# Optimization for Machine Learning

## Stochastic Gradient pt 2

October 19, 2023

Today: Theory + Exercise

Monday: Advanced methods + Lab (bring laptops if you can)

# THEORY OF STOCHASTIC GRADIENT

**Setup**

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

$f_i : \mathbb{R}^d \to \mathbb{R}$ $C^1$ and $f_i$ depends on the $i$th point in a dataset of size $m$ (with $m \gg 1$)

Stochastic gradient iteration:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$$

$i_k$ random index in $\{1, \dots, m\}$

$\alpha_k > 0$ stepsize / learning rate

Gradient descent (GD)

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) = x_k - \frac{\alpha_k}{m} \sum_{i=1}^{m} \nabla f_i(x_k)$$

↳ For GD, we can prove convergence rates when the function $f$ is $C^{1,1}_L$ ($C^1$ + $L$-Lipschitz continuous gradient):

$$\forall K \geq 1, \quad \min_{0 \leq k \leq K-1} \|\nabla f(x_k)\| \leq O\left(\frac{1}{\sqrt{K}}\right)$$

If in addition $f$ is $\mu$-strongly convex, can show

$$\forall K \geq 1, \quad f(x_K) - \min_{x \in \mathbb{R}^d} f(x) \leq O\left(\left(1 - \frac{\mu}{L}\right)^K\right)$$

Q) Can we prove convergence rates for SG?

## (1) Analysis in the strongly convex case

Assumption: $f$ is $\underbrace{C^{1,1}_L}_{\text{"L-smooth"}}$ and $\mu$-strongly convex for some $\mu > 0$, $L > 0$

↳ Since $f$ is $C^1$ and $\mu$-strongly convex,

(1) $\forall (x,z) \in (\mathbb{R}^d)^2, \quad f(z) \geq f(x) + \nabla f(x)^T (z-x) + \frac{\mu}{2} \|z-x\|^2$

↳ Since $f$ is $C^{1,1}_L$

(2) $\forall (x,z) \in (\mathbb{R}^d)^2, \quad f(z) \leq f(x) + \nabla f(x)^T (z-x) + \frac{L}{2} \|z-x\|^2$

$\Rightarrow L \geq \mu$

↳ Applying (2) with $x = x_k$ and $z = x_{k+1}$ (two iterates from the stochastic gradient method) gives

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L}{2} \alpha_k^2 \|\nabla f_{i_k}(x_k)\|^2$$

Assumptions on the stochastic gradients

the random indices $\{i_0, i_1, \dots\}$ are drawn so that $\forall k \geq 1$

i) $i_k$ is drawn independently of $i_0, \dots, i_{k-1}$

ii) $\mathbb{E}_{i_k}\left[\nabla f_{i_k}(x_k)\right] = \nabla f(x_k)$

iii) $\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2\right] \leq \|\nabla f(x_k)\|^2 + \sigma^2, \quad \sigma^2 \geq 0$

Ex) If $\{i_k\}$ are drawn uniformly at random, then i) and ii) are satisfied (and iii) is under additional assumptions)

Under these assumptions on $\{i_k\}$, we obtain for any $k$:

$$\mathbb{E}_{i_k}\left[f(x_{k+1})\right] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{L}{2}\alpha_k^2 \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\sigma^2$$

↳ To turn this inequality into a convergence rate, we specify the choice of the stepsize

Th) Suppose that we run $K$ iterations of SG with $\alpha_k = \frac{1}{L}$ $\forall k$

Then,

$$\mathbb{E}\left[f(x_K) - \underbrace{f^*}_{\min_{x\in\mathbb{R}^d} f(x)}\right] \leq \underbrace{\frac{\sigma^2}{2\mu}}_{\text{constant w.r.t. } K} + \underbrace{\underbrace{\left(1-\frac{\mu}{L}\right)^K}_{\text{rate of convergence for GD}}\left[f(x_0) - f^* - \frac{\sigma^2}{2\mu}\right]}_{\xrightarrow[K\to\infty]{} 0}$$

Proof: $\forall\, k \leq K$, we have shown

$$\mathbb{E}_{i_k}\left[f(x_{k+1})\right] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\sigma^2$$

Using $\alpha_k = \frac{1}{L}$, the inequality becomes

$$\mathbb{E}_{i_k}\left[f(x_{k+1})\right] \leq f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\sigma^2}{2L}$$

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\sigma^2}{2L}$$

Subtracting $f^*$ on both sides, we get

$$\mathbb{E}_{i_k}\left[ f(x_{k+1}) - f^* \right] \leq f(x_k) - f^* - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\sigma^2}{2L}$$

Using that $f$ is $\mu$-strongly convex, we have

$$\|\nabla f(x_k)\|^2 \geq 2\mu\left( f(x_k) - f^* \right)$$

(as a consequence of (1))

Hence,

$$\mathbb{E}_{i_k}\left[ f(x_{k+1}) - f^* \right] \leq f(x_k) - f^* - \frac{\mu}{L}\left( f(x_k) - f^* \right) + \frac{\sigma^2}{2L}$$

$$\begin{aligned}
\mathbb{E}_{i_k}\left[ f(x_{k+1}) - f^* \right] &\leq \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* \right) + \frac{\sigma^2}{2L} \\
&= \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* - \frac{\sigma^2}{2\mu} \right) \\
&\quad + \left(1 - \frac{\mu}{L}\right)\frac{\sigma^2}{2\mu} + \frac{\sigma^2}{2L} \\
&= \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* - \frac{\sigma^2}{2\mu} \right) \\
&\quad + \frac{\sigma^2}{2\mu} - \frac{\sigma^2}{2L} + \frac{\sigma^2}{2L} \\
&= \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* - \frac{\sigma^2}{2\mu} \right) + \frac{\sigma^2}{2\mu}
\end{aligned}$$

$$\mathbb{E}_{i_k}\left[ f(x_{k+1}) - f^* \right] \leq \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* - \frac{\sigma^2}{2\mu} \right) + \frac{\sigma^2}{2\mu}$$

$$\mathbb{E}_{i_k}\Big[ \underbrace{f(x_{k+1}) - f^* - \frac{\sigma^2}{2\mu}}_{\text{"Lyapunov function"} \xrightarrow[k\to\infty]{} 0} \Big] \leq \left(1 - \frac{\mu}{L}\right)\left( f(x_k) - f^* - \frac{\sigma^2}{2\mu} \right) \qquad \forall \text{ iteration } k \geq 1$$

We can apply this inequality recursively by taking the appropriate expected value

$$\mathbb{E}_{i_{k-1}}\left[\mathbb{E}_{i_k}\left[f(x_{k+1}) - f^* - \frac{\sigma^2}{2\mu}\right]\right] \leq \mathbb{E}_{i_{k-1}}\left[\left(1-\frac{\mu}{L}\right)\left(f(x_k) - f^* - \frac{\sigma^2}{2\mu}\right)\right]$$

$$= \left(1-\frac{\mu}{L}\right)\mathbb{E}_{i_{k-1}}\left[f(x_k) - f^* - \frac{\sigma^2}{2\mu}\right]$$

$$\leq \left(1-\frac{\mu}{L}\right)\left(1-\frac{\mu}{L}\right)\left(f(x_{k-1}) - f^* - \frac{\sigma^2}{2\mu}\right)$$

$$= \left(1-\frac{\mu}{L}\right)^2\left(f(x_{k-1}) - f^* - \frac{\sigma^2}{2\mu}\right)$$

As a result, for any $K \geq 1$, we have

$$\mathbb{E}\left[f(x_K) - f^* - \frac{\sigma^2}{2\mu}\right] \leq \left(1-\frac{\mu}{L}\right)\mathbb{E}\left[f(x_{K-1}) - f^* - \frac{\sigma^2}{2\mu}\right]$$

Expected value over all $\{i_k\}$

$$\leq \left(1-\frac{\mu}{L}\right)^2\mathbb{E}\left[f(x_{K-2}) - f^* - \frac{\sigma^2}{2\mu}\right]$$

$$\leq \left(1-\frac{\mu}{L}\right)^K\mathbb{E}\left[f(x_0) - f^* - \frac{\sigma^2}{2\mu}\right]$$

$$= \left(1-\frac{\mu}{L}\right)^K\left(f(x_0) - f^* - \frac{\sigma^2}{2\mu}\right)$$

Adding $\frac{\sigma^2}{2\mu}$ on both sides gives the final result

Interpretation:

- For GD, the convergence rate guarantees that $f(x_k) - f^* \xrightarrow[k\to\infty]{} 0$ deterministically and that the quantity $f(x_k) - f^*$ decreases at least as fast as $\left(1-\frac{\mu}{L}\right)^k$
- For SG

$$\mathbb{E}\left[f(x_K) - f^*\right] \leq \frac{\sigma^2}{2\mu} + \left(1-\frac{\mu}{L}\right)^K\left(f(x_0) - f^* - \frac{\sigma^2}{2\mu}\right)$$

↳ Does not show convergence of $f(x_K) - f^*$ (even in expectation)

↳ Does show that $\{f(x_k)\}$ converges to a neighborhood of the optimal value in expectation

$$\frac{\sigma^2}{2\mu} + \left(1-\frac{\mu}{L}\right)^k \left(f(x_0) - f^* - \frac{\sigma^2}{2\mu}\right) \xrightarrow[k\to\infty]{} \frac{\sigma^2}{2\mu}$$

$$0 \leq \mathbb{E}\left[f(x_k) - f^*\right] \xrightarrow[k\to\infty]{} \left[0, \frac{\sigma^2}{2\mu}\right)$$

"convergence within an interval"

$$\mathbb{E}\left[f(x_k)\right] \to \left[f^*, f^* + \frac{\sigma^2}{2\mu}\right)$$

Comments on this result

- Large $\sigma^2$ $\Rightarrow$ Large interval (possible convergence to a value far from the optimum)
- $\sigma^2 \approx 0$ $\Rightarrow$ CV in expectation to a value close to $f^*$
- $\mathbb{E}\left[f(x_k)\right]$ $\Rightarrow$ In practice, the values $\{f(x_k)\}$ will oscillate around the limit of $\mathbb{E}\left[f(x_k)\right]$

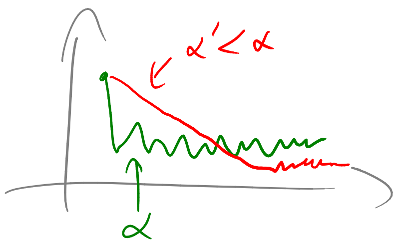
Remark: The result of the theorem generalizes to any fixed stepsize $\alpha_k = \alpha \in \left(0, \frac{1}{L}\right]$:

$$\mathbb{E}\left[f(x_K) - f^*\right] \leq \frac{\alpha L \sigma^2}{2\mu} + (1-\alpha\mu)^K \left[f(x_0) - f^* - \frac{\alpha L \sigma^2}{2\mu}\right]$$

- There $\mathbb{E}\left[f(x_k)\right] \to \left[f^*, f^* + \frac{\alpha L \sigma^2}{2\mu}\right)$ at a rate $(1-\alpha\mu)^K$
- Choosing $\alpha$ small means that $\left[f^*, f^* + \frac{\alpha L \sigma^2}{2\mu}\right)$ will be small (good!) but the convergence will occur at a rate $(1-\alpha\mu)^K$ with $1-\alpha\mu \approx 1$ (bad)

↳ The analysis above partly explains one strategy in learning rate scheduling which consists in running SG with fixed $\alpha > 0$ until the average function value appears to stall

then decreasing $\alpha$ (e.g. by a factor of 2) and run SG again until the same phenomenon is observed




⇒ Such strategies can be analyzed using similar tools as those used to analyze SG with decreasing stepsizes

$\{\alpha_k\}$ fixed in advance

$\alpha_k \to 0$

**Th)** Suppose that we run $K$ iterations of SG with $\alpha_k = \frac{\beta}{k+\gamma}$ where $\alpha_0 = \frac{\beta}{\gamma} \leq \frac{1}{L}$ and $\beta > \frac{1}{\mu}$. Then,

$$\mathbb{E}\left[f(x_K) - f^*\right] \leq O\left(\frac{1}{K+\gamma}\right) = O\left(\frac{1}{K}\right)$$

↳ Unlike in the constant stepsize case, this result guarantees that $f(x_K) - f^*$ converges to $0$ in expectation

↳ The rate of convergence is $\frac{1}{K}$, which is worse than $\left(1-\frac{\mu}{L}\right)^K$ that we had for SG with constant step size and for GD

Comparing the two rates :

SG $\quad \mathbb{E}\left[f(x_k) - f^*\right] \leq O\left(\frac{1}{k}\right)$

GD $\quad f(x_k) - f^* \leq O\left(\left(1-\frac{\mu}{L}\right)^k\right)$

↳ If we compare the rates of SG and GD with the same number of iterations, then the results are better for GD (deterministic + better rate)

↳ BUT an iteration of SG is less expensive than an iteration of GD in terms of accesses to data points

⟹ A fair comparison should use a metric that involves the number of accesses to data points

↳ Epochs

1 epoch = cost of $n$ accesses to a data point

1 iteration of GD costs 1 epoch

SG costs $\frac{1}{n}$ epoch

Consider now that we run GD and SG for $N_E \geq 1$ epochs

$N_E$ epochs $\equiv N_E$ GD iterations

$$f(x_{N_E}) - f^* \leq O\left(\left(1-\frac{\mu}{L}\right)^{N_E}\right)$$

$N_E$ epochs $\equiv m N_E$ SG iterations

$$\mathbb{E}\left[f(x_{mN_E}) - f^*\right] \leq O\left(\frac{1}{mN_E}\right)$$

If $m \gg N_E$, then $\frac{1}{mN_E} \ll \left(1-\frac{\mu}{L}\right)^{N_E}$

SG has a better rate (in expectation) than GD in that setting

## ② Extensions

→ To the convex and nonconvex cases

• Convex, $C_L^{1,1}$ $f$: similar conclusions than in the strongly convex case

$\mathbb{E}\left[f(x_K) - f^*\right]$ $\begin{cases} \text{Fixed } \alpha_k : & \text{CV to an interval in } O\left(\frac{1}{K}\right) \text{ (same rate than GD)} \\ \text{Decreasing } \alpha_k : & \text{CV to } f^* \text{ in } \underbrace{O\left(\frac{1}{\sqrt{K}}\right)}_{\text{better rate than GD when } n \gg \text{number of epochs}} \text{ (worse than GD)} \end{cases}$

GD: $\min_{0 \le k \le K-1} \|\nabla f(x_k)\| \le O\left(\frac{1}{\sqrt{K}}\right)$

⇕

$\min_{0 \le k \le K-1} \|\nabla f(x_k)\|^2 \le O\left(\frac{1}{K}\right)$

• Nonconvex, $C_L^{1,1}$ $f$

Guarantee on $\underbrace{\mathbb{E}\left[\frac{1}{\sum_{k=0}^{K-1}\alpha_k} \sum_{k=0}^{K-1} \alpha_k \|\nabla f(x_k)\|^2\right]}_{\text{weighted average of the gradient norms}}$

↳ For constant $\alpha_k$:

$\frac{1}{K} \sum_{k=0}^{K} \|\nabla f(x_k)\|^2$ converges to an interval.

rate $O\left(\frac{1}{K}\right)$

↳ For decreasing $\alpha_k$:

Rate is $O\left(\frac{1}{\sqrt{K}}\right)$

$\mathbb{E}\left[\frac{1}{\sum \alpha_k} \sum \alpha_k \|\nabla f(x_k)\|^2\right] \to 0$

Corollary of the result for decreasing $\alpha_k$:

• $\|\nabla f(x_{k(K)})\| \to 0$ in probability as $K \to \infty$

$\forall K \geq 1$, $k(K)$ is a random index in $0, \dots, K-1$

$$P(k(K) = j) = \frac{\alpha_j}{\sum_{k=0}^{K-1} \alpha_k}$$

↳ For SG on nonconvex functions, you can get guarantees on a random sequence drawn from the iterates

→ To batch SG methods

$$x_{k+1} = x_k - \frac{\alpha_k}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k)$$

$S_k$ is a set indices drawn randomly in $\{1, \dots, n\}$ (with or without replacement)

- By adapting the assumptions on $\{i_k\}$ for SG to assumptions on $\{S_k\}$, can prove similar rates for the batch variants

⇒ The results still differ from SG, and they help in explaining how batch methods have lower variance.

Ex) Constant step $\alpha_k = \frac{1}{L}$, batch size of $n_b \in [\![1, n]\!]$, $f$ $C^{1,1}_L$, $\mu$-strongly convex, $S_k = n_b$ iid indices drawn so as to satisfy i) ii) iii)

Then, after $K$ iterations,

$$\mathbb{E}[f(x_K) - f^*] \leq \frac{\sigma^2}{2\mu n_b} + \left(1 - \frac{\mu}{L}\right)^K \left[ f(x_0) - f^* - \frac{\sigma^2}{2\mu n_b} \right]$$

(Essentially $\sigma^2 \to \frac{\sigma^2}{m_b}$)

1) What does this result imply on the convergence of $\mathbb{E}[f(x_k) - f^*]$? Is the result better than that of SG?

2) Suppose that we run SG with $\alpha_k = \frac{1}{m_b L}$, where $m_b$ is the batch size used in batch SG. What convergence rate do we obtain? Is it better than that of batch SG?