DSC 190/291 · Assignment 3 Solutions

UCSD · Spring 2026

Part A: A Mini-Course on Concentration Inequalities (65 points)

In the Week 3 lecture we proved the following i.i.d. growth-function ERM guarantee. Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. Then with high probability over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

The learning-theoretic structure of the proof was covered in lecture, but several concentration inequalities were used without proof. The inequalities to learn in this mini-course are:

- Hoeffding's inequality,
- Hoeffding's inequality for sampling without replacement,
- McDiarmid's inequality,
- Bernstein's inequality.

Solution.

The following is a complete mini-course.

Concentration

A concentration inequality says that a random quantity is unlikely to be far from a typical value, usually its expectation. The most basic example is an empirical average. If $Z_1, ..., Z_n$ are independent random variables in $[0, 1]$ with common mean μ, then $|(Z) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ should be close to μ because no single variable can move the average very much and the independent fluctuations tend to cancel. For example, if $Z_i = \mathbf{1}[h(x_i) \neq y_i]$ for a fixed classifier h, then $|(Z) = L_n(h)$ and $\mu = L_{\mathcal{D}}(h)$. Concentration is exactly the statement that empirical risk tracks population risk for a fixed h.

The common intuition is: many small, bounded, independent contributions produce a stable aggregate. Different inequalities formalize this in different settings. Hoeffding controls bounded independent sums. Hoeffding without replacement controls averages from a fixed finite population. McDiarmid controls any function of independent inputs when changing one input has a bounded effect. Bernstein improves Hoeffding when the variance is small.

The MGF Method

The moment-generating-function method begins with Markov's inequality applied to an exponential. If X is any random variable and $\lambda > 0$, then

$$\mathbb{P}(X - \mathbb{E}X \geq t) = \mathbb{P}(\exp(\lambda(X - \mathbb{E}X)) \geq \exp(\lambda t)) \leq \exp(-\lambda t)\mathbb{E}\exp(\lambda(X - \mathbb{E}X)).$$

Thus a tail bound follows once we upper-bound the centered MGF $\mathbb{E}\exp(\lambda(X-\mathbb{E}X))$.

For sums of independent variables, MGFs factor. If $X=\sum_{i=1}^n X_i$ and the X_i are independent, then

$$\mathbb{E}\exp\left(\lambda\sum_{i=1}^n(X_i-\mathbb{E}X_i)\right)=\prod_{i=1}^n\mathbb{E}\exp(\lambda(X_i-\mathbb{E}X_i)).$$

The proof of Hoeffding uses a uniform bounded-range MGF bound. Bernstein uses a variance-sensitive MGF bound. McDiarmid uses the same exponential method after replacing the random variable by a Doob martingale.

A useful lemma behind Hoeffding is Hoeffding's lemma: if $X\in[a,b]$ almost surely, then for every $\lambda\in\mathbb{R}$,

$$\mathbb{E}\exp(\lambda(X-\mathbb{E}X))\le\exp\left(\lambda^2\frac{(b-a)^2}{8}\right).$$

For $X\in[0,1]$, this becomes $\mathbb{E}\exp(\lambda(X-\mathbb{E}X))\le\exp\left(\frac{\lambda^2}{8}\right)$.

Hoeffding's Inequality

Statement. Let $X_1,...,X_n$ be independent random variables with $X_i\in[a_i,b_i]$ almost surely. Then for every $t>0$,

$$\mathbb{P}\left(\sum_{i=1}^n(X_i-\mathbb{E}X_i)\ge t\right)\le\exp\left(-2\frac{t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right).$$

The same bound holds for the lower tail. Therefore, if $X_i\in[0,1]$ and $|(X)=\frac{1}{n}\sum_i X_i$, then

$$\mathbb{P}(||(X)-\mathbb{E}|(X)|\ge\varepsilon)\le 2\exp(-2n\varepsilon^2).$$

Proof. For $\lambda>0$, Chernoff's bound and independence give

$$\mathbb{P}\left(\sum_i(X_i-\mathbb{E}X_i)\ge t\right)\le\exp(-\lambda t)\prod_{i=1}^n\mathbb{E}\exp(\lambda(X_i-\mathbb{E}X_i)).$$

Applying Hoeffding's lemma to each factor,

$$\mathbb{P}\left(\sum_i(X_i-\mathbb{E}X_i)\ge t\right)\le\exp\left(-\lambda t+\frac{\lambda^2}{8}\sum_{i=1}^n(b_i-a_i)^2\right).$$

Optimizing over λ gives $\lambda=4\frac{t}{\sum_i}(b_i-a_i)^2$, which yields the displayed upper-tail bound. The lower tail follows by applying the same argument to $-X_i$.

Sampling Without Replacement

Statement. Let $z_1,...,z_N\in[0,1]$ be fixed, and let T be drawn uniformly from all size-n subsets of $[N]$. Then for every $\varepsilon>0$,

$$\mathbb{P}_T\left(\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| > \varepsilon\right) \le 2\exp(-2n\varepsilon^2).$$

Proof. Let $I_1, ..., I_n$ be sampled without replacement from $[N]$, and let $J_1, ..., J_n$ be sampled independently and uniformly from $[N]$. Hoeffding's comparison theorem says that for every convex function φ,

$$\mathbb{E}\varphi\left(\sum_{r=1}^{n} z_{I_r}\right) \le \mathbb{E}\varphi\left(\sum_{r=1}^{n} z_{J_r}\right).$$

Taking $\varphi(s) = \exp(\lambda s)$ and then centering by the common mean shows that the MGF of the without-replacement sum is at most the MGF of the corresponding with-replacement sum. The with-replacement variables z_{J_r} are independent and lie in $[0, 1]$, so ordinary Hoeffding gives

$$\mathbb{P}\left(\frac{1}{n}\sum_{r=1}^{n} z_{I_r} - \frac{1}{N}\sum_{i=1}^{N} z_i \ge \varepsilon\right) \le \exp(-2n\varepsilon^2).$$

The lower tail is identical, and combining the two tails gives the stated two-sided bound.

The conclusion is that sampling without replacement is no less concentrated than independent sampling from the same finite population.

McDiarmid's Inequality

Statement. Let $X_1, ..., X_n$ be independent random variables, and let $f(X_1, ..., X_n)$ be a function satisfying the bounded-differences condition: for every i, changing only the ith input can change f by at most c_i. Then for every $t > 0$,

$$\mathbb{P}(f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) \ge t) \le \exp\left(-2\frac{t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

The two-sided version has an extra factor 2.

Proof. Define the Doob martingale

$$M_i = \mathbb{E}[f(X_1, ..., X_n) \mid X_1, ..., X_i], \quad i = 0, ..., n.$$

Then $M_0 = \mathbb{E}f$ and $M_n = f$. The martingale difference $D_i = M_i - M_{i-1}$ has conditional mean zero, and the bounded-differences condition implies that, conditional on $X_1, ..., X_{i-1}$, its range has length at most c_i. Hoeffding's lemma applied conditionally gives

$$\mathbb{E}[\exp(\lambda D_i) \mid X_1, ..., X_{i-1}] \le \exp\left(\lambda^2 \frac{c_i^2}{8}\right).$$

Iterating conditional expectations,

$$\mathbb{E}\exp(\lambda(f - \mathbb{E}f)) \le \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2\right).$$

Chernoff's bound and optimization over λ give the result.

Bernstein's Inequality

Statement. Let $X_1, ..., X_n$ be independent random variables with $\mathbb{E}X_i = 0$, $|X_i| \leq b$ almost surely, and $\sigma^2 = \sum_{i=1}^n \mathbb{E}X_i^2$. Then for every $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\left(\sigma^2 + b\frac{t}{3}\right)}\right).$$

Proof. The key MGF estimate is that, for $0 < \lambda < \frac{3}{b}$,

$$\mathbb{E}\exp(\lambda X_i) \leq \exp\left(\lambda^2 \mathbb{E}\frac{X_i^2}{2\left(1 - \lambda\frac{b}{3}\right)}\right).$$

One proves this from the Taylor expansion of e^x and the bound $|X_i| \leq b$: higher moments satisfy $\mathbb{E}|X_i|^k \leq b^{k-2}\mathbb{E}X_i^2$ for $k \geq 2$. Multiplying the MGF bounds over i gives

$$\mathbb{E}\exp\left(\lambda \sum_i X_i\right) \leq \exp\left(\lambda^2 \frac{\sigma^2}{2\left(1 - \lambda\frac{b}{3}\right)}\right).$$

Chernoff's bound gives

$$\mathbb{P}\left(\sum_i X_i \geq t\right) \leq \exp\left(-\lambda t + \lambda^2 \frac{\sigma^2}{2\left(1 - \lambda\frac{b}{3}\right)}\right).$$

Choosing $\lambda = \frac{t}{\sigma^2 + b\frac{t}{3}}$ yields Bernstein's bound.

Bernstein improves Hoeffding when the variance term σ^2 is much smaller than the worst-case range bound. For empirical averages of bounded variables, Hoeffding only uses the fact that each summand lies in an interval; Bernstein also uses the variance.

Comparison

Hoeffding applies to independent bounded sums and is range-based. Hoeffding without replacement applies to averages from a fixed finite population and gives the same rate. McDiarmid applies to general functions of independent inputs, provided one coordinate cannot change the value too much. Bernstein applies to independent bounded sums and is sharper when the variance is small.

The common template is:

1. convert a tail event into an exponential moment using Chernoff;
2. control the MGF using boundedness, variance, or bounded differences;
3. optimize the exponential parameter.

Connection to the Week 3 ERM Proof

Let $L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$ and $L_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[h(x_i) \neq y_i]$. Define the uniform generalization gap

$$\Delta(S) = \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_n(h)|.$$

If, with probability at least $1 - \delta$, $\Delta(S) \le \varepsilon_n$, then ERM satisfies

$$\begin{aligned} L_{\mathcal{D}}\big(\hat{h}\big) &\le L_n\big(\hat{h}\big) + \varepsilon_n \\ &\le L_n(h^*) + \varepsilon_n \\ &\le L_{\mathcal{D}}(h^*) + 2\varepsilon_n, \end{aligned}$$

where h^* is any hypothesis satisfying $L_{\mathcal{D}}(h^*) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \eta$. Since $\eta > 0$ is arbitrary, this gives

$$L_{\mathcal{D}}\big(\hat{h}\big) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\varepsilon_n.$$

It remains to explain why $\varepsilon_n = O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log\left(\frac{1}{\delta}\right)}{n}}\right)$. The proof has three steps.

First, McDiarmid controls the deviation of the supremum from its expectation. Changing one example in S changes each empirical risk $L_n(h)$ by at most $\frac{1}{n}$, so it changes the supremum by at most $\frac{1}{n}$. Hence McDiarmid gives a high-probability deviation term of order $\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}$.

Second, symmetrization replaces population risk by an independent ghost sample. Let $S' \sim \mathcal{D}^n$ be independent of S, and let $L'_n(h)$ be empirical risk on S'. A standard symmetrization argument gives

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)) \le \mathbb{E}_{(S,S')} \sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)),$$

The same argument applied with $L_n(h) - L_{\mathcal{D}}(h)$ in place of $L_{\mathcal{D}}(h) - L_n(h)$ gives the opposite direction, so the two one-sided bounds combine into the absolute-value bound.

Third, condition on the combined sample $S \cup S'$. The split into S and S' is a random split of a fixed pool of size $2n$. On this fixed pool, only the restriction set

$$\mathcal{H}|_C$$

matters, and its size is at most $\Gamma_{\mathcal{H}}(2n)$. Applying the sampling-without-replacement Hoeffding bound to each label pattern and union bounding over at most $\Gamma_{\mathcal{H}}(2n)$ patterns gives the transductive tail bound

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right) \le \Gamma_{\mathcal{H}}(2n) \exp\left(-n\frac{t^2}{2}\right).$$

Integrating this tail bound yields an expectation of order $\sqrt{\log \Gamma_{\mathcal{H}} \frac{2n}{n}}$.

Combining the McDiarmid deviation term with the symmetrized expectation bound gives, with probability at least $1 - \delta$,

$$\Delta(S) \le C\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log\left(\frac{1}{\delta}\right)}{n}}$$

for a universal constant C. Plugging this into the ERM chain proves

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

References

Checkable references include Hoeffding, "Probability inequalities for sums of bounded random variables" (1963), for Hoeffding's inequality and sampling without replacement; McDiarmid, "On the method of bounded differences" (1989), for bounded differences; Boucheron, Lugosi, and Massart, **Concentration Inequalities** (2013), Chapters 2 and 6, for the MGF method, Bernstein's inequality, and bounded differences; Wainwright, **High-Dimensional Statistics** (2019), Chapter 2, for Hoeffding and Bernstein inequalities; and Shalev-Shwartz and Ben-David, **Understanding Machine Learning** (2014), Chapters 3–6 and 28, for ERM, uniform convergence, VC dimension, and the No-Free-Lunch theorem.

Part B: The No-Free-Lunch Theorem and the Fundamental Theorem

(20 points)

The Week 3 lecture proved the growth-function ERM guarantee displayed in Part A and, via Sauer-Shelah, the VC corollary

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{d \log(2en/d) + \log(1/\delta)}{n}}\right),$$

where $d = \mathrm{VCdim}(\mathcal{H})$. This establishes one direction of the Fundamental Theorem of PAC learning.

Theorem (No-Free-Lunch)

Let A be any learning algorithm for binary classification with the 0-1 loss over a domain $\mathcal{X}$, and let n be any integer with $n < |\mathcal{X}|/2$. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that (i) some function $f^* : \mathcal{X} \to \{0,1\}$ satisfies $L_{\mathcal{D}}(f^*) = 0$, and (ii) with probability at least $1/7$ over $S \sim \mathcal{D}^n$, the learner's output satisfies $L_{\mathcal{D}}(A(S)) \geq 1/8$.

1. **Proof.**

   Prove the NFL theorem above. Before writing the proof, study the statement carefully and identify which objects are universally quantified and which are chosen by the adversary.

   Solution.

   The universally quantified objects are the learner A and the sample size n, subject to $n < |\mathcal{X}|/2$. The adversary may choose a distribution $\mathcal{D}$ after seeing A.

Since $n < |\mathcal{X}|/2$, choose a subset $C \subseteq \mathcal{X}$ of size $2n$. For every labeling $f : C \to \{0,1\}$, define $\mathcal{D}_f$ to be the uniform distribution on $\{(x, f(x)) : x \in C\}$. This distribution is realizable by the extension of f to all of $\mathcal{X}$, so $L_{\mathcal{D}_f}(f^*) = 0$ for some $f^* : \mathcal{X} \to \{0,1\}$.

Now choose f uniformly at random from all 2^{2n} labelings of C, and then draw $S \sim \mathcal{D}_f^n$. Let $U(S)$ be the set of points in C that do not appear among the n sampled inputs. Conditional on the sampled inputs and on the labels of the sampled points, the labels $f(x)$ for $x \in U(S)$ are still independent fair bits. Therefore, for each unseen x, the learner's prediction $A(S)(x)$ is wrong with conditional probability $\frac{1}{2}$.

Thus

$$\mathbb{E}_{f,S} L_{\mathcal{D}_f}(A(S)) \geq \mathbb{E}_S\left[\frac{|U(S)|}{4n}\right].$$

It remains to lower-bound the expected number of unseen points. For each fixed $x \in C$, $\mathbb{P}(x \text{ is unseen}) = \left(1 - \frac{1}{2n}\right)^n \geq \frac{1}{2}$. Hence

$$\mathbb{E}|U(S)| = \sum_{x \in C} \mathbb{P}(x \text{ is unseen}) \geq 2n \cdot \frac{1}{2} = n.$$

Therefore

$$\mathbb{E}_{f,S} L_{\mathcal{D}_f}(A(S)) \geq \frac{1}{4}.$$

By averaging over f, there exists one labeling f_0 such that

$$\mathbb{E}_{S \sim \mathcal{D}_{f_0}^n} L_{\mathcal{D}_{f_0}}(A(S)) \geq \frac{1}{4}.$$

Let $Z = L_{\mathcal{D}_{f_0}}(A(S))$. Since $0 \leq Z \leq 1$,

$$\mathbb{E}Z \leq \mathbb{P}\left(Z \geq \frac{1}{8}\right) \cdot 1 + \mathbb{P}\left(Z < \frac{1}{8}\right) \cdot \frac{1}{8} = \frac{1}{8} + \frac{7}{8}\mathbb{P}\left(Z \geq \frac{1}{8}\right).$$

Combining this with $\mathbb{E}Z \geq \frac{1}{4}$ gives $\mathbb{P}\left(Z \geq \frac{1}{8}\right) \geq \frac{1}{7}$. This is the required distribution.

2. **Application: closing the Fundamental Theorem.**

   State the corollary that NFL implies the lower-bound direction of the Fundamental Theorem for PAC learning. Then make the application concrete.

Solution.

The corollary is: if $\mathrm{VCdim}(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC learnable in the realizable setting, and therefore is not agnostically PAC learnable either.

Here is the proof. Suppose toward a contradiction that $\mathcal{H}$ is realizably PAC learnable by some rule A. Apply the PAC guarantee with $\varepsilon = \frac{1}{16}$ and $\delta = \frac{1}{8}$, and let $n_{\mathcal{H}}\left(\frac{1}{16}, \frac{1}{8}\right)$ be the corresponding sample size. Choose $n \geq n_{\mathcal{H}}\left(\frac{1}{16}, \frac{1}{8}\right)$. Since $\mathrm{VCdim}(\mathcal{H}) = \infty$, there is a set $C \subseteq \mathcal{X}$ of size at least $2n + 1$ shattered by $\mathcal{H}$.

Run the NFL theorem on the domain C with sample size n. It gives a distribution supported on $C \times \{0,1\}$ and a target labeling f^* such that $\mathbb{P}\big(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\big) \geq \frac{1}{7}$. Since C is shattered by $\mathcal{H}$, some $h^* \in \mathcal{H}$ agrees with f^* on all of C, so the distribution is realizable by $\mathcal{H}$. But PAC learnability at $(\varepsilon, \delta) = \big(\frac{1}{16}, \frac{1}{8}\big)$ requires $\mathbb{P}\big(L_{\mathcal{D}}(A(S)) \leq \frac{1}{16}\big) \geq \frac{7}{8}$. This contradicts the NFL lower bound, since $\frac{1}{7} > \frac{1}{8}$ and the event $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$ is disjoint from $L_{\mathcal{D}}(A(S)) \leq \frac{1}{16}$.

A concrete class with infinite VC dimension is the class of all binary functions on an infinite domain, for instance $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{x \to \mathbf{1}[x \in A] : A \subseteq \mathbb{R}\}$. Every finite set is shattered: for any finite C and any labeling of C, choose A to be exactly the positively labeled subset of C. Hence $\mathrm{VCdim}(\mathcal{H}) = \infty$, and the corollary says this class is not PAC learnable.

3. **Worked construction.**

   Provide a concrete example of a learning rule and a domain, and explicitly construct a bad distribution and target labeling for which the rule fails with at least constant probability.

   Solution.

   Let $\mathcal{X} = \{1,2,3,4\}$ and consider the learning rule A that memorizes labels it has seen and predicts 0 on every point not seen in the training sample. More precisely, if x appears in S, then $A(S)(x)$ is its observed label; otherwise $A(S)(x) = 0$.

   Let the target labeling be $f^*(x) = 1$ for all $x \in \mathcal{X}$, and let $\mathcal{D}$ be the uniform distribution on $\{(x,1) : x \in \mathcal{X}\}$. This distribution is realizable by f^*.

   Draw a sample of size 2. The learner sees at most two distinct points, and it predicts 0 on every unseen point. Since all true labels are 1, every unseen point is misclassified. Therefore, for every possible sample,

   $$L_{\mathcal{D}}(A(S)) \geq \frac{2}{4} = \frac{1}{2}.$$

   In particular,

   $$\mathbb{P}_{S \sim \mathcal{D}^2}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) = 1.$$

   This is a concrete bad distribution for this particular learning rule.

4. **Comparison to the Week 1 No-Free-Lunch theorem.**

   Compare the Week 1 and Week 3 versions. Identify the learning setting, the kind of adversary, and the form of the conclusion.

   Solution.

   The Week 1 theorem is an online mistake-bound statement. The learner predicts labels one at a time on a finite domain, and performance is measured by the number of mistakes on a sequence. The proof constructs an adversarial sequence against a deterministic learner by presenting points and choosing labels that make the learner

wrong; after the construction, these labels define a fixed function $f : \mathcal{X} \to \{0, 1\}$. The conclusion is deterministic: there exist a target function and an ordering of the domain on which the learner makes $|\mathcal{X}|$ mistakes.

The Week 3 theorem is an i.i.d. PAC lower-bound statement. The adversary chooses a distribution $\mathcal{D}$, equivalently a finite support and a target labeling on that support, before the sample is drawn. The learner receives an i.i.d. sample, outputs a hypothesis, and performance is measured by population loss. The conclusion is probabilistic: with probability at least $\frac{1}{7}$ over the training sample, the learner's population loss is at least $\frac{1}{8}$.

The adversary in Week 1 is naturally viewed as online and adaptive during the construction of the bad sequence. The adversary in Week 3 is oblivious to the realized sample: it chooses the bad distribution first, and the randomness comes from the i.i.d. draw.

Both results are called No-Free-Lunch because both say that, without a structural restriction on the class of possible labelings, no learner can guarantee success uniformly. Week 1 rules out deterministic mistake guarantees in the online setting. Week 3 rules out distributional PAC guarantees when the class can realize arbitrary labelings on arbitrarily large finite sets.