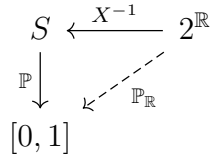
Physics of Learning Theory
Lecture 1
Probability and Learning Theory review

January 29, 2025
Nilava Metya

1 Introduction

We will recall some probability theory and look at useful *deviation* or *concentration* bounds which are frequently used in analyzing algorithms (in learning theory). Recall that a (real-valued) random variable on a probability space $(\Omega, S, \mathbb{P})$ is nothing but a 'measurable function' $X : \Omega \to \mathbb{R}$. Here Ω is the universal or sample space where we think of events in, S is a collection of events in Ω and $\mathbb{P} : S \to [0, 1]$ assigns probability to each event in S. The space of events S is constrained to satisfy some obvious rules like Ω is an event, if A is an event then so is $\Omega \setminus A$ and that a countable union of events is an event which makes it sensible to work with the concept of assigning probabilities to each event. We will often say that $\mathbb{P}[A]$ is the probability that event A occurs. If $A = \{a\}$ is a singleton, we always write $\mathbb{P}[a]$ instead of $\mathbb{P}[\{a\}]$. The probability function $\mathbb{P}$ is also constrained to a couple of rules, namely, that the probability of the union of a mutually disjoint collection of events, which is an event, is the same as the sum of the probabilities of each of those events and that the probability that Ω occurs is 1. Roughly a random variable is to be thought of as a way of assigning points of the sample space to real numbers which are *really real* and are more tangible to work with, while respecting the rules of S. Such a random variable induces a map $X^{-1} : 2^{\mathbb{R}} \to S$ by $X^{-1}(A) \coloneqq \{x \in \Omega \mid X(x) \in S\}$ for any $A \subseteq \mathbb{R}$, and hence induces a probability on $\mathbb{R}$ given by $\mathbb{P}_{\mathbb{R}}[A] = \mathbb{P}[X^{-1}(A)]$ where A is any 'measurable' subset of $\mathbb{R}$. The random variable being a 'measurable function' precisely means that $X^{-1}(A)$ always lies in S.



1.1 Mean

The average or mean of a random variable X, often denoted as $\mathbb{E}[X]$, $\mu(X)$, or simply μ when the context is clear, is $\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\,\mathbb{P}$. For the discrete case, which we will mostly be interested in, this boils down to $\mathbb{E}[X] = \sum_{i\in\Omega} X(i)\mathbb{P}[i]$. Note that if X is an indicator random variable for event A, that is, $X = 1$ if A occurs and 0 otherwise, then $\mathbb{E}[X] = \mathbb{P}[A]$.

Example 1. Consider tossing a fair coin. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The probability function is $\mathbb{P}[\varnothing] = 0, \mathbb{P}[\mathrm{H}] = \mathbb{P}[\mathrm{T}] = 0.5, \mathbb{P}[\{\mathrm{H}, \mathrm{T}\}] = 1$. A natural random variable to consider is $X(i) = \mathbf{1}_{\mathrm{H}} \coloneqq \begin{cases} 1 & \text{if } i = \mathrm{H} \\ 0 & \text{if } i = \mathrm{T} \end{cases}$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[A] = \begin{cases} 0 & \text{if } 0 \notin A, 1 \notin A \\ 0.5 & \text{if } 0 \in A, 1 \notin A \\ 0.5 & \text{if } 0 \notin A, 1 \in A \\ 1 & \text{if } 0 \in A, 1 \in A \end{cases}$. In this case, $\mathbb{E}[X] = 1 \cdot \mathbb{P}[\mathrm{H}] + 0 \cdot \mathbb{P}[\mathrm{T}] = 0.5$

Example 2. Consider tossing n fair coins sequentially and independently. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}^n$. So the singleton outcomes are tuples of H, T. The probability function is given by $\mathbb{P}[\boldsymbol{x}] = 2^{-n}$ for any element $x \in \Omega$ and then extending by countable additivity of $\mathbb{P}$. Consider n random variables $X_1, \cdots, X_n$ where $X_i(\boldsymbol{x}) \coloneqq \begin{cases} 1 & \text{if } x_i = \mathrm{H} \\ 0 & \text{if } x_i = \mathrm{T} \end{cases}$. Each X_i is the same random variable as the previous example after looking at the i^{th} coordinate. A natural variable to consider is the total number of heads obtained in one round of tossing, that is $X = X_1 + \cdots + X_n$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[k] = \begin{cases} \binom{n}{k} 2^{-n} & \text{if } k \in \{0, \cdots, n\} \\ 0 & \text{otherwise} \end{cases}$ and extend by countable additivity. Here $\mathbb{E}[X] = \frac{n}{2}$.

One useful result used for calculating expectations of sums of random variables is that if $a, b \in \mathbb{R}$ and X, Y are random variables then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. It's worthy to note that sums and scalings of random variables are random variables. This result does **not** depend on 'independence' of X, Y. Independence plays an important role for the average of products of random variables (which is a random variable). We say random variables $X_1, \cdots, X_n$ are (*mutually*) *independent* if $\mathbb{P}[\bigcap_{i=1}^n \{X_i \le a_i\}] = \prod_{i=1}^n \mathbb{P}[X_i \le a_i] \;\forall\; a_i \in \mathbb{R}$. This is a stronger notion than *pairwise* independence where we demand that only every pair of them are independent. Note that mutual independence implies pairwise independence. If X, Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

For a random variable $X \ge 0$ and $a \in \mathbb{R}$ let Y be the indicator random variable indicating whether $X \ge a$, that is, Y is 1 if $X \ge a$ and 0 otherwise. Then clearly $X \ge aY$. Indeed if $X \ge a$ then $Y = 1$ so $X \ge aY$ and if $X < a$ then $Y = 0$ so that $X \ge 0 = aY$. Expectation preserves inequalities, so $\mathbb{E}[X] \ge a\mathbb{E}[Y] = a\mathbb{P}[X \ge a]$. This establishes

Theorem 1 (Markov's inequality)
If X is a non-negative random variable and $a \in \mathbb{R}$ then $\mathbb{P}[X \geq a] \leq \dfrac{\mathbb{E}[X]}{a}$.

1.2 Variance

Let's come to deviation now. One natural way to measure *deviation* is to look how on average much a random variable deviates either way from its mean (behavior). To look for deviation in either direction of $\mathbb{E}[X]$ we consider the random variable $(X - \mathbb{E}[X])^2$. Define the variance of a random variable X as $\text{Var}[X] \coloneqq \mathbb{E}[(X - \mathbb{E}[X])^2]$. One useful result to compute variance is that if X, Y are independent then $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y]$. This extends to n pairwise independent random variables. Another useful result is $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Applying Theorem 1 to $(X - \mathbb{E}[X])^2 \geq 0$ gives

Theorem 2 (Chebyshev's inequality)
If X is a random variable and $a \in \mathbb{R}_{\geq 0}$ then $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \dfrac{\text{Var}[X]}{a^2}$.

1.3 Higher moments

One might just ask why stop at the second power to measure deviation. What about the random variable $(X - \mathbb{E}[X])^k$ for $k \geq 2$? These are called higher centeral moments. Note that $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] = 0$ when k is odd and the distribution of X is symmetric about $\mathbb{E}[X]$. So it makes sense to consider the random variables $X_k \coloneqq |X - \mathbb{E}[X]|^k$ instead. If we have access to such numbers, we can use the same trick as the proof of Chebyshev's inequality and get $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \dfrac{\mu_k}{a^k}$. Knowing all higher moments means that we know something known as the 'characteristic function' (not yet defined) of X which uniquely determines X. But our aim was the study deviations using small information. Generally, higher moments are not known.

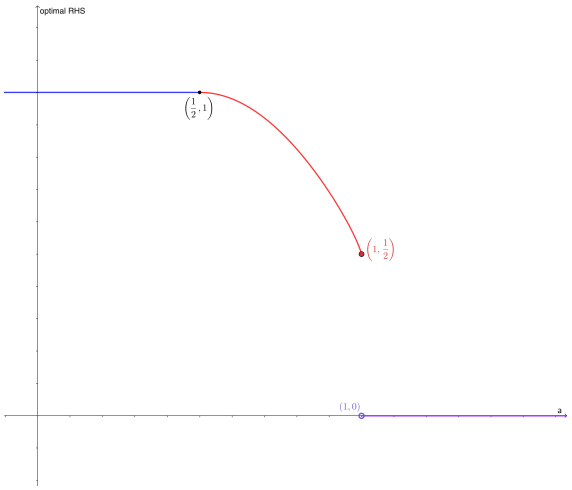
Here's a small trick to optimally apply Markov to a non-homogeneous function. Let's just take the 'best polynomial' ever known. It's non-homogeneous, positive, monotonic (but not *monotonous*) and has values at all points. We want to study the concentration of $e^{X-\mu}$. Take $f(x) = e^x \geq 0$. Chebyshev's inequality do this for $f = x^2$ but this was homegeneous so scaling the random variables had no effect on the inequalities obtained. Consider the random variable $Y_t = f(t(X - \mu))$ where t is a real variable. Then applying Markov on $|Y_t| = Y_t$ gives $\mathbb{P}\left[Y_t = e^{t(X-\mu)} \geq e^{ta}\right] \leq \frac{\mathbb{E}[Y_t]}{e^{ta}} \ \forall \ t \geq 0, a \in \mathbb{R}$. This is equivalent to $\mathbb{P}[X - \mu \geq a] \leq \frac{\mathbb{E}\left[e^{t(X-\mu)}\right]}{e^{ta}}$. Since this is true for every $t \geq 0$, we conclude that $\mathbb{P}[X \geq a + \mu] \leq \inf\limits_{t \geq 0} \dfrac{\mathbb{E}\left[e^{t(X-\mu)}\right]}{e^{ta}}$. One issue with this argument is that $M_X(t) \coloneqq \mathbb{E}[\exp(tX)]$ may not always exist. Let's say they exist for $t \in [0, b]$ for some $b \geq 0$ (sanity check: $b = 0$

always works). Then we can modify our inequality to $\mathbb{P}[X \geq a+\mu] \leq \inf_{t\in[0,b]} \frac{\mathbb{E}\left[e^{tX}\right]}{e^{t(a+\mu)}}$. The *moment generating function* (mgf, in short) of a random variable X is $M_X(t) = \mathbb{E}[\exp(tX)]$.

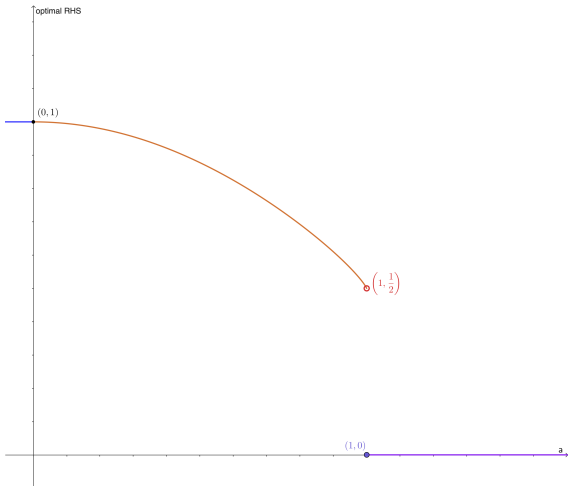
Example 3 (Bernouli). Say X takes values $0, 1$ with probability $\frac{1}{2}$ each. Then $M_X(t) = \mathbb{E}[\exp(tX)] = \frac{1}{2}\exp t + \frac{1}{2}$ always exists. Our above inequality takes the form $\mathbb{P}[X \geq a] \leq \frac{1}{2}\inf_{t\geq 0} \frac{\exp t + 1}{e^{ta}}$. If $a \leq \frac{1}{2}$ then the RHS is 1 at $t = 0$. If $\frac{1}{2} < a < 1$ then the RHS is $\frac{1}{2(1-a)^{1-a}a^a}$ at $t = \ln(a) - \ln(1-a)$. Taking $a \to 1^-$ gives that if $a = 1$, the RHS is $\frac{1}{2}$ attained at $''t = +\infty''$ (can also be checked directly by plugging in $a = 1$ directly). If $a > 1$ the RHS is 0 again at $''t = +\infty''$.

Example 4 (Rademacher). Say X takes values ± 1 with probability $\frac{1}{2}$ each. Such a random variable is called a *Rademacher random variable*. Then $M_X(t) = \mathbb{E}[\exp(tX)] = \frac{1}{2}(e^t + e^{-t})$ always exists. Our above inequality takes the form $\mathbb{P}[X \geq a] \leq \frac{1}{2}\inf_{t\geq 0} \frac{e^t + e^{-t}}{e^{ta}}$. The RHS looks like

$$\begin{cases} 1 & \text{at } t^* = 0 \text{ if } a \leq 0 \\ \sqrt{\frac{1}{(1-a)^{1-a}(1+a)^{1+a}}} & \text{at } t^* = \frac{1}{2}\ln\left(\frac{1+a}{1-a}\right) \text{ if } a \in (0,1]. \\ 0 & \text{at } t^* = \infty \text{ if } a > 1 \end{cases}$$



h



1.4 Sub-Gaussian random variables

Now let's apply it to our favorite distribution – the Gaussian. Recall that the Gaussian distribution Z with mean μ and variance σ^2 has the density $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$. The moment generating function of this Gaussian is $M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$ and exists $\forall\, t \in \mathbb{R}$. Substituting this into our 'moment-based Markov inequality' gives $\mathbb{P}[Z \geq \mu + a] \leq \inf_{t\geq 0} \exp\left(\frac{\sigma^2 t^2}{2} - at\right) = \exp\left(-\frac{a^2}{2\sigma^2}\right)$. This means $\mathbb{P}[|Z-\mu| \geq a] \leq 2\exp\left(-\frac{a^2}{2\sigma^2}\right)$ for any $a \geq 0$.

This calculation let's us study the deviation of a large class of random variables, if this class is defined properly. If we revisit the calculation done for the Gaussian, we see that the only necessary property of any random variable X that can get the same bound is the existence of some σ^2 such that we can get a similar function as an upper bound on the mgf of X. More precisely, we demand that there exist a real number $\sigma > 0$ such that $\mathbb{E}\left[\exp(t(X - \mathbb{E}[X]))\right] \leq \exp\left(\frac{t^2\sigma^2}{2}\right) \ \forall\ t \in \mathbb{R}$. Alternately, instead of using the proof and calculation details, one might suggest to study those class of random variables whose deviations are bounded by those of the Gaussian. They turn out to be the same.

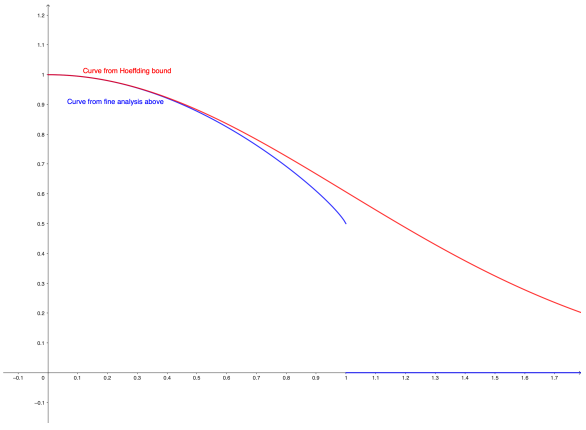
Definition 3
A random variable X with mean μ is said to be *sub-Gaussian* if there is a constant $c > 0$ and a Gaussian $Z \sim \mathcal{N}(0, \tau^2)$ such that $\mathbb{P}\left[|X - \mu| \geq a\right] \leq c\,\mathbb{P}\left[|Z| \geq a\right]\ \forall\ a \geq 0$.

Alternately, a random variable X with mean μ is said to be *sub-Gaussian* if there exists $\sigma > 0$ such that $\mathbb{E}\left[\exp(t(X - \mathbb{E}[X]))\right] \leq \exp\left(\frac{t^2\sigma^2}{2}\right) \ \forall\ t \in \mathbb{R}$. This σ^2 is said to be the sub-Gaussian parameter and acts as a proxy for variance.

This is quite a nice class because sub-Gaussianity is preserved under linear combinations. In particular, if X_1, X_2 are independent sub-Gaussians with parameters σ_1^2, σ_2^2 respectively then $X_1 + X_2$ is also a sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$. In other words, the variance proxies add up just like the Gaussian. Using this property we immediately get

Theorem 4 (Hoeffding)
If $\{X_i\}_{i=1}^m$ are independent sub-Gaussians with means $\{\mu_i\}_{i=1}^m$ and variance proxies $\{\sigma^2\}_{i=1}^m$ respectively. Then $\mathbb{P}\left[\sum_{i=1}^m (X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_i \sigma_i^2}\right\}$ for all $t \geq 0$.

At our current discussion stage, sub-Gaussians seem quite useless. But, a lot of the 'good' random variables we see are actually sub-Gaussian. In fact if X is a bounded random variable taking values in $[a, b]$ then X is sub-Gaussian with parameter $\left(\frac{b-a}{2}\right)^2$. Here's a comparison of the bounds obtained with fine analysis as in Example 4 vs what the Hoeffding bound gives us. Notice the smoothness difference.



Corollary 5
Let $X_1, \cdots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ (almost surely) and sample mean $\overline{X}$. Then $\mathbb{P}\left[\overline{X} - \mathbb{E}\left[\overline{X}\right] \geq t\right] \leq \exp\left\{-\frac{2n^2t^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $t \geq 0$.

2 Supervised Learning

The setup. In supervised learning, we are to design a *learner* that learns the 'labels' of certain 'objects' or 'data' and then we can use it to *predict* unlabelled objects. An example could be a coin-sorting machine that understands (with human help) the sizes of various coins (data) and what size associates with what denomination (labels), and then when this model is released as a commercial product, the machine can speed up the process of sorting coins into different labelled stacks.
Formally, the data or inputs belong to some space $\mathcal{X}$, and the labels are in some space $\mathcal{Y}$. For the above example, $\mathcal{X}$ is the set of all coins and $\mathcal{Y}$ contains the string of labels of these coins like 'dime', 'nickel', 'penny' and so on. We are interested in a certain joint probability distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. A *training set* is a finite (multi-)set of elements of $\mathcal{X} \times \mathcal{Y}$ chosen independently and identically according to the distribution $\mathbb{P}$. We always denote this training set as $\{z_i = (x_i, y_i)\}_{i=1}^n$. Our goal is to design a *model* $h : \mathcal{X} \to \mathcal{Y}$, based on this training data, which has certain properties according to our needs. Such an h is oftentimes also referred to as a *hypothesis* or a *predictor*. Note that a model can be **any** function $\mathcal{X} \to \mathcal{Y}$.

Loss function. How do we quantify the predictors which satisfy our needs. More precisely, when is a model better than another? For this, we have something called a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ which is to be thought of as penalizing a predicted label against the actual label. For example, the loss suffered by a model h on a data point x with label y is $\ell(h(x), y)$ because $h(x)$ is the predicted label whereas y is the actual label. Such a loss function is assumed to be non-negative. A 'best' model is one which suffers the least expected loss. The expected loss of a model h is $L(h) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}} [\ell(h(x), y)]$, also called the *population risk*. We want to find $\inf_h L(h)$.

Hypothesis class. One question one might wonder is that what is the $\arg\min$ being taken over. In practice, we do not have a way of optimizing over arbitrary functions. We instead want to focus on a more specific subclass of functions which either make more sense in the context we are working on or are easier to work with. Such a constrained set of functions $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is called a *hypothesis class*. Now we can clearly state a goal that we want to find $\inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}} [\ell(h(x), y)]$ (and find the minimizer if feasible or approximate it). This completes the formal setup of a supervised learning problem. This is impossible in general because we will not have access to $\mathbb{P}$ on the entire $\mathcal{X} \times \mathcal{Y}$ but only to a finite sample. So we aim to design some $h \in \mathcal{H}$ with minimum possible *empirical loss*. In practice, we need to make assumptions and lots of restrictions on $\mathcal{H}, \mathbb{P}, \ell$ to get 'good' results (whatever that means).

Examples.

Example 5 (Binary classification). In this case we want to classify objects in $\mathcal{X}$ into two categories, so the label space is $\mathcal{Y} = \{\square, \times\}$. The usual penalization is given by the function $\ell(\square, \times) = \ell(\times, \square) = 1, \ell(\square, \square) = \ell(\times, \times) = 0$. There is the classical problem of support vector machines. We describe a very simple but related problem. If $\mathcal{X} \subseteq \mathbb{R}^n$, take $\mathcal{H} = \{\operatorname{sgn}(\langle \boldsymbol{a}, \cdot \rangle - b) \mid \boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}\}$ where $\operatorname{sgn}(x) = \begin{cases} \square & \text{if } x \geq 0 \\ \times & \text{otherwise} \end{cases}$.

Example 6 (Regression). In the regression problem, we would like to predict continuous outputs $\mathcal{Y} = \mathbb{R}$ from a continuous input space $\mathcal{X} = \mathbb{R}^n$. A popular loss function used in this case is $\ell(y', y) = (y' - y)^2$. Other possible loss functions are $\ell(y', y) = |y' - y|^p$ for any $p \geq 0$ but $p = 2$ is used in practice due to smoothness, convexity and low integer power. The hypothesis class depends on what kind of functions one thinks are fit for the model, again a choice to be made. Let's just focus on $\mathcal{H}_d = \mathbb{R}[x_1, \cdots, x_n]_d$ as the real polynomials in n variables of degree at most d. If $d = 1$, we call it *linear regression*. If $d = 2$ we call it quadratic regression. And so on.

Empirical risk minimization. Let's recall that our goal was to minimize the population risk, namely, $\inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathbb{P}} [\ell(h(x), y)]$. In practice we do not have access to the entire population; we only have a training set of n data points, drawn independently from the same distribution as the entire population. To achieve our main goal, we can instead focus on the *empirical risk* or *sample risk* $\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$. *Empirical risk minimization*, or ERM in short, refers to finding $\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$. It is an unbiased estimator of the population risk. In other words, $\mathbb{E}_{z_i \overset{\text{iid}}{\sim} \mathbb{P}} \left[\hat{L}(h) \right] = L(h)$. The hope with ERM is that minimizing the empirical error will lead to small population error. So we are interested in the excess risk $L\left(\hat{h}\right) - \inf_{h \in \mathcal{H}} L(h)$. In other words, we are *generalizing* the empirical risk minimizer to the population risk minimizer. One way to make this rigorous is by showing that the ERM minimizer's excess risk is bounded. If n is quite large, it makes sense to hope this intuitively due to the law of large numbers.

2.1 Non-asymptotic analysis

We do want non-asymptotic results when we have limited number of data points (that is, n is relatively small). The LLN roughly states that the empirical average of a large number iid data behave as expected. In order to do the same for smaller-ish n, we study the concentration around the mean and hence we want to use concentration inequalities. Fortunately a lot of the distributions we deal with are, in real life, sub-Gaussian (or Lipschitz mappings of sub-Gaussians). But what we lose is that we can no longer make statements which are guaranteed to be true, but only bounds which hold 'with high probability'. There is no clear definition of this term in literature but often used to refer to probabilities which are $\geq 1 - \frac{1}{\text{poly}(n)}$.

Let me take a small detour and introduce a trick I learnt in my CS courses. Let $X_{j,1}, \cdots, X_{j,n} \in [0,1]$ be independent for $k \in [K]$. Think of j as the index of the person performing a repeated task. Denote their sample means by $Y_j \coloneqq \frac{1}{n}\sum_i X_{ji}$ and $\mu_j \coloneqq \mathbb{E}[Y_j]$. Then $\mathbb{P}\left[\underbrace{|Y_j - \mu_j| \geq t}_{E_j \coloneqq}\right] \leq 2\exp\{-2nt^2\}\ \forall\, t \geq 0, j \in [K]$ by Hoeffding. If I want to find out the chance that even one of these random variables has large deviation from mean, I would consider the event $E \coloneqq \bigcup\limits_{j\in[K]} \{|Y_j - \mu_j| \geq t\} = \bigcup_j E_j$. Let's find the probability of this 'bad' event. $\mathbb{P}[E] \leq \sum_j \mathbb{P}[E_j] \leq 2\sum_j \exp(-2nt^2) = 2K\exp(-2nt^2)$. $t = \sqrt{\frac{\ln(2K/\delta)}{2n}}$ gives $\mathbb{P}\left[\exists j \in [K] \text{ s.t. } |Y_j - \mu_j| \geq \sqrt{\frac{\ln(2K/\delta)}{2n}}\right] \leq \delta$. Taking complements,

$$\mathbb{P}\left[|Y_j - \mu_j| < \sqrt{\frac{\ln(2K/\delta)}{2n}}\ \forall\, j \in [K]\right] \geq 1 - \delta.$$

In other words, we can make statements of the form "with high probability, each person remains close to their expected behavior on average." Alternately taking $n = \frac{\ln(2K/\delta)}{2\varepsilon^2}$, $t = \varepsilon$ gives $\mathbb{P}[|Y_j - \mu_j| < \varepsilon\ \forall\, j \in [K]] \geq 1 - \delta$. In other words, if every person performs $\frac{\ln(2K/\delta)}{2\varepsilon^2}$ experiments, each of their average behavior is expected to be within ε distance of the expected behavior with probability $1 - \delta$.

Let's now consider this in the context of learning theory where the people are replaced with models $h \in \mathcal{H}$. Recall that our main goal was to reach that the minimizer of ERM approximately minimizes the actual loss, that is, the excess risk $L\left(\hat{h}\right) - \min\limits_{h\in\mathcal{H}} L(h)$ is quite small. If we can say with high certainty that every predictor is penalized almost as much on the population as the empirical data, we can conclude with high probability that the ERM minimizer also approximately minimizes the population risk. This is seen as follows.

2.1.1 Finite hypothesis class

Proposition 6
If every model $h \in \mathcal{H}$ has almost the same penalization on the population as the sample, that is $\left|L(h) - \hat{L}(h)\right| \leq \frac{\varepsilon}{2}$, then an ERM $\hat{h} \in \arg\min_{h\in\mathcal{H}} \hat{L}(h)$ minimizes L upto ε accuracy.

Proof. Denote $h^* := \arg\min_{h\in\mathcal{H}} L(h)$. We want an upper bound on $L\left(\hat{h}\right) - L\left(h^*\right)$. Let's write it a little differently. $L\left(\hat{h}\right) - L\left(h^*\right) = L\left(\hat{h}\right) - \hat{L}\left(\hat{h}\right) + \hat{L}\left(\hat{h}\right) - \hat{L}(h^*) + \hat{L}(h^*) - L(h^*)$. Note $\hat{L}\left(\hat{h}\right) - \hat{L}(h^*) \leq 0$. So $L\left(\hat{h}\right) - L\left(h^*\right) \leq \left|L\left(\hat{h}\right) - \hat{L}\left(\hat{h}\right)\right| + \left|\hat{L}(h^*) - L(h^*)\right|$. Using the hypothesis gives $L\left(\hat{h}\right) - L\left(h^*\right) \leq 2\sup_{h\in\mathcal{H}}\left|L(h) - \hat{L}(h)\right| \leq \varepsilon$. ■

In simpler terms, a uniform upper bound on $\left|L - \hat{L}\right|$ implies generalization of the ERM to the population risk minimizer. Note that if we did not know a uniform upper bound on $\left|L - \hat{L}\right|$, we could have still bounded $\left|\hat{L}(h^*) - L(h^*)\right|$ via Hoeffding bound (with high probability). However, $\left|L\left(\hat{h}\right) - \hat{L}\left(\hat{h}\right)\right|$ is data dependent (due to the data dependency of $\hat{h}$). It is quite possible that this term is quite big. In fact it's often practically encountered if $\mathcal{H}$ is not chosen carefully – even with small training error, there can be large testing error.

Corollary 7
For a finite hypothesis class $\mathcal{H}$, a loss function $\ell \in [0,1]$ with n training data points and $\delta \in (0, 0.5)$, we have $\mathbb{P}\left[\left|L(h) - \hat{L}(h)\right| < \sqrt{\frac{1}{2n}\ln\left(\frac{2\left|\mathcal{H}\right|}{\delta}\right)}\ \forall\ h \in \mathcal{H}\right] \geq 1 - \delta$.
Consequently, $\mathbb{P}\left[\left|L\left(\hat{h}\right) - L(h^*)\right| < \sqrt{\frac{2}{n}\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}\right] \geq 1 - \delta$.

Proof. The first part is proven the same way as the trick discussed in the previous page with people being replaced by models h, $K = |\mathcal{H}|$ and the random variables being the evaluation of ℓ on the training data. The second part is immediate by Proposition 6. ■

Corollary 8
For a finite hypothesis class $\mathcal{H}$, a loss function $\ell \in [0,1]$, $\delta \in (0, 0.5)$, and (additive) error tolerance $\varepsilon > 0$, it is enough to have $n = \mathcal{O}\left(\frac{2}{\varepsilon^2}\ln\left(\frac{2\left|\mathcal{H}\right|}{\delta}\right)\right)$ training data points to achieve ε-generalization of ERM to population risk minimum with probability $1 - \delta$.

Corollary 9
For a finite hypothesis class $\mathcal{H}$, a loss function $\ell \in [0,1]$, (additive) error tolerance $\varepsilon > 0$ and n samples, $\left|L(h) - \hat{L}(h)\right| < \varepsilon\ \forall\ h \in \mathcal{H}$ with probability $\geq 1 - 2\left|\mathcal{H}\right|\exp(-2n\varepsilon^2)$.

2.1.2 Infinite hypothesis class

The above analysis relied heavily on the size of $\mathcal{H}$. This cannot be done when $\mathcal{H}$ is infinite, which it usually is. Unless we assume some structure on $\mathcal{H}$, it's quite difficult to do the analysis for infinite $\mathcal{H}$. So we will assume that $\mathcal{H}$ is bounded with some bounded parameters, usually taken to be vectors in $\mathbb{R}^p$. That is we will have $B > 0$ such that $\mathcal{H} = \{h_\theta \mid \theta \in \mathbb{R}^p, \|\theta\|_2 \leq B\}$. $\Theta \coloneqq \{\theta \in \mathbb{R}^p, \|\theta\|_2 \leq B\}$ is the parameter space for θ's. The technique to be used here is called *brute-forced discretization*. Here's the main idea.

Let's abuse notation and write $L(\theta), \hat{L}(\theta)$ for $L(h_\theta), \hat{L}(h_\theta)$ respectively. As before, we name the 'bad' events $E_\theta \coloneqq \left\{L(\theta) - \hat{L}(\theta) \geq \varepsilon\right\}$. If we want to use the previous technique, we end up in the situation $\mathbb{P}\left[\bigcup_{\theta\in\Theta} E_\theta\right] \leq \sum_{\theta\in\Theta} \mathbb{P}\left[E_\theta\right]$ where the sum is an infinite sum of finite quantities whose known upper bounds (via the Hoeffding bound) are all equal. However, if we know that 'nearby' θ's give 'nearly the same' losses, we can choose some prototype candidates $\theta_1, \cdots, \theta_N \in \Theta$ so that every $\theta \in \Theta$ is 'near' some θ_i. This way, we have discretized Θ. Now a standard union bound + Hoeffding trick on these prototype θ_i's will do the job because there's only finitely many of them and they approximate the global behavior of the loss for all $\theta \in \Theta$. Let's make these precise.

The 'nearness' of the prototype θ_i's is made rigorous through what is called an $r-$net.

Definition 10 ($r-$net)
Let $\varepsilon > 0$ and S a subset of a metric space (X, d). The closed ball of radius r around $x \in X$ will be denoted by $D_r(x) = \{y \in X \mid d(x, y) \leq r\}$. An *$r-$net of S* is a subset $T_r \subseteq S$ such that for each $x \in S$ there is some $y \in T_r$ satisfying $d(x, y) \leq r$. In other words, $S \subseteq \bigcup_{x\in T_r} D_r(x)$.

Now we need to find an $r-$net of Θ which is not only finite, but also not too large in size, otherwise the union bound + Hoeffding trick would not work. We are in luck because there is an $r-$net of $\Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$ of size $\leq \left(\frac{3B}{r}\right)^p$. We'll prove this later.

Now let's make the idea of "nearby θ's give nearly the same loss" precise, which will be an added assumption on the loss function. Recall that the loss of h_θ on (x, y) is $\ell(h_\theta(x), y)$. This value actually depends on three things, namely θ, x, y. We would like that for the same data point (x, y), changing the parameters of the model only a little bit does not change the loss by much. That is, for any (x, y), if we change the parameters only slightly, the change in the penalty is controlled. This is captured by something called Lipschitz-ness. A Lipschitz function is continuous, but not necessarily differentiable. All we can say is that Lipschitz functions are almost everywhere differentiable.

Definition 11 ($\kappa-$Lipschitz)
A real. valued function $f : X \to \mathbb{R}$ on a metric space (X, d) is said to be *$\kappa-$Lipschitz* if

$|f(x) - f(y)| \leq \kappa d(x, y)$ for every $x, y \in X$.

Let's now try to imitate the calculations as before, incorporating the Lipschitzness of the loss function and see how things turn out. Say, the loss ℓ takes values in $[0, 1]$ and is $\kappa-$Lipschitz in θ with the usual ℓ_2 norm on $\mathbb{R}^p$, that is, $|\ell(h_\theta(x), y) - \ell(h_{\theta'}(x), y)| \leq \kappa \|\theta - \theta'\|_2$. Consequently, $L, \hat{L}$ are also $\kappa-$Lipschitz. Since we already have an $r-$net T for $\Theta \subseteq \mathbb{R}^p$ of size $N \leq \left(\frac{3B}{r}\right)^p$, let's just focus on the loss values on these points, which approximate the other points in its neighborhood. Say $T = \{\theta_1, \cdots, \theta_N\}$. So we are interested in the good event $E = \left\{\left|L(\theta_i) - \hat{L}(\theta_i)\right| < \frac{\varepsilon}{2} \; \forall \; i \in [N]\right\}$. We put $\varepsilon/2$ instead of ε in order to account for the errors caused by approximation of points in Θ outside T. By Corollary 9, $\mathbb{P}[E] \geq 1 - 2N \exp(-n\varepsilon^2/2)$. We have thus obtained a uniform upper bound for $\left|L - \hat{L}\right|$ on T with high probability. Let's extend this to Θ. Indeed for any $\theta \in \Theta$, there is some $x \in T$ such that $\|\theta - x\|_2 \leq r$. Recall that $L, \hat{L}$ are $\kappa-$Lipschitz. Thus conditioned on E, $\left|L(\theta) - \hat{L}(\theta)\right| \leq |L(\theta) - L(x)| + \left|L(x) - \hat{L}(x)\right| + \left|\hat{L}(x) - \hat{L}(\theta)\right| \leq 2\kappa r + \frac{\varepsilon}{2}$.

Theorem 12
Suppose we are given an hypothesis class $\mathcal{H}$ parameterized by $\Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$, a loss function ℓ taking values in $[0, 1]$ and $\kappa-$Lipschitz on model parameters θ, n training samples and an (additive) error tolerance $\varepsilon > 0$.
Then $\mathbb{P}\left[\left|L(\theta) - \hat{L}(\theta)\right| < \varepsilon \; \forall \; \theta \in \Theta\right] \geq 1 - 2\left(\frac{18B\kappa}{\varepsilon}\right)^p \exp(-2n\varepsilon^2)$.

Proof. Choose $r = \frac{\varepsilon}{5\kappa}$ in the above discussion. ■

Theorem 13
Suppose we are given an hypothesis class $\mathcal{H}$ parameterized by $\Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$, a loss function ℓ taking values in $[0, 1]$ and $\kappa-$Lipschitz on model parameters θ and n training samples.
Then $\mathbb{P}\left[\left|L(\theta) - \hat{L}(\theta)\right| < \mathcal{O}\left(\sqrt{\frac{p \max\{1, \ln(\kappa B n)\}}{n}}\right) \; \forall \; \theta \in \Theta\right] \geq 1 - \mathcal{O}(\exp(-\Omega(p)))$.

Corollary 14
Suppose we are given an hypothesis class $\mathcal{H}$ parameterized by $\Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$, a loss function ℓ taking values in $[0, 1]$ and $\kappa-$Lipschitz on model parameters θ, $\delta \in (0, 0.5)$ and an (additive) error tolerance $\varepsilon > 0$. Then it is enough to have a training sample set of size $n = \mathcal{O}\left(\frac{\ln(2/\delta) + p \max\{1, \ln(18B\kappa/\varepsilon)\}}{2\varepsilon^2}\right)$ to guarantee $\left|L(\theta) - \hat{L}(\theta)\right| < \varepsilon \; \forall \; \theta \in \Theta$ with probability at least $1 - \delta$.

We end with an upper bound for an $r-$net of Θ for $r \leq B$. In what follows, B, B^o will respectively denote closed and open balls.

Consider the following algorithm for any given $r > 0$ to find a set $T_r \subseteq \Theta$.

Input: $r > 0$, dimension p

Output: a number N and points $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N \in \Theta$ such that every point in Θ is $r-$close to some $\boldsymbol{v}_i$.

1: **begin**
2: $N \leftarrow 1$
3: $\boldsymbol{v}_1 \leftarrow (1, 0, \cdots, 0) \in \Theta$
4: $T \leftarrow B_r^o(\boldsymbol{v}_1) \cap \Theta$ ▷ points in Θ which are at distance $< \varepsilon$ from $\boldsymbol{v}_1$
5: **while** $N \geq 1$ **do**
6: $\boldsymbol{v}_N \leftarrow$any point in $\Theta \setminus T$
7: $T \leftarrow T \cup (B_r(\boldsymbol{v}_2) \cap \Theta)$
8: **if** $S = \Theta$ **then** ▷ check if Θ has been covered
9: **break**
10: **else**
11: $N \leftarrow N + 1$
12: **end if**
13: **end while**
14: **return** $N, T_r = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$
15: **end**

Now we prove that this algorithm actually gives T_r and size N as desired. If the above algorithm terminates with answer N, T_r, then $\Theta \subseteq \bigcup_{i=1}^{N} B_r^o(\boldsymbol{v}_i) \subseteq \bigcup_{i=1}^{N} B_r(\boldsymbol{v}_i)$.

Claim 15
The above algorithm terminates.

Proof. Suppose the algorithm goes on forever. So we get a sequence of points $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots$ such that $\Theta \subseteq \bigcup_{i \in \mathbb{N}} B_r^o(\boldsymbol{v}_i)$. Since Θ is compact there is a finite N such that $\Theta \subseteq \bigcup_{i=1}^{N} B_r^o(\boldsymbol{v}_i)$. This is a contradiction to our original assumption. ■

Next we note that just by how our algorithm is designed, if $\boldsymbol{x}, \boldsymbol{y} \in T_r$ then $\|\boldsymbol{x} - \boldsymbol{y}\|_2 \geq r$. This is because a new point (line 6) is always chosen so that it is not in the $r-$ball around any of the previously chosen points, and distance is symmetric.

Further T_r is maximal in the sense that if $T' \supsetneq T_r$ is a collection of points in Θ, there will be two points in T' which are at most $r-$close to each other. This is by our breaking criterion on line 8. Simply put, T_r covers Θ with $\varepsilon-$balls.

Claim 16
If $\boldsymbol{x}, \boldsymbol{y} \in T_r$ are distinct, then $B^o_{\frac{r}{2}}(\boldsymbol{x}) \cap B^o_{\frac{r}{2}}(\boldsymbol{y}) \cap \Theta = \varnothing$.

Proof. Suppose $\boldsymbol{p} \in \Theta \cap B^o_{\frac{r}{2}}(\boldsymbol{x}) \cap B^o_{\frac{r}{2}}(\boldsymbol{y})$ and say $\boldsymbol{y}$ was picked after $\boldsymbol{x}$ in the algorithm. Then $\|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq \|\boldsymbol{x}-\boldsymbol{p}\|_2 + \|\boldsymbol{p}-\boldsymbol{y}\|_2 \leq r$. Moreover equality here occurs only when $\|\boldsymbol{p}-\boldsymbol{x}\|_2 = \|\boldsymbol{p}-\boldsymbol{y}\|_2 = \frac{r}{2}$ which means $p \notin B^o_{\frac{r}{2}}(\boldsymbol{x})$ which is a contradiction. So it must happen that $\|\boldsymbol{x}-\boldsymbol{y}\|_2 < r$ which contradicts the constructive step in line 6 because this indicated that $\boldsymbol{y}$ was picked in the $r-$ball around $\boldsymbol{x}$. ■

Claim 17
For any $t \geq 0$, if $\boldsymbol{x} \in \bigcup_{i\in[N]} B_t(\boldsymbol{v}_i) \subseteq \mathbb{R}^n$ then $\|\boldsymbol{x}\|_2 \leq B + t$.

Proof. Say $\boldsymbol{x} \in B_t(\boldsymbol{v}_i)$ for some i. Then $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{v}\|_2 + \|\boldsymbol{x}-\boldsymbol{v}\|_2 \leq B + t$. ■

The last two claims show that $X \coloneqq \bigcup_{i\in[N]} B_{r/2}(\boldsymbol{v}_i)$ is an almost disjoint union and is contained in the closed ball of radius $B + \frac{r}{2}$. The volume of a ball of radius t in $\mathbb{R}^p$ is $c_p t^p$ where c_p is a constant depending only on p. Thus $N(r/2)^p \leq (B + r/2)^p$ whence $N \leq \left(\frac{2B}{r} + 1\right)^p \leq \left(\frac{3B}{r}\right)^p$ whenever $r \leq B$.