Classification-based Policy Iteration with a Critic

V. Gabillon[1], A. Lazaric[1], M. Ghavamzadeh[1] & B. Scherrer[2]

[1] INRIA Lille - Nord Europe, Team Sequel,
[2] INRIA Nancy Grand Est, Team MAIA.

INRIA

ICML, Bellevue, June 30, 2011.

Outline

Outline

1. Classification-based Policy Iteration
2. Introduction of a Critic
3. Theoretical Analysis
4. Experiments

Definitions I

Discounted Markov Decision Process MDP

A discounted Markov Decision Process $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{A}, r, p, \gamma \rangle$, where

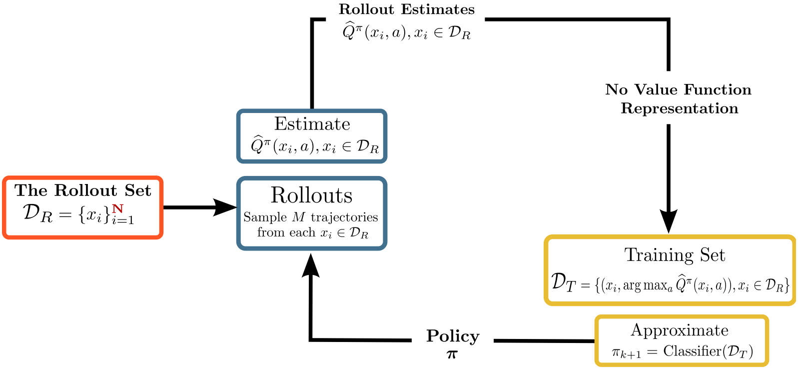
- The state space $\mathcal{X}$,
- The set of actions $\mathcal{A}$ is **finite** ($|\mathcal{A}| < \infty$),
- The transition model $p(\cdot|x, a)$ is a distribution over $\mathcal{X}$,
- The reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$,
- $\gamma \in (0, 1)$ is a discount factor.

π is a deterministic **policy** $\pi : \mathcal{X} \to \mathcal{A}$.

Classification-based Policy Iteration (CbPI)

Fern et.al. (2004), Lagoudakis & Parr (2003)

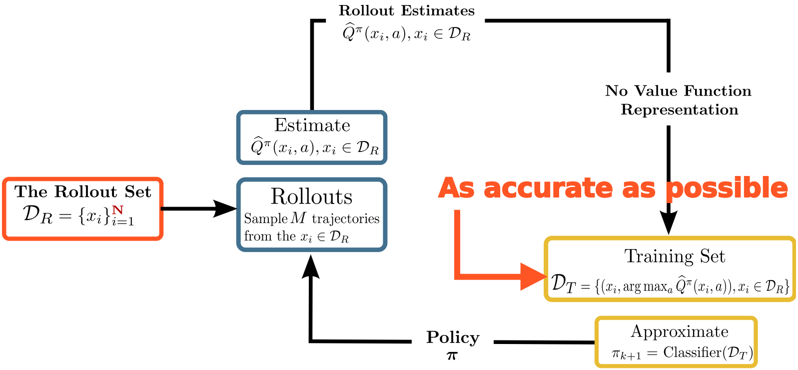
Without Value Function representation!



The policy improvement step is cast as a *classification* problem.

Classification-based Policy Iteration (CbPI)

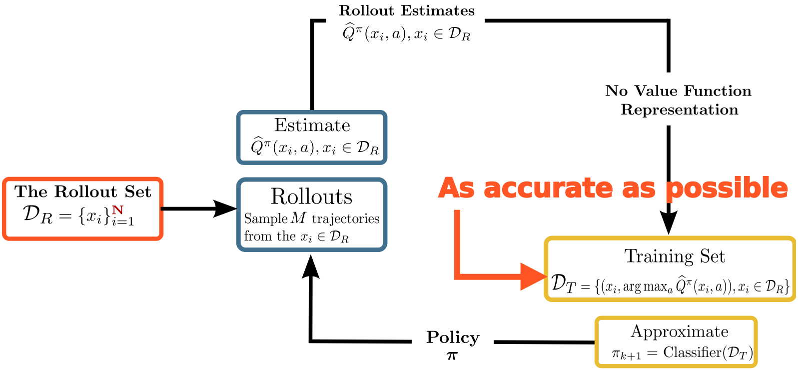
Without Value Function representation!



The policy improvement step is cast as a *classification* problem.

Classification-based Policy Iteration (CbPI)

Without Value Function representation?



The policy improvement step is cast as a *classification* problem.

Definitions II

Rollout

Sampling a trajectory starting from a state-action pair (x_0, a_0) and following policy π with return $R^\pi(x_0, a_0)$ bounded by $V_{\max} = \frac{R_{\max}}{1-\gamma}$:

$$R^\pi(x_0, a_0) = r(x_0, a_0) + \sum_{t>0} \gamma^t r(x_t, \pi(x_t)),$$
$$x_{t+1} \sim p(\cdot | x_t, a_t)$$

$$\mathbf{E}[R^\pi(x, a)] = Q^\pi(x, a), \qquad \forall x \in \mathcal{X} \text{ and } \forall a \in \mathcal{A},$$

$$\widehat{Q}^\pi(x_i, a) = \frac{1}{M} \sum_{j=1}^{M} R_j^\pi(x_i, a), \qquad \forall x \in \mathcal{D}_R \text{ and } \forall a \in \mathcal{A}.$$

Outline

Limitations of CbPI algorithms

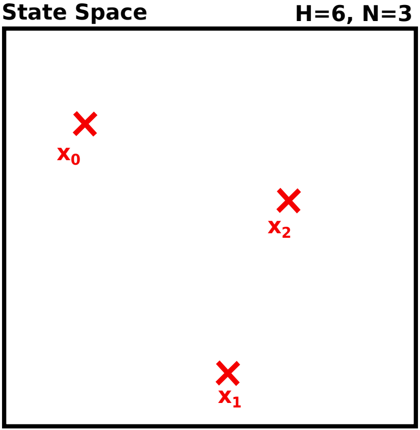
Limited budget In practice, the total number of calls B to the generative model (for each iteration) is limited. Therefore, the rollouts are truncated after H steps.

$$R^\pi(x_0, a_0) = \underbrace{r(x_0, a_0) + \sum_{t=1}^{H-1} \gamma^t r(x_i^t, \pi(x_t))}_{R^{\pi,H}(x_0,a_0)} + \sum_{t \geq H} \gamma^t r(x_i^t, \pi(x_t))$$

Bias The return after the truncation is implicitly set to 0.

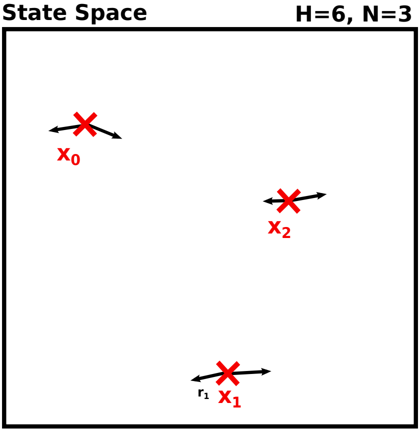
Variance High values of H leads to high variance estimates.

Introduction of a Value function approximation (Critic)

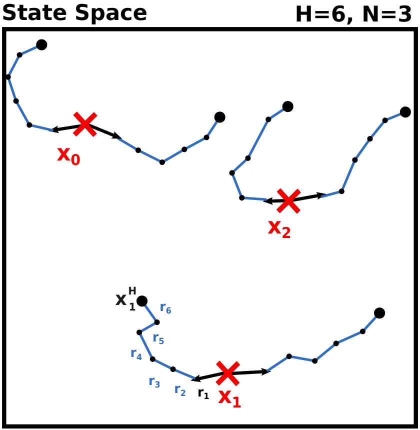


Rollout set distribution: $x_i \overset{iid}{\sim} \rho$

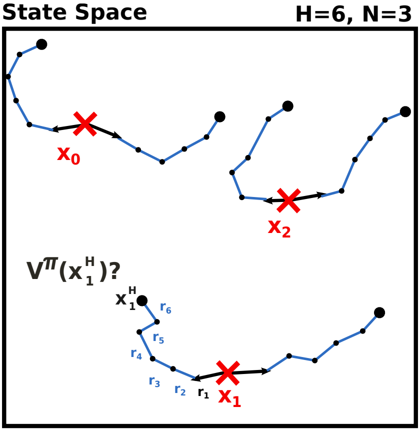
Introduction of a Value function approximation (Critic)



Introduction of a Value function approximation (Critic)

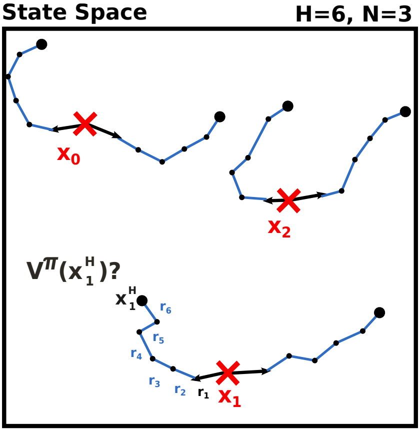


Introduction of a Value function approximation (Critic)



$$R^{\pi}(x_i, a) = R^{\pi,H}(x_i, a) + 0 \quad \leftarrow \textbf{ biased!}$$

Introduction of a Value function approximation (Critic)



New return: $R^{\pi}(x_i, a) = R^{\pi,H}(x_i, a) + \gamma^H \widehat{V}^{\pi}(x_i^H)$

DPI-Critic Algorithm

Input: policy space Π, value-function space $\mathcal{F}$, state distribution ρ
for $k = 0, 1, 2, \ldots$ **do**

- **Critic:** **(Extra budget n!)**

Construct the set S_k of n samples drawn from σ

$\widehat{V}^{\pi_k} \leftarrow \text{VF-APPROX}_{\mathcal{F}}(S_k)$

- **Rollout:** **(Budget B)**

Construct the rollout set $\mathcal{D}_R = \{x_i\}_{i=1}^N,\ x_i \overset{\text{iid}}{\sim} \rho$
for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**
  for $j = 1$ to M **do**
    Perform a rollout according to π_k and return
    $R_j(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x_{i,j}^t, \pi_k(x_{i,j}^t)\big)$
    $+ \gamma^H \widehat{V}^{\pi_k}(x_{i,j}^H)$,
    $x_{i,j}^t \sim p\big(\cdot \,|x_{i,j}^{t-1}, \pi_k(x_{i,j}^{t-1})\big)$ and $x_{i,j}^1 \sim p(\cdot|x_i, a)$
  end for
  $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^M R_j(x_i, a)$
end for

$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}; \pi)$ **(classifier)**
end for

Why DPI-Critic?

Bias The value function approximation provided by the critic reduces the bias.

Variance Small H reduces the estimation variance.

Compared to...

Classification-based We expect an improvement as soon as the accuracy of the value function is better than predicting 0.

VF-based No need for a perfect critic.
DPI-Critic still relies on a classifier as its policy improvement step.

Outline

Definition of the **weighted** losses and errors

Expected loss and expected error

$$\ell_{\pi_k}(x;\pi) = \max_{a\in\mathcal{A}} \; Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big), \qquad \forall x \in \mathcal{X},$$

$$\mathcal{L}_{\pi_k}(\rho;\pi) = \int_{\mathcal{X}} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\, \rho(dx)\;.$$

Empirical loss and empirical error

$$\widehat{\ell}_{\pi_k}(x;\pi) = \max_{a\in\mathcal{A}} \; \widehat{Q}^{\pi_k}(x,a) - \widehat{Q}^{\pi_k}\big(x,\pi(x)\big), \quad \forall x \in \mathcal{X},$$

$$\widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho};\pi) = \frac{1}{N} \sum_{i=1}^{N} \Big[\max_{a\in\mathcal{A}} \widehat{Q}^{\pi_k}(x_i,a) - \widehat{Q}^{\pi_k}(x_i,\pi(x_i))\Big].$$

The theorem

Theorem (Bound on the error at each iteration)

Let Π be a policy space with finite VC-dimension $h = VC(\Pi) < \infty$. Then, for any $\delta > 0$, we have

$$
\begin{aligned}
\mathcal{L}_{\pi_k}(\rho; \pi_{k+1}) \leq & \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho; \pi) \quad \textbf{(} \leftarrow \textbf{Approximation error)} \\
& + \quad \ldots \\
& + \quad \ldots
\end{aligned}
$$

with probability $1 - \delta$.

The theorem

Theorem (**Bound on the error at each iteration**)

Let Π *be a policy space with finite VC-dimension* $h = VC(\Pi) < \infty$. *Then, for any* $\delta > 0$, *we have*

$$\mathcal{L}_{\pi_k}(\rho; \pi_{k+1}) \leq \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho; \pi) \quad (\downarrow \textbf{ from empirical to expected loss})$$
$$+ 2\left(16 V_{\max} \sqrt{\frac{2}{N}\left(h \log \frac{eN}{h} + \log \frac{32}{\delta}\right)} \quad + \ldots\right.$$
$$+ \quad \ldots$$

with probability $1 - \delta$.

Theorem (**Bound on the error at each iteration**)

Let Π *be a policy space with finite VC-dimension* $h = VC(\Pi) < \infty$. *Then, for any* $\delta > 0$, *we have*

$$\mathcal{L}_{\pi_k}(\rho; \pi_{k+1}) \leq \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho; \pi) + 2\Bigg(\epsilon_0 + \ldots + \quad \ldots$$

with probability $1 - \delta$.

The theorem

Theorem (Bound on the error at each iteration)

Let Π be a policy space with finite VC-dimension $h = VC(\Pi) < \infty$. Then, for any $\delta > 0$, we have

$$
\begin{aligned}
\mathcal{L}_{\pi_k}(\rho; \pi_{k+1}) \leq & \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho; \pi) \quad \textbf{(} \downarrow \textbf{H-step rollouts)} \\
& + 2\Bigg(\epsilon_0 + (1-\gamma^H) V_{\max} \sqrt{\frac{2 \log(4|\mathcal{A}|/\delta)}{MN}} \\
& + \quad \ldots
\end{aligned}
$$

with probability $1 - \delta$.

The theorem

Theorem (Bound on the error at each iteration)

Let Π be a policy space with finite VC-dimension $h = VC(\Pi) < \infty$. Then, for any $\delta > 0$, we have

$$\mathcal{L}_{\pi_k}(\rho; \pi_{k+1}) \leq \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho; \pi) + 2\Big(\epsilon_0 + \epsilon_1 + \quad \ldots$$

with probability $1 - \delta$.

The theorem

Theorem (Bound on the error at each iteration)

Let Π *be a policy space with finite VC-dimension* $h = VC(\Pi) < \infty$. *Then, for any* $\delta > 0$, *we have*

$$\begin{aligned}\mathcal{L}_{\pi_k}(\rho;\pi_{k+1}) \leq & \inf_{\pi\in\Pi}\mathcal{L}_{\pi_k}(\rho;\pi) \\ & + 2\Bigg(\epsilon_0 + \epsilon_1 \\ & + \gamma^H\big(\underbrace{\widetilde{\mathcal{O}}(\frac{1}{\sqrt{MN}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{N}})}_{\textit{Rollout estimation}} + \underbrace{K_{approx} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})}_{\textit{Critic estimation}}\big)\Bigg)\end{aligned}$$

with probability $1 - \delta$.

At least as good as DPI...

Terms of the bound of DPI-Critic:

$$2\left(\epsilon_0 + \epsilon_1 + \gamma^H \big(\underbrace{\widetilde{\mathcal{O}}(\frac{1}{\sqrt{MN}}) + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{N}})}_{\text{Rollout estimation}} + \underbrace{K_{approx} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})}_{\text{Critic estimation}} \big)\right).$$

The bound of DPI in **Lazaric et. al. (2010)**

$$2(\epsilon_0 + \epsilon_1 + \gamma^H V_{\max}).$$

- ϵ_0 and ϵ_1 are the same DPI & DPI-Critic.
- The last terms share the same term $\gamma^H V_{\max}$ multiplied by a factor which decreases with N, M, n.
- Comparison depending on the approximation error of the function space $\mathcal{F}$.

...but unfair comparison.

DPI-Critic uses n samples more than DPI. Let's fix the budget to B and split it in two:

$$B = \underbrace{B_R}_{\text{Rollout budget}} + \underbrace{B_C}_{\text{Critic budget}}$$

with $B_C = n = Bp$ and $B_R = B(1-p)$. $p \in (0,1)$ is the critic ratio.

- Can we expect DPI-Critic to outperform DPI?

Whenever the advantage obtained by relying on the critic is larger than the loss in having a smaller number of rollouts.

Outline

Comparisons

We compare DPI-Critic to:

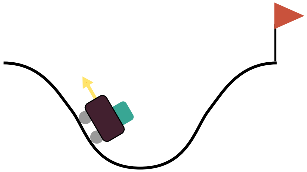
Classification-based algorithm: DPI

Value function-based algorithm: LSPI

Domains

We used classical setting of Mountain Car and Inverted Pendulum.

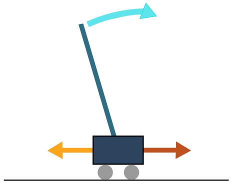
Mountain Car (MC)



Formulation:
Sutton & Barto(1998)

Noise to: 1

Inverted Pendulum (IP)



Formulation:
Lagoudakis & Parr (2003)

Noise to: 10

Experiments

The budget is split in two:

$$B = \underbrace{B_R}_{\text{Rollout budget}} + \underbrace{B_C}_{\text{Critic budget}}$$

with $B_C = Bp$ and $B_R = B(1-p)$. p is the critic ratio.

Critic trade-off

- Low values of p lead to a poor critic.
- High values of p lead to very small or inaccurate training set.

Find p^*!

Rollout budget

$$B_R = MN|A|H$$

Rollout trade-off

- With large H the rollouts are more likely to be *informative*.
- With larger N, the estimation error is reduced.

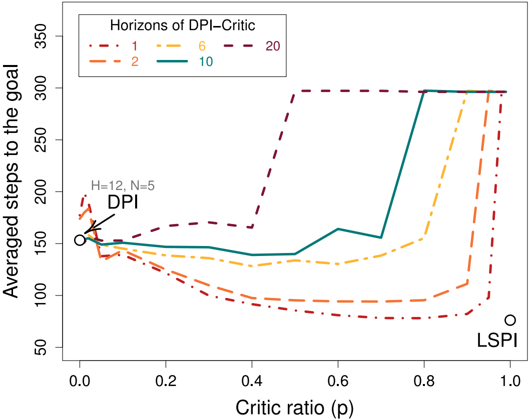
Performance of DPI We report the best combination of (H, M, N) among those tested.

Performance of DPI-Critic We report the best combination of (M, N) for different values of H.

In all the experiments, $M = 1$ provided us with the best results.

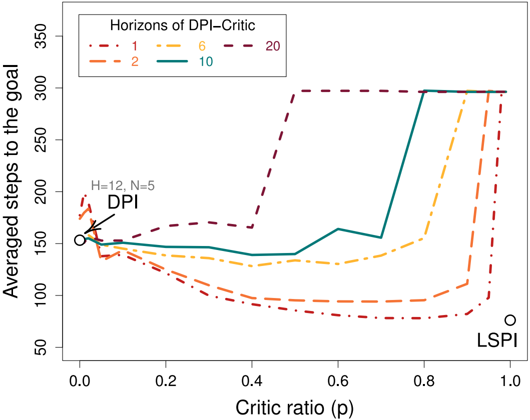
Mountain car: $3 * 3$ RBFs

$B = 200$. The objective is to **minimize** the number of steps to the goal.



Mountain car: $3 * 3$ RBFs

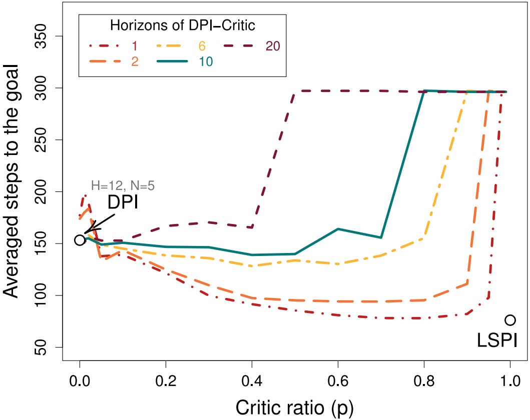
$B = 200$. The objective is to **minimize** the number of steps to the goal.



The introduction of a Critic helps DPI-Critic reach the performance of LSPI.

Mountain car: $3 * 3$ RBFs

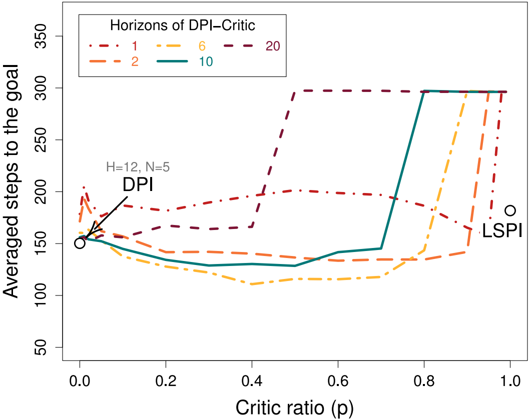
$B = 200$. The objective is to **minimize** the number of steps to the goal.



What happens if the VF approximation space is poorer?

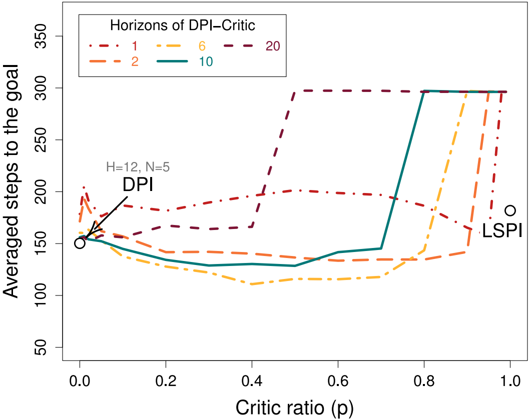
Mountain car: $2 * 2$ RBFs

$B = 200$. The objective is to **minimize** the number of steps to the goal.



Mountain car: $2 * 2$ RBFs

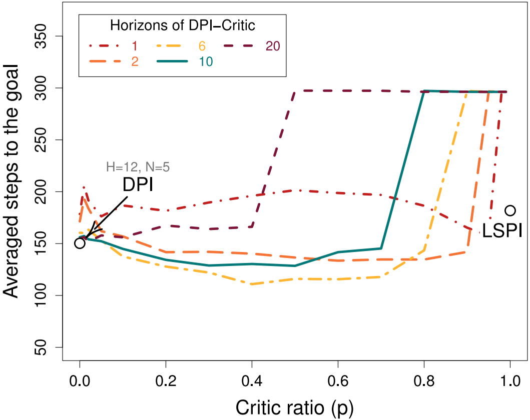
$B = 200$. The objective is to **minimize** the number of steps to the goal.



For $0.1 < p < 0.8$, DPI-Critic outperforms both LSPI and DPI.

Mountain car: $2 * 2$ RBFs

$B = 200$. The objective is to **minimize** the number of steps to the goal.



Optimal H depends on the critic and on the dynamics.

Conclusion

- We theoretically analysed the performance of DPI-Critic.
- This analysis is also supported by the experimental results which confirm the capability of DPI-Critic to take advantage of both rollouts and critic.
- In some environment, either DPI or LSPI might still be the better choice, still DPI-Critic is a promising alternative that introduces additional flexibility in the design of the algorithm.

Future Work

- A more detailed comparison of DPI-Critic and LSPI, including more challenging domains.
- Finding optimal or good rollout allocation strategies.

Thank You!

Assumptions

We introduce the following assumptions.

Assumption 1

At each iteration k of DPI-Critic, a data-set $S_k = \{(X_i, R_i)\}_{i=1}^n$ is built, where X_i's are obtained by following a single trajectory generated by a stationary β-mixing process with parameters $\hat{\beta}, b, \kappa$, and a stationary distribution σ_k equal to the stationary distribution of the Markov chain induced by policy π_k, and $R_i = r\big(X_i, \pi_k(X_i)\big)$.

Assumption 2

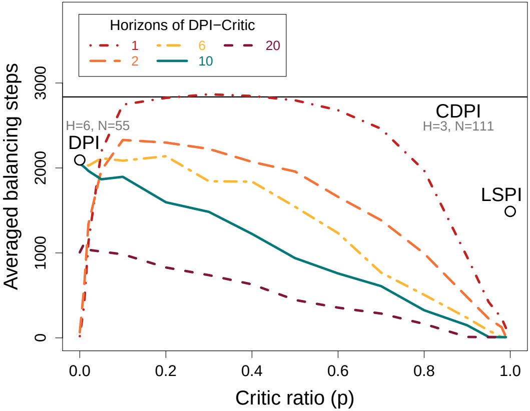
The rollout set sampling distribution ρ is such that for any policy $\pi \in \Pi$ and any action $a \in \mathcal{A}$, $\mu = \rho P^a (P^\pi)^{H-1} \leq C\sigma$, where $C < \infty$ is a constant and σ is the stationary distribution of π.

Lazaric et. al. (2010)'s proof

$$
\begin{aligned}
\mathcal{L}_{\pi_k}(\rho;\pi_{k+1}) &\overset{(a)}{\leq} \mathcal{L}_{\pi_k}(\widehat{\rho};\pi_{k+1}) + \epsilon_0 && \text{w.p. } 1-\delta' \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[Q^{\pi_k}(x_i,a^*) - Q^{\pi_k}\left(x_i,\pi_{k+1}(x_i)\right)\right] + \epsilon_0 \\
&\overset{(b)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\left[Q^{\pi_k}(x_i,a^*) - \widehat{Q}^{\pi_k}\left(x_i,\pi_{k+1}(x_i)\right)\right] + \epsilon_0 + \epsilon_T && \text{w.p. } 1-2\delta' \\
&\overset{(c)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\left[Q^{\pi_k}(x_i,a^*) - \widehat{Q}^{\pi_k}\left(x_i,\pi^*(x_i)\right)\right] + \epsilon_0 + \epsilon_T \\
&\overset{(d)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\left[Q^{\pi_k}(x_i,a^*) - Q^{\pi_k}\left(x_i,\pi^*(x_i)\right)\right] + \epsilon_0 + 2(\epsilon_T) && \text{w.p. } 1-3\delta' \\
&= \mathcal{L}_{\pi_k}(\widehat{\rho};\pi^*) + \epsilon_0 + 2(\epsilon_T) \\
&\overset{(e)}{\leq} \mathcal{L}_{\pi_k}(\rho;\pi^*) + 2(\epsilon_0 + \epsilon_T) && \text{w.p. } 1-4\delta'
\end{aligned}
$$

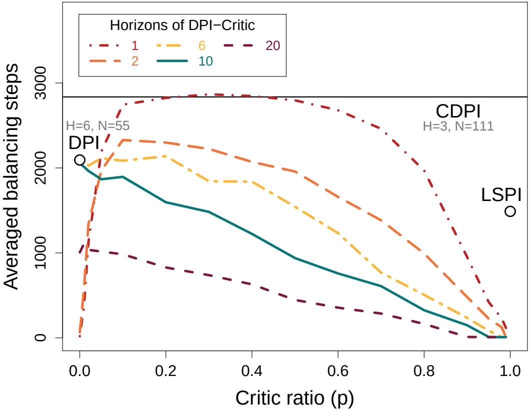
Inverted pendulum: $3 * 3$ RBFs

$B = 200$. The objective is to **maximize** the number of steps to the goal.



Inverted pendulum: $3 * 3$ RBFs

$B = 200$. The objective is to **maximize** the number of steps to the goal.



For $0.1 < p < 0.6$, DPI-Critic outperforms both LSPI and DPI.

DPI-Critic Algorithm

for $k = 0, 1, 2, \ldots$ **do**

S_k and $\mathcal{D}_k$ are independent. The budget is split.

- **Critic with budget B_C:**

Construct the set S_k of n samples (e.g., by following a trajectory or by using the generative model)

$\widehat{V}^{\pi_k} \leftarrow$ VF-APPROX(S_k)

- **Rollout: with budget B_R:**

Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N$, $x_i \overset{\text{iid}}{\sim} \rho$

for all states $x_i \in \mathcal{D}_k$ and actions $a \in$ **do**

  for $j = 1$ to M **do**

    Perform a rollout according to π_k and return

    $R_j(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x_{i,j}^t, \pi_k(x_{i,j}^t)\big)$
    $+ \gamma^H \widehat{V}^{\pi_k}(x_{i,j}^H)$,

  end for

  $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^M R_j(x_i, a)$

end for

$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}; \pi)$ **(classifier)**

end for

DPI-Critic Algorithm

for $k = 0, 1, 2, \ldots$ **do**

S_k and $\mathcal{D}_k$ are independent. The budget is split.

- **Critic with budget B_C:**
  Construct the set S_k of n samples (e.g., by following a trajectory or by using the generative model)

  $\widehat{V}^{\pi_k} \leftarrow$ VF-APPROX(S_k)

- **Rollout: with budget B_R:**
  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N$, $x_i \stackrel{\text{iid}}{\sim} \rho$

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in$ **do**

  for $j = 1$ to M **do**

  Perform a rollout according to π_k and return $R_j(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x_{i,j}^t, \pi_k(x_{i,j}^t)\big) + \gamma^H \widehat{V}^{\pi_k}(x_{i,j}^H)$,

  end for

  $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^{M} R_j(x_i, a)$

  end for

$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}; \pi)$ **(classifier)**

end for

Can we merge the collection of samples for the critic and the rollouts?

Combined DPI-Critic Algorithm

for $k = 0, 1, 2, \ldots$ **do**

- **1) Rollout: with budget B:**

  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^{N}$, $x_i \overset{\text{iid}}{\sim} \rho$
  for all states $x_i \in \mathcal{D}_k$ and actions $a \in$ **do**
    for $j = 1$ to M **do**
      Perform a rollout according to π_k and return
      $R_j(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x_{i,j}^t, \pi_k(x_{i,j}^t)\big)$
    end for
  end for

- **2) Critic**

  Construct the set S_k **with the trajectories followed by the rollouts**.

  $\widehat{V}^{\pi_k} \leftarrow$ VF-APPROX(S_k)

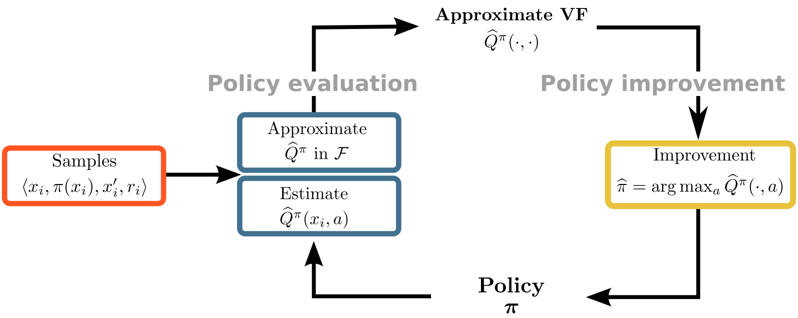
- **3) Merge :**

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in$ **and** $j = 1$ to M **do**
    $R_j(x_i, a) = R_j(x_i, a) + \gamma^H \widehat{V}^{\pi_k}(x_{i,j}^H)$,
  end for

$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}; \pi)$ **(classifier)**

end for

Value function-based Policy Iteration



The performance depends on $\mathcal{F}$.

Setting

Classifier

- We use a multi-layer perceptron with 10 hidden units.

Sampling distribution

Rollout set Uniform in both domains.

Critic training set Uniform in MC, like in **Lagoudakis & Parr** in IP.

Setting II

Approximation space for the value function

- Linear combination of RBFs plus a constant offset.
- RBFs are placed uniformly over the state space.

Algorithms to approximate the value function

- We use LSTD in DPI-Critic.