Algorithmic Foundations of Learning

Problem Sheet 3

Lecturer: Patrick Rebeschini *Due date: 48 hours before class*

3.1 U-statistics (Question type: A)

Let $h : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function, i.e., $h(x, y) = h(y, x)$, and assume that $\|h\|_\infty = \sup_{x,y\in\mathbb{R}^2} |h(x, y)| \leq c$. Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables, and define

$$U := \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j).$$

Show that $\mathbf{P}(|U - \mathbf{E}U| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{8c^2}}$.

Remark. U-statistics refers to a family of unbiased (hence the "U" term) estimators of interest. For instance, taking $h(x, y) = \frac{1}{2}(x - y)^2$ it can be showed that $U = \frac{1}{n-1}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{j=1}^n X_j)^2$, which is an unbiased estimatore for the variance, namely, $\mathbf{E}U = \mathbf{Var}X_1$.

3.2 Lipschitz Concentration for Gaussian Random Variables (Question type: B)

Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a vector of i.i.d. standard Gaussian random variables (mean 0 and variance 1), and let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, γ-Lipschitz with respect to the Euclidean norm. Prove that $f(X)$ is sub-Gaussian with variance proxy $\pi^2\gamma^2/4$, hence

$$\mathbf{P}(|f(X) - \mathbf{E}f(X)| \geq \varepsilon) \leq 2e^{-\frac{2\varepsilon^2}{\pi^2\gamma^2}}.$$

Hint: Use that if a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and $X, Y \in \mathbb{R}^d$ are two independent vectors of i.i.d. standard Gaussian random variables, then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbf{E}\phi(f(X) - \mathbf{E}f(X)) \leq \mathbf{E}\phi\left(\frac{\pi}{2}\nabla f(X)^\top Y\right).$$

Remark. This result can be improved, and it can be shown that even if f is not differentiable, $f(X)$ is sub-Gaussian with variance proxy γ^2, hence leading to

$$\mathbf{P}(|f(X) - \mathbf{E}f(X)| \geq \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2\gamma^2}}.$$

This is remarkable, as it shows that any γ-Lipschitz function of a standard Gaussian vector exhibits the same concentration as a one dimensional Gaussian random variable with variance γ^2.

3.3 Sub-exponential Random Variables and χ^2 Concentration (Question type: B)

A random variable X is said to be *sub-exponential* with non-negative parameters (ν^2, c) if

$$\boxed{\mathbf{E}e^{\lambda(X-\mathbf{E}X)} \leq e^{\nu^2\lambda^2/2} \qquad \text{for any } \lambda \in (-1/c, 1/c)}$$

1. Show that if a random variable X satisfies the two-sided Bernstein's condition with parameter $b > 0$ (see property (7.1) in Section 7.3 in the Lecture Notes), then it belongs to the class of sub-exponential random variables with parameters (ν^2, c), where $c = 2b$ and ν^2 is a parameter you should state.

2. Let Z be a standard Gaussian random variable, i.e., Gaussian with $\mathbf{E}Z = 0$ and $\mathbf{Var}Z = 1$. Then, Z^2 is a chi-squared random variable with 1 degree of freedom. Show that Z^2 is sub-exponential with parameters $\nu^2 = 4$ and $c = 4$.

3. Let $Z_1, \ldots, Z_n$ be independent standard Gaussian random variables. Then, $Y = Z_1^2 + \ldots + Z_n^2$ is a chi-squared random variable with n degrees of freedom. Show that

$$\mathbf{P}(|Y/n - 1| \geq \varepsilon) \leq 2e^{-n\varepsilon^2/8} \qquad \text{for all } \varepsilon \in (0, 1).$$

4. Assume that we are given n d-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^d$, and we are in a situation when d is so large that even storing this dataset into memory is very expensive. We would like to design a compression map $f : \mathbb{R}^d \to \mathbb{R}^p$ with $p \ll d$ that preserves some important characteristics of the data, so that we can store the compressed data $f(x_1), \ldots, f(x_n)$ and not "lose much" by doing so. As many algorithms in machine learning are based on computing pairwise distances, we are interested in finding a map f that preserves Euclidean distances, i.e., a map f such that for any $i, j \in [n]$ we have

$$(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2$$

for a tolerance parameter $\varepsilon \in (0, 1)$. Let $\mathbf{Z} \in \mathbb{R}^{p\times d}$ be a matrix formed by i.i.d. standard Gaussian random variables, and consider the (random) mapping $x \in \mathbb{R}^d \to F(x) := \mathbf{Z}x/\sqrt{p}$. Show that if $p > \frac{16}{\varepsilon^2}\log(n/\delta)$, then the mapping F satisfies the property above with high probability, namely:

$$\mathbf{P}\left(\frac{\|F(x_i) - F(x_j)\|_2^2}{\|x_i - x_j\|_2^2} \in (1 - \varepsilon, 1 + \varepsilon) \text{ for all } i \neq j\right) \geq 1 - \delta.$$

Hint: What is the distribution of $\frac{\|F(x)\|_2^2}{\|x\|_2^2}$?

Remark. This is surprising! The map F achieves an arbitrarily good compression with a projection dimension p that is independent of the original dimension d and that scales only logarithmically with the number of data points n.

3.4 Stochastic Mirror Descent (Question type: B)

Prove Theorem 11.2 in the Lecture Notes on the convergence of projected stochastic mirror descent.

Hint: Consider the proof of Theorem 10.11 and the proof of Theorem 11.1. Recall from Lecture 0 the properties of conditional expectations. Also, recall that Hölder's inequality for vectors reads $|x^\top y| \leq \|x\|\|y\|_*$ and that Hölder's inequality for expectations reads $\mathbf{E}|XY| \leq (\mathbf{E}[X^p])^{1/p}(\mathbf{E}[Y^q])^{1/q}$, for $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

3.5 Boosting (Question type: C)

You are competing in a Kaggle competition involving binary classification. Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^m \times \{-1, 1\}$ be the training data you are given. Assume that you have already computed d classifiers $h_k : \mathbb{R}^m \to \{-1, 1\}$, for $k \in [d]$, and that you want to combine them to design the final classifier for the competition. In particular, you want to compute the convex combination of weights $W^\star = (W_1^\star, \dots, W_d^\star) \in \Delta_d$ that solves the empirical risk minimization problem:

$$\min_{w=(w_1,\dots,w_d)\in\Delta_d} R(w) := \frac{1}{n}\sum_{i=1}^n \varphi\bigg(\bigg(\sum_{k=1}^d w_k h_k(X_i)\bigg) Y_i\bigg), \tag{3.1}$$

where $\varphi : [-1, 1] \to \mathbb{R}_+$ is a loss function to be chosen between the exponential loss, the hinge loss, or the logistic loss. Recall the following definitions:

- $\mathcal{B} = \{x \in \mathbb{R}^m \to a(x) \in \{-1, 1\}\}$
  $\mathcal{A}_{\text{soft}} = \{x \in \mathbb{R}^m \to a(x) = \sum_{k=1}^d w_k h_k(x) \in [-1, 1] : w = (w_1, \dots, w_d) \in \Delta_d\}$
- $r(a) = \mathbf{P}(a(X) \neq Y)$ for any $a \in \mathcal{B}$
  $r_\varphi(a) = \mathbf{E}\varphi(a(X)Y)$ for any $a \in \mathcal{A}_{\text{soft}}$
- $a^{\star\star} = \operatorname{argmin}_{a\in\mathcal{B}} r(a)$ (Bayes classifier)
  $a_\varphi^\star = \operatorname{argmin}_{a\in\mathcal{A}_{\text{soft}}} r_\varphi(a)$

Answer the following questions.

1. Let $A_\varphi^\star = \sum_{k=1}^d W_k^\star h_k \in \mathcal{A}_{\text{soft}}$ be the soft classifier obtained with the weights that solve problem (3.1). Show that with probability at least $1 - \delta$ the excess risk of the corresponding hard classifier $\operatorname{sign}(A_\varphi^\star)$ is upper-bounded as follows

   $$r(\operatorname{sign}(A_\varphi^\star)) - r(a^{\star\star}) \leq 2c\bigg(4\gamma\sqrt{\frac{2\log d}{n}} + \tilde{c}\sqrt{2\frac{\log(1/\delta)}{n}}\bigg)^\nu + 2c(r_\varphi(a_\varphi^\star) - r_\varphi(a_\varphi^{\star\star}))^\nu.$$

   Define the constants c, $\tilde{c}$, ν, and γ when the loss function φ is the exponential loss, the hinge loss, and the logistic loss, respectively. Hint: For $\nu \in [0, 1]$ and $a, b \geq 0$ we have $(a + b)^\nu \leq a^\nu + b^\nu$.

2. Based on the bound above, does the statistical performance of the final classifier increase or decrease with d? Why?

3. Give the full implementation (pseudocode) of a computationally-efficient algorithm to approximately solve problem (3.1) when the loss function φ is the exponential loss, the hinge loss, and the logistic loss, respectively. If $\overline{W}_t \in \mathbb{R}^d$ denotes the output of this algorithm at time step t, give a bound for the quantity

   $$\mathbf{E}[R(\overline{W}_t) - R(W^\star)].$$

   How long would you run the algorithm for? Why? What is the total computational complexity as a function of n and d?

Remark. Most algorithms that end up winning Kaggle competitions are indeed obtained by using ensemble meta-algorithms such as boosting, as a way to aggregate a variety of different methods and "boost up" their performances!

3.6 Algorithmic Stability: Strongly Convex Functions (Question type: B)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function that is α-strongly convex and β-smooth (with respect to the ℓ_2 norm), for given $\alpha < \beta$.

1. Prove that the function $x \in \mathbb{R}^d \to g(x) := f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex and c-smooth, for a constant $c > 0$ that you should state.

   Hint: What happens if the function f is twice-differentiable?

2. Let $\eta \leq 2/(\beta + \alpha)$. Show that, for any $x, y \in \mathbb{R}$, we have

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\|_2 \leq \Big(1 - \frac{\eta\beta\alpha}{\beta + \alpha}\Big)\|x - y\|_2.$$

   Hint: If a function h is convex and β-smooth with respect to the ℓ_2 norm, then its gradients are co-coercive, that is, $\langle \nabla h(x) - \nabla h(y), x - y \rangle \geq \frac{1}{\beta}\|\nabla h(x) - \nabla h(y)\|_2^2$.

3. Given the training data $Z_1, \dots, Z_n$, a loss function ℓ, and a convex set $\mathcal{C}$, consider the following empirical risk minimization problem

$$\begin{aligned} &\underset{w}{\text{minimize}} \quad R(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i) \\ &\text{subject to} \quad w \in \mathcal{C} \end{aligned}$$

   Assume that for any z the function $w \in \mathcal{C} \to \ell(w, z)$ is α-strongly convex, β-smooth, and γ-Lipschitz (with respect to the Euclidean norm $\|\cdot\|_2$). Consider the projected stochastic gradient descent algorithm (multiple passes through the data) with initial condition $W_1 = 0$ and learning rate $\eta_s \equiv \eta$ satisfying $\eta \leq 2/(\beta + \alpha)$. Use algorithmic stability to prove that for any $t \geq 1$ we have

$$\mathbf{E}[r(W_t) - R(W_t)] \leq \frac{2\eta\gamma^2}{n}\frac{\alpha + \beta}{\eta\alpha\beta}.$$