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	Determinic	Constant	$\log(1/\varepsilon)$
			Geometrically	$\log(1/\varepsilon)$
		Stochastic	Constant	–
			Geometrically	$\log(1/\varepsilon)$
	Weakly	Deterministic	Constant	$\log(1/\varepsilon)$
			Geometrically	$\log(1/\varepsilon)$
		Stochastic	Constant	–
			Geometrically	$\log(1/\varepsilon)$
SPL/SPP	Convex	Deterministic	Constant	$\log\log(1/\varepsilon)$
			Geometrically	Needed
		Stochastic	Constant	$\log(1/\varepsilon)^\dagger$
			Geometrically	$\log(1/\varepsilon)$
	Weakly	Deterministic	Constant	$\log\log(1/\varepsilon)$
			Geometrically	Needed
		Stochastic	Constant	Needed
			Geometrically	$\log(1/\varepsilon)$

Table 1: Literature over optimization with sharpness

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

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. $\{\xi_1, \ldots, \xi_n\}$.

Assumption 2. f is λ-weakly convex.

We assume that $f + \frac{\lambda}{2}\|x\|^2$ is convex.

Assumption 3. f is sharp. In other words,

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

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.,

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

It is already known that in the convex case, the proximal point method converges quadratically [1] and its stochastic variant has linear convergence when using a geometrically decaying stepsize [2]. Hence there is space for acceleration.

3 Convex Optimization

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

3.1 Restarting Strategy with Decaying Stepsize

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

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

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*

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

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

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

where the inequality is by Markov's inequality.

Then we consider

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

Next we combine the above and obtain that

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

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

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

□

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

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

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 [*Dmitri*] to recover the quadratic convergence.

Now we analyze the way to choose $(g, \{m_t\})$ for faster convergence.

Consider taking $m_t = m(t)$ and we get

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

For brevity we first consider the proximal point method with $\tau = 0$ and we get the bound

$$\mathbb{P}(E_T) \quad \geq \quad 1 - \sum_{t=0}^{T-1} \left(\frac{4\sqrt{6}L}{3\mu\sqrt{K_t+1}} \cdot \frac{g(t+1)}{\sqrt{m_t} g(t)} \right).$$

Super-linear Batchsize

Take $g(t) = 2^{t^2}$ and we have

$$\mathbb{P}(E_T) \quad \geq \quad 1 - \sum_{t=0}^{T-1} \left(\frac{8\sqrt{6}L}{3\mu\sqrt{K_t+1}} \cdot \frac{4^t}{\sqrt{m_t}} \right).$$

Take $m_t = 16^t, T = \left\lceil \sqrt{\log_2\left(\frac{R_0}{\varepsilon}\right)} \right\rceil$ and $K_t \equiv \left\lfloor \frac{128T^2}{3} \cdot \left(\frac{L}{\delta\mu}\right)^2 \right\rfloor$, we have the total sample complexity of

$$\begin{aligned}
\sum_{t=0}^{T-1} m_t K_t \quad &= \quad \frac{128T^2}{3} \left(\frac{L}{\delta\mu}\right)^2 \sum_{t=0}^{T-1} 16^t \\
&\leq \quad \frac{128T^2}{45} \left(\frac{L}{\delta\mu}\right)^2 \exp\left(4\sqrt{\log_2\left(\frac{R_0}{\varepsilon}\right)}\right) \\
&\leq \quad \frac{128\log_2\left(\frac{R_0}{\varepsilon}\right)}{45} \left(\frac{L}{\delta\mu}\right)^2 \exp\left(4\sqrt{\log_2\left(\frac{R_0}{\varepsilon}\right)}\right)
\end{aligned}$$

Optimal Choice for Parameters

Last 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

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

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

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

Now we consider the following cases.

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

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

Constant Sample per Iteration

In this case we assume that $u(t) \equiv u$ and we have

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

Or more abstractly, we have to solve

$$\begin{array}{rc} \min_{f} & f^{-1}(\alpha) \\ \text{subject to} & \int_0^{f^{-1}(\alpha)} \frac{f(x+1)}{f(x)} dx \leq 1 \end{array}$$

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

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

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.

Super-linear $\exp(\mathcal{P}(t))$

In this case we consider a special case of super-linear convergence with $g(t) = e^{\beta t^p}$. In this case we have $\frac{g(t+1)}{g(t)} = \exp(\beta(t+1)^p - \beta t^p)$ and $T = g^{-1}(\alpha) = [\log(\alpha)]^{1/p}$. Hence we have the optimization problem given by

$$\begin{array}{rc} \min_{p,u} & \sum_{t=0}^{[\log(\alpha)]^{1/p}-1} u(t) \\ \text{subject to} & \sum_{t=0}^{[\log(\alpha)]^{1/p}-1} \frac{1}{\sqrt{u(t)}} \cdot \exp((t+1)^p - t^p) \leq \theta. \end{array}$$

A trivial selection is $p=2$ and $\frac{g(t+1)}{g(t)} = \exp(2\beta t+\beta)$. Then we have $[\log(\alpha)]^{1/p} - 1 = \sqrt{\log(\alpha)} - 1$, giving

$$\begin{array}{rc} \min_{u} & \sum_{t=0}^{\sqrt{\log(\alpha)}-1} u(t) \\ \text{subject to} & \sum_{t=0}^{\sqrt{\log(\alpha)}-1} \frac{\exp(2\beta t)}{\sqrt{u(t)}} \leq \theta e^{-\beta}. \end{array}$$

Then by writing the Lagrangian function

$$\mathcal{L}(\{u(t)\}, \lambda) \quad := \quad \sum_{t=0}^{\sqrt{\log(\alpha)}-1} u(t) - \lambda \left(\theta e^{-\beta} - \sum_{t=0}^{\sqrt{\log(\alpha)}-1} \frac{\exp(2\beta t)}{\sqrt{u(t)}}\right)$$

we have

$$\partial_{u(t)}\mathcal{L} \quad = \quad 1 - \frac{\lambda \exp(2\beta t)}{2} u(t)^{-3/2}$$

and that

$$
\begin{aligned}
u(t) &= \lambda^{2/3}\left(\frac{\exp(2\beta t)}{2}\right)^{2/3}
\end{aligned}
$$

$$
\begin{aligned}
\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\frac{\exp(2\beta t)}{\sqrt{u(t)}} &= \left[\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\frac{\exp(2\beta t)^{2/3}}{2^{-1/3}}\right]\lambda^{-1/3} \\
&= \theta e^{-\beta}.
\end{aligned}
$$

Hence we have $\lambda^{*2/3} = \frac{\left(\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\frac{\exp(2\beta t)^{2/3}}{2^{-1/3}}\right)^2}{\theta^2 e^{-2\beta}}$ and

$$
u(t) = \frac{\left(\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\exp(2\beta t)^{2/3}\right)^2}{\theta^2 e^{-2\beta}}(\exp(2\beta t))^{2/3},
$$

giving

$$
\begin{aligned}
\sum_{t=0}^{\sqrt{\log(\alpha)}-1} u(t) &= \frac{1}{\theta^2 e^{-2\beta}}\left(\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\exp(4\beta t/3)\right)^2 \\
&= \frac{1}{\theta^2 e^{-2\beta}}\left(\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\exp(4\beta/3)^t\right)^2 \\
&= \frac{1}{\theta^2 e^{-2\beta}}\left(\sum_{t=0}^{\sqrt{\log(\alpha)}-1}\exp(4\beta/3)^t\right)^2 \\
&= \frac{1}{\theta^2 e^{-2\beta}}\left(\frac{\exp(4\beta/3)^{\sqrt{\log(\alpha)}}-1}{\exp(4\beta/3)-1}\right)^2 \\
&= \frac{1}{\theta^2}\frac{\exp(2\beta)\left(\exp(4\beta/3)^{\sqrt{\log(\alpha)}}-1\right)^2}{[\exp(4\beta/3)-1]^2}.
\end{aligned}
$$

For some given β, we get the total complexity of

$$
\frac{128L^2}{3\mu^2\delta^2}\cdot\frac{\exp(2\beta)\left(\exp(4\beta/3)^{\sqrt{\log(\alpha)}}-1\right)^2}{[\exp(4\beta/3)-1]^2} = \mathcal{O}\left(\frac{128L^2}{3\mu^2\delta^2}e^{\sqrt{\log(R_0/\varepsilon)}}\right)
$$

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

[2] Davis, D. , D. Drusvyatskiy , and V Charisopoulos. “Stochastic algorithms with geometric step decay converge linearly on sharp functions.” *arXiv* (2019).

[3] Davis, Damek, et al. “Subgradient methods for sharp weakly convex functions.” *Journal of Optimization Theory and Applications* 179.3 (2018): 962-982.