SGD: General Analysis and Improved Rates

Robert M. Gower[1] Nicolas Loizou[3] Xun Qian[2] Alibek Sailanbayev[2] Egor Shulgin[2,4] Peter Richtárik[2,3,4]

[1]Télécom ParisTech [2]KAUST [3]University of Edinburgh [4]MIPT

The Problem

$$x^* = \arg\min_{x\in\mathbb{R}^d} \left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x) \right] \quad (1)$$

We assume f_i are differentiable and f is quasi strongly convex.

Stochastic Reformulation

Stochastic reformulation of (1) is the problem:

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_{v\sim\mathcal{D}} \left[f_v(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n v_i f_i(x) \right]. \quad (2)$$

where $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ ("sampling vector") is any random vector for which

$$\mathbb{E}_{v\sim\mathcal{D}}[v_i] = 1, \quad \forall i \in \{1, 2, \ldots, n\}. \quad (3)$$

- Equivalence: (2) is equivalent to (1) since $\mathbb{E}_{v\sim\mathcal{D}}[f_v] = f$. Also note that $\mathbb{E}_{v\sim\mathcal{D}}[\nabla f_v] = \nabla f$, which can be seen via

$$\mathbb{E}_{v\sim\mathcal{D}}[\nabla f_v] \overset{(2)}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{v\sim\mathcal{D}}[v_i] \nabla f_i = \nabla f. \quad (4)$$

- We propose to solve (1) by applying **SGD** to (2):

$$\boxed{x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)} \quad (5)$$

where $v^k \sim \mathcal{D}$ is sampled i.i.d. and $\gamma^k > 0$ is a stepsize.

Example: Arbitrary Sampling

A sampling is a random set-valued mapping S with values being subsets of $\{1, \ldots, n\}$. A sampling is defined by assigning probabilities to all 2^n subsets of $\{1, \ldots, n\}$.

- A sampling is proper if $p_i \overset{\text{def}}{=} \mathbb{P}[i \in S] > 0$ for all $i \in \{1, \ldots, n\}$.
- Each proper sampling S gives rise to a sampling vector v:

$$v = \mathrm{Diag}(p_1^{-1}, \ldots, p_n^{-1}) \sum_{i\in S} e_i,$$

where e_i is the ith standard unit basis vector in $\mathbb{R}^n$. It is easy to see that $\mathbb{E}[v_i] = 1$. Indeed, just notice that $v_i = p_i^{-1}$ if $i \in S$ and $v_i = 0$ if $i \notin S$.

Main Contributions

- We introduce and study a flexible stochastic reformulation (see (2)) of the finite-sum problem (1), and study SGD applied to this reformulation (see (5)). This way we obtain a wide array of existing and many new variants of SGD for (1).
- We establish linear convergence of SGD applied to the stochastic reformulation. As a by-product, we establish linear convergence of SGD under the arbitrary sampling paradigm [2].
- Our results require very weak assumptions. In particular, we *do not* assume bounded second moment of the gradients for every x (only at x^*; see (8)). We rely on the expected smoothness assumption (7) [3, 4].
- Optimal mini-batch size: We establish formulas for the optimal dependence of the stepsize on the mini-batch size.
- Learning schedule: We provide a formula for when SGD should switch from a constant stepsize to a decreasing stepsize (see (9)).
- Interpolated models. We extend the findings in [5]; and show that optimal mini-batch size is 1 for independent sampling and sampling with replacement.

Assumptions

- **Quasi strong convexity:** f is quasi μ–strongly convex [1]:

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2, \ \forall x \quad (6)$$

- **Expected Smoothness:** There exists $\mathcal{L} \geq 0$ such

$$\mathbb{E}_{v\sim\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] \leq 2\mathcal{L}(f(x) - f(x^*)), \ \forall x. \quad (7)$$

As $\mathcal{L}$ depends on both f and $\mathcal{D}$, we will write $(f, \mathcal{D}) \sim ES(\mathcal{L})$.

- **Finite Gradient Noise**

$$\sigma^2 \overset{\text{def}}{=} \mathbb{E}_{v\sim\mathcal{D}}\left[\|\nabla f_v(x^*)\|^2\right] < \infty. \quad (8)$$

Assumptions (7) and (8) include also some non-convex functions!

Linear Convergence with Fixed Step Size

Assumptions (7) and (8) lead to a bound on the 2nd moment of the stochastic gradient:

Lemma: 2nd moment

If $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and $\sigma < +\infty$ (i.e., if (7) and (8) hold), then

$$\mathbb{E}_{v\sim\mathcal{D}}\left[\|\nabla f_v(x)\|^2\right] \leq 4\mathcal{L}(f(x) - f(x^*)) + 2\sigma^2.$$

The above lemma can now be used to establish a linear convergence result:

Theorem 1

Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$, then SGD (5) satisfies:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}.$$

In particular, with stepsize $\gamma = \min\left\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\right\}$, we have

$$k \geq \max\left\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\right\} \log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right) \Rightarrow \mathbb{E}\|x^k - x^*\|^2 \leq \epsilon.$$

Proof. Let $r^k \overset{\text{def}}{=} x^k - x^*$ and $g^k \overset{\text{def}}{=} \mathbb{E}_k\left[\|\nabla f_{v^k}(x^k)\|^2\right]$.

$$\begin{aligned}\|r^{k+1}\|^2 &\overset{(5)}{=} \|x^k - x^* - \gamma\nabla f_{v^k}(x^k)\|^2 \\ &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2\|\nabla f_{v^k}(x^k)\|^2\end{aligned}$$

Taking expectation conditioned on x^k we obtain:

$$\begin{aligned}\mathbb{E}_k\|r^{k+1}\|^2 &\overset{(4)}{=} \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle + \gamma^2 g^k \\ &\overset{(6)}{\leq} (1 - \gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] + \gamma^2 g^k.\end{aligned}$$

Taking expectations again and using the lemma :

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq (1 - \gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2 \\ &\quad + 2\gamma(2\gamma\mathcal{L} - 1)\mathbb{E}\left[f(x^k) - f(x^*)\right] \\ &\leq (1 - \gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2,\end{aligned}$$

since $2\gamma\mathcal{L} \leq 1$ and $\gamma \leq \frac{1}{2\mathcal{L}}$. Recursively applying the above and summing up the resulting geometric series gives

$$\begin{aligned}\mathbb{E}\|r^k\|^2 &\leq (1 - \gamma\mu)^k \|r^0\|^2 + 2\sum_{j=0}^{k-1}(1 - \gamma\mu)^j \gamma^2\sigma^2 \\ &\leq (1 - \gamma\mu)^k \|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu}.\end{aligned}$$

□

Example: Mini-batch SGD Without Replacement (τ-nice sampling)

- Consider sampling S which picks from all subsets of $\{1, \ldots, n\}$ of cardinality τ, uniformly at random. Then $p_i = \frac{\tau}{n}$ for all i and the sampling vector v is given by:

$$v_i = \begin{cases} \frac{n}{\tau} & i \in S \\ 0 & \text{otherwise.}\end{cases}$$

- SGD (5) then takes the form

$$\boxed{x^{k+1} = x^k - \gamma^k \frac{n}{\tau}\sum_{i\in S^k} \nabla f_i(x^k)}$$

- If each f_i is L_i–smooth and convex, $L_{\max} \overset{\text{def}}{=} \max_i L_i$, and f is L-smooth, then $(f, \mathcal{D}) \sim ES(\mathcal{L})$, where

$$\mathcal{L} \leq \mathcal{L}(\tau) \overset{\text{def}}{=} \frac{n(\tau - 1)}{\tau(n-1)} L + \frac{n - \tau}{\tau(n-1)} L_{\max}$$

- Let $h^* \overset{\text{def}}{=} \frac{1}{n}\sum_i \|\nabla f_i(x^*)\|^2$. Then the gradient noise is

$$\sigma^2 = \sigma^2(\tau) \overset{\text{def}}{=} \frac{h^*}{\tau} \cdot \frac{n - \tau}{n - 1}.$$

- Applying Theorem 1,

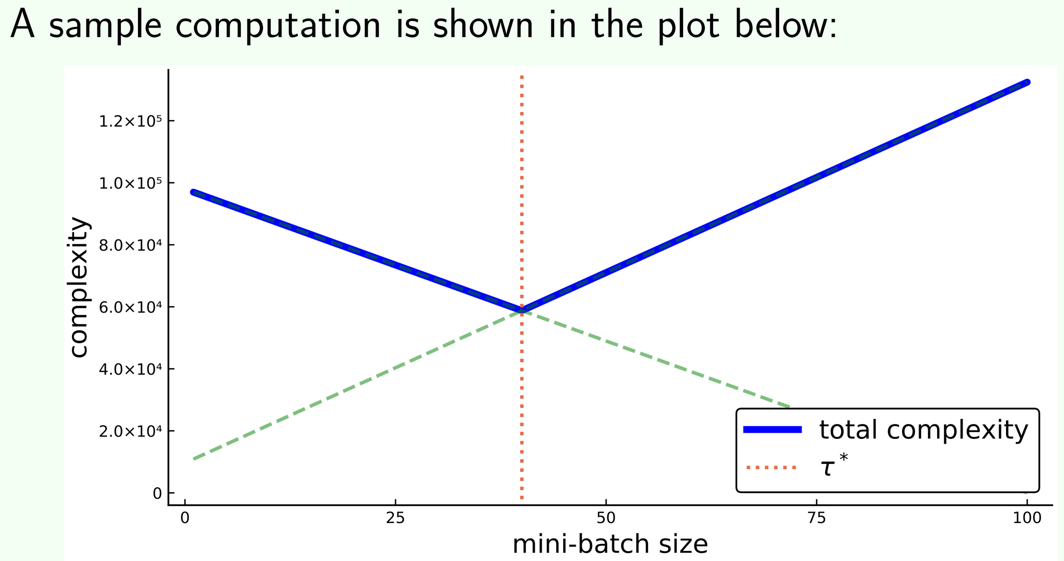
$$k \geq \frac{2(n-\tau)}{\tau(n-1)} \max\left\{\frac{n(\tau-1)}{n-\tau}\frac{L}{\mu} + \frac{L_{\max}}{\mu}, \frac{2h^*}{\epsilon\mu^2}\right\} \log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right),$$

implies $\mathbb{E}\|x^k - x^*\|^2 \leq \epsilon$.

- Theoretically optimal mini-batch size is obtained by minimizing the above bound on k in τ:

$$\tau^* = n\frac{L - L_{\max} + \frac{2}{\epsilon\mu} \cdot h^*}{nL - L_{\max} + \frac{2}{\epsilon\mu} \cdot h^*}.$$

A sample computation is shown in the plot below:



Sublinear Convergence with Constant and Later Decreasing Step Size

In the next theorem we propose a stepsize switching strategy: first use a constant stepsize, and at some point switch to $\mathcal{O}(1/k)$ stepsize. This leads to $\mathcal{O}(1/k)$ rate.

Theorem 2

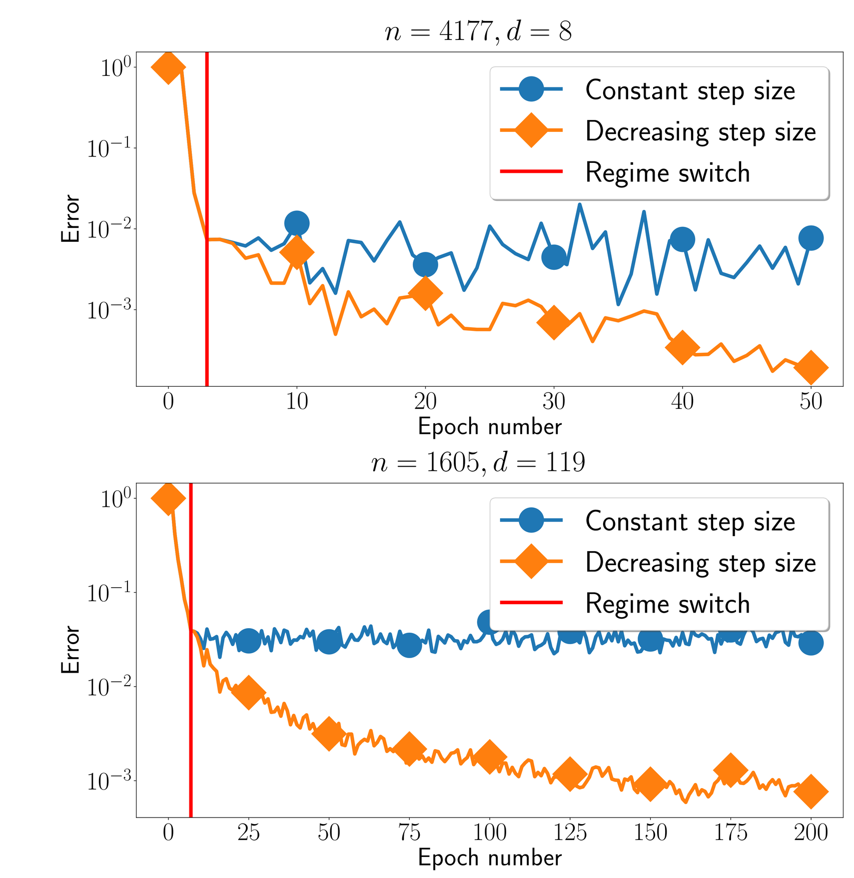
Let $\mathcal{K} \overset{\text{def}}{=} \mathcal{L}/\mu$ and

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for } k \leq 4\lceil\mathcal{K}\rceil \\ \frac{2k+1}{(k+1)^2\mu} & \text{for } k > 4\lceil\mathcal{K}\rceil.\end{cases} \quad (9)$$

If $k \geq 4\lceil\mathcal{K}\rceil$, then SGD iterates given by (5) satisfy:

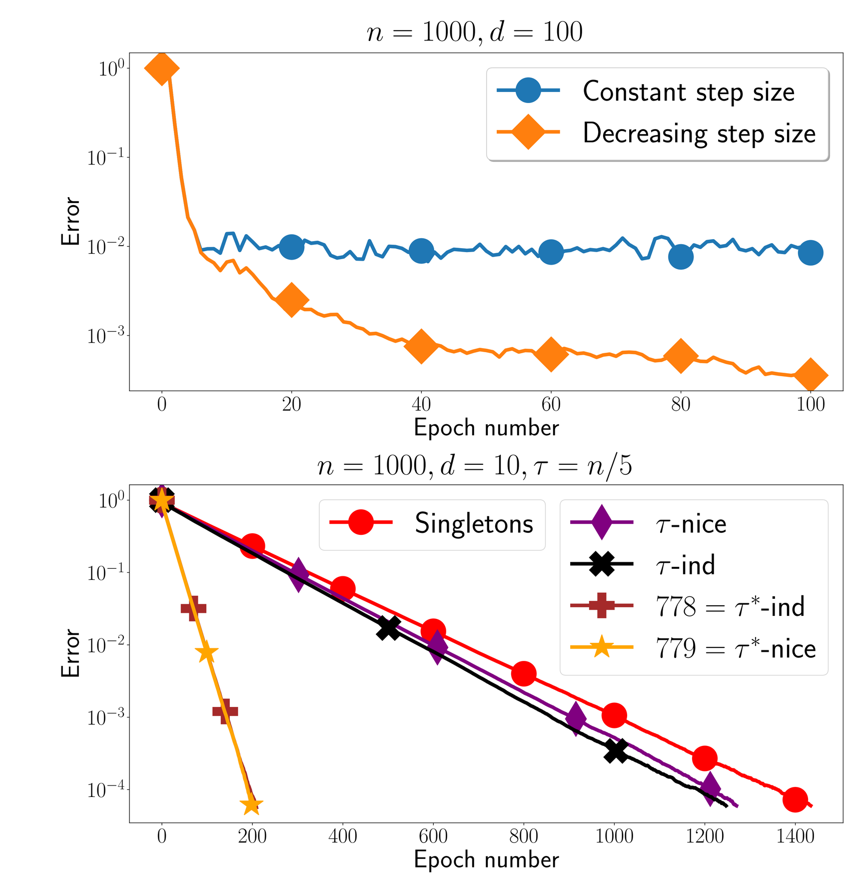
$$\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\sigma^2 8}{\mu^2 k} + \frac{16\lceil\mathcal{K}\rceil^2}{e^2 k^2}\|x^0 - x^*\|^2. \quad (10)$$

Learning Schedule



Constant vs decreasing step size regimes of SGD with $\lambda = 1/n$. *Top:* Ridge regression problem with `abalone`. *Bottom:* Logistic regression with `a1a`. Data from LIBSVM.

PCA (Sum-of-non-convex functions)



Top: Comparison between constant and decreasing step size regimes of SGD for PCA. *Bottom:* Comparison of different sampling strategies of SGD for PCA.

References

[1] Ion Necoara, Yurii Nesterov, and Francois Glineur.
Linear convergence of first order methods for non-strongly convex optimization.
Mathematical Programming, pages 1–39, 2018.

[2] Peter Richtárik and Martin Takáč.
On optimal probabilities in stochastic coordinate descent methods.
Optimization Letters, 10(6):1233–1243, 2016.

[3] Robert M. Gower, Peter Richtárik, and Francis Bach.
Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching.
arxiv:1805.02632, 2018.

[4] Nidham Gazagnadou, Robert Mansel Gower, and Joseph Salmon.
Optimal mini-batch and step sizes for saga.
In *36th International Conference on Machine Learning*, 2019.

[5] Siyuan Ma, Raef Bassily, and Mikhail Belkin.
The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning.
In *35th International Conference on Machine Learning*, 2018.