

FINITE-TIME BOUNDS FOR SAMPLING-BASED FITTED VALUE ITERATION

Rémi Munos¹ Csaba Szepesvári²

¹Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau Cedex, France

E-mail: remi.munos@polytechnique.fr

²Computer and Automation Research Institute of the
Hungarian Academy of Sciences

Kende u. 13-17, Budapest 1111, Hungary

E-mail: szcsaba@sztaki.hu

International Conference on Machine Learning
Bonn, 2005

CONTENTS

- 1 FITTED VALUE ITERATION
 - Markovian Decision Problems
 - Least Squares Value Iteration
 - Previous Results
- 2 ALGORITHMS AND RESULTS
 - Algorithm
 - Finite-time Bounds
 - Intuitions on the Proof
 - Single-sample Variant
 - How to Use the Result?
- 3 ILLUSTRATION
- 4 CONCLUSIONS

PROBLEM SETUP

Problem Setup:

- Markovian Decision Problems, continuous (or very large) state-spaces
- Generative model (“planning”)
- $\Rightarrow$ Value function approximation
- $\Rightarrow$ Approximate Dynamic Programming (ADP)

Main problem:

- Standard analysis uses L^∞ bounds
- Function fitting uses L^2 (L^p) bounds: No L^∞ guarantees!

GOAL

Finite sample bounds for a practical algorithm (FVI)

PRELIMINARIES – NORMS

- Supremum-norm:

$$\|f\|_{\infty} \stackrel{\text{def}}{=} \sup_{x \in \mathcal{X}} |f(x)|$$

- Space of bounded functions: $B(\mathcal{X})$
- $L^p(\mu)$ -norms: μ distribution over $\mathcal{X}$, $p \geq 1$:

$$\|f\|_{p,\mu} \stackrel{\text{def}}{=} \left(\int |f(x)|^p \mu(dx) \right)^{1/p}.$$

- Space of $L^p(\mu)$ -norm bounded functions: $L^p(\mathcal{X}; \mu)$

MARKOVIAN DECISION PROBLEMS

$(\mathcal{X}, \mathcal{A}, P, r)$: State space $\mathcal{X}$ ($\subset \mathbb{R}^d$, closed, compact), action space $\mathcal{A}$ (finite), transition probabilities $P(\cdot|x, a)$, reward function $r(x, a)$.

Definitions:

- (stationary) **policy**: a mapping $\pi : \mathcal{X} \rightarrow \mathcal{A}$,
- The **value function** V^π defines the performance of a policy π , for example (in the infinite horizon, expected discounted reward case):

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \mid X_0 = x, A_t = \pi(X_t)\right].$$

Optimal control problem: find (an) optimal policy π^* , i.e., $V^{\pi^*} = \sup_{\pi} V^\pi$, V^{π^*} written V^* , called the **optimal value function**

DYNAMIC PROGRAMMING

Proposition: The optimal value function V^* solves the Dynamic Programming (or Bellman) Equation:

$$V^* = TV^*$$

where $T : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ is the **Bellman operator**:

$$(TW)(x) \stackrel{\text{def}}{=} \max_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \int W(y) P(dy|x, a) \right\}.$$

Definition: A policy π is **greedy** w.r.t. $W \in B(\mathcal{X})$ if $\forall x \in \mathcal{X}$,

$$\pi(x) \in \operatorname{argmax}_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \int W(y) P(dy|x, a) \right\}.$$

VALUE ITERATION

- **Property:** T is a contraction mapping in L^∞ -norm Banach Fixed Point Theorem $\Rightarrow$ the optimal value function is the unique solution of the DP equation and may be computed by **value iteration**:

$$V_{k+1} = TV_k$$

with any initial V_0 . Then $V_k \rightarrow V^*$.

When the state space is large or infinite (e.g. continuous), the iterates cannot be computed exactly any more $\Rightarrow$ use function approximation.

- **Fitted Value Iteration** (Boyan (1995), Gordon (1995),...)

$$V_{k+1} = \operatorname{argmin}_{f \in \mathcal{F}} \|f - TV_k\|$$

for some appropriate norm $\|\cdot\|$, and where $\mathcal{F}$ is an appropriate function-space (finite dim. parametrizations)

EXAMPLE: LEAST SQUARES VALUE ITERATION

Let the norm be $L^2(\mu)$ for some distribution μ defined over $\mathcal{X}$.
Let $\mathcal{X}$ be finite, e.g. $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ ($\sum_{i=1}^N \mu(x_i) = 1$)
e.g. $\mathcal{F} = \{\theta^T \phi(x) \mid \theta \in \mathbb{R}^m\}$

Examples: regression with polynomials, cosines, wavelets

Algorithm (stage k):

- 1 Calculate the backed up values $v_i = TV_k(x_i)$, $i = 1, 2, \dots, N$.
- 2 Solve the least-squares problem:

$$V_{k+1} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N (f(x_i) - v_i)^2 \mu(x_i)$$

Other possibilities: non-linear approximation (neural networks, adaptive wavelets), non-parametric methods (locally weighted learning, support vector regression, kernel methods, Kriging interpolations, ...).

PREVIOUS RESULTS: I

Fitted value iteration is a *special case* of approximate value iteration:

$$V_{k+1} = TV_k + \epsilon_k.$$

L^∞ -bound Theorem [Bertsekas & Tsitsiklis, 1996]: Let π_k be the greedy policy w.r.t. V_k . Then

$$\limsup_{k \rightarrow \infty} \|V^* - V^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k \rightarrow \infty} \|\epsilon_k\|_\infty$$

PREVIOUS RESULTS: II

$$\limsup_{k \rightarrow \infty} \|V^* - V^{\pi_k}\|_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k \rightarrow \infty} \|\epsilon_k\|_{\infty}$$

Problems:

- **Problem 1:** TV can be discontinuous, in which case $\limsup_{k \rightarrow \infty} \|\epsilon_k\|_{\infty}$ will be big! However, this does not mean that the algorithm does not work in these cases (e.g. when it is not important to know precisely the values at the discontinuity!).
- **Problem 2:** Difficult to guarantee uniformly small errors (i.e. that $\|\epsilon_k\|_{\infty}$ is small) – algorithms optimize for a different norm.
- **Result:** Algorithm may work, analysis breaks down. **Idea:** Use L^p bounds that tolerate local approximation errors.

PREVIOUS RESULTS: III

L^p -bound Theorem: [Munos, 2005] Assume $\mathcal{X}$ is finite. Let ρ be a distribution used to evaluate the performance of the greedy policies π_k . Consider FVI with $L^p(\mu)$ -norm fitting. Then

$$\limsup_{k \rightarrow \infty} \|V^* - V^{\pi_k}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C^{1/p} \limsup_{k \rightarrow \infty} \|\epsilon_k\|_{p,\mu}$$

Advantages:

- $\|\epsilon_k\|_{p,\mu}$ directly controlled in the algorithm
- Capable of tolerating discontinuities

Problems:

- ① $C = ?$ Might be difficult to estimate C (the estimate might be too crude)
- ② Result proved for finite state spaces, knowledge of model

Goal here: Remove Problem 2.

ALGORITHM

Input: $\mathcal{F}$ – function space, N, M, K integers, μ – distribution over the state space.

Algorithm (stage k):

- ① Sample “basis points”: $X_1, \dots, X_N \in \mathcal{X}$, $X_i \sim \mu$
- ② For each action $a \in \mathcal{A}$ and state X_i , sample next states and rewards: $Y_j^{X_i, a} \sim P(\cdot | X_i, a)$, $R_j^{X_i, a} \sim S(\cdot | X_i, a)$, $j = 1, \dots, M$
- ③ Calculate the Monte-Carlo approximation of backed up values:

$$v_i = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^M \left[R_j^{X_i, a} + \gamma V_k(Y_j^{X_i, a}) \right], \quad i = 1, 2, \dots, N.$$

- ④ Solve the least-squares problem:

$$V_{k+1} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N (f(x_i) - v_i)^2$$

FINITE-TIME BOUNDS

Theorem 3:

Fix $\delta > 0$, $\epsilon > 0$, $\mathcal{F}$, ρ , μ . Assume that $\mathcal{V}$, the “capacity” of $\mathcal{F}$ (e.g. metric entropy) is finite. Assume that

$$\sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|f - Tg\|_{\rho, \mu} \leq \epsilon.$$

Then, it is possible to select N, M, K such that after K iterations of the sampling based FVI algorithm run with (μ, N, M)

$$\|V^* - V^{\pi_K}\|_{\rho, \rho} \leq \frac{4C^{1/\rho}}{(1-\gamma)^2} \epsilon$$

with probability at least $1 - \delta$. Further, N, M, K are polynomial in $\mathcal{V}$, $R_{\max}$, $1/\epsilon$, $\log |\mathcal{A}|$, $\log(1/\delta)$, $1/(1-\gamma)$.

Here C is a constant related to how quickly future state distributions can **concentrate** starting from ρ and relative to μ .

DEFINITION OF C

- μ – distribution used in the optimization step of the algorithm
- ρ – distribution in the performance bounds
- m -step (worst-case) concentration of future state distribution:

$$c(m) = \sup_{\pi_1, \dots, \pi_m, \|V\|=1} \frac{\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m} V}{\mu V}$$

- Average (discounted) concentration:

$$C = (1 - \gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c(m).$$

- Relation to **Lyapunov exponents**: If $\limsup_{m \rightarrow \infty} \frac{1}{m} \log^+ \|\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m}\| \leq 0$ holds for all non-stationary policies $\pi = (\pi_1, \dots, \pi_m, \dots)$ then the growth rate of $c(m)$ is polynomial: Hence, C is finite.

INTUITION: WHY DOES C ENTER THE BOUND?

Simple!

We are measuring performance w.r.t. ρ , whilst we are optimizing w.r.t. μ .

C relates ρ and μ : Optimization w.r.t. μ is not a good idea if future state distributions (starting from ρ) can concentrate “away from μ ”.

How to select μ ? (..and ρ)

PROOF – MAIN STEPS

- Single-iteration PAC Bound
- L^P bounds for AVI
- Putting it all together

WANT TO KNOW MORE?

⇒ Come to the poster!

THE SINGLE-SAMPLE VARIANT

- Imagine that it is **expensive** to generate the samples (e.g. expensive simulation). Why not generate a single set of samples (initially), save it and use it throughout all the iterations?
- Problem:** In the previous result one bounds

$$\mathbb{P}(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon | D_k),$$

where D_k is the sample used up to iteration k . If $D_k = D$, V_{k+1} becomes measurable w.r.t. D_k and $\mathbb{P}(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon | D)$ degenerates.

- Question:** Will this method still work?
- Answer:** Yes!
- Idea:** Strengthen main lemma, use

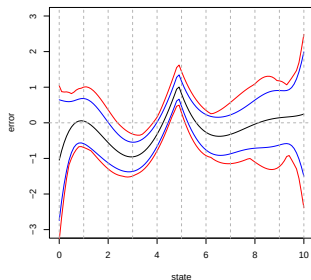
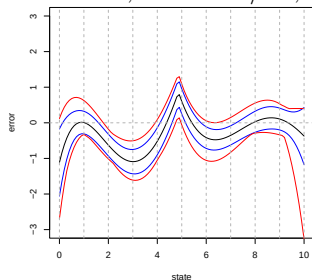
$$\mathbb{P}(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon) \leq \mathbb{P}(\sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|f - Tg\|_{p,\mu} > \epsilon)$$

HOW TO USE THE RESULT?

- **Problem 1:** V_K does not itself lead to a policy: Given V_K , we still need to compute a greedy policy w.r.t. V_K . How?
- **Good news:** Can do it using Monte-Carlo.
- This works (similar bounds to the previous ones).
- **Problem 2: How to select $\mathcal{F}$?** E.g. $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \mathcal{F}_s \dots$, increasingly richer parameterizations. Capacity and approximation power both grow. **Procedure:**
 - 1 Select target precision
 - 2 Select s large enough as a function of the target precision
 - 3 Select N, M, K as in the theorem.
- **Note:** Given a finite amount of data, the capacity and approximation power of $\mathcal{F}$ needs to be traded off.

ILLUSTRATION

- **Optimal replacement problem** (e.g. Rust, 1996)
- X_t – accumulated utilization of a durable ($X_t = 0$: new)
 - 'keep': $X_{t+1} - X_t \sim \exp(-\beta(X_{t+1} - X_t))$, $X_{t+1} - X_t \geq 0$
 - 'replace': $X_{t+1} \sim \exp(-\beta X_{t+1})$, $X_{t+1} \geq 0$
- $r(x, \text{'keep'}) = -4x$, $r(x, \text{'replace'}) = -30$
- Chebysev-polynomials with $d = 5$,
 $N = 100$, $M = 100/10$, $K = 10$, #runs= 50



CONCLUSIONS

- Continuous (or infinite, or very large) state space, generative model of the environment
- Main condition: Future state distributions do not concentrate fast
- Result: Error of multi/single-sample FVI bounded with high prob, in terms of approximation power and capacity of underlying function space, and the so-called concentration coefficient

FUTURE WORK

- Improve bound by introducing uneven distribution of error (and the error probability), exploit independence of samples between iterates to improve bound
- Sharpen definition of C or estimate it from data
- Rademacher-averages
- Data-dependent error bounds
- Is the single-sample variant indeed more sample-efficient?
- Consider other optimization criterion (regularization, ..)
- Single trajectory learning: Policy iteration (mostly done)

QUESTIONS?

???

PROOF: SINGLE-ITERATION ERROR BOUND

Lemma: Given V ,

$$v_i = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^M \left[R_j^{X_i, a} + \gamma V(Y_j^{X_i, a}) \right], \quad i = 1, 2, \dots, N,$$

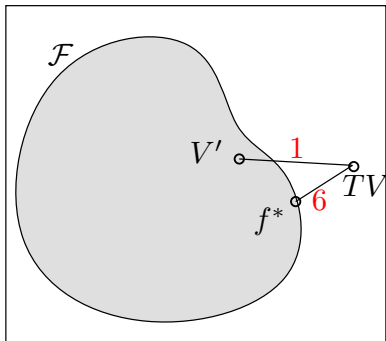
for N, M sufficiently large,

$$V' = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N (f(x_i) - v_i)^2$$

will not be much larger than the approximation error of TV with $\mathcal{F}$, with high probability.

A GRAPHICAL PROOF

$$(B(\mathcal{X}), \|\cdot\|_{p,\mu})$$



$$(\mathbb{R}^N, \|\cdot\|_p)$$

