Generalized Classification-based Approximate Policy Iteration

Amir-massoud Farahmand and **Doina Precup**
School of Computer Science, McGill University, Montreal, Canada

Mohammad Ghavamzadeh
INRIA Lille - Team SequeL, Lille, France

Editors: Marc Deisenroth, Csaba Szepesvári, Jan Peters

Abstract

Classification-based approximate policy iteration is very useful when the optimal policy is easier to represent and learn than the optimal value function. We theoretically analyze a general algorithm of this type. The analysis extends existing work by allowing the policy evaluation to be performed by any reinforcement learning algorithm, by handling nonparametric representations of policies, and by providing tighter convergence bounds. A small illustration shows that this approach can be faster than purely value-based methods.

Keywords: Reinforcement Learning, Classification-based Algorithms, Convergence Rate, Approximate Policy Iteration

1. Introduction

Solving reinforcement learning (RL) problems with large state spaces can be difficult, unless one exploits some form of regularity or structure of the problem in hand. Recently, several [nonparametric] algorithms have been suggested that benefit from regularities of the *value* function, e.g., Farahmand et al. (2009); Kolter and Ng (2009); Taylor and Parr (2009); Ghavamzadeh et al. (2011). Nevertheless, this is only one type of regularities of RL problems and one may also benefit from the regularities of the *policy* as well. Benefiting from this less-studied type of regularities has been the motivation behind classification-based RL algorithms (e.g., Lagoudakis and Parr 2003b; Fern et al. 2006; Lazaric et al. 2010).

The main idea of classification-based approaches is to get a rough estimate of the value function, find the [noisy] greedy action at several states, and use this information to train a classifier that generalizes to the whole state space while smoothing out the noise in greedy actions. In many problems this approach is helpful because of two main reasons. The first is that at each state even a rough estimate of the value function is often sufficient to separate the best action from the rest. For example, consider an RL problem with only two available actions (e.g., go over Atlantic by ship or by airplane). The precise estimation of the value function may be difficult (e.g., it might be hard to estimate precisely the value of being seasick); however, if the true values of these two actions are different enough, even an inaccurate estimate will suffice to tell which one is better. The second is that good policies are sometimes simpler to represent and learn than good value functions.

Nevertheless, the current classification-based approaches have some drawbacks and cannot benefit from the present regularities very well. Firstly, most previous works use 0/1 classification loss, which does not consider the relative importance of different regions of the state space and may lead to surprisingly bad policies (cf. Section 5). Secondly, existing approaches estimate the action-value function using rollouts, which is not satisfactory because it does not benefit from the possible regularities the value function, and hence the data is not used efficiently.

This paper studies a generic Classification-based Approximate Policy Iteration (CAPI) framework. In standard API, the algorithm iteratively evaluates a policy (i.e., finding the action-value function) and then improves it by computing the greedy policy w.r.t. (with respect to) the most recent value function. In CAPI, however, we fit a policy from a restricted policy space to the greedy policy obtained at sample points. The error function is weighted according to the difference between the value of the greedy action and those of the other actions. This ensures that the resulting policy closely follows the greedy policy in regions of the state space where the difference between the best action and the rest is considerable (so choosing the wrong action leads to a large performance loss) but pays less attention to regions where the value of all actions is almost the same. When the state space is large, the policy evaluation step cannot be done exactly, so the use of function approximation is inevitable. CAPI can use any policy evaluation method including, but not restricted to, rollout-based estimates (as in previous works), LSTD, Fitted Q-Iteration, and their regularized variants. Note that this is a strict generalization of existing algorithms, which become special cases of CAPI.

The main theoretical contribution is the finite sample error analysis of CAPI, which allows general policy evaluation algorithms, handles nonparametric policy spaces, and provides a faster convergence rate than existing results. Using nonparametric policies is a significant extension of the work by Fern et al. (2006), which is limited to finite policy spaces, and of Lazaric et al. (2010) and Gabillon et al. (2011), which are limited to policy spaces with finite VC dimension. We also provide a new error propagation result for classification-based RL algorithms that shows the errors at the later iterations play a more important role in the quality of the final policy. We obtain much faster rates of convergence than existing results (even when one uses rollouts), because we use a concentration inequality that is based on the powerful notion of local Rademacher complexity (Bartlett et al., 2005), which is known to lead to fast rates in the supervised learning scenarios. We also benefit from the action-gap regularity of the problem, introduced by Farahmand (2011), which means that choosing the right action at each state may not require a precise estimate of the action-value function. Whenever this (quite common) regularity is present, the convergence rate of the performance loss is faster than the convergence rate of the action-value function. This work exploits the action-gap regularity in the analysis of classification-based RL.

2. Background and Notation

In this section, we define our notation and summarize necessary definitions. For more information, the reader is referred to Bertsekas and Tsitsiklis (1996); Sutton and Barto (1998); Szepesvári (2010).

For a space Ω, with σ-algebra σ_Ω, $\mathcal{M}(\Omega)$ denotes the set of all probability measures over σ_Ω. $B(\Omega)$ denotes the space of bounded measurable functions w.r.t. σ_Ω and $B(\Omega, L)$ denotes the subset of $B(\Omega)$ with bound $0 < L < \infty$.

A *finite-action discounted MDP* is a 5-tuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{X}$ is a measurable state space, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is a discount factor. Let $r(x, a) = \mathbb{E}[\mathcal{R}(\cdot|x, a)]$, and assume that r is uniformly bounded by $R_{\max}$. For simplicity, we focus on MDPs with two actions, i.e., $|\mathcal{A}| = 2$, but with significant extra work, the analysis can be done for the general case as well. A measurable mapping $\pi : \mathcal{X} \to \mathcal{A}$ is called a deterministic Markov stationary policy, or just *policy* for short. Following a policy π means that at each time step, $A_t = \pi(X_t)$.

A policy π induces the transition probability kernel $P^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$. For a measurable subset S of $\mathcal{X}$, we define $(P^\pi)(S|x) \triangleq \int P(dy|x, \pi(x))\mathbb{I}_{\{y \in S\}}$, in which $\mathbb{I}_{\{\cdot\}}$ is the indicator function. The m-step transition probability kernels $(P^\pi)^m : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ for $m = 2, 3, \cdots$ are inductively defined as $(P^\pi)^m(S|x) \triangleq \int_{\mathcal{X}} P(dy|x, \pi(x))(P^\pi)^{m-1}(S|y, \pi(y))$.

Given a transition probability kernel $P : \mathcal{X} \to \mathcal{M}(\mathcal{X})$, define the right-linear operator $P\cdot : B(\mathcal{X}) \to B(\mathcal{X})$ by $(PV)(x) \triangleq \int_{\mathcal{X}} P(dy|x)V(y)$. For a probability measure $\rho \in \mathcal{M}(\mathcal{X})$ and a measurable subset S of $\mathcal{X}$, define the left-linear operators $\cdot P : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by $(\rho P)(S) = \int \rho(dx)P(dy|x)\mathbb{I}_{\{y \in S\}}$. A typical choice of P is $(P^\pi)^m : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$.

The value function V^π and the action-value function Q^π of a policy π are defined as follows: Let $(R_t; t \geq 1)$ be the sequence of rewards when the Markov chain is started from state X_1 (or state-action (X_1, A_1) for Q^π) drawn from a positive probability distribution over $\mathcal{X}$ ($\mathcal{X} \times \mathcal{A}$) and the agent follows the policy π. Then $V^\pi(x) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \,\middle|\, X_1 = x\right]$ and $Q^\pi(x, a) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \,\middle|\, X_1 = x, A_1 = a\right]$. The value of V^π and Q^π are uniformly bounded by $Q_{\max} = R_{\max}/(1 - \gamma)$, independent of the choice of π.

The *optimal value* and *optimal action-value* functions are defined as $V^*(x) = \sup_\pi V^\pi(x)$ for all $x \in \mathcal{X}$ and $Q^*(x, a) = \sup_\pi Q^\pi(x, a)$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. A policy π^* is *optimal* if $V^{\pi^*} = V^*$. A policy π is *greedy* w.r.t. an action-value function Q, denoted $\pi = \hat{\pi}(\cdot; Q)$, 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). Note that a greedy policy w.r.t. the optimal action-value function Q^* is an optimal policy.

The $L_\infty(\mathcal{X} \times \mathcal{A})$-norm 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 on $\mathcal{X}$. Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ with $X_i \in \mathcal{X}$; then, $\|Q\|_{\infty, \mathcal{D}_n} \triangleq \sup_{x \in \mathcal{D}_n, a \in \mathcal{A}} |Q(x, a)|$.

3. Framework

CAPI is an approximate policy iteration framework that takes a policy space Π and a distribution over states $\nu \in \mathcal{M}(\mathcal{X})$ as input, and returns a policy whose performance should be close to the best policy in Π (Algorithm 1). CAPI starts with an arbitrary policy $\pi_{(0)} \in \Pi$ and at each iteration k, it **1**) constructs a dataset $\mathcal{D}_n^{(k)}$ by drawing n i.i.d. samples from ν, **2**) calculates an estimate of the action-value function of the current policy $\hat{Q}^{\pi_{(k)}}$ (subroutine PolicyEval), and **3**) computes the new policy $\pi_{(k+1)} \leftarrow \operatorname{argmin}_{\pi \in \Pi} \hat{L}_n^{\pi_{(k)}}(\pi)$ by minimizing

Algorithm 1 CAPI(Π, ν, K)

```
**Input:** Policy space $\Pi$, State distribution $\nu$, 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 $\mathcal{D}_n^{(k)} = \{X_i\}_{i=1}^{n},\ X_i \overset{\text{i.i.d.}}{\sim} \nu$
    $\hat{Q}^{\pi_{(k)}} \leftarrow \mathrm{PolicyEval}(\pi_{(k)})$
    $\pi_{(k+1)} \leftarrow \mathrm{argmin}_{\pi\in\Pi}\, \hat{L}_n^{\pi_{(k)}}(\pi)$          (classification)
**end for**
```

the empirical loss

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

where ν_n is the empirical distribution induced by the samples in $\mathcal{D}_n^{(k)}$ and $\mathbf{g}_Q$ is the *action-gap* function defined as follows: for any $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, $\mathbf{g}_Q(x) \triangleq |Q(x,1) - Q(x,2)|$ for all $x \in \mathcal{X}$. Note that the loss (1) is also used by Lazaric et al. (2010).

PolicyEval can be any algorithm that takes a policy π as input and returns an estimate $\hat{Q}^{\pi}$ of Q^{π}. This includes rollout estimation (in which case CAPI reduces to the DPI algorithm Lazaric et al. 2010), LSTD-Q (Lagoudakis and Parr, 2003a) and Fitted Q-Iteration (Ernst et al., 2005), or a combination of both rollout and function approximation (in which case CAPI becomes the DPI-Critic algorithm Gabillon et al. 2011). The only important requirement is that $\hat{Q}^{\pi}$ should be a good approximation of Q^{π} at the states in $\mathcal{D}_n^{(k)}$.

4. Theoretical Analysis

In this section we analyze the theoretical properties of the CAPI algorithm and provide an upper bound on the *performance loss* (or *regret*) of the resulting policy $\pi_{(K)}$. 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 ρ, i.e.,

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

The value of $\mathrm{Loss}(\pi_{(K)}; \rho)$ is the main quantity of interest and indicates how much worse it would be to follow $\pi_{(K)}$, on average, instead of π^*. The choice of ρ enables the user to specify the relative importance of different states.

The analysis of the performance loss has two main steps. First, in Section 4.1 we study the behaviour of one iteration of the algorithm and provide an error bound on the expected loss $L^{\pi_{(k)}}(\pi_{(k+1)}) \triangleq \int_{\mathcal{X}} g_{Q^{\pi_{(k)}}}(x)\mathbb{I}\{\pi_{(k+1)}(x) \neq \mathrm{argmax}_{a\in\mathcal{A}}\, Q^{\pi_{(k)}}(x,a)\}\, \mathrm{d}\nu$, as a function of number of samples in $\mathcal{D}_n^{(k)}$, the quality of the estimate $\hat{Q}^{\pi_{(k)}}$, the complexity of the policy space Π, and the policy approximation error. In Section 4.2, we analyze how the loss sequence $\left(L^{\pi_{(k)}}(\pi_{(k+1)})\right)_{k=0}^{K-1}$ influence $\mathrm{Loss}(\pi_{(K)}; \rho)$.

4.1. Approximate Policy Improvement Error

Policy $\pi_{(k)}$ 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 $\mathcal{D}_n^{(0)}, \dots, \mathcal{D}_n^{(k-1)}$. To avoid clutter, we will 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 $\sigma(\mathcal{D}_n^{(0)}, \dots, \mathcal{D}_n^{(k-1)})$-measurable policy and is independent of $\mathcal{D}_n$, which denotes a set of n i.i.d. samples from the distribution $\nu \in \mathcal{M}(\mathcal{X})$. We also assume that we are given $\hat{Q}^{\pi'}$, an approximation of the action-value function $Q^{\pi'}$, that is independent of $\mathcal{D}_n$.

For any $\pi \in \Pi$, we define two pointwise loss functions:

$$l(\pi) = l^{\pi'}(x;\pi) = \mathbf{g}_{Q^{\pi'}}(x)\mathbb{I}\{\pi(x) \neq \operatorname*{argmax}_{a\in\mathcal{A}} Q^{\pi'}(x,a)\},$$
$$\hat{l}(\pi) = \hat{l}^{\pi'}(x;\pi) = \mathbf{g}_{\hat{Q}^{\pi'}}(x)\mathbb{I}\{\pi(x) \neq \operatorname*{argmax}_{a\in\mathcal{A}} \hat{Q}^{\pi'}(x,a)\}.$$

For a function $l : \mathcal{X} \to \mathbb{R}$, define $P_n l = \frac{1}{n}\sum_{i=1}^n l(X_i)$ and $Pl = \mathbb{E}\left[l(X)\right]$. Here $X, X_i \sim \nu$. Now we can define the expected loss $L(\pi) = Pl(\pi)$ and the empirical loss $L_n(\pi) = P_n l(\pi)$ (both w.r.t. the true action-value function $Q^{\pi'}$) and the distorted empirical loss $\hat{L}_n(\pi) = P_n \hat{l}$ (w.r.t. the estimate $\hat{Q}^{\pi'}$). Given the dataset $\mathcal{D}_n$ and the action-value function estimate $\hat{Q}^{\pi'}$, define

$$\hat{\pi}_n \leftarrow \operatorname*{argmin}_{\pi\in\Pi} \hat{L}_n(\pi). \tag{2}$$

Here and in the rest of the paper we assume that the minimizer above exists. We make the following action-gap assumption.

Assumption A1 (Action-Gap). For a fixed MDP $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$ with $|\mathcal{A}| = 2$, there exist constants $c_g > 0$ and $\zeta \geq 0$ such that for any $\pi' \in \Pi$ and all $t > 0$, we have

$$\mathbb{P}_\nu\left(0 < \mathbf{g}_{Q^{\pi'}}(X) \leq t\right) \triangleq \int_{\mathcal{X}} \mathbb{I}\{0 < \mathbf{g}_{Q^{\pi'}}(x) \leq t\}\, \mathrm{d}\nu(x) \leq c_g\, t^{\zeta}.$$

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 and vice versa. 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 its corresponding greedy policy would still be the same as the greedy policy w.r.t. $Q^{\pi'}$. Note that this assumption is not restrictive as by setting $\zeta = 0$ and $c_g = 1$, we are effectively disabling it. The action-gap regularity is inspired by the low-noise condition in the classification literature (Audibert and Tsybakov, 2007) and was introduced to RL problems by Farahmand (2011).

Next in Lemma 1, we quantify the error caused by using $\hat{Q}^{\pi'}$ instead of $Q^{\pi'}$ in calculating the empirical loss function. Subsequently, Theorem 2 relates the quality of the minimizer of the empirical loss function to that of the expected loss function. To save space, we omitted the proofs from this paper.

Lemma 1 (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 (2) and define $\pi_n^* \leftarrow \text{argmin}_{\pi \in \Pi} L_n(\pi)$. 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*

$$L_n(\hat{\pi}_n) \leq 3L_n(\pi_n^*) + c_1 \left\| \hat{Q}^{\pi'} - Q^{\pi'} \right\|_{\infty,\mathcal{D}_n}^{1+\zeta} + c_2 \frac{z}{n},$$

with probability at least $1 - e^{-z}$.

To upper bound the expected loss $L(\hat{\pi}_n)$, we need to define a notion of complexity for the policy space Π. Among possible choices (such as the the VC dimension, metric entropy, etc), we use localized Rademacher complexity since it has favourable properties that often lead to tight upper bounds. Moreover, as opposed to the VC dimension, it can be used to describe the complexity of nonparametric (infinite dimensional) policy spaces. Another nice property of Rademacher complexity is that it can be empirically estimated. This might be of great importance in the task of model selection. In this paper, however, we do not discuss the empirical Rademacher complexity and how it can be used in model selection.

We briefly define the Rademacher complexity (Bartlett et al., 2005; Bartlett and Mendelson, 2002). 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 $R_n G = \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i)$. The Rademacher average of $\mathcal{G}$ is $\mathbb{E}[R_n G]$, in which the expectation is w.r.t. both σ and X_i. 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 $r > 0$ Bartlett et al. (2005). The following theorem is the main result of this subsection.

Theorem 2 *Fix a policy π' and assume that $\mathcal{D}_n$ consists of n i.i.d. samples drawn from distribution ν. Let $\hat{\pi}_n$ be defined as (2). Suppose that Assumption A1 holds. Let Ψ be a sub-root function with a fixed point of r^* such that for $r \geq r^*$,*

$$\Psi(r) \geq 2Q_{max}\mathbb{E}\left[R_n \left\{ l^{\pi'}(\pi) \, : \, \pi \in \Pi, P[l^{\pi'}(\pi)]^2 \leq r \right\}\right]. \tag{3}$$

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 $0 < \delta < 1$,

$$L(\hat{\pi}_n) \leq 12 \inf_{\pi \in \Pi} L(\pi) + c_1 r^* + c_2 \left\| \hat{Q}^{\pi'} - Q^{\pi'} \right\|_{\infty,\mathcal{D}_n}^{1+\zeta} + c_3 \frac{\ln(1/\delta)}{n},$$

with probability at least $1 - \delta$.

The upper bound has three important terms. The first term is $\inf_{\pi \in \Pi} L(\pi)$, which is the policy approximation error. For a rich enough policy space (e.g., a nonparametric one), this term can be zero. The constant multiplier 12 is by no means optimal and can be chosen arbitrarily close to 1 at the price of increasing other constants. The second important term is the estimation error of the classifier, which is mainly determined by the behaviour of the fixed point r^*, whose existence and uniqueness is proved in Lemma 3.2 of Bartlett et al.

(2005). The value of r^* captures the local complexity of the space $G_\Pi = \{l^{\pi'}(\pi) : \pi \in \Pi\}$ around its minimizer $\operatorname{argmin}_{\pi\in\Pi} L^{\pi'}(\pi)$. This complexity is indirectly related to the complexity of the policy space, but is not the same as it is possible to have a complex policy space but a very simple G_Π, e.g., in the extreme case in which the reward function is constant everywhere, G_Π has only a single function. Nevertheless, even a conservative analysis leads to fast rates: 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 Bartlett et al. 2005). This rate is considerably faster than $O(\sqrt{d/n})$ behaviour of the estimation error term in the result of Lazaric et al. (2010). The last important term is $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}^{1+\zeta}$, whose size depends on 1) the quality of $\hat{Q}^{\pi'}$ at points $\mathcal{D}_n$, which in turn depends on whether the policy evaluation benefits from regularities of the action-value function (such as its smoothness) and 2) the action-gap regularity of the problem characterized by ζ. Note that when there is no action-gap assumption ($\zeta = 0$), the policy evaluation error $\|\hat{Q}^{\pi'} - Q^{\pi'}\|_{\infty,\mathcal{D}_n}$ is not dampened, but when $\zeta > 0$, the rate improves. The analysis of Lazaric et al. (2010) does not benefit from this regularity. Currently, the result is only stated when the quality of policy evaluation is quantified by the supremum norm, but it could be extended to other L_p-norms as well.

4.2. Error Propagation for CAPI

In this section, we state the main result of this paper, Theorem 5, which upper bounds the performance loss $\mathrm{Loss}(\pi_{(K)}; \rho)$ as a function of the expected loss $L^{\pi_{(k)}}(\pi_{(k+1)})$ at iterations $k = 0, 1, \ldots, K-1$ and some other properties of the MDP and the policy space Π. First we introduce two definitions.

Definition 3 (Inherent Greedy Policy Error) *For a policy space* Π*, the inherent greedy policy error is* $d(\Pi) = \sup_{\pi'\in\Pi} \inf_{\pi\in\Pi} L^{\pi'}(\pi)$.

This definition can be understood as follows: Consider a policy π' belonging to Π. It induces an action-value function $Q^{\pi'}$ and consequently a greedy policy w.r.t. $Q^{\pi'}$. This greedy policy may not belong to the policy space Π, so there will be a policy approximation error $\inf_{\pi\in\Pi} L^{\pi'}(\pi)$. The inherent greedy policy error is the supremum of this error over all possible $\pi' \in \Pi$.

Next, we define a concentrability coefficient, required for the error propagation analysis; it is similar in spirit to those previously defined by Munos (2003, 2007); Farahmand et al. (2010).

Definition 4 (Concentrability Coefficient) *Given* $\rho, \nu \in \mathcal{M}(\mathcal{X})$*, a policy* π*, and two integers* $m_1, m_2 \geq 0$*, let* $\rho(P^*)^{m_1}(P^\pi)^{m_2}$ *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* $c_{\rho,\nu}(m_1; m_2; \pi) \triangleq \|\frac{\mathrm{d}(\rho(P^*)^{m_1}(P^\pi)^{m_2})}{\mathrm{d}\nu}\|_\infty$*. If* $\rho(P^*)^{m_1}(P^\pi)^{m_2}$ *is not absolutely continuous w.r.t.* ν*, we set* $c(m_1, m_2; \pi) = \infty$*. For an integer* $K \geq 1$ *and a real* $s \in [0, 1]$*, define* $C_{\rho,\nu}(K, s) \triangleq \frac{1-\gamma}{2} \sum_{k=0}^{K-1} \gamma^{(1-s)k} \sum_{m\geq 0} \gamma^m \sup_{\pi'\in\Pi} c_{\rho,\nu}(k, m; \pi')$.

We are now ready to state the main result of this paper.

Theorem 5 *Consider the sequence of independent datasets* $(\mathcal{D}_n^{(k)})_{k=1}^K$*, 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 that are obtained by solving* (1) *using estimate* $\hat{Q}^{\pi_{(k)}}$ *of* $Q^{\pi_{(k)}}$*. We assume that* $\hat{Q}^{\pi_{(k)}}$ *is independent of* $\mathcal{D}_n^{(k)}$*. Suppose that Assumption A1 holds and* r^* *is the fixed point of a sub-root function* Ψ *that for any* $\pi' \in \Pi$ *and* $r \geq r^*$ *satisfies* $\Psi(r) \geq 2Q_{max}\mathbb{E}\left[R_n\left\{ l^{\pi'}(\pi) : \pi \in \Pi, P[l^{\pi'}(\pi)]^2 \leq r \right\}\right]$*. Then there exist constants* $c_1, c_2, c_3 > 0$ *such that for any* $0 < \delta < 1$*, for* $\mathcal{E}(s)$ *defined as* $(0 \leq s \leq 1)$

$$\mathcal{E}(s) \triangleq 12d(\Pi) + c_1 r^* + c_2 \max_{0 \leq k \leq K-1}\left[\gamma^{(K-k-1)s}\left\|\hat{Q}^{\pi_{(k)}} - Q^{\pi_{(k)}}\right\|_{\infty,\mathcal{D}_n}^{1+\zeta}\right] + c_3\frac{\ln(K/\delta)}{n},$$

we have with probability at least $1 - \delta$*,*

$$\text{Loss}(\pi_{(K)};\rho) \leq \frac{2}{1-\gamma}\left[\inf_{s\in[0,1]} C_{\rho,\nu}(K,s)\,\mathcal{E}(s) + \gamma^K R_{max}\right].$$

All the discussion after Theorem 2 applies here too. Moreover, the new error propagation result shows improvement compared to Lazaric et al. (2010). The current result indicates that the error $\|\hat{Q}^{\pi_{(k)}} - Q^{\pi_{(k)}}\|_{\infty,\mathcal{D}_n}$ is weighted proportional to $\gamma^{(K-k-1)s}$, which means that the errors at earlier iterations are geometrically discounted. So, if one has finite resources (samples or computational time), it is better to obtain better estimates of $Q^{\pi_{(k)}}$ at later iterations. The same advice holds for the classifier too: using more samples at later iterations is beneficial (though this is not apparent from the bound, as we fixed n throughout all iterations). See Farahmand et al. (2010) for more discussion on this type of error propagation results.

5. Illustration

The goal of this illustration is to show that using action-gap weighted loss can lead to significantly better performance compared to 1) pure value-based approaches and 2) classification-based API with 0/1 loss. Here we compare CAPI with Value Iteration (VI), Policy Iteration (PI), and a modified CAPI that uses 0/1-loss on a simple 1D chain walk problem (based on the example in Section 9.1 of Lagoudakis and Parr 2003a). The problem has 200 states, the reward function is zero everywhere except at states $10-15$ (where it is $+1$ for both actions) and $180-190$ (where it is $+0.1$ for both actions), and $\gamma = 0.99$.

Note that the model is known. Moreover, CAPI is run when the measure ν_n in the loss function (1) is the uniform distribution over states, so there is no "sampling" from the states. The value of $\hat{Q}^{\pi_{(k)}}$ at iteration k of CAPI is obtained by running just one iteration of evaluation, i.e., $\hat{Q}^{\pi_{(k)}} = T^{\pi_{(k)}}\hat{Q}^{\pi_{(k-1)}}$, in which $T^{\pi_{(k)}}$ is the Bellman operator for policy $\pi_{(k)}$. This makes the number of times CAPI queries the model similar to VI. The policy space Π used in CAPI is defined as the space of the indicator functions of the set of all half spaces, i.e., the set of policies that choose action 1 (or 2) on $\{1,\ldots,p\}$ and action 2 (or 1) on $\{p+1,\ldots,200\}$ for $1 \leq p \leq 200$. This is a very small subset of all possible policies. We intentionally designed the reward function such that the optimal policy is **not** in Π, so CAPI will be subject to policy approximation error.

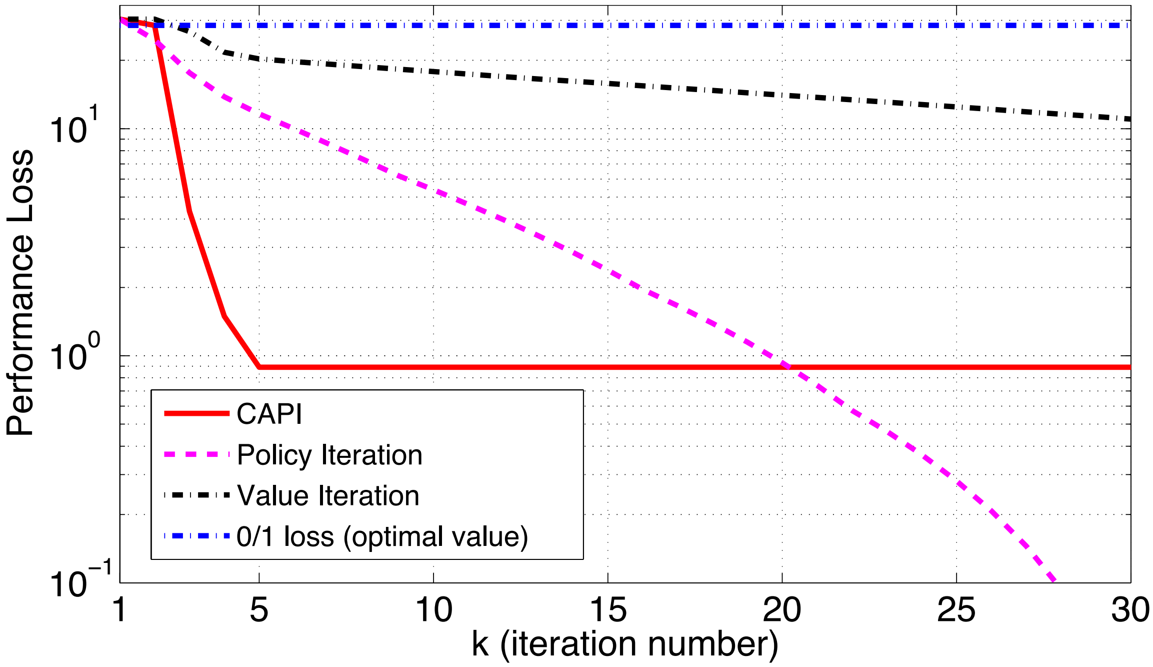

Figure 1: Performance loss for CAPI, Value Iteration, Policy Iteration, and 0/1 loss given Q^*. The problem is a 1D random walk with 200 states and $\gamma = 0.99$.

Figure 1 shows that the performance loss of CAPI converges to this policy approximation error, which is the minimum achievable given Π. The convergence rate is considerably faster than that of VI and PI. This speedup is due to the fact that CAPI searches in a much smaller policy space compared to VI or PI. The comparison of CAPI and VI is especially striking since both of them use the same number of queries to the model. We also report the performance loss of a modified CAPI that uses the 0/1 loss and the *exact* Q^* (so there will be no estimation error). The result is quite poor. To understand this behaviour, note that the minimizer of the 0/1 loss is a policy that approximates the greedy policy (in this case, the optimal policy) without paying attention to the action-gap function. Here the minimizer policy is such that it fits the greedy policy in a large region of the state space where the action-gap is small and differs from the greedy policy in a smaller region where the action-gap is large. This selection is poor as it ignores the relative importance of choosing the wrong action in different regions of the state space.

6. Conclusion

We proposed a general family of classification-based RL algorithms (which has some existing algorithms as special cases). Our approach uses any policy evaluation method of choice, defines an action-gap weighted loss function and then minimizes the loss. We provided an error upper bound that is tighter than existing results and applies to nonparametric policy spaces as well. An open question is how to efficiently solve the optimization problem. The use of surrogate losses seems to be the answer, but theoretical properties should be investigated further. Another question is how to choose the sampling distribution ν, which can greatly affect performance. One may even change the sampling distribution at each iteration to actively obtain more informative samples. How this should be done remains to be answered.

References

Jean-Yves Audibert and Alexander B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific, 1996.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS - 24)*, 2011.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 441–448. MIT Press, 2009.

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 568–576. 2010.

Alan Fern, Sungwook Yoon, and Robert Givan. Approximate policy iteration with a policy language bias: Solving relational markov decision processes. *Journal of Artificial Intelligence Research*, 25:85–118, 2006.

Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, and Bruno Scherrer. Classification-based policy iteration with a critic. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, Washington, USA,, 2011. Omnipress.

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matthew Hoffman. Finite-sample analysis of lasso-TD. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1177–1184, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5.

J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 521–528. ACM, 2009.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003a.

Michail G. Lagoudakis and Ronald Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *ICML '03: Proceedings of the 20th international conference on Machine learning*, pages 424–431, 2003b.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of a classification-based policy iteration algorithm. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 607–614. Omnipress, 2010.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML 2003: Proceedings of the 20th Annual International Conference on Machine Learning*, pages 560–567, 2003.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010.

Gavin Taylor and Ronald Parr. Kernelized value function approximation for reinforcement learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1017–1024, New York, NY, USA, 2009. ACM.