MDPs

In supervised learning, a well-established method for controlling generalization is to add a regularization term, such as the L_2 norm of the parameters, to the objective function that is minimized. The works of Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010) showed that for convex loss functions, adding a strongly convex regularizer such as the L_2 norm leads to algorithmic stability, which can be used to derive generalization bounds that are controlled by the amount of regularization. In this work, we ask whether a similar approach of adding regularization to the BRL objective (3) can be used to control generalization.

We focus on the following regularization scheme. For some $\lambda > 0$, consider a regularized ERM of the form:

$$\hat{\mathcal{L}}^{\lambda}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi;M_i}\left[\sum_{t=0}^{T} C_{M_i}(s_t, a_t) + \lambda \mathcal{R}(\pi(\cdot|h_t))\right],$$

where $\mathcal{R}$ is some regularization function applied to the policy. In particular, we will be interested in L_2 regularization, $\mathcal{R}(\pi(\cdot|h_t)) = \|\pi(\cdot|h_t)\|_2$, and in the remainder of this paper $\mathcal{R}$ corresponds to this form. We also define the regularized population risk,

$$\mathcal{L}^{\lambda}(\pi) = \mathbb{E}_{M\sim P}\mathbb{E}_{\pi;M}\left[\sum_{t=0}^{T} C_M(s_t, a_t) + \lambda \mathcal{R}(\pi(\cdot|h_t))\right].$$

In standard (non-Bayesian) RL, regularized MDPs have been studied extensively (Neu, Jonsson, and Gómez 2017). A popular motivation has been to use the regularization to induce exploration (Fox, Pakman, and Tishby 2015; Schulman et al. 2017). Recently, Shani, Efroni, and Mannor (2020) showed that for optimizing a policy using k iterations of mirror descent (equivalent to trust region policy optimization Schulman et al. 2015) with L_2 or entropy regularization enables a fast $O(1/k)$ convergence rate, similarly to convergence rates for strongly convex functions, although the MDP objective is not convex. In our work, we build on these results to show a stability property for regularization in the BRL setting described above. We begin by adapting a central result in Shani, Efroni, and Mannor (2020), which was proved for discounted MDPs, to our finite horizon and history dependent policy setting.

The BRL objectives in Eq. (1) (similarly, Eq. (3)) can be interpreted as follows: we first choose a history dependent policy $\pi(h_t)$, and then nature draws an MDP $M \sim P(M)$ (similarly, $M \sim \hat{P}_N$), and we then evaluate $\pi(h_t)$ on M. The expected cost (over the draws of M), is the BRL performance. In the following discussion, for simplicity, the results are given for the prior $P(M)$, but they hold for $\hat{P}_N$ as well.

Let $P(M|h_t;\pi)$ denote the posterior probability of nature having drawn the MDP M, given that we have seen the history h_t under policy π. From Bayes rule, we have that

$$P(M|h_t;\pi) \propto P(h_t|M;\pi)P(M).$$

Let us define the regularized expected cost,

$$C_\lambda(h_t, a_t;\pi) = \mathbb{E}_{M|h_t;\pi} C_M(s_t, a_t) + \lambda \mathcal{R}(\pi_t(\cdot|h_t)),$$

and the value function,

$$V_t^\pi(h_t) = \mathbb{E}_{\pi;M|h_t}\left[\sum_{t'=t}^{T} C_\lambda(h_{t'}, a_{t'};\pi)\middle| h_t\right].$$

The value function satisfies Bellman's equation. Let

$$P(c_t, s_{t+1}|h_t, a_t) = \sum_M P(M|h_t)P(c_t|M, s_t, a_t)P(s_{t+1}|M, s_t, a_t)$$

denote the posterior probability of observing c_t, s_{t+1} at time t. Then

$$V_T^\pi(h_T) = \sum_{a_T} \pi(a_T|h_T) C_\lambda(h_T, a_T;\pi),$$

and, letting $h_{t+1} = \{h_t, a_t, c_t, s_{t+1}\}$,

$$V_t^\pi(h_t) = \sum_{a_t}\pi(a_t|h_t)\Big(C_\lambda(h_t, a_t;\pi) + \sum_{c_t, s_{t+1}} P(c_t, s_{t+1}|h_t, a_t)V_{t+1}^\pi(\{h_t, a_t, c_t, s_{t+1}\})\Big).$$

Consider two histories $h_t, \bar{h}_{\bar{t}} \in \mathcal{H}$, and let

$$\mathbf{P}^\pi(\bar{h}_{\bar{t}}|h_t) = \begin{cases}\sum_{a_t}\pi(a_t|h_t)P(\bar{c}_t, \bar{s}_{t+1}|h_t, a_t), & \text{if } \bar{t} = t+1\\ 0, & \text{else}\end{cases}$$

denote the transition probability between histories. Also, define

$$C^\pi(h_t) = \sum_{a_t}\pi(a_t|h_t)C_\lambda(h_t, a_t;\pi).$$

We can write the Bellman equation in matrix form as follows

$$V^\pi = C^\pi + \mathbf{P}^\pi V^\pi, \tag{4}$$

where V^π and C^π are vectors in $\mathbb{R}^{|\mathcal{H}|}$, and $\mathbf{P}^\pi$ is a matrix in $\mathbb{R}^{|\mathcal{H}|\times|\mathcal{H}|}$.

The uniform trust region policy optimization algorithm of Shani, Efroni, and Mannor (2020) is a type of mirror descent algorithm applied to the policy in regularized MDPs. An adaptation of this algorithm for our setting is given in Sec. B.4 of the supplementary material. The next result provides a fundamental inequality that the policy updates of this algorithm satisfy, in the spirit of an inequality that is used to establish convergence rates for mirror descent (cf. Lemma 8.11 in Beck 2017). The proof follows Lemma 10 in Shani, Efroni, and Mannor (2020), with technical differences due to the finite horizon setting; it is given in Sec. B.4.

Proposition 3. Let $\{\pi_k\}$ be the sequence generated by uniform trust region policy optimization with step sizes $\{\alpha_k\}$ and L_2 regularization. Then for every π and $k \geq 0$,

$$\alpha_k(I - \mathbf{P}^\pi)(V^{\pi_k} - V^\pi) \leq \frac{(1-\alpha_k\lambda)}{2}\|\pi - \pi_k\|_2^2 - \frac{1}{2}\|\pi - \pi_{k+1}\|_2^2 + \frac{\lambda\alpha_k}{2}(\|\pi_k\|_2^2 - \|\pi_{k+1}\|_2^2) + \frac{\alpha_k^2 L^2}{2}e,$$

where e is a vector of ones, $L = C_{\max}T|A|$, and $\|\pi\|_2 \in \mathbb{R}^{|\mathcal{H}|}$ denotes the L_2 norm of the policy element-wise, for each history.

In the following, we shall show that Proposition 3 can be used to derive stability bounds in the regularized BRL setting. To simplify our presentation, we first present a key technique that our approach builds on in a general optimization setting, and only then come back to MDPs.

6 Stability based on the Fundamental Inequality for Mirror Descent

Standard stability results, such as in Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010), depend on convexity of the loss function, and strong convexity of the regularizing function (Bousquet and Elisseeff 2002). While our L_2 regularization is strongly convex, the MDP objective is not convex in the policy.[2] In this work, we show that nevertheless, algorithmic stability can be established. To simplify our presentation, we first present the essence of our technique in a general form, without the complexity of MDPs. In the next section, we adapt the technique to the BRL setting.

Our key insight is that the fundamental inequality of mirror descent (cf. Prop. 3), actually prescribes a quadratic growth condition. The next lemma shows this for a general iterative algorithm, but it may be useful to think about mirror descent when reading it. In the sequel, we will show that similar conditions hold for regularized MDPs.

Lemma 1. Let $f : \mathcal{X} \to \mathbb{R}$ be some function that attains a minimum $f(x^*) \leq f(x) \quad \forall x \in \mathcal{X}$. Consider a sequence of step sizes $\alpha_0, \alpha_1, \dots \in \mathbb{R}^+$ and corresponding sequence of iterates $x_0, x_1, \dots \in \mathcal{X}$. Assume that $f(x_{k+1}) \leq f(x_k)$ for all $k \geq 0$. Also consider a sequence of values $z_0, z_1, \dots \in \mathbb{R}^+$ that satisfy $|z_k - z_0| \leq B$ for all $k \geq 0$. Assume that there exists $\lambda > 0$ and $L \geq 0$ such that the following holds for any step size sequence, all $k \geq 0$, and any $x \in \mathcal{X}$:

$$\alpha_k\left(f(x_k) - f(x)\right) \leq (1 - \lambda\alpha_k)\|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \lambda\alpha_k\left(z_k - z_{k+1}\right) + \frac{\alpha_k^2 L^2}{2}. \tag{5}$$

Then the following statements hold true.

1. For step sizes $\alpha_k = \frac{1}{\lambda(k+2)}$, the sequence converges to x^* at rate
$$f(x_k) - f(x^*) \leq \frac{L^2 \log k}{\lambda k}.$$
2. Quadratic growth: $\lambda\|x^* - x_0\|^2 \leq f(x_0) - f(x^*)$.

Proof. The first claim is similar to Theorem 2 of Shani, Efroni, and Mannor (2020); for completeness we give a full proof in Sec. A of the supplementary. We prove the second claim. Let $\alpha_k = \frac{1}{\lambda(k+2)}$, and multiply (5) by $\lambda(k+2)$:

$$f(x_k) - f(x_0) \leq \lambda(k+1)\|x_k - x_0\|^2 - \lambda(k+2)\|x_{k+1} - x_0\|^2 + \lambda\left(z_k - z_{k+1}\right) + \frac{L^2}{2\lambda(k+2)}.$$

Summing over k, and observing the telescoping sums:

$$\sum_{k=0}^{N}\left(f(x_k) - f(x_0)\right) \leq -\lambda(N+2)\|x_{N+1} - x_0\|^2 + \lambda\left(z_0 - z_{N+1}\right) + \frac{L^2}{2\lambda}\sum_{k=0}^{N}\frac{1}{(k+2)} \leq -\lambda(N+2)\|x_{N+1} - x_0\|^2 + \lambda B + \frac{L^2\log(N+2)}{2\lambda}.$$

Since $f(x_k)$ is decreasing, $\sum_{k=0}^{N}\left(f(x_N) - f(x^*)\right) \leq \sum_{k=0}^{N}\left(f(x_k) - f(x^*)\right)$, and

$$N\left(f(x_N) - f(x_0)\right) \leq -\lambda(N+2)\|x_{N+1} - x_0\|^2 + \lambda B + \frac{L^2\log(N+2)}{2\lambda}.$$

Dividing by N, taking $N \to \infty$, and using the first part of the lemma:

$$f(x^*) - f(x_0) \leq -\lambda\|x^* - x_0\|^2.$$

Rearranging give the result. □

We now present a stability result for a regularized ERM objective. The proof resembles Shalev-Shwartz et al. (2010), but replaces strong convexity with quadratic growth.

Proposition 4. Let $z_0, z_1 \dots \in \mathcal{Z}$ denote a sequence of independent samples, and let $\ell : \mathcal{X} \times \mathcal{Z} \to \mathcal{R}$ be a loss for a predictor $x \in \mathcal{X}$ and sample $z \in \mathcal{Z}$. Consider a regularized ERM objective $L_N(x) = \frac{1}{N}\sum_{i=1}^{N}\ell(x, z_i) + \lambda\mathcal{R}(x)$, and let $L_N^{\setminus j} = \frac{1}{N}\sum_{\substack{i=1 \\ i \neq j}}^{N}\ell(x, z_i) + \lambda\mathcal{R}(x)$. Assume that ℓ is β-Lipschitz: for any $z \in \mathcal{Z}$, and any x, x', $|\ell(x,z) - \ell(x',z)| \leq \beta\|x - x'\|$. Assume that $L_N(x)$ and $L_N^{\setminus j}(x)$ have unique minimizers, and denote them x^*

[2] The linear programming formulation is not suitable for establishing stability in our BRL setting, as changing the prior would change the constraints in the linear program.

and $x^{*,\setminus j}$, respectively. Further assume quadratic growth: $\lambda\|x^* - x\|^2 \leq L_N(x) - L_N(x^*)$ for any $x \in \mathcal{X}$. Then, we have that

$$\|x^* - x^{*,\setminus j}\| \leq \frac{\beta}{\lambda N},$$

and $\forall z \in \mathcal{Z}$

$$\ell(x^*, z) - \ell(x^{*,\setminus j}, z) \leq \frac{\beta^2}{\lambda N}.$$

Proof. (sketch; full proof in Sec. A.) Let $\Delta = L_N(x^{*,\setminus j}) - L_N(x^*)$. From quadratic growth, we have that

$$\Delta \geq \lambda\|x^* - x^{*,\setminus j}\|^2.$$

On the other hand, by taking out the j'th element from the loss terms L_N, and observing that $x^{*,\setminus j}$ minimizes $L_N^{\setminus j}$, we have that

$$\begin{aligned}\Delta =& \frac{1}{N}\sum_{\substack{i=1\\i\neq j}}^{N}\ell(x^{*,\setminus j}, z_i) + \lambda\mathcal{R}(x^{*,\setminus j}) - \frac{1}{N}\sum_{\substack{i=1\\i\neq j}}^{N}\ell(x^*, z_i) - \lambda\mathcal{R}(x^*)\\ &+ \frac{\ell(x^{*,\setminus j}, z_j) - \ell(x^*, z_j)}{N}\\ \leq& \frac{\ell(x^{*,\setminus j}, z_j) - \ell(x^*, z_j)}{N},\end{aligned}$$

and from the Lipschitz condition, $\Delta \leq \frac{\beta\|x^* - x^{*,\setminus j}\|}{N}$. Combining the above inequalities for Δ gives $\|x^* - x^{*,\setminus j}\| \leq \frac{\beta}{\lambda N}$, and the final result is obtained by using the Lipschitz condition one more time. □

7 Stability for Regularized Bayesian RL

We are now ready to present stability bounds for the L_2-regularized Baysian RL setting. Let $\mu \in \mathbb{R}^{|\mathcal{H}|}$ denote the distribution over h_0, the initial history (we assume that all elements in the vector that correspond to histories of length greater than 0 are zero). Recall the regularized ERM loss $\hat{\mathcal{L}}^\lambda(\pi)$, and let π^* denote its minimizer. Define the leave-one-out ERM loss,

$$\hat{\mathcal{L}}^{\lambda,\setminus j}(\pi) = \frac{1}{N}\sum_{\substack{i=1\\i\neq j}}^{N}\mathbb{E}_{\pi;M_i}\left[\sum_{t=0}^{T} C(s_t, a_t) + \lambda\mathcal{R}(\pi(\cdot|h_t))\right],$$

and let $\pi^{\setminus j,*}$ its minimizer. Recall the definition of $\mathbf{P}^\pi$ – the transition probability between histories under policy π, which depends on the prior. In the following, we use the following notation: $\mathbf{P}^\pi$ refers to the empirical prior $\hat{P}_N$, while $\mathbf{P}^\pi_{M_j}$ refers to a prior that has all its mass on a single MDP M_j. The following theorem will be used to derive our stability results. The proof is in Sec. C.

Theorem 3. Let $\Delta = \hat{\mathcal{L}}^\lambda(\pi^{\setminus j,*}) - \hat{\mathcal{L}}^\lambda(\pi^*)$. We have that

$$\Delta \geq \frac{\lambda}{2}\mu^\top(I - \mathbf{P}^{\pi^{\setminus j,*}})^{-1}\|\pi^{\setminus j,*} - \pi^*\|_2^2,$$

and

$$\Delta \leq \frac{1}{N}C_{\max}T\sqrt{|A|}\mu^\top(I - \mathbf{P}^{\pi^{\setminus j,*}}_{M_j})^{-1}\left\|\pi^{\setminus j,*} - \pi^*\right\|_2.$$

Following the proof of Proposition 4, we would like to use the two expressions in Theorem 3 to bound $\left\|\pi^{\setminus j,*} - \pi^*\right\|_2$, which would directly lead to a stability result. This is complicated by the fact that different factors $(I - \mathbf{P}^{\pi^{\setminus j,*}})^{-1}$ and $(I - \mathbf{P}^{\pi^{\setminus j,*}}_{M_j})^{-1}$ appear in the two expressions. Our first result assumes that these expressions cannot be too different; a discussion of this assumption follows.

Assumption 1. For any two MDPs $M, M' \in \mathcal{M}$, and any policy π, let $\mathbf{P}^\pi_M$ and $\mathbf{P}^\pi_{M'}$ denote their respective history transition probabilities. There exists some $D < \infty$ such that for any $x \in \mathbb{R}^{|\mathcal{H}|}$

$$\mu^\top(I - \mathbf{P}^\pi_M)^{-1}x \leq D\mu^\top(I - \mathbf{P}^\pi_{M'})^{-1}x.$$

Let us define the regularized loss for MDP M, $\mathcal{L}^\lambda_M(\pi) = \mathbb{E}_{\pi;M}\left[\sum_{t=0}^{T} C_M(s_t, a_t) + \lambda\mathcal{R}(\pi(\cdot|h_t))\right]$. We have the following result.

Corollary 1. Let Assumption 1 hold, and let $\kappa = 2D^2C^2_{\max}T^2|A|$. Then, for any MDP $M' \in \mathcal{M}$,

$$\mathcal{L}^\lambda_{M'}(\pi^{\setminus j,*}) - \mathcal{L}^\lambda_{M'}(\pi^*) \leq \frac{\kappa}{\lambda N},$$

and with probability at least $1 - \delta$,

$$\mathcal{R}_T(\hat{\pi}^*) \leq 2\lambda T + \frac{2\kappa}{\lambda N} + \left(\frac{4\kappa}{\lambda} + 3C_{\max}T\right)\sqrt{\frac{\ln(1/\delta)}{2N}}.$$

Note that each element that corresponds to history h_t in the vector $\mu^\top(I - \mathbf{P}^\pi_M)^{-1}$ is equivalent to $P(h_t|M;\pi)$, the probability of observing h_t under policy π and MDP M (see Sec. B.1 for formal proof). Thus, Assumption 1 essentially states that two different MDPs under the prior cannot visit completely different histories given the same policy. With our regularization scheme, such an assumption is required for uniform stability: if the test MDP can reach completely different states than possible during training, it is impossible to guarantee anything about the performance of the policy in those states. Unfortunately, the constant D can be very large. Let

$$q = \sup_{M,M'\in\mathcal{M}, s,s'\in S, a\in A, c\in\mathcal{C}}\frac{P_M(s', c|s, a)}{P_{M'}(s', c|s, a)},$$

where we assume that $0/0 = 1$. Then, $P(h_t|M;\pi)/P(h_t|M';\pi) = \Pi_t\frac{P_M(s_{t+1}, c_t|s_t, a_t)}{P_{M'}(s_{t+1}, c_t|s_t, a_t)}$ is at most q^T, and therefore D can be in the order of q^T. One way to guarantee that D is finite, is to add a small amount of noise to any state transition. The following example estimates q is such a case.

Example 1. Consider modifying each MDP M in $\mathcal{M}$ such that $P_M(s', c|s, a) \to (1 - \alpha)P_M(s', c|s, a) + \alpha/|S||\mathcal{C}|$, where $\alpha \in (0, 1)$. In this case, $q \leq \frac{(1-\alpha)|S||\mathcal{C}|}{\alpha}$.

Let us now compare Corollary 1 with the trivial bound in Proposition 1. First, Corollary 1 allows to control generalization by increasing the regularization λ. The term $2\lambda T$ is a bias, incurred by adding the regularization to the objective, and can be reduced by decreasing λ. Comparing the constants of the $\mathcal{O}(1/\sqrt{N})$ term, the dominant terms are D^2 vs. $(|S||\mathcal{C}||A|)^T$. Since D does not depend on $|A|$, the bound in Corollary 1 is important for problems with large $|A|$. The example above shows that in the worst case, D^2 can be $\mathcal{O}((|S||\mathcal{C}|)^{2T})$. There are, of course, more favorable case, where the structure of $\mathcal{M}$ is such that D is better behaved.

Example 2. Consider an hypothesis set $\mathcal{M}$ such that all MDPs in $\mathcal{M}$ differ only on a set of states that cannot be visited more than k times in an episode. In this case, D would be in the order of $q^{kT/H}$.

Another case is where the set $\mathcal{M}$ is finite. In this case, we can show that the pointwise hypothesis stability does not depend on D, and we obtain a bound that does not depend exponentially on T, as we now show.

Corollary 2. Let $\mathcal{M}$ be a finite set, and let $P_{\min} = \min_{M\in\mathcal{M}} P(M)$. Then

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}^{\lambda}_{M_j}(\pi^{\setminus j,*}) - \mathcal{L}^{\lambda}_{M_j}(\pi^*)\right]\\ &\leq \frac{4C_{\max}^2 T^2|A|}{\lambda N P_{\min}} + \exp\left(\frac{-NP_{\min}}{8}\right)C_{\max}T,\end{aligned}$$

and with probability at least $1-\delta$, (ignoring exponential terms)

$$\mathcal{R}_T(\hat{\pi}^*) \leq 2\lambda T + \sqrt{\frac{C_{\max}^2 T^2}{2N\delta} + \frac{48 C_{\max}^3 T^3 |A|}{2\delta\lambda N P_{\min}}}.$$

In the generalization bounds of both Corollary 1 and Corollary 2, reducing λ and increasing N at a rate such that the stability is $o(1/\sqrt{N})$ will guarantee learnability, i.e., convergence to π_{BO} as $N\to\infty$.

Example 3. Under the setting of Corollary 2, letting $\lambda = N^{-1/3}$ gives that, with probability at least $1-\delta$, (ignoring exponential terms)

$$\mathcal{R}_T(\hat{\pi}^*) \leq \frac{2T}{N^{1/3}} + \sqrt{\frac{C_{\max}^2 T^2}{2N\delta} + \frac{48 C_{\max}^3 T^3 |A|}{2\delta N^{2/3} P_{\min}}}.$$

For a finite N, and when there is structure the hypothesis space $\mathcal{M}$, as displayed for example in D, the bounds in Corollaries 1 and 2 allow to set λ to obtain bounds that are more optimistic than the trivial bound in Proposition 1. In these cases, our results show that regularization allows for improved generalization.

Example 4. Set $\lambda = 1$. Then the bound in Corollary 2 becomes

$$\mathcal{R}_T(\hat{\pi}^*) \leq 2T + \sqrt{\frac{C_{\max}^2 T^2}{2N\delta} + \frac{48 C_{\max}^3 T^3 |A|}{2\delta N P_{\min}}},$$

while the naive bound is

$$\mathcal{R}_T(\hat{\pi}^*) \leq \sqrt{\frac{\ln(2/\delta) + (|S||A||\mathcal{C}|)^T C_{\max}^2 T^2}{N}}.$$

For a finite N that is much smaller than $(|S||A||\mathcal{C}|)^T$, and for reasonable values of $P_{\min}$ and δ, the naive bound can be completely vacuous (larger than $C_{\max}T$ – the maximum regret possible), while the bound in Corollary 2 can be significantly smaller.

8 Discussion

In this work, we analyzed generalization in Bayesian RL, focusing on algorithmic stability and a specific form of policy regularization. The bounds we derived can be controlled by the amount of regularization, and under some structural assumptions on the space of possible MDPs, compare favorably to a trivial bound based on the finite policy space. We next outline several future directions.

Specialized regularization for k-shot learning: One can view our results as somewhat pessimistic – at worst, they require that every history has a non-zero probability of being visited, and even then, the dependence on T can be exponential. One may ask whether alternative regularization methods could relax the dependence on T. We believe this is true, based on the following observation. Recall the example in the lower bound of Proposition 2. Let $T = kH$, and consider the policy that at time step $t = H$ chooses an action, and based on the observed cost chooses which action to use at time steps $t = 2H, t = 3H, \ldots$. Note that after observing the first cost, it is clear which action is optimal, and therefore the policy obtains at most a $\frac{k-1}{k}$ fraction of the optimal total cost on both training and test, *regardless of the training data*. More generally, there exist deterministic policies, such as the Q-learning algorithm of Jin et al. (2018), that achieve $\mathcal{O}\left(\sqrt{H^3|S||A|T}\right)$ regret for *any MDP*. Thus, we believe that in the k-shot learning setting, regularization methods that *induce efficient exploration* can be devised. We leave this as an open problem.

Continuous MDPs: Another important direction is developing PAC algorithms for continuous state, cost, and action spaces. It is clear that without the finite hypothesis space assumption, overfitting is a much more serious concern; in Sec. F of the supplementary material we provide a simple example of this, when only the costs do not belong to a finite set. We hypothesize that regularization techniques can be important in such settings, in concordance with known results for supervised learning. We believe that the tools for stability analysis in MDPs that we developed in this work may be useful for this problem, which we leave to future work.

Implicit regularization: Finally, all the results in this work considered *optimal* solutions of the regularized Bayesian RL problem. In practice, due to the size of the state space that scales exponentially with the horizon, computing such policies is intractable even for medium-sized problems. Interestingly, approximate solutions do not necessarily hurt generalization: *implicit* regularization, for example as implied by using stochastic gradient descent for optimization, is known to improve generalization, at least in supervised learning (Hardt, Recht, and Singer 2016; Zou et al. 2021). We hypothesize that stability results similar to Hardt, Recht, and Singer (2016) may be developed for Bayesian RL as well, using the quadratic growth property established here.

Acknowledgements

Aviv Tamar thanks Kfir Levy for insightful discussions on optimization and stability, and Shie Mannor for advising on the presentation of this work. The research of Aviv Tamar was partly funded by the Israel Science Foundation (grant No. 759/19). The research of Daniel Soudry was supported by the Israel Science Foundation (grant No. 1308/18), and by the Israel Innovation Authority (the Avatar Consortium).

References

Amit, R.; and Meir, R. 2018. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, 205–214. PMLR.

Beck, A. 2017. *First-order methods in optimization*. SIAM.

Bertsekas, D. P. 2006. *Dynamic programming and optimal control*, volume 2. Athena Scientific.

Bousquet, O.; and Elisseeff, A. 2002. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526.

Boutilier, C.; Hsu, C.-w.; Kveton, B.; Mladenov, M.; Szepesvari, C.; and Zaheer, M. 2020. Differentiable Meta-Learning of Bandit Policies. *Advances in Neural Information Processing Systems*, 33.

Charles, Z.; and Papailiopoulos, D. 2018. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, 745–754. PMLR.

Cobbe, K.; Hesse, C.; Hilton, J.; and Schulman, J. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, 2048–2056. PMLR.

Dorfman, R.; Shenfeld, I.; and Tamar, A. 2020. Offline Meta Learning of Exploration. *arXiv preprint arXiv:2008.02598*.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Farid, A.; and Majumdar, A. 2021. PAC-BUS: Meta-Learning Bounds via PAC-Bayes and Uniform Stability. *arXiv preprint arXiv:2102.06589*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Fox, R.; Pakman, A.; and Tishby, N. 2015. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*.

Ghavamzadeh, M.; Mannor, S.; Pineau, J.; and Tamar, A. 2016. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*.

Grover, D.; Basu, D.; and Dimitrakakis, C. 2020. Bayesian reinforcement learning via deep, sparse sampling. In *International Conference on Artificial Intelligence and Statistics*, 3036–3045. PMLR.

Guez, A.; Silver, D.; and Dayan, P. 2012. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in neural information processing systems*, 1025–1033.

Hallak, A.; Di Castro, D.; and Mannor, S. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 1225–1234. PMLR.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-learning provably efficient? *arXiv preprint arXiv:1807.03765*.

Kostrikov, I.; Yarats, D.; and Fergus, R. 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*.

Mendelson, A. F.; Zuluaga, M. A.; Hutton, B. F.; and Ourselin, S. 2016. What is the distribution of the number of unique original items in a bootstrap sample? *arXiv preprint arXiv:1602.05822*.

Neu, G.; Jonsson, A.; and Gómez, V. 2017. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*.

Rivlin, O.; Hazan, T.; and Karpas, E. 2020. Generalized planning with deep reinforcement learning. *arXiv preprint arXiv:2005.02305*.

Rothfuss, J.; Fortuin, V.; Josifoski, M.; and Krause, A. 2021. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *International Conference on Machine Learning*, 9116–9126. PMLR.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11: 2635–2670.

Shani, L.; Efroni, Y.; and Mannor, S. 2020. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5668–5675.

Strehl, A. L.; Li, L.; Wiewiora, E.; Langford, J.; and Littman, M. L. 2006. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, 881–888.

Tamar, A.; Wu, Y.; Thomas, G.; Levine, S.; and Abbeel, P. 2016. Value iteration networks. In *Advances in Neural Information Processing Systems*, 2154–2162.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer science & business media.

Yao, H.; Zhang, L.; and Finn, C. 2021. Meta-Learning with Fewer Tasks through Task Interpolation. *arXiv preprint arXiv:2106.02695*.

Zhang, Y. 2015. Multi-task learning and algorithmic stability. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zintgraf, L.; Shiarlis, K.; Igl, M.; Schulze, S.; Gal, Y.; Hofmann, K.; and Whiteson, S. 2020. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representation (ICLR)*.

Zou, D.; Wu, J.; Braverman, V.; Gu, Q.; Foster, D. P.; and Kakade, S. M. 2021. The Benefits of Implicit Regularization from SGD in Least Squares Problems. *arXiv preprint arXiv:2108.04552*.