Physics of Learning Theory
Lecture 2
Introduction to Learning Theory

February 5, 2025
Nilava Metya

1 Variations and applications of the Hoeffding bound

Recall the Hoeffding bound.

Theorem 1 (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 $t \geq 0$.

The first variation is obtained by replacing $\sum_{i=1}^k X_i$ with its sample mean $\frac{1}{k}\sum_{i=1}^k X_i$ and recalling that if $X \in [a,b]$ almost surely then X is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$.

Corollary 2
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 $t \geq 0$.

One can change parameters from t to $\varepsilon := t + \mathbb{E}\left[\overline{X}\right]$ to get $\mathbb{P}\left[\overline{X} \geq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mathbb{E}\left[\overline{X}\right]\right)^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $\varepsilon \geq \mathbb{E}\left[\overline{X}\right]$.

Using the above with all X_i's (hence a_i, b_i's) negated gives a lower tail bound, that is, $\mathbb{P}\left[\overline{X} \leq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mathbb{E}\left[\overline{X}\right]\right)^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $\varepsilon \leq \mathbb{E}\left[\overline{X}\right]$.

Combining we have

Corollary 3
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 $\overline{X}$ and $\mu = \mathbb{E}\left[\overline{X}\right]$. Then

$$\mathbb{P}\left[\overline{X} \geq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon-\mu\right)^2}{\sum_i (b_i-a_i)^2}\right\} \ \forall\, \varepsilon \geq \mu$$

$$\mathbb{P}\left[\overline{X} \leq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon-\mu\right)^2}{\sum_i (b_i-a_i)^2}\right\} \ \forall\, \varepsilon \leq \mu.$$

That is, we have a symmetric tail bound on either side of μ.

Recall that the Hoeffding bound gives the same bounds for a Bernouli random variable as a random variable taking values in $[0,1]$. Somehow this extra information about Bernouli random variables can be incorporated to get the stronger Chernoff bound.

Theorem 4 (Chernoff Bound)
Let $X_1, \cdots, X_n$ be independent $\{0,1\}$ values random variables such that $p_i = \mathbb{E}\left[X_i\right]$, with $X = \sum_i X_i$ and $\mu = \mathbb{E}\left[X\right] = \sum_i p_i$. Then $\mathbb{P}\left[X \geq (1+\varepsilon)\mu\right] \leq \exp\left\{-\frac{\varepsilon^2\mu}{2+\varepsilon}\right\}$ for $\varepsilon > 0$ and $\mathbb{P}\left[X \leq (1-\varepsilon)\mu\right] \leq \exp\left\{-\frac{\varepsilon^2\mu}{2}\right\}$ for $\varepsilon \in (0,1)$.

To understand why the Chernoff bound is slightly stronger, let's fix a probability parameter $\delta \in (0,1)$ (to be thought of as the failure probability). Say $X_1, \cdots, X_n$ are $\{0,1\}$ valued random variables with $p = \mathbb{E}\left[X_i\right]$ for each i. Then using Corollary 2 with $t = \sqrt{\frac{-\ln\delta}{2n}}$ gives $\mathbb{P}\left[\overline{X} \geq p + \sqrt{\frac{-\ln\delta}{2n}}\right] \leq \delta$ and using Theorem 4 with $\varepsilon = \sqrt{\frac{-3\ln\delta}{pn}}$ gives $\mathbb{P}\left[\overline{X} \geq p + \sqrt{\frac{-6p\ln\delta}{2n}}\right] \leq \delta$ as long as $p > \frac{3\ln(1/\delta)}{n}$. Note that this scenario happens only when δ is exponentially small (in terms of n). If p is constant, the Chernoff bound gives no useful information for the rate. However, in certain scenarios the iid Bernouli parameters $p \equiv p_n$ depend on the number of samples and $p_n \to 0$ so Chernoff speaks louder.

Now we look at some examples where we apply the Hoeffding (or Chernoff bound) to analyze algorithms.

Example 1 (Boosting in two sided errors). Suppose we designed a randomized algorithm f to answer a $0/1$ question and on any given input x, it answers correctly with probability $\frac{2}{3}$. How can we use f to correctly predict its actual answer of input x with very high confidence. Of course, we may or may not get the correct answer if we run f once on x. Intuitively, if we run f on x 3000 times, we expect to get about 2000 correct answers and 1000 wrong answers. Of course, then with high confidence we predict that the answer which is reported most number of times (that is, more than half the times) is the correct one. Intuitively, this makes sense. But, how do we quantify this confidence? We want to answer the question that how many times should we run f on x so that we succeed with probability $1 - \frac{1}{n}$.

Let's run the algorithm n times on x and let the outputs be $X_1, \cdots, X_k \in \{0,1\}$. Suppose the actual answer of x on the actual question was $a \in \{0,1\}$ (a is not random, but X_i's are). Our reported answer is $Y = \mathbf{1}_{\overline{X} \geq \frac{1}{2}}$. This is also a random variable and we will show that the probability of Y not being a is very small. Note that X_i are all Bernouli$((1+a)/3)$, so Hoeffding bound is good enough. Corollary 3 gives a the same tail bound on $\mathbb{P}\left[\overline{X} \geq \frac{1}{2}\right]$ and $\mathbb{P}\left[\overline{X} \leq \frac{1}{2}\right]$ (corresponding to the 'bad' event $\{Y \neq a\}$ for $a = 0, 1$ respectively). Thus $\mathbb{P}[Y \neq a] \leq \exp\{-2k((1-2a)/6)^2\} = \exp\{-k/18\}$ irrespective of whether a is 0 or 1. Hence $k \geq 18 \ln n$ trials gives us a confidence of $\geq 1 - \frac{1}{n}$.

Example 2 (Johnson-Lindenstrauss lemma). Say we are a dimension d, a probability parameter $\delta \in (0, 1/2)$, fault tolerance $\varepsilon \in (0,1)$, a positive integer $m > \frac{-\ln \delta}{\varepsilon^2}$ and any vector $\boldsymbol{x} \in \mathbb{R}^d$. We pick a matrix $M \in \mathbb{R}^{m \times d}$ whose entries are independent $\mathcal{N}(0,1)$'s and consider $\Pi = \frac{1}{\sqrt{m}} M$. Then $\mathbb{P}\left[(1-\varepsilon)\|\boldsymbol{x}\|_2 \leq \|\Pi \boldsymbol{x}\|_2 \leq (1+\varepsilon)\|\boldsymbol{x}\|_2\right] \geq 1 - \delta$. This is known as the famous Johnson-Lindenstrauss dimensionality reduction. In fact, if we are given n points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, by a union bound argument (because the above was for one $\boldsymbol{x}$) we can show that for $m = \mathcal{O}(\ln(n/\delta)/\varepsilon^2)$, all the distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ are preserved under a random such Π with probability $1 - \delta$. For instance, with probability 0.99, we can reduce the dimension to $m = \mathcal{O}(\ln n/\varepsilon^2)$ upto ε error. In other words, the existence of such a Π has positive probability for small enough δ. Since the Π's were combinatorial (i.e., chosen from a finite set), we conclude that such a dimension-reducing Π always exists.

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 $L(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}}[\ell(h(x), y)]$, 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 3 (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 4 (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 $E \coloneqq \bigcup_{j\in[K]}\{|Y_j-\mu_j|\geq t\} = \bigcup_j E_j$. 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 $\mathbb{P}\left[|Y_j - \mu_j| < \varepsilon \; \forall \; j \in [K]\right] \geq 1 - \delta$. 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 $L\left(\hat{h}\right) - \min_{h \in \mathcal{H}} L(h)$ 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 5
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 6
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 5. ■

Corollary 7
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 8
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 9 ($r-$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 considerable size. A detailed proof of this can be found in Appendix A.1.

Lemma 10
$\Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$ has an $r-$net of size $\leq \left(\frac{3B}{r}\right)^p$.

Proof. Greedy. ■

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 8, $\mathbb{P}[E] \geq 1 - 2N \exp(-n\varepsilon^2/2)$. 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 $1 - \delta$.

A Appendix

A.1 Proof of Lemma 10

Proof. In what follows, $\mathcal{B}, \mathcal{B}^o$ will respectively denote closed and open balls.

Consider the following algorithm for any given $r > 0$ to find a set $T_r \subseteq \Theta = \{\theta \in \mathbb{R}^p \mid \|\theta\|_2 \leq B\}$.

Input: $r > 0$, dimension p, radius B (of Θ).

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: $T \leftarrow \mathcal{B}_r^o(\boldsymbol{v}_1) \cap \Theta$ ▷ 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: $T \leftarrow T \cup (\mathcal{B}_r(\boldsymbol{v}_2) \cap \Theta)$
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 $\Theta \subseteq \bigcup_{i=1}^{N} \mathcal{B}_r^o(\boldsymbol{v}_i) \subseteq \bigcup_{i=1}^{N} \mathcal{B}_r(\boldsymbol{v}_i)$.

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 $\Theta \subseteq \bigcup_{i \in \mathbb{N}} \mathcal{B}_r^o(\boldsymbol{v}_i)$. Since Θ is compact there is a finite N such that $\Theta \subseteq \bigcup_{i=1}^{N} \mathcal{B}_r^o(\boldsymbol{v}_i)$. 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 $r-$balls.

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

Proof. Suppose $\boldsymbol{p} \in \Theta \cap \mathcal{B}^o_{\frac{r}{2}}(\boldsymbol{x}) \cap \mathcal{B}^o_{\frac{r}{2}}(\boldsymbol{y})$ 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 $\boldsymbol{p} \notin \mathcal{B}^o_{\frac{r}{2}}(\boldsymbol{x})$ 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 $\boldsymbol{x} \in \bigcup_{i\in[N]} \mathcal{B}_t(\boldsymbol{v}_i) \subseteq \mathbb{R}^n$ then $\|\boldsymbol{x}\|_2 \leq B + t$.

Proof. Say $\boldsymbol{x} \in \mathcal{B}_t(\boldsymbol{v}_i)$ 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 $X \coloneqq \bigcup_{i\in[N]} \mathcal{B}_{r/2}(\boldsymbol{v}_i)$ 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 $r \leq B$. ■