

Required reading: Wright and Recht, Ch. 5.

Throughout the problem set, $\|\cdot\|$ stands for the 2-norm $\|\cdot\|_2$.

1 (Exercise 4(b) from Wright and Recht, Ch. 5) Consider the general additive Gaussian noise model

$$g(x; \xi) = \nabla f(x) + \xi,$$

where ξ is an n -dimensional Gaussian random vector with mean 0 and covariance matrix $\sigma^2 I_n$. Suppose we estimate the gradient $\nabla f(x)$ using a minibatch of size $s \geq 1$:

$$g(x; \xi_1, \dots, \xi_s) := \nabla f(x) + \frac{1}{s} \sum_{j=1}^s \xi_j,$$

where ξ_1, \dots, ξ_s are i.i.d. Gaussian random vectors as above. Show that

$$\mathbf{E}_{\xi_1, \dots, \xi_s} (\|g(x; \xi_1, \dots, \xi_s)\|^2) = \|\nabla f(x)\|^2 + \frac{n\sigma^2}{s}.$$

2 (Exercise 6 from Wright and Recht, Ch. 5) Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be an m -strongly convex function that can be expressed as an expectation of the form $f(x) = \mathbf{E}_{\xi}[F(x; \xi)]$. Assume that there exist constants $L_g > 0$ and $B \geq 0$, such that

$$\mathbf{E}\|\nabla F(x; \xi)\|^2 \leq L_g^2 \|x - x^*\|^2 + B^2, \quad \text{for all } x \in \mathbb{R}^n$$

where x^* is the unique global minimizer of f . Suppose we run the stochastic gradient method on f by sampling ξ and taking steps along $\nabla F(x; \xi)$ using an epoch-doubling approach. That is, we run for T steps with steplength α , then for $2T$ steps with steplength $\alpha/2$, then for $4T$ steps with steplength $\alpha/4$, and so on. Let \hat{x}_t be the average of all the iterates in the t th epoch. How many epochs are required to guarantee that $\mathbf{E}\|\hat{x}_t - x^*\|^2 \leq \varepsilon$?

3 Consider the finite-sum objective

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad x \in \mathbb{R}^n$$

A simple mini-batching strategy is to sample $p \geq 1$ indices $i_1, \dots, i_p \in \{1, \dots, N\}$ uniformly at random with replacement and estimate the gradient $\nabla f(x)$ by

$$g(x; \xi) := \frac{1}{p} \sum_{k=1}^p \nabla f_{i_k}(x),$$

where $\xi := (i_1, \dots, i_p)$. Prove that, for each $x \in \mathbb{R}^n$,

$$\mathbf{E}_{\xi}[g(x; \xi)] = \nabla f(x)$$

and

$$\mathbf{E}_{\xi}\|g(x; \xi)\|^2 = \frac{1}{pN} \sum_{i=1}^N \|\nabla f_i(x)\|^2 + \frac{p-1}{p} \|\nabla f(x)\|^2.$$

4 Consider the finite-sum objective

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad x \in \mathbb{R}^n$$

satisfying the following:

- The functions f_i are all nonnegative and L -smooth, and there exists a point x^* such that $f_i(x^*) = 0$ for all i .
- There exists a constant $m > 0$, such that

$$\|\nabla f(x)\|^2 \geq 2mf(x), \quad \text{for all } x \in \mathbb{R}^n.$$

That is, f satisfies the Polyak–Łojasiewicz (PL) condition, see Section 3.8 of Wright and Recht.

Suppose that we run the following stochastic gradient method: Starting with an arbitrary initialization $x^0 \in \mathbb{R}^n$, at each iteration $k = 0, 1, 2, \dots$ we generate

$$x^{k+1} = x^k - \alpha g(x^k; \xi^k),$$

where $\alpha > 0$ is a constant steplength and where $g(x^k; \xi^k)$ is an estimate of $\nabla f(x^k)$ using a mini-batch of size p , as in Problem 3.

(a) Prove that, for each k ,

$$\begin{aligned} & \mathbf{E}[f(x^k) - f(x^{k+1}) | \xi^0, \dots, \xi^{k-1}] \\ & \geq \alpha \|\nabla f(x^k)\|^2 - \frac{\alpha^2 L}{2} \left(\frac{1}{pN} \sum_{i=1}^N \|\nabla f_i(x^k)\|^2 + \frac{p-1}{p} \|\nabla f(x^k)\|^2 \right). \end{aligned}$$

Hint: Use smoothness and the result of Problem 3.

(b) Assume that $\alpha \leq \frac{2}{L}$. Prove that, under our assumptions on f , it follows from the result in (a) that

$$\mathbf{E}[f(x^k) - f(x^{k+1}) | \xi^0, \dots, \xi^{k-1}] \geq \alpha \left(1 - \frac{\alpha L}{2} \frac{p-1}{p} \right) 2mf(x^k) - \frac{\alpha^2 L^2}{p} f(x^k).$$

Hint: Since each f_i is L -smooth, $\|\nabla f_i(x)\|^2 \leq 2Lf_i(x)$ for all $x \in \mathbb{R}^n$ (recall Problem 4(a) in Homework 1).

(c) Starting from the result in (b), show that

$$\mathbf{E}[f(x^{k+1})] \leq \left(1 - 2m\alpha + \frac{L}{p} (L + m(p-1)) \alpha^2 \right) \mathbf{E}[f(x^k)], \quad k = 0, 1, 2, \dots$$

(d) By optimizing over the choice of α in part (c), prove that our stochastic gradient method converges at an exponential rate:

$$\mathbf{E}[f(x^k)] \leq (1 - m\alpha^*)^k \mathbf{E}[f(x^0)],$$

where $\alpha^* := \frac{mp}{L(L+m(p-1))}$.