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, June 18, 2024 (Variance Reduction)

1 Bound of Variance Lemma

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

Hint for 9.2: For any $i \in \{1, \dots, n\}$, convexity and L_i-smoothness of f_i imply

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

Hint for 9.3: Use that

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

2 Loopless SVRG method

We now consider removing the outer loop present in the SVRG method and instead use a probabilistic update of the full gradient. The resulting loopless SVRG method is presented in Algorithm 1.

Algorithm 1 Loopless SVRG

Require: stepsize $\eta > 0$, probability $p \in (0, 1]$
1: **set** $\mathbf{x}_0 = \mathbf{w}_0 \in \mathbb{R}^d$
2: **for** $t = 0, 1, 2...$ **do**
3: sample $i \in \{1, 2, ..., n\}$ uniformly at random
4: $g_t = \nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)$
5: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta g_t$
6: $\mathbf{w}_{t+1} = \begin{cases} \mathbf{x}_t & \text{with probability } p \\ \mathbf{w}_t & \text{with probability } 1-p \end{cases}$
7: **end for**

As we shall see in this exercise, the simple choice $p = 1/n$ leads to complexity identical to that of the original SVRG method, while the proof is much simpler. A key role in the analysis is played by the *gradient learning* quantity defined as

$$D_t := \frac{4\eta^2}{pn} \sum_{i=1}^{n} ||\nabla f_i(\mathbf{w}_t) - \nabla f_i(\mathbf{x}^\star)||^2 \,.$$

We assume $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ is μ-strongly convex and each f_i is L-smooth henceforth.

2.1 Decrease Lemma

1. Prove that

$$\mathbb{E}[||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2] \le (1 - \mu\eta)||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\eta(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + \eta^2 \mathbb{E}[||g_t||^2] \,.$$

2. Show that

$$\mathbb{E}[||g_t||^2] \leq 4L\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) + \frac{p}{2\eta^2} D_t \ .$$

2.2 Decrease of the Lyapunov function

1. Prove that

$$\mathbb{E}[D_{t+1}] \leq (1-p)D_t + 8L\eta^2\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \ .$$

2. Define the Lyapunov function $\Phi_t := ||\mathbf{x}_t - \mathbf{x}^\star||^2 + D_t$. Show that with a properly chosen stepsize, it holds

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

2.3 Complexity

1. Prove that with a properly chosen stepsize, it holds

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

2. What can you say about the complexity of the total gradient computations?