Classification-based Approximate Policy Iteration

Amir-massoud Farahmand, Doina Precup, André M.S. Barreto, Mohammad Ghavamzadeh

***Abstract*—Tackling large approximate dynamic programming or reinforcement learning problems requires methods that can exploit regularities of the problem in hand. Most current methods are geared towards exploiting the regularities of either the value function or the policy. We introduce a general classification-based approximate policy iteration (CAPI) framework that can exploit regularities of both. We establish theoretical guarantees for the sample complexity of CAPI-style algorithms, which allow the policy evaluation step to be performed by a wide variety of algorithms, and can handle nonparametric representations of policies. Our bounds on the estimation error of the performance loss are tighter than existing results.**[1]

***Index Terms*—Approximate Dynamic Programming, Reinforcement Learning, Approximate Policy Iteration, Classification, Finite-Sample Analysis**

I. Introduction

We consider the problem of finding a near-optimal policy (i.e., controller) for discounted Markov Decision Processes (MDPs) with large state space and finite action space [4] with an unknown model. For problems with large state spaces (e.g., when the state space is $\mathbb{R}^d$ with large d), finding a close-to-optimal policy is difficult (due to the so-called curse of dimensionality) unless one benefits from regularities, or special structure, of the problem in hand. One group of algorithms developed in reinforcement learning (RL) and approximate dynamic programming (ADP) focuses on exploiting regularities of the *value* function [5, 6, 7, 8, 9], while another group tries to benefit from regularities of the *policy* [10, 11, 12]. The goal of this paper is to introduce and analyze a class of algorithms, which we call Classification-based Approximate Policy Iteration (CAPI), that can potentially benefit simultaneously from both types of regularities.

Our approach is inspired by existing classification-based RL algorithms [13, 14, 15]. These methods use rollout (i.e., Monte Carlo trajectories) to roughly estimate the action-value function of the current policy at several states. The estimates define a set of (noisy) greedy actions (positive examples) and non-greedy actions (negative examples), which are then fed to a classifier. The classifier "generalizes" the greedy action choices over the state space. The procedure is repeated.

Classification-based methods can be interpreted as variants of Approximate Policy Iteration (API) that use rollouts to estimate the action-value function (policy evaluation) and then *project* the greedy policy obtained at those points onto a given space of controllers (policy improvement).

Although classification-based RL methods can benefit from regularities of the policy, the use of rollouts prevents generalization through the value function, which reduces data efficiency. This lack of generalization makes rollout-based estimators data-inefficient. This is a concern in real problems, in which new samples may be expensive, e.g., in adaptive treatment strategies. Moreover, one cannot easily use rollouts when only access to a batch of data is allowed and a generative model or simulator of the environment is not available.

To address the limitation of rollout-based estimators, we propose the CAPI framework. CAPI generalizes the current classification-based algorithms by allowing the use any policy evaluation method including, but not limited to, rollout-based estimators (as in previous work [13, 15]), LSTD [16], the policy evaluation version of Fitted Q-Iteration [17], and their regularized variants [5, 7], as well as online methods for policy evaluation such as Temporal Difference learning. This is a significant generalization of the existing classification-based RL algorithms, which become special cases of CAPI. Our theoretical results indicate that this extension is indeed sound. CAPI uses a weighted loss instead of the conventional $0/1$-loss of classification, which may lead to surprisingly bad policies [3].

The main theoretical contribution of this paper is the finite-sample error analysis of CAPI-style algorithms, which allows *general policy evaluation* algorithms, handles *nonparametric* (in the sense used by e.g., [18, 19]) policy spaces, and provides a *faster convergence rate for the estimation error* than existing results. Using nonparametric policies is a significant extension of the work by Fern et al. [14], which is limited to finite policy spaces, and of Lazaric et al. [15] and Gabillon et al. [20], which are limited to policy spaces with finite Vapnik-Chervonenkis (VC) dimension. Our faster convergence rates are due to using a concentration inequality based on the powerful notion of *local Rademacher complexity* [21], which is known to lead to fast rates in supervised learning.

We also leverage the notion of *action-gap regularity* [22], which implies that choosing the right action at each state may not require a precise estimate of the action-value function. When the action-gap regularity of a problem is favourable, the convergence rate of CAPI is faster than the convergence rate of the estimate of the action-value function (and without any such assumption, the convergence rate is the same).

A.M. Farahmand is affiliated with the School of Computer Science, McGill University, Montreal, Canada, the Robotics Institute, Carnegie Mellon University, Pittsburgh, USA, and Mitsubishi Electric Research Laboratories, Cambridge, USA (webpage: academic.SoloGen.net).

D. Precup is with the School of Computer Science, McGill University, Montreal, Canada (email: dprecup@cs.mcgill.ca).

A.M.S. Barreto is affiliated with the School of Computer Science, McGill University, Montreal, Canada and the National Laboratory for Scientific Computing (LNCC), Petrópolis, Brazil (e-mail: amsb@lncc.br).

M. Ghavamzadeh is with Adobe Research, USA on leave of absence from INRIA Lille, France (email: mohammad.ghavamzadeh@inria.fr).

Manuscript received on November 18, 2013, and revised and resubmitted on July 3, 2014 and November 29, 2014.

[1]The CAPI framework was presented at the European Workshop on Reinforcement Learning (no proceedings) [1] and the Multidisciplinary Conference on Reinforcement Learning and Decision Making (extended abstract) [2]. This version includes the proofs and a significantly more detailed discussion of the results. An extended version, including experimental results, is available [3].

Another theoretical contribution of this work is a new *error propagation* result that shows that the errors at later iterations of CAPI play a more important role on the performance of the resulting policy.

II. Background and Notation

We consider a *finite-action discounted MDP* $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{X}$ is a measurable state space, $\mathcal{A}$ is a finite set of actions, $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R})$ is the reward kernel (with expected reward uniformly bounded by $R_{\max}$), and $\gamma \in [0, 1)$ is a discount factor. We use rather standard notations and definitions (see e.g., [3, 4]): $\pi : \mathcal{X} \to \mathcal{A}$ is a (deterministic Markov stationary) policy, V^π and Q^π are its value and action-value functions, and V^* and Q^* are the optimal value and action-value functions (bounded by $Q_{\max}$). A policy π is *greedy* w.r.t. an action-value function Q, denoted by $\pi = \hat{\pi}(\cdot; Q)$, if $\pi(x) = \operatorname{argmax}_{a \in \mathcal{A}} Q(x, a)$ holds for all $x \in \mathcal{X}$ (if there exist multiple maximizers, one of them is chosen in an arbitrary deterministic manner).

Our theoretical analysis will rely on the notion of action-gap regularity of an MDP [22], which characterizes the complexity of a control problem. For simplicity, we define and analyze the two-action case, but the CAPI framework naturally accommodates MDPs with more actions, as we explain below.

Consider an MDP with two actions. For any $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the action-gap function is defined as $\mathbf{g}_Q(x) \triangleq |Q(x, 1) - Q(x, 2)|$ for all $x \in \mathcal{X}$. To understand why the action-gap function is informative, suppose that we have an estimate $\hat{Q}^\pi$ of Q^π and we want to perform policy improvement based on $\hat{Q}^\pi$. The greedy policy w.r.t. $\hat{Q}^\pi$, i.e., $\hat{\pi}(\cdot; \hat{Q}^\pi)$, should ideally be close to the greedy policy w.r.t. Q^π, i.e., $\hat{\pi}(\cdot; Q^\pi)$. If the action-gap $\mathbf{g}_{Q^\pi}(x)$ is large for some state x, the regret of choosing an action different from $\hat{\pi}(x; Q^\pi)$, roughly speaking, is large; however, confusing the best action with the other one is also less likely. If the action-gap is small, a confusion is more likely to arise, but the regret stemming from the wrong choice will be small. To characterize how difficult a problem is, we need to summarize the behaviour of the action-gap function over the entire state space. This is done in the following assumption.

Assumption A1 (Action-Gap). For a fixed MDP $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ with $|\mathcal{A}| = 2$ and a fixed distribution over states $\nu \in \mathcal{M}(\mathcal{X})$, there exist constants $c_g > 0$ and $\zeta \geq 0$ such that for any $\pi \in \Pi$ and all $\varepsilon > 0$, we have $\mathbb{P}_\nu(0 < \mathbf{g}_{Q^\pi}(X) \leq \varepsilon) \triangleq \int_{\mathcal{X}} \mathbb{I}\{0 < \mathbf{g}_{Q^\pi}(x) \leq \varepsilon\} \, \mathrm{d}\nu(x) \leq c_g \, \varepsilon^\zeta$.

The value of ζ controls the distribution of the action-gap $\mathbf{g}_{Q^\pi}(X)$. A large value of ζ indicates that the probability of $Q^\pi(X, 1)$ being very close to $Q^\pi(X, 2)$ is small. This implies that the estimate $\hat{Q}^\pi$ can be quite inaccurate in a large subset of the state space (measured according to ν), but $\hat{\pi}(\cdot; \hat{Q}^\pi)$ would still be the same as $\hat{\pi}(\cdot; Q^\pi)$. Note that any MDP satisfies the inequality when $\zeta = 0$ and $c_g = 1$, so the class of MDPs satisfying this property is not restricted in any way.

Finally, the L_∞-norm on $\mathcal{X} \times \mathcal{A}$ is defined as $\|Q\|_\infty \triangleq \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |Q(x, a)|$. We also use a definition of supremum norm that holds only on a set of points from $\mathcal{X}$. Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$; then, $\|Q\|_{\infty, \mathcal{D}_n} \triangleq \max_{x \in \mathcal{D}_n, a \in \mathcal{A}} |Q(x, a)|$.

Algorithm CAPI(Π, ν, K)

Input: Policy space Π, State distribution ν, Number of iterations K
Initialize: Let $\pi_{(0)} \in \Pi$ be an arbitrary policy
for $k = 0, 1, \ldots, K - 1$ **do**
  Construct a dataset $\mathcal{D}_n^{(k)} = \{X_i\}_{i=1}^n$, $X_i \overset{\text{i.i.d.}}{\sim} \nu$
  $\hat{Q}^{\pi_k} \leftarrow \text{PolicyEval}(\pi_k)$
  $\pi_{k+1} \leftarrow \operatorname{argmin}_{\pi \in \Pi} \hat{L}_n^{\pi_k}(\pi)$ (action-gap-weighted classification)
end for

Fig. 1. CAPI pseudocode

III. CAPI Framework

CAPI is an approximate policy iteration framework that takes a policy space Π, a distribution over states $\nu \in \mathcal{M}(\mathcal{X})$, and the number of iterations K as inputs, and returns a policy whose performance should be close to the best policy in Π (Figure 1). PolicyEval can be any algorithm that computes an estimate $\hat{Q}^\pi$ of Q^π, including all policy evaluation methods mentioned in the Introduction.

Exploiting the intuition given by the action-gap phenomenon [22], which entails that when $\mathbf{g}_{Q^\pi}(x)$ is large at some state x, the regret of choosing an action different from $\hat{\pi}(x; Q^\pi)$ is also large, the approximate policy improvement step of CAPI at each iteration k is performed by minimizing the following action-gap-weighted empirical loss function in policy space Π:

$$\hat{L}_n^{\pi_k}(\pi) \triangleq \int_{\mathcal{X}} \mathbf{g}_{\hat{Q}^{\pi_k}}(x) \mathbb{I}\{\pi(x) \neq \operatorname*{argmax}_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(x, a)\} \, \mathrm{d}\nu_n \quad (1)$$
$$= \sum_{X_i \in \mathcal{D}_n^{(k)}} \mathbf{g}_{\hat{Q}^{\pi_k}}(X_i) \mathbb{I}\{\pi(X_i) \neq \operatorname*{argmax}_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(X_i, a)\},$$

where ν_n is the empirical distribution induced by the samples in $\mathcal{D}_n^{(k)} = \{X_i\}_{i=1}^n$ with $X_i \sim \nu$, i.e., $\nu_n = \frac{1}{n} \sum_{X_i \in \mathcal{D}_n^{(k)}} \delta_{X_i}$ with δ_{X_i} being a point mass at X_i for $i = 1, \ldots, n$. This loss function emphasizes states in which the regret of choosing a non-greedy action is large. The policy improvement step of CAPI is defined by

$$\pi_{k+1} \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \hat{L}_n^{\pi_k}(\pi) \quad (2)$$

Policy π_{k+1} is the projection of the greedy policy $\hat{\pi}(\cdot; \hat{Q}^{\pi_k})$, defined only at points $\mathcal{D}_n^{(k)}$, onto policy space Π when the distance measure is weighted according to the estimated action-gap function $\mathbf{g}_{\hat{Q}^{\pi_k}}$. This should be contrasted with the conventional classification-based approaches [13], which use a uniform weight for all states, i.e., they minimize $\int_{\mathcal{X}} \mathbb{I}\{\pi(x) \neq \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(x, a)\} \mathrm{d}\nu_n$. Note that the loss (1) is also used by [15, 20]. In [3], we discuss why uniformly weighted loss might lead to a bad choice of policies and provide some empirical evidence too.

The flexibility in the choice of policy space Π and PolicyEval allows benefitting from regularities of both policy and value function. The policy space can be a parametric function space, which is described by a fixed finite number

of parameters, or a nonparametric space, which grows with data [18, 19]. Some examples are described by [3]. The flexibility in the choice of PolicyEval enables CAPI to exploit regularities of the value function, such as smoothness, which is impossible with a rollout-based estimator. The optimal choices for PolicyEval and Π are problem-dependent and should ideally be determined by a model selection method [23].

The dataset used by PolicyEval to generate $\hat{Q}^{\pi_k}$, in general, is different from $\mathcal{D}_n^{(k)}$ used in (1). In practice, however, one might use the same dataset for both. It is also possible to change the sampling distribution ν at each iteration, e.g., similar to [24]. Reusing the same dataset or changing the sampling distribution is not analyzed here.

To extend the current loss function to problems with $|\mathcal{A}| > 2$, one can define the action-gap function as $\mathbf{g}_Q(x,a) \triangleq \max_{a' \in \mathcal{A}} Q(x,a') - Q(x,a)$. The empirical loss function would be $\hat{L}_n^{\pi_k}(\pi) \triangleq \int_{\mathcal{X}} \mathbf{g}_{\hat{Q}^{\pi_k}}(x,\pi(x)) \mathbb{I}\{\pi(x) \neq \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}^{\pi_k}(x,a)\} \, \mathrm{d}\nu_n$. Our theoretical analysis, however, does not cover this case.

Since the loss function (1) is non-convex, solving (2) is computationally difficult for some policy spaces. But for local methods such as action-gap-weighted K-Nearest Neighbour or decision trees, one can get simple and computationally efficient rules [3]. Another possibility is to relax the non-convex loss with a convex surrogate such as action-gap-weighted hinge or exponential loss.

IV. Theoretical Analysis

In this section we analyze the theoretical properties of CAPI-style algorithms and provide an upper bound on the *performance loss* (or *regret*) of the resulting policy π_K. The performance loss of a policy π is the expected difference between the value of the optimal policy π^* and the value of π when the initial state distribution is $\rho \in \mathcal{M}(\mathcal{X})$, i.e.,

$$\mathrm{Loss}(\pi;\rho) \triangleq \int_{\mathcal{X}} (V^*(x) - V^{\pi}(x)) \, \mathrm{d}\rho(x).$$

The choice of ρ enables the user to specify the relative importance of different states.

The analysis has two main steps. First, in Section IV-A we study the behaviour of one iteration of the algorithm and provide an error bound on the expected loss $L^{\pi_k}(\pi_{k+1}) \triangleq \int_{\mathcal{X}} g_{Q^{\pi_k}}(x) \mathbb{I}\{\pi_{k+1}(x) \neq \operatorname{argmax}_{a \in \mathcal{A}} Q^{\pi_k}(x,a)\} \, \mathrm{d}\nu$, as a function of the number of samples in $\mathcal{D}_n^{(k)}$, the quality of the estimate $\hat{Q}^{\pi_k}$, the complexity of Π, and the policy approximation error. In Section IV-B, we analyze how the loss sequence $(L^{\pi_k}(\pi_{k+1}))_{k=0}^{K-1}$ affects $\mathrm{Loss}(\pi_K;\rho)$.

A. Approximate Policy Improvement Error

Policy π_k depends on data used in earlier iterations, but is independent of $\mathcal{D}_n^{(k)}$, so we will work on the probability space conditioned on $\mathcal{D}_n^{(0)}, \ldots, \mathcal{D}_n^{(k-1)}$. To avoid clutter, we omit the conditional probability symbol and the dependence of the loss function, policy, and dataset on the iteration number. In the rest of this section, π' refers to a $\sigma(\mathcal{D}_n^{(0)}, \ldots, \mathcal{D}_n^{(k-1)})$-measurable policy and is independent of $\mathcal{D}_n$, which denotes a set of n independent and identically distributed (i.i.d.) samples from the distribution $\nu \in \mathcal{M}(\mathcal{X})$. We also assume that we have a $\mathcal{D}_n$-independent approximation $\hat{Q}^{\pi'}$ of the action-value function $Q^{\pi'}$.

For any $\pi \in \Pi$, we define two pointwise loss functions: $l^{\pi'}(\pi)(x) = \mathbf{g}_{Q^{\pi'}}(x) \mathbb{I}\{\pi(x) \neq \operatorname{argmax}_{a \in \mathcal{A}} Q^{\pi'}(x,a)\}$ and $\hat{l}^{\pi'}(\pi)(x) = \mathbf{g}_{\hat{Q}^{\pi'}}(x) \mathbb{I}\{\pi(x) \neq \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}^{\pi'}(x,a)\}$. Note that $l^{\pi'}(\pi)$ is defined as a function of $Q^{\pi'}$, which is not accessible to the algorithm. On the other hand, $\hat{l}^{\pi'}(\pi)$ is defined as a function of $\hat{Q}^{\pi'}$, which is available to the algorithm. The latter pointwise loss is a distorted version of the former. To simplify the notation, we may use $l(\pi)$ and $\hat{l}(\pi)$ to refer to $l^{\pi'}(\pi)$ and $\hat{l}^{\pi'}(\pi)$, respectively.

For a function $l : \mathcal{X} \to \mathbb{R}$, let $\mathbf{P}_n l = \frac{1}{n} \sum_{i=1}^n l(X_i)$ and $\mathbf{P} l = \mathbb{E}[l(X)]$, where $X, X_i \overset{\text{i.i.d.}}{\sim} \nu$ and X_is are from $\mathcal{D}_n$. Now we can define the expected loss $L(\pi) = \mathbf{P} l(\pi)$ and the empirical loss $L_n(\pi) = \mathbf{P}_n l(\pi)$ (both w.r.t. the true action-value function $Q^{\pi'}$) and the distorted empirical loss $\hat{L}_n(\pi) = \mathbf{P}_n \hat{l}(\pi)$ (w.r.t. the estimate $\hat{Q}^{\pi'}$). Given $\mathcal{D}_n$ and $\hat{Q}^{\pi'}$, let

$$\hat{\pi}_n \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \hat{L}_n(\pi), \tag{3}$$

(cf. (2)). Here and in the rest of the paper we make the standard assumption that the minimum in (3) exists.

To study the behaviour of $L(\hat{\pi}_n)$, we need to take care of two main issues. First we should relate the empirical loss of the minimizer of the distorted empirical loss $\hat{L}_n$, that is $L_n(\hat{\pi}_n)$, to the (unavailable) minimum of the empirical loss, $\min_{\pi \in \Pi} L_n(\pi)$ (Lemma 3 in Appendix A). We also should relate the expected loss $L(\hat{\pi}_n)$ to the empirical loss $L_n(\hat{\pi}_n)$. Making this relation requires define a notion of complexity (or capacity) of policy space Π. Among common choices in the machine learning/statistics literature (such as VC-dimension, metric entropy, etc., see e.g., [18]), we use localized Rademacher complexity since it has favourable properties that often lead to tight upper bounds [21]. The use of localized Rademacher complexity to analyze an RL/ADP algorithm is a novel aspect of this work.

Let $\sigma_1, \ldots, \sigma_n$ be independent random variables with $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$. For a function space $\mathcal{G} : \mathcal{X} \to \mathbb{R}$, define $R_n \mathcal{G} = \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i)$ with $X_i \sim \nu$. The Rademacher complexity (or average) of $\mathcal{G}$ is $\mathbb{E}[R_n \mathcal{G}]$, in which the expectation is w.r.t. both σ and X_i [21]. One can interpret the Rademacher complexity as a measure that quantifies the extent that a function from $\mathcal{G}$ can fit a noise sequence of length n.

In order to benefit from the localized version of Rademacher complexity, we need to define a sub-root function. A non-negative and non-decreasing function $\Psi : [0,\infty) \to [0,\infty)$ is called sub-root if $r \mapsto \frac{\Psi(r)}{\sqrt{r}}$ is non-increasing for $r > 0$ [21]. The following theorem is the main result of this subsection.

Theorem 1. *Fix a policy π' and assume that $\mathcal{D}_n$ consists of n i.i.d. samples drawn from distribution ν and $\hat{Q}^{\pi'}$ is independent of $\mathcal{D}_n$. Let $\hat{\pi}_n$ be defined by (3). Suppose that Assumption A1 holds with a particular value of (ζ, c_g). Let Ψ be a sub-root function with a fixed point of r^* such that for $r \geq r^*$,*

$$\Psi(r) \geq 2 Q_{max} \mathbb{E}\left[R_n \left\{ l^{\pi'}(\pi) \, : \, \pi \in \Pi, \mathbf{P}[l^{\pi'}(\pi)]^2 \leq r \right\} \right]. \tag{4}$$

Then there exist $c_1, c_2, c_3 > 0$, which are independent of n, $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$, and r^*, so that for any $0 < \delta < 1$, $L(\hat{\pi}_n) \leq 12 \inf_{\pi\in\Pi} L(\pi) + c_1 r^* + c_2 \|\hat{Q}^{\pi'} - Q^{\pi'}\|^{1+\zeta}_{\infty,\mathcal{D}_n} + c_3 \frac{\ln(1/\delta)}{n}$, with probability at least $1-\delta$.

The upper bound has three important terms. The $\inf_{\pi\in\Pi} L(\pi)$ term is the *policy approximation error*. For a rich enough policy space, this term can be zero.

The second important term is the *estimation error* of the classifier, which is mainly determined by the behaviour of the fixed point r^* of (4). Condition (4) implies that the estimation error is not determined by the global complexity of the function space, but by its complexity in the neighbourhood of the minimizer $\operatorname{argmin}_{\pi\in\Pi} L^{\pi'}(\pi)$. If Π is a space with VC-dimension d, one can show that r^* behaves as $O(d\log(n)/n)$ (cf. proof of Corollary 3.7 of [21]). This rate is considerably faster than the $O(\sqrt{d/n})$ behaviour of the estimation error term in the result of [15, 20]. Similar local Rademacher complexity results exist for nonparametric spaces.

The last important term is $\|\hat{Q}^{\pi'} - Q^{\pi'}\|^{1+\zeta}_{\infty,\mathcal{D}_n}$, whose size depends on 1) the quality of $\hat{Q}^{\pi'}$ at the points in $\mathcal{D}_n$, and 2) the action-gap regularity of the problem, characterized by ζ. When $\zeta > 0$, the policy evaluation error $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$ is dampened and the rate improves geometrically. The analysis of [15, 20] does not benefit from this regularity. Finally note that $\hat{Q}^{\pi'}$ is often estimated using data, so $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$ would be random. As we assumed that $\hat{Q}^{\pi'}$ is independent of $\mathcal{D}_n$, the source of randomness of $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$ in the upper bound is different from $\mathcal{D}_n$. The high probability guarantee of the theorem is on the randomness due to $\mathcal{D}_n$.

B. Performance Loss of CAPI

Here we state the main result of this paper, which upper bounds $\mathrm{Loss}(\pi_K;\rho)$ as a function of $L^{\pi_k}(\pi_{k+1})$ at iterations $k = 0,\dots,K-1$ and some other properties of the MDP and policy space Π. First we introduce two definitions.

Definition 1 (Worst-Case Greedy Policy Error)**.** *For a policy space Π, the worst-case greedy policy error is $d(\Pi) = \sup_{\pi'\in\Pi} \inf_{\pi\in\Pi} L^{\pi'}(\pi)$.*

Definition 2 (Concentrability Coefficient)**.** *Given $\rho,\nu \in \mathcal{M}(\mathcal{X})$, a policy π, and two integers $m_1, m_2 \geq 0$, let $\rho(\mathcal{P}^*)^{m_1}(\mathcal{P}^{\pi})^{m_2}$ denote the future-state distribution obtained when the first state is drawn from ρ, then the optimal policy π^* is followed for m_1 steps and policy π for m_2 steps. Denote the supremum of the Radon-Nikodym derivative of the resulting distribution w.r.t. ν by $c_{\rho,\nu}(m_1;m_2;\pi) \triangleq \|\frac{\mathrm{d}(\rho(\mathcal{P}^*)^{m_1}(\mathcal{P}^{\pi})^{m_2})}{\mathrm{d}\nu}\|_\infty$. If $\rho(\mathcal{P}^*)^{m_1}(\mathcal{P}^{\pi})^{m_2}$ is not absolutely continuous w.r.t. ν, then $c(m_1,m_2;\pi) = \infty$. For an integer $K \geq 1$ and a real $s \in [0,1]$, define $C_{\rho,\nu}(K,s) \triangleq \frac{1-\gamma}{2}\sum_{k=0}^{K-1}\gamma^{(1-s)k}\sum_{m\geq 0}\gamma^m \sup_{\pi'\in\Pi} c_{\rho,\nu}(k,m;\pi')$.*

The intuition behind Definition 1 is discussed by [3]. For a discussion of Definition 2 and similar concentrability coefficients, refer to [25, 26]. We are now ready to state the main result.

Theorem 2. *Consider the sequence of independent datasets $(\mathcal{D}_n^{(k)})_{k=0}^{K-1}$, each with n i.i.d. samples drawn from $\nu \in \mathcal{M}(\mathcal{X})$. Let $\pi_0 \in \Pi$ be a fixed initial policy and $(\pi_k)_{k=1}^K$ be a sequence of policies obtained by solving (1), using estimate $\hat{Q}^{\pi_k}$ of Q^{π_k}. Suppose that $\hat{Q}^{\pi_k}$ is independent of $\mathcal{D}_n^{(k)}$ and Assumption A1 holds with a particular value of (ζ, c_g). Let r^* be the fixed point of a sub-root function Ψ such that for any $\pi' \in \Pi$ and $r \geq r^*$, $\Psi(r) \geq 2Q_{max}\mathbb{E}\left[R_n\left\{ l^{\pi'}(\pi) : \pi\in\Pi, \mathbf{P}[l^{\pi'}(\pi)]^2 \leq r\right\}\right]$. Then there exist constants $c_1, c_2, c_3 > 0$ such that for any $0 < \delta < 1$, for $\mathcal{E}(s)$ $(0 \leq s \leq 1)$ defined as $\mathcal{E}(s) \triangleq 12d(\Pi) + c_1 r^* + c_2 \max_{0\leq k\leq K-1}\left[\gamma^{(K-k-1)s}\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|^{1+\zeta}_{\infty,\mathcal{D}_n^{(k)}}\right] + c_3\frac{\ln(K/\delta)}{n}$, we have with probability at least $1-\delta$, $\mathrm{Loss}(\pi_K;\rho) \leq \frac{2}{1-\gamma}\left[\inf_{s\in[0,1]} C_{\rho,\nu}(K,s)\,\mathcal{E}(s) + \gamma^K R_{max}\right]$.*

All discussions after Theorem 1 regarding the policy approximation error, the estimation error, and the role of the action-gap regularity apply here too. Moreover, the new error propagation result used in the proof is an improvement over the previous results [15, 20]. The result indicates that the error $\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_{\infty,\mathcal{D}_n}$ is weighted proportional to $\gamma^{(K-k-1)s}$, i.e., the errors at earlier iterations are geometrically discounted.

V. Conclusion and Future Work

We proposed CAPI, a general family of algorithms that exploits regularities of both the value function and the policy. CAPI uses any policy evaluation method, defines an action-gap-weighted loss function, and finds the policy minimizing this loss from a desired policy space. We provided an error upper bound that is tighter than existing results and applies to general policy evaluation algorithms and nonparametric policy spaces. The experiments reported in [3] show that CAPI using a powerful PolicyEval outperforms a rollout-based classification-based algorithm as well as a state-of-the-art purely value-based approach.

Analyzing CAPI with a convex surrogate loss is an interesting question, as is extending CAPI to continuous action spaces. The sampling distribution ν can have a significant effect on the performance; how to choose it is an open question.

Appendix
Proofs

Lemma 3 (Loss Distortion Lemma)**.** *Fix a policy π'. Suppose that $\hat{Q}^{\pi'}$ is an approximation of the action-value function $Q^{\pi'}$. Given the dataset $\mathcal{D}_n$, let $\hat{\pi}_n$ be defined as (3) and define $\pi_n^* \leftarrow \operatorname{argmin}_{\pi\in\Pi} L_n(\pi)$. Let Assumption A1 hold. There exist finite $c_1, c_2 > 0$, which depend only on ζ, c_g, and Q_{max}, such that for any $z > 0$, we have $L_n(\hat{\pi}_n) \leq 3L_n(\pi_n^*) + c_1\|\hat{Q}^{\pi'} - Q^{\pi'}\|^{1+\zeta}_{\infty,\mathcal{D}_n} + c_2\frac{z}{n}$, with probability at least $1-e^{-z}$.*

In the proofs, $c_1, c_2, \dots$ are constants whose values may change from line to line – unless specified otherwise.

Proof of Lemma 3. Let $\varepsilon = \|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$ and define the set $A_\varepsilon = \{x : 0 < \mathbf{g}_{Q^{\pi'}}(x) \leq 4\varepsilon\}$. Denote $p = \mathbb{P}_\nu(X \in A_\varepsilon)$. For any $z > 0$, Bernstein inequality (Theorem 6.12 of [27]) shows that $\mathbb{P}_{\nu_n}(X \in A_\varepsilon) - \mathbb{P}_\nu(X \in A_\varepsilon) \leq \sqrt{\frac{2p(1-p)z}{n}} + \frac{2z}{3n}$ with probability at least $1-e^{-z}$. By the arithmetic mean–geometric mean inequality $\sqrt{[p(1-p)]\frac{2z}{n}} \leq \frac{p(1-p)}{2} + \frac{2z}{2n} \leq$

$\frac{p}{2}+\frac{z}{n}$, so we get

$$\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right)\leq\frac{3}{2}\mathbb{P}_{\nu}\left(X\in A_\varepsilon\right)+\frac{5z}{3n} \tag{5}$$

with probability at least $1-e^{-z}$. From now on, we focus on the event that this inequality holds.

Define the new auxiliary loss $\tilde{L}_n(\pi)=\int_{\mathcal{X}}\mathbf{g}_{Q^{\pi'}}(x)\mathbb{I}\{\pi(x)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\mathrm{d}\nu_n$. Notice that unlike $\hat{L}_n(\pi)$, it uses the weighting function $\mathbf{g}_{Q^{\pi'}}$ (instead of $\mathbf{g}_{\hat{Q}^{\pi'}}$). In the following, for any π, we first relate $L_n(\pi)$ to $\tilde{L}_n(\pi)$, and then relate $\tilde{L}_n(\pi)$ to $\hat{L}_n(\pi)$.

Upper bounding $|L_n(\pi)-\tilde{L}_n(\pi)|$. For any π, $|L_n(\pi)-\tilde{L}_n(\pi)|=|\int_{\mathcal{X}}\mathbf{g}_{Q^{\pi'}}(x)[\mathbb{I}\{\pi(x)\neq\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\}-\mathbb{I}\{\pi(x)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}]\mathrm{d}\nu_n|\leq\int_{\mathcal{X}}\mathbf{g}_{Q^{\pi'}}(x)\cdot\mathbb{I}\{\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\mathrm{d}\nu_n=\int_{A_\varepsilon}\mathbf{g}_{Q^{\pi'}}(x)\mathbb{I}\{\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\cdot\mathrm{d}\nu_n+\int_{A_\varepsilon^c}\mathbf{g}_{Q^{\pi'}}(x)\times\mathbb{I}\{\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\mathrm{d}\nu_n$.

Whenever $|\hat{Q}^{\pi'}(x,a)-Q^{\pi'}(x,a)|<\frac{1}{2}\mathbf{g}_{Q^{\pi'}}(x)$ (for $x\in\mathcal{D}_n$ and $a\in\{1,2\}$), the maximizer action is the same. So on the set A_ε^c, where $\mathbf{g}_{Q^{\pi'}}(x)>4\varepsilon\geq4|\hat{Q}^{\pi'}(x,a)-Q^{\pi'}(x,a)|$, the value of $\mathbb{I}\{\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\neq\operatorname{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}$ is always zero. Thus for any $z>0$, we have

$$\begin{aligned}
&\left|L_n(\pi)-\tilde{L}_n(\pi)\right|\leq(4\varepsilon)\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right)\leq\\
&4\varepsilon\left[\frac{3}{2}\mathbb{P}_{\nu}\left(X\in A_\varepsilon\right)+\frac{5z}{3n}\right]\leq\\
&6\times2^{2\zeta}\left\|\hat{Q}^{\pi'}-Q^{\pi'}\right\|_{\infty,\mathcal{D}_n}^{1+\zeta}+\frac{20}{3}\left\|\hat{Q}^{\pi'}-Q^{\pi'}\right\|_{\infty,\mathcal{D}_n}\frac{z}{n}\leq\\
&c_1(\zeta)\left\|\hat{Q}^{\pi'}-Q^{\pi'}\right\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_2(Q_{\max})\frac{z}{n}.
\end{aligned}\tag{6}$$

Here we used (5) in the second inequality, Assumption A1 in the third inequality, and $\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}\leq2Q_{\max}$ in the last one.

Relation of $\hat{L}_n(\pi)$ to $\tilde{L}_n(\pi)$. First note that $|\mathbf{g}_{\hat{Q}^{\pi'}}(x)-\mathbf{g}_{Q^{\pi'}}(x)|\leq2\varepsilon$ (for all $x\in\mathcal{D}_n$). We also have $\max_{x\in A_\varepsilon^c\cap\mathcal{D}_n}\frac{\mathbf{g}_{\hat{Q}^{\pi'}}(x)-\mathbf{g}_{Q^{\pi'}}(x)}{\mathbf{g}_{Q^{\pi'}}(x)}\leq\frac{2\varepsilon}{4\varepsilon}=\frac{1}{2}$ and $\max_{x\in A_\varepsilon^c\cap\mathcal{D}_n}\frac{\mathbf{g}_{Q^{\pi'}}(x)-\mathbf{g}_{\hat{Q}^{\pi'}}(x)}{\mathbf{g}_{\hat{Q}^{\pi'}}(x)}\leq\frac{2\varepsilon}{2\varepsilon}=1$. Thus,

$$\begin{aligned}
&\hat{L}_n(\pi)-\tilde{L}_n(\pi)=\\
&\int_{A_\varepsilon}(\mathbf{g}_{\hat{Q}^{\pi'}}(x)-\mathbf{g}_{Q^{\pi'}}(x))\mathbb{I}\{\pi(x)\neq\operatorname*{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\mathrm{d}\nu_n+\\
&\int_{A_\varepsilon^c}\frac{\mathbf{g}_{\hat{Q}^{\pi'}}(x)-\mathbf{g}_{Q^{\pi'}}(x)}{\mathbf{g}_{Q^{\pi'}}(x)}\mathbf{g}_{Q^{\pi'}}(x)\mathbb{I}\{\pi(x)\neq\operatorname*{argmax}_{a\in\mathcal{A}}\hat{Q}^{\pi'}(x,a)\}\mathrm{d}\nu_n\\
&\leq(2\varepsilon)\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right)+\frac{1}{2}\tilde{L}_n(\pi).
\end{aligned}$$

After re-arranging, we get

$$\hat{L}_n(\pi)\leq\frac{3}{2}\tilde{L}_n(\pi)+2\varepsilon\,\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right). \tag{7}$$

Likewise, by writing $\mathbf{g}_{Q^{\pi'}}(x)-\mathbf{g}_{\hat{Q}^{\pi'}}(x)$ as $\frac{\mathbf{g}_{Q^{\pi'}}(x)-\mathbf{g}_{\hat{Q}^{\pi'}}(x)}{\mathbf{g}_{\hat{Q}^{\pi'}}(x)}\mathbf{g}_{\hat{Q}^{\pi'}}(x)$ and doing a similar decomposition of the state space into A_ε and A_ε^c, we get

$$\hat{L}_n(\pi)\geq\frac{1}{2}\tilde{L}_n(\pi)-\varepsilon\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right). \tag{8}$$

We use the optimizer property of $\hat{\pi}_n$ (which implies that $\hat{L}_n(\hat{\pi}_n)\leq\hat{L}_n(\pi_n^*)$), apply (7), and finally use inequalities (6) and (5) to get $\hat{L}_n(\hat{\pi}_n)\leq\hat{L}_n(\pi_n^*)\leq\frac{3}{2}\tilde{L}_n(\pi_n^*)+(2\varepsilon)\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right)\leq\frac{3}{2}[L_n(\pi_n^*)+c_1\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_2\frac{z}{n}]+(2\varepsilon)[\frac{3}{2}\mathbb{P}_{\nu}\left(X\in A_\varepsilon\right)+\frac{5}{3}\frac{z}{n}]$. From (8) and by applying (6), we also have $\hat{L}_n(\hat{\pi}_n)\geq\frac{1}{2}\tilde{L}_n(\hat{\pi}_n)-\varepsilon\mathbb{P}_{\nu_n}\left(X\in A_\varepsilon\right)\geq\frac{1}{2}[L_n(\hat{\pi}_n)-c_1\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}-c_2\frac{z}{n}]-\varepsilon[\frac{3}{2}\mathbb{P}_{\nu}\left(X\in A_\varepsilon\right)+\frac{5z}{3n}]$. These two inequalities imply that $L_n(\hat{\pi}_n)\leq3L_n(\pi_n^*)+c_1\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_2\frac{z}{n}$ in the event that (5) holds, which has probability at least $1-e^{-z}$. □

Proof of Theorem 1. We use Theorem 3.3 by [21]. For function $l(\pi)(x)=\mathbf{g}_{Q^{\pi'}}(x)\mathbb{I}\{\pi(x)\neq\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(x,a)\}$, we have $\operatorname{Var}\left[l(\pi)(X)\right]\leq\mathbb{E}\left[|\mathbf{g}_{Q^{\pi'}}(X)\mathbb{I}\{\pi(X)\neq\operatorname{argmax}_{a\in\mathcal{A}}Q^{\pi'}(X,a)\}|^2\right]\leq2Q_{\max}\mathbb{E}\left[l(\pi)(X)\right]$, so the variance condition of that theorem is satisfied. If we have a function Ψ as defined in (4), the theorem states that there exist $c_1,c_2>0$ such that for any $z>0$ and any $\pi\in\Pi$ (including $\hat{\pi}_n\in\Pi$),

$$L(\pi)=\mathbf{P}l(\pi)\leq2P_nl(\pi)+c_1r^*+c_2\frac{z}{n}, \tag{9}$$

with probability at least $1-e^{-z}$ (c_1 can be chosen as $704/Q_{\max}$ and c_2 can be chosen as $126\,Q_{\max}$).

Let $\pi_\Pi^*\leftarrow\operatorname{argmin}_{\pi\in\Pi}L(\pi)$ be the minimizer of the expected loss in policy space Π. Consider (9) with the choice of $\pi=\hat{\pi}_n$, and add and subtract $6P_nl(\pi_\Pi^*)$ and $6\mathbf{P}l(\pi_\Pi^*)$ and then use Lemma 3. With probability at least $1-2e^{-z}$, we get $L(\hat{\pi}_n)\leq2\mathbf{P}_nl(\hat{\pi}_n)-6\left[\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}_nl(\pi_\Pi^*)\right]-6\left[\mathbf{P}l(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)\right]+c_1r^*+c_2\frac{z}{n}\leq6\left[\mathbf{P}_nl(\pi_n^*)-\mathbf{P}_nl(\pi_\Pi^*)\right]+6\left[\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)\right]+6\mathbf{P}l(\pi_\Pi^*)+c_1r^*+c_2\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_3\frac{z}{n}\leq6\left[\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)\right]+6\mathbf{P}l(\pi_\Pi^*)+c_1r^*+c_2\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_3\frac{z}{n}$, where in the last inequality we used the minimizing property of π_n^*, i.e., $\mathbf{P}_nl(\pi_n^*)-\mathbf{P}_nl(\pi_\Pi^*)\leq0$. Here c_2 can be chosen as $36\times2^{2\zeta}$.

To upper bound $\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)$, we apply Bernstein inequality to get that for any $z>0$, $\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)\leq\sqrt{\frac{2\operatorname{Var}\left[l(\pi_\Pi^*)\right]z}{n}}+\frac{4Q_{\max}z}{3n}$, with probability at least $1-e^{-z}$. Since $\operatorname{Var}\left[l(\pi_\Pi^*)\right]\leq2Q_{\max}\mathbf{P}l(\pi_\Pi^*)$ (as shown above), by the application of arithmetic mean–geometric mean inequality we obtain $\mathbf{P}_nl(\pi_\Pi^*)-\mathbf{P}l(\pi_\Pi^*)\leq\mathbf{P}l(\pi_\Pi^*)+\frac{7Q_{\max}z}{3n}$ with the same probability. This and the inequality in the previous paragraph result in $L(\hat{\pi}_n)\leq12\mathbf{P}l(\pi_\Pi^*)+c_1r^*+c_2\|\hat{Q}^{\pi'}-Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}+c_3\frac{z}{n}$, with probability at least $1-3e^{-z}$ as desired. □

Proof of Theorem 2. It is shown by [15] that $V^*-V^{\pi_K}\leq\sum_{k=0}^{K-1}\gamma^{K-k-1}(\mathcal{P}^{\pi^*})^{K-k-1}\sum_{m\geq0}\gamma^m(\mathcal{P}^{\pi_{k+1}})^ml^{\pi_k}(\pi_{k+1})+(\gamma\mathcal{P}^{\pi^*})^K(V^*-V^{\pi_0})$. We apply ρ to both sides and use the definition of $c_{\rho,\nu}(m_1;m_2;\pi)$ to get $\rho(V^*-V^{\pi_K})\leq\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{m\geq0}\gamma^mc_{\rho,\nu}(K-k-1,m;\pi_{k+1})\nu l^{\pi_k}(\pi_{k+1})+\gamma^K(2Q_{\max})$. Recall that $\nu l^{\pi_k}(\pi_{k+1})=L^{\pi_k}(\pi_{k+1})$. We decompose γ to $\gamma^s\gamma^{(1-s)}$ (for $0\leq s\leq1$) and separate terms involving the concentrability coefficients and those related to $L^{\pi_k}(\pi_{k+1})$. We then have for any $0\leq s\leq1$, $\rho(V^*-V^{\pi_K})\leq\max_{0\leq k\leq K-1}\left\{\gamma^{s(K-k-1)}L^{\pi_k}(\pi_{k+1})\right\}\times\sum_{k'=0}^{K-1}\gamma^{(1-s)k'}\sum_{m\geq0}\gamma^m\sup_{\pi'\in\Pi}c_{\rho,\nu}(k',m;\pi')+$

$\gamma^K(2Q_{\max})$. Taking the infimum w.r.t. s and using the definition of $C_{\rho,\nu}(K)$, we get that $\text{Loss}(\pi_K;\rho) = \rho(V^* - V^{\pi_K}) \leq \frac{2}{1-\gamma}[\inf_{s\in[0,1]} C_{\rho,\nu}(K,s) \max_{0\leq k\leq K-1}[\gamma^{s(K-k-1)} L^{\pi_k}(\pi_{k+1})] + \gamma^K R_{\max}]$.

Fix $0 < \delta < 1$. For each iteration $k = 0, \ldots, K-1$, by invoking Theorem 1 with the confidence parameter δ/K, we get $L^{\pi_k}(\pi_{k+1}) \leq 12 \inf_{\pi\in\Pi} L^{\pi_k}(\pi) + c_1 r^* + c_2 \|\hat{Q}^{\pi_k} - Q^{\pi_k}\|^{1+\zeta}_{\infty,\mathcal{D}_n^k} + c_3 \frac{\ln(K/\delta)}{n}$, which holds with probability at least $1 - \delta/K$. Since $\inf_{\pi\in\Pi} L^{\pi_k}(\pi) \leq d(\Pi)$, the previous set of inequalities alongside the upper bound on $\text{Loss}(\pi_K;\rho)$ imply the desired result. □

We would like to remark that to extend the analysis to $|\mathcal{A}| > 2$, Lemma 3 is the main result that should be modified. The proofs of Theorems 1 and 2 remain intact.

Acknowledgment

We thank the anonymous reviewers for their comments. This work is financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] A-m. Farahmand, D. Precup, and M. Ghavamzadeh. Generalized classification-based approximate policy iteration. In *European Workshop on Reinforcement Learning (EWRL)*, 2012. 1

[2] A-m. Farahmand, D. Precup, A.M.S Barreto, and M. Ghavamzadeh. CAPI: Generalized classification-based approximate policy iteration. In *Multidisciplinary Conference on Reinforcement Learning and Decision Making*, October 2013. 1

[3] A-m. Farahmand, D. Precup, A.M.S Barreto, and M. Ghavamzadeh. Classification-based approximate policy iteration: Experiments and extended discussions. *arXiv e-print: 1407.0449*, 2014. URL http://arxiv.org/abs/1407.0449.

[4] Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010. 1, 2

[5] A-m. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor. Regularized policy iteration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 441–448, 2009. 1

[6] G. Taylor and R. Parr. Kernelized value function approximation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 1017–1024, 2009. 1

[7] A-m. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian Decision Problems. In *American Control Conference (ACC)*, pages 725–730, 2009. 1

[8] M. Ghavamzadeh, A. Lazaric, R. Munos, and M. Hoffman. Finite-sample analysis of Lasso-TD. In *International Conference on Machine Learning (ICML)*, pages 1177–1184, 2011. 1

[9] A-m. Farahmand and D. Precup. Value pursuit iteration. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1

[10] P. Marbach and J.N. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Trans. on Automatic Control*, 46(2):191–209, 2001. 1

[11] X.-R. Cao. A basic formula for online policy gradient algorithms. *IEEE Trans. on Automatic Control*, 50(5): 696–699, 2005. 1

[12] M. Ghavamzadeh and Y. Engel. Bayesian policy gradient algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 457–464, 2007. 1

[13] M.G. Lagoudakis and R. Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *International Conference on Machine Learning (ICML)*, pages 424–431, 2003. 1, 2

[14] A. Fern, S. Yoon, and R. Givan. Approximate policy iteration with a policy language bias: Solving relational Markov Decision Processes. *Journal of Artificial Intelligence Research*, 25:85–118, 2006. 1

[15] A. Lazaric, M. Ghavamzadeh, and R. Munos. Analysis of a classification-based policy iteration algorithm. In *International Conference on Machine Learning (ICML)*, pages 607–614, 2010. 1, 2, 4, 5

[16] M.G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4: 1107–1149, 2003. 1

[17] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. 1

[18] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, 2002. 1, 3

[19] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2007. 1, 3

[20] V. Gabillon, A. Lazaric, M. Ghavamzadeh, and B. Scherrer. Classification-based policy iteration with a critic. In *International Conference on Machine Learning (ICML)*, 2011. 1, 2, 4

[21] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005. 1, 3, 4, 5

[22] A-m. Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. 1, 2

[23] A-m. Farahmand and Cs. Szepesvári. Model selection in reinforcement learning. *Machine Learning Journal*, 85 (3):299–332, 2011. 3

[24] S. Ross, G. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Artifical Intelligence and Statistics (AISTATS)*, April 2011. 3

[25] R. Munos. Performance bounds in L_p norm for approximate value iteration. *SIAM Journal on Control and Optimization*, pages 541–561, 2007. 4

[26] A-m. Farahmand, R. Munos, and Cs. Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2010. 4

[27] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008. 4