ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 2

Assigned November 3, 2015; due November 12, 2015

---

1. **Some facts about convex functions.**

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

   (a) Prove that if $\varphi : \mathscr{F} \to \mathbb{R}$ is σ-strongly convex (with $\sigma > 0$), then the minimizer of φ on $\mathscr{F}$ is unique.

   (b) Let $\varphi : \mathscr{F} \to \mathbb{R}$ be a function which is σ-strongly convex and β-smooth. Prove that $\beta \geq \sigma$.

   Hint: First prove that, for a σ-strongly convex function,

   $$\langle \nabla\varphi(f) - \nabla\varphi(f'), f - f' \rangle \geq \sigma \|f - f'\|^2, \qquad \forall f, f' \in \mathscr{F}.$$

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

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

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

   $$L(f) \triangleq \mathbb{E}[\ell(f, Z)]$$

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

   In general, a direct approach may be too much to ask for, especially if both $\mathscr{F}$ 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 $\mathscr{H}$, such that $\mathbb{E}[G(f, Z)] = \nabla L(f)$. Then we can use the following iterative scheme to approximately minimize $L(f)$:

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

     $$f_t = P_{\mathscr{F}}\left(f_{t-1} - \alpha_t G(f_{t-1}, Z_t)\right).$$

Here, $\{\alpha_t\}_{t=1}^{\infty}$ is a nonincreasing sequence of positive step sizes, $\nabla \ell(f,z)$ denotes the gradient of $\ell(\cdot,z)$ at f, and $P_{\mathscr{F}} : \mathscr{H} \to \mathscr{H}$ denotes the projection operator

$$P_{\mathscr{F}}(h) \triangleq \operatorname*{arg\,min}_{f \in \mathscr{F}} \|f - h\|.$$

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

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

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

Hint: Use the fact that the projection operator $P_{\mathscr{F}}$ is *contractive*:

$$\|P_{\mathscr{F}}(h) - P_{\mathscr{F}}(h')\| \le \|h - h'\|, \qquad \forall h, h' \in \mathscr{H}.$$

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

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

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

(c) Suppose that there is a positive constant $M > 0$, such that $\mathbb{E}\|G(f,Z)\|^2 \le M^2$ for all $f \in \mathscr{F}$. Use the result of part (b) to prove that, with the constant step size α, for any $T \ge 1$ we have

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

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

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

satisfies

$$\mathbb{E}\left[L(\bar{f}^{(T)}) - \inf_{f \in \mathscr{F}} L(f)\right] \le \frac{DM}{\sqrt{T}}.$$

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

$$\mathbb{E}[V_t] \le (1 - 2\sigma\alpha_t)\,\mathbb{E}[V_{t-1}] + \frac{M^2\alpha_t^2}{2}.$$

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

$$\mathbb{E}[V_t] \le \frac{K}{t},$$

where K is some constant that depends on D, M, and σ. Moreover, prove that if L is also β-smooth, then, for any $T \ge 1$,

$$\mathbb{E}\left[L(f_T) - \inf_{f \in \mathscr{F}} L(f)\right] \le \frac{K\beta}{2T}.$$

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 $Z_1, \ldots, Z_n$ 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 $\mathscr{F}$. The SGD algorithm with random selection works as follows:

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

$$f_t = P_{\mathscr{F}} \left(f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, Z_{I_t}) \right),$$

  where $\{\alpha_t\}_{t=0}^{\infty}$ is a nonincreasing sequence of positive stepsizes.

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

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

satisfies

$$\mathbb{E} L_n(\bar{f}^{(T)}) - \inf_{f \in \mathscr{F}} L_n(f) \leq \frac{DM}{\sqrt{T}},$$

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

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

$$\mathbb{E} L_n(f_T) - \inf_{f \in \mathscr{F}} L_n(f) \leq \frac{K\beta}{T},$$

for some constant K that depends only on D, M, σ, and the expectation is only with respect to $I_1, I_2, \ldots, I_T$. Use this result to demonstrate that SGD yields an AERM algorithm under the above assumptions on ℓ and $\mathscr{F}$.