SGD: General Analysis and Improved Rates

Ref: Gower R M, Loizou N, Qian X, et al. SGD: General Analysis and Improved Rates[J]. arXiv: Learning, 2019.

- Linear convergence on SGD
  - General sampling
  - Expected smoothness assumption
  - Linear convergence rates with strong quasi-convexity (this class includes some non-convex functions as well).
  - Furthermore, do not require the functions f_i to be convex.
- Gradient noise assumption Our analysis does not directly assume a growth condition. Instead, we make use of the remarkably weak expected smoothness assumption.
- Optimal mini-batch size We prove (see Section 4) that this is the case, upto a certain optimal mini-batch size, and provide exact formulas for the dependency of the stepsizes on the mini-batch sizes.
- Learning schedules
  - a closed-form formula for when should SGD switch from a constant stepsize to a decreasing stepsize (see Theorem 3.2).
  - Further, we clearly show how the optimal stepsize (learning rate) increases and the iteration complexity decreases as the mini-batch size increases for both independent sampling and sampling with replacement.
  - We also recover the well known $\frac{L}{\mu}\log(\frac{1}{\epsilon})$ convergence rate of gradient descent (GD) when the mini-batch size is n; this is the first time a generic SGD analysis recovers the correct rate of GD.
- Over-parameterized
  - In the case of over-parametrized models, we extend the findings of Ma et al. (2018) to independent sampling and sampling with replacement by showing that the optimal mini-batch size is 1.
  - Moreover, we provide results in the more general setting where the model is not necessarily over-parametrized.

Stochastic reformulation

Optimization problem

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

where each f_i is smooth.

- **Assumption** Further, assume that f has a unique global minimizer x^* and is μ-strongly quasi-convex ($\mu > 0$):
  $$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2 \tag{2}$$
  for all $x \in \mathbb{R}^d$

- **Definition 1.1** We say that a random vector $v \in \mathbb{R}^n$ drawn from some distribution $\mathcal{D}$ is a sampling vector if its mean is the vector of all ones $\mathbb{E}_{\mathcal{D}}[v_i] = 1, \quad \forall i \in \{1, 2, \ldots, n\}$
    - Based on Definition 1.1, we introduce a stochastic reformulation of (1)
$$\arg\min_{x \in \mathbb{R}^d} \mathbb{E}_{\mathcal{D}} \left[f_v(x) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(x) \right] \tag{4}$$
    - $f_v(x)$ and $\nabla f_v(x)$ are unbiased estimators of $f(x)$ and $\nabla f(x)$

SGD step $$x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k) \tag{6}$$ where $v^k \sim \mathcal{D}$ is sampled i.i.d. at each iteration and $\gamma^k > 0$ is a stepsize.

Expected smoothness

- **Assumption 2.1** (Expected smoothness) We say that f is $\mathcal{L}$-smooth in expectation with respect to distribution $\mathcal{D}$ if there exists $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$ such that
$$\mathbb{E}_{\mathcal{D}} \left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2 \right] \le 2\mathcal{L}(f(x) - f(x^*)) \tag{7}$$
for all $x \in \mathbb{R}^d$. For simplicity, we denote it by $(f, \mathcal{D}) \sim ES(\mathcal{L})$.
- This assumption contains some non-convex cases.

Finite gradient noise

- **Assumption 2.3** (Finite gradient noise) The gradient noise $\sigma = \sigma(f, \mathcal{D})$, defined by
$$\sigma^2 = \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2] \tag{8}$$
is finite.

Key lemma

- **Lemma 2.4** If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then
$$\mathbb{E}_{\mathcal{D}} \left[\|\nabla f_v(x)\|^2 \right] \le 4\mathcal{L}(f(x) - f(x^*)) + 2\sigma^2. \tag{9}$$
    - This Lemma can be proved directly by combining equations (7) and (8).

Main results

- **Theorem 3.1** Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$ for all k. Then iterates of SGD given by (6) satisfy:
$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \le (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu} \tag{10}$$
Hence, given any $\epsilon > 0$, choosing stepwise $\gamma = \min\left\{ \frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2} \right\}$ and $k \ge \max\left\{ \frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2} \right\} \log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon} \right)$, implies $\mathbb{E}\left[\|x^k - x^*\|^2 \right] \le \epsilon$.
    - Proof of Theorem 3.1 Let $r^k = x^k - x^*$. From (6), we have
$$\begin{aligned} \|r^{k+1}\|^2 &= \|x^k - x^* - \gamma^k \nabla f_{v^k}(x^k)\|^2 \\ &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2 \|\nabla f_{v^k}(x^k)\|^2 \end{aligned}$$
Take expectation conditioned on x^k
$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\|r^{k+1}\|^2 &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle + \gamma^2 \mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \\ &\le (1 - \gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] + \gamma^2 \mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \end{aligned}$$
Taking expectation again and using (9)

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\le (1-\gamma\mu)\mathbb{E}\|r^k\|^2 - 2\gamma\mathbb{E}[f(x^k)-f(x^*)] + 4\gamma^2\mathcal{L}\mathbb{E}(f(x)-f(x^*)) + 2\gamma^2\sigma^2 \\ &= (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma(2\gamma\mathcal{L}-1)\mathbb{E}[f(x^k)-f(x^*)] + 2\gamma^2\sigma^2 \\ &\le (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2\end{aligned}$$

Note that $\gamma \le \frac{1}{2\mathcal{L}}$. Recursively we obtain

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

- We control $\mathcal{L}$ and σ via controlling $\mathcal{D}$.
- Furthermore, we can control the additive constant by carefully choosing the step size, as show in Theorem 3.2.

- **Theorem 3.2** (Decreasing stepsizes). Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Let $\mathcal{K} := \frac{\mathcal{L}}{\mu}$ and

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & k \le 4\lceil\mathcal{K}\rceil \\ \frac{2k+1}{(k+1)^2\mu} & k > 4\lceil\mathcal{K}\rceil \end{cases} \tag{14}$$

If $k > 4\lceil\mathcal{K}\rceil$, then SGD iterates given by (6) satisfy:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le \frac{\gamma^2}{\mu^2}\frac{8}{k} + \frac{16\lceil\mathcal{K}\rceil^2}{e^2k^2}\|x^0 - x^*\|^2 \tag{15}$$

- Proof of Theorem 3.2 Let $\gamma_k := \frac{2k+1}{(k+1)^2\mu}$ and let k^* be an integer that satisfies $\gamma_k^* \le \frac{1}{2\mathcal{L}}$. In particular this holds for $k^* \ge \rceil 4\mathcal{K} - 1\rceil$. Note that γ_k is decreasing in k and consequently $\gamma_k \le \frac{1}{2\mathcal{L}}$ for all $k \ge k^*$. This in turn guarantees that (13) holds for all $k \ge k^*$ with γ_k, that is

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\le (1-\gamma\mu)^k\|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu} \\ &= \frac{k^2}{(k+1)^2}\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2k+1)^2}{(k+1)^4}\end{aligned} \tag{51}$$

Then

$$\begin{aligned}(k+1)^2\mathbb{E}\|r^{k+1}\|^2 &\le k^2\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2k+1)^2}{(k+1)^2} \\ &\le k^2\mathbb{E}\|r^k\|^2 + \frac{8\sigma^2}{\mu^2}\end{aligned}$$

Summing from $t = k^*, \ldots, k$

$$\sum_{t=k^*}^{k}[(t+1)^2\mathbb{E}\|r^{t+1}\|^2 - t^2\mathbb{E}\|r^t\|^2] \le \sum_{t=k^*}^{k}\frac{8\sigma^2}{\mu^2} \tag{52}$$

Then

$$(k+1)^2\mathbb{E}\|r^{t+1}\|^2 - (k^*)^2\mathbb{E}\|r^{k^*}\|^2 \le \sum_{t=k^*}^{k}[(t+1)^2\mathbb{E}\|r^{t+1}\|^2 - t^2\mathbb{E}\|r^t\|^2] \le \frac{8\sigma^2(k-k^*)}{\mu^2}$$

We obtain

$$\mathbb{E}\|r^{t+1}\|^2 \le \frac{(k^*)^2}{(k+1)^2}\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2(k+1)^2} \tag{53}$$

For $k \le k^*$ we have that (13) holds, which combined with (53), gives

$$\begin{aligned}\mathbb{E}\|r^{t+1}\|^2 &\le \frac{(k^*)^2}{(k+1)^2}\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2(k+1)^2} \\ &\le \frac{(k^*)^2}{(k+1)^2}\left((1-\gamma\mu)^{k^*}\|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu}\right) + \frac{8\sigma^2(k-k^*)}{\mu^2(k+1)^2} \\ &= \frac{(k^*)^2}{(k+1)^2}\left((1-\frac{\mu}{2\mathcal{L}})^{k^*}\right)\|r^0\|^2 + \frac{\sigma^2}{\mu^2(k+1)^2}\left(8(k-k^*) + \frac{(k^*)^2}{\mathcal{K}}\right)\end{aligned}$$

Choosing k^* that minimizes the second term of above gives $k^* = 4\lceil\mathcal{K}\rceil$, which gives

$$\begin{aligned}\mathbb{E}\|r^{t+1}\|^2 &\le \frac{16\lceil\mathcal{K}\rceil^2}{(k+1)^2}\left(1-\frac{1}{2\mathcal{K}}\right)^{4\lceil\mathcal{K}\rceil}\|r^0\|^2 + \frac{8\sigma^2(k-2\lceil\mathcal{K}\rceil)}{\mu^2(k+1)^2} \\ &\le \frac{16\lceil\mathcal{K}\rceil^2}{e^2(k+1)^2}\|r^0\|^2 + \frac{8\sigma^2}{\mu^2(k+1)}\end{aligned}$$

Specific $\mathcal{D}$

- Notations
  - $e_C := \sum_{i\in C} e_i$ for $C \subseteq \{1,2,\ldots,n\}$
  - A sampling map S (to choose C): $\mathbb{P}[S=C] = p_C, \quad \forall C \subset \{1,2,\ldots,n\}$ where $p_C \ge 0$ and $\sum_{C\subseteq\{1,2,\ldots,n\}} p_C = 1$.
  - A proper sampling S $p_i := \mathbb{P}[i \in S] = \sum_{C:i\in C} p_C \ge 0, \quad \forall I$

We now define practical sampling vector $v = v(S)$ as followings:

- **Lemma 3.3** Let S be a proper sampling, and let $\hat{P} = \mathrm{Diag}(p_1,\ldots,p_n)$. Then the random vector v = v(S) given by
  $v = \hat{P}^{-1} e_S$ (17) is a sampling vector.
- Samplings **Independent sampling**. The sampling S includes every i, independently, with probability $p_i > 0$.
  Partition sampling. A partition $\mathcal{G}$ of $[n]$ is a set consisting of subsets of $[n]$ such that $\cup_{C\in\mathcal{G}} C = [n]$ and $C_i \cap C_j = \emptyset$ for any $C_i, C_j \in \mathcal{G}$ with$i \ne j$. A partition sampling S is a sampling such that $p_C = \mathbb{P}[S=C] > 0$ for all $C \in \mathcal{G}$ and $\sum_{C\in\mathcal{G}} p_C = 1$. **$\tau$-nice sampling**. We say that S is a τ-nice if S samples from all subsets of $[n]$ of cardinality τ uniformly at random. In this case we have that $p_i = \tau$ for all $i \in [n]$. So,$\mathbb{P}[v(S) = \frac{n}{\tau} e_C] = \frac{1}{C_n^\tau}$ for all subsets $C \subseteq \{1,\ldots,n\}$ with τ elements.

Bounding $\mathcal{L}$ and σ^2

- **Assumption 3.4** There exists a symmetric positive definite matrix $M_i \in \mathbb{R}^{d\times d}$ such that
  $$f_i(x+h) \ge f_i(x) + \langle \nabla f_i(x), h\rangle + \frac{1}{2}\|h\|^2_{M_i} \tag{18}$$
  for all $x, h \in \mathbb{R}^d$, and $I \in [n]$, where $\|h\|_{M_i} := \langle M_i h, h\rangle$. In this case we say that f_i is M_i-smooth. Furthermore, we assume that each f_i is convex.
- **Theorem 3.6** Let S be a proper sampling, and $v = v(S)$ (i.e., v is defined by (17). Let f_i be M_i-smooth, and $P \in \mathbb{R}^{n\times n}$ be defined by $P_{ij} = \mathbb{P}[i \in S \,\&j \in\ S]$. Then $(f, \mathcal{D}) \sim ES(\mathcal{L})$, where
  $$\begin{aligned}\mathcal{L} \le \mathcal{L}_{\max} &:= \max_{i\in[n]}\left\{\sum_{C:i\in C}\frac{p_C}{p_i}L_C\right\} \\ &\le \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]} P_{ij}\frac{\lambda_{\max}(M_j)}{p_i p_j}\right\}\end{aligned}$$
  and $L_C := \frac{1}{n}\lambda_{\max}(\sum_{j\in C}\frac{1}{p_j}M_j)$. If $|S| = \tau$, then $L \le \mathcal{L}_{\max} \le L_{\max} = \max_{i\in[n]} \lambda_{\max}(M_i)$
- **Theorem 3.9** Let $h_i = \nabla f_i(x^*)$. Then $\sigma^2 = \frac{1}{n^2}\sum_{i,j\in[n]}\frac{P_{ij}}{p_i p_j}\langle h_i, h_j\rangle$.