Linear Programming for Large-Scale Markov Decision Problems

Yasin Abbasi-Yadkori[1]   Peter Bartlett[1 2]   Alan Malek[2]

[1]Queensland University of Technology
Brisbane, QLD, Australia

[2]University of California, Berkeley
Berkeley, CA

June 24th, 2014

Outline

1. Introduce MDPs and the Linear Program formulation
2. Algorithm
3. Oracle inequality
4. Experiments

Markov Decision Processes

A Markov Decision Process is specified by:

- State space $\mathcal{X} = \{1, \ldots, X\}$
- Action space $\mathcal{A} = \{1 \ldots, A\}$
- Transition Kernel $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$
- Loss function $\ell : \mathcal{X} \times \mathcal{A} \to [0, 1]$

Let P^{π} be the state transition kernel under policy $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$.
Our goal is to choose π to minimize the average loss when X and A are very large.
Aim for optimality *within a restricted family of policies.*

Linear Program Formulation

- LP formulation (Manne 1960):

$$\max_{\lambda,h} \lambda \,, \qquad (1)$$
$$\text{s.t.} \quad B^\top(\lambda \mathbf{1} + h) \leq \ell + P^\top h \,,$$

  where $B \in \{0,1\}^{(X \times XA)}$ is the marginalization matrix.
- Primal variables: h is the cost-to-go, λ is the average cost
- Dual:

$$\min_{\mu \in \mathbb{R}^{XA}} \ell^\top \mu \,, \qquad (2)$$
$$v\text{s.t.} \quad \mathbf{1}^\top \mu = 1,\ \mu \geq \mathbf{0},\ (P - B)\mu = \mathbf{0} \ .$$

- Define policy via $\pi(a|x) \propto \mu_{(x,a)}$.
- Dual variables: μ is a stationary distribution over $\mathcal{X} \times \mathcal{A}$
- Still a problem when X, A very large

The Dual ALP

- Feature matrix $\Phi \in \mathbb{R}^{XA \times d}$; constrain $\mu = \Phi\theta$

$$\min_{\mu \in \mathbb{R}^{XA}} \ell^\top \Phi\theta \,, \tag{3}$$
$$\text{s.t.} \quad \mathbf{1}^\top \Phi\theta = 1,\ \Phi\theta \geq \mathbf{0},\ (P - B)^\top \Phi\theta = \mathbf{0} \,.$$

- $[\cdot]_+$ is positive part
- Define policy via $\pi_\theta(a|x) \propto [(\Phi\theta)(x,a)]_+$,
- μ_θ is the stationary distribution of P^{π_θ}
- $\mu_\theta \approx \Phi\theta$
- $\ell^\top \mu_\theta$ is the average loss of policy π_θ
- Want to compete with $\min_\theta \ell^\top \mu_\theta$

Reducing Constraints

- Still intractable: d-dimensional problem but $O(XA)$ constraints
- Form the convex cost function:

$$\begin{aligned} c(\theta) &= \ell^\top \Phi\theta + \|[\Phi\theta]_-\|_1 + \left\|(P-B)^\top \Phi\theta\right\|_1 \\ &= \ell^\top \Phi\theta + \sum_{(x,a)} \left|[\Phi_{(x,a),:}\theta]_-\right| + \sum_{x'} \left|(\Phi\theta)^\top (P-B)_{:,x'}\right| \end{aligned}$$

- Sample $(x_t, a_t) \sim q_1$ and $y_t \sim q_2$
- Unbiased subgradient estimate:

$$\begin{aligned} g_t(\theta) =& \ell^\top \Phi - \frac{\Phi_{(x_t,a_t),:}}{q_1(x_t, a_t)} \mathbb{I}_{\{\Phi_{(x_t,a_t),:}\theta<0\}} \\ &+ \frac{(\Phi^\top (P-B)_{:,y_t})^\top}{q_2(y_t)} \operatorname{sgn}\left((\Phi\theta)^\top (P-B)_{:,y_t}\right) \end{aligned} \tag{4}$$

The Stochastic Subgradient Method for MDPs

Input: Constants $S, H > 0$, number of rounds T.
Let Π_Θ be the Euclidean projection onto S-radius 2-norm ball.
Initialize $\theta_1 \propto 1$.
for $t := 1, 2, \dots, T$ **do**
  Sample $(x_t, a_t) \sim q_1$ and $x'_t \sim q_2$.
  Compute subgradient estimate g_t
  Update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t g_t)$.
end for
$\widehat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \theta_t$.
Return policy $\pi_{\widehat{\theta}_T}$.

Theorem

Given some $\epsilon > 0$, the $\widehat{\theta}_T$ produced by the stochastic subgradient method after $T = 1/\epsilon^4$ steps satisfies

$$\ell^\top \mu_{\widehat{\theta}_T} \leq \min_{\theta \in \Theta} \left(\ell^\top \mu_\theta + \frac{V(\theta)}{\epsilon} + O(\epsilon) \right)$$

with probability at least $1 - \delta$, where $V = O(V_1 + V_2)$ is a violation function defined by

$$V_1(\theta) = \|[\Phi\theta]_-\|_1$$
$$V_2(\theta) = \left\| (P - B)^\top \Phi\theta \right\|_1 .$$

The big-O notation hides polynomials in S, d, C_1, C_2, and $\log(1/\delta)$.

Comparison with previous techniques

- We bound performance of found policy directly (not through J)
- Previous bounds were of the form $\inf_\theta \|J^* - \Psi\theta\|$
- Our bounds: performance w.r.t. best in class w.o. near optimality of class
- No knowledge of optimal policy assumed
- First method to make approximations in the dual

Discussion

- Can remove the awkward $V(\theta)/\epsilon + O(\epsilon)$ by taking a grid of ϵ
- Recall

$$C_1 = \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_1(x,a)}, \qquad C_2 = \max_{x\in\mathcal{X}} \frac{\left\|(P-B)_{:,x}^\top \Phi\right\|}{q_2(x)}$$

- We also pick Φ and q_1, so we can make C_1 small
- Making C_2 may require knowledge of P (such as sparsity or some stability assumption)
- Natural selection: state aggregation

Comparison with Constraint Sampling

- Use the constraint sampling of (de Farias and Van Roy, 2004)
- Must assume feasibility
- Need a vector $v(x) \geq |(P - B)^\top \Phi\theta|$ as envelope to constraint violations
- Bound includes $||v(x)||_1$; could be very large
- Requires specific knowledge about problem

Analysis

- Assume fast mixing: for every policy π, $\exists \tau(\pi) > 0$ s.t. $\forall d, d' \in \triangle_{\mathcal{X}}$,

$$\left\| dP^{\pi} - d' P^{\pi} \right\|_1 \leq e^{-1/\tau(\pi)} \left\| d - d' \right\|_1$$

- Define

$$C_1 = \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{\left\| \Phi_{(x,a),:} \right\|}{q_1(x,a)}, \qquad C_2 = \max_{x\in\mathcal{X}} \frac{\left\| (P-B)_{:,x}^{\top} \Phi \right\|}{q_2(x)}.$$

- The proof has three main parts
    1. $V_1(\theta) \leq \epsilon_1$ and $V_2(\theta) \leq \epsilon_2 \Rightarrow \left\| \mu_\theta - \Phi\theta \right\|_1 \leq O(\epsilon_1 + \epsilon_2)$
    2. Bounding gradient of $c(\theta)$; checking it is unbiased
    3. Applying stochastic gradient descent theorem: $\ell^{\top} \Phi \widehat{\theta} \leq \min_{\theta\in\Theta} c(\theta) + O(\epsilon)$

Proof part 1

Lemma

Let $u \in \mathbb{R}^{XA}$ be a vector with

$$\mathbf{1}^\top u = 1, \|u\| \leq 1 + \epsilon_1, \left\| u^\top (P - B) \right\|_1 \leq \epsilon_2$$

For the stationary distribution μ_u of policy $u^+ = [u]_+ / \|[u]_+\|_1$, we have

$$\|\mu_u - u\|_1 \leq \tau(\mu_u) \log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \ .$$

Proof:

- Two bounds give $\left\|(P - B)^\top u^+\right\|_1 \leq 2\epsilon_1 + \epsilon_2 := \epsilon'$
- Also, $\|u^+ - u\|_1 \leq 2\epsilon_1$
- Define $M^{u^+} \in \mathbb{R}^{X \times XA}$ as the matrix that encodes policy u^+, e.g. $M^{u^+} P = P^{u^+}$

Proof (continued):

- Let $\mu_0 = u^+$, $\mu_t^\top = \mu_{t-1}^\top P M^{u^+}$, $v_t = \mu_t^\top (P - B) = v_{t-1} M^{u^+} P$
- μ_t is the state-action distribution after running the policy for t steps
- By previous bound, $\|v_0\|_1 \leq \epsilon' \Rightarrow \|v_t\|_1 \leq \epsilon'$
- $\mu_t^\top = \mu_{t-1}^\top P M^{u^+} = (\mu_{t-1}^\top B + v_{t-1}) M^{u^+} = \mu_{t-1}^\top + v_{t-1} M^{u^+}$
- Telescoping: $\mu_k^\top = \mu_0^\top + \sum_{t=0}^{k} v_t M^{u^+}$
- Thus, $\|\mu_k - u^+\|_1 \leq k\epsilon$
- By mixing assumption: $\|\mu_k - \mu_u\|_1 \leq e^{-1/\tau(u^+)}$
- Take $k = \tau(u^+) \log(1/\epsilon')$ and use triangle inequality

Applying SGD theorem

Theorem (Lemma 3.1 of (Flaxman et al., 2005))

Assume we have

- *Convex set* $\mathcal{Z} \subseteq B_2(Z, 0)$ *and* $(f_t)_{t=1,2,\ldots,T}$ *convex functions on* $\mathcal{Z}$.
- *Gradient estimates* f'_t *with* $\mathbb{E}\left[[]\, f'_t | z_t\right] = \nabla f(z_t)$ *and bound* $\|f'_t\|_2 \leq F$
- *Sample Path* $z_1 = 0$ *and* $z_{t+1} = \Pi_{\mathcal{Z}}(z_t - \eta f'_t)$ *(*$\Pi_{\mathcal{Z}}$ *Euclidean projection)*

Then, for $\eta = Z/(F\sqrt{T})$ *and any* $\delta \in (0, 1)$*, the following holds with probability at least* $1 - \delta$*:*

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z \in \mathcal{Z}} \sum_{t=1}^{T} f_t(z) \leq ZF\sqrt{T} + \sqrt{(1 + 4Z^2T)\left(2\log\frac{1}{\delta} + d\log\left(1 + \frac{Z^2T}{d}\right)\right)} . \tag{5}$$

checking conditions of theorem

Recall gradient: for $(x_t, a_t) \sim q_1$ and $y_t \sim q_2$,

$$
\begin{aligned}
g_t(\theta) = & \ell^\top \Phi - H \frac{\Phi_{(x_t,a_t),:}}{q_1(x_t, a_t)} \mathbb{I}_{\{\Phi_{(x_t,a_t),:}\theta < 0\}} \\
& + H \frac{(P - B)^\top_{:,y_t} \Phi}{q_2(y_t)} \operatorname{sgn}\left((\Phi\theta)^\top (P - B)_{:,y_t} \right).
\end{aligned}
$$

We can bound

$$
\begin{aligned}
\|g_t(\theta)\|_2 & \le \left\| \ell^\top \Phi \right\|_2 + H \frac{\left\| \Phi_{(x_t,a_t),:} \right\|_2}{q_1(x_t, a_t)} + \frac{\left\| (P - B)^\top_{:,y_t} \Phi \right\|_2}{q_2(y_t)} \\
& \le \sqrt{d} + H(C_1 + C_2) := F.
\end{aligned}
$$

and $\mathbb{E}\left[[]\, g_t(\theta)\right] = \nabla c(\theta)$.

proof conclusion

The SGD theorem gives us:

$$\ell^\top \Phi \widehat{\theta}_T + H(V_1(\widehat{\theta}) + V_2(\widehat{\theta})) \leq \ell^\top \Phi \theta^* \leq H(V_1(\theta^*) + V_2(\theta^*)) + b_T$$

where b_T is the regret bound from the theorem:

$$b_T = \frac{SF}{\sqrt{T}} + \sqrt{\frac{1+4S^2T}{T^2}\left(2\log(\frac{1}{\delta}) + d\log(\frac{d+S^2T}{d})\right)}.$$

We take

$$V_1(\widehat{\theta}), V_2(\widehat{\theta}) \leq \frac{1}{H}(2(1+S) + HV_1(\theta^*) + HV_2(\theta^*) + b_T) := \epsilon'$$

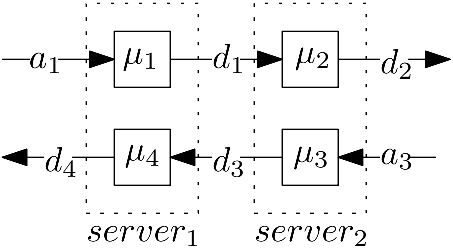
Applying the lemma twice:

$$\ell^\top \mu_{\widehat{\theta}_T} - \ell^\top \mu_{\theta^*} \leq HV_1(\theta^*) + HV_2(\theta^*) + b_T + \tau(\mu_{\widehat{\theta}}) \log(1/\epsilon') 3\epsilon' + 3\epsilon' \\ + \tau(\mu_{\theta*}) \log(1/V(\theta^*))(2V_1(\theta^*) + V_2(\theta^*)) + 3V_1(\theta)$$

Since $b_T = O(H/\sqrt{T}$, taking $H = 1/\epsilon$ and $T = 1/\epsilon^4$ yields:

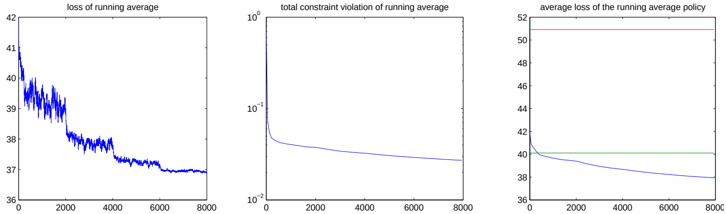
$$\ell^\top \mu_{\widehat{\theta}_T} - \ell^\top \mu_{\theta^*} \leq \frac{1}{\epsilon}(V_1(\theta^*) + V_1(\theta^*)) + O(\epsilon).$$

Queueing network example (Rybko-Stolyar)



- Customers arrive at μ_1/μ_3 then move to μ_2/μ_4
- Server 1 processes μ_1 or μ_4, server 2 processes μ_2 or μ_3
- Features: indicators of sub-blocks in state-action space, stationary distribution of LONGER and LBSF heuristics
- Loss is the total queue size
- $a_1 = a_3 = .08$, $d_1 = d_2 = .12$, and $d_3 = d_4 = .28$, $X = 902500$

Results



- The left plot: linear objective of the running average, i.e. $\ell^\top \Phi \widehat{\theta}_t$.
- The center plot: sum of the two constraint violations of $\widehat{\theta}_t$
- The right plot: $\ell^\top \mu_{\widehat{\theta}_t}$. The two horizontal lines correspond to the loss of two heuristics, LONGER and LBFS.

Conclusion

- Presented an algorithm to solve average-cost large-scale MDPs
  - Restricted the dual LP to a subspace to reduce dimension
  - Used Stochastic Gradient Descent to sample constraints
- Presented oracle inequality guaranteeing we perform well w.r.t. best policy in the subspace.
- Demonstrated algorithm on a queueing network
- Visit us at poster T75

Bibliography

D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29, 2004.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, 2005.