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

5. Stochastic and adversarial analysis of stochastic gradient descent

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

(Problems and solutions)

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?

   Solution: Since $E_{\xi_t}[g(f, \xi_t)] = \nabla\Gamma(f)$, property (i) is true with $\mu = 1$. The second condition can be written as

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla\ell(f_t, z_i)\|^2 \le B + B_G \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(f_t, z_i) \right\|^2$$

   A natural possibility is to assume $\|\nabla\ell(f, z)\|^2 \le B$ for all f, z (i.e. $\nabla\ell(f, z)$ is L-Lipschitz continuous and $B = L^2$) and take $B_G = 0$. (Note: That assumption is pretty strong and would preclude ℓ being strongly convex in f unless the domain of possible f is bounded. The algorithm can be modified to work over a bounded domain by adding a projection onto the domain at each iteration.)

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 $t \ge s > 0$ with the initial value $x_s = 1$.

   Solution: The equation can be written as $\frac{d \ln x_t}{dt} = -\frac{\mu c\beta}{t}$ so that $\ln h(t, s) = -\int_s^t \frac{\mu c\beta}{u} du = \mu c\beta \ln \frac{s}{t}$, yielding $h(t, s) = \left(\frac{s}{t}\right)^{\mu c\beta}$ for $0 < s < t$.

   (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 $0 < \mu c\beta < 1$.

Solution:

$$
\begin{aligned}
x_t &= x_1 \left(\frac{1}{t}\right)^{\mu c\beta} + \int_1^t \left(\frac{s}{t}\right)^{\mu c\beta} \left(\frac{\beta}{s}\right)^2 ds \\
&= x_1 \left(\frac{1}{t}\right)^{\mu c\beta} + \frac{\beta^2}{t^{\mu c\beta}} \int_1^t s^{\mu c\beta - 2} ds \\
&= x_1 \left(\frac{1}{t}\right)^{\mu c\beta} + \frac{\beta^2}{\mu c\beta - 1} \left\{ \frac{1}{t} - \frac{1}{t^{\mu c\beta}} \right\} \\
&\sim \begin{cases} \frac{\beta^2}{\mu c\beta - 1} \frac{1}{t} & \mu c\beta > 1 \\ \frac{\beta^2}{1 - \mu c\beta} \frac{1}{t^{\mu c\beta}} & 0 < \mu c\beta < 1. \end{cases}
\end{aligned}
$$

The solution converges to zero as long as $\mu c\beta > 0$, but more quickly if $\mu c\beta > 1$. (Though it was't asked for, $x_t \sim \frac{\beta^2 \ln t}{t}$ in case $\mu c\beta = 1$.)

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 $\dot{\mathcal{E}}_g(t) \le 0$. Therefore, $\mathcal{E}_g(t) \le \mathcal{E}_g(0)$ for all $t \ge 0$. Secondly, use that to derive an upper bound on $\Gamma(f(t)) - \Gamma^*$ for $t > 0$.

Solution: Since $\nabla \|u\|^2/2 = u$, the chain rule yields

$$
\begin{aligned}
\dot{\mathcal{E}}_g(t) &= \Gamma(f(t)) - \Gamma^* + t\langle \nabla\Gamma(f(t)), \dot{f}(t)\rangle + \langle f(t) - f^*, \dot{f}(t)\rangle \\
&= \Gamma(f(t)) - \Gamma^* - t\|\nabla\Gamma(f(t))\|^2 - \langle f(t) - f^*, \nabla\Gamma(f(t))\rangle \\
&\le \Gamma(f(t)) - \Gamma^* - \langle f(t) - f^*, \nabla\Gamma(f(t))\rangle \\
&= \Gamma(f(t)) + \langle f^* - f(t), \nabla\Gamma(f(t))\rangle - \Gamma^* \le 0,
\end{aligned}
$$

where the last step follows from the convexity of Γ. Therefore, $\mathcal{E}_g(t) \le \mathcal{E}_g(0)$ or

$$t(\Gamma(f(t)) - \Gamma^*) + \|f(t) - f^*\|^2/2 \le \|f(0) - f^*\|^2/2,$$

which implies $\Gamma(f(t)) - \Gamma^* \le \|f(0) - f^*\|^2/2t$.

(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 $\mathcal{E}_a(t) = t^2(\Gamma(f(t)) - \Gamma^*) + \|f(t) + t\dot{f}(t)/2 - f^*\|^2/2$. (see Su, Boyd, and Candés, *JMLR*, 2016) yielding the bound $\Gamma(f(t)) - \Gamma^* \le \frac{2\|f(0) - f^*\|^2}{t^2}$. (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.

Solution: The point is that $f(t)$ depends on Γ. For example, if for the gradient ode in part (a), Γ were changed to $\widetilde{\Gamma} = 10\Gamma$, then f would change to $\widetilde{f}(t) = f(10t)$. In other words, the trajectory of f would be traced out 10 times faster. Then,

$$\widetilde{\Gamma}(\widetilde{f}(t)) = 10\Gamma(f(10t)) \le \frac{10\|f(0) - f^*\|^2}{10t} = \frac{\|f(0) - f^*\|^2}{t}.$$

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.)

Solution: Equation (11.3) with $m = 0$ simplifies to

$$2(\ell_t(f_t) - \ell_t(f^*)) \leq \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha} + \alpha L^2 \tag{1}$$

Summing each side of (1) from $t = 1$ to T and dropping the negative term yields

$$2(J_T((f_t)) - J_T(f^*)) \leq \frac{1}{\alpha}\|f_1 - f^*\|^2 + \alpha L^2 T \leq \frac{D^2}{\alpha} + \alpha L^2 T.$$

Selecting $\alpha = \sqrt{\frac{D^2}{L^2 T}}$ yields $J_T((f_t)) - J_T(f^*) \leq DL\sqrt{T}$.

(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.)

Solution: Let $\triangle_t = f_{t+1}^* - f_t^*$. Then

$$\begin{aligned}\|f_{t+1} - f_{t+1}^*\|^2 &= \|f_{t+1} - f_t^* - \triangle_t\|^2 \\ &= \|f_{t+1} - f_t^*\|^2 - 2\langle f_{t+1} - f_t^*, \triangle_t\rangle + \|\triangle_t\|^2\end{aligned}$$

By the Schwartz inequality and the definition of D, $|\langle f_{t+1} - f_t^*, \triangle_t\rangle| \leq \|f_{t+1} - f_t^*\|\|\triangle_t\| \leq D\|\triangle_t\|$ and $\|\triangle_t\|^2 \leq D\|\triangle_t\|$. Thus, $\left|\|f_{t+1} - f_{t+1}^*\|^2 - \|f_{t+1} - f_t^*\|^2\right| \leq 3D\|\triangle_t\|$. Or slightly tighter:

$$\begin{aligned}\|f_{t+1} - f_{t+1}^*\|^2 &= \|f_{t+1} - f_t^* - \triangle_t\|^2 \\ &= \|f_{t+1} - f_t^*\|^2 - 2\langle f_{t+1} - f_t^*, \triangle_t\rangle + \|\triangle_t\|^2 \\ &= \|f_{t+1} - f_t^*\|^2 - \langle f_{t+1} - f_t^*, \triangle_t\rangle - \langle f_{t+1} - f_t^* - \triangle_t, \triangle_t\rangle \\ &= \|f_{t+1} - f_t^*\|^2 - \langle f_{t+1} - f_t^*, \triangle_t\rangle - \langle f_{t+1} - f_{t+1}^*, \triangle_t\rangle\end{aligned}$$

Thus, $\left|\|f_{t+1} - f_{t+1}^*\|^2 - \|f_{t+1} - f_t^*\|^2\right| \leq 2D\|\triangle_t\|$. Using this bound and the triangle inequality in (1) with $f^* = f_t^*$ yields

$$2(\ell_t(f_t) - \ell_t(f_t^*)) \leq \frac{\|f_t - f_t^*\|^2 - \|f_{t+1} - f_{t+1}^*\|^2}{\alpha} + \frac{2D\|\triangle_t\|}{\alpha} + \alpha L^2 \tag{2}$$

Summing each side of (2) from $t = 1$ to T, dropping a negative term and dividing by 2 yields

$$R_{T,W} = J_T((f_t)) - J_T((f_t^*)) \leq \frac{1}{2\alpha}\|f_1 - f_1^*\|^2 + \frac{DW}{\alpha} + \frac{\alpha L^2 T}{2} \leq \frac{D^2 + 2DW}{2\alpha} + \frac{\alpha L^2 T}{2}.$$

The minimum over α (not asked for in problem) is $L\sqrt{(D^2 + 2DW)T}$.

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?
Solution: For some initial state $f_1 \in [-1, 1]$, $f_{t+1} = \Pi(f_t - \alpha_t z_t)$ where $\Pi(f^\flat) = \max\{-1, \min\{1, f^\flat\}\}$.

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

(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.)
Solution: Given an online algorithm, a possible value for (z_t) is the binary sequence such that $z_t = \mathsf{sgn}(\widetilde{f}_t)$ for all t, and, for such sequence, $J_T((f_t), z^T) = T + \sum_{t=1}^T |f_t| \ge T$. Therefore, the minimum, over all online algorithms, of $\max_{z^T \in \mathsf{Z}^T} J_T((f_t), z^T)$, is at least T. Furthermore, equality holds if and only if the algorithm always produces the sequence $\widetilde{f}_t \equiv 0$ for all t.

(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 independent, 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 $\epsilon > 0$, $\sup_{z^T} R_T((f_t), z^T) \ge (1 - \epsilon)c\sqrt{T}$ for all sufficiently large T and any online algorithm.
Solution: By part (b),

$$\mathbb{E}\left[\min_{f^* \in \mathcal{F}} J_T(f^*, Z^T)\right] = T - \mathbb{E}\left[\left|\sum_{t=1}^T Z_t\right|\right]$$

In contrast, since f_t is determined by the deterministic learning algorithm and $(Z_1, \ldots, Z_{t-1})$, f_t is independent of Z_t. Also, $\mathbb{E}[Z_t] = 0$, so

$$\mathbb{E}\left[J_T((f_t), Z^T)\right] = \sum_{t=1}^T 1 + \mathbb{E}[f_t Z_t] = T + \sum_{t=1}^T \mathbb{E}[f_t]\,\mathbb{E}[Z_t] = T.$$

Taking the difference of the respective sides of the last two centered equations yields:

$$\mathbb{E}\left[R_T((f_t), Z^T)\right] = \mathbb{E}\left[\left|\sum_{t=1}^T Z_t\right|\right].$$

Since there is at least one possible realization of z^T such that $R_T((f_t), z^T)$ is greater than or equal to the mean, it follows that

$$\sup_{z^T} R_T((f_t), z^T) \ge \mathbb{E}\left[\left|\sum_{t=1}^T Z_t\right|\right].$$

By the central limit theorem, $\frac{\sum_{t=1}^T Z_t}{\sqrt{T}}$ converges in distribution to W, where W is a standard normal random variable. Thus, $\left|\frac{\sum_{t=1}^T Z_t}{\sqrt{T}}\right|$ converges in distribution to $|W|$. Since $\mathbb{E}\left[\left(\frac{\sum_{t=1}^T Z_t}{\sqrt{T}}\right)^2\right]$ is bounded for all t (it is equal to one for all t) it follows that

$$\mathbb{E}\left[\left|\frac{\sum_{t=1}^T Z_t}{\sqrt{T}}\right|\right] \xrightarrow{t \to \infty} \mathbb{E}[|W|] = c = \sqrt{\frac{2}{\pi}}.$$

Thus, 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.