ECE 434: Problem Set 4

Due on November 29, 2023

Professor Chi Jin

Evan Dogariu
Collaborators: None

Problem 2

In this problem, we will show that we can transform any algorithms with convergence guarantees for strongly convex functions to new algorithms with provable convergence guarantees for convex functions.

Suppose we have an algorithm $\mathcal{A}$ (which is not necessarily gradient descent). The algorithm takes an initial point $x_1 \in X$, and an integer $T \in \mathbb{N}$ as input, and has the following guarantee: for any ℓ-smooth, α-strongly convex function f, after querying the gradient oracle T times, the output x_T satisfies:

$$f(x_T) - f(x^\star) \leq \ell D^2 \exp(-T/\kappa)$$

where $\kappa := \ell/\alpha$ is the condition number, and D is the diameter of domain X, i.e., $D := \max_{x,y \in X} \|x - y\|$.

Prove that, for any ℓ-smooth, convex function f, to find a point $\hat{x}$ such that $f(\hat{x}) - f(x^\star) \leq \varepsilon$, it suffices to query the gradient oracle $\tilde{O}(\ell D^2/\varepsilon)$ times, by smart uses of the algorithm $\mathcal{A}$. Here $\tilde{O}(\cdot)$ hides both constant and logarithmic factors.

Solution

Proof. Let $\epsilon > 0$. We will define

$$\tilde{f}(x) := f(x) + \frac{\epsilon}{4D^2}\|x - x_1\|^2$$

Note that this is $\frac{\epsilon}{2D^2}$-strongly convex since $f(x) = \tilde{f}(x) - \frac{\epsilon}{4D^2}\|x - x_1\|^2$ is convex. Furthermore, we see that for all x, y, we have

$$\begin{aligned}
\|\nabla \tilde{f}(x) - \nabla \tilde{f}(y)\| &= \left\| \nabla f(x) - \nabla f(y) + \frac{\epsilon}{D^2}\left(\|x - x_1\| - \|y - x_1\|\right) \right\| \\
&\leq \|\nabla f(x) - \nabla f(y)\| + \frac{\epsilon}{2D^2}\|x - y\| \\
&\leq \left(\ell + \frac{\epsilon}{2D^2}\right)\|x - y\|
\end{aligned}$$

So, $\tilde{f}$ is $\left(\ell + \frac{\epsilon}{2D^2}\right)$-smooth. Then we get that $\kappa = \frac{\ell + \frac{\epsilon}{2D^2}}{\frac{\epsilon}{2D^2}} = 1 + \frac{2D^2\ell}{\epsilon}$. If we let x^* be the optimizer of f and w be the optimizer of $\tilde{f}$, then running algorithm $\mathcal{A}$ for T steps gives

$$\tilde{f}(x_T) - \tilde{f}(w) \leq \left(\ell + \frac{\epsilon}{2D^2}\right) D^2 \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right)$$

We know that $\tilde{f}(w) \leq \tilde{f}(x^*)$ by selection of w, and also that $f(x_T) \leq \tilde{f}(x_T)$ since $f \leq \tilde{f}$ always. So,

$$f(x_T) - \tilde{f}(x^*) \leq \tilde{f}(x_T) - \tilde{f}(w) \leq \left(\ell + \frac{\epsilon}{2D^2}\right) D^2 \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right)$$

Plugging in our expression of $\tilde{f}$,

$$\begin{aligned}
f(x_T) - f(x^*) &\leq \left(\ell + \frac{\epsilon}{2D^2}\right) D^2 \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right) + \frac{\epsilon}{4D^2}\|x^* - x_1\|^2 \\
&\leq \left(\ell + \frac{\epsilon}{2D^2}\right) D^2 \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right) + \frac{\epsilon}{4} \\
&\leq \ell D^2 \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right) + \frac{\epsilon}{2} + \frac{\epsilon}{4},
\end{aligned}$$

where we used that $\|x^* - x_1\| \leq D$ and $\exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right) \leq 1$. So, if we set

$$T \geq \ln\left(\frac{4\ell D^2}{\epsilon}\right)\left(\frac{2\ell D^2}{\epsilon} + 1\right),$$

then we get that

$$\frac{\epsilon T}{\epsilon + 2D^2\ell} \geq \ln\left(\frac{4\ell D^2}{\epsilon}\right) \implies \exp\left(-\frac{\epsilon T}{\epsilon + 2D^2\ell}\right) \leq \frac{\epsilon}{4\ell D^2}$$

So, for such a selection of T, we have that

$$f(x_T) - f(x^*) \leq \ell D^2 \frac{\epsilon}{4\ell D^2} + \frac{3\epsilon}{4} = \epsilon$$

Since such a selection of T can be made to be $\widetilde{O}\left(\frac{\ell D^2}{\epsilon}\right)$ using O-tilde notation, we are done. ■

Problem 3

Consider SGD of update form

$$x_{t+1} = x_t - \eta g(x_t)$$

We assume that the stochastic gradient satisfies the following conditions:

(a) $\forall x,\ \mathbb{E}g(x) = \nabla f(x)$;

(b) $\forall x,\ \mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq \sigma^2$.

In this question, we consider the unconstrained problem, and aim to prove the following theorem. Consider a fixed time horizon T.

Theorem 1. *For any α-strongly convex and ℓ-smooth function f, SGD with learning rate $\eta = \min\left\{\frac{1}{\ell}, \frac{\iota}{\alpha T}\right\}$ and $\iota = \max\{1, 2\ln\frac{\alpha T\|x_1 - x^*\|}{\sigma}\}$ satisfies the following:*

$$\mathbb{E}f\left(\sum_{t=2}^{T+1}\lambda_t x_t\right) - f(x^*) \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{2\sigma^2\iota}{\alpha T}$$

where $\lambda_t := (1-\eta\alpha)^{T+1-t}/\sum_{s=2}^{T+1}(1-\eta\alpha)^{T+1-s}$ and x^* is the minimizer of f.

(a) Prove that for any $t \in [T]$, we have

$$\mathbb{E}\|x_{t+1} - x^*\|^2 \leq (1-\eta\alpha)\mathbb{E}\|x_t - x^*\|^2 - 2\eta\mathbb{E}[f(x_{t+1}) - f(x^*)] + 2\eta^2\sigma^2$$

(b) Prove the following inequality

$$\mathbb{E}\left[\sum_{t=2}^{T+1}\lambda_t(f(x_t) - f(x^*))\right] \leq \frac{e^{-\eta\alpha T}}{2\eta}\|x_1 - x^*\|^2 + \eta\sigma^2$$

(c) Use above results to prove Theorem 1.

Solution

Proof. We will use that $\mathbb{E}\|g(x)\|^2 = \|\nabla f(x)\|^2 + \sigma^2$ for all x, which can be found in the notes for Lecture 7. Throughout the problem below, we will at times condition on x_t, and so we must be careful about the difference between $\|\mathbb{E}\nabla f(x_t)\|$ and $\mathbb{E}\|\nabla f(x_t)\|$; luckily, by Jensen's inequality and convexity of the norm we see that $\|\mathbb{E}\nabla f(x_t)\| \leq \mathbb{E}\|\nabla f(x_t)\|$. So, when we apply the tower rule and remove the conditioning on x_t, we may bound both by $\mathbb{E}\|\nabla f(x_t)\|$ and avoid any confusion.

(a) By strong convexity, we know for all t that, conditioned on the value of x_t, it holds that

$$\begin{aligned} f(x_t) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \\ &= -\frac{1}{\eta}\langle -\eta\nabla f(x_t), x_t - x^* \rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \end{aligned}$$

By linearity of the inner product and the fact that $-\eta\nabla f(x_t) = \mathbb{E}[-\eta g(x_t)] = \mathbb{E}[x_{t+1} - x_t]$, we see that

$$\begin{aligned} f(x_t) - f(x^*) &\leq -\frac{1}{\eta}\mathbb{E}[\langle x_{t+1} - x_t, x_t - x^* \rangle] - \frac{\alpha}{2}\|x_t - x^*\|^2 \\ &= \frac{1}{2\eta}\left(\mathbb{E}\|x_{t+1} - x_t\|^2 + \|x_t - x^*\|^2 - \mathbb{E}\|x_{t+1} - x^*\|^2\right) - \frac{\alpha}{2}\|x_t - x^*\|^2 \end{aligned}$$

Note that $\|x_{t+1} - x_t\|^2 = \eta^2 \|g(x_t)\|^2 \implies \mathbb{E}\|x_{t+1} - x_t\|^2 = \eta^2(\|\nabla f(x_t)\|^2 + \sigma^2)$. So, applying an outer expectation over x_t as well,

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{\eta}{2}\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{\eta\sigma^2}{2} + \frac{1}{2\eta}\left((1-\eta\alpha)\mathbb{E}\|x_t - x^*\|^2 - \mathbb{E}\|x_{t+1} - x^*\|^2\right) \tag{1}$$

Next, we may take directly from the first part of the proof of Theorem 3 in Lecture 7 that by convexity, smoothness, and the SGD update rule,

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq -\frac{\eta}{2}\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{\eta\sigma^2}{2} \tag{2}$$

For notation, let $r := 1 - \eta\alpha$ and $\delta_t := \mathbb{E}\|x_t - x^*\|^2$. Then, adding inequalities (1) and (2) yields

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \eta\sigma^2 + \frac{1}{2\eta}(r\delta_t - \delta_{t+1})$$

Rearranging,

$$\delta_{t+1} \leq r\delta_t - 2\eta\mathbb{E}[f(x_{t+1} - f(x^*)] + 2\eta^2\sigma^2,$$

which is precisely the result of part (a).

(b) From part (a), we know that for all $t > 1$,

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \eta\sigma^2 + \frac{1}{2\eta}(r\delta_{t-1} - \delta_t)$$

Note that $r \geq 0$ for large enough T by selection of η; if $\iota = 1$ then $\eta \leq \frac{1}{\alpha T} \leq \frac{1}{\alpha} \implies 1 - \eta\alpha \geq 0$ and if ι takes the other value then $\eta\alpha \leq \frac{2\ln(\alpha T\sqrt{\delta_1})}{\sigma T}$, which becomes ≤ 1 for large enough T. Define $\gamma_t := r^{T+1-t}$ and $Z := \sum_{s=2}^{T+1} \gamma_s$ such that $\lambda_t = \frac{\gamma_t}{Z}$. Then, multiplying the above equation by γ_t,

$$\mathbb{E}[\gamma_t(f(x_t) - f(x^*))] \leq \eta\sigma^2\gamma_t + \frac{1}{2\eta}(\gamma_{t-1}\delta_{t-1} - \gamma_t\delta_t),$$

where we used that $r\gamma_t = \gamma_{t-1}$. Summing this inequality, we see that it telescopes into

$$\sum_{t=2}^{T+1} \mathbb{E}[\gamma_t(f(x_t) - f(x^*))] \leq \eta\sigma^2 Z + \frac{1}{2\eta}(\gamma_1\delta_1 - \gamma_{T+1}\delta_{T+1})$$

Since $\gamma_{T+1}\delta_{T+1} \geq 0$ and $\gamma_1 = r^T = (1-\eta\alpha)^T \leq e^{\eta\alpha T}$ since $1 - b \leq e^{-b}\ \forall b$, we see that

$$\sum_{t=2}^{T+1} \mathbb{E}[\gamma_t(f(x_t) - f(x^*))] \leq \eta\sigma^2 Z + \frac{e^{-\eta\alpha T}}{2\eta}\|x_1 - x^*\|^2$$

Dividing by Z,

$$\sum_{t=2}^{T+1} \mathbb{E}[\lambda_t(f(x_t) - f(x^*))] \leq \frac{e^{-\eta\alpha T}}{2\eta Z}\|x_1 - x^*\|^2 + \eta\sigma^2$$

If we can show that $Z \geq 1$ then the result follows. To see this, we note that Z increases as a function of r and that as $r \to 0$, we have $Z \to 1$ since $\sum_{s=2}^{T+1} 0^{T+1-t} = 0^0 = 1$. Therefore, (b) is proven.

(c) We first note that since $\sum_{t=2}^{T+1} \lambda_t = 1$ this is a convex combination, and so convexity of f guarantees that

$$f\left(\sum_{t=2}^{T+1} \lambda_t\right) \leq \sum_{t=2}^{T+1} \lambda_t f(x_t)$$

Problem 3 continued on next page...

Applying this to the result of part (b),

$$\mathbb{E}f\left(\sum_{t=2}^{T+1}\lambda_t\right) - f(x^*) \leq \frac{e^{-\eta\alpha T}}{2\eta}\|x_1 - x^*\|^2 + \eta\sigma^2$$

By our choice of η we know that $\eta\sigma^2 \leq \frac{\sigma^2\iota}{\alpha T}$. Also, we know that $\eta \leq \frac{1}{\ell}$, and so

$$\frac{e^{-\eta\alpha T}}{2\eta} \leq \frac{\ell e^{-\eta\alpha T}}{2} \implies \mathbb{E}f\left(\sum_{t=2}^{T+1}\lambda_t\right) - f(x^*) \leq \frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 + \frac{\sigma^2\iota}{\alpha T}$$

We will split the last analysis into cases; in each case we want to show that

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{\sigma^2\iota}{\alpha T}$$

as the result of part (c) will then follow. We proceed.

- Suppose that $\eta = \frac{1}{\ell}$. Then, $\eta\alpha T = \frac{T}{\kappa}$, and so

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 = \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{\sigma^2\iota}{\alpha T}$$

- Suppose now that $\eta = \frac{\iota}{\alpha T}$ and $\iota = 1$, and so $\eta\alpha T = 1$. Note that since $\frac{\iota}{\alpha T} \leq \frac{1}{\ell}$, we have

$$-\frac{T}{\kappa} = -\frac{\alpha T}{\ell} \geq -\frac{\alpha T\iota}{\alpha T} = -1 = -\eta\alpha T$$

  Thus,

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{\sigma^2\iota}{\alpha T}$$

  as desired.

- Suppose now that $\eta = \frac{\iota}{\alpha T}$ and $\iota = 2\ln\frac{\alpha T\|x_1 - x^*\|}{\sigma}$. Then, $\eta\alpha T = \iota$, and so

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 = \frac{\ell\|x_1 - x^*\|^2}{2}\frac{\sigma^2}{\alpha^2 T^2\|x_1 - x^*\|^2} = \frac{\ell\sigma^2}{2\alpha^2 T^2}$$

  Since $\eta = \frac{\iota}{\alpha T}$ we know that $\ell \leq \frac{\alpha T}{\iota}$, and so

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 \leq \frac{\sigma^2}{2aT\iota}$$

  Lastly, we know that $\iota \geq 1$ and so $\frac{1}{\iota} \leq \iota$, yielding that

$$\frac{\ell e^{-\eta\alpha T}}{2}\|x_1 - x^*\|^2 \leq \frac{\sigma^2\iota}{2aT} \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{\sigma^2\iota}{\alpha T}$$

  as desired.

So, in any case we see that

$$\mathbb{E}f\left(\sum_{t=2}^{T+1}\lambda_t\right) - f(x^*) \leq \frac{\ell e^{-T/\kappa}}{2}\|x_1 - x^*\|^2 + \frac{2\sigma^2\iota}{\alpha T}$$

and (c) is proven. ■

Problem 4

Consider binary classification problems with $0 - 1$ loss. In the lecture, we proved that for a finite class of classifiers $\mathcal{F}$, under the realizable assumption

$$\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}\{f(x_t) \neq y_t\} = 0,$$

the Halving algorithm (an improper learner) achieves a regret of $O(\ln |\mathcal{F}|)$, where O hides absolute constant factors.

Prove that, under the same setting, Hedge (a proper learner) with learning rate $\eta = 1/2$ achieves an expected regret that is upper bounded by $4 \ln |\mathcal{F}|$.

Solution

Proof. Let $K := \ln |\mathcal{F}|$. For each timestep t, let $p_t \in \Delta(K)$ be our probabilities for each classifier in $\mathcal{F}$ and $\ell_t \in \{0,1\}^K$ be the loss vector that assigns a loss value to each classifier. Then, we see that for the p_t's provided by the Hedge algorithm, the lemma from the USC notes grants that

$$\text{expected regret} = \sum_{t=1}^{n} \langle p_t, \ell_t \rangle - \sum_{t=1}^{n} \ell_t(i^*) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{n} \sum_{i=1}^{K} p_t(i) \ell_t(i)^2,$$

where we used the reasoning in the proof of Theorem 4 from the USC notes to observe that the LHS is indeed the expected regret. Since we are in the realizable setting, $\ell_t(i^*) = 0$ for all t. Since we are using the $0 - 1$ loss, $\ell_t(i)^2 = \ell_t(i)$, and so $\sum_{i=1}^{K} p_t(i) \ell_t(i)^2 = \langle p_t, \ell_t \rangle$. Combining these facts, we see that

$$\text{expected regret} = \sum_{t=1}^{n} \langle p_t, \ell_t \rangle \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{n} \langle p_t, \ell_t \rangle$$

Rearranging,

$$\text{expected regret} \leq \frac{\ln K}{\eta(1-\eta)}$$

Plugging in $\eta = \frac{1}{2}$, we see that

$$\text{expected regret} \leq 4 \ln K = 4 \ln |\mathcal{F}|$$

as desired. ■