

# ECE 543: Statistical Learning Theory

Maxim Raginsky

## Homework 4

Assigned March 29, 2018; due April 5, 2018

---

**Required reading:** lecture notes of Hajek and Raginsky, Sections 11.1–11.5

---

### 1. Some facts about convex functions.

Throughout the problem,  $\mathcal{F}$  is a nonempty, closed, convex subset of a Hilbert space  $\mathcal{H}$ .

(a) Prove that if  $\varphi : \mathcal{F} \rightarrow \mathbb{R}$  is  $m$ -strongly convex (with  $m > 0$ ), then the minimizer of  $\varphi$  on  $\mathcal{F}$  is unique.

(b) Let  $\varphi : \mathcal{F} \rightarrow \mathbb{R}$  be a function which is  $m$ -strongly convex and  $M$ -smooth. Prove that  $M \geq m$ .

(c) Prove that if there exists a function  $\varphi : \mathcal{F} \rightarrow \mathbb{R}$  which is simultaneously  $m$ -strongly convex and  $L$ -Lipschitz with  $m, L > 0$ , then  $\mathcal{F}$  must be norm-bounded.

*Hint:* First prove that  $\|\nabla\varphi(f) - \nabla\varphi(f')\| \geq m\|f - f'\|$ , then use this together with the Lipschitz assumption to argue that  $\|\nabla\varphi(f_n)\| \rightarrow \infty$  as  $n \rightarrow \infty$  along any sequence  $\{f_n\}_{n=1}^\infty$  of points of  $\mathcal{F}$  escaping to infinity.

2. **Stochastic optimization.** The stochastic optimization problem can be stated as follows: Let  $\mathcal{F}$  be a closed, convex, norm-bounded subset of a Hilbert space  $\mathcal{H}$ , and let  $Z$  be an arbitrary space. Given a function  $\ell : \mathcal{F} \times Z \rightarrow \mathbb{R}$  and a probability distribution  $P$  on  $Z$ , the goal is to minimize the function

$$L(f) := \mathbf{E}[\ell(f, Z)]$$

over  $f \in \mathcal{F}$ . We assume that  $L$  is convex and differentiable on  $\mathcal{F}$ .

In general, a direct approach may be too much to ask for, especially if both  $\mathcal{F}$  and  $Z$  are high-dimensional. Instead, assume that we can freely obtain samples  $Z \sim P$ , and for each sample  $Z \sim P$  we can get an unbiased estimate  $G(f, Z)$  of  $\nabla L(f)$  — that is,  $G(f, Z)$  is a random object taking values in  $\mathcal{H}$ , such that  $\mathbf{E}[G(f, Z)] = \nabla L(f)$ . Then we can use the following iterative scheme to approximately minimize  $L(f)$ :

- Pick an initial point  $f_0 \in \mathcal{F}$ .
- At discrete time steps  $t = 1, 2, \dots$ , draw an independent sample  $Z_t \sim P$ , and let

$$f_t = \Pi(f_{t-1} - \alpha_t G(f_{t-1}, Z_t)).$$

Here,  $\{\alpha_t\}_{t=1}^{\infty}$  is a nonincreasing sequence of positive step sizes, and  $\Pi : \mathcal{H} \rightarrow \mathcal{F}$  denotes the projection operator

$$\Pi(h) := \operatorname{argmin}_{f \in \mathcal{F}} \|f - h\|.$$

In this problem, we will analyze the convergence properties of this Projected Gradient Descent (PGD) scheme.

(a) Fix an arbitrary point  $f^* \in \mathcal{F}$ , and define  $V_t := \frac{1}{2} \|f_t - f^*\|^2$ ,  $t = 0, 1, \dots$ . Prove that

$$V_t \leq V_{t-1} + \frac{\alpha_t^2}{2} \|G(f_{t-1}, Z_t)\|^2 - \alpha_t \langle G(f_{t-1}, Z_t), f_{t-1} - f^* \rangle$$

(b) From the result of part (a), deduce that

$$\mathbf{E}[V_t] \leq \mathbf{E}[V_{t-1}] + \frac{\alpha_t^2}{2} \mathbf{E} \|G(f_{t-1}, Z_t)\|^2 - \alpha_t \mathbf{E} [\langle \nabla L(f_{t-1}), f_{t-1} - f^* \rangle].$$

Now we will derive convergence estimates for our PGD scheme under various further assumptions on  $\ell$ .

(c) Suppose that there is a positive constant  $\bar{G} > 0$ , such that  $\mathbf{E} \|G(f, Z)\|^2 \leq \bar{G}^2$  for all  $f \in \mathcal{F}$ . Use the result of part (b) to prove that, with the constant step size  $\alpha$ , for any  $T \geq 1$  we have

$$\mathbf{E} \left[ \sum_{t=0}^{T-1} (L(f_t) - L(f^*)) \right] \leq \frac{D^2}{2\alpha} + \frac{T\alpha\bar{G}^2}{2},$$

where  $D := \sup_{f, f' \in \mathcal{F}} \|f - f'\|$  is the *diameter* of  $\mathcal{F}$ . Use this to show that, with an appropriate choice of  $\alpha$  as a function of  $T$ , the average of the iterates

$$\bar{f}^{(T)} := \frac{1}{T} \sum_{t=0}^{T-1} f_t$$

satisfies

$$\mathbf{E} \left[ L(\bar{f}^{(T)}) - \inf_{f \in \mathcal{F}} L(f) \right] \leq \frac{D\bar{G}}{\sqrt{T}}.$$

(d) Suppose, in addition, that  $L$  is  $m$ -strongly convex. Let  $f^* \in \mathcal{F}$  be the (unique) minimizer of  $L$  on  $\mathcal{F}$ . Use the result of part (b) to prove that

$$\mathbf{E}[V_t] \leq (1 - 2m\alpha_t) \mathbf{E}[V_{t-1}] + \frac{\bar{G}^2 \alpha_t^2}{2}.$$

Deduce from this, that with a diminishing stepsize  $\alpha_t = \theta/t$  with a suitably chosen  $\theta > 0$ , we have

$$\mathbf{E}[V_t] \leq \frac{K}{t},$$

where  $K$  is some constant that depends on  $D$ ,  $\bar{G}$ , and  $m$ . Moreover, prove that if  $L$  is also  $M$ -smooth, then, for any  $T \geq 1$ ,

$$\mathbf{E} \left[ L(f_T) - \inf_{f \in \mathcal{F}} L(f) \right] \leq \frac{KM}{2T}.$$

3. **Stochastic gradient descent and the AERM property.** We will now use the results of Problem 2 to analyze the Stochastic Gradient Descent (SGD) algorithm with random selection. Recall the rationale behind SGD: We have  $n$  i.i.d. samples  $Z_1, \dots, Z_n$  from some distribution on  $Z$ , and our goal is to approximately minimize the empirical loss

$$L_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, Z_i)$$

over  $\mathcal{F}$ . The SGD algorithm with random selection works as follows:

- At time  $t = 0$ , pick an arbitrary initial point  $f_0 \in \mathcal{F}$ .
- At time  $t = 1, 2, \dots$ , pick an index  $I_t$  uniformly at random from the set  $\{1, \dots, n\}$ , independently of all past realizations, and compute

$$f_t = \Pi(f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, Z_{I_t})),$$

where  $\{\alpha_t\}_{t=0}^\infty$  is a nonincreasing sequence of positive stepsizes and  $\nabla \ell(f, z)$  is the gradient of  $\ell(f, z)$  with respect to  $f$  (for a fixed  $z$ ).

- (a) Assume that, for each  $z \in Z$ , the loss function  $f \mapsto \ell(f, z)$  is convex and  $L$ -Lipschitz (where the Lipschitz constant  $L$  is the same for all  $z$ ). Prove that, for an arbitrary  $T \geq 1$  and with an appropriately tuned constant stepsize  $\alpha$ , the average of the SGD updates

$$\bar{f}^{(T)} = \frac{1}{T} \sum_{t=0}^{T-1} f_t$$

satisfies

$$\mathbf{E}L_n(\bar{f}^{(T)}) - \inf_{f \in \mathcal{F}} L_n(f) \leq \frac{DL}{\sqrt{T}},$$

where  $D$  is the diameter of  $\mathcal{F}$ , and the expectation is only with respect to the random selection of the indices  $I_1, I_2, \dots, I_T$ . Use this result to demonstrate that SGD can be used to develop an AERM algorithm under the above assumptions on  $\ell$  and  $\mathcal{F}$ .

- (b) Assume, in addition, that the function  $f \mapsto \ell(f, z)$  is  $m$ -strongly convex and  $M$ -smooth (again,  $m$  and  $M$  do not depend on  $z$ ). Prove that, for appropriately tuned diminishing stepsizes  $\alpha_t = \theta/t$ , the SGD updates satisfy

$$\mathbf{E}L_n(f_T) - \inf_{f \in \mathcal{F}} L_n(f) \leq \frac{KM}{T},$$

for some constant  $K$  that depends only on  $D, L, m$ , and the expectation is only with respect to  $I_1, I_2, \dots, I_T$ . Use this result to demonstrate that SGD yields an AERM algorithm under the above assumptions on  $\ell$  and  $\mathcal{F}$ .