Taming Nonconvex Stochastic Mirror Descent with General Bregman Divergence

Ilyas Fatkhullin   Niao He

Nonconvex Stochastic Optimization

$$\min_{x\in\mathcal{X}} \quad \mathbb{E}\left[f(x,\xi)\right] + r(x).$$

$F(x) := \mathbb{E}\left[f(x,\xi)\right]$ differentiable $r(x)$ convex
$\xi \sim \mathcal{D}$ unknown distribution $\mathcal{X} \subset \mathbb{R}^d$ closed, convex

SMD [1] $x_{t+1} = \mathrm{argmin}_{x\in\mathcal{X}} \langle \nabla f(x_t,\xi_t), x\rangle + r(x) + \frac{1}{\eta_t} D_\omega(x, x_t)$

Distance generating function: $\omega(x)$ is 1-strongly convex w.r.t. $\|\cdot\|$.
Bregman divergence: $D_\omega(x,y) := \omega(x) - \omega(y) - \langle \nabla\omega(y), x-y\rangle$.

Examples:

	$\omega(x)$	$D_\omega(x,y)$	Smooth?
1. Euclidean	$\frac{1}{2}\|x\|_2^2$	$\frac{1}{2}\|x-y\|_2^2$	✔
2. Entropy	$\sum_{i=1}^d x^i \log(x^i)$	$\sum_{i=1}^d x^i \log\left(x^i/y^i\right)$	✗
3. Polynomial	$\frac{1}{2}\|x\|_2^2 + \frac{1}{q+2}\|x\|_2^{q+2}$	–	✗

Convergence Measures

(i) Bregman Forward-Backward Envelope

$$Q_\rho(x,y) := \langle \nabla F(x), y-x\rangle + \rho D_\omega(y,x) + r(y) - r(x),$$
$$\mathcal{D}_\rho(x) := -2\rho \min_{y\in\mathcal{X}} Q_\rho(x,y).$$

(ii) Bregman Gradient Mapping

$$x^+ := \mathrm{argmin}_{y\in\mathcal{X}}\, Q_\rho(x,y),$$
$$\Delta_\rho^+(x) := \rho^2(D_\omega(x^+,x) + D_\omega(x,x^+)).$$

Remark: $\mathcal{D}_\rho(x) = \Delta_\rho^+(x) = \|\nabla F(x)\|^2$ if $\omega(x) = \frac{1}{2}\|x\|_2^2$, $r=0$.

Lemma 1.
a. $2\mathcal{D}_{\rho/2}(x) \geq \Delta_\rho^+(x) \geq \rho^2\|x^+ - x\|^2$, $\forall x\in\mathcal{X}, \rho > 0$.
b. It can be $\mathcal{D}_\rho(x) \gg \Delta_{\rho_1}^+(x)$, e.g., for $r(x) = |x|$, $F(x) = x^2$

$$\mathcal{D}_\rho(x) \geq \frac{2}{|x|}\Delta_{\rho_1}^+(x) \qquad \forall x\in(0,1], \forall \rho_1 \in [\rho, 2\rho].$$

Claim 1. $\mathcal{D}_\rho(x)$ is the strongest FOSP measure we know for **SMD**.

Assumptions

A.1. Relative smoothness w.r.t. $\omega(\cdot)$.

$$-\ell D_\omega(x,y) \leq F(x) - F(y) - \langle \nabla F(y), x-y\rangle \leq \ell D_\omega(x,y).$$

Remark: A.1. is implied by $\|\nabla F(x) - \nabla F(y)\|_* \leq \ell\|x-y\|$.

A.2. Bounded variance w.r.t. dual $\|\cdot\|_*$.

$$\mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla F(x), \qquad \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla F(x)\|_*^2\right] \leq \sigma^2.$$

Limitations in Prior Work

✗ [2] Large mini-batch $\Omega(\varepsilon^{-2})$, Euclidean norms in **A.1.** and **A.2.**

$$\lambda_{t,1} := \Phi(x_t) - \Phi^*, \qquad \Phi(x) := F(x) + r(x).$$

✗ [3,4] Smooth $\omega(\cdot)$ and bounded gradient assumption.

$$\lambda_{t,2} := \Phi_{1/\rho}(x_t) - \Phi^*, \qquad \Phi_{1/\rho}(x) := \min_{y\in\mathcal{X}}\left[\Phi(y) + \rho D_\omega(y,x)\right].$$

Contributions.

✔ New Lyapunov function:

$$\lambda_t := \eta_{t-1}\rho\lambda_{t,1} + \lambda_{t,2}.$$

✔ Analysis with general non-smooth $\omega(\cdot)$.

✔ Stronger measure, $\mathcal{D}_\rho(x)$, and assume mild **A.1.**, **A.2.**

Main Results

Convergence in-expectations

Theorem 1. Let **A.1.** and **A.2.** hold, $\bar{x}_T \sim \mathcal{U}(x_0,\ldots,x_{T-1})$, $\eta_t := \min\left\{\frac{1}{2\ell}, \sqrt{\frac{\lambda_0}{\sigma^2\ell T}}\right\}$,

$$\mathbb{E}\left[\mathcal{D}_{3\ell}(\bar{x}_T)\right] = \mathcal{O}\left(\frac{\ell\lambda_0}{T} + \sqrt{\frac{\sigma^2\ell\lambda_0}{T}}\right).$$

High probability convergence.

Theorem 2. Let **A.1.**, **A.2.** hold and $\|\nabla f(x,\xi) - \nabla F(x)\|_*$ be σ-sub-Gaussian. Then with probability $1-\beta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{D}_{5\ell}(x_t) \leq \mathcal{O}\left(\frac{\ell\widetilde{\lambda}_0}{T} + \sqrt{\frac{\sigma^2\ell\widetilde{\lambda}_0}{T}}\right),$$

where $\widetilde{\lambda}_0 := \Phi(x_0) - \Phi^* + \eta_0\sigma^2\log\left(1/\beta\right)$.

Global convergence under Generalized Proximal PŁ.
A.3. There exists $\alpha\in[1,2]$, $\mu > 0$ such that for some $\rho \geq 3\ell$ and all $x\in\mathcal{X}$

$$\mathcal{D}_\rho(x) \geq 2\mu(\Phi(x) - \Phi^*)^{2/\alpha}.$$

Theorem 3. Let **A.1.**, **A.2.**, **A.3.** hold. Then for any $\varepsilon > 0$, we have $\min_{t\leq T}\mathbb{E}\left[\Phi(x_t^+) - \Phi^*\right] \leq \varepsilon$ after

$$T = \mathcal{O}\left(\frac{\ell\lambda_0}{\mu}\frac{1}{\varepsilon^{\frac{2-\alpha}{\alpha}}}\log\left(\frac{\ell\lambda_0}{\mu\varepsilon}\right) + \frac{\ell\lambda_0\sigma^2}{\mu^2}\frac{1}{\varepsilon^{\frac{4-\alpha}{\alpha}}}\right).$$

Implications for Machine Learning

I. Differentially Private Learning in ℓ_1 setup.
Definition 1. Algorithm $\mathcal{M}$ is (ϵ,δ)-DP if for any $\mathcal{Y} \subseteq Range(\mathcal{M})$

$$\Pr\left(\mathcal{M}(S)\in\mathcal{Y}\right) \leq e^\epsilon \Pr\left(\mathcal{M}(S')\in\mathcal{Y}\right) + \delta.$$

Let $S := \{\xi^1,\ldots,\xi^n\}$, $\nabla F(x) := \sum_{i=1}^n \nabla f(x,\xi^i)$, $\omega(x) = \sum_{i=1}^d x^{(i)}\log x^{(i)}$, and inject Gaussian noise $b_t \sim \mathcal{N}(0, \sigma_G^2 I_d)$, $\sigma_G > 0$.

DP-MD: $x_{t+1} = \mathrm{argmin}_{y\in\mathcal{X}}\eta_t(\langle \nabla F(x_t) + b_t, y\rangle + r(y)) + D_\omega(y, x_t)$,

Corollary 1. Let $\mathcal{X}$ be a unit simplex, and $\|\nabla F(x)\|_2 \leq G$ for all $x\in\mathcal{X}$. Then **DP-MD** is (ϵ,δ)-DP and with probability $1-\beta$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{D}_{5\ell}(x_t) = \mathcal{O}\left(\frac{G\sqrt{\ell\lambda_0 \log(d)\log\left(1/\delta\right)\log\left(1/\beta\right)}}{n\epsilon}\right).$$

Implication: This replaces d by $\log(d)$ compared to **DP-GD**, due to dual norm in **A.2.**

II. Policy Optimization in Reinforcement Learning.
MDP $M = \{\mathcal{S},\mathcal{A},\mathcal{P},R,\gamma,p\}$ with finite $|\mathcal{S}|$ and $|\mathcal{A}|$. $\Delta(\mathcal{A})$ is a probability simplex for each $s\in\mathcal{S}$. Minimize over π

$$V_p(\pi) := -\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h R(s_h,a_h)\right], \quad \text{s.t. } \pi\in\mathcal{X} := \Delta(\mathcal{A})^{|\mathcal{S}|}, \quad s_0\sim p.$$

Fact 1. $\|\nabla V_p(\pi) - \nabla V_p(\pi')\|_{2,\infty} \leq \frac{2\gamma}{(1-\gamma)^3}\|\pi - \pi'\|_{2,1}$ $\forall \pi,\pi'\in\mathcal{X}$.

SMPG : $\pi_{t+1} = \pi_t \odot E_t, \quad E_t^s := \frac{\exp\left(-\eta_t\widehat{\nabla}_s V_\mu(\pi_t)\right)}{\sum_{a\in\mathcal{A}}\exp\left(-\eta_t\widehat{\nabla}_{s,a}V_\mu(\pi_t)\right)} \quad \forall s\in\mathcal{S}$,

where $\widehat{\nabla}_s V_\mu(\pi_t) \approx \nabla_s V_\mu(\pi_t)$ with variance $\sigma_{2,\infty}^2$ in $\|\cdot\|_{2,\infty}$ norm.

Corollary 2. $\forall\varepsilon > 0$, **SMPG** gives $\min_{0\leq t\leq T-1}\mathbb{E}\left[\mathcal{D}_\rho(\pi_t)\right] \leq \varepsilon^2$ after

$$T = \mathcal{O}\left(\frac{\mathbf{1}}{(1-\gamma)^3\varepsilon^2} + \frac{\boldsymbol{\sigma}_{\mathbf{2},\boldsymbol{\infty}}^{\mathbf{2}}}{(1-\gamma)^3\varepsilon^4}\right).$$

Implication: Improves the Euclidean version: $\mathcal{O}\left(\frac{|\mathcal{A}|}{(1-\gamma)^3\varepsilon^2} + \frac{|\mathcal{A}|\sigma_F^2}{(1-\gamma)^3\varepsilon^4}\right)$ without access to Q-function, due to **A.1.**& **Fact 1.**

References

[1] A. Nemirovskij and D. Yudin. *Efficient methods of solving convex programming problems of high dimensionality.* Econom.&math. methods (in russian), 1979.
[2] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization.* Mathematical Programming, 2016.
[3] Damek Davis, Dmitriy Drusvyatskiy, and Kellie J MacPhee. *Stochastic model-based minimization under high-order growth.* preprint arXiv:1807.00255, 2018.
[4] Siqi Zhang and Niao He. *On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization.* preprint arXiv:1806.04781, 2018.