SGD without Replacement: Sharper Rates for General Smooth Convex Functions

Dheeraj Nagaraj

Massachusetts Institute of Technology

June 12, 2019

Joint work with Praneeth Netrapalli and Prateek Jain (MSR India)

Overview

1. Introduction
2. Current Results
3. Our Results
4. Main Techniques

SGD with Replacement (SGD)

Consider observations $\xi_1, \dots, \xi_n$. Convex loss function $f(, \xi_i) : \mathbb{R}^d \to \mathbb{R}$.
Empirical Risk Minimization :

$$x^* = \arg\min_{x \in D} \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i) := \arg\min_{x \in D} \nabla \hat{F}(x, \xi_i), .$$

- **SGD with replacement (SGD)**: fix step size sequence $\alpha_t \geq 0$. Start at $x_0 \in D$. For every time step generate independent random variable $I_t \sim \mathrm{unif}([n])$.

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t, \xi_{I_t})$$

- Easy to analyze since independence of I_t ensures that $\mathbb{E}_{I_t} \nabla f(x_t, \xi_{I_t}) = \hat{F}(x_t)$.
- Sharp non-asymptotic guarantees available but seldom used in practice.

SGD without Replacement (SGDo)

In practice, the order of data is fixed (say $\xi_1, \ldots, \xi_n$) and the data is selected in this order, one after the other. One such pass is called an epoch. The algorithm is run for K epochs. A randomized version of this 'gets rid' of the bad orderings.

- **SGD without Replacement (SGDo)** At the beginning of the k th epoch, draw an independent uniformly random permutation σ_k.

$$x_{k,i} = x_{k,i-1} - \alpha_{k,i} \nabla f(x_{k,i}; \xi_{\sigma_k(i)})$$

- This is closer to the algorithm implemented in practice.
- Harder to analyze since $\mathbb{E}\nabla f(x_{k,i}; \xi_{\sigma_k(i)}) \neq \mathbb{E}\nabla \hat{F}(x_{k,i})$

Experimental Observations

- Experiments[1] found that on many problems SGDo converges as $O(1/K^2)$, which is faster than SGD which converges at $O(1/K)$. (K = number of epochs)
- Theoretically, it wasn't even shown that SGDo 'matches' the rate of SGD for all K.

[1] Léon Bottou. "Curiously fast convergence of some stochastic gradient descent algorithms". In: *Proceedings of the symposium on learning and data science, Paris*. 2009.

Currently Known Bounds

Paper	Guarantee	Assumptions	Step Sizes
Gürbüzbalaban et al. 2015	$O\left(\frac{C(n,d)}{K^2}\right)$	Lipschitz, Strong convexity Smoothness, **Hessian Lipschitz** $K > \kappa^{1.5}\sqrt{n}$	$\frac{1}{K}$
HaoChen and Sra 2018	$\tilde{O}\left(\frac{1}{n^2K^2} + \frac{1}{K^3}\right)$		$\frac{\log nK}{\mu nK}$
This paper	$\tilde{O}\left(\frac{1}{nK^2}\right)$	Lipschitz, Strong convexity Smoothness, $K > \kappa^2$	$\frac{\log nK}{\mu nK}$
Shamir 2016	$O\left(\frac{1}{nK}\right)$	Lipschitz, Strong convexity, Smoothness **Generalized Linear Function**, $K = 1$	$\frac{1}{\mu nK}$
This Paper	$O\left(\frac{1}{nK}\right)$	Lipschitz, Strong convexity, Smoothness	$\min\left(\frac{2}{L}, \frac{\log nK}{\mu nK}\right)$
Shamir 2016	$O\left(\frac{1}{\sqrt{nK}}\right)$	Lipschitz **Generalized Linear Function**, $K = 1$	$\frac{1}{\sqrt{nK}}$
This Paper	$O\left(\frac{1}{\sqrt{nK}}\right)$	Lipschitz, **Smoothness**	$\min\left(\frac{2}{L}, \frac{1}{\sqrt{nK}}\right)$

Table 1. Comparison of our results with previously known results in terms of number of functions n and number of epochs K. For simplicity, we suppress the dependence on other problem dependent parameters such as Lipschitz constant, strong convexity, smoothness etc.

Small number of Epochs

- Assumptions: $f(\cdot;\xi_i)$ is L smooth, $\|\nabla f(\cdot;\xi_i)\| \leq G$, $\text{diam}(\mathcal{W}) \leq D$.
- Suboptimality $O\left(\frac{GD}{\sqrt{nK}}\right)$ (leading order, General case)
- Suboptimality $O\left(\frac{G^2 \log nK}{\mu nK}\right)$ (leading order, μ strongly convex)
- Shamir's result[2] only works for generalized linear functions and when $K = 1$.
- All other "acceleration" results hold only when K is very large.

[2] Ohad Shamir. "Without-replacement sampling for stochastic gradient methods". In: *Advances in Neural Information Processing Systems*. 2016, pp. 46–54.

Large number of Epochs

- Assumptions: $f(\cdot;\xi_i)$ is L smooth, $\|\nabla f(\cdot;\xi_i)\| \leq G$ and $\hat{F}$ is μ strongly convex.
- When $K \gtrsim \kappa^2$, Suboptimality: $O\left(\frac{\kappa^2 G^2}{\mu}\frac{(\log nK)^2}{nK^2}\right)$
- Previous results[3] require Hessian smoothness and $K \geq \kappa^{1.5}\sqrt{n}$ to give suboptimality of $O\left(\frac{\kappa^4}{n^2K^2} + \frac{\kappa^4}{K^3}\right)$.
- Without smoothness assumption, there can be no acceleration.

[3] Jeffery Z HaoChen and Suvrit Sra. "Random Shuffling Beats SGD after Finite Epochs". In: *arXiv preprint arXiv:1806.10077* (2018).

Main Techniques

- Main bottleneck in analysis: $\mathbb{E}\nabla f(x_{k,i};\xi_{\sigma_k(i)}) \neq \mathbb{E}\nabla \hat{F}(x_{k,i})$.
- If σ'_k is independent of σ_k,

$$\mathbb{E}\nabla f(x_{k,i};\xi_{\sigma'_k(i)}) = \mathbb{E}\nabla \hat{F}(x_{k,i})\,.$$

- Therefore,
  $\mathbb{E}\nabla f(x_{k,i};\xi_{\sigma_k(i)}) = \mathbb{E}\nabla \hat{F}(x_{k,i}) + O(d_{\mathrm{W}}\left(x_{k,i}\middle|\sigma_k(i)=r, x_{k,i}\right)$
- Through coupling arguments: $d_{\mathrm{W}}\left(x_{k,i}\middle|\sigma_k(i)=r, x_{k,i}\right) \lesssim \alpha_{k,0} G$

Automatic Variance Reduction and Acceleration

- For the smooth and strongly convex case, $\nabla \hat{F}(x^*) = 0 = \frac{1}{n}\sum_{i=1}^{n} f(x^*, \xi_{\sigma_k(i)})$. (Note that this doesn't hold with independent sampling).
- Therefore, when $x_{k,0} \approx x^*$ we show by coupling arguments that:

$$0 \approx \nabla \hat{F}(x_{k,0}) \approx \frac{1}{n}\sum_{i=1}^{n} f(x_{i,k}, \xi_{\sigma_k(i)})$$

- This is similar to the variance reduction as seen in modifications of SGD like SAGA, SVRG etc.

References

- Bottou, Léon. "Curiously fast convergence of some stochastic gradient descent algorithms". In: *Proceedings of the symposium on learning and data science, Paris*. 2009.
- Gürbüzbalaban, Mert, Asu Ozdaglar, and Pablo Parrilo. "Why random reshuffling beats stochastic gradient descent". In: *arXiv preprint arXiv:1510.08560* (2015).
- HaoChen, Jeffery Z and Suvrit Sra. "Random Shuffling Beats SGD after Finite Epochs". In: *arXiv preprint arXiv:1806.10077* (2018).
- Shamir, Ohad. "Without-replacement sampling for stochastic gradient methods". In: *Advances in Neural Information Processing Systems*. 2016, pp. 46–54.

Questions?