Planning in Large-Scale Markov Decision Problems

Yasin Abbasi-Yadkori [1], Peter Bartlett [1,2], Xi Chen [3], **Alan Malek**[2]

[1]Queensland University of Technology

[2]University of California, Berkeley

[3]NYU Stern School of Business

June 27th, 2016

Map

- Problem: planning in an MDP with large state space
- Goal: find near-optimal policy in low dimensional family
- Average Cost
  - Parameterize dual LP
  - Obtain “agnostic” guarantee
  - Queueing network
- KL-cost
  - Exploit Linearly Solvable MDPs
  - Parameterize log of loss function
  - Crowdsourcing

Motivation

- Markov decision process: modeling sequential decisions
- Decouple learning and planning, e.g. [**?**]
- E.g. queueing network, robot planning
- Can solve for small state spaces
- Large state spaces: “curse of dimensionality”

Outline

MDPs

A Markov Decision Process is specified by:

- State space $\mathcal{X} = \{1, \ldots, X\}$
- Action space $\mathcal{A}$
- Transition Kernel $\boldsymbol{P} : \mathcal{X} \times \mathcal{A} \to \triangle_{\mathcal{X}}$
- Loss function $\ell : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$

Planning problem:

- Find policy $\pi : \mathcal{X} \to \triangle_{\mathcal{A}}$ to minimize value function

$$\boldsymbol{J}_\pi^\gamma(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(X_t, \pi) \middle| X_0 = x\right] \qquad \text{(discounted cost)}$$

$$\boldsymbol{J}_\pi(x) = \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{t=0}^{n} \ell(X_t, \pi) \middle| X_0 = x\right] \qquad \text{(average cost)}$$

$$\boldsymbol{J}_\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \ell(X_t, \pi) \middle| X_0 = x\right] \qquad \text{(total cost)}$$

Discounted Cost

- Earliest attempt to solve planning problem [**?**]
- Define the *Bellman Operator* L^{γ}

$$(L^{\gamma}J)(x) = \min_{a} \ell(x,a) + \gamma \mathbb{E}\left[J(X_1) | X_0 = x, A_0 = a\right]$$

- L^{γ} is an L_{∞} contraction: $||J - J'||_{\infty} \leq \gamma ||L^{\gamma}J - L^{\gamma}J'||_{\infty}$
- If $\gamma < 1$, there is a unique solution $J^{*} = \lim_{k\to\infty} {L^{\gamma}}^{k} J$
- J is optimal iff $L^{\gamma}J = J$
- Optimal policy is greedy:

$$\pi^{*}(a|x) = \mathbb{I}\{a = \arg\min_{a} \ell(x,a) + \gamma \mathbb{E}\left[J^{*}(X_1) | X_0 = x, A_0 = a\right]\}$$

- Unfortunately, Bellman iteration is $O(X^2 A)$

Average Cost

- More complicated: L^1 not a contraction
- Need to measure w.r.t. the average cost, $\lambda \in \mathbb{R}$ and rely on Markov Chain stationarity
- Define the differential cost function $h \in \mathbb{R}^X$ and Bellman operaton

$$Lh(x) := \min_{a \in \mathcal{A}} \left[\ell(x,a) + \sum_y P(y|x,a)h(y) \right]$$

- Bellman optimality: $Lh = h + \lambda \mathbf{1}$
- [?] Thm. 8.4.1: Suppose λ and h satisfy $Lh \geq h + \lambda \mathbf{1}$. Then $\lambda \leq \lambda^*$.
- Motivates exact average-cost LP [?]

$$\max_{\lambda, h} \lambda \,,$$
$$\text{s.t.} \quad h + \lambda \mathbf{1} \leq Lh$$

- Always has a solution [Thm. 8.4.3] for recurrent chains

Linear Programming Formulation

- Define $B \in \mathbb{R}^{XA \times X}$ by $B_{(x,a),y} = \{x = y\}$. We can write:

$$h + \lambda \mathbf{1} \leq Lh = \min_a \left[\ell(x,a) + \sum_y P(y|x,a) h(y) \right]$$

$$\Leftrightarrow$$

$$h(x) + \lambda \leq \ell(x,a) + \sum_y P(y|x,a) h(y) \quad \forall x, a$$

$$\Leftrightarrow$$

$$B(\lambda \mathbf{1} + h) \leq \ell + Ph$$

- Average-cost LP equivalent too

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

- Dimension X, number of constraints $O(XA)$. Intractable!

Large state space

- Parametric class of value functions J_θ or h_θ for $\theta \in \Theta \subset \mathbb{R}^d$
- For any value function J or h, there is a greedy policy π_J or π_h (the argmax in L^γ)

Problem (Large-State Planning Problem)

Assume:

- $\mathcal{X}$ *is very large*
- *We have entrywise access to* P *and* ℓ
- *Goal: find* θ *to minimize*

$$\underbrace{J_{\underbrace{\pi_{J_\theta}}_{\text{Greedy policy using value function } J_\theta}}}_{\text{Value function of running this greedy policy}}$$

Approximate solutions

- Approximate Dynamic programming
  - Attempt to minimize θ directly, e.g. OGD
  - Approximate policy iteration; e.g. LSPI [**?**]
- Approximate Linear program
  - For a feature matrix $\Psi \in \mathbb{R}^{X \times d}$ for some $d \ll X$

$$\max_{\lambda, h} \lambda \,,$$
$$\text{s.t.} \quad B(\lambda \mathbf{1} + \Psi\theta) \leq \ell + \boldsymbol{P}\Psi\theta$$

Previous work

- Approximate Dynamic Programming (linear approximation of the value function): [**??**]
- Approximate Linear Programming: (approximately solving LP) [**?????????**].
- Solving LMDPs (with no theoretical guarantees): [**?**] and [**??**]

Previous work: average cost

- Average cost suffers from a new set of problems
  - State-relevance vectors are not in average cost LP
  - Lyapunov function ideas are hard to extend
- First algorithms studied: [**?**]
  - Awkward: had one LP to estimate λ and a second to estimate h^*
  - Requires feasibility
- [**?**] first looked at minimizing the dual LP, but provided no performance bounds (described versions of DP algorithms in the dual space)

Outline

The Dual

- Recall: average cost LP

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

- Dual is

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

- μ is a stationary distribution of P^π for $\pi(a|x) \propto \mu_{x,a}$

The Dual ALP

- Feature matrix $\Phi \in \mathbb{R}^{XA \times d}$; constrain $\mu = \Phi\theta$, $\theta \in B_2(0, S)$

$$\min_{\theta \in B_2(0,S)} \ell^\top \Phi\theta\,,$$
$$\text{s.t.} \quad \mathbf{1}^\top \Phi\theta = 1,\ \Phi\theta \geq \mathbf{0},\ (\boldsymbol{P} - \boldsymbol{B})^\top \Phi\theta = \mathbf{0}\,.$$

- Policy $\pi_\theta(a|x) \propto [(\Phi\theta)(x, a)]_+$
- μ_θ is the stationary distribution of P^{π_θ}
- Intuition: $\mu_\theta \approx \Phi\theta$, so $\min_\theta \ell^\top \mu_\theta \approx \min_\theta \ell^\top \Phi\theta$

Interpretation

- $\Phi\theta$ is an approximate stationary distributions
- The primal of the dual ALP is:

$$\min_{\lambda, h} \lambda$$
$$\text{s.t.} \quad \Phi^\intercal(\ell + (P - B)h - \lambda \mathbf{1}) \in \Phi^+$$

where $\Phi^+ = \{x \in \mathbb{R}^d | \exists \nu \geq 0 \text{ s.t. } x = \Phi^\intercal \nu\}$

- Similar to weighted constraint aggregation

Reducing Constraints

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

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

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

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

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 $t := 1, 2, \ldots, T$ **do**
  Sample $(x_t, a_t) \sim q_1$ and $y_t \sim q_2$.
  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

$$\ell^\intercal \mu_{\widehat{\theta}_T} \leq \min_{\theta \in B(0,S)} \left(\ell^\intercal \mu_\theta + \frac{V(\theta)}{\epsilon} + O(\epsilon) \right)$$

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

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

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

Discussion

- Previous bounds were of the form $\inf_{\boldsymbol{r}} \|h^* - \Psi \boldsymbol{r}\|$
- Can remove the awkward $V(\theta)/\epsilon + O(\epsilon)$ by taking a grid of ϵ
- Constants:

$$C_1 = \underbrace{\max_{x,a} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_1(x,a)}}_{\text{Control via } \Phi \text{ and } q_1}, \qquad C_2 = \underbrace{\max_{x} \frac{\|\Phi^{\top}(\boldsymbol{P} - \boldsymbol{B})_{:,x}\|}{q_2(x)}}_{\text{control via structure of } \boldsymbol{P}}$$

- $V(\theta^*)$ measures the difficulty of the problem
- Assume fast mixing: for every policy π, $\exists \tau(\pi) > 0$ s.t. $\forall d, d' \in \triangle_{\mathcal{X}}$,

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

Proof Outline

- First,

$$\mathbf{1}^\mathsf{T}\mu = 1, \|\mu\|_1 \le 1 + \epsilon_1, \|\mu^\mathsf{T}(\boldsymbol{P} - \boldsymbol{B})\|_1 \le \epsilon_2 \Rightarrow$$
$$\left\|\mu_{\pi_{\mu^+}} - \mu\right\|_1 \le \tau(\mu_\mu)\log(1/\epsilon_1)O(\epsilon_1 + \epsilon_2)$$

- SGD theorem

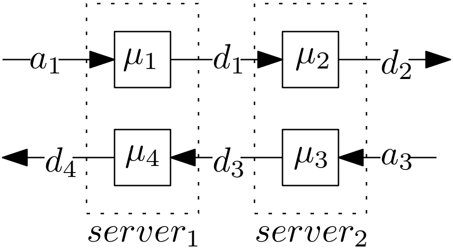
$$\ell^\mathsf{T}\Phi\widehat{\theta}_T + H(V(\widehat{\theta})) \le \ell^\mathsf{T}\Phi\theta^* + H(V(\theta^*)) + O\left(\frac{SH(C_1 + C_2)}{\sqrt{T}}\right)$$

- Use $\Phi\theta \approx \mu_\theta$

$$\ell^\mathsf{T}\mu_{\widehat{\theta}_T} - \ell^\mathsf{T}\mu_{\theta^*} \le \quad HO(V_1(\theta^*) + V_2(\theta^*)) + O\left(\frac{H(C_1 + C_2)}{\sqrt{T}}\right)$$

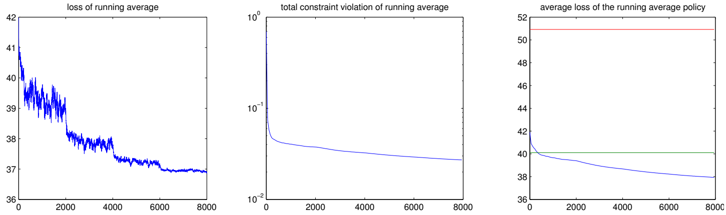
- Optimize H and T

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 $d_3 = d_4 = .28$, $X = 902500$

Experiments: 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.

Dual ALP Summary

- 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

KL-cost

- Introduced in [**?**]
- $\mathcal{A} = \triangle_{\mathcal{X}}$: we are playing polcies
- Loss: $\ell(x, \boldsymbol{P}) = q(x) + D_{KL}(\underbrace{\boldsymbol{P}}_{\text{Learner's action}} \,||\, \underbrace{\boldsymbol{P}_0}_{\text{Base dynamics}}(\cdot|x))$
- Infinite loss unless $\boldsymbol{P} \ll \boldsymbol{P}_0$
- Terminal state z: $q(z) = 0$ and $\boldsymbol{P}_0(z|z) = 1$
- Obejective is total cost:

$$\boldsymbol{J}_{\boldsymbol{P}}(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \ell(X_t, \boldsymbol{P}) \middle| X_0 = x\right]$$

Details

- $\mathcal{L}J_P(x) = \min_P \left[\ell(x,P) + \sum_y P(y|x) J_P(y)\right]$
- Greedy action is:

$$P_J(\cdot|x) = \arg\min_{p \in \triangle_{\mathcal{X}}} \sum_y p(y) \log \frac{p(y)}{P_0(y|x)e^{-J(y)}} = \frac{P_0(x'|x)e^{-J(x')}}{Z(x)}$$

  with $Z = P_0 e^{-J}$
- This implies $\mathcal{L}J = q - \log(Z)$
- Value function is the solution to:

$$J = \mathcal{L}J \Leftrightarrow J = q - \log(P_0 e^{-J})$$

- Exponentiating: $\mathcal{L}J = J \Leftrightarrow e^{-q} P_0 e^{-J} = e^{-J}$

Parameterizing J_θ

- Previous approaches: $J_\theta = \Psi\theta$
- Instead: $J_\theta = -\log(\Psi\theta)$
- Surrogate optimization:

$$\min_\theta c^\mathsf{T} J_\theta + \underbrace{\|LJ_\theta - J_\theta\|}_{\text{Bellman error}} \tag{1}$$

- $\|LJ_\theta - J_\theta\|$ not convex in θ, but

$$e^{-\max\{LJ_\theta, J_\theta\}} \|LJ_\theta - J_\theta\| \leq \left\| e^{-LJ_\theta} - e^{-J_\theta} \right\|$$

- Plugging $\Psi\theta = e^{-J\theta}$ into (**??**):

$$\min_\theta -c^\mathsf{T} \log(\Psi\theta) + || e^{-q} P_0 \Psi\theta - \Psi\theta||$$

Parameterizing J_θ

- Previous approaches: $J_\theta = \Psi\theta$
- Instead: $J_\theta = -\log(\Psi\theta)$
- Surrogate optimization:

$$\min_\theta c^\mathsf{T} J_\theta + \underbrace{\|LJ_\theta - J_\theta\|}_{\text{Bellman error}} \tag{1}$$

- $\|LJ_\theta - J_\theta\|$ not convex in θ, but

$$e^{-\max\{LJ_\theta, J_\theta\}} \|LJ_\theta - J_\theta\| \leq \left\| e^{-LJ_\theta} - e^{-J_\theta} \right\|$$

- Plugging $\Psi\theta = e^{-J\theta}$ into (**??**):

$$\min_\theta -c^\mathsf{T} \log(\Psi\theta) + || \underbrace{e^{-q} P_0}_{\substack{\text{Bellman} \\ \text{operator}}} \Psi\theta - \Psi\theta ||$$

Our algorithm

- Let $\mathcal{T}$ be the set of trajectories with $x_1 \sim c$ with distribution $Q(\cdot)$
- Recall relaxed optimization:

$$\min_{\theta} -c^{\intercal} \log(\Psi\theta) + \left\| e^{-q} \boldsymbol{P}_0 \Psi\theta - \Psi\theta \right\|_Q$$

- Optimization is equal to:

$$\min_{\theta} -c^{\intercal} \log(\Psi\theta) + \sum_{T \in \mathcal{T}} Q(T) \sum_{x \in \mathcal{T}} \left| e^{-q(x)} \boldsymbol{P}_0 \Psi\theta(x) - \Psi\theta(x) \right|$$

- Use stochastic gradient descent by sampling trajectories

Theorem

Let $\widehat{\theta}$ be an ϵ-optimal solution returned by SGD. Then,

$$J_{P_{J_{\widehat{\theta}}}}(x_1) \le \inf_{\theta \in \Theta} \left\{ J_{P_{J_\theta}}(x_1) + \mathcal{E}(J_\theta) \right\} + \epsilon$$

$$+ \underbrace{\left\| P_{J_{\widehat{\theta}}} - Q \right\|_1}_{\text{Off-policy error}} \max_{T \in \mathcal{T}} \sum_{x \in T} \left| J_{\widehat{\theta}}(x) - LJ_{\widehat{\theta}}(x) \right|$$

Penalty function:

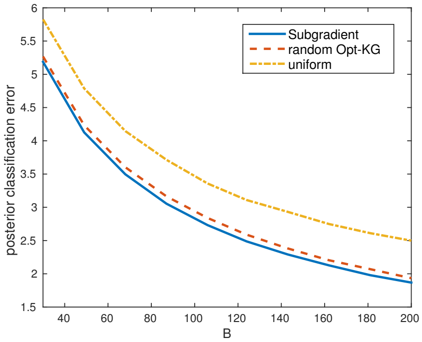
$$\mathcal{E}(J_\theta) = \sum_{T \in \mathcal{T}} \sum_{x \in T} \left(Q(T) e^{-\min(J_\theta, LJ_\theta)} + P_{J_\theta}(T) \right) \underbrace{\left| J_\theta(x) - LJ_\theta(x) \right|}_{\substack{\text{Small if } J_\theta \text{ is} \\ \text{close to the} \\ \text{optimal value}}}$$

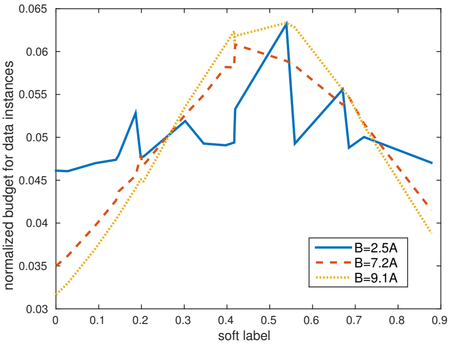
Proof outline of main theorem

- $\left| J_{P_{J_{\theta^*}}}(x_1) - J_{\theta^*}(x_1) \right| = O(\|LJ_{\theta^*} - J_{\theta^*}\|)$
- Similarly bounding $\left| J_{P_{J\widehat{\theta}}}(x_1) - J_{\widehat{\theta}}(x_1) \right| = O\left(\left\|LJ_{\widehat{\theta}} - J_{\widehat{\theta}}\right\|\right)$
- J_{θ^*} and $J_{\widehat{\theta}}$ are close by the optimization

Crowdsourcing

- Need to label A items.
- Each item has soft label $\mu_i \in [0, 1]$
- Guess if $\mu_i \geq \frac{1}{2}$ for as many i as we can
- For $t = 1, \ldots, T$:
  - Pick $a \in \{1, \ldots, A\}$
  - Receive $X_t \sim \text{Bern}(\mu_i)$
- Use Beta prior $\Rightarrow$ MDP dynamics equivalent to Bayesian updates
- P_0 limits transitions
- $q(x)$ rewards correct labels





- Average error of three policies
- Our method requires 10% fewer samples for same accuracy

- Portion of budget vs. soft label
- Harder soft labels receive more budget
- Larger difference as B grows

Thanks!

Crowdsourcing details

- Objective: posterior classification error
- Prior for label i: Beta(a_0^i, b_0^i)
- State space: all possible integer increments for (a_0^i, b_0^i)
- Define: $I(a,b) = \Pr(\theta > .5 | \theta \sim \mathrm{Beta}(a,b))$, $h(x) = x \wedge (1-x)$
- Opt-KG: $p_i \propto [h(I(a_i+1, b_i)) \wedge h(I(a_i, b_i+1))] - h(I(a_i, b_i))$
- Base policy: Opt-KG
- Features: For each state $\{a_i, b_i\}$, all $\mathbb{E}[X_i]$, $1 - \mathbb{E}[X_i]$, and $\mathbb{E}[X_i^2]$ for $X_i \sim \mathrm{Beta}(a_i, b_i)$

Proof part 1

Lemma

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

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

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

$$\|\mu_u - u\|_1 \leq \tau(\mu_u) \log(1/\epsilon_1) O(\epsilon_1 + \epsilon_2)$$

- Let μ_t be u^+ after t steps
- $\|\mu_t - u^+\|_1 = O(t(\epsilon_1 + \epsilon_2))$
- Mixing assumption: $\|\mu_t - \mu_u\|_1 \leq e^{-t/\tau(u^+)}$
- Take $t = \tau(u^+) \log(1/\epsilon')$ and use triangle inequality

Applying SGD theorem

Theorem (Lemma 3.1 of [**?**])

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* $\mathbb{E}[f'_t | z_t] = \nabla f(z_t)$ *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* $1 - \delta$*:*

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z \in \mathcal{Z}} \sum_{t=1}^{T} f_t(z) \leq O\left(Z\sqrt{T}\left(F + \sqrt{\log \frac{1}{\delta}} \right) \right)$$

checking conditions of theorem

- Recall gradient: for $(x, a) \sim q_1$ and $y \sim q_2$, $g_t(\theta) =$

$$\ell^\top \Phi - H \frac{\Phi_{(x,a),:}}{q_1(x,a)} \mathbb{I}\{\Phi_{(x,a),:}\theta < 0\} + H \frac{(P - B)_{:,y}^\top \Phi}{q_2(y)} \operatorname{sgn}\left((\Phi\theta)^\top (P - B)_{:,y}\right.$$

- We can bound $\|g_t(\theta)\|_2 \le \sqrt{d} + H(C_1 + C_2) := F$
- $\mathbb{E}[g_t(\theta)] = \nabla c(\theta)$.
- The SDG theorem gives

$$\begin{aligned} \ell^\top \Phi \widehat{\theta}_T + H(V_1(\widehat{\theta}) + V_2(\widehat{\theta})) \le & \ell^\top \Phi \theta^* + H(V_1(\theta^*) + V_2(\theta^*)) \\ & + O\left(\frac{SH(C_1 + C_2)}{\sqrt{T}}\right) \end{aligned}$$

proof conclusion

- We take, $i = 1, 2$

$$V_i(\widehat{\theta}) \leq \frac{2 + 2S}{H} + V_1(\theta^*) + V_2(\theta^*) + O\left(\frac{C_1 + C_2}{\sqrt{T}}\right) := \epsilon'$$

- Apply the lemma to $\Phi\widehat{\theta}$ and $\Phi\theta^*$:

$$\begin{aligned}\ell^{\mathsf{T}}\mu_{\widehat{\theta}_T} - \ell^{\mathsf{T}}\mu_{\theta^*} &\leq H(V_1(\theta^*) + V_2(\theta^*)) + O\left(\frac{H(C_1 + C_2)}{\sqrt{T}}\right) \\ &+ \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')O(\epsilon') \\ &+ \tau(\mu_{\theta^*})\log(1/(V_1(\theta^*) + V_2(\theta^*))O(V_1(\theta^*) + V_2(\theta^*)) \\ &= HO(V_1(\theta^*) + V_2(\theta^*)) + O\left(\frac{H(C_1 + C_2)}{\sqrt{T}}\right)\end{aligned}$$

- Taking $H = 1/\epsilon$ and $T = 1/\epsilon^4$:

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