Stochastic Model-based Algorithm can be Accelerated by Minibatching for Sharp Functions

September 2, 2021

1 Literature Review

Algorithm	Convexity	Randomness	Stepsize	Complexity
SGD	Convex	Deterministic	Constant	$\log(1/\varepsilon)$ [Ber15]
			Geometrically	$\log(1/\varepsilon)$ [DDMP18]
		Stochastic	Constant	–
			Geometrically	$\log(1/\varepsilon)$ [DDC19]
	Weakly	Deterministic	Constant	$\log(1/\varepsilon)$ [DDMP18]
			Geometrically	$\log(1/\varepsilon)$ [DDMP18]
		Stochastic	Constant	–
			Geometrically	$\log(1/\varepsilon)$ [DDC19]
SPL/SPP	Convex	Deterministic	Constant	$\log\log(1/\varepsilon)$ [Ber15]
			Geometrically	Needed
		Stochastic	Constant	$\log(1/\varepsilon)^\dagger$ [AD19]
			Geometrically	$\log(1/\varepsilon)$ [DDC19]
	Weakly	Deterministic	Constant	$\log\log(1/\varepsilon)$ [CCD+21]
			Geometrically	Needed
		Stochastic	Constant	Needed
			Geometrically	$\log(1/\varepsilon)$ [DDC19]

Table 1: Literature over optimization with sharpness

$\dagger$: minibatch acceleration is already proven for easy problems ($\arg\min_x f(x,\xi) = x^*, \forall \xi$) in [ACCD20].

2 Preliminaries

Consider the following optimization problem

$$\min_{x\in\mathcal{X}} \quad \mathbb{E}_\xi[f(x,\xi)]$$

Assumption 1. It is possible to sample i.i.d. $\{\xi_1, \dots, \xi_n\}$.

Assumption 2. f is λ-weakly convex. i.e., $f + \frac{\lambda}{2}\|x\|^2$ is convex.

Assumption 3. f is sharp. In other words,

$$\mu \cdot \mathrm{dist}(x, \mathcal{X}^*) \le f(x) - f^*, \forall x \in \mathcal{X}^*,$$

where $\mathcal{X}^*$ is the set of optimal solutions to the problem.

Assumption 4. f is locally Lipschitz-continuous.

Define the tube $\mathcal{T}_\gamma := \left\{x \in \mathcal{X} : \operatorname{dist}(x, \mathcal{X}^*) \le \frac{\gamma\mu}{\tau}\right\}$ and we have

$$\min_{g \in \partial f_x(x,\xi)} \|g\| \le L, \forall x \in \mathcal{T}_2, \xi.$$

Assumption 5. Two-sided accuracy is available. i.e.,

$$|f(y) - f_x(y, \xi)| \le \frac{\tau}{2}\|x - y\|^2.$$

3 Convex Optimization

To analyze the case of convex optimization, we specially let $\lambda = 0$ and further assume that global Lipchitzness of the model $f_x(\cdot, \xi)$ holds.

3.1 Restarting Strategy with Decaying Stepsize

Lemma 1 *The algorithm in [DG21] initialized with y_0 satisfies*

$$\mathbb{E}[f(x^{K+1}) - f^*] \le \frac{2\tau \operatorname{dist}^2(y_0, \mathcal{X}^*)}{(K+1)(K+2)} + \frac{4\sqrt{2}L \operatorname{dist}(y_0, \mathcal{X}^*)}{\sqrt{3m(K+1)}}.$$

Lemma 2 *For some growth function $g > 0$, denote $E_t := \left\{\operatorname{dist}(x_t, \mathcal{X}^*) \le \frac{R_0}{g(t)}\right\}$ and we have the following relation holds*

$$\mathbb{P}(E_T) \quad \ge \quad 1 - \sum_{t=0}^{T-1} \left[\frac{2\tau R_0}{\mu K^2} \cdot \frac{g(t+1)}{g(t)^2} + \frac{4\sqrt{6}L}{3\sqrt{m(K+1)}} \cdot \frac{g(t+1)}{g(t)}\right].$$

Proof Without loss of generality we have

$$\begin{aligned} & \mathbb{P}(E_{t+1}) \\ = \ & \mathbb{P}(E_{t+1}|\overline{E_t})\mathbb{P}(\overline{E_t}) + \mathbb{P}(E_t)\mathbb{P}(E_{t+1}|E_t)\mathbb{P}(E_t) \\ \ge \ & \mathbb{P}(E_t)\mathbb{P}(E_{t+1}|E_t) \end{aligned}$$

and that

$$\begin{aligned} \mathbb{P}(E_{t+1}|E_t) &= 1 - \mathbb{P}(\overline{E_{t+1}}|E_t) \\ &= 1 - \mathbb{P}\left(\operatorname{dist}(x_{t+1}, \mathcal{X}^*) \ge \frac{R_0}{g(t+1)}|E_t\right) \\ &\ge 1 - \frac{\mathbb{E}[\operatorname{dist}(x_{t+1}, \mathcal{X}^*)|E_t]}{R_0/g(t+1)} \\ &= 1 - \frac{\mathbb{E}[\operatorname{dist}(x_{t+1}, \mathcal{X}^*)\mathbb{I}\{E_t\}]}{R_0/g(t+1)} \frac{1}{\mathbb{P}(E_t)}, \end{aligned}$$

where the inequality is by Markov's inequality. Then we consider

$$\begin{aligned} \mathbb{E}[\operatorname{dist}(x_{t+1}, \mathcal{X}^*)\mathbb{I}\{E_t\}] &\le \frac{1}{\mu}\mathbb{E}[(f(x_{t+1}) - f^*)\mathbb{I}\{E_t\}] \\ &\le \frac{1}{\mu}\left\{\frac{2\tau\mathbb{E}[\operatorname{dist}^2(x_t, \mathcal{X}^*)\mathbb{I}\{E_t\}]}{(K+1)(K+2)} + \frac{4\sqrt{2}L\mathbb{E}[\operatorname{dist}(x_t, \mathcal{X}^*)\mathbb{I}\{E_t\}]}{\sqrt{3m(K+1)}}\right\} \\ &\le \frac{2\tau R_0^2}{\mu K^2} \cdot \frac{1}{g(t)^2} + \frac{4\sqrt{6}LR_0}{3\sqrt{m(K+1)}} \cdot \frac{1}{g(t)}. \end{aligned}$$

Next we combine the above and obtain that

$$\begin{aligned}
& \mathbb{P}(E_{t+1}) \\
\geq\ & \mathbb{P}(E_t)\left\{1-\frac{\mathbb{E}[\operatorname{dist}(x_{t+1},\mathcal{X}^*)\mathbb{I}\{E_t\}]}{R_0/g(t+1)}\frac{1}{\mathbb{P}(E_t)}\right\} \\
=\ & \mathbb{P}(E_t)-\frac{\mathbb{E}[\operatorname{dist}(x_{t+1},\mathcal{X}^*)\mathbb{I}\{E_t\}]}{R_0/g(t+1)} \\
\geq\ & \mathbb{P}(E_t)-\left[\frac{2\tau R_0}{\mu K^2}\cdot\frac{g(t+1)}{g(t)^2}+\frac{4\sqrt{6}L}{3\sqrt{m(K+1)}}\cdot\frac{g(t+1)}{g(t)}\right].
\end{aligned}$$

Summing over $t=0,\ldots,T-1$ gives

$$\mathbb{P}(E_T)\quad\geq\quad 1-\sum_{t=0}^{T-1}\left[\underbrace{\frac{2\tau R_0}{\mu K^2}\cdot\frac{g(t+1)}{g(t)^2}}_{\text{Quadratic}}+\underbrace{\frac{4\sqrt{6}L}{3\sqrt{m(K+1)}}\cdot\frac{g(t+1)}{g(t)}}_{\text{Linear}}\right]$$

□

Remark 1 For SPP algorithm we have $\tau=0$ and the quadratic acceleration term is not present and we hence have

$$\mathbb{P}(E_T)\quad\geq\quad 1-\frac{4\sqrt{6}L}{3\sqrt{m(K+1)}}\sum_{t=0}^{T-1}\frac{g(t+1)}{g(t)}$$

Remark 2 To recover the deterministic quadratic convergence, we let $m\to\infty$ and get

$$\mathbb{P}(E_T)\quad\geq\quad 1-\frac{2\tau R_0}{\mu K^2}\sum_{t=0}^{T-1}\frac{g(t+1)}{g(t)^2}$$

and this allows us to take growth function to $g(t)=2^{2^t}$ such that $\frac{g(t+1)}{g(t)^2}=2=\mathcal{O}(1)$. Then we can follow [DDC19] to recover the quadratic convergence.

From now on we assume that $\tau = 0$ (proximal point) and carry out the analysis.

3.2 Optimal Choice for Parameters

Now we consider the general choice of $g(t), m_t$ and K_t. For brevity we use $m(t)$ and $K(t)$ as functions of discrete values t. Then due to monotonicity of g we have $T = g^{-1}(t)$ and that

$$\mathbb{P}(E_T) \quad \geq \quad 1 - \sum_{t=0}^{g^{-1}(R_0/\varepsilon)-1} \left(\frac{8\sqrt{6}L}{3\mu\sqrt{K(t)+1}} \cdot \frac{g(t+1)}{g(t)\sqrt{m(t)}} \right).$$

Also, we have the total sample complexity given by

$$\sum_{t=0}^{g^{-1}(R_0/\varepsilon)-1} m(t)K(t).$$

Then we use $K(t)+1$ to replace $K(t)$ and get an abstract optimization problem

$$\begin{array}{ll} \min\limits_{g,m,K} & \displaystyle\sum_{t=0}^{g^{-1}(R_0/\varepsilon)-1} m(t)K(t) \\ \text{subject to} & \displaystyle\sum_{t=0}^{g^{-1}(R_0/\varepsilon)-1} \left(\frac{8\sqrt{6}L}{3\mu} \cdot \frac{g(t+1)}{g(t)\sqrt{m(t)K(t)}} \right) \leq \delta. \end{array}$$

To solve the problem, we first denote $\alpha := R_0/\varepsilon, \theta := \frac{\sqrt{6}\mu\delta}{16L}, u(t) := m(t)K(t)$ and get

$$\begin{array}{ll} \min\limits_{g,u} & \displaystyle\sum_{t=0}^{g^{-1}(\alpha)-1} u(t) \\ \text{subject to} & \displaystyle\sum_{t=0}^{g^{-1}(\alpha)-1} \frac{1}{\sqrt{u(t)}} \cdot \frac{g(t+1)}{g(t)} \leq \theta. \end{array}$$

Linear Convergence

In this case we have $\frac{g(t+1)}{g(t)} = \beta$ and by optimality condition we know that it is optimal to let $u(t_1) = u(t_2), \forall t_1, t_2$ and the constraint is transformed into

$$\frac{\log_\beta(\alpha)}{\sqrt{u(0)}} \leq \theta/\beta \Rightarrow u(0) \geq \frac{\beta^2 \log_\beta^2(\alpha)}{\theta^2} = \frac{128L^2\beta^2 \log_\beta^2(\alpha)}{3\mu^2\delta^2}.$$

Also the objective is into

$$\sum_{t=0}^{g^{-1}(\alpha)-1} u(t) = \log_\beta(\alpha) u(0) \geq \left(\frac{\beta}{\log^3(\beta)}\right) \left(\frac{128L^2}{3\mu^2\delta^2}\right) \log^3(\alpha).$$

Hence the best bound in terms of linear convergence is attained by $\beta = e^3 \Rightarrow \frac{\beta}{\log^3(\beta)} = \frac{e^3}{27}$ with constant batchsize and this gives the best sample complexity

$$\frac{128e^3}{81} \left(\frac{L^2}{\mu^2\delta^2}\right) \log^3\left(\frac{R_0}{\varepsilon}\right).$$

Super-linear $\exp(t\log(t+1))$

In this case we have $\frac{g(t+1)}{g(t)} = \left(1 + \frac{1}{t+1}\right)^t (t+2)$ and in this case we have

$$\begin{aligned} \min_{g,u} \quad & \sum_{t=0}^{W(R_0/\varepsilon)-1} u(t) \\ \text{subject to} \quad & \sum_{t=0}^{W(eR_0/\varepsilon)-2} \frac{1}{\sqrt{u(t)}} \cdot \left(1 + \frac{1}{t}\right)^t (t+1) \le \theta, \end{aligned}$$

where $W(x)$ is the Lambert-W function. By taking $m(t) \equiv m, K(t) = \frac{512L^2e^2}{3m\mu^2\delta^2} \log^4\left(\frac{R_0}{\varepsilon}\right)$ we have the sample complexity of $o\left(\frac{512L^2}{3\mu^2\delta^2} \log^5\left(\frac{R_0}{\varepsilon}\right)\right)$. Hence we achieve super-linear convergence.

Remark 3 Currently for $\exp(t\log(t+1))$ we can preserve the order of the $\log\left(\frac{R_0}{\varepsilon}\right)$ term in sample complexity.

Constant Sample per Iteration
In this case we assume that $u(t) \equiv u$ and we have

$$\begin{aligned} \min_{g,u} \quad & g^{-1}(\alpha) \\ \text{subject to} \quad & \sum_{t=0}^{g^{-1}(\alpha)-1} \frac{g(t+1)}{g(t)} \le \theta\sqrt{u}. \end{aligned}$$

Or more abstractly, we have to solve

$$\begin{aligned} \min_{f} \quad & f^{-1}(\alpha) \\ \text{subject to} \quad & \int_0^{f^{-1}(\alpha)} \frac{f(x+1)}{f(x)} dx \le 1. \end{aligned}$$

References

[ACCD20] Hilal Asi, Karan Chadha, Gary Cheng, and John C Duchi. Minibatch stochastic approximate proximal point methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[AD19] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[Ber15] Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

[CCD+21] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

[DDC19] Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[DDMP18] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

[DG21] Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic non-smooth non-convex optimization. *arXiv preprint arXiv:2106.03034*, 2021.