Homework 12

Due May 7 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.

This homework is a two-part problem on ridge regression. In the first part, we study ridge regression in the framework of risk minimization with an appropriate loss. The goal of the second part is to assess the performance of stochastic gradient descent on this problem.

1. We consider the regression setting

$$\mathbf{y} = \langle \theta_0, \mathbf{x} \rangle + \varepsilon$$

where $\theta_0 \in \mathbb{R}^d$ with $\|\theta_0\| \leq R$, $\mathbf{x} \in \mathbb{R}^d$ and ε is a noise term independent from $\mathbf{x}$. We assume that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Id}_d$ (we then say that $\mathbf{x}$ is isotropic). We also assume that the noise is centered $\mathbb{E}[\varepsilon] = 0$ and that $\mathbb{E}[\varepsilon^2] \leq \sigma^2$. We call P the joint distribution of $\mathbf{x}$ and $\mathbf{y}$.

(a) Let $\lambda > 0$. We define the loss function $\ell_\lambda(y, y') = (y - y')^2 + \lambda y'^2$. The P-risk of a predictor $f : \mathbb{R}^d \to \mathbb{R}$ is by definition equal to

$$\mathcal{R}_P(f) = \mathbb{E}_P[\ell_\lambda(\mathbf{y}, f(\mathbf{x}))] = \mathbb{E}_P[(\mathbf{y} - f(\mathbf{x}))^2] + \lambda \mathbb{E}[f(\mathbf{x})^2].$$

Show that

$$\mathcal{R}_P(f) = \mathbb{E}_P[(\langle \theta_0, \mathbf{x} \rangle - f(\mathbf{x}))^2] + \lambda \mathbb{E}[f(\mathbf{x})^2] + \sigma^2.$$

Show that the Bayes predictor is given by $f_P^\star(x) = \langle x, \frac{\theta_0}{1+\lambda} \rangle$. (Hint: minimize the quantity $z \mapsto (\langle \theta_0, \mathbf{x} \rangle - z)^2 + \lambda z^2$.)

(b) Show that the Bayes risk is equal to

$$\mathcal{R}_P^\star = \mathcal{R}_P(f_P^\star) = \frac{\lambda \|\theta_0\|^2}{1 + \lambda} + \sigma^2.$$

(c) For $\theta \in \mathbb{R}^d$, we let f_θ be the linear predictor $x \mapsto \langle \theta, x \rangle$. Show that

$$\mathcal{R}_P(f_\theta) = \|\theta - \theta_0\|^2 + \lambda \|\theta\|^2 + \sigma^2.$$

2. Let $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ be a sample of n i.i.d. observations from distribution P.

(a) Show that the function $\theta \mapsto \mathcal{R}_P(f_\theta)$ is α-strongly convex for $\alpha = 2(\lambda + 1)$.

(b) Show that $\mathbf{v_i} = 2\mathbf{x_i}(\langle \mathbf{x_i}, \theta \rangle - \mathbf{y_i}) + 2\lambda\theta$ is an unbiased estimate of $\nabla \mathcal{R}_P(f_\theta)$ (that is prove that $\mathbb{E}[\mathbf{v_i}] = \nabla \mathcal{R}_P(f_\theta)$).

(c) Assume that $\mathbf{x}$ is bounded: there exists $M > 0$ such that $|\mathbf{x}| \leq M$ almost surely. Show that for every θ with $\|\theta\| \leq R$, it holds that

$$\mathbb{E}[\|\mathbf{v_i}\|^2] \leq C_1 \lambda^2 R^2 + C_2 R^2 + C_3 \sigma^2$$

for some constants C_1, C_2, C_3 that may depend on M. Show that one can apply Theorem 5 in the lecture notes in this setting. How small does the excess of risk get with stochastic gradient descent (with projection step on $B(0; R)$) using the n samples? (You may only give the order of convergence with respect to n, not the exact constant appearing in the bound.) What is the time complexity of this method (depending on d and n)?

The last question is completely optional and will not get you any additional points. It consists in comparing SGD with the "traditional" way of computing ridge regression.

3. **(Optional)** We let $\mathbf{Y} = (\mathbf{y_1}, \dots, \mathbf{y_n}) \in \mathbb{R}^n$ and $\mathbf{X}$ be the $n \times d$ matrix with rows given by $\mathbf{x_1}, \dots, \mathbf{x_n}$. We recall that the ridge regression estimator is given by

$$\hat{\theta}_{\mathrm{RR}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{n} + \lambda \mathrm{Id}_d \right)^{-1} \frac{\mathbf{X}^\top \mathbf{Y}}{n}.$$

   (a) Show that the excess of risk of f_θ is equal to

$$\mathcal{R}_P(f_\theta) - \mathcal{R}_P^\star = (1+\lambda)\|\theta - \frac{\theta_0}{1+\lambda}\|^2.$$

   (b) Using the law of large number, show that $\hat{\theta}$ converges to $\theta_0/(1+\lambda)$. (Hint: What is $\mathbb{E}[\mathbf{x_i}\mathbf{y_i}]$? What is $\mathbb{E}[\mathbf{x_i}\mathbf{x_i}^\top]$?)

   (c) Show that the excess of risk of $f_{\hat{\theta}_{\mathrm{RR}}}$ is of order at most $1/n$ (up to constants depending on σ, R and M).

   (d) What is the time complexity required to compute $\hat{\theta}_{\mathrm{RR}}$? If $d \gg 1$, is it faster to compute $\hat{\theta}_{\mathrm{RR}}$ or to apply SGD for n steps?