COMP8601: Advanced Topics in Theoretical Computer Science
Lecture 11: Dudley's Integral
Lecturer: Hubert Chan
Date: 19 Nov 2013

These lecture notes are supplementary materials for the lectures. They are by no means substitutes for attending lectures or replacement for your own notes!

1 Upper Bound for the Rademacher Averages

Recall that given a class C of functions from $S = \{x_1, \ldots, x_m\}$ to $\mathbb{R}$, the Rademacher averages of C is defined as $R_S(C) = \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^m \sigma_i f(x_i)\right]$, where the σ_i's are independent and uniform random variables taken from $\{-1, +1\}$. In this lecture we give an upper bound for $R_S(C)$ that is diminishing as m increases. We denote by $\mathbb{R}^S$ the collection of functions from S to $\mathbb{R}$.

The following lemma gives an upper bound for $R_S(C)$ when C is finite.

Lemma 1.1 (Massart's Lemma) *Let $\mathcal{V}$ be a finite subset of $\mathbb{R}^S$ with $|S| = m$ where each member v of $\mathcal{V}$ is denoted by $v = (v_1, \ldots, v_m)$. Let $\sigma_1, \ldots, \sigma_m$ be random variables chosen from $\{-1, +1\}$ uniformly at random such that all σ_i's are independent. Let $r := \max_{v \in \mathcal{V}} \sqrt{\sum_{i=1}^m v_i^2}$. Then we have*

$$\mathbf{E}\left[\max_{v \in \mathcal{V}} \frac{1}{m} \sum_{i=1}^m \sigma_i v_i\right] \leq \frac{r\sqrt{2 \ln |\mathcal{V}|}}{m}.$$

When $C \subseteq \{0,1\}^S$ is a class of boolean functions, the size of C is at most 2^m, which is finite. Also we have $\max_{f \in C} \sqrt{\sum_{i=1}^m (f(x_i))^2} \leq \sqrt{m}$. Then, by Massart's lemma we can give an upper bound for $R_S(C)$ as

$$R_S(C) \leq \frac{\sqrt{m} \cdot \sqrt{2 \ln 2^m}}{m} = \sqrt{2 \ln 2}.$$

However, this upper bound is a constant, which is not small enough for large m. In the next section we give a tighter upper bound for $R_S(C)$ using Dudley's integral.

2 Dudley's Integral

Definition 2.1 (Cover and Covering Number) *For a metric space $(\mathcal{A}, \rho)$ and a subset $C \subseteq \mathcal{A}$, we say $T \subseteq \mathcal{A}$ is an ϵ-*cover *of (C, ρ) if for all $f \in C$, there exists $t \in T$ such that $\rho(f, t) \leq \epsilon$. The ϵ-*covering number *of (C, ρ) is the minimum cardinality of ϵ-covers of (C, ρ), which we denote by $N(\epsilon, C, \rho) = \min\{|T| : T \text{ is an } \epsilon\text{-cover of } (C, \rho)\}$.*

Given $S = \{x_1, \ldots, x_m\}$, we consider the metric space $(\mathbb{R}^S, L_2^S)$, where the metric L_2^S is defined as for all $f, g \in \mathbb{R}^S$, we have $L_2^S(f, g) := \sqrt{\frac{1}{m} \sum_{i=1}^m (f(x_i) - g(x_i))^2}$.

Theorem 2.2 (Dudley's Integral [1]) *Let C be a class of functions from $S = \{x_1, \ldots, x_m\}$ to $\mathbb{R}$. Let h be the zero function such that $h(x) = 0$ for all $x \in S$. Suppose $B := \sup_{f \in C} L_2^S(f, h)$ is finite and $N(\epsilon, C, L_2^S)$ is the ϵ-covering number of (C, L_2^S). Then, $R_S(C) \leq 12 \int_0^B \sqrt{\frac{\ln N(\epsilon, C, L_2^S)}{m}} d\epsilon$.*

Proof: Let k be a positive integer. For all $0 \leq j \leq k$, define $\epsilon_j := B \cdot 2^{-j}$ and let T_j be a minimum ϵ_j-cover of (C, L_2^S). It follows that $|T_j| = N(\epsilon_j, C, L_2^S)$. We let $T_0 := \{h\}$ since $L_2^S(f, h) \leq B = \epsilon_0$ for all $f \in C$. Note that $N(\epsilon, C, L_2^S)$ is non-increasing with respect to ϵ, hence $|T_{j-1}| = N(\epsilon_{j-1}, C, L_2^S) \leq N(\epsilon_j, C, L_2^S) = |T_j|$ for $0 < j \leq k$. Without loss of generality we assume T_k is a finite set (and hence all T_j's are finite sets), since otherwise $N(\epsilon, C, L_2^S)$ is unbounded for $0 \leq \epsilon \leq \epsilon_k$, in which case the integral is also unbounded and the inequality is trivially true.

For each $f \in C$ and $0 \leq j \leq k$, let $f_j \in T_j$ be a function such that f_j covers f in T_j, that is, $L_2^S(f, f_j) \leq \epsilon_i$. Then we can represent each f by $f = f - f_k + \sum_{j=1}^{k}(f_j - f_{j-1})$, where $f_0 = h$. Then we have

$$\begin{aligned} R_S(C) &= \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(x_i) - f_k(x_i) + \sum_{j=1}^{k} (f_j(x_i) - f_{j-1}(x_i)) \right) \right] \\ &\leq \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(x_i) - f_k(x_i) \right) \right] + \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{j=1}^{k} \sum_{i=1}^{m} \sigma_i \left(f_j(x_i) - f_{j-1}(x_i) \right) \right] \\ &\leq \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(x_i) - f_k(x_i) \right) \right] + \sum_{j=1}^{k} \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f_j(x_i) - f_{j-1}(x_i) \right) \right]. \quad (2.1) \end{aligned}$$

We consider the first and second terms in the last expression respectively. For the first term, recall that the σ_i's are random variables taken from $\{-1, +1\}$. Applying the Cauchy-Schwartz inequality we obtain

$$\begin{aligned} &\mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(x_i) - f_k(x_i) \right) \right] \leq \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sqrt{\sum_{i=1}^{m} \sigma_i^2 \sum_{i=1}^{m} (f(x_i) - f_k(x_i))^2} \right] \\ =&\mathbf{E}_\sigma \left[\sup_{f \in C} \sqrt{\frac{1}{m} \sum_{i=1}^{m} (f(x_i) - f_k(x_i))^2} \right] = \mathbf{E}_\sigma \left[\sup_{f \in C} L_2^S(f, f_k) \right] \leq \epsilon_k. \quad (2.2) \end{aligned}$$

Now we consider the second term $\sum_{j=1}^{k} \mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f_j(x_i) - f_{j-1}(x_i) \right) \right]$. We fix j, and define $g_f := f_j - f_{j-1}$. That is, we define a new function g_f for each $f \in C$. Let $\mathcal{G} := \{g_f : f \in C\}$ be the collection of g functions. It follows that

$$\mathbf{E}_\sigma \left[\sup_{f \in C} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f_j(x_i) - f_{j-1}(x_i) \right) \right] = \mathbf{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i) \right] \quad (2.3)$$

Since $f_j \in T_j$ and $f_{j-1} \in T_{j-1}$, we have $|\mathcal{G}| \leq |T_j||T_{j-1}| \leq |T_j|^2$. Since T_j is finite, the set $\mathcal{G}$ is also

finite. Also note that for each $g = g_f \in \mathcal{G}$ for some $f \in C$,

$$\sqrt{\sum_{i=1}^m g_f^2(x_i)} = L_2^S(f_j, f_{j-1})\sqrt{m} \leq (L_2^S(f, f_j) + L_2^S(f, f_{j-1}))\sqrt{m} \leq (\epsilon_j + \epsilon_{j-1})\sqrt{m} = 3\epsilon_j\sqrt{m},$$

that is, $\sup_{g\in\mathcal{G}} \sqrt{\sum_{i=1}^m (g(x_i))^2} \leq 3\epsilon_j\sqrt{m}$. Applying Massart's Lemma to the functions $\mathcal{G}$, we obtain

$$\mathbf{E}_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \sigma_i g(x_i)\right] \leq \frac{3\epsilon_j\sqrt{m}\cdot\sqrt{2\ln|\mathcal{G}|}}{m} \leq 6\epsilon_j\sqrt{\frac{\ln|T_j|}{m}}. \tag{2.4}$$

Combining (2.1), (2.2), (2.3) and (2.4) we get

$$\begin{aligned}
R_S(C) &\leq \epsilon_k + 6\sum_{j=1}^k \epsilon_j\sqrt{\frac{\ln|T_j|}{m}} \\
&= \epsilon_k + 12\sum_{j=1}^k (\epsilon_j - \epsilon_{j+1})\sqrt{\frac{\ln N(\epsilon_j, C, L_2^S)}{m}} \\
&= \epsilon_k + 12\sum_{j=1}^k \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\ln N(\epsilon_j, C, L_2^S)}{m}}d\epsilon \\
&\leq \epsilon_k + 12\sum_{j=1}^k \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\frac{\ln N(\epsilon, C, L_2^S)}{m}}d\epsilon \\
&= \epsilon_k + 12\int_{\epsilon_{k+1}}^{\epsilon_1} \sqrt{\frac{\ln N(\epsilon, C, L_2^S)}{m}}d\epsilon,
\end{aligned}$$

where the second inequality follows from $N(\epsilon, C, L_2^S) \geq N(\epsilon_j, C, L_2^S)$ for all $\epsilon_{j+1} \leq \epsilon \leq \epsilon_j$. Taking $k \to \infty$ implies $R_S(C) \leq 12\int_0^{\frac{B}{2}} \sqrt{\frac{\ln N(\epsilon,C,L_2^S)}{m}}d\epsilon \leq 12\int_0^B \sqrt{\frac{\ln N(\epsilon,C,L_2^S)}{m}}d\epsilon$. ■

Note that Lemma 1.5 in notes 9 holds as a special case of Theorem 2.2. Also, if we can further give an upper bound for $N(\epsilon, C, L_2^S)$ that is independent of m, then the bound for $R_S(C)$ is diminishing with respect to m. In the next lecture we give an upper bound for $N(\epsilon, C, L_2^S)$ independent of m.

3 Homework Preview

Massart's Lemma. Let $\mathcal{V}$ be a finite subset of $\mathbb{R}^S$ with $|S| = m$ where each member v of $\mathcal{V}$ is denoted by $v = (v_1, \ldots, v_m)$. Let $\sigma_1, \ldots, \sigma_m$ be random variables chosen from $\{-1, +1\}$ uniformly at random such that all σ_i's are independent.

(a) **Jensen's Inequality.** Suppose X is a random variable and $f : \mathbb{R} \mapsto \mathbb{R}$ is a differentiable convex function. Prove that $\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$.

(Hint: A differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}$, it holds that $f(x) \geq f(y) + f'(y)(x - y)$.)

(b) Let $\mu := \mathbf{E}[\max_{v\in\mathcal{V}} \sum_{i=1}^{m} \sigma_i v_i]$. Suppose $\lambda > 0$ is some constant. Prove that $e^{\lambda\mu} \leq \sum_{v\in\mathcal{V}} \prod_{i=1}^{m} \mathbf{E}\left[e^{\lambda\sigma_i v_i}\right]$.
(Hint: The function $f(x) := e^{\lambda x}$ is convex.)

(c) Let $r := \max_{v\in\mathcal{V}} \sqrt{\sum_{i=1}^{m} v_i^2}$. Prove that $\mu \leq r\sqrt{2\ln|\mathcal{V}|}$.
(Hint: For $x \in \mathbb{R}$, it holds that $\frac{e^x+e^{-x}}{2} \leq e^{\frac{x^2}{2}}$.)

References

[1] R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.