Convex MDPs

1 Preliminaries

The convex MDP problem can be written as

$$\min_{d^\pi \in \mathcal{K}} f(d_\pi),$$

where $f : \Delta(\mathcal{A} \mid \mathcal{S}) \mapsto \mathbb{R}$ is a convex function, $d_\pi \in \Delta(\mathcal{A} \mid \mathcal{S})$ is the state-action stationary distribution induced by policy π, and $\mathcal{K}$ is the set consisting of all the admissible occupancy measures. The objective can be equivalently written using Fenchel duality as

$$\begin{aligned}
\min_{d_\pi \in \mathcal{K}} f(d_\pi) &\overset{\text{(i)}}{=} \min_{d^\pi \in \mathcal{K}} f^{\star\star}(d_\pi) \\
&\overset{\text{(ii)}}{=} \min_{d^\pi \in \mathcal{K}} \max_{\lambda \in \Lambda} (\langle \lambda, d_\pi \rangle - f^\star(\lambda)) \\
&\overset{\text{(iii)}}{=} \max_{\lambda \in \Lambda} \min_{d^\pi \in \mathcal{K}} (\langle \lambda, d_\pi \rangle - f^\star(\lambda)).
\end{aligned}$$

(i) follows that the biconjugate $f^{\star\star} := f^\star(f) = f$ if f is convex and lower semicontinuous. (ii) follows the definition of Fenchel conjugate: $f^\star(x) := \sup_y x \cdot y - f(y)$, and Λ is the closure of the (sub-)gradient space $\{\partial f(d_\pi) \mid d_\pi \in \mathcal{K}\}$. (iii) follows from the minmax theorem.

Therefore, we can define the Lagrangian as $\mathcal{L}(d_\pi, \lambda) := \langle \lambda, d_\pi \rangle - f^\star(\lambda)$.

2 Linear Function Approximation

We can tackle the above formulation with a primal-dual learning framework. Suppose we can run the algorithm for K iterations, and for each iteration $k \in [K]$, we have that

$$\begin{aligned}
d_\pi^k &= \arg\min_{d_\pi \in \mathcal{K}} \left\langle \lambda^k, d_\pi \right\rangle, \\
\lambda^k &= \arg\max_{\lambda \in \Lambda} \left\langle \lambda, \sum_{j=1}^k d_\pi^j \right\rangle - k \cdot f^\star(\lambda).
\end{aligned}$$

Dual-Variable Parameterization. Assume that the dual variable λ can be parameterized by a linear function. That is, $\lambda_\theta = \langle \phi, \theta \rangle$ where $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is the feature map, and $\theta : \mathbb{R}^d$ is the learnable parameter.

Given the above parameterization, we can rewrite the dual update as follows. Note that in the discounted setting, $d_\pi(s, a) = (1 - \gamma)\, \mathbb{E}_\pi[\sum_{t=1}^\infty \gamma^t \mathbb{1}(s_t = s, a_t = a)]$. Then, we have that

$$\begin{aligned}
\lambda^k &= \arg\max_{\lambda \in \Lambda} \left\langle \lambda, \sum_{j=1}^k d_\pi^j \right\rangle - k \cdot f^\star(\lambda) \\
&= \arg\max_{\theta \in \mathbb{R}^d} \sum_{s,a} \lambda_\theta(s, a) \cdot \sum_{j=1}^k d_\pi^j(s, a) - k \cdot f^\star(\langle \phi, \theta \rangle)
\end{aligned}$$

$$
\begin{aligned}
&= \arg\max_{\theta\in\mathbb{R}^d} \sum_{j=1}^{k} \sum_{s,a} \lambda_\theta(s,a) \cdot d_\pi^j(s,a) - k \cdot f^\star(\langle \phi, \theta \rangle) \\
&= \arg\max_{\theta\in\mathbb{R}^d} \sum_{j=1}^{k} \left\langle \sum_{s,a} d_\pi^j(s,a) \cdot \phi(s,a), \theta \right\rangle - k \cdot f^\star(\langle \phi, \theta \rangle) \\
&= \arg\max_{\theta\in\mathbb{R}^d} (1-\gamma) \sum_{j=1}^{k} \left\langle \sum_{s,a} \mathbb{E}_{\pi_j} \left[\sum_{t=1}^{\infty} \gamma^t \mathbb{1}(s_t = s, a_t = a) \right] \cdot \phi(s,a), \theta \right\rangle - k \cdot f^\star(\langle \phi, \theta \rangle) \\
&= \arg\max_{\theta\in\mathbb{R}^d} (1-\gamma) \sum_{j=1}^{k} \left\langle \mathbb{E}_{\pi_j} \left[\sum_{t=1}^{\infty} \gamma^t \phi(s_t, a_t) \right], \theta \right\rangle - k \cdot f^\star(\langle \phi, \theta \rangle)
\end{aligned}
$$

We can then use sampling-based approach to estimate $\mathbb{E}_{\pi_j}[\sum_{t=1}^{\infty} \gamma^t \phi(s_t, a_t)]$ using π_j where $\pi_j(s,a) = \frac{d_\pi^j(s,a)}{\sum_{a'} d_\pi^j(s,a')}$.

3 Exploration

References

Contents of Appendix