5. Stability, Smoothness, and Stochastic Gradient Descent

**Assigned reading:** Chapter 11, Sections 1-5 of Prof. Raginsky’s notes, available at http://maxim.ece.illinois.edu/teaching/SLT/, and [1, Chapter 4].

1. **[Existence and uniqueness of the minimum of a strongly convex function]**
   Suppose $\mathcal{F}$ is a closed convex subset\(^{1}\) of a Hilbert space $\mathcal{H}$ and $\varphi : \mathcal{F} \to \mathbb{R}$ is a $\sigma$-strongly convex function for some $\sigma > 0$. (See revised definition at https://courses.engr.illinois.edu/ece543/sp2017/supplement5.pdf). This problem establishes that $\varphi$ has a unique minimizer.

   (a) Show $\varphi$ is bounded below. That is, $m \triangleq \inf_{f \in \mathcal{F}} \varphi(f)$ is finite.

   (b) By the definition of $m$, there exists a sequence $(f_n)_{n \geq 1}$ such that $\varphi(f_n) \to m$. Show $(f_n)_{n \geq 1}$ is a Cauchy sequence.

   (c) Explain why the sequence $(f_n)_{n \geq 1}$ has a limit $f^* \in \mathcal{F}$, and show that $f^*$ minimizes $\varphi$ over $\mathcal{F}$.

   (d) Show that if $f^*$ and $\tilde{f}^*$ both minimize $\varphi$ over $\mathcal{F}$, then $f^* = \tilde{f}^*$.

2. **[Stability of minimizers of strongly convex functions under Lipschitz perturbations]**
   Suppose $\mathcal{F}$ is a closed convex subset of a Hilbert space $\mathcal{H}$, $\varphi : \mathcal{F} \to \mathbb{R}$ is a $\sigma$-strongly convex function for some $\sigma > 0$, and $B$ is an $L$-Lipschitz continuous function on $\mathcal{F}$. Let $f^* = \arg\min_{f \in \mathcal{F}} \varphi(f)$ and suppose $\tilde{f}^*$ is a minimizer of $\varphi + B$. The goal of this problem is to show $\|f^* - \tilde{f}^*\| \leq \frac{L}{\sigma}$. Hint: Suppose $f' \in \mathcal{F}$ such that $\|f' - f^*\| > \frac{L}{\sigma}$. For $0 < \lambda < 1$, let $f_{\lambda} = \lambda f^* + (1 - \lambda) f'$. You can show that $\varphi(f_{\lambda}) = \varphi(f') + B(f')$ for $\lambda$ sufficiently small. The following facts are relevant:

   \[
   \varphi(f_{\lambda}) \leq \lambda \varphi(f^*) + (1 - \lambda) \varphi(f') - \frac{\lambda(1 - \lambda)\sigma}{2} \|f' - f^*\|^2
   \]

   \[
   \varphi(f') \geq \varphi(f^*) + \frac{\sigma\|f' - f^*\|^2}{2}.
   \]

3. **[Stability and consistency]**
   As shown in Chapter 11 of the notes (and Section 3 of the supplementary notes), in considerable generality, an ERM (even an asymptotic ERM) algorithm that is stable on average, with respect to replace one sample perturbations of the data, is consistent. The converse is true as well. That is, an ERM algorithm that is consistent must be stable on average [2, Theorem 9 and Lemma 16]. An important assumption in this result is that the set of probability distributions $\mathcal{P}$ should include all probability distributions on $\mathcal{Z}$. To illustrate the later point, adopt the setup $(\mathcal{Z}, \mathcal{P}, \mathcal{F}, \ell)$ of the first six lines in Section 2 of https://courses.engr.illinois.edu/ece543/sp2017/supplement5.pdf, so $\mathcal{F}$ is a Hilbert space and $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ is such that $\ell(f, z) \geq 0$ for all $f, z$. Give an example such that the ERM algorithm is consistent, but it does not generalize because it does better on the training data than any fixed estimator. (Hint: You have a lot of flexibility in selecting $\ell$, and $\mathcal{P}$ should not be the set of all probability distributions on $\mathcal{Z}$.)

4. **[Some basic properties of smooth convex functions]**
   Throughout this problem suppose $\varphi : \mathcal{F} \to \mathbb{R}$ is a convex function defined on a Hilbert space $\mathcal{F}$. By definition, $\varphi$ is $\beta$-smooth if $\nabla \varphi$ is $\beta$-Lipschitz continuous. Such smoothness is very helpful in case a gradient descent algorithm is used to minimize the function.

   (a) For this part only consider the case $\mathcal{F}$ is some $d$-dimensional Euclidean space, and $\varphi$ is twice continuously differentiable. By convexity, the Hessian satisfies, $\nabla^2 \varphi(f) \succeq 0$ for all $f$, where $A \succeq B$ for symmetric matrices $A$ and $B$ means $A - B$ is positive semidefinite. Assuming $\varphi$ is convex, show that $\varphi$ is $\beta$-smooth if $\nabla^2 \varphi(f) \preceq \beta I$ for all $f$. (Hint: $\nabla \varphi(f') - \nabla \varphi(f) = \int_0^1 \frac{d}{dt} (\nabla \varphi(f_t)) \, dt$ where

---

\(^{1}\)It is also assumed that $\mathcal{F}$ is nonempty.
(Remark: The same argument can be used to show that $\varphi$ is $\sigma$-strongly convex if $\nabla^2 \varphi \succeq \sigma I$. These conditions are necessary as well. $\beta$-smoothness and $\sigma$-strong convexity are closely related dual notions recurring in the theory of convex functions.)

(b) For the remaining parts of this problem suppose $\varphi$ is a $\beta$-smooth convex function. Show that for any $f, f' \in \mathcal{F}$, $\varphi(f') - \varphi(f) \leq \langle \nabla \varphi(f), f' - f \rangle + \frac{\beta}{2} \|f' - f\|^2$. (Hint: One approach: $\varphi(f') - \varphi(f) = \int_0^1 \frac{d\varphi(f + (1-t)f')}{dt} dt$ and chain rule applies, where $f_t = (1-t)f + tf'$.)

(c) Show that if $\nabla \varphi(f) = 0$ (i.e. if $f$ is a minimizer of $\varphi$), then $\varphi(f') - \varphi(f) \geq \frac{1}{2\beta} \|\nabla \varphi(f')\|^2$, for any $f' \in \mathcal{F}$. This interesting inequality shows that if $\varphi(f')$ is only slightly larger than the minimum value $\varphi(f)$, then $\|\nabla \varphi(f')\|$ must be small, even if $\|f' - f\|$ is very large. (Hint: Use $\varphi(f') - \varphi(f) \geq \varphi(f') - \varphi(f) - \frac{1}{\beta} \nabla \varphi(f')$, and apply part (b) for a suitable choice of $f$ and $f'$ in (b)).

(d) Show that $\varphi(f') - \varphi(f) \geq \langle \nabla \varphi(f), f' - f \rangle + \frac{\beta}{2} \|\nabla \varphi(f') - \nabla \varphi(f)\|^2$ for any $f, f' \in \mathcal{F}$. (Hint: What does the inequality say if $\nabla \varphi(f) = 0$? What is the effect of changing $\varphi$ by subtracting a linear term of the from $\langle g, f \rangle$ for a fixed vector $g$?)

(e) Prove the so-called co-coercive property of the gradient of a $\beta$-smooth convex function: $\langle \nabla \varphi(f') - \nabla \varphi(f), f' - f \rangle \geq \frac{1}{\beta} \|\nabla \varphi(f') - \nabla \varphi(f)\|^2$. (Hint: Use part (d) twice)

(f) Show that if $G_{f, \alpha} \triangleq f - \alpha \nabla \varphi(f)$ as in the course notes, and if $\alpha < 2/\beta$, then $\|G_{f, \alpha}(f) - G_{f, \alpha}(f')\| \leq \|f' - f\|$ for all $f, f' \in \mathcal{F}$. (Hint: Expand left-hand side and apply previous part.)

5. **SGD recursion for error bound on expected excess loss**

To better understand the meaning of the recursion for the convergence of the SGD algorithm described in [1, Chapter 4], we analyze the asymptotic behavior of the continuous approximation to the recursion, namely, the following ordinary differential equation (ode):

$$\dot{x}_t = -\left(\frac{\mu c \beta}{t}\right) x_t + \left(\frac{\beta}{t}\right)^2$$

where $\mu, c,$ and $\beta$ are positive constants.

(a) This is a linear ode. Find the impulse response function (aka propagator) $h(t, s)$ defined to equal $x_t$ for the ode $\dot{x} = -\left(\frac{\mu c \beta}{t}\right) x_t$ for $t \geq s$ with the initial value $x_s = 1$.

(b) The variation of parameters formula for the solution to the original ode with initial condition $x_1$ at $t = 1$ is:

$$x_t = h(t, 1) x_1 + \int_1^t h(t, s) \left(\frac{\beta}{s}\right)^2 ds.$$  

Simplify this expression for $x$ and identify the asymptotic behavior of $x_t$ as $t \to \infty$ in case $\mu c \beta > 1$ and in case $0 < \mu c \beta < 1$.

6. **Exploring stochastic gradient descent**

The python programming problem for this problem set is explained within the .ipynb file. You can see a static version at http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ee543/ee543_PythonProblem5.ipynb?flush_cache=true and download the ipynb file from the static version or directly from https://courses.engr.illinois.edu/ee543//ee543_PythonProblem5.ipynb.

References
