On the Stochastic (Variance-Reduced) Proximal Gradient Method for Regularized Expected Reward Optimization

Ling Liang

Joint work with: Haizhao Yang
University of Maryland, College Park

February 22, 2024, Brin MRC

BRIN MATHEMATICS RESEARCH CENTER

Workshop on Scientific Machine Learning: Theory and Algorithms

Outline

Markov Decision Process

Reward Optimization

$$\max_{\theta} \mathcal{J}(\theta) := \mathbb{E}_{x \sim \pi_\theta}[\mathcal{R}(x)]$$

- Motivation: convergence of the finite expression method (Liang and Yang, 2022)
- Can be solved by the stochastic policy gradient method (Williams, 1992).

Markov Decision Process

Reward Optimization

$$\max_{\theta} \mathcal{J}(\theta) := \mathbb{E}_{x\sim\pi_\theta}[\mathcal{R}(x)]$$

- Motivation: convergence of the finite expression method (Liang and Yang, 2022)
- Can be solved by the stochastic policy gradient method (Williams, 1992).
- $x := \{s_t, a_t, r_{t+1}\}_{t=0}^{\infty}$: trajectory.
- $\mathcal{R}(x) := \sum_{t=0}^{\infty} \gamma^t r_{t+1}$.
- $\pi_\theta(x) := \rho(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t)\tilde{\pi}_\theta(a_t|s_t)$.

- S: State space.
- A: Action space.
- $R : S \times A \to [-U, U]$: reward function.
- $P(s'|s, a)$ state transition probability.
- $\tilde{\pi}_\theta(\cdot|\cdot) : A \times S \to [0, 1]$: policy parameterized by θ.
- $\gamma \in [0, 1)$: discount factor.
- ρ : initial state distribution.

Performative Prediction

Performative Prediction (Perdomo et al., 2020)

$$\min_{x} \mathcal{J}(x) := \mathbb{E}_{z \sim \mathcal{D}(x)}[\ell(z, x)]$$

- Stochastic optimization with decision-dependent distributions.
- ℓ: loss function is assumed to be smooth and strongly convex.

Performative Prediction

Performative Prediction (Perdomo et al., 2020)

$$\min_{x} \ \mathcal{J}(x) := \mathbb{E}_{z\sim\mathcal{D}(x)}[\ell(z,x)]$$

- Stochastic optimization with decision-dependent distributions.
- ℓ: loss function is assumed to be smooth and strongly convex.

Theorem

If the loss is smooth, strongly convex, and the mapping $\mathcal{D}(\cdot)$ is sufficiently Lipschitz, then the repeated risk minimization:

$$x_{t+1} = \mathrm{argmin}_x \ \mathbb{E}_{z\sim\mathcal{D}(x_t)}[\ell(z;x)], \quad t \geq 0, \quad (\textit{not practical})$$

converges to the performative stationary point:

$$x_{PS} := \mathrm{argmin} \ \mathbb{E}_{z\sim\mathcal{D}(x_{PS})}[\ell(z,x)]$$

at a linear rate.

Thinking

- The model can not handle constraints on the decision variable.
- The repeated risk minimization is not practical.
- Global convergence.
- The inner loss function ℓ needs to be strongly convex.

Regularized Expected Reward Optimization

Regularized Performative Prediction (Drusvyatskiy and Xiao, 2023)

$$\min_{x} \mathbb{E}_{z \sim \mathcal{D}(x)}[\ell(z, x)] + r(x)$$

- r: convex regularizer (e.g., indicator functions), could be nonsmooth.
- Classical stochastic algorithms, originally designed for static problems, can be applied directly for finding such performative stability with little loss in efficiency.

Algorithms	Iterate update with $z_t \sim \mathcal{D}(x_t)$
Proximal point	$x_{t+1} = \operatorname*{arg\,min}_x \ \ell(x, z_t) + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2$
Prox-gradient	$x_{t+1} = \text{prox}_{\eta_t r}\left(x_t - \eta_t \nabla \ell(x_t, z_t)\right)$
Accel. prox-grad.	$\begin{Bmatrix} x_t = \text{prox}_{\eta_t r}\left(y_{t-1} - \eta_t \nabla \ell(y_{t-1}, z_t')\right) \\ y_t = x_t + \beta_t(x_t - x_{t-1}) \end{Bmatrix}$ with $z_t' \sim \mathcal{D}(y_{t-1})$
Clipped gradient	$x_{t+1} = \operatorname*{arg\,min}_x \left(\ell(x_t, z_t) + \langle \nabla \ell(x_t, z_t), x - x_t \rangle\right)^+ + r(x) + \frac{1}{2\eta_t}\|x - x_t\|^2$
Dual averaging	$x_{t+1} = \operatorname*{arg\,min}_x \left\langle \frac{1}{t}\sum_{i=1}^{t} \nabla \ell(x_t, z_t), x \right\rangle + r(x) + \frac{1}{2\eta_t}\|x - x_0\|^2$

Table 1: Stochastic algorithms with state-dependent distributions.

Regularized Expected Reward Optimization

Our model

$$\max_{\theta} \ \mathcal{F}(\theta) := \underbrace{\mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\right]}_{\mathcal{J}(\theta)} - \mathcal{G}(\theta)$$

- $\mathcal{J}$ can be non-concave while $\mathcal{G}$ is assumed to be a convex regularizer.
- Can the Stochastic Proximal Gradient Method:

$$\theta^{t+1} = \mathrm{Prox}_{\eta\mathcal{G}}\left(\theta^t + \eta g^t\right), \quad g^t \approx \nabla\mathcal{J}(\theta^t).$$

  be applied to the nonconcave maximization problem?
- What are the convergence properties?
- Can the Variance Reduction be applied to get better results?

Policy Gradient

Conditions on $\mathcal{R}_\theta$

1. $\mathcal{R}_\theta(\cdot)$ is π_θ-integrable for any $\theta \in \mathbb{R}^n$ and $\sup_{\theta,x} |\mathcal{R}_\theta(x)| \leq U$.
2. $\mathcal{R}_\theta(\cdot)$ is twice continuously differentiable with respect to θ, and there exist positive constants $\widetilde{C}_g$ and $\widetilde{C}_h$ such that

$$\sup_{\theta,x} \|\nabla_\theta \mathcal{R}_\theta(x)\| \leq \widetilde{C}_g, \quad \sup_{\theta,x} \left\|\nabla_\theta^2 \mathcal{R}_\theta(x)\right\|_2 \leq \widetilde{C}_h.$$

Conditions on π_θ

The function $\log \pi_\theta(x)$ is twice differential with respect to $\theta \in \mathbb{R}^n$ and there exist positive constants C_g and C_h such that

$$\sup_{x\in\mathbb{R}^d,\ \theta\in\mathbb{R}^n} \|\nabla_\theta \log \pi_\theta(x)\| \leq C_g, \qquad \sup_{x\in\mathbb{R}^d,\ \theta\in\mathbb{R}^n} \left\|\nabla_\theta^2 \log \pi_\theta(x)\right\|_2 \leq C_h.$$

The policy gradient (Sutton and Barto, 2018)

$$\nabla_\theta \mathcal{J}(\theta) := \mathbb{E}_{x\sim\pi_\theta}\left[\mathcal{R}_\theta(x)\nabla_\theta \log \pi_\theta(x) + \nabla_\theta \mathcal{R}_\theta(x)\right].$$

- L-Smoothness: $\left\|\nabla_\theta \mathcal{J}(\theta) - \nabla_\theta \mathcal{J}(\theta')\right\| \leq L \left\|\theta - \theta'\right\|$, $L > 0$.

Stochastic Proximal Gradient Method

Algorithm 1 The stochastic proximal gradient method

Input: initial point θ^0, sample size N and the learning rate $\eta > 0$.
for $t = 0, \ldots, T-1$ **do**
  Compute the stochastic gradient estimator:

$$g^t := \frac{1}{N} \sum_{j=1}^{N} g(x^{t,j}, \theta^t),$$

  where $\{x^{t,1}, \ldots, x^{t,N}\}$ are sampled independently according to π_{θ^t}.
  Update

$$\theta^{t+1} = \mathrm{Prox}_{\eta\mathcal{G}}\left(\theta^t + \eta g^t\right).$$

end for
Output: $\hat{\theta}^T$ selected randomly from the generated sequence $\{\theta^t\}_{t=1}^{T}$.

- Stochastic gradient estimator: g^t.
- Proximal gradient update: Prox.
- Output strategy.

Convergence Properties

First-order Stationary Point

$$
\begin{aligned}
& 0 \in -\nabla_\theta \mathcal{J}(\theta) + \partial \mathcal{G}(\theta) \\
\Leftrightarrow\ & \mathrm{dist}(0, -\nabla_\theta \mathcal{J}(\theta) + \partial \mathcal{G}(\theta)) = 0 \\
\Leftrightarrow\ & 0 = G_\eta(\theta) := \frac{1}{\eta}\left[\mathrm{Prox}_{\eta\mathcal{G}}\left(\theta + \eta \nabla_\theta \mathcal{J}(\theta)\right) - \theta\right]
\end{aligned}
$$

Theorem

Under suitable conditions, let $\epsilon > 0$ be a given accuracy. Running the Algorithm 1 for $T = O(\epsilon^{-2})$ iterations with the learning rate $\eta < \frac{1}{2L}$ and the sample size $N := O(\epsilon^{-2})$ outputs a point $\hat{\theta}^T$ satisfying

$$
\mathbb{E}_T\left[\mathrm{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial \mathcal{G}(\hat{\theta}^T)\right)^2\right] \le \epsilon^2.
$$

Moreover, the sample complexity is $O(\epsilon^{-4})$.

Global Convergence

Gradient Domination

$$\|G_\eta(\theta))\| \geq 2\sqrt{\omega}\left(\mathcal{F}^* - \mathcal{F}(\theta)\right), \quad \forall\, \theta \in \mathbb{R}^n,$$

- Gradient Domination is related to PL condition and KL condition in the field of optimization.
- Running Algorithm 1 for $T = O(\epsilon^{-2})$ iterations:

$$\mathbb{E}_T\left[\mathcal{F}^* - \mathcal{F}(\hat{\theta}^T)\right] \leq \frac{1}{2\sqrt{\omega}}\epsilon.$$

- MDP satisfies the gradient domination (Agarwal et al., 2021).
- Our model: no explicit structures as in MDP, remains open.

PAGE: ProbAbilistic Gradient Estimator

Importance Sampling Based PAGE (Li et al., 2021)

$$g^{t+1} = \begin{cases} \frac{1}{N_1}\sum_{j=1}^{N_1} g(x^{t+1,j}, \theta^{t+1}), & \text{w.p. } p, \\ \frac{1}{N_2}\sum_{j=1}^{N_2} g(x^{t+1,j}, \theta^{t+1}) - \frac{1}{N_2}\sum_{j=1}^{N_2} g_w(x^{t+1,j}, \theta^{t}, \theta^{t+1}) + g^{t}, & \text{w.p. } 1-p, \end{cases},$$

where $g_w(x, \theta, \theta') = \frac{\pi_\theta(x)}{\pi_{\theta'}(x)} g(x, \theta)$.

- Strong conditions on g_w is needed.

Improved Complexity via Variance Reduction

Theorem

Under suitable conditions. For a given $\epsilon \in (0,1)$*, we set* $p := \frac{N_2}{N_1+N_2}$ *with* $N_1 := O(\epsilon^{-2})$ *and* $N_2 := \sqrt{N_1} = O(\epsilon^{-1})$*. Choose a learning rate* η *satisfying*

$$\eta \in \left(0, \frac{L}{2C + 2L^2}\right].$$

Then, running the algorithm for $T := O(\epsilon^{-2})$ *iterations outputs a point* $\hat{\theta}^T$ *satisfying*

$$\mathbb{E}_T\left[\mathrm{dist}\left(0, -\nabla_\theta \mathcal{J}(\hat{\theta}^T) + \partial \mathcal{G}(\hat{\theta}^T)\right)^2\right] \le \epsilon^2.$$

Moreover, the total expected sample complexity is $O(\epsilon^{-3})$*.*

Summary

- Stochastic proximal gradient method for (nonconcave) regularized expected reward optimization.
- Improve sample complexity via variance reduction.
- How to obtain global convergence? More applications?
- Convergence to performative stability?
- How to relax the employed conditions?
- Acceleration?
- Practical performance?

Questions