ECE 543 STATISTICAL LEARNING THEORY **SPRING 2019**
PROBLEM SET 5 **Due beginning of class, Tuesday, April 9**

5. Stability, Smoothness, and Stochastic Gradient Descent

Assigned reading: Section 10.5 and Chapter 11 of the course notes.

1. **[On the analysis of stochastic gradient descent]**
Example 10.1 of the notes gives the following standard example of a function and a stochastic gradient: $\Gamma(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i)$ and $g(f, \xi_t) = \nabla \ell(f, z_{\xi_t})$ where the variables ξ_t are independent, and uniformly distributed over the set of indices, $[n]$. Assume $f \mapsto \ell(f, z)$ is convex and differentiable for any z, so Γ is also convex and differentiable. What additional assumption(s) on the function ℓ could be used so that Assumption 10.1 of the notes holds, and what are the corresponding choices of μ, B and B_G?

2. **[SGD recursion for error bound on expected excess loss]**
To better understand the meaning of the recursion for the convergence of the SGD algorithm with diminishing step size (last part of proof of Thm. 10.10 in the notes) 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 μ, c, and β 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 $\dot{x}_t = -\left(\frac{\mu c \beta}{t}\right) x_t$ for $t \geq s > 0$ 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 $0 < \mu c \beta < 1$.

3. **[Convex optimization using gradient only]**
Insight into the performance of iterative optimization algorithms can sometimes be provided by ordinary differential equations (odes) which can be derived by taking step size to zero while speeding up time. Suppose Γ is a differentiable convex function defined on a Hilbert space $\mathcal{H}$. Suppose $f^* \in \mathcal{H}$ and $\Gamma^* \in \mathbb{R}$ such that $\min_{f \in \mathcal{H}} \Gamma(f) = \Gamma^* = \Gamma(f^*)$.

   (a) The gradient ordinary differential equation is given by

$$\dot{f} = -\nabla \Gamma(f)$$

   with $f(0) = f_0$. Consider the energy function $\mathcal{E}_g(t) = t(\Gamma(f(t)) - \Gamma^*) + \|f(t) - f^*\|^2/2$. First, show that $\dot{\mathcal{E}}_g(t) \leq 0$. Therefore, $\mathcal{E}_g(t) \leq \mathcal{E}_g(0)$ for all $t \geq 0$. Secondly, use that to derive an upper bound on $\Gamma(f(t)) - \Gamma^*$ for $t > 0$.

   (b) Part (a) can be repeated for the ode version of Nesterov's accelerated gradient method. Such ode is given by:

$$\ddot{f} + \frac{3}{t}\dot{f} + \nabla \Gamma(f) = 0.$$

   with the initial condition $f(0) = f_0$ and $\dot{f}(0) = 0$, using the energy function $\mathcal{E}_g(t) = t^2(\Gamma(f(t)) - \Gamma^*) + \|f(t) + t\dot{f}(t)/2 - f^*\|^2/2$. (see Su, Boyd, and Candés, *JMLR*, 2016) yielding the bound

$\Gamma(f(t)) - \Gamma^* \leq \frac{2\|f(0)-f^*\|^2}{t^2}$. (You don't need to repeat that.) A curious thing about this bound and the bound you found in part (a) is that the bounds are still valid if Γ is replaced by 10Γ, whereas the righthand sides of the bounds don't depend on Γ. Is that a contradiction? Explain.

4. **[Bounds on dynamic regret for online adversarial learning]**
Consider the framework of online convex optimization of Section 11.1. In particular, $\mathcal{F}$ is a closed, convex subset of a Hilbert space with diameter at most D. Suppose $\ell(\cdot, z)$ is an L-Lipschitz continuous, convex function for each z.

   (a) Suppose the projected gradient algorithm is run with a fixed step size α. Find a regret bound analogous to the one of Theorem 11.1 of the notes that depends on α, and then minimize it with respect to α. (You might be surprised to get a bound smaller than the one of Theorem 11.1, but the decreasing step size in Theorem 11.1 has the advantage it doesn't depend on the time horizon.)

   (b) The original paper of Zinkevich (2003) considers not only regret compared to fixed strategies, but also dynamic regret. The *path length* of a dynamic strategy $(f_t^*)_{1\leq t\leq T}$ is $\sum_{t=1}^{T-1}\|f_{t+1}^* - f_t^*\|$. The dynamic regret for an algorithm producing (f_t) is $R_{T,W} = J_T((f_t)) - J_T((f_t^*))$, where (f_t^*) minimizes J_T subject to having path length less than or equal to W. Assuming a constant step size α is used, derive a bound on dynamic regret that involves D, T, α, L, and W that reduces to the bound you found in (a) when $W = 0$. (Hint: Equation (11.3) in the notes is true for any f^*, so consider it with f^* replaced by f_t^*. The term $\|f_{t+1} - f_t^*\|^2$ will appear, whereas the sum will telescope if instead the term were $\|f_{t+1} - f_{t+1}^*\|^2$. Show that $\left|\|f_{t+1} - f_{t+1}^*\|^2 - \|f_{t+1} - f_t^*\|^2\right| \leq C\|f_{t+1}^* - f_t^*\|$ for some constant C depending on D.)

5. **[Optimality of O(sqrt T) regret bound for on-line convex function minimization]**
In the notation of Section 10.1, suppose $\mathcal{F} = \mathsf{Z} = [-1, 1]$ and $\ell(f, z) = 1 + fz$.

   (a) What does the gradient descent algorithm reduce to for this example?

   (b) Express $\min_{f^*\in\mathcal{F}} J_T(f^*, z^T)$ in terms of $z^T = (z_1, \ldots, z_T)$. Here, $J_T(f^*, z^T) = \sum_{t=1}^{T}\ell(f^*, z_t)$.

   (c) Suppose an online algorithm $\widetilde{A}$ (i.e. an algorithm of the form $(\widetilde{f}_t = \widetilde{A}(\widetilde{f}_1, \ldots, \widetilde{f}_{t-1}, z_1, \ldots, z_{t-1}))$) minimizes $\max_{z^T\in\mathsf{Z}^T} J_T((f_t), z^T)$ over all online algorithms. Is the sequence $(\widetilde{f}_1, \ldots, \widetilde{f}_T)$ produced by $\widetilde{A}$ uniquely determined? (This part shows that there is a difference between minimizing maximum loss, and minimizing maximum regret against all fixed strategies.)

   (d) Suppose for this part that the sequence $Z^T = (Z_1, \ldots, Z_T)$ is a Rademacher sequence (i.e. the Z_t's are iid, each equally likely to be ± 1.). Show that

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[\min_{f^*\in\mathcal{F}} J_T(f^*, Z^T)\right] - T}{\sqrt{T}} = -c,$$

   and identify the constant $c > 0$. (Hint: Apply the central limit theorem.) In contrast, find $\mathbb{E}\left[J_T((f_t), Z^T)\right]$ for (f_t) produced by an arbitrary online algorithm. Finally, explain why, for any $\epsilon > 0$, $\sup_{z^T} R_T((f_t), z^T) \geq (1-\epsilon)c\sqrt{T}$ for all sufficiently large T and any online algorithm.

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/ece543/sp2019/ece543_PythonProblem5.ipynb?flush_cache=true and download the ipynb file from the static version or directly from https://courses.engr.illinois.edu/ece543/sp2019/ece543_PythonProblem5.ipynb.