CSCI699: Machine Learning Theory Fall 2023

Lecture 2: Agnostic Learning, Uniform Convergence, Concentration

Instructor: Vatsal Sharan

These lecture notes are based on an initial version scribed by Ta-Yang Wang, Yingxiao Ye and Berk Tinaz.

1 Agnostic PAC learning

PAC learning requires the *realizability* assumption, meaning that it cannot handle noise/error within the labels.

Agnostic PAC learning relaxes the realizability distribution, and also allows the labels to be noisy. Consider an arbitrary distribution P over $\mathcal{X} \times \mathcal{Y}$. The conditional distribution of y given x is given by $P(y|x)$, which may not be deterministic.

We will continue with the zero/one loss, and our previous definitions of population and empirical risk are still preserved,

$$R(h) = \mathbb{E}_{(x,y)\sim P}\left[\ell(h(x), y)\right] = \Pr_{(x,y)\sim P}[h(x) \neq y]m \tag{1}$$

$$\hat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(x_i) \neq y_i). \tag{2}$$

With the realizability assumption, it was possible to get zero error (since the true hypothesis h^* gets zero error). A quick calculation tells us the best possible risk (i.e. the Bayes optimal risk) in the agnostic case.

$$\begin{aligned}
&\min_{h:\mathcal{X}\to\mathcal{Y}} \Pr_{(x,y)\sim P}[h(x) \neq y] \\
&= \sum_{x\in\mathcal{X}} P(x) \left(\min_{h(x)\in\{0,1\}} \{\Pr_{(x,y)\sim P}(y=1|x)\mathbb{1}\{h(x)=0\}, \Pr_{(x,y)\sim P}(y=0|x)\mathbb{1}\{h(x)=1\}\} \right) \\
&= \sum_{x\in\mathcal{X}} P(x) \min\{\Pr_{(x,y)\sim P}(y=1|x), \Pr_{(x,y)\sim P}(y=0|x)\}.
\end{aligned}$$

This calculation also tells us that the Bayes-optimal predictor h^* on a datapoint x is given by:

$$h^*(x) = \mathbb{1}\left(\Pr_{(x,y)\sim P}(y = 1 \mid x) \geq 1/2\right).$$

We are now ready to define agnostic PAC learning.

Definition 1 (Agnostic PAC Learnability)**.** *A hypothesis $\mathcal{H}$ is agnostic PAC learnable if for every $\epsilon, \delta \in (0,1)$, there exists a function $n_{\mathcal{H}}(\epsilon, \delta)$ and a learning algorithm such that for every distribution D over $\mathcal{X} \times \mathcal{Y}$, if the algorithm is run on $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ samples drawn i.i.d. from D, then the algorithm returns a hypothesis $\hat{h}$ with $R(\hat{h}) \leq \min_{h\in\mathcal{H}} R(h) + \epsilon$, except with probability δ.*

We remark that PAC learning is a special case of agnostic PAC learning where $\min_{h\in\mathcal{H}} R(h) = 0$ (due to realizability). Therefore agnostic PAC learnablity implies PAC learnablity.

As an example of agnostic PAC learnability, consider the hypothesis class

$$\mathcal{H} = \{h_w(x) : 1(w^T x > 0), w \in \mathbb{R}^d\} \quad \text{(the class of linear classifiers)}.$$

Learning $\mathcal{H}$ agnostically means that the algorithm works except with a failure probability δ in finding a linear classifier which is at most ϵ-suboptimal compared to the best linear classifier on the data.

2 Uniform Convergence

Even with the agnostic learning framework, it makes sense to find some ERM $\hat{h} \in \mathcal{H}$ on the training set. We want to show that $\hat{h}$ is close to the best it is possible to do with respect to population risk using the hypothesis class $\mathcal{H}$ ($\min_{h\in\mathcal{H}} R(h)$).

Uniform convergence is a powerful framework to show this. The key idea is that it suffices to show that **all** empirical risks of **all** members of $\mathcal{H}$ are close to their population risk.

Recall our definitions of population risk (1) and empirical risk (2). Let $h_{\text{ERM}} = \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$ (we were calling this $h_{S,\text{ERM}}$ last lecture but will drop the S in the subscript now for brevity) and $\widetilde{h} = \arg\min_{h\in\mathcal{H}} R(h)$. Note that $\widetilde{h}$ is not the same as the Bayes optimal predictor h^* since we do not assume realizability.

We can decompose the differeence between the population risk of h_{ERM} and $\widetilde{h}$ as follows:

$$R(h_{\text{ERM}}) - R(\widetilde{h}) = \underbrace{R(h_{\text{ERM}}) - \hat{R}_S(h_{\text{ERM}})}_{\text{trickier}} + \underbrace{\hat{R}_S(h_{\text{ERM}}) - \hat{R}_S(\widetilde{h})}_{\leq 0} + \underbrace{\hat{R}_S(\widetilde{h}) - R(\widetilde{h})}_{\text{easy to bound}}. \tag{3}$$

In (3), $\hat{R}_S(h_{\text{ERM}}) - \hat{R}_S(\widetilde{h}) \leq 0$ due to the definition of h_{ERM}.

Note that by definition,

$$\hat{R}_S(\widetilde{h}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\widetilde{h}(x_i), y_i)}_{\text{average of } n \text{ i.i.d. random variable}},$$

and $\forall\, i \in [n]$,

$$R(\widetilde{h}) = \mathbb{E}_{(x_i,y_i)\sim D}[\ell(\widetilde{h}(x_i), y_i)].$$

Therefore, by concentration inequalities which we will see soon, it is not difficult to get

$$\Pr[|\hat{R}_S(\widetilde{h}) - R(\widetilde{h})| \geq \epsilon] \leq \delta$$

for any fixed hypothesis $\tilde{h}$. Uniform convergence asks for such a deviation bound for every hypothesis in the hypothesis class.

Definition 2 (Uniform Convergence)**.** *A hypothesis class $\mathcal{H}$ has the* ***uniform convergence property*** *if for every $\epsilon, \delta \in (0,1)$, there exists a function $n_{\mathcal{H}}^{UC}(\epsilon, \delta)$ such that for every distribution D over $\mathcal{X} \times \mathcal{Y}$, if S is a training set of $n \geq n_{\mathcal{H}}^{UC}(\epsilon, \delta)$ samples drawn i.i.d. from D, then with probability $1-\delta$,*

$$\forall\, h \in \mathcal{H}, |\hat{R}_S(h) - R(h)| \leq \epsilon.$$

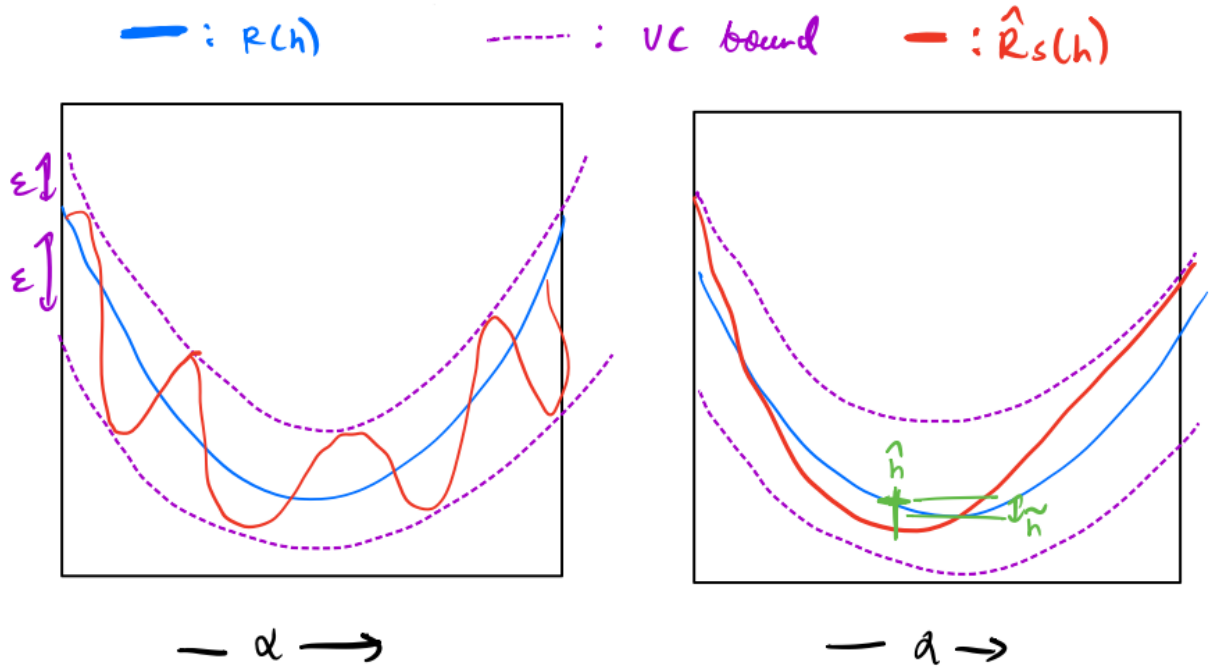

Figure 1: Consider the hypothesis class $\mathcal{H}$ is parameterized by single parameter $\alpha \in \mathbb{R}$. For e.g. $\mathcal{H}$ could be thresholds on the first coordindate $x(1)$ of the datapoint x, $\{\mathcal{H} : y = \mathbb{1}(x(1) > \alpha)\}$. Suppose $\mathcal{H}$ has the UC property. Then the deviation of the empirical risk from the true risk is bounded by ϵ. The figure on the right consider a stronger case: $\hat{R}(h)$ is a convex function with respect to the parameter α. In this case, there exists a unique local minimizer (and thus it is the global minimizer) so optimization method e.g. stochastic gradient descent algorithm can find the optimal solution. In contrast, in the figure on the left, there are multiple local minimizers. SGD can converge to anyone of them, which may lead to a suboptimal solution. Uniform convergence does not distinguish between these landscapes, but there is recent work on doing this [1, 2].

Proposition 3 (UC $\implies$ Agnostic PAC learning)**.** *If $\mathcal{H}$ has the UC property with $n_{\mathcal{H}}^{UC}(\epsilon, \delta)$, then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq n_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Moreover, ERM is an algorithm which achieves this sample complexity.*

Proof. Let S be a sample of size $n \geq n_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta)$. By definition,

$$\forall\, h \in \mathcal{H}, |\hat{R}_S(h) - R(h)| \leq \epsilon/2.$$

Consider ERM h_{ERM} and let $\widetilde{h} = \arg\min_{h \in \mathcal{H}} R(h)$.

$$R(h_{\mathrm{ERM}}) - R(\widetilde{h}) = \underbrace{R(h_{\mathrm{ERM}}) - \hat{R}_S(h_{\mathrm{ERM}})}_{\leq \epsilon/2} + \underbrace{\hat{R}_S(\hat{h}_{s,\mathrm{ERM}}) - \hat{R}_S(\widetilde{h})}_{\leq 0} + \underbrace{\hat{R}_S(\widetilde{h}) - R(\widetilde{h})}_{\leq \epsilon/2}.$$

■

Last lecture we show a PAC bound for finite hypothesis classes. Using uniform convergence we can show a agnostic PAC bound for finite hypothesis classes.

Theorem 4 (Agnostic PAC for finite classes)**.** *Let $\mathcal{H}$ be a class $|\mathcal{H}| < \infty$. Then $\mathcal{H}$ is agnostic-PAC learnable with*

$$n_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2\log(2|H|/\delta)}{\epsilon^2} \right\rceil.$$

Proof. We will show:

(1) For any fixed $h \in \mathcal{H}$ and $\epsilon > 0$,

$$\Pr\left[\left|\hat{R}_S(h) - R(h)\right| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

(2) For any $\epsilon > 0$,

$$\Pr[\forall\, h \in \mathcal{H}, \left|\hat{R}_S(h) - R(h)\right| \leq \epsilon] \geq 1 - 2|\mathcal{H}|e^{-2n\epsilon^2}.$$

(3) For $n \geq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$, with probability $1 - \delta$,

$$|\hat{R}_S(h) - R(h)| < \epsilon \quad \forall\, h \in \mathcal{H}.$$

(4) By UC, $\mathcal{H}$ is agnostic PAC learnable with

$$n \geq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

samples.

Proof of part (1)

We will use the following result, which we will prove later.

Lemma 5 (Hoeffdings inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq x_i \leq b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Given Hoeffding's, we prove part (1): take each $X_i = \ell(h(x_i), y_i)$. Since $\ell(h(x_i), y_i) = 1(h(x_i) \neq y_i)$, $X_i \in \{0, 1\}$, and thus $a_i = 0, b_i = 1$. Therefore, we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right| \geq \epsilon\right] \leq 2\exp(-2n\epsilon^2)$$

Proof of part (2)

$$\begin{aligned}\Pr[\forall\, h \in \mathcal{H}, |\hat{R}_S(h) - R(h)| \leq \epsilon] &= 1 - \Pr\left[\bigcup_{i=1}^{|\mathcal{H}|} |\hat{R}_S(h) - R(h)| > \epsilon\right] \\ &\geq 1 - \sum_{i=1}^{|\mathcal{H}|} \Pr[|\hat{R}_S(h_i) - R(h_i)| > \epsilon] \\ &\geq 1 - |\mathcal{H}|(2\exp(-2n\epsilon^2))\end{aligned}$$

Proof of part (3)

Set $n \geq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$ in the previous step, then we have the failure probability is bounded by δ for any $h \in \mathcal{H}$ i.e. $\Pr[\forall\, h \in \mathcal{H}, |\hat{R}_S(h) - R(h)| \leq \epsilon] \geq 1 - \delta$.

Proof of part (4)

In part (3), we have shown that $\mathcal{H}$ has the uniform convergence property, and thus by Proposition 3, we have $\mathcal{H}$ is agnostic PAC learnable with $n \geq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$ samples. ■

Therefore the proof of agnostic PAC learnability here boils down to Lemma 5 (Hoeffdings inequality). Results such as Hoeffdings inequality are known as *concentration inequalities* and lie at the heart of proving generalization bounds for ML algorithms. We will now spend some time understanding them.

3 Concentration inequalities

Let $\{X_1, X_2, \ldots\}$ be sequence of random variables $\mathbb{E}[X_i] = \mu, \mathbf{Var}(X_i) = \sigma^2 < \infty$. Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. How close is $\overline{X_n}$ to μ?

The Central Limit Theorem gives an asymptotic answer to this question.

Theorem 6 (Central Limit Theorem)**.** *As* $n \to \infty$

$$\left(\overline{X_n} - \mu\right) \xrightarrow{\textit{in distribution}} \mathcal{N}(0, \sigma^2/n).$$

In words, this means that $\overline{X_n}$ converges to its mean μ at a rate of $O\left(\frac{\sigma}{\sqrt{n}}\right)$. However, the central limit theorem is an asympototic result which applies when $n \to \infty$. We are interested in bounds which hold non-asymptotically for finite n. Concentration inequalities provide this, the simplest of which is Markov's inequality. Markov's inequality can be used whenever the random variable is non-negative and has a finite expectation.

3.1 Markov, Chebyshev and Chernoff

We will consider a series of bounds which analyze the 'tail' behavior of a random variable. Tail behavior refers to the proabability of a random variable deviating too far from its mean.

Proposition 7 (Markov's inequality)**.** *Let X be a non-negative random variable. Then for any $t > 0$, we have*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \Pr[X = i] \\
&\geq \sum_{i=t}^{\infty} i \Pr[X = i] \quad \text{(follows from non-negativity)} \\
&\geq t \cdot \sum_{i=1}^{t} \Pr[X = i] \\
\implies \Pr[X \geq t] &\leq \mathbb{E}[X]/t.
\end{aligned}$$

■

Without further assumptions, Markov's inequality is tight (*Exercise: Prove that this is the case by showing that for any $t > 0$ there exists some distribution such that* $\Pr[X \geq t] = \mathbb{E}[X]/t$.)

Markov's inequality only applies to non-negative random variables. But if we have a bound on the variance of the random variable, then we can work with the square of the random variables to still apply Markov's. Since we have a bound on the variance we can actually get a stronger tail bounds ($1/t^2$ instead of $1/t$ for Markov).

Proposition 8 (Chebyshev's inequality)**.** *Let X be a random variable. Then for any $t > 0$,*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Proof. Note that

$$\Pr[|X - \mathbb{E}[X]| \geq t] = \Pr[|X - \mathbb{E}[X]|^2 \geq t^2].$$

We now apply Markov's inequality to the non-negative random variable $(X - \mathbb{E}(X))^2$,

$$\Pr[(X - \mathbb{E}(X))^2 \geq t^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{t^2} = \frac{\mathbf{Var}[X]}{t^2}.$$

■

We can ask if this is already enough to get Hoeffding's inequality 5.

- Consider $a_i = 0, b_i = 0$, and $\mathbb{E}[X_i] = \mu, \ \ \forall\, i$.

- By definition, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and thus $\mathbb{E}[\overline{X}_n] = \mu$ (by linearity of expectations).
- Also, note that since the variance of the sum of indepdent random variables is the sum of their individual variances, $\mathbf{Var}[\overline{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{Var}[X_i] \leq \frac{1}{n}$ where the last step follows because $\mathbf{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \leq 1$.

Let us see what Chebyshev gives us,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq \frac{\mathbf{Var}[\overline{X}_n]}{\epsilon^2} \leq \frac{1}{n\epsilon^2}.$$

But Hoeffding's inequality says,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq 2\exp(-2n\epsilon^2),$$

which is saying that the tails have exponentially small probability, so Chebyshev clearly isn't strong enough. The crucial idea to get stronger concentration bounds is that there is no need to stop at the second moment. In fact for any positive integer k,

$$\Pr[|X - \mathbb{E}[X]| \geq t] = \Pr[|X - \mathbb{E}[X]|^k \geq t^k] \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^k)}{t^k}$$

We can even consider functions other than polynomials. For any $\lambda > 0$,

$$\begin{aligned}\Pr[X - \mathbb{E}(X) \geq t] &= \Pr\left[e^{\lambda(X-\mathbb{E}(X))} \geq e^{\lambda t}\right] \\ &\leq \frac{\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]}{e^{\lambda t}}.\end{aligned}$$

Since the inequality holds for any $\lambda > 0$, to get the best possible bound, we have

$$\Pr[X - \mathbb{E}(X) \geq t] \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]}{e^{\lambda t}}.$$

This brings us to **Chernoff-style bounds** which take the following form.

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0} \left(\log\left(\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]\right) - \lambda t\right) \tag{4}$$

As a remark, the quantity that appears inside the logarithm of the Chernoff bound above is known as the *moment-generating function*. Formally, for any random variable X, the moment generating function $M_X(t)$ is defined as $M_X(t) = \mathbb{E}(e^{tx})$.

Exercise: Verify that

$$\left.\frac{\partial^k M_X(t)}{\partial t^k}\right|_{t=0} = \mathbb{E}[X^k]$$

which is where the name moment-generating function comes from.

The moment generating function exists if all finite moments exist for the random varibale. By taking the exponentiation, Chernoff-bounds consdier all moments of X simultaneously which will allow us to prove *exponential* tail bounds.

3.2 Gaussian tail bounds

We will now use Chernoff-style bounds to derive concentration bounds for sums of random variables. We start with just deriving a tail bound for a Gaussian using (4).

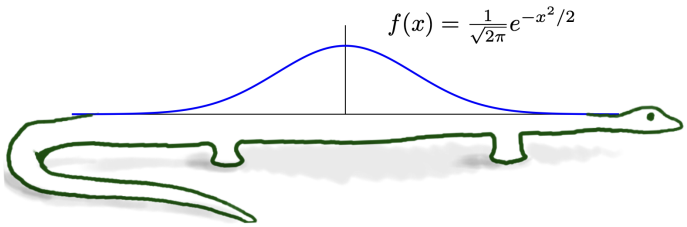

Figure 2: Why is it called a tail bound? Perhaps because the Gaussian density function looks a bit like a Brontosaurus, and we are interested in the probability mass in its 'tail'? :) Picture directly copied from Mary Wootters' beautifully written and illustrated Ph.D. thesis [3].

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$
- Goal: Bound $\Pr[X - \mu \geq t]$.

By using (4) and applying the exponential function on both sides,

$$\Pr[X - \mu \geq t] \leq \inf_{\lambda \geq 0} \frac{e^{(\lambda(X - \mathbb{E}[X]))}}{e^{\lambda t}}.$$

Let $X - \mathbb{E}[X] = y$. Note that $y \sim \mathcal{N}(0, \sigma^2)$. We can now derive the moment generating function of the Gaussian,

$$\begin{aligned}
\mathbb{E}(e^{\lambda(X - \mathbb{E}[X])}) &= \mathbb{E}(e^{\lambda y}) \\
&= \int_{\mathbb{R}} e^{\lambda y} \frac{e^{-y^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\
&= \int_{\mathbb{R}} \frac{e^{-(y - \lambda/\sigma^2)^2 \cdot \frac{1}{2\sigma^2}} e^{\lambda^2\sigma^2/2}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\
&= e^{\lambda^2\sigma^2/2} \int_{\mathbb{R}} \frac{e^{-(y - \lambda/\sigma^2) \cdot \frac{1}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\
&= e^{\frac{\lambda^2\sigma^2}{2}}.
\end{aligned}$$

Using this, we can write,

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0} \left(\frac{\lambda^2\sigma^2}{2} - \lambda t \right) = \frac{-t^2}{2\sigma^2} \quad \left\{ \begin{array}{l} \textit{Exercise: minimize the quadratic} \\ \textit{to show the equality.} \end{array} \right.$$

$$\implies \Pr[X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2}.$$

For the case of a Gaussian distribution, we could have derived a tail bound directly without using the moment generating function or the Chernoff-bound. But we didn't loose much, the bound above is in fact tight up to polynomial factors, i.e., the best bound one could hope to show is $\Pr[X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2} \cdot (\text{poly}(t, \sigma))^{-1}$.

In Hoeffding's inequality 5, the random variable of interest is a sum of random variables. To get tail bounds for many 'nice' random variables (which we will formally define later), one can pretent that the random variables are Gaussian and see what the tail bound are in that case. Since Gaussian have many nice properties, it is often much easier to work with Gaussian random variables. This gives a back-of-the-envelope calculation which can often be made rigorous, and we will see some of the machinery involved. But let us first see what bound we would get for sums of Gaussian random variables.

Let

$$\overline{X}_n = \sum_{i=1}^{n} X_i/n, \quad \mathbb{E}[X_i] = \mu, \quad \mathbf{Var}[X_i] = \sigma^2.$$

Since the Gaussian distribution is a stable distribution, the sum of Gaussians is also a Gaussian, therefore $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. Therefore using the tail bound we derived for a single Gaussian,

$$\Pr[\overline{X}_n - \mu \geq t] \leq e^{-nt^2/2\sigma^2}.$$

If we now want to set the failure probability to be δ, then the deviation expected from the true mean given n samples is about $\mathcal{O}\left(\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right)$:

$$\Pr\left[\overline{X}_n - \mu \geq \sigma\sqrt{\frac{2\log(1/\delta)}{n}}\right] \leq \delta.$$

We will next see that similar bounds can be obtained by sums of 'nice' non-Gaussian random variables as well.

3.3 Sub-Gaussian random variables

Notice that we didn't really use too many special properties of the Gaussian to get the previous tail bound. All we really need is a bound on the moment generating function, and the notion of *sub-Gaussian random variables* formalizes this.

Definition 9. *A random variable X with mean $\mu = \mathbb{E}(X)$ is **sub-Gaussian** if there exists a positive number σ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \quad \forall\, \lambda \in \mathbb{R}.$$

σ is known as the sub-Gaussian parameter (think of this as the standard deviation proxy for the distribution).

- Any Gaussian random variable with standard deviation σ is sub-Gaussian with parameter σ.
- Many non-Gaussian random variables also have this property! Let's see one example.

Rademacher random variable

Let X be the random variable which is $\{-1, +1\}$ with equal probability. This random variable is known as a *Rademacher random variable*.

Claim 10. *The Rademacher random variable is sub-Gaussian with parameter $\sigma = 1$.*

Proof. Note that $\mathbb{E}(X) = 0$. We can write,

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) \\
&= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) \\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \quad ((2k)! \geq 2^k k!) \\
&= e^{\lambda^2/2}.
\end{aligned}$$

■

3.4 Concentration bound for sub-Gaussian random variables

We now show that a sub-Gaussian random variable with sub-Gaussian parameter σ has similar tail upper bounds as a Gaussian random variable with standard deviation σ,

Lemma 11. *If X is a sub-Gaussian with parameter σ, then for any $t > 0$,*

(1) $\Pr[X > \mathbb{E}[X] + t] \leq e^{-t^2/2\sigma^2}$.

(2) $\Pr[X < \mathbb{E}[X] - t] \leq e^{-t^2/2\sigma^2}$.

(3) $\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-t^2/2\sigma^2}$.

Proof. (1) Using (4),

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0}(\log \mathbb{E}(e^{\lambda(X - \mathbb{E}(X))}) - \lambda t).$$

By sub-Gaussianity,

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\sigma^2\lambda^2/2}.$$

Therefore,

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0}\left(\frac{\sigma^2\lambda^2}{2} - \lambda t\right) = \frac{-t^2}{2\sigma^2}$$

and part (1) follows.

(2) Take $X' = -X$. Note that X' is also sub-Gaussian with parameter σ (by symmetry of the definition with respect to λ, i.e. the definition requires the inequality to be satisfied both for $\lambda \geq 0$ and $\lambda \leq 0$). We can now write,

$$\Pr[X < \mathbb{E}[X] - t] = \Pr[X' > \mathbb{E}[X'] + t].$$

The result now follows by part (1).

(3) Follows from a union bound.

■

3.5 Concentration bound for sum of sub-Gaussian random variables

As we did for Gaussians, we can now get a tail-bound for sums of independent sub-Gaussian random variables. The crucial property we will use here is that *sums of independent sub-Gaussian random variables are sub-Gaussian.*

Theorem 12. *Suppose that the random variable $\{X_i\}_{i=1}^n$ are independent and $\mathbb{E}[X_i] = \mu_i$ and X_i is sub-Gaussian with parameter σ_i. Then for all $t \geq 0$, we have*

$$\Pr\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right)$$

Proof. We show that $Z = \sum_{i=1}^{n} X_i$ is sub-Gaussian with parameter $\sigma_Z^2 = \sum_{i=1}^{n}\sigma_i^2$.

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda(Z-\mathbb{E}[Z])}\right) &= \mathbb{E}\left(e^{\lambda(\sum X_i - \mathbb{E}[\sum X_i])}\right) \\
&= \prod_{i=1}^{n}\mathbb{E}\left(e^{\lambda(X_i-\mathbb{E}[X_i])}\right) \quad \text{(this uses independence)} \\
&\leq \prod_{i=1}^{n} e^{\sigma_i^2\lambda^2/2} = e^{(\sum\sigma_i^2)\lambda^2/2}.
\end{aligned}$$

The Theorem now follows from Lemma 11. ■

We state the following direct Corollary of the above theorem for the case where the random variables X_i have the same mean and variance.

Corollary 13 (Corollary of Theorem 12 for identical random variables)**.** *Consider n i.i.d. random variable $X_1, \ldots, X_n$ each with $\mathbb{E}[X_i] = \mu$ and sub-Gaussian parameter σ. Then for all $t \geq 0$,*

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

Not coincidentally, this is exactly the same as the bound we got earlier for sums of independent Gaussian random variables.

3.6 Hoeffding's inequality

We now use the tools we have developed to show Hoeffding's inequality, which is a concentration bound for bounded random variables. We first show that bounded random variables are sub-Gaussian.

Claim 14 (Bounded random variable is sub-Gaussian)**.** *Let X be a random variable with mean zero and supported on the interval $[a, b]$. Then X is sub-Gaussian with sub-Gaussian parameter $(b-a)$ (in fact, it is possible to show sub-Gaussianity with parameter $(b-a)/2$, but we won't do that here).*

Proof. The proof uses the idea of *symmetrization*, which we will see in future lectures. Let X' be an independent copy of X. Note that since $\mathbb{E}[X'] = 0$, we can write,

$$\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_X[\exp(\lambda(X - \mathbb{E}[X']))] \leq \mathbb{E}_{X,X'}[\exp(\lambda(X - X'))]$$

where the last step uses Jensen's inequality, $f(\mathbb{E}[X]) \leq \mathbb{E}(f(X))$ for any convex f. Note that $X - X'$ and $X' - X$ have the same distribution. This implies

$$\mathbb{E}_{X,X'}[\exp(\lambda(X - X'))] = \mathbb{E}_{X,X'}[\exp(\lambda(X' - X))].$$

Therefore, we can introduce a Rademacher random variable ϵ, which $\{\pm 1\}$ with probability $1/2$ each, without changing the expectation,

$$\mathbb{E}_{X,X'}[\exp(\lambda(X - X'))] = \mathbb{E}_{X,X',\epsilon}[\exp(\lambda\epsilon(X - X'))].$$

Now fixing X, X' and just taking the expectation over ϵ, we can use the moment generating function bound for Rademacher random variables in Claim 10 to write,

$$\mathbb{E}_\epsilon[\exp(\lambda\epsilon(X - X'))] \leq \exp(\lambda^2(X - X')^2/2).$$

Plugging this back,

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}_{X,X'}[\exp(\lambda^2(X - X')^2/2)] \leq \exp(\lambda^2(b-a)^2/2)$$

where last inequality follows since X, X' lie in the interval $[a, b]$. ■

Using Theorem 12 we now get a slightly weaker version of Hoeffding's inequality.

Lemma 15 (weaker version of Hoeffding's)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq x_i \leq b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \leq \epsilon\right] \geq 1 - 2\exp\left(\frac{-n^2\epsilon^2}{2\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Remark: This bound is only loose compared to Hoeffding's inequality stated earlier in terms of a factor of 2 in the denominator of the exponent, which should instead be in the numerator. If we use the right sub-Gaussian parameter $(b_i - a_i)/2$ for the random variables, we will recover Hoeffding's inequality as stated earlier.

So far, we've only worked with sums of random variables. There is a very rich and vast literature on showing that various other functions of random variables are also close to their expectation with high probability. A statement of this form is McDiarmid's inequality, which we will use in a few lectures.

Theorem 16 (McDiarmid's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values over some domain $\mathcal{X}$. A function $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the* ***bounded differences property*** *if:*

$$\forall\, i \in [n], \quad \forall\, x_1, \ldots, x_n, x_i' \in \mathcal{X},$$
$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

If f satisfies this property, then

$$\Pr(f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)] > \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Remark: Note that this gives us back Hoeffding's inequality by taking $f = \dfrac{\sum X_i}{n}$, and realizing that f satisfies the bounded differences property with $c_i = \dfrac{b_i - a_i}{n} \quad \forall\, i$.

4 Vapnik-Chervonenkis (VC) Dimension

So far, we have shown that for hypothesis classes $\mathcal{H}$ with finite size $|\mathcal{H}|$,

- $\mathcal{H}$ is PAC learnable with $n_{\mathcal{H}}(\epsilon, \delta) = \mathcal{O}\left(\dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon}\right)$ samples (with the realizability assumption).
- $\mathcal{H}$ is agnostic-PAC learnable with $n_{\mathcal{H}}(\epsilon, \delta) = \mathcal{O}\left(\dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ samples.

What can we say when$\mathcal{H}$ is infinite? (This is the case for many common hypothesis classes, such as linear classifiers.)

- One quick hack is to do discretization: Think about a linear classifier in $\mathbb{R}^d$. For a 32-bit system $\implies (2^{32})^d$ possible classifiers which is large but still finite. Moreover, since our finite hypothesis classes bound only depends logarithmically on the cardinality of the hypothesis class, the number of samples needed is $\approx \log(|\mathcal{H}|) = O(d)$, which does not seem too bad.
- Discretization can be good for back-of-the-envelope calculations, but is a bit limited since it only counts the number of functions within the class and not their complexity. For example, what if all the different hypothesis in the hypothesis class actually make the same prediction? In this case, the number of samples required for learning should not depend on the number of hypotheses in the class. **VC dimension** gives a more complete solution than discretization and can handle ∞-sized classes.

5 Further reading

You can read more about Agnostic PAC learning in Section 3.2 of [4]. You can read Chapter 4 for uniform convergence and agnostic PAC learnability for finite hypothesis classes.

There are many good references for concentration inequalities, a very good reference one is Martin Wainwright's notes https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf. We covered the first few pages of this (up to Proposition 2.1) in class, but I'd encourage you to read the rest as well if interested, it's very nice and useful material. Everything we talked about in class was for independent random variables, but we often have to deal with dependent random variables too. Some of the bounds we showed can be obtained without independence and rhere is a lot of interesting work on this, for example see [5, 6, 7]. If you're interested in bounds for dependent random variable, you can also check out Central Limit Theorems for Martingales (https://en.wikipedia.org/wiki/Martingale_central_limit_theorem) and Markov chains (https://en.wikipedia.org/wiki/Markov_chain_central_limit_theorem). If you're interested in the proof of McDiarmid's inequality you can have at Martin Wainwright's notes linked above or Madhur Tulsiani's notes which also have some cool applications: https://home.ttic.edu/~madhurt/courses/toolkit2015/l15.pdf Someone asked about Lipschitz functions in the context of McDiarmid's in class, you might also want to have a look at Theorem 2.4 in Section 2.3 of Wainwright's notes, since it shows concentration for any Lipschitz function.

References

[1] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[2] Tengyu Ma. Why do local methods solve nonconvex problems?, 2020.

[3] Mary Katherine Wootters. *Any Errors in this Dissertation are Probably Fixable: Topics in Probability and Error Correcting Codes*. PhD thesis, 2014.

[4] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[5] Christos Pelekis and Jan Ramon. Hoeffdings inequality for sums of weakly dependent random variables, 2015.

[6] Leonid Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. 2008.

[7] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. 2015.