Labs
Optimization for Machine Learning
Spring 2024

Saarland University
CISPA Helmholtz Center for Information Security
Sebastian Stich
TAs: Yuan Gao & Xiaowen Jiang
https://cms.cispa.saarland/optml24/

Problem Set 9 — Solutions (Variance Reduction)

In two steps of the solutions, we use the following inequality on the squared Euclidean norm.

Lemma 1 (Inequality on the squared norm). For any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$\|\mathbf{a}+\mathbf{b}\|_2^2 \le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2 .$$

Proof.

$$\begin{aligned}
\|\mathbf{a}+\mathbf{b}\|_2^2 &= \langle \mathbf{a}+\mathbf{b}, \mathbf{a}+\mathbf{b}\rangle \\
&= \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle \\
&\le \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 + 2\|\mathbf{a}\|_2\|\mathbf{b}\|_2 && \text{By Cauchy-Schwarz inequality} \\
&\le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2. && \text{By AM-GM inequality}
\end{aligned}$$

□

1 Bound of Variance Lemma

Prove Lemma 9.2 (Property of smoothness) and Lemma 9.3 (Bound of variance) from the slides.

Lemma 9.2 (Property of Smoothness). Let $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex and L_i-smooth function and F has a global minimum $\mathbf{x}^\star$. Let $L_{max} = \max\{L_1, \dots, L_n\}$. Then, for any $\mathbf{x} \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|_2^2 \le 2L_{max}\left(F(\mathbf{x}) - F(\mathbf{x}^\star)\right).$$

Proof. For any $i \in \{1, \dots, n\}$, convexity and L_i-smoothness of f_i imply

$$f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) \le f_i(\mathbf{x}) \le f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) + \frac{L_i}{2}\|\mathbf{x}-\mathbf{x}^\star\|_2^2. \tag{1}$$

We consider the function $g_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\mathbf{x}^\star) - \nabla f_i(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star)$. The convexity of f_i implies $g_i \ge 0$. Additionally, g_i is the sum of f_i and an affine function and thus also L_i-smooth[1]. Applying sufficient decrease to g_i shows that

$$g_i\left(\mathbf{x} - \frac{1}{L_i}\nabla g_i(\mathbf{x})\right) \le g_i(x) - \frac{1}{2L_i}\|\nabla g_i(\mathbf{x})\|_2^2.$$

By the non-negativity of g_i and the definition of L_{max} we then have

$$g_i(\mathbf{x}) \ge g_i\left(\mathbf{x} - \frac{1}{L_i}\nabla g_i(\mathbf{x})\right) + \frac{1}{2L_i}\|\nabla g_i(\mathbf{x})\|_2^2 \ge \frac{1}{2L_i}\|\nabla g_i(\mathbf{x})\|_2^2 \ge \frac{1}{2L_{max}}\|\nabla g_i(\mathbf{x})\|_2^2$$

Reinserting the definition of $g_i(\mathbf{x})$ shows that

$$f_i(\mathbf{x}) - f_i(\mathbf{x}^\star) - \nabla f_i(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) \ge \frac{1}{2L_{max}}\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|_2^2.$$

Summing these inequalities over $i = 1, \dots, n$ and dividing by n yields

$$F(\mathbf{x}) - F(\mathbf{x}^\star) - \nabla F(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) \ge \sum_{i=1}^n \frac{1}{2L_{max}n}\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|_2^2.$$

By assumption, $\mathbf{x}^\star$ is a global minimum of F and thus $\nabla F(\mathbf{x}^\star) = 0$. The result then follows, by multiplying the above inequality with $2L_{max}$ □

[1] An affine function is 0-smooth by Lemma 3.4 and L_i-smoothness of the sum follows by Lemma 3.5.

Lemma 9.3 (Bound on Variance). Let $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$, where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex and L_i-smooth function and F has a global minimum $\mathbf{x}^\star$. Let $L_{max} = \max\{L_1, \ldots, L_n\}$ and $\tilde{\mathbf{x}}, \mathbf{x}_t \in \mathbb{R}^d$. Denote $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$, where i_t is sampled uniformly from $\{1, \ldots, n\}$. Then

$$\mathbb{E}_{i_t}\left[\|\mathbf{g}_t\|_2^2\right] \leq 4L_{max}(F(\mathbf{x}_t) - F(\mathbf{x}^\star)) + 4L_{max}(F(\tilde{\mathbf{x}}) - F(\mathbf{x}^\star))$$

Proof. We have

$$\begin{aligned}
\|\mathbf{g}_t\|_2^2 &= \|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})\|_2^2 \\
&= \|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^\star) + \nabla f_{i_t}(\mathbf{x}^\star) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})\|_2^2 \\
&\leq 2\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^\star)\|_2^2 + 2\|\nabla f_{i_t}(\mathbf{x}^\star) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})\|_2^2,
\end{aligned}$$

by Lemma 1. Lemma 9.2 allows us to directly bound the expectation of the first term by

$$\mathbb{E}_{i_t}\left[2\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}^\star)\|_2^2\right] = \frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}^\star)\|_2^2 \leq 4L_{max}(F(\mathbf{x}_t) - F(\mathbf{x}^\star))$$

For the second term, we apply the following result from probability theory[2]

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2^2\right] \leq \mathbb{E}\left[\|\mathbf{X}\|_2^2\right]$$

with $\mathbf{X} = \nabla f_{i_t}(\mathbf{x}^\star) - \nabla f_{i_t}(\tilde{\mathbf{x}})$. We compute

$$\mathbb{E}_{i_t}[\mathbf{X}] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}^\star) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{\mathbf{x}}) = \nabla F(\mathbf{x}^\star) - \nabla F(\tilde{\mathbf{x}}) = 0 - \nabla F(\tilde{\mathbf{x}}).$$

So the second term is exactly of the form $\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2^2$ and we can bound its expectation by

$$\begin{aligned}
\mathbb{E}_{i_t}\left[2\|\nabla f_{i_t}(\mathbf{x}^\star) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})\|_2^2\right] &\leq 2\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\mathbf{x}^\star) - \nabla f_{i_t}(\tilde{\mathbf{x}})\|_2^2\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{\mathbf{x}}) - \nabla f_i(\mathbf{x}^\star)\|_2^2 \\
&\leq 4L_{max}(F(\tilde{\mathbf{x}}) - F(\mathbf{x}^\star)),
\end{aligned}$$

where the last inequality follows again by Lemma 10.2. Combining the two bounds proves the statement. □

2 Loopless SVRG

2.1 Decrease Lemma

1. Plugging in the definition of the update, we get

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2] &= \mathbb{E}[||\mathbf{x}_t - \mathbf{x}^\star - \eta g_t||^2] \\
&= ||\mathbf{x}_t - \mathbf{x}^\star||^2 + \mathbb{E}[2\eta\langle g_t, \mathbf{x}^\star - \mathbf{x}_t\rangle] + \eta^2\mathbb{E}[||g_t||^2]\ .
\end{aligned}$$

Note that g_t is unbiased, i.e. $\mathbb{E}[g_t] = \nabla f(\mathbf{x}_t)$. We get

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2] &= ||\mathbf{x}_t - \mathbf{x}^\star||^2 + 2\eta\langle\nabla f(\mathbf{x}_t), \mathbf{x}^\star - \mathbf{x}_t\rangle + \eta^2\mathbb{E}[||g_t||^2] \\
&\overset{\text{strong convexity}}{\leq} ||\mathbf{x}_t - \mathbf{x}^\star||^2 + 2\eta(f^\star - f(\mathbf{x}_t) - \frac{\mu}{2}||\mathbf{x}_t - \mathbf{x}^\star||^2) + \eta^2\mathbb{E}[||g_t||^2] \\
&= (1 - \mu\eta)||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\eta(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + \eta^2\mathbb{E}[||g_t||^2]\ .
\end{aligned}$$

[2] A possible proof of this inequality is

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2^2\right] &= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right] \\
&= \mathbb{E}\left[\mathbf{X}^\top\mathbf{X} - 2\mathbb{E}[\mathbf{X}]^\top\mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{X}]^\top\mathbb{E}[\mathbf{X}]\right] \\
&= \mathbb{E}\left[\|\mathbf{X}\|_2^2\right] - \|\mathbb{E}[\mathbf{X}]\|_2^2 \\
&\leq \mathbb{E}\left[\|\mathbf{X}\|_2^2\right]
\end{aligned}$$

2. With the same proof procedure for Lemma 9.3, we get

$$\mathbb{E}[||g_t||^2] \leq 4L\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) + 2\mathbb{E}[||\nabla f_i(\mathbf{w}_t) - \nabla f_i(\mathbf{x}^\star)||^2] \; .$$

Plugging the definition of D_t, we get the claim.

2.2 Decrease of the Lyapunov function

1. Note that $\mathbb{E}[\mathbf{w}_{t+1}] = p\mathbf{x}_t + (1-p)\mathbf{w}_t$. It follows that

$$\begin{aligned}\mathbb{E}[D_{t+1}] &= (1-p)D_t + p\frac{4\eta^2}{pn}\sum_{i=1}^{n} ||\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^\star)||^2 \\ &\leq (1-p)D_t + 8L\eta^2\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \; .\end{aligned}$$

The last inequality is due to the smoothness of f.

2. Combine the previous statements together, we get

$$\begin{aligned}\mathbb{E}[||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2 + D_{t+1}] &\leq (1-\mu\eta)||\mathbf{x}_t - \mathbf{x}^\star||^2 + 2\eta\big(f^\star - f(\mathbf{x}_t)\big) + \eta^2\mathbb{E}[||g_t||^2] \\ &\quad + (1-p)D_t + 8L\eta^2\big(f(\mathbf{x}_t) - f^\star\big) \\ &\leq (1-\mu\eta)||\mathbf{x}_t - \mathbf{x}^\star||^2 + (1-p)D_t + (2\eta - 8L\eta^2)\big(f^\star - f(\mathbf{x}_t)\big) \\ &\quad + \eta^2\Big(4L\big(f(\mathbf{x}_t) - f^\star\big) + \frac{p}{2\eta^2}D_t\Big) \\ &= (1-\mu\eta)||\mathbf{x}_t - \mathbf{x}^\star||^2 + (1-\frac{p}{2})D_t + (2\eta - 12L\eta^2)\big(f^\star - f(\mathbf{x}_t)\big)\end{aligned}$$

By picking $\eta \leq \frac{1}{6L}$, we get according to the definition of Φ_t,

$$\mathbb{E}[\Phi_{t+1}] \leq (1-\eta\mu)||\mathbf{x}_t - \mathbf{x}^\star||^2 + (1-\frac{p}{2})D_t \; .$$

2.3 Complexity

1. From the previous display, we get

$$\mathbb{E}[\Phi_t] \leq \max\{1-\eta\mu, 1-\frac{p}{2}\}^t\Phi_0 \; .$$

Clearly, the optimal choice of η is $\frac{1}{6L}$. In terms of total number of stochastic gradient calls, Loopless SVRG calls the stochastic gradient oracle in expectation $2+pn$ times in each iteration. Combining it with the iteration complexity, we get the total complexity $\mathcal{O}\big([(1+pn)*(\frac{L}{\mu}+\frac{1}{p})]\log(1/\epsilon)\big)$. Note that a simple choice of $p=\frac{1}{n}$ gives the optimal complexity $\mathcal{O}\big((n+\frac{L}{\mu})\log(1/\epsilon)\big)$.