A Statistical Learning approach to Approximate Dynamic Programming

Rémi Munos
Centre de Mathématiques Appliquées
Ecole Polytechnique

Csaba Szepesvari
Computer and Automation Research Institute of the Hungarian Academy of Sciences

Outline

1. L_p-norm error bounds in Approximate Dynamic Programming (Rémi)
2. PAC performance bounds in RL using Statistical Learning results (Csaba)

L_p-analysis for Approximate Dynamic Programming

Extend usual L_∞-norm analysis.

Benefits :

- Performance bounds for ADP in terms of approximation capacity of the function space
- Combine resuts from Statistical Learning theory, eg.
  - Complexity analysis, PAC performance bounds for RL, ...
  - Data-based function approximation (SVM, Kernels, ...)

Statistical Learning

L_p-analysis in DP $\quad -> \quad$ **RL and ADP analysis with function approximation**

Example : value iteration

Markov Decision Problem : state space X, action space A, transition kernel $P(dy|x,a)$, reward function $r(x,a)$.

Policy $\pi : X \to A$. **Value function** V^π = the performance of π (eg. discounted with $\gamma < 1$, infinite hozizon) :

$$V^\pi(x) = \mathbb{E}\big[\sum_{t \geq 0} \gamma^t r(x_t, a_t) \,|\, x_0 = x,\, a_t = \pi(x_t)\big]$$

The optimal value function $V^* = \max_\pi V^\pi$ is the fixed-point $V^* = \mathcal{T} V^*$ of the **Bellman operateur** :

$$\mathcal{T} f(x) = \max_{a \in A} \left[r(x,a) + \gamma \int P(dy|x,a) f(y) \right].$$

$\mathcal{T}$ is a contraction mapping in L_∞, thus V^* may be computed by value iteration $V_{n+1} = \mathcal{T} V_n$.

Approximate value iteration

Continuous (or large discrete) space –> need to use representations.

Approximate value iteration algorithm :

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n,$$

where $\mathcal{A}$ is an *approximation operator.*

Example : $\mathcal{F}$ is finite-dimensional linear subspace of a Hilbert space, and $\mathcal{A}$ the orthogonal projection (wrt. L_2) onto $\mathcal{F}$.

Properties :

– $\mathcal{T}$ is a contraction mapping in L_∞,

– $\mathcal{A}$ is non-expansive in L_2

Problem : we can't say anything about $\mathcal{A}\mathcal{T}$!

L_∞-analysis of AVI

Write $\epsilon_n = V_{n+1} - \mathcal{T} V_n$ the approximation error. Performance bound for AVI [Bertsekas & Tsitsiklis, 1996]

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{n\to\infty} ||\epsilon_n||_\infty .$$

Nice bound, but :

– How does the uniform error $||\epsilon_n||_\infty$ relates to the empirical error $\max_i |\epsilon_k(x_i)|$ (based on the data $\{x_i\}$) minimized by a real algorithm ?
– Well... actually, a real algorithm performs a L_p empirical minimization ! (except for exceptions... like *averagers* [Gordon, 1995]), ie.

$$V_{n+1} = \arg\min_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} |f(x_i) - \mathcal{T} V_n(x_i)|^p$$

(think about least squares regression, neural networks, SVM, kernel...)

L_p-analysis of AVI

Let μ be a distribution on X. Write : $||f||_{p,\mu} = \big(\int \mu(dx)|f(x)|^p\big)^{1/p}$.

Assume $P(\cdot|x,a)$ has a density wrt. μ (uniformly for $x \in X$, $a \in A$), ie, there exists $C(\mu) < \infty$ such that,

$$P(\cdot|x,a) \leq C(\mu)\mu(\cdot)$$

Then :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C(\mu)^{1/p} \limsup_{n\to\infty} ||\epsilon_n||_{p,\mu}.$$

Bound in terms of the L_p approximation errors.

Statistical Learning theory gives us (see Csaba's part) :

$$||\epsilon_n||_{p,\mu} \leq \Big[\frac{1}{K}\sum_{k=1}^{K} |\epsilon_n(x_k)|^p\Big]^{1/p} + E\big(K, \mathrm{VC}(\mathcal{F}), \ldots\big)$$

L_p-analysis of AVI (continued)

Assumption 1 : for all $x \in X$, $a \in A$,

$$P(\cdot|x,a) \leq C(\mu)\mu(\cdot)$$

Then :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C(\mu)^{1/p} \limsup_{n\to\infty} ||\epsilon_n||_{p,\mu}.$$

We recover the usual L_∞ bounds when $p \to \infty$.

Assumption 2 : for all sequence of policies $\pi_1, \pi_2, \ldots,$

$$(1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} \Pr\big(x_m \in dy \big| x_0 \sim \rho, \pi_1, \ldots, \pi_m\big) \leq C(\rho,\mu)\mu(dy).$$

Then :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C(\rho,\mu)^{1/p} \limsup_{n\to\infty} ||\epsilon_n||_{p,\mu}.$$

Other ADP algorithms

It seems that all usual L_∞ analysis in DP generalizes to L_p-norm.

– **Policy Iteration** [Munos, 2003]. Performance bound

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \le \frac{2\gamma}{(1-\gamma)^2} C(\mu)^{1/p} \varepsilon_{\mathcal{F}}.$$

in terms of the representation power of the value functions $\{V^{\pi_n}\}$ in the function space

$$\varepsilon_{\mathcal{F}} = \limsup_{n\to\infty} \inf_{f\in\mathcal{F}} ||V^{\pi_n} - f||_{p,\mu}.$$

– **Bellman residual minimization**

$$||V^* - V^\pi||_\infty \le \frac{2}{1-\gamma} C(\mu)^{1/p} ||\mathcal{T}V - V||_{p,\mu}.$$

Some insights : pointwise bounds

Assume that for $u, v \geq 0$ one has $u \leq Qv$, with Q a transition kernel.

– Then, $||u||_\infty \leq ||v||_\infty$ (since $||Q||_\infty = 1$)

– But also, if ρ and μ are probability distributions on X s.t. $\rho Q = \mu$, then

$$||u||_{p,\rho} \leq ||v||_{p,\mu}.$$

Indeed :

$$\begin{aligned} ||u||^p_{p,\rho} = \int \rho(dx)|u(x)|^p &\leq \int \rho(dx)\Big|\int Q(x,dy)v(y)\Big|^p \\ &\leq \int \rho(dx)\int Q(x,dy)|v(y)|^p \\ &= \int \mu(y)|v(y)|^p = ||v||^p_{p,\mu}, \end{aligned}$$

using Jensen's inequality.

Example : for $\rho = \mu$ stationary distribution for Q (ie. $\rho Q = \rho$).

Example : the Bellman residual bound

Let V a function on X. Let π be the greedy policy wrt. V, and V^π its performance. We have, pointwise,

$$V^* - V^\pi \quad \leq \quad \Big[(I - \gamma P^{\pi^*})^{-1} - (I - \gamma P^\pi)^{-1}\Big](\mathcal{T}V - V)$$

Thus :

In L_∞−norm, [Williams & Baird, 1993] :

$$||V^* - V^\pi||_\infty \leq \frac{2}{1-\gamma}||\mathcal{T}V - V||$$

In L_p−norm,

$$\begin{aligned} ||V^* - V^\pi||_\infty &\leq \frac{2}{(1-\gamma)} C(\mu)^{1/p} ||\mathcal{T}V - V||_{p,\mu}, \\ ||V^* - V^\pi||_{p,\rho} &\leq \frac{2}{(1-\gamma)} \big[C(\rho,\mu)\big]^{1/p} ||\mathcal{T}V - V||_{p,\mu}. \end{aligned}$$

Perspectives

– ADP analysis in the same L_p-norm as the one used in the approximation operation –> tight and useful bounds.

– Control generalization error

– Combine with results in approximation theory and statistical learning theory, eg. kernel methods in RKHS