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 4 — Solutions (Stochastic Gradient Descent)

Stochastic Gradient Descent

Exercise (Sigma-star Assumption).

Solution:

1. According to the hint, let us denote $g_i := f_i - \langle \mathbf{x}, \nabla f_i(\mathbf{y}) \rangle$. Since g_i is convex and $\nabla g_i(\mathbf{y}) = \nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{y}) = 0$, it holds that $\mathbf{y}$ is a minimizer of g_i. Note that g_i is L-smooth. Recall that by sufficient decrease property, we have $g_i(\mathbf{y}) \leq g_i(\mathbf{x} - \frac{1}{L}\nabla g_i(\mathbf{x})) \leq g_i(\mathbf{x}) - \frac{1}{2L}||\nabla g_i(\mathbf{x})||^2$. It remains to plug in the definition of g_i and ∇g_i.

2. $$\begin{aligned}
\mathbb{E}_i||\nabla f_i(\mathbf{x})||^2 &= \mathbb{E}_i[||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)||^2] \\
&\leq 2\mathbb{E}_i[||\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)||^2] + 2\mathbb{E}_i[||\nabla f_i(\mathbf{x}^\star)||^2] \\
&\leq 2\mathbb{E}_i[2L(f_i(\mathbf{x}) - f_i(\mathbf{x}^\star) - \langle \nabla f_i(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle)] + 2\sigma_\star^2 \\
&= 4L(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 2\sigma_\star^2
\end{aligned}$$

   where in the first inequality, we use $||\mathbf{a} + \mathbf{b}||^2 \leq 2||\mathbf{a}||^2 + 2||\mathbf{b}||^2$, and in the second inequality, we use the one given in the first question.

3. From the recursion of SGD, we have:

   $$\begin{aligned}
\mathbb{E}_{i_t}[||\mathbf{x}_{t+1} - \mathbf{x}^\star||^2] &= ||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\gamma\mathbb{E}_{i_t}[\langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle] + \gamma^2\mathbb{E}_{i_t}[||\nabla f_{i_t}(\mathbf{x}_t)||^2] \\
&= ||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\gamma\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \gamma^2\mathbb{E}_{i_t}[||\nabla f_{i_t}(\mathbf{x}_t)||^2] \\
&\leq ||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}||\mathbf{x}_t - \mathbf{x}^\star||^2) + \gamma^2\mathbb{E}_{i_t}[||\nabla f_{i_t}(\mathbf{x}_t)||^2] \\
&\leq (1 - \mu\gamma)||\mathbf{x}_t - \mathbf{x}^\star||^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) + \gamma^2(4L(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 2\sigma_\star^2)
\end{aligned}$$

   where in the first inequality, we use the fact that f is μ-strongly convex and in the last inequality, we plug in the upper bound derived from above. The last step is to choose γ such that $-1 + 4L\gamma \leq 0$.

4. No bounded gradient or bounded variance assumption needed!

Practical Implementation of SGD

Follow the Python notebook provided here:

github.com/epfml/OptML_course/tree/master/labs/ex05/solution/