Introduction to Statistical Learning

Exercise sheet n°4

Exercise 1 - Rademacher complexity and multiclass classification

A. We recall that given $X_1, \ldots, X_n$ random vectors on $\mathbb{R}^d$ and $\mathcal{F}$ being a class of bounded real-valued functions, the empirical Rademacher average is defined as the random quantity :

$$\widehat{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \middle| X_1, \ldots, X_n \right)$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are IID random sign variables such that $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = +1\} = 1/2$. We also admit the following result : let $\Psi \; : \; \mathbb{R} \to \mathbb{R}$ be a Lipschitz function (*i.e.* $\exists L \; : \; \forall u, v \in \mathbb{R}, \; |\Psi(u) - \Psi(v)| \leq L|u - v|$), then we have :

$$\widehat{R}_n(\Psi \circ \mathcal{F}) \leq L\widehat{R}_n(\mathcal{F})$$

(a) What is the Lipschitz constant L for the function $u \mapsto |u|$?

(b) Consider two classes of bounded real-valued functions $\mathcal{F}_1$, $\mathcal{F}_2$. Find a simple upper bound of the following quantity :

$$\frac{1}{n}\mathbb{E}\left(\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2} \sum_{i=1}^{n} \varepsilon_i |f_1(X_i) - f_2(X_i)| \middle| X_1, \ldots, X_n \right)$$

depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

(c) Express $\max\{f_1, f_2\}$ as a linear relation involving $|f_1 - f_2|$.

(d) Consider the class $\mathcal{F} = \{\max\{f_1, f_2\} \; : \; f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and provide a simple upper bound of $\widehat{R}_n(\mathcal{F})$ depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

B. Consider a multiclass classification problem with observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ IID copies of the random pair (X, Y) where the output variable Y takes values in $\{1, \ldots, K\}$. The decision rules are functions g_h of the form :

$$g_h \; : \; x \mapsto \operatorname*{arg\,max}_{y\in\{1,\ldots,K\}} h(x, y)$$

where h is a real-valued function in a class $\mathcal{H}$ of functions over the set $\mathbb{R}^d \times \{1, \ldots, K\}$. The complexity of learning in the multiclass classification setup relies on the complexity of the class $\mathcal{H}$ that will be considered here under the margin approach. We thus define the margin ρ_h of function h as :

$$(x, y) \mapsto \rho_h(x, y) = h(x, y) - \max_{y'\neq y} h(x, y') \; ,$$

and ρ_h belongs to the class $\mathcal{H}_\rho$ of functions induced by $\mathcal{H}$.

(a) Set the empirical Rademacher complexity of the class $\mathcal{H}_\rho$ to be :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq \frac{1}{n}\mathbb{E}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i \rho_h(X_i, Y_i) \middle| (X_1, Y_1), \ldots, (X_n, Y_n) \right) .$$

Note that, for any Λ, we have that :

$$\Lambda(X_i, Y_i) = \sum_{y=1}^K \Lambda(X_i, y)\mathbb{I}\{y = Y_i\} = \sum_{y=1}^K \Lambda(y) \left(\frac{2\mathbb{I}\{y = Y_i\} - 1}{2} + \frac{1}{2} \right) ,$$

and show that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq \frac{1}{n}\sum_{y=1}^K \mathbb{E}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i \rho_h(X_i, y) \middle| X_1, \ldots, X_n \right)$$

(b) Set $\mathcal{H}_X = \{x \mapsto h(x, y) \ : \ y \in \{1, \ldots, K\}, \ h \in \mathcal{H}\}$. Using the definition of ρ_h and the main result of Part A, prove that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq K^\alpha \widehat{R}_n(\mathcal{H}_X)$$

where α will be made explicit.

(c) Set $\varphi_\gamma(u) = (1 - u/\gamma)\mathbb{I}\{u \in (0, \gamma]\} + \mathbb{I}\{u \geq 0\}$ and compute its Lipschitz constant.

(d) Relate the multiclass classification error $L(g_h) = \mathbb{P}\{Y \neq g_h(X)\}$ to the multi-class margin error $L_\gamma(h) = \mathbb{E}\{\varphi_\gamma\big(\rho_h(x, y)\big)\}$.

(e) We introduce $\widehat{L}_\gamma(h) = \frac{1}{n}\sum_{i=1}^n \varphi_\gamma(\rho_h(X_i, Y_i))$. Use a concentration inequality to derive an upper bound on the quantity :

$$\sup_{h\in\mathcal{H}} (L_\gamma(h) - \widehat{L}_\gamma(h)) .$$

(f) Give a sketch of proof that the following inequality holds, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$:

$$L(g_h) \leq \widehat{L}_\gamma(h) + c_1(K, \gamma)\mathbb{E}(\widehat{R}_n(\mathcal{H}_X)) + c_2(n, \delta)$$

where c_1 and c_2 will have to be computed explicitly.

Exercise 2 - Mirror descent algorithm for Online Convex Risk Minimization

A. We consider E a metric space with norm $\|\cdot\|$ and $\mathcal{D}$ an open and convex set in E. We first introduce some definitions :

- **[Strong convexity]** Fix $\alpha > 0$. A convex function $V : \mathcal{D} \to \mathbb{R}$ is said to be α-strongly convex with respect to norm $\|\cdot\|$ if :

$$V(sx + (1-s)y) \leq sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x-y\|^2$$

  for all $x, y \in \mathcal{D}$ and any $s \in [0,1]$.

  We assume in the sequel of the exercise that $V : \mathcal{D} \to \mathbb{R}$ is differentiable and α-strongly convex with respect to norm $\|\cdot\|$.

- **[Bregman divergence]** The Bregman divergence of V is defined as :

$$\mathcal{B}_V(y,x) = V(y) - V(x) - \nabla V(x)^T(y-x)$$

- **[Bregman projection]** For any $x \in \mathcal{D}$ and any closed convex set $\mathcal{C}$ in $\overline{\mathcal{D}}$, we define the Bregman projection as :

$$\Pi_{\mathcal{C},V}(x) = \underset{z \in \mathcal{C} \cap \mathcal{D}}{\arg\min}\, \mathcal{B}_V(z,x)$$

(a) Prove that, for any $x, y, z \in \mathcal{D}$, we have :

$$\big(\nabla V(x) - \nabla V(y)\big)^T(x-z) = \mathcal{B}_V(x,y) + \mathcal{B}_V(z,x) - \mathcal{B}_V(z,y)$$

(b) Take $z \in \mathcal{C} \cap \mathcal{D}$ and prove that, for any $y \in \mathcal{D}$, we have :

$$\big(\nabla V(\Pi_{\mathcal{C},V}(y)) - \nabla V(y)\big)^T(\Pi_{\mathcal{C},V}(y) - z) \leq 0$$

(c) Show that, for any $z \in \mathcal{C} \cap \mathcal{D}$ and any $y \in \mathcal{D}$, we have :

$$\mathcal{B}_V(z, \Pi_{\mathcal{C},V}(y)) \leq \mathcal{B}_V(z,y)$$

B. We consider the problem of the minimization of a convex function f which is assumed to be Lipschitz wrt $\|\cdot\|$ with Lipschitz constant equal to L. We denote by $\|\cdot\|_*$ the dual norm of $\|\cdot\|$ in the sense of convex conjugates. We introduce the following algorithm, known as the mirror descent algorithm, for given sets $\mathcal{C}$, $\mathcal{D}$, and potential function V :

Algorithm 1 Mirror descent algorithm

Require: $\eta > 0$, $x_1 \in \mathcal{C} \cap \mathcal{D}$ and $\zeta : (E, \|\cdot\|) \to (E, \|\cdot\|_*)$ with $\zeta(x) = \nabla V(x)$.
for $t = 1, \ldots, T$ **do**
  $\zeta(y_{t+1}) = \zeta(x_t) - \eta g_t$ with $g_t \in \partial f(x_t)$
  $x_{t+1} = \Pi_{\mathcal{C},V}(y_{t+1})$
end for
return either $\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ or $x^\circ \in \underset{x \in \{x_1, \ldots, x_T\}}{\arg\min} f(x)$

(a) Show that :

$$\mathcal{B}_V(x_t, y_{t+1}) \leq \frac{\eta^2 L^2}{2\alpha}$$

We will make use of Hölder's inequality : $g^T x \leq \|g\|_* \cdot \|x\|$.

(b) Show that : for any $x \in \mathcal{C} \cap \mathcal{D}$, we have :

$$\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - f(x)\big) \leq \frac{\eta L^2}{2\alpha} + \frac{\mathcal{B}_V(x, x_1)}{\eta T}$$

(c) For $x_1 = \underset{z \in \mathcal{C} \cap \mathcal{D}}{\arg\min}\, V(x)$ and any $x \in \mathcal{C} \cap \mathcal{D}$, show that :

$$\mathcal{B}_V(x, x_1) \leq \sup_{\mathcal{C} \cap \mathcal{D}} V - \inf_{\mathcal{C} \cap \mathcal{D}} V = R^2$$

(d) Find an upper bound for the rate of convergence of the mirror descent algorithm (for both estimates $\overline{x}_T$ and x°) to the minimum x^* of f, expressed in terms of R, L, α, and T.

C. Describe explicitly the Mirror Descent Algorithm in the following cases :

(a) [**Euclidean case**] $\mathcal{D} = \mathbb{R}^d$ and $V(x) = \frac{1}{2}\|x\|^2$.

(b) [ℓ_1 **case**] $\mathcal{D} = \mathbb{R}_+^d - 0$, $\mathcal{C} = \{x \in \mathbb{R}_+^d \ : \ \|x\|_1 = 1\}$ (simplex), and $V(x) = \sum_{i=1}^{d} x^{(i)} \log\big(x^{(i)}\big)$

(c) Apply the latter result to a finite convex combination of weak classifiers to minimize the convex risk in classification.