

Lecture 19 — April 1

Lecturer: Martin Wainwright

Scribe: Peter Dimitrov

Note: These lecture notes are still rough, and have only have been mildly proofread.

Outline

- Empirical Rademacher complexity and Lipschitz classes
- Empirical Rademacher complexity and Kernels

Recap

Given class \mathcal{F} and samples $\{X^{(i)}\}$, the *empirical Rademacher complexity* $\hat{R}_n(\mathcal{F})$ is defined as:

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} f(X^{(i)}) \right| \right]$$

$\hat{R}_n(\mathcal{F})$ comes in the proof of ULLN (Glivenko-Cantelli). Last time observed that:

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2}{n} \log [s(\mathcal{F}, \{X^{(i)}\})]}, \quad (19.1)$$

where $s(\mathcal{F}, \{X^{(i)}\})$ is the empirical shatter coefficient for set $\{X^{(i)}\}$ and function class \mathcal{F} . $s(\mathcal{F}, \{X^{(i)}\})$ is a random quantity which is always \leq worst-case shatter coefficient $\max_{\{X^{(i)}\}} s(\mathcal{F}, \{X^{(i)}\})$. In general, bound (19.1) is sharper than the worst-case one, which is defined by the VC dimension of the function class \mathcal{F} .

Useful result for finite classes is the following lemma.

Lemma 19.1. *For a finite set $\mathbb{A} \in \mathbb{R}^n$ with $\frac{1}{n} \sum_i a_i^2 \leq B, \forall a \in \mathbb{A}$.*

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} a_i \right| \right] \leq B \sqrt{\frac{2}{n} \log [\mathbb{A}]}.$$

Proof. left as an exercise. Hint: consider the worst-case scenario for a vector (a_i) with bounded norm, and use a bound for sub-Gaussian RVs from previous homework. \square

19.1 Behavior with respect to Lipschitz functions

Normally, we work with loss function \mathcal{L} , which in turn induces a loss function class: $\mathcal{F} \rightarrow \mathcal{L}(\mathcal{F})$.

Lemma 19.2. *(Ledoux-Talagrand contraction inequality) Let $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz functions with parameter $L, \forall i = 1, \dots, n$, i.e. $|\phi_i(a) - \phi_i(b)| \leq L|a - b|, \forall a, b \in \mathbb{R}$. Then,*

$$\hat{R}_n(\phi \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \phi_i(f(X^{(i)})) \right| \right] \leq L \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} f(X^{(i)}) \right| \right] = L \hat{R}_n(\mathcal{F}),$$

where $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$ is the loss function class $\mathcal{L}(\mathcal{F})$.

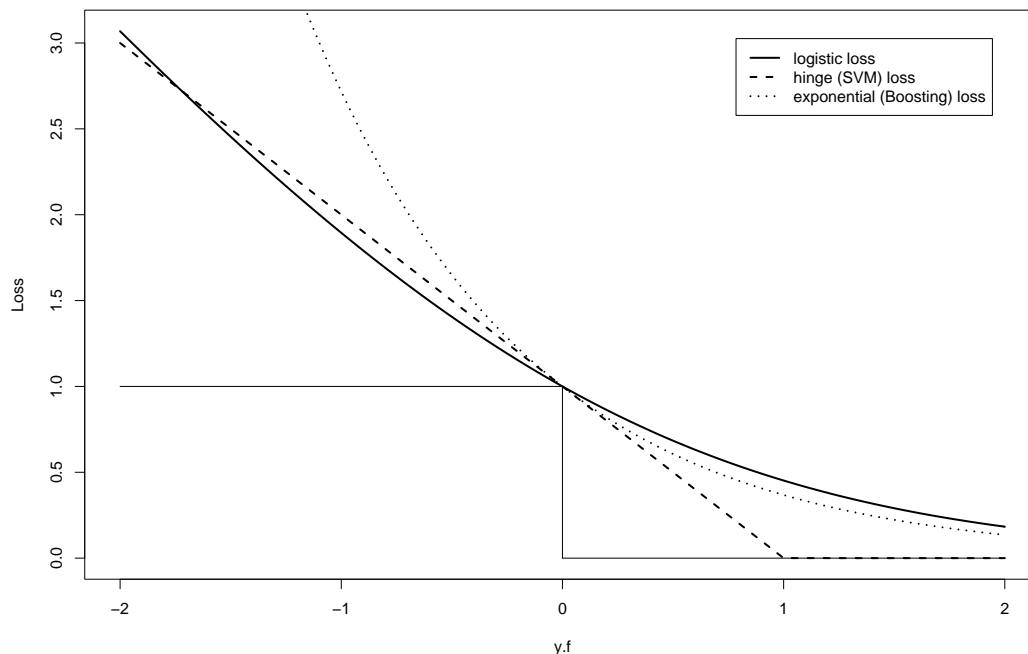


Figure 19.1. Surrogate convex loss (y -axis) vs. margin (x -axis).

Proof. Quite technical; skipped. □

19.1.1 Consequences for classification

Let ϕ be one of the surrogate convex losses (illustration shown in Figure 19.1):

1. hinge loss, used in SVM: $\phi(t) = \max(0, 1 - t)$ - loss function class is Lipschitz with $L = 1$;
2. logistic loss used in, for example, in logistic regression: $\phi(t) = \log(1 + e^{-t})$ - loss function class is Lipschitz with $L = \sup_t \left| \phi'(t) \right| = \sup_t \left| \frac{-e^{-t}}{1+e^{-t}} \right| = 1$;
3. exponential loss used in boosting: $\phi(t) = e^{-t}$ - corresponding loss function class is not globally Lipschitz. Usually, one has $\forall f \in \mathcal{F}$ that $\|f\|_\infty \leq B$, for some $B > 0$, i.e. \mathcal{F} is uniformly bounded class; in boosting, margin $|t| = |Yf(X)| \leq 1 = B$. Hence, the exponential loss class is locally Lipschitz $L = e^B$, $|t| \leq B$.

How can the Ledoux-Talagrand contraction inequality be used in practice? If a surrogate loss ϕ is L -Lipschitz, then the following bound holds with high probability $\geq 1 - \delta$, $\delta > 0$:

$$\mathbb{E}_{\text{test}} \left[\phi \left(Y \hat{f}_n(X) \right) \right] \leq \hat{\mathbb{E}} \left[\phi \left(Y \hat{f}_n(X) \right) \right] + 2\hat{R}_n(\phi \circ \mathcal{F}) + c \sqrt{\frac{\log \frac{1}{\delta}}{n}},$$

where c is constant, usually between 1 and 2; $\mathbb{E}_{\text{test}} \left[\phi \left(Y \hat{f}_n(X) \right) \right]$ is the population risk of the classifier \hat{f}_n ; $\hat{\mathbb{E}} \left[\phi \left(Y \hat{f}_n(X) \right) \right]$ is the training data risk of \hat{f}_n computed easily. Thus, by the contraction inequality,

$\hat{R}_n(\phi \circ \mathcal{F}) \leq L \hat{R}_n(\mathcal{F})$ with L being the corresponding Lipschitz constant of the loss function class. Note that the population risk of \hat{f}_n using surrogate loss function ϕ is an upper bound for the population risk under the 0 – 1 loss because of the convexity of ϕ .

19.2 Link to kernels

So far in this course, we have considered a broad class of “kernelized” methods (SVM, boosting, logistic) that choose f_n that minimize the penalized empirical risk over “nice” function classes \mathcal{F} whose elements f live in a Reproducible Kernel Hilbert Spaces (RKHS) endowed with a norm $\|f\|_{\mathcal{H}}$:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \phi \left(Y^{(i)} f \left(X^{(i)} \right) \right) + \lambda_n \|f\|_{\mathcal{H}}^2. \quad (19.2)$$

By Lagrangian duality, minimization of the objective (19.2) is equivalent to

$$\min_{\|f\|_{\mathcal{H}}^2 \leq B_n} \frac{1}{n} \sum_{i=1}^n \phi \left(Y^{(i)} f \left(X^{(i)} \right) \right),$$

where B_n depends on λ_n . Define, the effective class $\mathcal{F}_{B,\mathcal{H}} = \left\{ f \in \mathcal{H} : \|f\|_{\mathcal{H}}^2 \leq B \right\}$ which is a B -ball in the Hilbert space \mathcal{H} . Let’s try to understand $\hat{R}_n(\mathcal{F}_{B,\mathcal{H}})$. Since \mathcal{H} is a RKHS there exist a PSD kernel $\mathbb{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

Theorem 19.3. *[[The empirical Rademacher complexity of a B -ball in RKHS \mathcal{H} , $\hat{R}_n(\mathcal{F}_{B,\mathcal{H}})$, has the following properties:*

- $\hat{R}_n(\mathcal{F}_{B,\mathcal{H}}) \leq \frac{B}{\sqrt{n}} \sqrt{\text{tr}(K)} = \frac{B}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{K}(X^{(i)}, X^{(i)})}$, where $K \in \mathbb{R}^{n \times n}$ and $K_{ij} = \mathbb{K}(X^{(i)}, X^{(j)})$;
- Define kernel operator $T_{\mathbb{K}} : (T_{\mathbb{K}}f)(\cdot) = \int f(t) \mathbb{K}(\cdot, t) dP(t)$, $X \sim P$. If $T_{\mathbb{K}}$ satisfies the conditions of Mercer’s theorem then $\mathbb{E}_X \left[\hat{R}_n(\mathcal{F}_{B,\mathcal{H}}) \right] \leq \frac{B}{\sqrt{n}} \sqrt{\sum_{i=1}^{\infty} \mu_i}$, where $\mu_1 \geq \mu_2 \geq \dots$ are the eigenvalues of $T_{\mathbb{K}}$.

Before we prove theorem 19.3, let’s note that λ_n (and thus $B_n = B_n(\lambda_n)$) controls the trade-off between the empirical risk and the richness of $\mathcal{F}_{B_n,\mathcal{H}}$; as the empirical risk goes down when $n \rightarrow \infty$, the norm penalty will go up. Hence, choosing λ_n sufficiently small will ensure proper balance between risk and penalty; consequently, $\lambda_n \rightarrow 0$, as $n \rightarrow \infty$. Moreover, from previous lectures we know that the spectrum of the kernel operator $T_{\mathbb{K}}$ determines the complexity of the kernel; when $T_{\mathbb{K}}$ has finite number of eigenvalues, the functional class $\mathcal{F}_{B,\mathcal{H}}$ is not very rich.

Proof. a) Using properties of RKHS,

$$\begin{aligned} \sup_{f \in \mathcal{F}_{B,\mathcal{H}}} \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} f \left(X^{(i)} \right) &= \\ \text{(by representer theorem)} &= \sup_{f \in \mathcal{F}_{B,\mathcal{H}}} \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \left\langle f, \mathbb{K} \left(\cdot, X^{(i)} \right) \right\rangle \\ \text{(by inner product linearity)} &= \sup_{f \in \mathcal{F}_{B,\mathcal{H}}} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \mathbb{K} \left(\cdot, X^{(i)} \right), f \right\rangle \\ &= B \left\| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \mathbb{K} \left(\cdot, X^{(i)} \right) \right\|_{\mathcal{H}}, \end{aligned}$$

where last inequality follows from the observation that $\max_{x: \|x\|_{\mathcal{H}} \leq B} \langle x, a \rangle = \max_{x: \|x\|_{\mathcal{H}} = B} \langle x, a \rangle = B \|a\|_{\mathcal{H}}$ using Cauchy-Schwartz inequality. Hence,

$$\begin{aligned}
 \hat{R}_n(\mathcal{F}_{B,\mathcal{H}}) &= B \mathbb{E}_{\sigma} \left\| \frac{1}{n} \sum_{i=1}^n \sigma^{(i)} \mathbb{K}(\cdot, X^{(i)}) \right\|_{\mathcal{H}} \\
 &= B \mathbb{E}_{\sigma} \left[\sqrt{\frac{1}{n^2} \sum_{i,j=1}^n \sigma^{(i)} \sigma^{(j)} \mathbb{K}(X^{(i)}, X^{(j)})} \right] \\
 \text{(by Jensen's inequality)} &\leq B \left[\sqrt{\frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}_{\sigma} [\sigma^{(i)} \sigma^{(j)}] \mathbb{K}(X^{(i)}, X^{(j)})} \right] \\
 &= \frac{B}{\sqrt{n}} \left[\sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{K}(X^{(i)}, X^{(i)})} \right],
 \end{aligned}$$

since $\mathbb{E}_{\sigma} [\sigma^{(i)} \sigma^{(j)}] = \mathbb{I}(i = j)$.

b) Moreover,

$$\mathbb{E}_X [\hat{R}_n(\mathcal{F}_{B,\mathcal{H}})] \leq \frac{B}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E}_X [\mathbb{K}(X^{(i)}, X^{(i)})]} \leq \frac{B}{\sqrt{n}} \sqrt{\mathbb{E}_X [\mathbb{K}(X, X)]}.$$

By Mercer's theorem:

$$\mathbb{K}(x, y) = \sum_{i=1}^{\infty} \mu_i \psi_i(x) \psi_i(y),$$

where $\psi_i(x)$ are the eigenfunctions of $T_{\mathbb{K}}$, which are orthonormal: $\int \psi_i(x) \psi_j(x) dP(X) = \mathbb{I}(i = j)$. Hence,

$$\mathbb{E}_X [\mathbb{K}(X, X)] = \sum_{i=1}^{\infty} \mu_i \mathbb{E}_X [\psi_i^2(x)] = \sum_{i=1}^{\infty} \mu_i.$$

□