6. More on SGD, Online Convex Function Minimization


Problems and solutions

1. [Assumptions on the stochastic gradient]

Consider the assumptions imposed for the analysis of the SGD algorithm: (i) There exists $\mu > 0$ such that for all $t \geq 1$, $\langle \nabla \Gamma(f_t), E_{\xi_t}[g(f_t, \xi_t)] \rangle \geq \mu \|\nabla \Gamma(f_t)\|^2$, and (ii) there exist $M \geq 0$ and $M_G \geq 0$ such that $E_{\xi_t}[\|g(f_t, \xi_t)\|^2] \leq M + M_G \|\nabla \Gamma(f_t)\|^2$.

(a) Show that (i) and (ii) hold and if $\|\nabla \Gamma(f_t)\|$ can be arbitrarily large, then $M_G \geq \mu^2$.

**Solution:** By the Cauchy-Schwartz inequality and the first assumption,

$$\|\nabla \Gamma(f_t)\| E_{\xi_t}[g(f_t, \xi_t)] \geq \langle \nabla \Gamma(f_t), E_{\xi_t}[g(f_t, \xi_t)] \rangle \geq \mu \|\nabla \Gamma(f_t)\|^2.$$ 

so that $E_{\xi_t}[g(f_t, \xi_t)] \geq \mu \|\nabla \Gamma(f_t)\|$. For fixed $f_t$, the random vector $g(f_t, \xi_t) - E_{\xi_t}[g(f_t, \xi_t)]$ has mean zero, so that

$$E_{\xi_t}[\|g(f_t, \xi_t)\|^2] = E_{\xi_t}[\|g(f_t, \xi_t) - E_{\xi_t}[g(f_t, \xi_t)]\|^2] \geq \|E_{\xi_t}[g(f_t, \xi_t)]\|^2 \geq \mu^2 \|\nabla \Gamma(f_t)\|^2.$$ 

Hence, if both assumptions hold,

$$M + M_G \|\nabla \Gamma(f_t)\|^2 \geq \mu^2 \|\nabla \Gamma(f_t)\|^2$$

with probability one. Hence, if $M = 0$, or if $\|\nabla \Gamma(f_t)\|$ can be arbitrarily large, then $M_G \geq \mu^2$.

(b) For what choice(s) of $\mu$, $M_G$, and $M$ must $\mathbb{P}\{g(f_t, \xi_t) = \nabla \Gamma(f_t)\} = 1$?

**Solution:** Equality holds in the Schwartz inequality used in (a) if and only if $g(f_t, \xi_t) = \mu \nabla \Gamma(f_t)$ with probability one, where the randomness is with respect to $\xi_t$ with $f_t$ fixed. If $M = 0$ and $M_G = \mu^2$ then equality must hold in the Schwartz inequality, implying that $g(f_t, \xi_t) = \mu \nabla \Gamma(f_t)$ with probability one, and setting $\mu = 1$ yields the deterministic gradient descent with stepsize multipliers ($\alpha_t$).

2. [Contraction property of projection onto a closed, convex set]

Let $\mathcal{F}$ be a closed, convex subset of a Hilbert space $\mathcal{H}$ and let $\Pi : \mathcal{H} \to \mathcal{F}$ denote the projection mapping. That is, $\Pi(f)$ is the unique element in $\arg \min_{h \in \mathcal{F}} \|h - f\|^2$. (Existence and uniqueness follows by strong convexity of $h \mapsto \|h - f\|^2$.)

(a) Show that $\langle f - \Pi(f), h - \Pi(f) \rangle \leq 0$ for any $h \in \mathcal{F}$. (Hint: $\|f - (\Pi(f)(1 - \lambda) + \lambda h)\|^2$ is minimized with respect to $\lambda \in [0, 1]$ at $\lambda =$ 0.

**Solution:** By the hint, $\|f - \Pi(f)\|^2 + 2\lambda(f - \Pi(f), h - \Pi(f)) + \lambda^2\|h - \Pi(f)\|^2$ is minimized over $\lambda \in [0, 1]$ at $\lambda =$ 0 so the coefficient of $\lambda$ must be less than or equal to zero.

(b) Show that $\Pi$ is a contraction mapping: $\|\Pi(f) - \Pi(f')\| \leq \|f - f'\|$.

**Solution:** By part (a), $\langle f - \Pi(f), \Pi(f') - \Pi(f) \rangle \leq 0 \leq \langle f' - \Pi(f), \Pi(f') - \Pi(f) \rangle$ Thus,

$$\langle f' - f + \Pi(f) - \Pi(f'), \Pi(f') - \Pi(f) \rangle \geq 0$$

or $\|f' - f + \Pi(f) - \Pi(f')\| \geq \|\Pi(f') - \Pi(f)\|$. For the last step, apply the Cauchy-Schwarz inequality, $\|f' - f\| \times \|\Pi(f') - \Pi(f)\| \geq (f' - f, \Pi(f') - \Pi(f))$.

3. [Optimality of $O(\sqrt{T})$ regret bound for on-line convex function minimization]

In the notation of https://courses.engr.illinois.edu/ece543/sp2017/supplement6.pdf, suppose $\mathcal{F} = \mathcal{Z} = [-1, 1]$ and $\ell(f, z) = 1 + fz$. 


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

**Solution:** For some initial state \( f_1 \in [-1, 1] \), \( f_{t+1} = \Pi(f_t - \alpha_z t) \) where \( \Pi(f^p) = \max\{-1, \min\{1, f^p\}\} \).

(b) Express \( \min_{f \in X} J_T(f^*, z^T) \) 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:** \( \min_{f \in X} J_T(f^*, z^T) = \min_{-1 \leq f^* \leq 1} T + f^* \sum_{t=1}^T z_t = T - \| \sum_{t=1}^T z_t \| \).

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

**Solution:** Any online algorithm determines a sequence \( (z_t) \) such that \( z_t = \text{sgn}(\tilde{f}_t) \) for all \( t \). Then \( J_T((f_t), z^T) = T + \sum_{t=1}^T |f_t| \geq T \). Therefore, the minimum, over all online algorithms, of \( \max_{x^T \in X} J_T((f_t), z^T) \), is at least \( T \), with equality if and only if the algorithm produces the sequence \( f_t \equiv 0 \) 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 \( Z_i \)'s are iid, each equally likely to be \( \pm 1 \)). Show that

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

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.

**Solution:** By part (b),

\[
\mathbb{E} \left[ \min_{f^* \in X} J_T(f^*, z^T) \right] = T - \mathbb{E} \left[ \sum_{t=1}^T Z_t \right]
\]

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, \( \mathbb{E} \left[ Z_t \right] = 0 \), so

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

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

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

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

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

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

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

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.
4. [Continuation of previous problem]

Continue to consider the setting of problem 3, with \( \mathcal{F} = \mathcal{Z} = [-1, 1] \) and \( \ell(f, z) = 1 + fz \).

(a) Consider the projected GD algorithm run with step size \( \alpha_t = \frac{1}{\sqrt{T}} \) for \( t \geq 1 \) and initial state \( f_1 = 0 \). Suppose \( T \) is even and \( z_1 = \ldots = z_{T/2} = -1 \) and \( z_{T/2+1}, \ldots, z_T = 1 \). Show that there is a finite constant \( c > 0 \) such that \( R_T((f_t), z^T) \leq -T + c\sqrt{T} \) for all \( T \) sufficiently large. (Large negative maximum regret means the algorithm has much smaller loss than any fixed \( f^* \).)

**Solution:** For any \( f^* \), \( J_T(f^*) = 1 + f^* \sum_{t=1}^{T} z_t = T \). Thus, \( R_T((f_t), z^T) = J_T((f_t), z^T) - T \). It thus remains to show that for some constant \( c \), \( J_T((f_t), z^T) \leq c\sqrt{T} \) for all \( T \). The algorithm is

\[
f_{t+1} = \begin{cases} \Pi(f_t + \alpha_t) & 1 \leq t \leq \frac{T}{2} \\ \Pi(f_t - \alpha_t) & \frac{T}{2} + 1 \leq t \leq T - 1 \\ \end{cases}
\]

Since \( f_1 = 0 \) and \( \alpha_t = 1 \), \( f_t = 1 \) for \( 2 \leq t \leq \frac{T}{2} + 1 \). Thus, \( \ell_1(f_t) = 1 \) and \( \ell_t(f_t) = 0 \) for \( 2 \leq t \leq \frac{T}{2} \). At step \( \frac{T}{2} + 1 \) we still have \( f_{\frac{T}{2}+1} = 1 \) so that \( \ell_{\frac{T}{2}+1}(f_{\frac{T}{2}+1}) = 2 \). Subsequently, the iterates decrease until reaching -1, if \( T \) is large enough. The total distance from 1 to -1 is 2, and \( \alpha_t \geq \frac{1}{\sqrt{T}} \) for all \( t \in [T] \), so the number of steps after \( \frac{T}{2} + 1 \) til state -1 is reached is less than or equal to \( [2\sqrt{T}] \leq 2\sqrt{T} + 1 \). So \( f_1 = -1 \) and \( \ell(f_1, z_1) = 0 \) for \( \frac{T}{2} + 1 + 2\sqrt{T} \leq t \leq T \), which makes sense if \( \frac{T}{2} + 1 + 2\sqrt{T} \leq T \), or \( T \geq 20 \). To summarize, \( \ell_1(f_1) = 1, \ell_t(f_t) \leq 2 \) for \( \frac{T}{2} + 1 \leq t \leq \frac{T}{2} + 1 + [2\sqrt{T}] \), and \( \ell_t(f_t) = 0 \) for other \( t \in [T] \). So

\[
J_T((f_t), z^T) \leq 3 + 4\sqrt{T}
\]

for all \( T \geq 20 \). So the claim in the problem statement is true for any \( c < 4 \). (This argument is not tight.)

(b) A conclusion of problem 3(d) is that for any online algorithm, there is a constant \( c' > 0 \) so that \( \max_{z \in \mathbb{Z}^T} R_T((f_t), z^T) \geq c'\sqrt{T} \) for all \( T \) sufficiently large. The proof uses randomization and is thus nonconstructive. To gain more insight into this result, prove this same result for the projected gradient descent algorithm with fixed step size \( \alpha \) with \( \alpha \) by using deterministic sequences \( z^T \). For convenience you can assume \( T \) is even and consider only \( 0 < \alpha \leq 1 \). (Hint: The choice of \( z^T \) depends on \( \alpha \). The entire range \( \alpha > 0 \) can be covered by two choices of \( z^T \).)

**Solution:** Consider consider sequence \( z^T:A \) defined by \( z_t = (-1)^t \). Starting with \( f_1 = 0 \), the sequence of hypotheses is \( f_t = 0 \) for \( t \) odd and \( f_t = -\alpha \) for \( t \) even. Thus, \( \ell_t(f_t) = 1 + \alpha \) for \( t \) odd and \( \ell_t(f_t) = 0 \) for \( t \) even. So \( J_T((f_t), z^T:A) = T + \alpha T/2 \). For any fixed \( f^* \), \( J_T(f^*, z^T:A) = T \) because \( \sum z_t = 0 \). Thus, \( R_T((f_t), z^T:A) = \alpha T/2 \).

Consider next the sequence \( z^T:B \) defined by \( z_t = -1 \) for all \( t \). Then \( J_T(f^*, z^T:B) = 0 \) for \( f^* \) = 1, so \( R_T((f_t), z^T:B) = J_T((f_t), z^T:B) \). Then \( f_t = \min\{1, (t-1)\alpha\} \) for all \( t \). Thus, \( \ell_t(f_t, z_t) = 1 - \min\{1, (t-1)\alpha\} = (1 - (t-1)\alpha)_{+} \) for all \( t \geq 1 \). Thus,

\[
J_T((f_t), z^T:B) = \sum_{t=1}^{T} (1 - (t-1)\alpha)_{+} = \int_{0}^{T} (1 - (t-1)\alpha)_{+} dt = \begin{cases} T(1 - \frac{\alpha T}{2}) \geq \frac{T}{2} & \alpha T < 1 \\ \frac{1}{2\alpha} & \alpha T \geq 1 \end{cases}
\]

So if \( \alpha T \leq 1 \) then \( R_T((f_t), z^T:B) \geq \frac{T}{2} \). If \( \alpha T > 1 \) then \( \max\{}R_T((f_t), z^T:A), R_T((f_t), z^T:B)\right) \geq \max\{\frac{\alpha T}{2}, \frac{1}{2\alpha}\} \geq \sqrt{T}/2 \). Thus, in general, if \( T \geq 20 \), then for \( 0 \leq \alpha \leq 1 \),

\[
\max\{R_T((f_t), z^T:A), R_T((f_t), z^T:B)\right) \geq \sqrt{T}2.
\]

5. [Dynamic programming solution for minimizing maximum regret]

Continue to consider the setting of problem 3, with \( \mathcal{F} = \mathcal{Z} = [-1, 1] \) and \( \ell(f, z) = 1 + fz \). Write \( f_t \) to denote \( (f_1, \ldots, f_t) \), and \( f^T \) to denote \( (f_1, \ldots, f_T) \), and define \( z^T \) similarly. Let \( \mathcal{A} \) represent an arbitrary online algorithm of the learner, and \( \mathcal{B} \) represent an arbitrary online algorithm of the adversary. That is, \( \mathcal{A} \) determines \( f_1 \) and it has mappings of the form \( f_t = A(f_{t-1}, z_{t-1}) \) for \( 2 \leq t \leq T \), and \( \mathcal{B} \) has mappings of the form \( z_t = B(f_{t-1}, z_{t-1}) \) for \( 1 \leq t \leq T \). Together, a choice of \( \mathcal{A} \) and \( \mathcal{B} \) determine \( f^T \) and \( z^T \) uniquely. The goal of this problem is to determine the min max regret, \( R_T = \min_{\mathcal{A}} \max_{\mathcal{B}} R_T(f^T, z^T) \).
(a) Show by induction on \( k \) that for \( 0 \leq k \leq T \)

\[
R^*_T = \min_A \max_B \sum_{t=1}^{T-k} f_t z_t + V_k \left( \sum_{t=1}^{T-k} z_t \right),
\]

where \( V_0(s) = |s| \) for all \( s \in \mathbb{R} \), and

\[
V_{k+1}(s) = \min_{f \in [-1,1]} \left( \max_{z \in [-1,1]} fz + V_k(s+z) \right).
\]

In particular, \( R^*_T = V_T(0) \). (Hint: These are equations of dynamic programming, working backwards from the end of a problem by induction. First check the base case, \( k = 0 \). For the induction step, isolate the terms \( f_{T-k} \) and \( z_{T-k} \) on the righthand side of (1).)

**Solution:** The proof is by induction on \( k \). For the base case, we repeat from problem 3 that \( \min_{f \in \mathcal{F}} J_T(f^*, z^T) = \min_{-1 \leq f \leq 1} T + f^* \sum_{t=1}^{T} z_t = T - |\sum_{t=1}^{T} z_t| \). Subtracting from \( L_T((f_t),z^T) \) gives \( R_T((f_t), z^T) = \sum_{t=1}^{T} f_t z_t + |\sum_{t=1}^{T} z_t| \). Applying \( \min_A \max_B \) shows that (1) is true for \( k = 0 \).

For the induction step, suppose (1) is true for some \( k \) with \( 0 \leq k \leq T - 1 \). Following the hint, and using the notation \( s_t = \sum_{r=1}^{t} z_r \):

\[
\sum_{i=1}^{T-k} f_i z_i + V_k \left( \sum_{i=1}^{T-k} z_i \right) = \sum_{i=1}^{T-(k+1)} f_i z_i + \left( f_{T-k} z_{T-k} + V_k(s_{T-(k+1)} + z_{T-k}) \right)
\]

The idea is to apply \( \min_A \max_B \) to each side of (3) to get (1) with \( k \) replaced by \( k+1 \), completing the proof of the induction step. Notice that for the algorithm \( A \) fixed in (3), whatever choice it makes for \( f_{T-k} \), the algorithm \( B \) can use that choice to determine \( z_{T-k} \) to maximize the quantity in parentheses in (3). Just before that happens (i.e. going backwards in time a bit more), algorithm \( A \) can use knowledge of past choices \( f^{T-k-1} \), \( z^{T-k-1} \) to determine \( f_{T-k} \) by minimizing the maximum over \( z_{T-k} \) of the quantity in parentheses.

(b) Show \( V_k(s) = \mathbb{E}[|s + Z_1 + \ldots + Z_k|] \) for \( 0 \leq k \leq T \), where \( Z_1, \ldots, Z_T \) are independent Rademacher random variables. In particular, \( R^*_T = \mathbb{E}[|Z_1 + \ldots + Z_T|] \). How does this compare with the lower bound on maximum regret you found in problem 3? (Hint: Let \( V_k(s) = \mathbb{E}[|s + Z_1 + \ldots + Z_k|] \) and show by induction it indeed satisfies \( V_0(s) = |s| \) and (2) for \( k \geq 1 \).)

**Solution:** Following the hint, we let \( V_k(s) = \mathbb{E}[|s + Z_1 + \ldots + Z_k|] \). We show the dynamic programming equations are true by induction on \( k \). The base case \( k = 0 \) is true by definition. It remains to prove (2) for a given \( k \) with \( 0 \leq k \leq T - 1 \). Note that \( V_k \) is an average of convex functions so it is convex. Therefore,

\[
\max_{z \in [-1,1]} fz + V_k(s+z) = \max\{-f + V_k(s-1), f + V_k(s+1)\}
\]

Since \( V_k \) is an average of 1-Lipschitz functions, it is also 1-Lipschitz, so that \( |V_k(s-1), f - V_k(s+1)| \leq 2 \). Therefore, \( f \) can be selected in \([-1,1]\) to equalize the two terms in the max, yielding the minimum over \( f \). That yields,

\[
\min_{f \in [-1,1]} \max_{z \in [-1,1]} fz + V_k(s+z) = \frac{V_k(s-1) + V_k(s+1)}{2} = V_{k+1}(s),
\]

where the last equality follows easily from the definition of \( V_{k+1} \) by conditioning on \( Z_{k+1} \). Combining with part (a) then shows \( R^*_T = \mathbb{E}[|Z_1 + \ldots + Z_T|] \propto \sqrt{\frac{2T}{T}} \). This expression for \( R^*_T \) exactly matches to lower bound found in problem 3. Therefore, the Rademacher distribution for the adversary is a max min optimal randomized strategy for the adversary.

**Remark 1.** The analysis yields expressions for a min max optimal online algorithm for the player, and a max min optimal algorithm for the adversary. For \( T \) fixed, to give expressions for \( f_t \) and \( z_t \) we use the equations above for \( k \) such that \( t = T - k \), or equivalently, \( k = T - t \). To find \( f_t \)
we consider the value of \( f \) that equalizes the two terms in the max on the right-hand side of (4), resulting in the formula:

\[
f_t = \left( \mathbb{E} \left[ |z_1 + \cdots + z_{t-1} - 1 + Z_{t+1} + \cdots + Z_T| - \mathbb{E} \left[ |z_1 + \cdots + z_{t-1} + 1 + Z_{t+1} + \cdots + Z_T| \right] \right] \right) / 2.
\]

Similarly, to find \( z_t \), we examine the value of \( z \) achieving the maximum on the left-hand side of (4); it is given by \( z = \text{sgn}(2f + V_k(s + 1) - V_k(s - 1)) \). Therefore, a max min optimal strategy for the adversary is given by

\[
z_t = \text{sgn} \left( 2f + \mathbb{E} \left[ |z_1 + \cdots + z_{t-1} + 1 + Z_{t+1} + \cdots + Z_T| - \mathbb{E} \left[ |z_1 + \cdots + z_{t-1} - 1 + Z_{t+1} + \cdots + Z_T| \right] \right] \right)
\]

Note that \( z_t \) can be chosen to be either 1 or \(-1\) in case the argument of \( \text{sgn}(\cdot) \) is zero. Because of this indeterminism, the optimal strategy for the adversary is not unique. In fact, if the player uses the min max optimal strategy then the argument of \( \text{sgn}(\cdot) \) is zero for all \( t \), meaning that the play of the adversary is arbitrary. Since the player’s responses depend on the choices of the adversary, the trajectory of the game can vary considerably, but the regret for the player is always \( R_T^* \).