COMP8601: Advanced Topics in Theoretical Computer Science
Lecture 10: ϵ-Sample and Rademacher averages
Lecturer: Hubert Chan
Date: 12 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 Better Result for ϵ-Sample

Recall that we have a set X with some distribution D, and C is a class of boolean functions on X such that the VC-dimension is d. The ϵ-sample is defined as follows.

Definition 1.1 *An ϵ-sample S for a set X with distribution D under a class C of boolean functions on X is a bag (multi-set) of points from X that satisfies $\forall f \in C$, $|E_X[f] - \mathsf{Avg}_S[f]| \leq \epsilon$.*

We draw m independent samples from X to form a random multi-set S, and we wish to find out how large m has to be in order for S to be an ϵ-sample with high probability. The following is the result we proved last time.

Theorem 1.2 *Suppose (X, C) has VC-dimension at most d. Moroever, suppose S is a bag of points in X obtained by sampling from X under distribution D independently m times. If $m \geq \Omega(\frac{1}{\epsilon^2}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$, then with probability at least $1 - \delta$, S is an ϵ-sample.*

In this lecture, we are going to prove a better result for ϵ-sample using the concept of Rademacher averages. In particular, we show the following result.

Theorem 1.3 *Suppose (X, C) has VC-dimension at most d. Then, suppose S is a bag of points in X obtained by sampling from X under distribution D independently m times. If $m \geq \Omega(\frac{1}{\epsilon^2}(d + \log \frac{1}{\delta}))$, then with probability at least $1 - \delta$, S is an ϵ-sample.*

Remark. Notice that this result improves the former one by eliminating a factor of $\log \frac{1}{\epsilon}$. We see that more advanced techniques are required to prove this stronger result.

Proof Skeleton. Before presenting the detailed proof of Theorem 1.3, we first present some useful lemmas, whose combinations prove Theorem 1.3. Using the same notation C and S as above, we denote the Rademacher average of C on S by $R_S(C)$, which will be defined later. It will be convenient to view $S \in X^m$ as a vector with m coordinates by imposing an arbitrary order on the elements in S.

Lemma 1.4 (Measure Concentration via Rademacher Averages) *Suppose (X, C) has VC-dimension at most d. Then, suppose S is a bag of points in X obtained by sampling from X under distribution D independently m times. With probability $1 - \delta$, for any function $f \in C$, we have $|\mathsf{Avg}_S[f] - E_X[f]| \leq 2R_S(C) + 3\sqrt{\frac{1}{2m} \ln \frac{4}{\delta}}$.*

For a metric space (C, ρ) and a constant $\epsilon \geq 0$, let $N(\epsilon, C, \rho)$ be the ϵ-covering number of C, which will be defined later. Let L_2^S be a metric space on C with respect to $S = (x_1, x_2, \ldots, x_m)$ such that for all $f, g \in C$, $L_2^S(f, g) = \sqrt{\frac{1}{m} \sum_{i=1}^m (f(x_i) - g(x_i))^2}$. The following two lemmas upper bounds

the Rademacher averages.

Lemma 1.5 (Dudley's Integral) *For any class C of functions from X to $\{0,1\}$ and m points $x_1, x_2, \ldots, x_m \in X$, let $S = (x_1, x_2, \ldots, x_m)$, we have $R_S(C) \leq 12 \int_0^1 \sqrt{\frac{\log N(\epsilon, C, L_2^S)}{m}} d\epsilon$.*

Lemma 1.6 (Haussler's Theorem) *There exists a constant $c > 1$ such that the following holds. For any class C of functions from X to $\{0,1\}$ and m points $x_1, x_2, \ldots, x_m \in X$, suppose $S = \{x_1, x_2, \ldots, x_m\}$ and (X, C) has VC-dimension at most d, we have $N(\epsilon, C, L_2^S) \leq (\frac{c}{\epsilon})^{2d}$.*

Before proceeding to prove those lemmas, we combine them to give a proof of Theorem 1.3.

Proof of Theorem 1.3: By Lemmas 1.5 and 1.6, we can upper bound the Rademacher averages:

$R_S(C) \leq 12 \int_0^1 \sqrt{\frac{2d \log(\frac{c}{\epsilon})}{m}} d\epsilon = 12\sqrt{\frac{2d}{m}} \int_0^1 \sqrt{\log \frac{c}{\epsilon}} d\epsilon = c'\sqrt{\frac{d}{m}}$, for some constant c' depending on c.

Hence according to lemma 1.4, if $|S| = m$, we have, with probability $1 - \delta$, for all $f \in C$,

$|\mathsf{Avg}_S[f] - E_X[f]| \leq 2R_S(C) + 3\sqrt{\frac{1}{2m} \ln \frac{4}{\delta}} \leq 2c'\sqrt{\frac{d}{m}} + 3\sqrt{\frac{1}{2m} \ln \frac{4}{\delta}}$.

By setting $m \geq \max(\frac{16c'^2 d}{\epsilon^2}, \frac{18}{\epsilon^2} \ln \frac{4}{\delta}) = \Theta(\frac{1}{\epsilon^2}(d + \log \frac{1}{\delta}))$, we have, with probability at least $1 - \delta$, for all $f \in C$, $|\mathsf{Avg}_S[f] - E_X[f]| \leq 2c'\sqrt{\frac{d}{m}} + 3\sqrt{\frac{1}{2m} \ln \frac{4}{\delta}} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, which means that S is an ϵ-sample of X, by definition. ■

2 ϵ-Sample and Rademacher Averages

We prove in this section Lemma 1.4, which upper bounds the difference between the fraction of 1's in the bag of sampled points and that in the whole set, using the Rademacher averages of the class C of functions on X.

Definition 2.1 (Rademacher Averages) *For a class C of functions from X to $\mathbb{R}$ and $S = (x_1, x_2, \ldots, x_m) \in X^m$, define the Rademacher average of C with respect to S as*

$$R_S(C) = E_\sigma[\sup_{f \in C} \frac{1}{m} \sum_{i=0}^{m} (\sigma_i f(x_i))],$$

where $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$, and each σ_i is sampled independently and uniformly at random from $\{-1, 1\}$.

The following measure concentration result from McDiarmid is useful for the proof of lemma 1.4.

Lemma 2.2 (McDiarmid's Inequality) *Let $x_1, x_2, \ldots, x_m$ be independent random objects and let $h(x_1, x_2, \ldots, x_m)$ be a function on $(x_1, x_2, \ldots, x_m)$ that satisfies the "c_i-Lipschitz" property:*

$\forall i, \forall x_1, x_2, \ldots, x_m, x_i', \; |h(x_1, \ldots, x_i, \ldots, x_m) - h(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i,$

then for any $\epsilon > 0$, we have $\Pr[h - E[h] > \epsilon] \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2})$

Remark. Observe that if we consider the function $(-h)$, then we can show that:

$\Pr[|h - E[h]| > \epsilon] \leq 2\exp(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2})$

Proof of Lemma 1.4: We only prove one side of the inequality: with probability $1-\frac{\delta}{2}$, for any function f in C, $\mathsf{Avg}_S[f] - E_X[f] \le 2R_S(C) + 3\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}$. The other side can be proved similarly.

Define $M(S) = \sup_{f\in C}(\mathsf{Avg}_S[f] - E_X[f])$ to be the maximum difference of fractions among all functions in C. Notice that $M(S)$ satisfies the $\frac{1}{m}$-Lipschitz property on all m coordinates and hence by McDiarmid's Inequality, $\Pr[M(S) - E[M(S)] > \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}] \le \exp(-\frac{2\cdot\frac{1}{2m}\ln\frac{4}{\delta}}{m(\frac{1}{m})^2}) = \frac{\delta}{4}$.

Hence with probability at least $1-\frac{\delta}{4}$, $M(S) \le E[M(S)] + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}$.

Consider $-R_S(C)$ as a function of S, we can see that it also satisfies $\frac{1}{m}$-Lipschitz property on all m coordinates and hence by McDiarmid's Inequality, with probability at least $1-\frac{\delta}{4}$, $E_S[R_S(C)] \le R_S(C) + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}$.

Thus it suffices to show that $E_S[M(S)] \le 2E_S[R_S(C)]$, which implies that with probability $1-\frac{\delta}{2}$, for any function $f \in C$, $\mathsf{Avg}_S[f] - E_X[f] \le M(S) \le E[M(S)] + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}} \le 2E_S[R_S(C)] + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}} \le 2R_S(C) + 3\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}$.

Suppose we make another m samples independently and form $S' = (x'_1, x'_2, \ldots, x'_m)$, and observe that $E_X[f] = E_{S'}[\mathsf{Avg}_{S'}[f]]$. Hence we have

$E_S[M(S)] = E_S[\sup_{f\in C}(\mathsf{Avg}_S[f] - E_X[f])] \le E_{S,S'}[\sup_{f\in C}(\mathsf{Avg}_S[f] - \mathsf{Avg}_{S'}[f])]$.

The inequality holds since for any collection of random variables U_j's, $\sup_j(\mathbf{E}[U_j]) \le \mathbf{E}\big[\sup_j(U_j)\big]$.

By the definition of $\mathsf{Avg}_S[f]$, we have that

$$\begin{aligned} E_{S,S'}[\sup_{f\in C}(\mathsf{Avg}_S[f] - \mathsf{Avg}_{S'}[f])] &= E_{S,S'}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m f(x_i) - \frac{1}{m}\sum_{i=1}^m f(x'_i))] \\ &= E_{S,S'}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m (f(x_i) - f(x'_i)))] \\ &= E_{S,S',\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i(f(x_i) - f(x'_i)))], \end{aligned}$$

where $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ and each σ_i is independently uniformly distributed in $\{-1, 1\}$. The last equality holds since it does not affect the expectation if we flip fair coins independently to decide whether we swap x_i and x'_i in the ith coordinate when sampling S and S'. Notice that

$$\begin{aligned} E_{S,S',\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i(f(x_i) - f(x'_i)))] &\le E_{S,\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i))] + E_{S',\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m (-\sigma_i) f(x'_i))] \\ &= 2E_{S,\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i))] = 2E_S[R_S(C)]. \end{aligned}$$

The inequality holds since the right hand side allows us to choose different functions in C to maximize each sum and the first equality holds by the symmetry of σ_i.

Combining all the equalities and inequalities, we have $E_S[M(S)] \le 2E_{S,\sigma}[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i))] = 2E_S E_\sigma[\sup_{f\in C}(\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i))] = 2E_S[R_S(C)]$, which according to the analysis above, implies one side of Lemma 1.4. ■

3 Homework Preview

1. **Details in the Proof of Theorem 1.3.** Prove that for any constant $c \ge 1$, we have $\int_0^1 \sqrt{\log \frac{c}{x}}dx = \Theta(1)$.

2. **McDiarmid's Inequality.** Let $\mathcal{O}$ be a set of objects. Let $X_1, X_2, \ldots, X_m$ be independent random objects taken from $\mathcal{O}$. Let $h : \mathcal{O}^m \mapsto \mathbb{R}$ be a function that satisfies the "c_i-Lipschitz" property: for all $i \in [m]$ and objects $x_1, x_2, \ldots, x_m, x_i' \in \mathcal{O}$,

$$|h(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m) - h(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \le c_i.$$

   (a) Let Y be a real random variable with $\mathbf{E}[Y] = 0$ such that $a \le Y \le b$. Let $\lambda \in \mathbb{R}$. We prove that $\mathbf{E}\big[e^{\lambda Y}\big] \le e^{\frac{\lambda^2(b-a)^2}{8}}$.

      i. Suppose $a < b$ and $p := -\frac{a}{b-a}$. Define function $L(z) := -pz + \ln(1 - p + pe^z)$ for $z \in \mathbb{R}$. Prove that $\mathbf{E}\big[e^{\lambda Y}\big] \le e^{L(\lambda(b-a))}$.
      (Hint: The function $e^{\lambda z}$ with respect to z is convex.)

      ii. Prove that $L(\lambda(b-a)) \le \frac{\lambda^2(b-a)^2}{8}$.
      (Hint: If a function L is twice differentiable in $\mathbb{R}$, then for $y \in \mathbb{R}$ we have $L(y) = L(0) + L'(0)y + \frac{L''(z)}{2}y^2$ for some z between 0 and y.)

   (b) For $i \in \{0, 1, \ldots, m\}$ define $X^{(i)} := (X_1, \ldots, X_i)$, where $X^{(0)}$ can be considered as a random object independent of $X_1, \ldots, X_m$. Let $X := X^{(m)}$. Define $Z_i := \mathbf{E}\big[h(X)|X^{(i)}\big]$. Let $t > 0$. Prove that $\mathbf{E}\big[e^{t(Z_i - Z_{i-1})}|X^{(i-1)}\big] \le e^{\frac{t^2c_i^2}{8}}$ for $1 \le i \le m$.
      (Hint: Consider the random variable $(Z_i - Z_{i-1})|X^{(i-1)}$. What is $\mathbf{E}\big[Z_i - Z_{i-1}|X^{(i-1)}\big]$? Find a and b such that $a \le (Z_i - Z_{i-1})|X^{(i-1)} \le b$ and that $b - a \le c_i$. Try $a = \inf_{x\in\mathcal{O}}(\mathbf{E}\big[h(X)|X^{(i-1)}, X_i = x\big] - \mathbf{E}\big[h(X)|X^{(i-1)}\big])$ and $b = \sup_{x\in\mathcal{O}}(\mathbf{E}\big[h(X)|X^{(i-1)}, X_i = x\big] - \mathbf{E}\big[h(X)|X^{(i-1)}\big])$.)

   (c) Prove that for $1 \le i \le m$, we have $\mathbf{E}\big[e^{t(Z_i - Z_0)}\big] \le e^{\frac{t^2c_i^2}{8}}\mathbf{E}\big[e^{t(Z_{i-1} - Z_0)}\big]$.
      (Hint: For arbitrary functions f and g, we have $\mathbf{E}\big[\mathbf{E}\big[f(X^{(i-1)})g(X^{(i)})|X^{(i-1)}\big]\big] = \mathbf{E}\big[f(X^{(i-1)})\mathbf{E}\big[g(X^{(i)})|X^{(i-1)}\big]\big]$.)

   (d) Prove that for $\epsilon > 0$, we have $\Pr(h(X) - \mathbf{E}[h(X)] > \epsilon) \le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right)$.

3. **$\frac{1}{m}$-Lipschitz Property.** In this question we give some details on the proof of Lemma 1.4. Recall that given a set X of points and a class C of boolean functions on X, we denote by $S = (x_1, \ldots, x_m)$ a bag of points sampled from X. Let $M(S) := \sup_{f\in C}(\mathsf{Avg}_S[f] - E_X[f])$, where $\mathsf{Avg}_S[f]$ is the fraction of points in S that takes value 1 under function f. Also let $R_S(C) := E_\sigma[\sup_{f\in C}\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)]$, where $\sigma = (\sigma_1, \ldots, \sigma_m)$ and each σ_i is sampled independently and uniformly at random from $\{-1, 1\}$.

(a) Prove that $M(S)$ satisfies the $\frac{1}{m}$-Lipschitz property on all m coordinates.

(b) Prove that $R_S(C)$ as a function of S satisfies the $\frac{1}{m}$-Lipschitz property on all m coordinates.