Homework 8

Due April 10 at 11pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex and β-smooth function. In class, we studied gradient descent with constant step-size s, where s has to be chosen smaller than $1/\beta$. We here study an alternative method to find a step size s in the gradient descent algorithm called **backtracking line search**. Given an iterate x^t of the gradient descent, the next iterate is defined by

$$x^{t+1} = x^t - s\nabla f(x^t). \tag{1}$$

where s is chosen in the following iterative way. Fix two parameters $0 < \lambda < 1/2$ and $0 < \mu < 1$ and consider the following iterative scheme.

Step 1. Start with $s = 1$.

Step 2. While $f(x^t - s\nabla f(x^t)) > f(x^t) - \lambda s\|\nabla f(x^t)\|^2$, let $s := \mu s$ and reiterate.

(a) Assume that f is β-smooth. Show that if $s < 1/\beta$, then the backtracking line search stops. What is then the maximal number of iterations of the search?

(b) Show that if $\mu \leq \beta$, then the search stops at a value $s \geq \mu/\beta$. Conclude that

$$f(x^{t+1}) \leq f(x^t) - \lambda \min(1, \mu/\beta)\|\nabla f(x^t)\|^2.$$

(c) Use the previous question and argue as in the proof of Proposition 3.2 to conclude that after T steps of gradient descent with backtracking line search, we have

$$f(x^T) - f(x^\star) \leq (1 - 2\alpha\lambda \min(1, \mu/\beta))^T (f(x^0) - f(x^\star)).$$

Compare this rate of convergence to the rate obtained in Proposition 3.2. Is it better? Is it worse? What is an advantage of backtracking line search compared to the method proposed in Proposition 3.2 (constant step size)?

2. Consider $f : \mathbb{R}^d \to \mathbb{R}$ be a function in $\mathbb{R}^d$ that is twice differentiable in some point x^0 and has a unique minimizer $x^\star$.

   (a) Argue that Newton's method will converge quadratically in the following setting: assume that the restriction of f on $B(x^\star, \|x^0 - x^\star\|)$ (the ball centered at $x^\star$ with x^0 on its boundary) is α-strongly convex, β-smooth and has a Hessian that is γ-Lipschitz continuous, and assume also that $\|x^0 - x^\star\| \leq 2\alpha/\gamma$. Hint: in Proposition 4.2, we proved this result when f satisfies these conditions on $\mathbb{R}^d$, but do we really need those to hold on all of $\mathbb{R}^d$? You **do not** have to rewrite the proof of Proposition 4.2, only explain why the proof still holds with the weaker assumptions.

   (b) Consider $f : x \in \mathbb{R} \mapsto \log(e^x + e^{-x})$. What is the minimizer $x^\star$ of f? Find the minimal α such that α is α-strongly convex. Find the maximal β such that f is β-smooth.

(c) Consider the initialization $x^0 = 1$. According to question 1., should Newton's method converge with this initialization? Plot f and f' through the 5 first iterations of the method. Hint: the second derivative f'' is $4/(3\sqrt{3})$-Lipschitz continuous (you **do not** have to prove this).

(d) Same question with initialization $x^0 = 1.1$.

3. In this exercise, we give some properties of the AdaBoost method. The AdaBoost method consists in aggregating some "weak" classifiers to create a stronger one. Let $\mathcal{X}$ be a set of inputs and $\mathcal{Y} = \{-1, +1\}$ the set of outputs. We consider the exponential loss $\ell_{\exp}(y, y') = \exp(-yy')$. Let $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ be a set of n i.i.d. observations of distribution P. Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a set of T (possibly not very good) classifiers with values in $\{-1, +1\}$. The AdaBoost aims at finding a good classifier in the set

$$\mathrm{Span}(\mathcal{F}) = \{F = \sum_{j=1}^{m} \alpha_j f_j : \ \alpha_j \in \mathbb{R}\}.$$

To do so, we perform a greedy minimization of the associated empirical risk

$$\mathcal{R}_n(F) = \frac{1}{n} \sum_{i=1}^{n} \exp(-\mathbf{y_i} F(\mathbf{x_i})).$$

More precisely, we compute a sequence of classifiers $\hat{F}_t$ for $t = 1, \ldots, T$ with initialization $\hat{F}_0 = 0$. Given $\hat{F}_{t-1}$, we consider

$$\min_{\substack{j=1,\ldots,p \\ \alpha \in \mathbb{R}}} \mathcal{R}_n(\hat{F}_{t-1} + \alpha f_j),$$

If α_t and j_t attain this minimum, we let $\hat{F}_t = \hat{F}_{t-1} + \alpha_t f_{j_t}$. The final classifier is defined as $\hat{G}(x) = \mathrm{sgn}(\hat{F}_T(x))$.

(a) Let $w_i^{(t)} = n^{-1} \exp(-\mathbf{y_i} \hat{F}_{t-1}(\mathbf{x_i}))$ and define for $j = 1, \ldots, m$, the weighted empirical error of f_j at time t as

$$\varepsilon_t(j) = \frac{\sum_{i=1}^{n} w_i^{(t)} \mathbf{1}\{f_j(\mathbf{x_i}) \neq \mathbf{y_i}\}}{\sum_{i=1}^{n} w_i^{(t)}}.$$

Prove that

$$\mathcal{R}_n(\hat{F}_{t-1} + \alpha f_j) = e^{-\alpha} \sum_{i=1}^{n} w_i^{(t)} (1 + (e^{2\alpha} - 1)\varepsilon_t(j)).$$

Assume that we always have $\varepsilon_t(j) < 1/2$ for every j. Show that at a fixed j, the function $\alpha \mapsto \mathcal{R}_n(\hat{F}_{t-1} + \alpha f_j)$ is minimized at $\alpha = \frac{1}{2}\log\left(\frac{1-\varepsilon_t(j)}{\varepsilon_t(j)}\right)$ (Hint: compute the derivative). Conclude that the function $(\alpha, j) \mapsto \mathcal{R}_n(\hat{F}_{t-1} + \alpha f_j)$ is minimized at (α_t, j_t), where j_t is the index j minimizing $\varepsilon_t(j)$. Let $E_t := \min_{j=1,\dots,m} \varepsilon_t(j)$. Show that

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-E_t}{E_t}\right).$$

(b) Show that $\mathcal{R}_n(\hat{F}_T) = \prod_{t=1}^{T} \sqrt{4E_t(1-E_t)}$.

If the set of "weak" classifiers is reasonable, we can expect that the minimal empirical error $E(t)$ at time t is always smaller than $1/2 - \gamma$ for some $\gamma > 0$ (that is we do strictly better than "guessing at random" that would yield an empirical error of $1/2$ on average).

(c) Show that under this condition, $\mathcal{R}_n(\hat{F}_T) \leq \exp(-2T\gamma^2)$. To put it another way, the empirical risk of the AdaBoost predictor converges exponentially fast to 0.

The following result is true: the VC-dimension of the set of classifiers $\mathcal{G} = \{\mathrm{sgn} \circ F : F \in \mathrm{Span}(\mathcal{F})\}$ is smaller than $cT\log(m)$ for some absolute constant c. **You do not have to prove this result.**

(d) Let $\tilde{\mathcal{R}}_P(G)$ be the P-risk of a classifier G **for the** $0-1$ **loss**. Show that under the previous conditions

$$\mathbb{E}[\tilde{\mathcal{R}}_P(\hat{G})] \leq \exp(-2T\gamma^2) + 2\sqrt{\frac{2cT\log(m)}{n}\log\left(\frac{en}{cT\log(m)}\right)}.$$

(f) Let $\mathcal{X} = \mathbb{R}^d$. A class of "weak" classifiers that is used in practice is given by the class $\mathcal{F}_{\text{stump}}$ of so-called "stumps" classifiers whose boundary is given by a hyperplane of the form $\{x \in \mathbb{R}^d : x_k = c\}$

for some index $k \in \{1, \dots, d\}$ and constant $c \in \mathbb{R}$. Use the `AdaBoostClassifier` function from `sklearn` to implement AdaBoost on the `spambase.csv` dataset on the class $\mathcal{F}_{\text{stump}}$ of classifiers (this is the default parameter in the `AdaBoostClassifier` function). Plot the empirical risk for the $0-1$ loss on the training sample and on the testing sample as a function of the number of iterations T. What do you observe for $T = 1$? For T large? What is surprising?