18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 1
Scribe: Philippe Rigollet Sep. 9, 2015

1. WHAT IS MACHINE LEARNING (IN THIS COURSE)?

This course focuses on *statistical learning theory*, which roughly means understanding the amount of data required to achieve a certain prediction accuracy. To better understand what this means, we first focus on stating some differences between *statistics* and *machine learning* since the two fields share common goals.

Indeed, both seem to try to use data to improve decisions. While these fields have evolved in the same direction and currently share a lot of aspects, they were at the beginning quite different. Statistics was around much before machine learning and statistics was already a fully developed scientific discipline by 1920, most notably thanks to the contributions of R. Fisher, who popularized *maximum likelihood estimation (MLE)* as a systematic tool for statistical inference. However, MLE requires essentially knowing the probability distribution from which the data is draw, up to some unknown parameter of interest. Often, the unknown parameter has a physical meaning and its estimation is key in better understanding some phenomena. Enabling MLE thus requires knowing a lot about the data generating process: this is known as *modeling*. Modeling can be driven by physics or prior knowledge of the problem. In any case, it requires quite a bit of domain knowledge.

More recently (examples go back to the 1960's) new types of datasets (demographics, social, medical,...) have become available. However, modeling the data that they contain is much more hazardous since we do not understand very well the input/output process thus requiring a *distribution free* approach. A typical example is image classification where the goal is to label an image simply from a digitalization of this image. Understanding what makes an image a cat or a dog for example is a very complicated process. However, for the classification task, one does not need to understand the labelling process but rather to replicate it. In that sense, machine learning favors a blackbox approach (see Figure 1).

input X
blackbox
output Y
input X
$y = f(x) + \varepsilon$
output Y

Figure 1: The machine learning blackbox (left) where the goal is to replicate input/output pairs from past observations, versus the statistical approach that opens the blackbox and models the relationship.

These differences between statistics and machine learning have receded over the last couple of decades. Indeed, on the one hand, statistics is more and more concerned with finite sample analysis, model misspecification and computational considerations. On the other hand, probabilistic modeling is now inherent to machine learning. At the intersection of the two fields, lies *statistical learning theory*, a field which is primarily concerned with sample complexity questions, some of which will be the focus of this class.

2. STATISTICAL LEARNING THEORY

2.1 Binary classification

A large part of this class will be devoted to one of the simplest problem of statistical learning theory: binary classification (aka pattern recognition [DGL96]). In this problem, we observe $(X_1, Y_1), \dots, (X_n, Y_n)$ that are n independent random copies of $(X, Y) \in \mathcal{X} \times \{0, 1\}$. Denote by $P_{X,Y}$ the joint distribution of (X, Y). The so-called *feature* X lives in some abstract space $\mathcal{X}$ (think $\mathbb{R}^d$) and $Y \in \{0, 1\}$ is called *label*. For example, X can be a collection of gene expression levels measured on a patient and Y indicates if this person suffers from obesity. The goal of binary classification is to build a rule to predict Y given X using only the data at hand. Such a rule is a function $h : \mathcal{X} \to \{0, 1\}$ called a *classifier*. Some classifiers are better than others and we will favor ones that have low *classification error* $R(h) = \mathbb{P}(h(X) \neq Y)$. Let us make some important remarks.

Fist of all, since $Y \in \{0, 1\}$ then Y has a Bernoulli distribution: so much for distribution free assumptions! However, we will not make assumptions on the marginal distribution of X or, what matters for prediction, the conditional distribution of Y given X. We write, $Y|X \sim \mathsf{Ber}(\eta(X))$, where $\eta(X) = \mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X]$ is called the *regression function* of Y onto X.

Next, note that we did not write $Y = \eta(X)$. Actually we have $Y = \eta(X) + \varepsilon$, where $\varepsilon = Y - \eta(X)$ is a "noise" random variable that satisfies $\mathbb{E}[\varepsilon|X] = 0$. In particular, this noise accounts for the fact that X may not contain enough information to predict Y perfectly. This is clearly the case in our genomic example above: it not whether there is even any information about obesity contained in a patient's genotype. The noise vanishes if and only if $\eta(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$. Figure 2.1 illustrates the case where there is no noise and the the more realistic case where there is noise. When $\eta(x)$ is close to .5, there is essentially no information about Y in X as the Y is determined essentially by a toss up. In this case, it is clear that even with an infinite amount of data to learn from, we cannot predict Y well since there is nothing to learn. We will see what the effect of the noise also appears in the sample complexity.

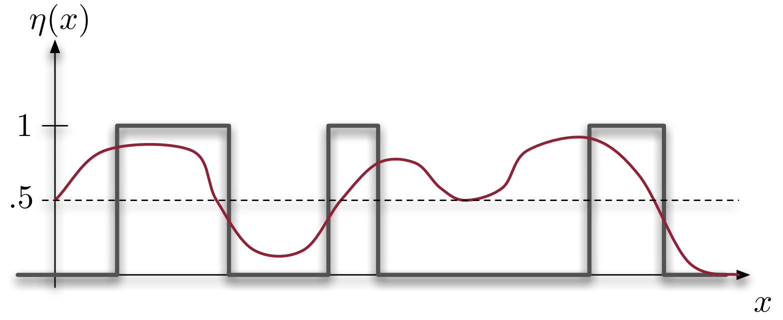


Figure 2: The thick black curve corresponds to the noiseless case where $Y = \eta(X) \in \{0, 1\}$ and the thin red curve corresponds to the more realistic case where $\eta \in [0, 1]$. In the latter case, even full knowledge of η does not guarantee a perfect prediction of Y.

In the presence of noise, since we cannot predict Y accurately, we cannot drive the classification error $R(h)$ to zero, regardless of what classifier h we use. What is the smallest value that can be achieved? As a thought experiment, assume to begin with that we have all the information that we may ever hope to get, namely we know the regression function $\eta(\cdot)$. For a given X to classify, if $\eta(X) = 1/2$ we may just toss a coin to decide our prediction and discard X since it contains no information about Y. However, if $\eta(X) \neq 1/2$, we have an edge over random guessing: if $\eta(X) > 1/2$, it means that $\mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X)$ or, in words, that 1 is more likely to be the correct label. We will see that the classifier $h^*(X) = \mathbb{1}(\eta(X) > 1/2)$ (called *Bayes classifier*) is actually the best possible classifier in the sense that

$$R(h^*) = \inf_{h(\cdot)} R(h)\,,$$

where the infimum is taken over all classifiers, i.e. functions from $\mathcal{X}$ to $\{0, 1\}$. Note that unless $\eta(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$ (noiseless case), we have $R(h^*) \neq 0$. However, we can always look at the *excess risk* $\mathcal{E}(h)$ of a classifier h defined by

$$\mathcal{E}(h) = R(h) - R(h^*) \geq 0\,.$$

In particular, we can hope to drive the excess risk to zero with enough observations by mimicking h^* accurately.

2.2 Empirical risk

The Bayes classifier h^*, while optimal, presents a major drawback: we cannot compute it because we do not know the regression function η. Instead, we have access to the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which contains some (but not all) information about η and thus h^*. In order to mimic the properties of h^* recall that it minimizes $R(h)$ over all h. But the function $R(\cdot)$ is unknown since it depends on the unknown distribution $P_{X,Y}$ of (X, Y). We estimate it by the empirical classification error, or simply *empirical risk* $\hat{R}_n(\cdot)$ defined for any classifier h by

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i)\,.$$

Since $\mathbb{E}[\mathbb{1}(h(X_i) \neq Y_i)] = \mathbb{P}(h(X_i) \neq Y_i) = R(h)$, we have $\mathbb{E}[\hat{R}_n(h)] = R(h)$ so $\hat{R}_n(h)$ is an *unbiased* estimator of $R(h)$. Moreover, for any h, by the law of large numbers, we have $\hat{R}_n(h) \to R(h)$ as $n \to \infty$, almost surely. This indicates that if n is large enough, $\hat{R}_n(h)$ should be close to $R(h)$.

As a result, in order to mimic the performance of h^*, let us use the *empirical risk minimizer (ERM)* $\hat{h}$ defined to minimize $\hat{R}_n(h)$ over all classifiers h. This is an easy enough task: define $\hat{h}$ such $\hat{h}(X_i) = Y_i$ for all $i = 1, \ldots, n$ and $h(x) = 0$ if $x \notin \{X_1, \ldots, X_n\}$. We have $\hat{R}_n(\hat{h}) = 0$, which is clearly minimal. The problem with this classifier is obvious: it does not *generalize* outside the data. Rather, it predicts the label 0 for any x that is not in the data. We could have predicted 1 or any combination of 0 and 1 and still get $\hat{R}_n(\hat{h}) = 0$. In particular it is unlikely that $\mathbb{E}[R(\hat{h})]$ will be small.

Important Remark: Recall that $R(h) = \mathbb{P}(h(X) \neq Y)$.

If $\hat{h}(\cdot) = \hat{h}(\{(X_1, Y_1), \dots, (X_n, Y_n)\}\,; \cdot)$ is constructed from the data, $R(\hat{h})$ denotes the *conditional probability*

$$R(\hat{h}) = \mathbb{P}(\hat{h}(X) \neq Y | (X_1, Y_1), \dots, (X_n, Y_n)) \,.$$

rather than $\mathbb{P}(\hat{h}(X) \neq Y)$. As a result $R(\hat{h})$ is a random variable since it depends on the randomness of the data $(X_1, Y_1), \dots, (X_n, Y_n)$. One way to view this is to observe that we compute the *deterministic* function $R(\cdot)$ and then plug in the random classifier $\hat{h}$.

This problem is inherent to any method if we are not willing to make any assumption on the distribution of (X, Y) (again, so much for distribution freeness!). This can actually be formalized in theorems, known as *no-free-lunch* theorems.

Theorem: For any integer $n \geq 1$, any classifier $\hat{h}$ built from $(X_1, Y_1), \dots, (X_n, Y_n)$ and any $\varepsilon > 0$, there exists a distribution $P_{X,Y}$ for (X, Y) such that $R(h^*) = 0$ and

$$\mathbb{E}R(\hat{h}_n) \geq 1/2 - \varepsilon \,.$$

To be fair, note that here the distribution of the pair (X, Y) is allowed to depend on n which is cheating a bit but there are weaker versions of the no-free-lunch theorem that essentially imply that it is impossible to learn without further assumptions. One such theorem is the following.

Theorem: For any classifier $\hat{h}$ built from $(X_1, Y_1), \dots, (X_n, Y_n)$ and any sequence $\{a_n\}_n > 0$ that converges to 0, there exists a distribution $P_{X,Y}$ for (X, Y) such that $R(h^*) = 0$ and

$$\mathbb{E}R(\hat{h}_n) \geq a_n \,, \quad \text{for all } n \geq 1$$

In the above theorem, the distribution of (X, Y) is allowed to depend on the whole sequence $\{a_n\}_n > 0$ but not on a specific n. The above result implies that the convergence to zero of the classification error may be arbitrarily slow.

2.3 Generative vs discriminative approaches

Both theorems above imply that we need to restrict the distribution $P_{X,Y}$ of (X, Y). But isn't that exactly what statistical modeling is? The is answer is not so clear depending on how we perform this restriction. There are essentially two schools: *generative* which is the statistical modeling approach and *discriminative* which is the machine learning approach.

Generative: This approach consists in restricting the set of candidate distributions $P_{X,Y}$. This is what is done in *discriminant analysis*[1]where it is assumed that the condition dis-

[1] Amusingly, the **generative** approach is called **discriminant** analysis but don't let the terminology fool you.

tributions of X given Y (there are only two of them: one for $Y = 0$ and one for $Y = 1$) are Gaussians on $\mathcal{X} = \mathbb{R}^d$ (see for example [HTF09] for an overview of this approach).

Discriminative: This is the machine learning approach. Rather than making assumptions directly on the distribution, one makes assumptions on what classifiers are likely to perform correctly. In turn, this allows to eliminate classifiers such as the one described above and that does not generalize well.

While it is important to understand both, we will focus on the discriminative approach in this class. Specifically we are going to assume that we are given a class $\mathcal{H}$ of classifiers such that $R(h)$ is small for some $h \in \mathcal{H}$.

2.4 Estimation vs. approximation

Assume that we are given a class $\mathcal{H}$ in which we expect to find a classifier that performs well. This class may be constructed from domain knowledge or simply computational convenience. We will see some examples in the class. For any candidate classifier $\hat{h}_n$ built from the data, we can decompose its excess risk as follows:

$$\mathcal{E}(\hat{h}_n) = R(\hat{h}_n) - R(h^*) = \underbrace{R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h)}_{\text{estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} R(h) - R(h^*)}_{\text{approximation error}} .$$

On the one hand, *estimation error* accounts for the fact that we only have a finite amount of observations and thus a partial knowledge of the distribution $P_{X,Y}$. Hopefully we can drive this error to zero as $n \to \infty$. But we already know from the no-free-lunch theorem that this will not happen if $\mathcal{H}$ is the set of all classifiers. Therefore, we need to take $\mathcal{H}$ small enough. On the other hand, if $\mathcal{H}$ is too small, it is unlikely that we will find classifier with performance close to that of h^*. A tradeoff between estimation and approximation can be made by letting $\mathcal{H} = \mathcal{H}_n$ grow (but not too fast) with n.

For now, assume that $\mathcal{H}$ is fixed. The goal of statistical learning theory is to understand how the estimation error drops to zero as a function not only of n but also of $\mathcal{H}$. For the first argument, we will use *concentration inequalities* such as Hoeffding's and Bernstein's inequalities that allow us to control how close the empirical risk is to the classification error by bounding the random variable

$$\Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i) - \mathbb{P}(h(X) \neq Y)\Big|$$

with high probability. More generally we will be interested in results that allow to quantify how close the average of independent and identically distributed (i.i.d) random variables is to their common expected value.

Indeed, since by definition, we have $\hat{R}_n(\hat{h}) \leq \hat{R}_n(h)$ for all $h \in \mathcal{H}$, the estimation error can be controlled as follows. Define $\bar{h} \in \mathcal{H}$ to be any classifier that minimizes $R(\cdot)$ over $\mathcal{H}$ (assuming that such a classifier exist).

$$\begin{aligned}
R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) &= R(\hat{h}_n) - R(\bar{h}) \\
&= \underbrace{\hat{R}_n(\hat{h}_n) - \hat{R}_n(\bar{h})}_{\leq 0} + R(\hat{h}_n) - \hat{R}_n(\hat{h}_n) + \hat{R}_n(\bar{h}) - R(\bar{h}) \\
&\leq \big|\hat{R}_n(\hat{h}_n) - R(\hat{h}_n)\big| + \big|\hat{R}_n(\bar{h}) - R(\bar{h})\big| .
\end{aligned}$$

Since $\bar{h}$ is deterministic, we can use a concentration inequality to control $\left|\hat{R}_n(\bar{h}) - R(\bar{h})\right|$. However,

$$\hat{R}_n(\hat{h}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{h}_n(X_i) \neq Y_i)$$

is **not** the average of independent random variables since $\hat{h}_n$ depends in a complicated manner on all of the pairs $(X_i, Y_i), i = 1, \dots, n$. To overcome this limitation, we often use a blunt, but surprisingly accurate tool: we "sup out" $\hat{h}_n$,

$$\left|\hat{R}_n(\hat{h}_n) - R(\hat{h}_n)\right| \le \sup_{h \in \mathcal{H}} \left|\hat{R}_n(\hat{h}_n) - R(\hat{h}_n)\right| .$$

Controlling this supremum falls in the scope of *suprema of empirical processes* that we will study in quite a bit of detail. Clearly the supremum is smaller as $\mathcal{H}$ is smaller but $\mathcal{H}$ should be kept large enough to have good approximation properties. This is the tradeoff between approximation and estimation. It is also know in statistics as the *bias-variance* tradeoff.

References

[DGL96] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Applications of Mathematics (New York), vol. 31, Springer-Verlag, New York, 1996. MR MR1383093 (97d:68196)

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, second ed., Springer Series in Statistics, Springer, New York, 2009, Data mining, inference, and prediction. MR 2722294 (2012d:62081)

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 2
Scribe: Jonathan Weed — Sep. 14, 2015

Part I
Statistical Learning Theory

1. BINARY CLASSIFICATION

In the last lecture, we looked broadly at the problems that machine learning seeks to solve and the techniques we will cover in this course. Today, we will focus on one such problem, *binary classification*, and review some important notions that will be foundational for the rest of the course.

Our present focus on the problem of binary classification is justified because both binary classification encompasses much of what we want to accomplish in practice and because the response variables in the binary classification problem are bounded. (We will see a very important application of this fact below.) It also happens that there are some nasty surprises in non-binary classification, which we avoid by focusing on the binary case here.

1.1 Bayes Classifier

Recall the setup of binary classification: we observe a sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ of n independent draws from a joint distribution $P_{X,Y}$. The variable Y (called the *label*) takes values in $\{0, 1\}$, and the variable X takes values in some space $\mathcal{X}$ representing "features" of the problem. We can of course speak of the marginal distribution P_X of X alone; moreover, since Y is supported on $\{0, 1\}$, the conditional random variable $Y|X$ is distributed according to a Bernoulli distribution. We write $Y|X \sim \text{Bernoulli}(\eta(X))$, where

$$\eta(X) = \mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X].$$

(The function η is called the *regression function*.)

We begin by defining an optimal classifier called the Bayes classifier. Intuitively, the Bayes classifier is the classifier that "knows" η—it is the classifier we would use if we had perfect access to the distribution $Y|X$.

Definition: The *Bayes classifier* of X given Y, denoted h^*, is the function defined by the rule

$$h^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{if } \eta(x) \le 1/2. \end{cases}$$

In other words, $h^*(X) = 1$ whenever $\mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X)$.

Our measure of performance for any classifier h (that is, any function mapping X to $\{0, 1\}$) will be the *classification error*: $R(h) = \mathbb{P}(Y \ne h(X))$. The Bayes risk is the value $R^* = R(h^*)$ of the classification error associated with the Bayes classifier. The following theorem establishes that the Bayes classifier is optimal with respect to this metric.

Theorem: For any classifier h, the following identity holds:

$$R(h) - R(h^*) = \int_{h \neq h^*} |2\eta(x) - 1| \, P_x(dx) = \mathbb{E}_X[|2\eta(X) - 1| \mathbf{1}(h(X) \neq h^*(X))] \quad (1.1)$$

Where $h = h^*$ is the (measurable) set $\{x \in \mathcal{X} \mid h(x) \neq h^*(x)\}$.

In particular, since the integrand is nonnegative, the classification error R^* of the Bayes classifier is the minimizer of $R(h)$ over all classifiers h.

Moreover,

$$R(h^*) = \mathbb{E}[\min(\eta(X), 1 - \eta(X))] \leq \frac{1}{2}. \quad (1.2)$$

Proof. We begin by proving Equation (1.2). The definition of $R(h)$ implies

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y = 1, h(X) = 0) + \mathbb{P}(Y = 0, h(X) = 1),$$

where the second equality follows since the two events are disjoint. By conditioning on X and using the tower law, this last quantity is equal to

$$\mathbb{E}[\mathbb{E}[\mathbf{1}(Y = 1, h(X) = 0)|X]] + \mathbb{E}[\mathbb{E}[\mathbf{1}(Y = 0, h(X) = 1)|X]]$$

Now, $h(X)$ is measurable with respect to X, so we can factor it out to yield

$$\mathbb{E}[\mathbf{1}(h(X) = 0)\eta(X) + \mathbf{1}(h(X) = 1)(1 - \eta(X))]], \quad (1.3)$$

where we have replaced $\mathbb{E}[Y|X]$ by $\eta(X)$.

In particular, if $h = h^*$, then Equation 1.3 becomes

$$\mathbb{E}[\mathbf{1}(\eta(X) \leq 1/2)\eta(X) + \mathbf{1}(\eta(x) > 1/2)(1 - \eta(X))].$$

But $\eta(X) \leq 1/2$ implies $\eta(X) \leq 1 - \eta(X)$ and conversely, so we finally obtain

$$\begin{aligned} R(h^*) &= \mathbb{E}[\mathbf{1}(\eta(X) \leq 1/2)\eta(X) + \mathbf{1}(\eta(x) > 1/2)(1 - \eta(X))] \\ &= \mathbb{E}[(\mathbf{1}(\eta(X) \leq 1/2) + \mathbf{1}(\eta(x) > 1/2)) \min(\eta(X), 1 - \eta(X))] \\ &= \mathbb{E}[\min(\eta(X), 1 - \eta(X))], \end{aligned}$$

as claimed. Since $\min(\eta(X), 1 - \eta(X)) \leq 1/2$, its expectation is also certainly at most $1/2$ as well.

Now, given an arbitrary h, applying Equation 1.3 to both h and h^* yields

$$\begin{aligned} R(h) - R(h^*) = \mathbb{E}[&\mathbf{1}(h(X) = 0)\eta(X) + \mathbf{1}(h(X) = 1)(1 - \eta(X))] \\ &- \mathbf{1}(h^*(X) = 0)\eta(X) + \mathbf{1}(h^*(X) = 1)(1 - \eta(X))]], \end{aligned}$$

which is equal to

$$\mathbb{E}[(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))\eta(X) + (\mathbf{1}(h(X) = 1) - \mathbf{1}(h^*(X) = 1))(1 - \eta(X))].$$

Since $h(X)$ takes only the values 0 and 1, the second term can be rewritten as $-(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))$. Factoring yields

$$\mathbb{E}[(2\eta(X) - 1)(\mathbf{1}(h(X) = 0) - \mathbf{1}(h^*(X) = 0))].$$

The term $\mathbf{1}(h(X)=0)-\mathbf{1}(h^*(X)=0)$ is equal to -1, 0, or 1 depending on whether h and h^* agree. When $h(X)=h^*(X)$, it is zero. When $h(X)\neq h^*(X)$, it equals 1 whenever $h^*(X)=0$ and -1 otherwise. Applying the definition of the Bayes classifier, we obtain

$$\mathbb{E}[(2\eta(X)-1)\mathbf{1}(h(X)\neq h^*(X))\,\mathrm{sign}(\eta-1/2)]=\mathbb{E}[|2\eta(X)-1|\mathbf{1}(h(X)\neq h^*(X))],$$

as desired. □

We make several remarks. First, the quantity $R(h)-R(h^*)$ in the statement of the theorem above is called the *excess risk* of h and denoted $\mathcal{E}(h)$. ("Excess," that is, above the Bayes classifier.) The theorem implies that $\mathcal{E}(h)\geq 0$.

Second, the risk of the Bayes classifier R^* equals $1/2$ if and only if $\eta(X)=1/2$ almost surely. This maximal risk for the Bayes classifier occurs precisely when Y "contains no information" about the feature variable X. Equation (1.1) makes clear that the excess risk weighs the discrepancy between h and h^* according to how far η is from $1/2$. When η is close to $1/2$, no classifier can perform well and the excess risk is low. When η is far from $1/2$, the Bayes classifier performs well and we penalize classifiers that fail to do so more heavily.

As noted last time, linear discriminant analysis attacks binary classification by putting some model on the data. One way to achieve this is to impose some distributional assumptions on the conditional distributions $X|Y=0$ and $X|Y=1$.

We can reformulate the Bayes classifier in these terms by applying Bayes' rule:

$$\eta(x)=\mathbb{P}(Y=1|X=x)=\frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)+\mathbb{P}(X=x|Y=0)\mathbb{P}(Y=0)}.$$

(In general, when P_X is a continuous distribution, we should consider infinitesimal probabilities $\mathbb{P}(X\in dx)$.)

Assume that $X|Y=0$ and $X|Y=1$ have densities p_0 and p_1, and $\mathbb{P}(Y=1)=\pi$ is some constant reflecting the underlying tendency of the label Y. (Typically, we imagine that π is close to $1/2$, but that need not be the case: in many applications, such as anomaly detection, $Y=1$ is a rare event.) Then $h^*(X)=1$ whenever $\eta(X)\geq 1/2$, or, equivalently, whenever

$$\frac{p_1(x)}{p_0(x)}\geq\frac{1-\pi}{\pi}.$$

When $\pi=1/2$, this rule amounts to reporting 1 or 0 by comparing the densities p_1 and p_0. For instance, in Figure 1, if $\pi=1/2$ then the Bayes classifier reports 1 whenever $p_1\geq p_0$, i.e., to the right of the dotted line, and 0 otherwise.

On the other hand, when π is far from $1/2$, the Bayes classifier is weighed towards the underlying bias of the label variable Y.

1.2 Empirical Risk Minimization

The above considerations are all *probabilistic*, in the sense that they discuss properties of some underlying probability distribution. The statistician does *not* have access to the true probability distribution $P_{X,Y}$; she only has access to i.i.d. samples $(X_1,Y_1),\ldots,(X_n,Y_n)$. We consider now this statistical perspective. Note that the underlying distribution $P_{X,Y}$ still appears explicitly in what follows, since that is how we measure our performance: we judge the classifiers we produced on *future* i.i.d. draws from $P_{X,Y}$.

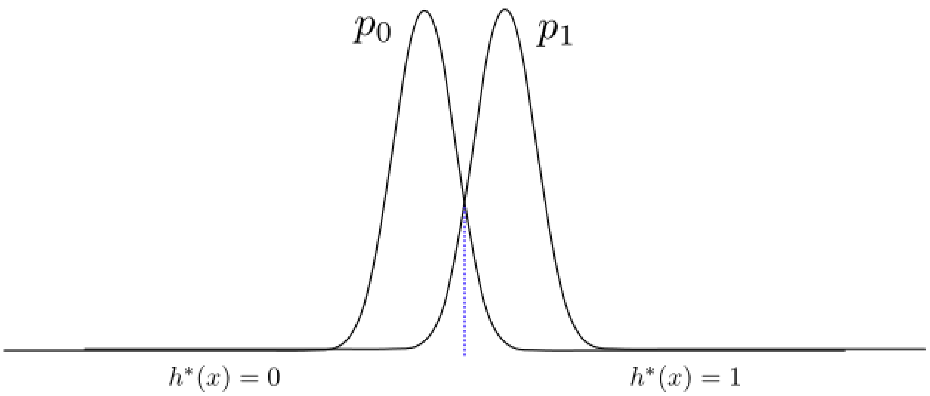


Figure 1: The Bayes classifier when $\pi = 1/2$.

Given data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we build a classifier $\hat{h}_n(X)$, which is random in two senses: it is a function of a random variable X and also depends implicitly on the random data $\mathcal{D}_n$. As above, we judge a classifier according to the quantity $\mathcal{E}(\hat{h}_n)$. This is a random variable: though we have integrated out X, the excess risk still depends on the data $\mathcal{D}_n$. We therefore will consider bounds both on its expected value and bounds that hold in high probability. In any case, the bound $\mathcal{E}(\hat{h}_n) \geq 0$ always holds. (This inequality does not merely hold "almost surely," since we proved that $R(h) \geq R(h^*)$ uniformly over all choices of classifier h.)

Last time, we proposed two different philosophical approaches to this problem. In particular, generative approaches make distributional assumptions about the data, attempt to learn parameters of these distributions, and then plug the resulting values into the model. The discriminative approach—the one taken in machine learning—will be described in great detail over the course of this semester. However, there is some middle ground, which is worth mentioning briefly. This middle ground avoids making explicit distributional assumptions about X while maintaining some of the flavor of the generative model.

The central insight of this middle approach is the following: since by definition $h^*(x) = \mathbf{1}(\eta(X) > 1/2)$, we estimate η by some $\hat{\eta}_n$ and thereby produce the estimator $\hat{h}_n = \mathbf{1}(\hat{\eta}_n(X) > 1/2)$. The result is called a *plug-in estimator.*

Of course, achieving good performance with a plug-in estimator requires some assumptions. (No-free-lunch theorems imply that we can't avoid making an assumption somewhere!) One possible assumption is that $\eta(X)$ is smooth; in that case, there are many nonparamteric regression techniques available (Nadaraya-Watson kernel regression, wavelet bases, etc.).

We could also assume that $\eta(X)$ is a function of a particular form. Since $\eta(X)$ is only supported on $[0, 1]$, standard linear models are generally inapplicable; rather, by applying the logit transform we obtain *logistic regression*, which assumes that η satisfies an identity of the form

$$\log\left(\frac{\eta(X)}{1 - \eta(X)}\right) = \theta^T X.$$

Plug-in estimators are called "semi-paramteric" since they avoid making any assumptions about the distribution of X. These estimators are widely used because they perform fairly well in practice and are very easy to compute. Nevertheless, they will not be our focus here.

In what follows, we focus here on the discriminative framework and empirical risk minimization. Our benchmark continues to be the risk function $R(h) = \mathbb{E}\mathbf{1}(Y \neq h(X))$, which is clearly not computable based on the data alone; however, we can attempt to use a naïve statistical "hammer" and replace the expectation with an average.

Definition: The *empirical risk* of a classifier h is given by

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \neq h(X_i)).$$

Minimizing the empirical risk over the family of all classifiers is useless, since we can always minimize the empirical risk by mimicking the data and classifying arbitrarily otherwise. We therefore limit our attention to classifiers in a certain family $\mathcal{H}$.

Definition: The *Empirical Risk Minimizer (ERM)* over $\mathcal{H}$ is any element[1] $\hat{h}^{\mathrm{erm}}$ of the set $\operatorname{argmin}_{h\in\mathcal{H}} \hat{R}_n(h)$.

In order for our results to be meaningful, the class $\mathcal{H}$ must be much smaller than the space of all classifiers. On the other hand, we also hope that the risk of $\hat{h}^{\mathrm{erm}}$ will be close to the Bayes risk, but that is unlikely if $\mathcal{H}$ is too small. The next section will give us tools for quantifying this tradeoff.

1.3 Oracle Inequalities

An oracle is a mythical classifier, one that is impossible to construct from data alone but whose performance we nevertheless hope to mimic. Specifically, given $\mathcal{H}$ we define $\bar{h}$ to be an element of $\operatorname{argmin}_{h\in\mathcal{H}} R(h)$—a classifier in $\mathcal{H}$ that minimizes the *true* risk. Of course, we cannot determine $\bar{h}$, but we can hope to prove a bound of the form

$$R(\hat{h}) \leq R(\bar{h}) + \text{something small}. \tag{1.4}$$

Since $\bar{h}$ is the best minimizer in $\mathcal{H}$ given perfect knowledge of the distribution, a bound of the form given in Equation 1.4 would imply that $\hat{h}$ has performance that is almost best-in-class. We can also apply such an inequality in the so-called *improper learning* framework, where we allow $\hat{h}$ to lie in a slightly larger class $\mathcal{H}' \supset \mathcal{H}$; in that case, we still get nontrivial guarantees on the performance of $\hat{h}$ if we know how to control $R(\bar{h})$

There is a natural tradeoff between the two terms on the right-hand side of Equation 1.4. When $\mathcal{H}$ is small, we expect the performance of the oracle $\bar{h}$ to suffer, but we may hope to approximate $\bar{h}$ quite closely. (Indeed, at the limit where $\mathcal{H}$ is a single function, the "something small" in Equation 1.4 is equal to zero.) On the other hand, as $\mathcal{H}$ grows the oracle will become more powerful but approximating it becomes more statistically difficult. (In other words, we need a larger sample size to achieve the same measure of performance.)

Since $R(\hat{h})$ is a random variable, we ultimately want to prove a bound in expectation or tail bound of the form

$$\mathbb{P}(R(\hat{h}) \leq R(\bar{h}) + \Delta_{n,\delta}(\mathcal{H})) \geq 1-\delta,$$

where $\Delta_{n,\delta}(\mathcal{H})$ is some explicit term depending on our sample size and our desired level of confidence.

[1] In fact, even an approximate solution will do: our bounds will still hold whenever we produce a classifier $\hat{h}$ satisfying $\hat{R}_n(\hat{h}) \leq \inf_{h\in\mathcal{H}} R_n(h) + \varepsilon$.

In the end, we should recall that

$$\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h^*) = (R(\hat{h}) - R(\bar{h})) + (R(\bar{h}) - R(h^*)).$$

The second term in the above equation is the approximation error, which is unavoidable once we fix the class $\mathcal{H}$. Oracle inequalities give a means of bounding the first term, the stochastic error.

1.4 Hoeffding's Theorem

Our primary building block is the following important result, which allows us to understand how closely the average of random variables matches their expectation.

Theorem (Hoeffding's Theorem): Let $X_1, \ldots, X_n$ be n independent random variables such that $X_i \in [0,1]$ almost surely.

Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i\right| > t\right) \leq 2e^{-2nt^2}.$$

In other words, deviations from the mean decay exponentially fast in n and t.

Proof. Define centered random variables $Z_i = X_i - \mathbb{E}X_i$. It suffices to show that

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-2nt^2},$$

since the lower tail bound follows analogously. (Exercise!)

We apply Chernoff bounds. Since the exponential function is an order-preserving bijection, we have for any $s > 0$

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) = \mathbb{P}\left(\exp\left(s\sum Z_i\right) > e^{stn}\right) &\leq e^{-stn}\mathbb{E}[e^{s\sum Z_i}] \quad \text{(Markov)}\\ &= e^{-stn}\prod \mathbb{E}[e^{sZ_i}], \end{aligned} \tag{1.5}$$

where in the last equality we have used the independence of the Z_i.

We therefore need to control the term $\mathbb{E}[e^{sZ_i}]$, known as the *moment-generating function* of Z_i. If the Z_i were normally distributed, we could compute the moment-generating function analytically. The following lemma establishes that we can do something similar when the Z_i are bounded.

Lemma (Hoeffding's Lemma): If $Z \in [a,b]$ almost surely and $\mathbb{E}Z = 0$, then

$$\mathbb{E}e^{sZ} \leq e^{\frac{s^2(b-a)^2}{8}}.$$

Proof of Lemma. Consider the log-moment generating function $\psi(s) = \log \mathbb{E}[e^{sZ}]$, and note that it suffices to show that $\psi(s) \leq s^2(b-a)^2/8$. We will investigate ψ by computing the first several terms of its Taylor expansion. Standard regularity conditions imply that we can interchange the order of differentiation and integration to obtain

$$\psi'(s) = \frac{\mathbb{E}[Ze^{sZ}]}{\mathbb{E}[e^{sZ}]},$$

$$\psi''(s) = \frac{\mathbb{E}[Z^2e^{sZ}]\mathbb{E}[e^{sZ}] - \mathbb{E}[Ze^{sZ}]^2}{\mathbb{E}[e^{sZ}]^2} = \mathbb{E}\left[Z^2\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}\right] - \left(\mathbb{E}\left[Z\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}\right]\right)^2.$$

Since $\frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}$ integrates to 1, we can interpret $\psi''(s)$ as the variance of Z under the probability measure $d\mathbb{F} = \frac{e^{sZ}}{\mathbb{E}[e^{sZ}]}d\mathbb{E}$. We obtain

$$\psi''(s) = \mathrm{var}_{\mathbb{F}}(Z) = \mathrm{var}_{\mathbb{F}}\left(Z - \frac{a+b}{2}\right),$$

since the variance is unaffected under shifts. But $|Z - \frac{a+b}{2}| \leq \frac{b-a}{2}$ almost surely since $Z \in [a, b]$ almost surely, so

$$\mathrm{var}_{\mathbb{F}}\left(Z - \frac{a+b}{2}\right) \leq \mathbb{F}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}.$$

Finally, the fundamental theorem of calculus yields

$$\psi(s) = \int_0^s \int_0^u \psi''(u)\, du \leq \frac{s^2(b-a)^2}{8}.$$

This concludes the proof of the Lemma. □

Applying Hoeffding's Lemma to Equation (1.5), we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-stn}\prod e^{s^2/8} = e^{ns^2/8 - stn},$$

for any $s > 0$. Plugging in $s = 4t > 0$ yields

$$\mathbb{P}\left(\frac{1}{n}\sum Z_i > t\right) \leq e^{-2nt^2},$$

as desired. □

Hoeffding's Theorem implies that, for any classifier h, the bound

$$|\hat{R}_n(h) - R(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

holds with probability $1 - \delta$. We can immediately apply this formula to yield a maximal inequality: if $\mathcal{H}$ is a finite family, i.e., $\mathcal{H} = \{h_1, \dots, h_M\}$, then with probability $1 - \delta/M$ the bound

$$|\hat{R}_n(h_j) - R(h_j)| \leq \sqrt{\frac{\log(2M/\delta)}{2n}}$$

holds. The event that $\max_j |\hat{R}_n(h_j) - R(h_j)| > t$ is the union of the events $|\hat{R}_n(h_j) - R(h_j)| > t$ for $j = 1, \ldots, M$, so the union bound immediately implies that

$$\max_j |\hat{R}_n(h_j) - R(h_j)| \leq \sqrt{\frac{\log(2M/\delta)}{2n}}$$

with probability $1 - \delta$. In other words, for such a family, we can be assured that the empirical risk and the true risk are close. Moreover, the logarithmic dependence on M implies that we can increase the size of the family $\mathcal{H}$ exponentially quickly with n and maintain the same guarantees on our estimate.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet | Lecture 3
Scribe: James Hirst | Sep. 16, 2015

1.5 Learning with a finite dictionary

Recall from the end of last lecture our setup: We are working with a finite dictionary $\mathcal{H} = \{h_1, \dots, h_M\}$ of estimators, and we would like to understand the scaling of this problem with respect to M and the sample size n. Given $\mathcal{H}$, one idea is to simply try to minimize the empirical risk based on the samples, and so we define the empirical risk minimizer, $\hat{h}^{\text{erm}}$, by

$$\hat{h}^{\text{erm}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h).$$

In what follows, we will simply write $\hat{h}$ instead of $\hat{h}^{\text{erm}}$ when possible. Also recall the definition of the oracle, $\bar{h}$, which (somehow) minimizes the true risk and is defined by

$$\bar{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

The following theorem shows that, although $\hat{h}$ cannot hope to do better than $\bar{h}$ in general, the difference should not be too large as long as the sample size is not too small compared to M.

Theorem: The estimator $\hat{h}$ satisfies

$$R(\hat{h}) \leq R(\bar{h}) + \sqrt{\frac{2\log(2M/\delta)}{n}}$$

with probability at least $1 - \delta$. In expectation, it holds that

$$\mathbb{E}[R(\hat{h})] \leq R(\bar{h}) + \sqrt{\frac{2\log(2M)}{n}}.$$

Proof. From the definition of $\hat{h}$, we have $\hat{R}_n(\hat{h}) \leq \hat{R}_n(\bar{h})$, which gives

$$R(\hat{h}) \leq R(\bar{h}) + [\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})].$$

The only term here that we need to control is the second one, but since we don't have any real information about $\bar{h}$, we will bound it by a maximum over $\mathcal{H}$ and then apply Hoeffding:

$$[\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})] \leq 2\max_j |\hat{R}_n(h_j) - R(h_j)| \leq 2\sqrt{\frac{\log(2M/\delta)}{2n}}$$

with probability at least $1 - \delta$, which completes the first part of the proof.

To obtain the bound in expectation, we start with a standard trick from probability which bounds a max by its sum in a slightly more clever way. Here, let $\{Z_j\}_j$ be centered random variables, then

$$\mathbb{E}\left[\max_j |Z_j|\right] = \frac{1}{s}\log\exp\left(s\mathbb{E}\left[\max_j |Z_j|\right]\right) \leq \frac{1}{s}\log\mathbb{E}\left[\exp\left(s\max_j |Z_j|\right)\right],$$

where the last inequality comes from applying Jensen's inequality to the convex function $\exp(\cdot)$. Now we bound the max by a sum to get

$$\leq \frac{1}{s}\log\sum_{j=1}^{2M}\mathbb{E}\left[\exp(sZ_j)\right] \leq \frac{1}{s}\log\left(2M\exp\left(\frac{s^2}{8n}\right)\right) = \frac{\log(2M)}{s} + \frac{s}{8n},$$

where we used $Z_j = \hat{R}_n(h_j) - R(h_j)$ in our case and then applied Hoeffding's Lemma. Balancing terms by minimizing over s, this gives $s = 2\sqrt{2n\log(2M)}$ and plugging in produces

$$\mathbb{E}\left[\max_j |\hat{R}_n(h_j) - R(h_j)|\right] \leq \sqrt{\frac{\log(2M)}{2n}},$$

which finishes the proof. □

2. CONCENTRATION INEQUALITIES

Concentration inequalities are results that allow us to bound the deviations of a function of random variables from its average. The first of these we will consider is a direct improvement to Hoeffding's Inequality that allows some dependence between the random variables.

2.1 Azuma-Hoeffding Inequality

Given a filtration $\{\mathcal{F}_i\}_i$ of our underlying space $\mathcal{X}$, recall that $\{\Delta_i\}_i$ are called *martingale differences* if, for every i, it holds that $\Delta_i \in \mathcal{F}_i$ and $\mathbb{E}\left[\Delta_i|\mathcal{F}_i\right] = 0$. The following theorem gives a very useful concentration bound for averages of bounded martingale differences.

Theorem (Azuma-Hoeffding): Suppose that $\{\Delta_i\}_i$ are margingale differences with respect to the filtration $\{\mathcal{F}_i\}_i$, and let $A_i, B_i \in \mathcal{F}_{i-1}$ satisfy $A_i \leq \Delta_i \leq B_i$ almost surely for every i. Then

$$\mathbb{P}\left[\frac{1}{n}\sum_i \Delta_i > t\right] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n \|B_i - A_i\|_\infty^2}\right).$$

In comparison to Hoeffding's inequality, Azuma-Hoeffding affords not only the use of non-uniform boundedness, but additionally requires no independence of the random variables.

Proof. We start with a typical Chernoff bound.

$$\mathbb{P}\left[\sum_i \Delta_i > t\right] \leq \mathbb{E}\left[e^{s\sum\Delta_i}\right]e^{-st} = \mathbb{E}\left[\mathbb{E}\left[e^{s\sum\Delta_i}|\mathcal{F}_{n-1}\right]\right]e^{-st}$$

$$= \mathbb{E}\left[e^{s\sum^{n-1}\Delta_i}\mathbb{E}[e^{s\Delta_n}|\mathcal{F}_{n-1}]\right]e^{-st} \leq \mathbb{E}[e^{s\sum^{n-1}\Delta_i}\cdot e^{s^2(B_n-A_n)^2/8}]e^{-st},$$

where we have used the fact that the Δ_i, $i < n$, are all $\mathcal{F}_n$ measureable, and then applied Hoeffding's lemma on the inner expectation. Iteratively isolating each Δ_i like this and applying Hoeffding's lemma, we get

$$\mathbb{P}\left[\sum_i \Delta_i > t\right] \leq \exp\left(\frac{s^2}{8}\sum_{i=1}^n \|B_i - A_i\|_\infty^2\right)e^{-st}.$$

Optimizing over s as usual then gives the result. □

2.2 Bounded Differences Inequality

Although Azuma-Hoeffding is a powerful result, its full generality is often wasted and can be cumbersome to apply to a given problem. Fortunately, there is a natural choice of the $\{\mathcal{F}_i\}_i$ and $\{\Delta_i\}_i$, giving a similarly strong result which can be much easier to apply. Before we get to this, we need one definition.

Definition (Bounded Differences Condition): Let $g : \mathcal{X} \to \mathbb{R}$ and constants c_i be given. Then g is said to satisfy the bounded differences condition (with constants c_i) if

$$\sup_{x_1,\ldots,x_n,x_i'} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_i',\ldots,x_n)| \leq c_i$$

for every i.

Intuitively, g satisfies the bounded differences condition if changing only one coordinate of g at a time cannot make the value of g deviate too far. It should not be too surprising that these types of functions thus concentrate somewhat strongly around their average, and this intuition is made precise by the following theorem.

Theorem (Bounded Differences Inequality): If $g : \mathcal{X} \to \mathbb{R}$ satisfies the bounded differences condition, then

$$\mathbb{P}\left[|g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)| > t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_i c_i^2}\right).$$

Proof. Let $\{\mathcal{F}_i\}_i$ be given by $\mathcal{F}_i = \sigma(X_1,\ldots,X_i)$, and define the martingale differences $\{\Delta_i\}_i$ by

$$\Delta_i = \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_i\right] - \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right].$$

Then

$$\mathbb{P}\left[|\sum_i \Delta_i| > t\right] = \mathbb{P}\left[\left|g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)\right| > t\right],$$

exactly the quantity we want to bound. Now, note that

$$\Delta_i \leq \mathbb{E}\left[\sup_{x_i} g(X_1,\ldots,x_i,\ldots,X_n)|\mathcal{F}_i\right] - \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right]$$

$$= \mathbb{E}\left[\sup_{x_i} g(X_1,\ldots,x_i,\ldots,X_n) - g(X_1,\ldots,X_n) | \mathcal{F}_{i-1}\right] =: B_i.$$

Similarly,

$$\Delta_i \geq \mathbb{E}\left[\inf_{x_i} g(X_1,\ldots,x_i,\ldots,X_n) - g(X_1,\ldots,X_n) | \mathcal{F}_{i-1}\right] =: A_i.$$

At this point, our assumption on g implies that $\|B_i - A_i\|_\infty \leq c_i$ for every i, and since $A_i \leq \Delta_i \leq B_i$ with $A_i, B_i \in \mathcal{F}_{i-1}$, an application of Azuma-Hoeffding gives the result.

□

2.3 Bernstein's Inequality

Hoeffding's inequality is certainly a powerful concentration inequality for how little it assumes about the random variables. However, one of the major limitations of Hoeffding is just this: Since it only assumes boundedness of the random variables, it is completely oblivious to their actual variances. When the random variables in question have some known variance, an ideal concentration inequality should capture the idea that variance controls concentration to some degree. Bernstein's inequality does exactly this.

Theorem (Bernstein's Inequality): Let $X_1,\ldots,X_n$ be independent, centered random variables with $|X_i| \leq c$ for every i, and write $\sigma^2 = n^{-1}\sum_i \mathrm{Var}(X_i)$ for the average variance. Then

$$\mathbb{P}\left[\frac{1}{n}\sum_i X_i > t\right] \leq \exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}tc}\right).$$

Here, one should think of t as being fixed and relatively small compared to n, so that strength of the inequality indeed depends mostly on n and $1/\sigma^2$.

Proof. The idea of the proof is to do a Chernoff bound as usual, but to first use our assumptions on the variance to obtain a slightly better bound on the moment generating functions. To this end, we expand

$$\mathbb{E}[e^{sX_i}] = 1 + \mathbb{E}[sX_i] + \mathbb{E}\left[\sum_{k=2}^\infty \frac{(sX_i)^k}{k!}\right] \leq 1 + \mathrm{Var}(X_i)\sum_{k=2}^\infty \frac{s^k c^{k-2}}{k!},$$

where we have used $\mathbb{E}[X_i^k] \leq \mathbb{E}[X_i^2 |X_i|^{k-2}] \leq \mathrm{Var}(X_i)c^{k-2}$. Rewriting the sum as an exponential, we get

$$\mathbb{E}[e^{sX_i}] \leq s^2 \mathrm{Var}(X_i) g(s), \quad g(s) := \frac{e^{sc} - sc - 1}{c^2 s^2}.$$

The Chernoff bound now gives

$$\mathbb{P}\left[\frac{1}{n}\sum_i X_i > t\right] \leq \exp\left(\inf_{s>0}[s^2(\sum_i \mathrm{Var}(X_i))g(s) - nst]\right) = \exp\left(n \cdot \inf_{s>0}[s^2\sigma^2 g(s) - st]\right),$$

and optimizing this over s (a fun calculus exercise) gives exactly the desired result. □

3. NOISE CONDITIONS AND FAST RATES

To measure the effectiveness of the estimator $\hat{h}$, we would like to obtain an upper bound on the excess risk $\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h^*)$. It should be clear, however, that this must depend significantly on the amount of noise that we allow. In particular, if $\eta(X)$ is identically equal to $1/2$, then we should not expect to be able to say anything meaningful about $\mathcal{E}(\hat{h})$ in general. Understanding this trade-off between noise and rates will be the main subject of this chapter.

3.1 The Noiseless Case

A natural (albeit somewhat naïve) case to examine is the completely noiseless case. Here, we will have $\eta(X) \in \{0,1\}$ everywhere, $\mathrm{Var}(Y|X) = 0$, and

$$\mathcal{E}(h) = R(h) - R(h^*) = \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(h(X) \neq h^*(X))] = \mathbb{P}[h(X) \neq h^*(X)].$$

Let us now denote

$$Z_i = \mathbb{1}(\bar{h}(X_i) \neq Y_i) - \mathbb{1}(\hat{h}(X_i) \neq Y_i),$$

and write $\bar{Z}_i = Z_i - \mathbb{E}[Z_i]$. Then notice that we have

$$|Z_i| = \mathbb{1}(\hat{h}(X_i) \neq \bar{h}(X_i)),$$

and

$$\mathrm{Var}(Z_i) \leq \mathbb{E}[Z_i^2] = \mathbb{P}[\hat{h}(X_i) \neq \bar{h}(X_i)].$$

For any classifier $h_j \in \mathcal{H}$, we can similarly define $Z_i(h_j)$ (by replacing $\hat{h}$ with h_j throughout). Then, to set up an application of Bernstein's inequality, we can compute

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(Z_i(h_j)) \leq \mathbb{P}[h_j(X_i) \neq \bar{h}(X_i)] =: \sigma_j^2.$$

At this point, we will make a (fairly strong) assumption about our dictionary $\mathcal{H}$, which is that $h^* \in \mathcal{H}$, which further implies that $\bar{h} = h^*$. Since the random variables Z_i compare to $\bar{h}$, this will allow us to use them to bound $\mathcal{E}(\hat{h})$, which rather compares to h^*. Now, applying Bernstein (with $c = 2$) to the $\{\bar{Z}_i(h_j)\}_i$ for every j gives

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \bar{Z}_i(h_j) > t\right] \leq \exp\left(-\frac{nt^2}{2\sigma_j^2 + \frac{4}{3}t}\right) =: \frac{\delta}{M},$$

and a simple computation here shows that it is enough to take

$$t \geq \max\left(\sqrt{\frac{2\sigma_j^2 \log(M/\delta)}{n}}, \frac{4}{3n}\log(M/\delta)\right) =: t_0(j)$$

for this to hold. From here, we may use the assumption $\bar{h} = h^*$ to conclude that

$$\mathbb{P}\left[\mathcal{E}(\hat{h}) > t_0(\hat{j})\right] \leq \delta, \quad h_{\hat{j}} = \hat{h}.$$

However, we also know that $\sigma_{\hat{j}}^2 \leq \mathcal{E}(\hat{h})$, which implies that

$$\mathcal{E}(\hat{h}) \leq \max\left(\sqrt{\frac{2\mathcal{E}(\hat{h})\log(M/\delta)}{n}}, \frac{4}{3n}\log(M/\delta)\right)$$

with probability $1-\delta$, and solving for $\mathcal{E}(\hat{h})$ gives the improved rate

$$\mathcal{E}(\hat{h}) \leq 2\frac{\log(M/\delta)}{n}\,.$$

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 4
Scribe: Cheng Mao — Sep. 21, 2015

In this lecture, we continue to discuss the effect of noise on the rate of the excess risk $\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h^*)$ where $\hat{h}$ is the empirical risk minimizer. In the binary classification model, noise roughly means how close the regression function η is from $\frac{1}{2}$. In particular, if $\eta = \frac{1}{2}$ then we observe only noise, and if $\eta \in \{0, 1\}$ we are in the noiseless case which has been studied last time. Especially, we achieved the fast rate $\frac{\log M}{n}$ in the noiseless case by assuming $h^* \in \mathcal{H}$ which implies that $\bar{h} = h^*$. This assumption was essential for the proof and we will see why it is necessary again in the following section.

3.2 Noise conditions

The noiseless assumption is rather unrealistic, so it is natural to ask what the rate of excess risk is when the noise is present but can be controlled. Instead of the condition $\eta \in \{0, 1\}$, we can control the noise by assuming that η is uniformly bounded away from $\frac{1}{2}$, which is the motivation of the following definition.

Definition (Massart's noise condition): The noise in binary classification is said to satisfy Massart's condition with constant $\gamma \in (0, \frac{1}{2}]$ if $|\eta(X) - \frac{1}{2}| \geq \gamma$ almost surely.

Once uniform boundedness is assumed, the fast rate simply follows from last proof with appropriate modification of constants.

Theorem: Let $cE(\hat{h})$ denote the excess risk of the empirical risk minimizer $\hat{h} = \hat{h}^{\text{erm}}$. If Massart's noise condition is satisfied with constant γ, then

$$\mathcal{E}(\hat{h}) \leq \frac{\log(M/\delta)}{\gamma n}$$

with probability at least $1 - \delta$. (In particular $\gamma = \frac{1}{2}$ gives exactly the noiseless case.)

Proof. Define $Z_i(h) = \mathbb{1}(\bar{h}(X_i) \neq Y_i) - \mathbb{1}(h(X_i) \neq Y_i)$. By the assumption $\bar{h} = h^*$ and the definition of $\hat{h} = \hat{h}^{\text{erm}}$,

$$\begin{aligned}
\mathcal{E}(\hat{h}) &= R(\hat{h}) - R(\bar{h}) \\
&= \hat{R}_n(\hat{h}) - \hat{R}_n(\bar{h}) + \hat{R}_n(\bar{h}) - \hat{R}_n(\hat{h}) - \big(R(\bar{h}) - R(\hat{h})\big) \qquad (3.1) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\big(Z_i(\hat{h}) - \mathbb{E}[Z_i(\hat{h})]\big). \qquad (3.2)
\end{aligned}$$

Hence it suffices to bound the deviation of $\sum_i Z_i$ from its expectation. To this end, we hope to apply Bernstein's inequality. Since

$$\mathrm{Var}[Z_i(h)] \leq \mathbb{E}[Z_i(h)^2] = \mathbb{P}[h(X_i) \neq \bar{h}(X_i)],$$

we have that for any $1 \leq j \leq M$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[Z_i(h_j)] \leq \mathbb{P}[h_j(X) \neq \bar{h}(X)] =: \sigma_j^2.$$

Bernstein's inequality implies that

$$\mathbb{P}\big[\frac{1}{n}\sum_{i=1}^{n}(Z_i(h_j) - \mathbb{E}[Z_i(h_j)]) > t\big] \leq \exp\big(-\frac{nt^2}{2\sigma_j^2 + \frac{2}{3}t}\big) =: \frac{\delta}{M}.$$

Applying a union bound over $1 \leq j \leq M$ and taking

$$t = t_0(j) := \max\big(\sqrt{\frac{2\sigma_j^2 \log(M/\delta)}{n}}, \frac{2\log(M/\delta)}{3n}\big),$$

we get that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i(h_j) - \mathbb{E}[Z_i(h_j)]) \leq t_0(j) \tag{3.3}$$

for all $1 \leq j \leq M$ with probability at least $1 - \delta$.

Suppose $\hat{h} = h_{\hat{j}}$. It follows from (3.2) and (3.3) that with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{h}) \leq t_0(\hat{j}).$$

(Note that so far the proof is exactly the same as the noiseless case.) Since $|\eta(X) - \frac{1}{2}| \geq \gamma$ a.s. and $\bar{h} = h^*$,

$$\mathcal{E}(\hat{h}) = \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(\hat{h}(X) \neq h^*(X))] \geq 2\gamma\mathbb{P}[h_{\hat{j}}(X) \neq \bar{h}(X)] = 2\gamma\sigma_{\hat{j}}^2.$$

Therefore,

$$\mathcal{E}(\hat{h}) \leq \max\big(\sqrt{\frac{\mathcal{E}(\hat{h})\log(M/\delta)}{\gamma n}}, \frac{2\log(M/\delta)}{3n}\big), \tag{3.4}$$

so we conclude that with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{h}) \leq \frac{\log(M/\delta)}{\gamma n}.$$

□

The assumption that $\bar{h} = h^*$ was used twice in the proof. First it enables us to ignore the approximation error and only study the stochastic error. More importantly, it makes the excess risk appear on the right-hand side of (3.4) so that we can rearrange the excess risk to get the fast rate.

Massart's noise condition is still somewhat strong because it assumes uniform boundedness of η from $\frac{1}{2}$. Instead, we can allow η to be close to $\frac{1}{2}$ but only with small probability, and this is the content of next definition.

Definition (Tsybakov's noise condition or Mammen-Tsybakov noise condition): The noise in binary classification is said to satisfy Tsybakov's condition if there exists $\alpha \in (0,1)$, $C_0 > 0$ and $t_0 \in (0, \frac{1}{2}]$ such that

$$\mathbb{P}[|\eta(X) - \frac{1}{2}| \le t] \le C_0 t^{\frac{\alpha}{1-\alpha}}$$

for all $t \in [0, t_0]$.

In particular, as $\alpha \to 1$, $t^{\frac{\alpha}{1-\alpha}} \to 0$, so this recovers Massart's condition with $\gamma = t_0$ and we have the fast rate. As $\alpha \to 0$, $t^{\frac{\alpha}{1-\alpha}} \to 1$, so the condition is void and we have the slow rate. In between, it is natural to expect fast rate (meaning faster than slow rate) whose order depends on α. We will see that this is indeed the case.

Lemma: Under Tsybakov's noise condition with constants α, C_0 and t_0, we have

$$\mathbb{P}[h(X) \neq h^*(X)] \le C\mathcal{E}(h)^\alpha$$

for any classifier h where $C = C(\alpha, C_0, t_0)$ is a constant.

Proof. We have

$$\begin{aligned}
\mathcal{E}(h) &= \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(h(X) \neq h^*(X))] \\
&\ge \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(|\eta(X) - \frac{1}{2}| > t)\mathbb{1}(h(X) \neq h^*(X))] \\
&\ge 2t\mathbb{P}[|\eta(X) - \frac{1}{2}| > t, h(X) \neq h^*(X)] \\
&\ge 2t\mathbb{P}[h(X) \neq h^*(X)] - 2t\mathbb{P}[|\eta(X) - \frac{1}{2}| \le t] \\
&\ge 2t\mathbb{P}[h(X) \neq h^*(X)] - 2C_0 t^{\frac{1}{1-\alpha}}
\end{aligned}$$

where Tsybakov's condition was used in the last step. Take $t = c\mathbb{P}[h(X) \neq h^*(X)]^{\frac{1-\alpha}{\alpha}}$ for some positive $c = c(\alpha, C_0, t_0)$ to be chosen later. We assume that $c \le t_0$ to guarantee that $t \in [0, t_0]$. Since $\alpha \in (0,1)$,

$$\begin{aligned}
\mathcal{E}(h) &\ge 2c\mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha} - 2C_0 c^{\frac{1}{1-\alpha}}\mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha} \\
&\ge c\mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha}
\end{aligned}$$

by selecting c sufficiently small depending on α and C_0. Therefore

$$\mathbb{P}[h(X) \neq h^*(X)] \le \frac{1}{c^\alpha}\mathcal{E}(h)^\alpha$$

and choosing $C = C(\alpha, C_0, t_0) := c^{-\alpha}$ completes the proof. □

Having established the key lemma, we are ready to prove the promised fast rate under Tsybakov's noise condition.

Theorem: If Tsybakov's noise condition is satisfied with constant α, C_0 and t_0, then there exists a constant $C = C(\alpha, C_0, t_0)$ such that

$$\mathcal{E}(\hat{h}) \leq C\big(\frac{\log(M/\delta)}{n}\big)^{\frac{1}{2-\alpha}}$$

with probability at least $1 - \delta$.

This rate of excess risk parametrized by α is indeed an interpolation of the slow ($\alpha \to 0$) and the fast rate ($\alpha \to 1$). Futhermore, note that the empirical risk minimizer $\hat{h}$ does not depend on the parameter α at all! It automatically adjusts to the noise level, which is a very nice feature of the empirical risk minimizer.

Proof. The majority of last proof remains valid and we will explain the difference. After establishing that

$$\mathcal{E}(\hat{h}) \leq t_0(\hat{\jmath}),$$

we note that the lemma gives

$$\sigma_{\hat{\jmath}}^2 = \mathbb{P}[\hat{h}(X) \neq \bar{h}(X)] \leq C\mathcal{E}(\hat{h})^\alpha.$$

It follows that

$$\mathcal{E}(\hat{h}) \leq \max\big(\sqrt{\frac{2C\mathcal{E}(\hat{h})^\alpha \log(M/\delta)}{n}}, \frac{2\log(M/\delta)}{3n}\big)$$

and thus

$$\mathcal{E}(\hat{h}) \leq \max\Big(\big(\frac{2C\log\frac{M}{\delta}}{n}\big)^{\frac{1}{2-\alpha}}, \frac{2\log(M/\delta)}{3n}\Big).$$

□

4. VAPNIK-CHERVONENKIS (VC) THEORY

The upper bounds proved so far are meaningful only for a finite dictionary $\mathcal{H}$, because if $M = |\mathcal{H}|$ is infinite all of the bounds we have will simply be infinity. To extend previous results to the infinite case, we essentially need the condition that only a finite number of elements in an infinite dictionary $\mathcal{H}$ really matter. This is the objective of the Vapnik-Chervonenkis (VC) theory which was developed in 1971.

4.1 Empirical measure

Recall from previous proofs (see (3.1) for example) that the key quantity we need to control is

$$2\sup_{h\in\mathcal{H}}\big(\hat{R}_n(h) - R(h)\big).$$

Instead of the union bound which would not work in the infinite case, we seek some bound that potentially depends on n and the complexity of the set $\mathcal{H}$. One approach is to consider some metric structure on $\mathcal{H}$ and hope that if two elements in $\mathcal{H}$ are close, then the quantity evaluated at these two elements are also close. On the other hand, the VC theory is more combinatorial and does not involve any metric space structure as we will see.

By definition

$$\hat{R}_n(h) - R(h) = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{I}(h(X_i) \neq Y_i) - \mathbb{E}[\mathbb{I}(h(X_i) \neq Y_i)]\big).$$

Let $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$, and let $\mathcal{A}$ denote the class of measurable sets in the sample space $\mathcal{X} \times \{0, 1\}$. For a classifier h, define $A_h \in \mathcal{A}$ by

$$\{Z_i \in A_h\} = \{h(X_i) \neq Y_i\}.$$

Moreover, define measures μ_n and μ on $\mathcal{A}$ by

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Z_i \in A) \quad \text{and} \quad \mu(A) = \mathbb{P}[Z_i \in A]$$

for $A \in \mathcal{A}$. With this notation, the slow rate we proved is just

$$\sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h) = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{\log(2|A|/\delta)}{2n}}.$$

Since this is not accessible in the infinite case, we hope to use one of the concentration inequalities to give an upper bound. Note that $\mu_n(A)$ is a sum of random variables that may not be independent, so the only tool we can use now is the bounded difference inequality.

If we change the value of only one z_i in the function

$$z_1, \dots, z_n \mapsto \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|,$$

the value of the function will differ by at most $1/n$. Hence it satisfies the bounded difference assumption with $c_i = 1/n$ for all $1 \leq i \leq n$. Applying the bounded difference inequality, we get that

$$\Big|\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|]\Big| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1 - \delta$. Note that this already precludes any fast rate (faster than $n^{-1/2}$). To achieve fast rate, we need Talagrand inequality and localization techniques which are beyond the scope of this section.

It follows that with probability at least $1 - \delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

We will now focus on bounding the first term on the right-hand side. To this end, we need a technique called symmetrization, which is the subject of the next section.

4.2 Symmetrization and Rademacher complexity

Symmetrization is a frequently used technique in machine learning. Let $\mathcal{D} = \{Z_1, \dots, Z_n\}$ be the sample set. To employ symmetrization, we take another independent copy of the sample set $\mathcal{D}' = \{Z'_1, \dots, Z'_n\}$. This sample only exists for the proof, so it is sometimes referred to as a ghost sample. Then we have

$$\mu(A) = \mathbb{P}[Z \in A] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Z'_i \in A)] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Z'_i \in A)|\mathcal{D}] = \mathbb{E}[\mu'_n(A)|\mathcal{D}]$$

where $\mu'_n := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z'_i \in A)$. Thus by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &= \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\mu_n(A) - \mathbb{E}[\mu'_n(A)|\mathcal{D}]\big|\big] \\
&\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} \mathbb{E}[|\mu_n(A) - \mu'_n(A)|\,|\mathcal{D}]\big] \\
&\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)|\big] \\
&= \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z'_i \in A)\big)\big|\big].
\end{aligned}$$

Since $\mathcal{D}'$ has the same distribution of $\mathcal{D}$, by symmetry $\mathbb{1}(Z_i \in A) - \mathbb{1}(Z'_i \in A)$ has the same distribution as $\sigma_i\big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z'_i \in A)\big)$ where $\sigma_1, \dots, \sigma_n$ are i.i.d. Rad($\frac{1}{2}$), i.e.

$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2},$$

and σ_i's are taken to be independent of both samples. Therefore,

$$\begin{aligned}
\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i\big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z'_i \in A)\big)\big|\big] \\
&\le 2\mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{1}(Z_i \in A)\big|\big]. \qquad (4.5)
\end{aligned}$$

Using symmetrization we have bounded $\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu(A)|]$ by a much nicer quantity. Yet we still need an upper bound of the last quantity that depends only on the structure of $\mathcal{A}$ but not on the random sample $\{Z_i\}$. This is achieved by taking the supremum over all $z_i \in \mathcal{X} \times \{0,1\} =: \mathcal{Y}$.

Definition: The Rademacher complexity of a family of sets $\mathcal{A}$ in a space $\mathcal{Y}$ is defined to be the quantity

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\dots,z_n\in\mathcal{Y}} \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{1}(z_i \in A)\big|\big].$$

The Rademacher complexity of a set $B \subset \mathbb{R}^n$ is defined to be

$$\mathcal{R}_n(B) = \mathbb{E}\big[\sup_{b\in B} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\big|\big].$$

We conclude from (4.5) and the definition that

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \le 2\mathcal{R}_n(\mathcal{A}).$$

In the definition of Rademacher complexity of a set, the quantity $\big|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\big|$ measures how well a vector $b \in B$ correlates with a random sign pattern $\{\sigma_i\}$. The more complex B is, the better some vector in B can replicate a sign pattern. In particular, if B is the full hypercube $[-1,1]^n$, then $\mathcal{R}_n(B) = 1$. However, if $B \subset [-1,1]^n$ contains only k-sparse vectors, then $\mathcal{R}_n(B) = k/n$. Hence $\mathcal{R}_n(B)$ is indeed a measurement of the complexity of the set B.

The set of vectors to our interest in the definition of Rademacher complexity of $\mathcal{A}$ is

$$T(z) := \{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A))^T, A \in \mathcal{A}\}.$$

Thus the key quantity here is the cardinality of $T(z)$, i.e., the number of sign patterns these vectors can replicate as A ranges over $\mathcal{A}$. Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is bounded by 2^n.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 5
Scribe: Vira Semenova and Philippe Rigollet Sep. 23, 2015

In this lecture, we complete the analysis of the performance of the empirical risk minimizer under a constraint on the VC dimension of the family of classifiers. To that end, we will see how to control Rademacher complexities using shatter coefficients. Moreover, we will see how the problem of controlling uniform deviations of the empirical measure μ_n from the true measure μ as done by Vapnik and Chervonenkis relates to our original classification problem.

4.1 Shattering

Recall from the previous lecture that we are interested in sets of the form

$$T(z) := \big\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\big\}, \quad z = (z_1, \dots, z_n). \tag{4.1}$$

In particular, the cardinality of $T(z)$, i.e., the number of binary patterns these vectors can replicate as A ranges over $\mathcal{A}$, will be of critical importance, as it will arise when controlling the Rademacher complexity. Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is always at most 2^n. When it is of the size 2^n, we say that $\mathcal{A}$ *shatters* the set $z_1, \dots, z_n$. Formally, we have the following definition.

Definition: A collection of sets $\mathcal{A}$ *shatters* the set of points $\{z_1, z_2, ..., z_n\}$

$$\text{card}\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\} = 2^n .$$

The sets of points $\{z_1, z_2, ..., z_n\}$ that we are interested are realizations of the pairs $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ and may, in principle take any value over the sample space. Therefore, we define the *shatter coefficient* to be the largest cardinality that we may obtain.

Definition: The *shatter coefficients* of a class of sets $\mathcal{A}$ is the sequence of numbers $\{\mathcal{S}_{\mathcal{A}}(n)\}_{n\geq 1}$, where for any $n \geq 1$,

$$\mathcal{S}_{\mathcal{A}}(n) = \sup_{z_1,\dots,z_n} \text{card}\{(\mathbb{1}(z_1 \in A), \dots, \mathbb{1}(z_n \in A)), A \in \mathcal{A}\}$$

and the suprema are taken over the whole sample space.

By definition, the nth shatter coefficient $\mathcal{S}_{\mathcal{A}}(n)$ is equal to 2^n if there exists a set $\{z_1, z_2, ..., z_n\}$ that $\mathcal{A}$ shatters. The largest of such sets is precisely the Vapnik-Chervonenkis or VC dimension.

Definition: The Vapnik-Chervonenkis dimension, or *VC-dimension* of $\mathcal{A}$ is the largest integer d such that $\mathcal{S}_{\mathcal{A}}(d) = 2^d$. We write $\mathsf{VC}(\mathcal{A}) = d$.

If $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for all positive integers n, then $\mathsf{VC}(\mathcal{A}) := \infty$

In words, $\mathcal{A}$ shatters *some* set of points of cardinality d but shatters *no* set of points of cardinality $d+1$. In particular, $\mathcal{A}$ also shatters no set of points of cardinality $d' > d$ so that the VC dimension is well defined.

In the sequel, we will see that the VC dimension will play the role similar to of cardinality, but on an exponential scale. For interesting classes $\mathcal{A}$ such that $\mathrm{card}(\mathcal{A}) = \infty$, we also may have $\mathsf{VC}(\mathcal{A}) < \infty$. For example, assume that $\mathcal{A}$ is the class of *half-lines*, $\mathcal{A} = \{(-\infty, a], a \in \mathbb{R}\} \cup \{[a, \infty), a \in \mathbb{R}\}$, which is clearly infinite. Then, we can clearly shatter a set of size 2 but we for three points $z_1, z_2, z_3, \in \mathbb{R}$, if for example $z_1 < z_2 < z_3$, we cannot create the pattern $(0, 1, 0)$ (see Figure 4.1). Indeed, half lines can can only create patterns with zeros followed by ones or with ones followed by zeros but not an alternating pattern like $(0, 1, 0)$.

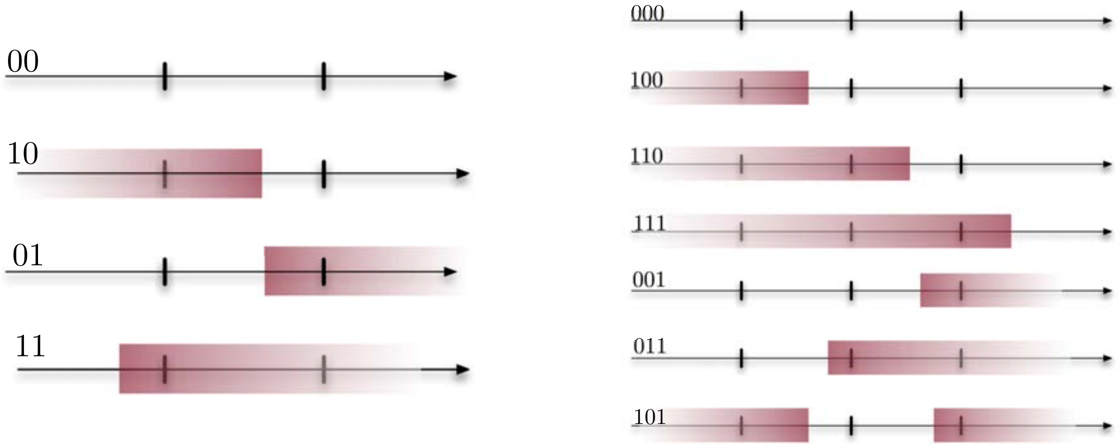


Figure 1: If $\mathcal{A} = \{\text{halflines}\}$, then any set of size $n = 2$ is shattered because we can create all $2^n = 4$ 0/1 patterns (left); if $n = 3$ the pattern $(0, 1, 0)$ cannot be reconstructed: $\mathcal{S}_A(3) = 7 < 2^3$ (right). Therefore, $\mathsf{VC}(\mathcal{A}) = 2$.

4.2 The VC inequality

We have now introduced all the ingredients necessary to state the main result of this section: the VC inequality.

Theorem (VC inequality): For any family of sets $\mathcal{A}$ with VC dimension $\mathsf{VC}(\mathcal{A}) = d$, it holds

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2d \log(2en/d)}{n}}$$

Note that this result holds even if $\mathcal{A}$ is infinite as long as its VC dimension is finite. Moreover, observe that $\log(|\mathcal{A}|)$ has been replaced by a term of order $d \log\big(2en/d\big)$.

To prove the VC inequality, we proceed in three steps:

1. Symmetrization, to bound the quantity of interest by the Rademacher complexity:
$$\mathbb{E}[\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|]\le 2\mathcal{R}_n(\mathcal{A}).$$
We have already done this step in the previous lecture.

2. Control of the Rademacher complexity using shatter coefficients. We are going to show that
$$\mathcal{R}_n(\mathcal{A})\le\sqrt{\frac{2\log\big(2\mathcal{S}_{\mathcal{A}}(n)\big)}{n}}$$

3. We are going to need the *Sauer-Shelah* lemma to bound the shatter coefficients by the VC dimension. It will yield
$$\mathcal{S}_{\mathcal{A}}(n)\le\Big(\frac{en}{d}\Big)^d,\quad d=\mathsf{VC}(\mathcal{A})\,.$$

Put together, these three steps yield the VC inequality.

Step 2: Control of the Rademacher complexity

We prove the following Lemma.

Lemma: For any $B\subset\mathbb{R}^n$, such that $|B|<\infty$:, it holds
$$\mathcal{R}_n(B)=\mathbb{E}\big[\max_{b\in B}\big|\frac{1}{n}\sum_{i=1}^n\sigma_ib_i\big|\big]\le\max_{b\in B}|b|_2\frac{\sqrt{2\log(2|B|)}}{n}$$
where $|\cdot|_2$ denotes the Euclidean norm.

Proof. Note that
$$\mathcal{R}_n(B)=\frac{1}{n}\mathbb{E}\big[\max_{b\in B}|Z_b|\big]\,,$$
where $Z_b=\sum_{i=1}^n\sigma_ib_i$. In particular, since $-|b_i|\le\sigma_i|b_i|\le|b_i|$, a.s., Hoeffding's lemma implies that the moment generating function of Z_b is controlled by
$$\mathbb{E}\big[\exp(sZ_b)\big]=\prod_{i=1}^n\mathbb{E}\big[\exp(s\sigma_ib_i)\big]\le\prod_{i=1}^n\exp(s^2b_i^2/2)=\exp(s^2|b|_2^2/2)\qquad(4.2)$$
Next, to control $\mathbb{E}\big[\max_{b\in B}|Z_b|\big]$, we use the same technique as in Lecture 3, section 1.5. To that end, define $\bar{B}=B\cup\{-B\}$ and observe that for any $s>0$,
$$\mathbb{E}\left[\max_{b\in B}|Z_b|\right]=\mathbb{E}\left[\max_{b\in\bar{B}}Z_b\right]=\frac{1}{s}\log\exp\left(s\mathbb{E}\left[\max_{b\in\bar{B}}Z_b\right]\right)\le\frac{1}{s}\log\mathbb{E}\left[\exp\left(s\max_{b\in\bar{B}}Z_b\right)\right],$$
where the last inequality follows from Jensen's inequality. Now we bound the max by a sum to get
$$\mathbb{E}\left[\max_{b\in B}|Z_b|\right]\le\frac{1}{s}\log\sum_{b\in\bar{B}}\mathbb{E}\left[\exp(sZ_b)\right]\le\frac{\log|\bar{B}|}{s}+\frac{s|b|_2^2}{2n},$$
where in the last inequality, we used (4.2). Optimizing over $s>0$ yields the desired result. □

We apply this result to our problem by observing that

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\ldots,z_n} \mathcal{R}_n(T(z))$$

where $T(z)$ is defined in (4.1). In particular, since $T(z) \subset \{0,1\}$, we have $|b|_2 \leq \sqrt{n}$ for all $b \in T(z)$. Moreover, by definition of the shatter coefficients, if $B = T(z)$, then $|\bar{B}| \leq 2|T(z)| \leq 2\mathcal{S}_{\mathcal{A}}(n)$. Together with the above lemma, it yields the desired inequality:

$$\mathcal{R}_n(\mathcal{A}) \leq \sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}}\,.$$

Step 3: Sauer-Shelah Lemma

We need to use a lemma from combinatorics to relate the shatter coefficients to the VC dimension. A priori, it is not clear from its definition that the VC dimension may be at all useful to get better bounds. Recall that steps 1 and 2 put together yield the following bound

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}} \tag{4.3}$$

In particular, if $\mathcal{S}_{\mathcal{A}}(n)$ is exponential in n, the bound (4.3) is not informative, i.e., it does not imply that the uniform deviations go to zero as the sample size n goes to infinity. The VC inequality suggest that this is not the case as soon as $\mathsf{VC}(\mathcal{A}) < \infty$ but it is not clear a priori. Indeed, it may be the case that $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for $n \leq d$ and $\mathcal{S}_{\mathcal{A}}(n) = 2^n - 1$ for $n > d$, which would imply that $\mathsf{VC}(\mathcal{A}) = d < \infty$ but that the right-hand side in (4.3) is larger than 2 for all n. It turns our that this can never be the case: if the VC dimension is finite, then the shatter coefficients are at most *polynomial* in n. This result is captured by the Sauer-Shelah lemma, whose proof is omitted. The reading section of the course contains pointers to various proofs, specifically the one based on *shifting* which is an important technique in enumerative combinatorics.

Lemma (Sauer-Shelah): If $\mathsf{VC}(\mathcal{A}) = d$, then $\forall n \geq 1$,

$$\mathcal{S}_{\mathcal{A}}(n) \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d.$$

Together with (4.3), it clearly yields the VC inequality. By applying the bounded difference inequality, we also obtain the following VC inequality that holds with high probability. This is often the preferred from for this inequality in the literature.

Corollary (VC inequality): For any family of sets $\mathcal{A}$ such that $\mathsf{VC}(\mathcal{A}) = d$ and any $\delta \in (0,1)$, it holds with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2d\log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}\,.$$

Note that the logarithmic term $\log(2en/d)$ is actually superfluous and can be replaced by a numerical constant using a more careful bounding technique. This is beyond the scope of this class and the interested reader should take a look at the recommending readings.

4.3 Application to ERM

The VC inequality provides an upper bound for $\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|$ in terms of the VC dimension of the class of sets $\mathcal{A}$. This result translates directly to our quantity of interest:

$$\sup_{h\in H}|\hat{R}_n(h)-R(h)| \le 2\sqrt{\frac{2\mathsf{VC}(\mathcal{A})\log\left(\frac{2en}{\mathsf{VC}(\mathcal{A})}\right)}{n}}+\sqrt{\frac{\log(2/\delta)}{2n}} \tag{4.4}$$

where $\mathcal{A}=\{A_h : h\in\mathcal{H}\}$ and $A_h=\{(x,y)\in\mathcal{X}\times\{0,1\} : h(x)\neq y\}$. Unfortunately, the VC dimension of this class of subsets of $\mathcal{X}\times\{0,1\}$ is not very natural. Since, a classifier h is a $\{0,1\}$ valued function, it is more natural to consider the VC dimension of the family $\mathcal{A}=\big\{\{h=1\} : h\in\mathcal{H}\big\}$.

Definition: Let $\mathcal{H}$ be a collection of classifiers and define

$$\bar{\mathcal{A}}=\big\{\{h=1\} : h\in\mathcal{H}\big\}=\big\{A : \exists\, h\in\mathcal{H}, h(\cdot)=\mathbb{1}(\cdot\in A)\big\}\,.$$

We define the VC dimension $\mathsf{VC}(\mathcal{H})$ of $\mathcal{H}$ to be the VC dimension of $\bar{\mathcal{A}}$.

It is not clear how $\mathsf{VC}(\bar{\mathcal{A}})$ relates to the quantity $\mathsf{VC}(\bar{\mathcal{A}})$, where $\mathcal{A}=\{A_h : h\in\mathcal{H}\}$ and $A_h=\{(x,y)\in\mathcal{X}\times\{0,1\} : h(x)\neq y\}$ that appears in the VC inequality. Fortunately, these two are actually equal as indicated in the following lemma.

Lemma: Define the two families for sets: $\mathcal{A}=\{A_h : h\in\mathcal{H}\}\in 2^{\mathcal{X}\times\{0,1\}}$ where $A_h=\{(x,y)\in\mathcal{X}\times\{0,1\} : h(x)\neq y\}$ and $\bar{\mathcal{A}}=\big\{\{h=1\} : h\in\mathcal{H}\big\}\in 2^{\mathcal{X}}$.
Then, $\mathcal{S}_{\bar{\mathcal{A}}}(n)=\mathcal{S}_{\bar{\mathcal{A}}}(n)$ for all $n\ge 1$. It implies $\mathsf{VC}(\bar{\mathcal{A}})=\mathsf{VC}(\mathcal{A})$.

Proof. Fix $x=(x_1,...,x_n)\in\mathcal{X}^n$ and $y=(y_1,y_2,...,y_n)\in\{0,1\}^n$ and define

$$T(x,y)=\{(\mathbb{1}(h(x_1)\neq y_1),\dots,\mathbb{1}(h(x_n)\neq y_n)), h\in\mathcal{H}\}$$

and

$$\bar{T}(x)=\{(\mathbb{1}(h(x_1)=1),\dots,\mathbb{1}(h(x_n)=1)), h\in\mathcal{H}\}$$

To that end, fix $v\in\{0,1\}$ and recall the XOR (exclusive OR) boolean function from $\{0,1\}$ to $\{0,1\}$ defined by $u\oplus v=\mathbb{1}(u\neq v)$. It is clearly[1] a bijection since $(u\oplus v)\oplus v=u$.

[1] One way to see that is to introduce the “spinned” variables $\tilde{u}=2u-1$ and $\tilde{v}=2v-1$ that live in $\{-1,1\}$. Then $\widetilde{u\oplus v}=\tilde{u}\cdot\tilde{v}$, and the claim follows by observing that $(\tilde{u}\cdot\tilde{v})\cdot\tilde{v}=\tilde{u}$. Another way is to simply write a truth table.

When applying XOR componentwise, we have

$$\begin{pmatrix} \mathbb{1}(h(x_1) \neq y_1) \\ \vdots \\ \mathbb{1}(h(x_i) \neq y_i) \\ \vdots \\ \mathbb{1}(h(x_n) \neq y_n) \end{pmatrix} = \begin{pmatrix} \mathbb{1}(h(x_1) = 1) \\ \vdots \\ \mathbb{1}(h(x_i) = 1) \\ \vdots \\ \mathbb{1}(h(x_n) = 1) \end{pmatrix} \oplus \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

Since XOR is a bijection, we must have $\operatorname{card}[T(x,y)] = \operatorname{card}[\bar{T}(x)]$. The lemma follows by taking the supremum on each side of the equality. □

It yields the following corollary to the VC inequality.

Corollary: Let $\mathcal{H}$ be a family of classifiers with VC dimension d. Then the empirical risk classifier $\hat{h}^{\text{erm}}$ over $\mathcal{H}$ satisfies

$$R(\hat{h}^{\text{erm}}) \leq \min_{h\in\mathcal{H}} R(h) + 4\sqrt{\frac{2d\log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1-\delta$.

Proof. Recall from Lecture 3 that

$$R(\hat{h}^{\text{erm}}) - \min_{h\in\mathcal{H}} R(h) \leq 2\sup_{h\in\mathcal{H}} \left|\hat{R}_n(h) - R(h)\right|$$

The proof follows directly by applying (4.4) and the above lemma. □

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 6
Scribe: Ali Makhdoumi — Sep. 28, 2015

5. LEARNING WITH A GENERAL LOSS FUNCTION

In the previous lectures we have focused on binary losses for the classification problem and developed VC theory for it. In particular, the risk for a classification function $h : \mathcal{X} \to \{0,1\}$ and binary loss function the risk was

$$R(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbb{1}(h(X) \neq Y)].$$

In this lecture we will consider a general loss function and a general regression model where Y is not necessarily a binary variable. For the binary classification problem, we then used the followings:

- Hoeffding's inequality: it requires boundedness of the loss functions.
- Bounded difference inequality: again it requires boundedness of the loss functions.
- VC theory: it requires binary nature of the loss function.

Limitations of the VC theory:

- Hard to find the optimal classification: the empirical risk minimization optimization, i.e.,
$$\min_h \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(X_i) \neq Y_i)$$
is a difficult optimization. Even though it is a hard optimization, there are some algorithms that try to optimize this function such as Perceptron and Adaboost.
- This is not suited for regression. We indeed know that classification problem is a subset of Regression problem as in regression the goal is to find $\mathbb{E}[Y|X]$ for a general Y (not necessarily binary).

In this section, we assume that $Y \in [-1,1]$ (this is not a limiting assumption as all the results can be derived for any bounded Y) and we have a regression problem where $(X,Y) \in \mathcal{X} \times [-1,1]$. Most of the results that we preset here are the analogous to the results we had in binary classification. This would be a good place to review those materials and we will refer to the techniques we have used in classification when needed.

5.1 Empirical Risk Minimization

5.1.1 Notations

Loss function: In binary classification the loss function was $\mathbb{1}(h(X) \neq Y)$. Here, we replace this loss function by $\ell(Y, f(X))$ which we assume is symmetric, where $f \in \mathcal{F}$, $f : \mathcal{X} \to [-1,1]$ is the regression functions. Examples of loss function include

- $\ell(a,b) = \mathbb{I}(a \neq b)$ (this is the classification loss function).
- $\ell(a,b) = |a-b|$.
- $\ell(a,b) = (a-b)^2$.
- $\ell(a,b) = |a-b|^p$, $p \geq 1$.

We further assume that $0 \leq \ell(a,b) \leq 1$.
Risk: risk is the expectation of the loss function, i.e.,

$$R(f) = \mathbb{E}_{X,Y}[\ell(Y, f(X))],$$

where the joint distribution is typically unknown and it must be learned from data.
Data: we observe a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent draws from a joint distribution $P_{X,Y}$, where $(X,Y) \in \mathcal{X} \times [-1,1]$. We denote the data points by $D_n = \{(X_1,Y_1), \ldots, (X_n,Y_n)\}$.
Empirical Risk: the empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)),$$

and the empirical risk minimizer denoted by $\hat{f}^{\text{erm}}$ (or $\hat{f}$) is defined as the minimizer of empirical risk, i.e.,

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \hat{R}_n(f).$$

In order to control the risk of $\hat{f}$ we shall compare its performance with the following oracle:

$$\bar{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, R(f).$$

Note that this is an oracle as in order to find it one need to have access to P_{XY} and then optimize $R(f)$ (we only observe the data D_n). Since $\hat{f}$ is the minimizer of the empirical risk minimizer, we have that $\hat{R}_n(\hat{f}) \leq \hat{R}_n(\bar{f})$, which leads to

$$\begin{aligned} R(\hat{f}) &\leq R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\hat{f}) - \hat{R}_n(\bar{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) + R(\bar{f}) \\ &\leq R(\bar{f}) + R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) \leq R(\bar{f}) + 2\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|. \end{aligned}$$

Therefore, the quantity of interest that we need to bound is

$$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|.$$

Moreover, from the bounded difference inequality, we know that since the loss function $\ell(\cdot,\cdot)$ is bounded by 1, $\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$ has the bounded difference property with $c_i = \frac{1}{n}$ for $i = 1, \ldots, n$, and the bounded difference inequality establishes

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)\,| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)\,|\right] \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right) = \exp\left(-2nt^2\right),$$

which in turn yields

$$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|\right] + \sqrt{\frac{\log\left(1/delta\right)}{2n}}, \text{ w.p. } 1-\delta.$$

As a result we only need to bound the expectation $\mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|]$.

5.1.2 Symmetrization and Rademacher Complexity

Similar to the binary loss case we first use symmetrization technique and then introduce Rademacher random variables. Let $D_n = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$ be the sample set and define an independent sample (ghost sample) with the same distribution denoted by $D'_n = \{(X'_1, Y'_1), \ldots (X'_n, Y'_n)\}$(for each i, (X'_i, Y'_i) is independent from D_n with the same distribution as of (X_i, Y_i)). Also, let $\sigma_i \in \{-1, +1\}$ be i.i.d. $\mathsf{Rad}(\frac{1}{2})$ random variables independent of D_n and D'_n. We have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}\left[\ell(Y_i, f(X_i))\right] |\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|D_n\right] |\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|D_n\right] |\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}\left[|\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|\ |D_n\right]\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\ell(Y_i, f(X_i)) - \ell(Y'_i, f(X'_i))\right) |\right] \\
&\overset{(c)}{\leq} 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Y_i, f(X_i))|\right] \\
&\leq 2\sup_{D_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i))|\right].
\end{aligned}$$

where (a) follows from Jensen's inequality with convex function $f(x) = |x|$, (b) follows from the fact that (X_i, Y_i) and (X'_i, Y'_i) has the same distributions, and (c) follows from triangle inequality.
Rademacher complexity: of a class $\mathcal{F}$ of functions for a given loss function $\ell(\cdot, \cdot)$ and samples D_n is defined as

$$\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{D_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i))|\right].$$

Therefore, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}[\ell(Y_i, f(X_i))]|\right] \leq 2\mathcal{R}_n(\ell \circ \mathcal{F})$$

and we only require to bound the Rademacher complexity.

5.1.3 Finite Class of functions

Suppose that the class of functions $\mathcal{F}$ is finite. We have the following bound.

Theorem: Assume that $\mathcal{F}$ is finite and that ℓ takes values in $[0,1]$. We have

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

Proof. From the previous lecture, for $B \subseteq \mathbb{R}^n$, we have that

$$\mathcal{R}_n(B) = \mathbb{E}\left[\max_{b\in B} |\frac{1}{n}\sum_{i=1}^n \sigma_i b_i|\right] \leq \max_{b\in B} |b|_2 \frac{\sqrt{2\log(2|B|)}}{n}.$$

Here, we have

$$B = \left\{ \begin{pmatrix} \ell(y_1, f(x_1)) \\ \vdots \\ \ell(y_n, f(x_n)) \end{pmatrix}, f \in \mathcal{F} \right\}.$$

Since ℓ takes values in $[0,1]$, this implies $B \subseteq \{b : |b|_2 \leq \sqrt{n}\}$. Plugging this bound in the previous inequality completes the proof. □

5.2 The General Case

Recall that for the classification problem, we had $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$. We have seen that the cardinality of the set $\{(f(x_1),\dots,f(x_n)), f \in \mathcal{F}\}$ plays an important role in bounding the risk of $\hat{f}^{\text{erm}}$ (this is not exactly what we used but the XOR argument of the previous lecture allows us to show that the cardinality of this set is the same as the cardinality of the set that interests us). In this lecture, this set might be uncountable. Therefore, we need to introduce a metric on this set so that we can treat the close points in the same manner. To this end we will define covering numbers (which basically plays the role of VC dimension in the classification).

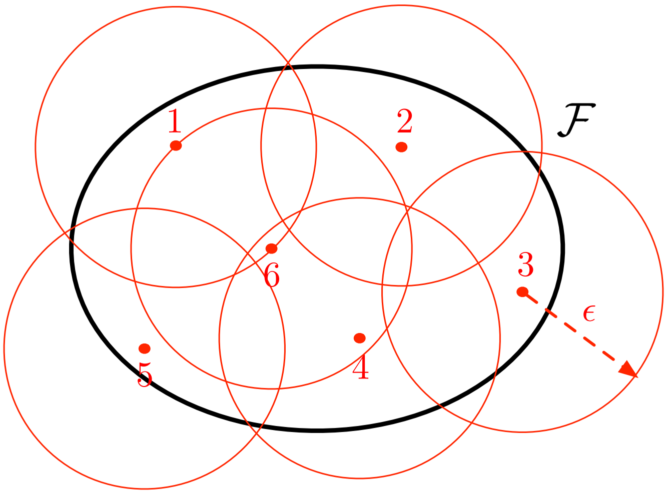
5.2.1 Covering Numbers

Definition: Given a set of functions $\mathcal{F}$ and a pseudo metric d on $\mathcal{F}$ ($(\mathcal{F}, d)$ is a metric space) and $\varepsilon > 0$. An *ε-net* of $(\mathcal{F}, d)$ is a set V such that for any $f \in \mathcal{F}$, there exists $g \in V$ such that $d(f,g) \leq \varepsilon$. Moreover, the *covering numbers* of $(\mathcal{F}, d)$ are defined by

$$N(\mathcal{F}, d, \varepsilon) = \inf\{|V| \, : \, V \text{ is an } \varepsilon\text{-net}\}.$$

For instance, for the $\mathcal{F}$ shown in the Figure 5.2.1 the set of points $\{1,2,3,4,5,6\}$ is a covering. However, the covering number is 5 as point 6 can be removed from V and the resulting points are still a covering.

Definition: Given $x = (x_1, \dots, x_n)$, the *conditional Rademacher average* of a class of



functions $\mathcal{F}$ is defined as

$$\hat{\mathcal{R}}_n^x = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right].$$

Note that in what follows we consider a general class of functions $\mathcal{F}$. However, for applying the results in order to bound empirical risk minimization, we take x_i to be (x_i, y_i) and $\mathcal{F}$ to be $\ell \circ \mathcal{F}$. We define the empirical l_1 distance as

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - g(x_i)|.$$

Theorem: If $0 \le f \le 1$ for all $f \in \mathcal{F}$, then for any $x = (x_1, \ldots, x_n)$, we have

$$\hat{R}_n^x(\mathcal{F}) \le \inf_{\varepsilon \ge 0} \{\varepsilon + \sqrt{\frac{2\log\left(2N(\mathcal{F}, d_1^x, \varepsilon)\right)}{n}}\}.$$

Proof. Fix $x = (x_1, \ldots, x_n)$ and $\varepsilon > 0$. Let V be a minimal ε-net of $(\mathcal{F}, d_1^x)$. Thus, by definition we have that $|V| = N(\mathcal{F}, d_1^x, \varepsilon)$. For any $f \in \mathcal{F}$, define $f^\circ \in V$ such that

$d_1^x(f, f^\circ) \leq \varepsilon$. We have that

$$\begin{aligned}
R_n^x(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i (f(x_i) - f^\circ(x_i))|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f^\circ(x_i)|\right] \\
&\leq \varepsilon + \mathbb{E}\left[\max_{f\in V} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)|\right] \\
&\leq \varepsilon + \sqrt{\frac{2\log(2|V|)}{n}} \\
&= \varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}.
\end{aligned}$$

Since the previous bound holds for any ε, we can take the infimum over all $\varepsilon \geq 0$ to obtain

$$R_n^x(\mathcal{F}) \leq \inf_{\varepsilon\geq 0} \{\varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}\}.$$

□

The previous bound clearly establishes a trade-off because as ε decreases $N(\mathcal{F}, d_1^x, \varepsilon)$ increases.

5.2.2 Computing Covering Numbers

As a warm-up, we will compute the covering number of the ℓ_2 ball of radius 1 in $\mathbb{R}^d$ denoted by B_2. We will show that the covering is at most $(\frac{3}{\varepsilon})^d$. There are several techniques to prove this result: one is based on a probabilistic method argument and one is based on greedily finding an ε-net. We will describe the later approach here. We select points in V one after another so that at step k, we have $u_k \in B_2 \setminus \cup_{j=1}^k B(u_j, \varepsilon)$. We will continue this procedure until we run out of points. Let it be step N. This means that $V = \{u_1, \ldots, u_N\}$ is an ε-net. We claim that the balls $B(u_i, \frac{\varepsilon}{2})$ and $B(u_j, \frac{\varepsilon}{2})$ for any $i, j \in \{1, \ldots, N\}$ are disjoint. The reason is that if $v \in B(u_i, \frac{\varepsilon}{2}) \cap B(u_j, \frac{\varepsilon}{2})$, then we would have

$$\|u_i - u_j\|_2 \leq \|u_i - v\|_2 + \|v - u_j\|_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which contradicts the way we have chosen the points. On the other hand, we have that $\cup_{j=1}^N B(u_j, \frac{\varepsilon}{2}) \subseteq (1 + \frac{\varepsilon}{2})B_2$. Comparing the volume of these two sets leads to

$$|V|(\frac{\varepsilon}{2})^d \mathsf{vol}(B_2) \leq (1 + \frac{\varepsilon}{2})^d \mathsf{vol}(B_2)\,,$$

where $\mathsf{vol}(B_2)$ denotes the volume of the unit Euclidean ball in d dimensions. It yields,

$$|V| \leq \frac{(1 + \frac{\varepsilon}{2})^d}{(\frac{\varepsilon}{2})^d} = \left(\frac{2}{\varepsilon} + 1\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d.$$

For any $p \geq 1$, define

$$d_p^x(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-g(x_i)|^p\right)^{\frac{1}{p}},$$

and for $p = \infty$, define

$$d_\infty^x(f,g) = \max_i |f(x_i)-g(x_i)|.$$

Using the previous theorem, in order to bound $\hat{\mathcal{R}}_n^x$ we need to bound the covering number with d_1^x norm. We claim that it is sufficient to bound the covering number for the infinity-norm. In order to show this, we will compare the covering number of the norms $d_p^x(f,g) = \left(\frac{1}{n}\sum_{i=1}^n |f(x_i)-g(x_i)|^p\right)^{\frac{1}{p}}$ for $p \geq 1$ and conclude that a bound on $N(\mathcal{F}, d_\infty^x, \varepsilon)$ implies a bound on $N(\mathcal{F}, d_p^x, \varepsilon)$ for any $p \geq 1$.

Proposition: For any $1 \leq p \leq q$ and $\varepsilon > 0$, we have that

$$N(\mathcal{F}, d_p^x, \varepsilon) \leq N(\mathcal{F}, d_q^x, \varepsilon).$$

Proof. First note that if $q = \infty$, then the inequality evidently holds. Because, we have

$$(\frac{1}{n}\sum_{i=1}^{n}|z_i|^p)^{\frac{1}{p}} \leq \max_i |z_i|,$$

which leads to $B(f, d_\infty^x, \varepsilon) \subseteq B(f, d_p^x, \varepsilon)$ and $N(f, d_\infty, \varepsilon) \geq N(f, d_p, \varepsilon)$. Now suppose that $1 \leq p \leq q < \infty$. Using Hölder's inequality with $r = \frac{q}{p} \geq 1$ we obtain

$$\left(\frac{1}{n}\sum_{i=1}^{n}|z_i|^p\right)^{\frac{1}{p}} \leq n^{-\frac{1}{p}}\left(\sum_{i=1}^{n}1\right)^{(1-\frac{1}{r})\frac{1}{p}}\left(\sum_{i=1}^{n}|z_i|^{pr}\right)^{\frac{1}{pr}} = \left(\frac{1}{n}\sum_{i=1}^{n}|z_i|^q\right)^{\frac{1}{q}}.$$

This inequality yeilds

$$B(f, d_q^x, \varepsilon) = \{g \,:\, d_q^x(f,g) \leq \varepsilon\} \subseteq B(f, d_p^x, \varepsilon),$$

which leads to $N(f, d_q, \varepsilon) \geq N(f, d_p, \varepsilon)$. □

Using this propositions we only need to bound $N(\mathcal{F}, d_\infty^x, \varepsilon)$.

Let the function class be $\mathcal{F} = \{f(x) = \langle f, x\rangle, f \in B_p^d, x \in B_q^d\}$, where $\frac{1}{p} + \frac{1}{q} = 1$. This leads to $|f| \leq 1$.

Claim: $N(\mathcal{F}, d_\infty^x, \varepsilon) \leq (\frac{2}{\varepsilon})^d$.

This leads to

$$\hat{R}_n^x(\mathcal{F}) \leq \inf_{\varepsilon>0}\{\varepsilon + \sqrt{\frac{2d\log(4/\varepsilon)}{n}}\}.$$

Taking $\varepsilon = O(\sqrt{\frac{d\log n}{n}})$, we obtain

$$\hat{R}_n^x(\mathcal{F}) \leq O(\sqrt{\frac{d\log n}{n}}).$$

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 7
Scribe: Zach Izzo — Sep. 30, 2015

In this lecture, we continue our discussion of covering numbers and compute upper bounds for specific conditional Rademacher averages $\hat{\mathcal{R}}_n^x(\mathcal{F})$. We then discuss chaining and conclude by applying it to learning.

Recall the following definitions. We define the risk function

$$R(f) = \mathbb{E}[\ell(X, f(X))], \qquad (X,Y) \in \mathcal{X} \times [-1,1]\,,$$

for some loss function $\ell(\cdot,\cdot)$. The conditiona Rademacher average that we need to control is

$$\mathcal{R}(\ell l \circ \mathcal{F}) = \sup_{(x_1,y_1),\dots,(x_n,y_n)} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(y_i, f(x_i))\right|\right].$$

Furthermore, we defined the conditional Rademacher average for a point $x = (x_1,\dots,x_n)$ to be

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right].$$

Lastly, we define the ε-covering number $N(\mathcal{F}, d, \varepsilon)$ to be the minimum number of balls (with respect to the metric d) of radius ε needed to cover $\mathcal{F}$. We proved the following theorem:

Theorem: Assume $|f| \le 1$ for all $f \in \mathcal{F}$. Then

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \inf_{\varepsilon>0}\left\{\varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}\right\},$$

where d_1^x is given by

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|.$$

We make use of this theorem in the following example. Define $B_p^d = \{x \in \mathbb{R}^d \,:\, |x|_p \le 1\}$. Then take $f(x) = \langle a, x\rangle$, set $\mathcal{F} = \{\langle a, \cdot\rangle \,:\, a \in B_\infty^d\}$, and $\mathcal{X} = B_1^d$. By Hölder's inequality, we have

$$|f(x)| \le |a|_\infty |x|_1 \le 1,$$

so the theorem above holds. We need to compute the covering number $N(\mathcal{F}, d_1^x, \varepsilon)$. Note that for all $a \in B_\infty^d$, there exists $v = (v_1,\dots,v_n)$ such that $v_i = g(x_i)$ and

$$\frac{1}{n}\sum_{i=1}^n |\langle a, x_i\rangle - v_i| \le \varepsilon$$

for some function g. For this case, we will take $g(x) = \langle b, x\rangle$, so $v_i = \langle b, x_i\rangle$. Now, note the following. Given this definition of g, we have

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^n |\langle a, x_1\rangle - \langle b, x_i\rangle| = \frac{1}{n}\sum_{i=1}^n |\langle a-b, x_i\rangle| \le |a-b|_\infty$$

by Hölder's inequality and the fact that $|x|_1 = 1$. So if $|a-b|_\infty \leq \varepsilon$, we can take $v_i = \langle b, x_i \rangle$. We just need to find a set of $\{b_1, \ldots, b_M\} \subset \mathbb{R}^d$ such that, for any a there exists b_j such that $|a - b_j|_\infty < \infty$. We can do this by dividing B^d_∞ into cubes with side length ε and taking the b_j's to be the set of vertices of these cubes. Then any $a \in B^d_\infty$ must land in one of these cubes, so $|a - b_j|_\infty \leq \varepsilon$ as desired. There are c/ε^d of such b_j's for some constant $c > 0$. Thus

$$N(B^d_\infty, d^x_1, \varepsilon) \leq c/\varepsilon^d.$$

We now plug this value into the theorem to obtain

$$\hat{\mathcal{R}}^x_n(\mathcal{F}) \leq \inf_{\varepsilon \geq 0} \left\{ \varepsilon + \sqrt{\frac{2\log(c/\varepsilon^d)}{n}} \right\}.$$

Optimizing over all choices of ε gives

$$\varepsilon^* = c\sqrt{\frac{d\log(n)}{n}} \qquad \Rightarrow \qquad \hat{\mathcal{R}}^x_n(\mathcal{F}) \leq c\sqrt{\frac{d\log(n)}{n}}.$$

Note that in this final inequality, the conditional empirical risk no longer depends on x, since we "sup'd" x out of the bound during our computations. In general, one should ignore x unless it has properties which will guarantee a bound which is better than the sup. Another important thing to note is that we are only considering one granularity of $\mathcal{F}$ in our final result, namely the one associated to ε^*. It is for this reason that we pick up an extra log factor in our risk bound. In order to remove this term, we will need to use a technique called *chaining.*

5.4 Chaining

We have the following theorem.

Theorem: Assume that $|f| \leq 1$ for all $f \in \mathcal{F}$. Then

$$\hat{\mathcal{R}}^x_n \leq \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^1 \sqrt{\log(N(\mathcal{F}, d^x_2, t))} dt \right\}.$$

(Note that the integrand decays with t.)

Proof. Fix $x = (x_1, \ldots, x_n)$, and for all $j = 1, \ldots, N$, let V_j be a minimal 2^{-j}-net of $\mathcal{F}$ under the d^x_2 metric. (The number N will be determined later.) For a fixed $f \in \mathcal{F}$, this process will give us a "chain" of points f°_i which converges to f: $d^x_2(f^\circ_i, f) \leq 2^{-j}$.

Define $F = \{(f(x_1), \ldots, f(x_n))^\top,\ f \in \mathcal{F}\} \subset [-1, 1]^n$. Note that

$$\hat{\mathcal{R}}^x_n(\mathcal{F}) = \frac{1}{n} \mathbb{E} \sup_{f \in F} \langle \sigma, f \rangle$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$. Observe that for all N, we can rewrite $\langle \sigma, f \rangle$ as a telescoping sum:

$$\langle \sigma, f \rangle = \langle \sigma, f - f^\circ_N \rangle + \langle \sigma, f^\circ_N - f^\circ_{N-1} \rangle + \ldots + \langle \sigma, f^\circ_1 - f^\circ_0 \rangle$$

where $f_0^\circ := 0$. Thus

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle\sigma, f - f_N^\circ\rangle| + \sum_{j=1}^N \frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle\sigma, f_j^\circ - f_{j-1}^\circ\rangle|\,.$$

We can control the two terms in this inequality separately. Note first that by the Cauchy-Schwarz inequality,

$$\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle\sigma, f - f_N^\circ\rangle| \le |\sigma|_2\frac{d_2^x(f, f_N^\circ)}{\sqrt{n}}.$$

Since $|\sigma|_2 = \sqrt{n}$ and $d_2^x(f, f_N^\circ) \le 2^{-N}$, we have

$$\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle\sigma, f - f_N^\circ\rangle| \le 2^{-N}\,.$$

Now we turn our attention to the second term in the inequality, that is

$$S = \sum_{j=1}^N \frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle\sigma, f_j^\circ - f_{j-1}^\circ\rangle|.$$

Note that since $f_j^\circ \in V_j$ and $f_{j-1}^\circ \in V_{j-1}$, there are at most $|V_j||V_{j-1}|$ possible differences $f_j^\circ - f_{j-1}^\circ$. Since $|V_{j-1}| \le |V_j|/2$, $|V_j||V_{j-1}| \le |V_j|^2/2$ and we find ourselves in the finite dictionary case. We employ a risk bound from earlier in the course to obtain the inequality

$$\mathcal{R}_n(B) \le \max_{b\in B}|b|_2\frac{\sqrt{2\log(2|B|)}}{n}\,.$$

In the present case, $B = \{f_j^\circ - f_{j-1}^\circ\,, f \in F\}$ so that $|B| \le |V_j|^2/2$. It yields

$$\mathcal{R}_n(B) \le r\cdot\frac{\sqrt{2\log(\frac{2|V_j|^2}{2})}}{n} = 2r\cdot\frac{\sqrt{\log|V_j|}}{n}\,,$$

where $r = \sup_{f\in F}|f_j^\circ - f_{j-1}^\circ|_2$. Next, observe that

$$|f_j^\circ - f_{j-1}^\circ|_2 = \sqrt{n}\cdot d_2^x(f_j^\circ, f_{j-1}^\circ) \le \sqrt{n}(d_2^x(f_j^\circ, f) + d_2^x(f, f_{j-1}^\circ)) \le 3\cdot 2^{-j}\sqrt{n}\,.$$

by the triangle inequality and the fact that $d_2^x(f_j^\circ, f) \le 2^{-j}$. Substituting this back into our bound for $\mathcal{R}_n(B)$, we have

$$\mathcal{R}_n(B) \le 6\cdot 2^{-j}\sqrt{\frac{\log|V_j||}{n}} = 6\cdot 2^{-j}\sqrt{\frac{\log(N(\mathcal{F}, d_2^x, 2^{-j}))}{n}}$$

since V_j was chosen to be a minimal 2^{-j}-net.

The proof is almost complete. Note that $2^{-j} = 2(2^{-j} - 2^{-j-1})$ so that

$$\frac{6}{\sqrt{n}}\sum_{j=1}^N 2^{-j}\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))} = \frac{12}{\sqrt{n}}\sum_{j=1}^N(2^{-j} - 2^{-j-1})\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))}\,.$$

Next, by comparing sums and integrals (Figure 1), we see that

$$\sum_{j=1}^N(2^{-j} - 2^{-j-1})\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))} \le \int_{2^{-(N+1)}}^{1/2}\sqrt{\log(N(\mathcal{F}, d_2^x, t))}dt.$$

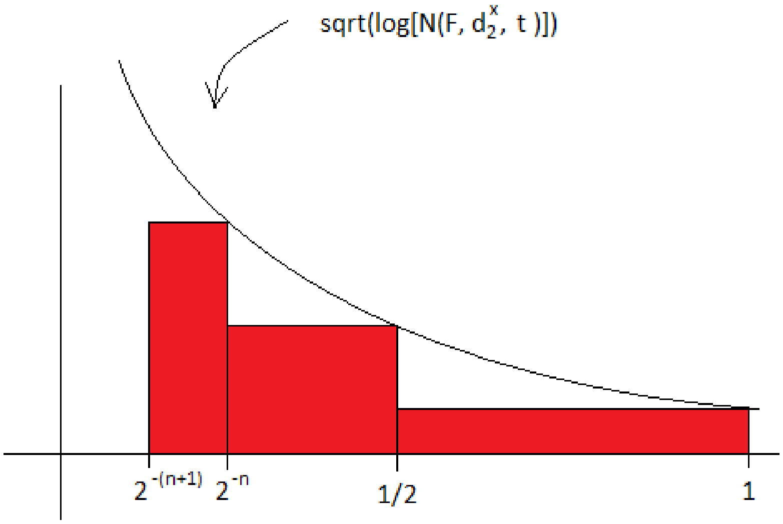


Figure 1: A comparison of the sum and integral in question.

So we choose N such that $2^{-(N+2)} \le \varepsilon \le 2^{-(N+1)}$, and by combining our bounds we obtain

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le 2^{-N} + \frac{12}{\sqrt{n}} \int_{2^{-(N+1)}}^{1/2} \sqrt{\log(N(\mathcal{F}, d_2^x, t))}dt \le 4\varepsilon + \int_{\varepsilon}^{1} \sqrt{\log(N, \mathcal{F}, t)}dt$$

since the integrand is non-negative. (Note: this integral is known as the "Dudley Entropy Integral.") □

Returning to our earlier example, since $N(\mathcal{F}, d_2^x, \varepsilon) \le c/\varepsilon^d$, we have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{1} \sqrt{\log((c'/t)^d)}dt \right\}.$$

Since $\int_0^1 \sqrt{\log(c/t)}dt = \bar{c}$ is finite, we then have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le 12\bar{c}\sqrt{d/n}.$$

Using chaining, we've been able to remove the log factor!

5.5 Back to Learning

We want to bound

$$\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i)) \right| \right].$$

We consider $\hat{\mathcal{R}}_n^x(\Phi \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \Phi \circ f(x_i)\right|\right]$ for some L-Lipschitz function Φ, that is $|\Phi(a) - \Phi(b)| \le L|a - b|$ for all $a, b \in [-1, 1]$. We have the following lemma.

Theorem: (Contraction Inequality) Let Φ be L-Lipschitz and such that $\Phi(0) = 0$, then

$$\hat{\mathcal{R}}_n^x(\Phi \circ \mathcal{F}) \leq 2L \cdot \hat{\mathcal{R}}_n^x(\mathcal{F}) \,.$$

The proof is omitted and the interested reader should take a look at [LT91, Kol11] for example.

As a final remark, note that requiring the loss function to be Lipschitz prohibits the use of $\mathbb{R}$-valued loss functions, for example $\ell(Y, \cdot) = (Y - \cdot)^2$.

References

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Lecture Notes in Mathematics 2033. Berlin: Springer. ix, 254 p. EUR 48.10 , 2011.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)].* Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 8
Scribe: Quan Li — Oct. 5, 2015

Part II
Convexity

1. CONVEX RELAXATION OF THE EMPIRICAL RISK MINIMIZATION

In the previous lectures, we have proved upper bounds on the excess risk $R(\hat{h}^{\text{erm}}) - R(h^*)$ of the Empirical Risk Minimizer

$$\hat{h}^{\text{erm}} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq h(X_i)). \tag{1.1}$$

However due to the nonconvexity of the objective function, the optimization problem (1.1) in general can not be solved efficiently. For some choices of $\mathcal{H}$ and the classification error function (e.g. $\mathbb{1}(\cdot)$), the optimization problem can be NP-hard. However, the problem we deal with has some special features:

1. Since the upper bound we obtained on the excess risk is $O(\sqrt{\frac{d \log n}{n}})$, we only need to approximate the optimization problem with error up to $O(\sqrt{\frac{d \log n}{n}})$.
2. The optimization problem corresponds to the average case problem where the data $(X_i, Y_i) \overset{i.i.d}{\sim} P_{X,Y}$.
3. $\mathcal{H}$ can be chosen to be some 'natural' classifiers, e.g. $\mathcal{H} = \{\text{half spaces}\}$.

These special features might help us bypass the computational issue. Computational issue in machine learning have been studied for quite some time (see, e.g. [Kea90]), especially in the context of PAC learning. However, many of these problems are somewhat abstract and do not shed much light on the practical performance of machine learning algorithms.

To avoid the computational problem, the basic idea is to minimize a convex upper bound of the classification error function $\mathbb{1}(\cdot)$ in (1.1). For the purpose of computation, we shall also require that the function class $\mathcal{H}$ be a convex set. Hence the resulting minimization becomes a convex optimization problem which can be solved efficiently.

1.1 Convexity

Definition: A set C is convex if for all $x, y \in C$ and $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in C$.

Definition: A function $f : D \to \mathbb{R}$ on a convex domain D is convex if it satisfies

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in D, \text{ and } \lambda \in [0,1].$$

1.2 Convex relaxation

The convex relaxation takes three steps.

Step 1: Spinning.
Using a mapping $Y \mapsto 2Y - 1$, the i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is transformed to lie in $\mathcal{X} \times \{-1, 1\}$. These new labels are called *spinned* labels. Correspondingly, the task becomes to find a classifier $h : \mathcal{X} \mapsto \{-1, 1\}$. By the relation

$$h(X) \neq Y \Leftrightarrow -h(X)Y > 0,$$

we can rewrite the objective function in (1.1) by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbb{1}}(-h(X_i)Y_i) \tag{1.2}$$

where $\varphi_{\mathbb{1}}(z) = \mathbb{1}(z > 0)$.

Step 2: Soft classifiers.
The set $\mathcal{H}$ of classifiers in (1.1) contains only functions taking values in $\{-1, 1\}$. As a result, it is non convex if it contains at least two distinct classifiers. Soft classifiers provide a way to remedy this nuisance.

Definition: A *soft classifier* is any measurable function $f : \mathcal{X} \to [-1, 1]$. The *hard classifier* (or simply "classifier") associated to a soft classifier f is given by $h = \text{sign}(f)$.

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a *convex* set soft classifiers. Several popular choices for $\mathcal{F}$ are:

- Linear functions:

  $$\mathcal{F} := \{\langle a, x\rangle : a \in \mathcal{A}\}.$$

  for some convex set $\mathcal{A} \in \mathbb{R}^d$. The associated hard classifier $h = \text{sign}(f)$ splits $\mathbb{R}^d$ into two half spaces.

- Majority votes: given weak classifiers $h_1, \ldots, h_M$,

  $$\mathcal{F} := \Big\{\sum_{j=1}^{M} \lambda_j h_j(x) : \lambda_j \geq 0, \sum_{j=1}^{M} \lambda_j = 1\Big\}.$$

- Let φ_j, $j = 1, 2, \ldots$ a family of functions, e.g., Fourier basis or Wavelet basis. Define

  $$\mathcal{F} := \{\sum_{j=1}^{\infty} \theta_j \varphi_j(x) : (\theta_1, \theta_2, \ldots) \in \Theta\},$$

  where Θ is some convex set.

Step 3: Convex surrogate.
Given a convex set $\mathcal{F}$ of soft classifiers, using the rewriting in (1.2), we need to solve that minimizes the empirical classification error

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varphi_{\mathbb{1}}(-f(X_i)Y_i),$$

However, while we are now working with a convex constraint, our objective is still not convex: we need a *surrogate* for the classification error.

Definition: A function $\varphi : \mathbb{R} \mapsto \mathbb{R}_+$ is called a convex surrogate if it is a convex non-decreasing function such that $\varphi(0) = 1$ and $\varphi(z) \geq \varphi_{\mathbb{1}}(z)$ for all $z \in \mathbb{R}$.

The following is a list of convex surrogates of loss functions.

- Hinge loss: $\varphi(z) = \max(1 + z, 0)$.
- Exponential loss: $\varphi(z) = \exp(z)$.
- Logistic loss: $\varphi(z) = \log_2(1 + \exp(z))$.

To bypass the nonconvexity of $\varphi_{\mathbb{1}}(\cdot)$, we may use a convex surrogate $\varphi(\cdot)$ in place of $\varphi_{\mathbb{1}}(\cdot)$ and consider the minimizing the *empirical φ-risk* $\hat{R}_{n,\varphi}$ defined by

$$\hat{R}_{n,\varphi}(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f(X_i))$$

It is the empirical counterpart of the *φ-risk* R_φ defined by

$$R_\varphi(f) = \mathbb{E}[\varphi(-Yf(X))].$$

1.3 φ-risk minimization

In this section, we will derive the relation between the φ-risk $R_\varphi(f)$ of a soft classifier f and the classification error $R(h) = \mathbb{P}(h(X) \neq Y)$ of its associated hard classifier $h = \text{sign}(f)$

Let

$$f^*_\varphi = \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} E[\varphi(-Yf(X))]$$

where the infimum is taken over all measurable functions $f : \mathcal{X} \to \mathbb{R}$.

To verify that minimizing the φ serves our purpose, we will first show that if the convex surrogate $\varphi(\cdot)$ is differentiable, then $\text{sign}(f^*_\varphi(X)) \geq 0$ is equivalent to $\eta(X) \geq 1/2$ where $\eta(X) = \mathbb{P}(Y = 1 \mid X)$. Conditional on $\{X = x\}$, we have

$$\mathbb{E}[\varphi(-Yf(X)) \mid X = x] = \eta(x)\varphi(-f(x)) + (1 - \eta(x))\varphi(f(x)).$$

Let

$$H_\eta(\alpha) = \eta(x)\varphi(-\alpha) + (1 - \eta(x))\varphi(\alpha) \tag{1.3}$$

so that

$$f^*_\varphi(x) = \operatorname*{argmin}_{\alpha\in\mathbb{R}} H_\eta(\alpha)\,, \qquad \text{and} \qquad R^*_\varphi = \min_{f\in\mathbb{R}^{\mathcal{X}}} R_\varphi(f) = \min_{\alpha\in\mathbb{R}} H_{\eta(x)}(\alpha)\,.$$

Since $\varphi(\cdot)$ is differentiable, setting the derivative of $H_\eta(\alpha)$ to zero gives $f^*_\varphi(x) = \bar\alpha$, where

$$H'_\eta(\bar\alpha) = -\eta(x)\varphi'(-\bar\alpha) + (1-\eta(x))\varphi'(\bar\alpha) = 0,$$

which gives

$$\frac{\eta(x)}{1-\eta(x)} = \frac{\varphi'(\bar\alpha)}{\varphi'(-\bar\alpha)}$$

Since $\varphi(\cdot)$ is a convex function, its derivative $\varphi'(\cdot)$ is non-decreasing. Then from the equation above, we have the following equivalence relation

$$\eta(x) \geq \frac{1}{2} \Leftrightarrow \bar\alpha \geq 0 \Leftrightarrow \operatorname{sign}(f^*_\varphi(x)) \geq 0. \tag{1.4}$$

Since the equivalence relation holds for all $x \in \mathcal{X}$,

$$\eta(X) \geq \frac{1}{2} \Leftrightarrow \operatorname{sign}(f^*_\varphi(X)) \geq 0.$$

The following lemma shows that if the *excess φ-risk* $R_\varphi(f) - R^*_\varphi$ of a soft classifier f is small, then the excess-risk of its associated hard classifier $\operatorname{sign}(f)$ is also small.

Lemma (Zhang's Lemma [Zha04]): Let $\varphi : \mathbb{R} \mapsto \mathbb{R}_+$ be a convex non-decreasing function such that $\varphi(0) = 1$. Define for any $\eta \in [0,1]$,

$$\tau(\eta) := \inf_{\alpha\in\mathbb{R}} H_\eta(\alpha).$$

If there exists $c > 0$ and $\gamma \in [0,1]$ such that

$$|\eta - \frac{1}{2}| \leq c(1-\tau(\eta))^\gamma\,, \quad \forall \eta \in [0,1]\,, \tag{1.5}$$

then

$$R(\operatorname{sign}(f)) - R^* \leq 2c(R_\varphi(f) - R^*_\varphi)^\gamma$$

Proof. Note first that $\tau(\eta) \leq H_\eta(0) = \varphi(0) = 1$ so that condition (2.5) is well defined.

Next, let $h^* = \operatorname{argmin}_{h\in\{-1,1\}^{\mathcal{X}}} \mathbb{P}[h(X) \neq Y] = \operatorname{sign}(\eta - 1/2)$ denote the Bayes classifier, where $\eta = \mathbb{P}[Y = 1|X = x]$, . Then it is easy to verify that

$$\begin{aligned} R(\operatorname{sign}(f)) - R^* &= \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(\operatorname{sign}(f(X)) \neq h^*(X))] \\ &= \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(f(X)(\eta(X) - 1/2) < 0)] \\ &\leq 2c\mathbb{E}[((1 - \tau(\eta(X)))\mathbb{1}(f(X)(\eta(X) - 1/2) < 0))^\gamma] \\ &\leq 2c\left(\mathbb{E}[(1 - \tau(\eta(X)))\mathbb{1}(f(X)(\eta(X) - 1/2) < 0)]\right)^\gamma\,, \end{aligned}$$

where the last inequality above follows from Jensen's inequality.

We are going to show that for any $x \in \mathcal{X}$, it holds

$$(1-\tau(\eta))\mathbb{1}(f(x)(\eta(x)-1/2)<0)] \le \mathbb{E}[\varphi(-Yf(x)) \mid X=x] - R^*_\varphi \,. \tag{1.6}$$

This will clearly imply the result by integrating with respect to x.

Recall first that

$$\mathbb{E}[\varphi(-Yf(x)) \mid X=x] = H_{\eta(x)}(f(x)) \qquad \text{and} \qquad R^*_\varphi = \min_{\alpha\in\mathbb{R}} H_{\eta(x)}(\alpha) = \tau(\eta(x))\,.$$

so that (2.6) is equivalent to

$$(1-\tau(\eta))\mathbb{1}(f(x)(\eta(x)-1/2)<0)] \le H_{\eta(x)}(\alpha) - \tau(\eta(x))$$

Since the right-hand side above is nonnegative, the case where $f(x)(\eta(x)-1/2)\ge 0$ follows trivially. If $f(x)(\eta(x)-1/2)<0$, (2.6) follows if we prove that $H_{\eta(x)}(\alpha)\ge 1$. The convexity of $\varphi(\cdot)$ gives

$$\begin{aligned}
H_{\eta(x)}(\alpha) &= \eta(x)\varphi(-f(x)) + (1-\eta(x))\varphi(f(x))\\
&\ge \varphi(-\eta(x)f(x) + (1-\eta(x))f(x))\\
&= \varphi((1-2\eta(x))f(x))\\
&\ge \varphi(0) = 1\,,
\end{aligned}$$

where the last inequality follows from the fact that φ is non decreasing and $f(x)(\eta(x)-1/2)<0$. This completes the proof of (2.6) and thus of the Lemma. □

IT is not hard to check the following values for the quantities $\tau(\eta)$, c and γ for the three losses introduced above:

- Hinge loss: $\tau(\eta) = 1-|1-2\eta|$ with $c=1/2$ and $\gamma=1$.
- Exponential loss: $\tau(\eta) = 2\sqrt{\eta(1-\eta)}$ with $c=1/\sqrt{2}$ and $\gamma=1/2$.
- Logistic loss: $\tau(\eta) = -\eta\log\eta - (1-\eta)\log(1-\eta)$ with $c=1/\sqrt{2}$ and $\gamma=1/2$.

References

[Kea90] Michael J Kearns. *The computational complexity of machine learning.* PhD thesis, Harvard University, 1990.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 9
Scribe: Xuhong Zhang — Oct. 7, 2015

Recall that last lecture we talked about convex relaxation of the original problem

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i)$$

by considering soft classifiers (i.e. whose output is in $[-1,1]$ rather than in $\{0,1\}$) and convex surrogates of the loss function (e.g. hinge loss, exponential loss, logistic loss):

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_{\varphi,n}(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f(X_i))$$

And $\hat{h} = \operatorname{sign}(\hat{f})$ will be used as the 'hard' classifier.

We want to bound the quantity $R_\varphi(\hat{f}) - R_\varphi(\bar{f})$, where $\bar{f} = \operatorname{argmin}_{f \in \mathcal{F}} R_\varphi(f)$.

(1) $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_{\varphi,n}(f)$, thus

$$\begin{aligned}
R_\varphi(\hat{f}) &= R_\varphi(\bar{f}) + \hat{R}_{\varphi,n}(\bar{f}) - \hat{R}_{\varphi,n}(\bar{f}) + \hat{R}_{\varphi,n}(\hat{f}) - \hat{R}_{\varphi,n}(\hat{f}) + R_\varphi(\hat{f}) - R_\varphi(\bar{f}) \\
&\leq R_\varphi(\bar{f}) + \hat{R}_{\varphi,n}(\bar{f}) - \hat{R}_{\varphi,n}(\hat{f}) + R_\varphi(\hat{f}) - R_\varphi(\bar{f}) \\
&\leq R_\varphi(\bar{f}) + 2 \sup_{f \in \mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|
\end{aligned}$$

(2) Let us first focus on $\mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|]$. Using the symmetrization trick as before, we know it is upper-bounded by $2\mathcal{R}_n(\varphi \circ \mathcal{F})$, where the Rademacher complexity

$$\mathcal{R}_n(\varphi \circ \mathcal{F}) = \sup_{X_1,\ldots,X_n,Y_1,\ldots,Y_n} \mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \sigma_i \varphi(-Y_i f(X_i))|]$$

One thing to notice is that $\varphi(0) = 1$ for the loss functions we consider (hinge loss, exponential loss and logistic loss), but in order to apply contraction inequality later, we require $\varphi(0) = 0$. Let us define $\psi(\cdot) = \varphi(\cdot) - 1$. Clearly $\psi(0) = 0$, and

$$\begin{aligned}
&\mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i f(X_i))])|] \\
&= \mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (\psi(-Y_i f(X_i)) - \mathbb{E}[\psi(-Y_i f(X_i))])|] \\
&\leq 2\mathcal{R}_n(\psi \circ \mathcal{F})
\end{aligned}$$

(3) The Rademacher complexity of $\psi \circ \mathcal{F}$ is still difficult to deal with. Let us assume that $\varphi(\cdot)$ is L-Lipschitz, (as a result, $\psi(\cdot)$ is also L-Lipschitz), apply the contraction inequality, we have

$$R_n(\psi \circ \mathcal{F}) \leq 2LR_n(\mathcal{F})$$

(4) Let $Z_i = (X_i, Y_i)$, $i = 1, 2, ..., n$ and

$$g(Z_1, Z_2, ..., Z_n) = \sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n}(\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i f(X_i))])|$$

Since $\varphi(\cdot)$ is monotonically increasing, it is not difficult to verify that $\forall Z_1, Z_2, ..., Z_n, Z_i'$

$$|g(Z_1, ..., Z_i, ..., Z_n) - g(Z_1, ..., Z_i', ..., Z_n)| \le \frac{1}{n}(\varphi(1) - \varphi(-1)) \le \frac{2L}{n}$$

The last inequality holds since g is L-Lipschitz. Apply Bounded Difference Inequality,

$$\mathbb{P}(|\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|]| > t) \le 2\exp(-\frac{2t^2}{\sum_{i=1}^{n}(\frac{2L}{n})^2})$$

Set the RHS of above equation to δ, we get:

$$\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| \le \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|] + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

(5) Combining (1) - (4), we have

$$R_\varphi(\hat{f}) \le R_\varphi(\bar{f}) + 8L\mathcal{R}_n(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

1.4 Boosting

In this section, we will specialize the above analysis to a particular learning model: Boosting. The basic idea of Boosting is to convert a set of weak learners (i.e. classifiers that do better than random, but have high error probability) into a strong one by using the weighted average of weak learners' opinions. More precisely, we consider the following function class

$$\mathcal{F} = \{\sum_{j=1}^{M}\theta_j h_j(\cdot) : |\theta|_1 \le 1, h_j : \mathcal{X} \mapsto [-1, 1], j \in \{1, 2, ..., M\} \text{ are classifiers}\}$$

and we want to upper bound $\mathcal{R}_n(\mathcal{F})$ for this choice of $\mathcal{F}$.

$$\mathcal{R}_n(\mathcal{F}) = \sup_{Z_1,...,Z_n} \mathbb{E}[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n}\sigma_i Y_i f(X_i)|] = \frac{1}{n}\sup_{Z_1,...,Z_n} \mathbb{E}[\sup_{|\theta|_1\le 1} |\sum_{j=1}^{M}\theta_j \sum_{i=1}^{n} Y_i\sigma_i h_j(X_i)|]$$

Let $g(\theta) = |\sum_{j=1}^{M}\theta_j \sum_{i=1}^{n} Y_i\sigma_i h_j(X_i)|$. It is easy to see that $g(\theta)$ is a convex function, thus $\sup_{|\theta|_1\le 1} g(\theta)$ is achieved at a vertex of the unit ℓ_1 ball $\{\theta : \|\theta\|_1 \le 1\}$. Define the finite set

$$B_{\mathbf{X},\mathbf{Y}} \triangleq \left\{ \pm\begin{pmatrix} Y_1 h_1(X_1) \\ Y_2 h_1(X_2) \\ \vdots \\ Y_n h_1(X_n)\end{pmatrix}, \pm\begin{pmatrix} Y_1 h_2(X_1) \\ Y_2 h_2(X_2) \\ \vdots \\ Y_n h_2(X_n)\end{pmatrix}, \dots, \pm\begin{pmatrix} Y_1 h_M(X_1) \\ Y_2 h_M(X_2) \\ \vdots \\ Y_n h_M(X_n)\end{pmatrix} \right\}$$

Then

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{X},\mathbf{Y}} R_n(B_{\mathbf{X},\mathbf{Y}})\,.$$

Notice $\max_{b\in B_{\mathbf{X},\mathbf{Y}}} |b|_2 \leq \sqrt{n}$ and $|B_{\mathbf{X},\mathbf{Y}}| = 2M$. Therefore, using a lemma from Lecture 5, we get

$$\mathcal{R}_n(B_{\mathbf{X},\mathbf{Y}}) \leq \big[\max_{b\in B_{\mathbf{X},\mathbf{Y}}} |b|_2\big]\frac{\sqrt{2\log(2|B_{\mathbf{X},\mathbf{Y}}|)}}{n} \leq \sqrt{\frac{2\log(4M)}{n}}$$

Thus for Boosting,

$$R_\varphi(\hat{f}) \leq R_\varphi(\bar{f}) + 8L\sqrt{\frac{2\log(4M)}{n}} + 2L\sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{with probability 1 - } \delta$$

To get some ideas of what values L usually takes, consider the following examples:

(1) for hinge loss, i.e. $\varphi(x) = (1+x)_+$, $L = 1$.

(2) for exponential loss, i.e. $\varphi(x) = e^x$, $L = e$.

(3) for logistic loss, i.e. $\varphi(x) = \log_2(1+e^x)$, $L = \frac{e}{1+e}\log_2(e) \approx 2.43$

Now we have bounded $R_\varphi(\hat{f}) - R_\varphi(\bar{f})$, but this is not yet the excess risk. Excess risk is defined as $R(\hat{f}) - R(f^*)$, where $f^* = \operatorname{argmin}_f R_\varphi(f)$. The following theorem provides a bound for excess risk for Boosting.

Theorem: Let $\mathcal{F} = \{\sum_{j=1}^M \theta_j h_j : \|\theta\|_1 \leq 1, h_j s \text{ are weak classifiers}\}$ and φ is an L-Lipschitz convex surrogate. Define $\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}} R_{\varphi,n}(f)$ and $\hat{h} = \operatorname{sign}(\hat{f})$. Then

$$R(\hat{h}) - R^* \leq 2c\big(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*)\big)^\gamma + 2c\left(8L\sqrt{\frac{2\log(4M)}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma$$

with probability $1-\delta$

Proof.

$$\begin{aligned} R(\hat{h}) - R^* &\leq 2c\big(R_\varphi(f) - R_\varphi(f^*)\big)^\gamma \\ &\leq 2c\left(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*) + 8L\sqrt{\frac{2\log(4M)}{n}} + 2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma \\ &\leq 2c\big(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*)\big)^\gamma + 2c\left(8L\sqrt{\frac{2\log(4M)}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma \end{aligned}$$

Here the first inequality uses Zhang's lemma and the last one uses the fact that for $a_i \geq 0$ and $\gamma \in [0,1]$, $(a_1+a_2+a_3)^\gamma \leq a_1^\gamma + a_2^\gamma + a_3^\gamma$. □

1.5 Support Vector Machines

In this section, we will apply our analysis to another important learning model: Support Vector Machines (SVMs). We will see that hinge loss $\varphi(x) = (1+x)_+$ is used and the associated function class is $\mathcal{F} = \{f : \|f\|_W \leq \lambda\}$ where W is a Hilbert space. Before analyzing SVMs, let us first introduce Reproducing Kernel Hilbert Spaces (RKHS).

1.5.1 Reproducing Kernel Hilbert Spaces (RKHS)

Definition: A function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called a *positive symmetric definite kernel* (PSD kernel) if

(1) $\forall x, x' \in \mathcal{X}$, $K(x, x') = K(x', x)$

(2) $\forall n \in \mathbb{Z}_+$, $\forall x_1, x_2, ..., x_n$, the $n \times n$ matrix with $K(x_i, x_j)$ as its element in i^{th} row and j^{th} column is positive semi-definite. In other words, for any $a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

Let us look at a few examples of PSD kernels.

Example 1 Let $\mathcal{X} = \mathbb{R}$, $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$ is a PSD kernel, since $\forall a_1, a_2, ..., a_n \in \mathbb{R}$

$$\sum_{i,j} a_i a_j \langle x_i, x_j \rangle_{\mathbb{R}^d} = \sum_{i,j} \langle a_i x_i, a_j x_j \rangle_{\mathbb{R}^d} = \langle \sum_i a_i x_i, \sum_j a_j x_j \rangle_{\mathbb{R}^d} = \| \sum_i a_i x_i \|^2_{\mathbb{R}^d} \geq 0$$

Example 2 The Gaussian kernel $K(x, x') = \exp(-\frac{1}{2\sigma^2} \|x - x'\|^2_{\mathbb{R}^d})$ is also a PSD kernel.

Note that here and in the sequel, $\| \cdot \|_W$ and $\langle \cdot, \cdot \rangle_W$ denote the norm and inner product of Hilbert space W.

Definition: Let W be a Hilbert space of functions $\mathcal{X} \mapsto \mathbb{R}$. A symmetric kernel $K(\cdot, \cdot)$ is called **reproducing kernel** of W if

(1) $\forall x \in \mathcal{X}$, the function $K(x, \cdot) \in W$.

(2) $\forall x \in \mathcal{X}$, $f \in W$, $\langle f(\cdot), K(x, \cdot) \rangle_W = f(x)$.

If such a $K(x, \cdot)$ exists, W is called a **reproducing kernel Hilbert space** (RKHS).

Claim: If $K(\cdot, \cdot)$ is a reproducing kernel for some Hilbert space W, then $K(\cdot, \cdot)$ is a PSD kernel.

Proof. $\forall a_1, a_2, ..., a_n \in \mathbb{R}$, we have

$$\begin{aligned}\sum_{i,j} a_i a_j K(x_i, x_j) &= \sum_{i,j} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle \quad \text{(since } K(\cdot, \cdot) \text{ is reproducing)} \\ &= \langle \sum_i a_i K(x_i, \cdot), \sum_j a_j K(x_j, \cdot) \rangle_W \\ &= \| \sum_i a_i K(x_i, \cdot) \|^2_W \geq 0\end{aligned}$$

□

In fact, the above claim holds both directions, i.e. if a kernel $K(\cdot,\cdot)$ is PSD, it is also a reproducing kernel.

A natural question to ask is, given a PSD kernel $K(\cdot,\cdot)$, how can we build the corresponding Hilbert space (for which $K(\cdot,\cdot)$ is a reproducing kernel)? Let us look at a few examples.

Example 3 Let $\varphi_1, \varphi_2, ..., \varphi_M$ be a set of orthonormal functions in $L_2([0,1])$, i.e. for any $j, k \in \{1, 2, ..., M\}$

$$\int_x \varphi_j(x)\varphi_k(x)dx = \langle \varphi_j, \varphi_k \rangle = \delta_{jk}$$

Let $K(x, x') = \sum_{j=1}^{M} \varphi_j(x)\varphi_j(x')$. We claim that the Hilbert space

$$W = \{\sum_{j=1}^{M} a_j \varphi_j(\cdot) : a_1, a_2, ..., a_M \in \mathbb{R}\}$$

equipped with inner product $\langle \cdot,\cdot \rangle_{L_2}$ is a RKHS with reproducing kernel $K(\cdot,\cdot)$.

Proof. (1) $K(x,\cdot) = \sum_{j=1}^{M} \varphi_j(x)\varphi_j(\cdot) \in W$. (Choose $a_j = \varphi_j(x)$).

(2) If $f(\cdot) = \sum_{j=1}^{M} a_j\varphi_j(\cdot)$,

$$\langle f(\cdot), K(x,\cdot)\rangle_{L_2} = \langle \sum_{j=1}^{M} a_j\varphi_j(\cdot), \sum_{k=1}^{M} \varphi_k(x)\varphi_k(\cdot)\rangle_{L_2} = \sum_{j=1}^{M} a_j\varphi_j(x) = f(x)$$

(3) $K(x, x')$ is a PSD kernel: $\forall a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) = \sum_{i,j,k} a_i a_j \varphi_k(x_i)\varphi_k(x_j) = \sum_k (\sum_i a_i \varphi_k(x_i))^2 \geq 0$$

□

Example 4 If $\mathcal{X} = \mathbb{R}^d$, and $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$, the corresponding Hilbert space is $W = \{\langle w, \cdot\rangle : w \in \mathbb{R}^d\}$ (i.e. all linear functions) equipped with the following inner product: if $f = \langle w, \cdot \rangle$, $g = \langle v, \cdot \rangle$, $\langle f, g\rangle \triangleq \langle w, v\rangle_{\mathbb{R}^d}$.

Proof. (1) $\forall x \in \mathbb{R}^d$, $K(x,\cdot) = \langle x, \cdot\rangle_{\mathbb{R}^d} \in W$.

(2) $\forall f = \langle w, \cdot\rangle_{\mathbb{R}^d} \in W$, $\forall x \in \mathbb{R}^d$, $\langle f, K(x,\cdot)\rangle = \langle w, x\rangle_{\mathbb{R}^d} = f(x)$

(3) $K(x, x')$ is a PSD kernel: $\forall a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) = \sum_{i,j} a_i a_j \langle x_i, x_j\rangle = \langle \sum_i a_i x_i, \sum_j a_j x_j\rangle_{\mathbb{R}^d} = \| \sum_i a_i x_i\|^2_{\mathbb{R}^d} \geq 0$$

□

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 10
Scribe: Aden Forrow — Oct. 13, 2015

Recall the following definitions from last time:

Definition: A function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called a *positive symmetric definite kernel* (PSD kernel) if

1. $\forall x, x' \in \mathcal{X}, K(x, x') = K(x', x)$
2. $\forall n \in \mathbb{Z}_+, \forall x_1, x_2, \ldots, x_n$, the $n \times n$ matrix with entries $K(x_i, x_j)$ is positive definite. Equivalently, $\forall a_1, a_2, \ldots, a_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$$

Definition: Let W be a Hilbert space of functions $\mathcal{X} \mapsto \mathbb{R}$. A symmetric kernel $K(\cdot, \cdot)$ is called a *reproducing kernel* of W if

1. $\forall x \in \mathcal{X}$, the function $K(x, \cdot) \in W$.
2. $\forall x \in \mathcal{X}$, $\forall f \in W$, $\langle f(\cdot), K(x, \cdot)\rangle_W = f(x)$.

If such a $K(x, \cdot)$ exists, W is called a *reproducing kernel Hilbert space* (RKHS).

As before, $\langle \cdot, \cdot \rangle_W$ and $\| \cdot \|_W$ respectively denote the inner product and norm of W. The subscript W will occasionally be omitted. We can think of the elements of W as infinite linear combinations of functions of the form $K(x, \cdot)$. Also note that

$$\langle K(x, \cdot), K(y, \cdot)\rangle_W = K(x, y)$$

Since so many of our tools rely on functions being bounded, we'd like to be able to bound the functions in W. We can do this uniformly over $x \in \mathcal{X}$ if the diagonal $K(x, x)$ is bounded.

Proposition: Let W be a RKHS with PSD K such that $\sup_{x \in \mathcal{X}} K(x, x) = k_{\max}$ is finite. Then $\forall f \in W$,

$$\sup_{x \in \mathcal{X}} |f(x)| \leq \|f\|_W \sqrt{k_{\max}}$$

.

Proof. We rewrite $f(x)$ as an inner product and apply Cauchy-Schwartz.

$$f(x) = \langle f, K(x, \cdot)\rangle_W \leq \|f\|_W \|K(x, \cdot)\|_W$$

Now $\|K(x, \cdot)\|_W^2 = \langle K(x, \cdot), K(x, \cdot)\rangle_W = K(x, x) \leq k_{\max}$. The result follows immediately. □

1.5.2 Risk Bounds for SVM

We now analyze support vector machines (SVM) the same way we analyzed boosting. The general idea is to choose a linear classifier that maximizes the margin (distance to classifiers) while minimizing empirical risk. Classes that are not linearly separable can be embedded in a higher dimensional space so that they are linearly separable. We won't go into that, however; we'll just consider the abstract optimization over a RKHS W.

Explicitly, we minimize the empirical φ-risk over a ball in W with radius λ:

$$\hat{f} = \min_{f \in W, \|f\|_W \leq \lambda} \hat{R}_{n,\varphi}(f)$$

The soft classifier $\hat{f}$ is then turned into a hard classifier $\hat{h} = \text{sign}(\hat{f})$. Typically in SVM φ is the hinge loss, though all our convex surrogates behave similarly. To choose W (the only other free parameter), we choose a PSD $K(x_1, x_2)$ that measures the similarity between two points x_1 and x_2.

As written, this is an intractable minimum over an infinite dimensional ball $\{f, \|f\|_W \leq \lambda\}$. The minimizers, however, will all be contained in a finite dimensional subset.

Theorem: Representer Theorem. Let W be a RKHS with PSD K and let $G : \mathbb{R}^n \mapsto \mathbb{R}$ be any function. Then

$$\begin{aligned}\min_{f \in W, \|f\| \leq \lambda} G(f(x_1), \ldots, f(x_n)) &= \min_{f \in \bar{W}_n, \|f\| \leq \lambda} G(f(x_1), \ldots, f(x_n)) \\ &= \min_{\alpha \in \mathbb{R}^n, \alpha^\top \mathbb{K} \alpha \leq \lambda^2} G(g_\alpha(x_1), \ldots, g_\alpha(x_n)),\end{aligned}$$

where

$$\bar{W}_n = \{f \in W | f(\cdot) = g_\alpha(\cdot) = \sum_{i=1}^n \alpha_i K(x_i, \cdot)\}$$

and $\mathbb{K}_{ij} = K(x_i, x_j)$.

Proof. Since $\bar{W}_n$ is a linear subspace of W, we can decompose any $f \in W$ uniquely as $f = \bar{f} + f^\perp$ with $\bar{f} \in \bar{W}_n$ and $f^\perp \in \bar{W}_n^\perp$. The Pythagorean theorem then gives

$$\|f\|_W^2 = \|\bar{f}\|_W^2 + \|f^\perp\|_W^2$$

Moreover, since $K(x_i, \cdot) \in \bar{W}_n$,

$$f^\perp(x_i) = \langle f^\perp, K(x_i, \cdot) \rangle_W = 0$$

So $f(x_i) = \bar{f}(x_i)$ and

$$G(f(x_1), \ldots, f(x_n)) = G(\bar{f}(x_1), \ldots, \bar{f}(x_n)).$$

Because $f^\perp$ does not contribute to G, we can remove it from the constraint:

$$\min_{f \in W, \|\bar{f}\|^2 + \|f^\perp\|^2 \leq \lambda^2} G(f(x_1), \ldots, f(x_n)) = \min_{f \in W, \|\bar{f}\|^2 \leq \lambda^2} G(\bar{f}(x_1), \ldots, \bar{f}(x_n)).$$

Restricting to $f \in \bar{W}_n$ now does not change the minimum, which gives us the first equality. For the second, we need to show that $\|g_\alpha\|_W \leq \lambda$ is equivalent to $\alpha^\top \mathbb{K} \alpha \leq \lambda^2$.

$$\begin{aligned}
\|g_\alpha\|^2 &= \langle g_\alpha, g_\alpha \rangle \\
&= \langle \sum_{i=1}^n \alpha_i K(x_i, \cdot), \sum_{j=1}^n \alpha_j K(x_j, \cdot) \rangle \\
&= \sum_{i,j=1}^n \alpha_i \alpha_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle \\
&= \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \\
&= \alpha^\top \mathbb{K} \alpha
\end{aligned}$$

□

We've reduced the infinite dimensional problem to a minimization over $\alpha \in \mathbb{R}^n$. This works because we're only interested in G evaluated at a finite set of points. The matrix $\mathbb{K}$ here is a Gram matrix, though we will not not use that. $\mathbb{K}$ should be a measure of the similarity of the points x_i. For example, we could have $W = \{\langle x, \cdot \rangle_{\mathbb{R}^d}, x \in \mathbb{R}^d\}$ with $K(x, y$ the usual inner product $K(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$.

We've shown that $\hat{f}$ only depends on K through $\mathbb{K}$, but does $\hat{R}_{n,\varphi}$ depend on $K(x, y)$ for $x, y \notin \{x_i\}$? It turns out not to:

$$\hat{R}_{n,\varphi} = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i g_\alpha(x_i)) = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i \sum_{j=1}^n \alpha_j K(x_j, x_i)).$$

The last expression only involves $\mathbb{K}$. This makes it easy to encode all the knowledge about our problem that we need. The hard classifier is

$$\hat{h}(x) = \text{sign}(\hat{f}(x)) = \text{sign}(g_{\hat{\alpha}}(x)) = \text{sign}(\sum_{j=1}^n \hat{\alpha}_j K(x_j, x))$$

If we are given a new point x_{n+1}, we need to compute a new column for $\mathbb{K}$. Note that x_{n+1} must be in some way comparable or similar to the previous $\{x_i\}$ for the whole idea of extrapolating from data to make sense.

The expensive part of SVMs is calculating the $n \times n$ matrix $\mathbb{K}$. In some applications, $\mathbb{K}$ may be sparse; this is faster, but still not as fast as deep learning. The minimization over the ellipsoid $\alpha^\top \mathbb{K} \alpha$ requires quadratic programming, which is also relatively slow. In practice, it's easier to solve the Lagrangian form of the problem

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i g_\alpha(x_i)) + \lambda' \alpha^\top \mathbb{K} \alpha$$

This formulation is equivalent to the constrained one. Note that λ and λ' are different.

SVMs have few tuning parameters and so have less flexibility than other methods.

We now turn to analyzing the performance of SVM.

Theorem: Excess Risk for SVM. Let φ be an L-Lipschitz convex surrogate and W a RKHS with PSD K such that $\max_x |K(x,x)| = k_{\max} < \infty$. Let $\hat{h}_{n,\varphi} = \text{sign}\, \hat{f}_{n,\varphi}$, where $\hat{f}_{n,\varphi}$ is the empirical φ-risk minimizer over $\mathcal{F} = \{f \in W. \|f\|_W \leq \lambda\}$ (that is, $\hat{R}_{n,\varphi}(\hat{f}_{n,\varphi}) \leq \hat{R}_{n,\varphi}(f) \forall f \in \mathcal{F}$). Suppose $\lambda\sqrt{k_{\max}} \leq 1$. Then

$$R(\hat{h}_{n,\varphi}) - R^* \leq 2c\left(\inf_{f\in\mathcal{F}}(R_\varphi(f) - R^*_\varphi)\right)^\gamma + 2c\left(8L\lambda\sqrt{\frac{k_{\max}}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{2\log(2/\delta)}{n}}\right)^\gamma$$

with probability $1-\delta$. The constants c and γ are those from Zhang's lemma. For the hinge loss, $c = \frac{1}{2}$ and $\gamma = 1$.

Proof. The first term comes from optimizing over a restricted set $\mathcal{F}$ instead of all classifiers. The third term comes from applying the bounded difference inequality. These arise in exactly the same way as they do for boosting, so we will omit the proof for those parts. For the middle term, we need to show that $R_{n,\varphi}(\mathcal{F}) \leq \lambda\sqrt{\frac{k_{\max}}{n}}$.

First, $|f(x)| \leq \|f\|_W\sqrt{k_{\max}} \leq \lambda\sqrt{k_{\max}} \leq 1$ for all $f \in \mathcal{F}$, so we can use the contraction inequality to replace $R_{n,\varphi}(\mathcal{F})$ with $R_n(\mathcal{F})$. Next we'll expand $f(x_i)$ inside the Rademacher complexity and bound inner products using Cauchy-Schwartz.

$$\begin{aligned}
R_n(\mathcal{F}) &= \sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \\
&= \frac{1}{n}\sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i \langle K(x_i,\cdot), f\rangle\right|\right] \\
&= \frac{1}{n}\sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\langle\sum_{i=1}^n \sigma_i K(x_i,\cdot), f\rangle\right|\right] \\
&\leq \frac{\lambda}{n}\sup_{x_1,\ldots,x_n} \sqrt{\mathbb{E}\left[\|\sum_{i=1}^n \sigma_i K(x_i,\cdot)\|_W^2\right]}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\|\sum_{i=1}^n \sigma_i K(x_i,\cdot)\|_W^2\right] &= \mathbb{E}\left[\langle\sum_{i=1}^n \sigma_i K(x_i,\cdot), \sum_{j=1}^n \sigma_j K(x_j,\cdot)\rangle_W\right] \\
&= \sum_{i,j=1}^n \langle K(x_i,\cdot), K(x_j,\cdot)\rangle \mathbb{E}[\sigma_i\sigma_j] \\
&= \sum_{i,j=1}^n K(x_i,x_j)\delta_{ij} \\
&\leq nk_{\max}
\end{aligned}$$

So $R_n(\mathcal{F}) \leq \lambda\sqrt{\frac{k_{\max}}{n}}$ and we are done with the new parts of the proof. The remainder follows as with boosting, using symmetrization, contraction, the bounded difference inequality, and Zhang's lemma. □

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 11
Scribe: Kevin Li — Oct. 14, 2015

2. CONVEX OPTIMIZATION FOR MACHINE LEARNING

In this lecture, we will cover the basics of convex optimization as it applies to machine learning. There is much more to this topic than will be covered in this class so you may be interested in the following books.

Convex Optimization by Boyd and Vandenberghe
Lecture notes on Convex Optimization by Nesterov
Convex Optimization: Algorithms and Complexity by Bubeck
Online Convex Optimization by Hazan

The last two are drafts and can be obtained online.

2.1 Convex Problems

A convex problem is an optimization problem of the form $\min_{x \in \mathcal{C}} f(x)$ where f and $\mathcal{C}$ are convex. First, we will debunk the idea that convex problems are easy by showing that virtually all optimization problems can be written as a convex problem. We can rewrite an optimization problem as follows.

$$\min_{\mathcal{X} \in \mathcal{X}} f(x) \quad \Leftrightarrow \quad \min_{t \geq f(x), x \in \mathcal{X}} t \quad \Leftrightarrow \quad \min_{(x,t) \in \mathrm{epi}(f)} t$$

where the epigraph of a function is defined by

$$\mathrm{epi}(f) = \{(x,t) \in \mathcal{X} \times \mathbb{R} : t \geq f(x)\}$$

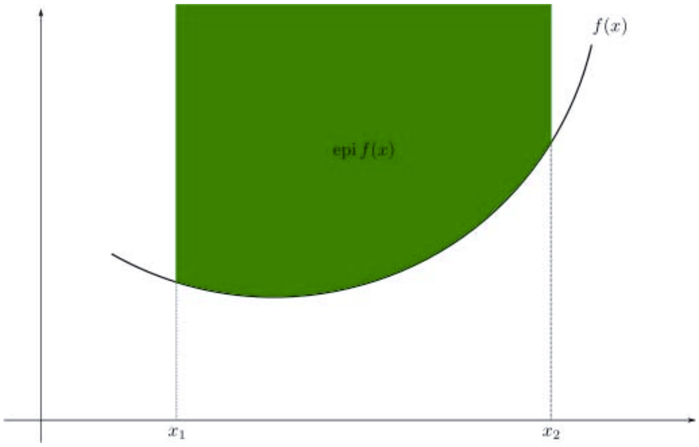


Figure 1: An example of an epigraph.
Source: https://en.wikipedia.org/wiki/Epigraph_(mathematics)

Now we observe that for linear functions,

$$\min_{x\in D} c^\top x = \min_{x\in \mathsf{conv}(D)} c^\top x$$

where the convex hull is defined

$$\mathsf{conv}(D) = \{y : \exists N \in \mathbb{Z}_+, x_1, \dots, x_N \in D, \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1, y = \sum_{i=1}^N \alpha_i x_i\}$$

To prove this, we know that the left side is a least as big as the right side since $D \subset \mathsf{conv}(D)$. For the other direction, we have

$$\begin{aligned}
\min_{x\in \mathsf{conv}(D)} c^\top x &= \min_N \min_{x_1,\dots,x_N\in D} \min_{\alpha_1,\dots,\alpha_N} c^\top \sum_{i=1}^N \alpha_i x_i \\
&= \min_N \min_{x_1,\dots,x_N\in D} \min_{\alpha_1,\dots,\alpha_N} \sum_{i=1}^N \alpha_i c^\top x_i \geq \min_{x\in D} c^\top x \\
&\geq \min_N \min_{x_1,\dots,x_N\in D} \min_{\alpha_1,\dots,\alpha_N} \sum_{i=1}^N \alpha_i \min_{x\in D} c^\top x \\
&= \min_{x\in D} c^\top x
\end{aligned}$$

Therefore we have

$$\min_{x\in\mathcal{X}} f(x) \Leftrightarrow \min_{(x,t)\in \mathsf{conv}(\mathsf{epi}(f))} t$$

which is a convex problem.

Why do we want convexity? As we will show, convexity allows us to infer global information from local information. First, we must define the notion of *subgradient*.

Definition (Subgradient): Let $\mathcal{C} \subset \mathbb{R}^d$, $f : \mathcal{C} \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is called a *subgradient* of f at $x \in \mathcal{C}$ if

$$f(x) - f(y) \leq g^\top (x - y) \qquad \forall y \in \mathcal{C}\,.$$

The set of such vectors g is denoted by $\partial f(x)$.

Subgradients essentially correspond to gradients but unlike gradients, they always exist for convex functions, even when they are not differentiable as illustrated by the next theorem.

Theorem: If $f : \mathcal{C} \to \mathbb{R}$ is convex, then for all x, $\partial f(x) \neq \emptyset$. In addition, if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

Proof. Omitted. Requires separating hyperplanes for convex sets. □

Theorem: Let $f, \mathcal{C}$ be convex. If x is a local minimum of f on $\mathcal{C}$, then it is also global minimum. Furthermore this happens if and only if $0 \in \partial f(x)$.

Proof. $0 \in \partial f(x)$ if and only if $f(x) - f(y) \leq 0$ for all $y \in \mathcal{C}$. This is clearly equivalent to x being a global minimizer.

Next assume x is a local minimum. Then for all $y \in \mathcal{C}$ there exists ε small enough such that $f(x) \leq f\Big((1-\varepsilon)x + \varepsilon y\Big) \leq (1-\varepsilon)f(x) + \varepsilon f(y) \implies f(x) \leq f(y)$ for all $y \in \mathcal{C}$. □

Not only do we know that local minimums are global minimums, looking at the subgradient also tells us where the minimum can be. If $g^\top(x-y) < 0$ then $f(x) < f(y)$. This means $f(y)$ cannot possibly be a minimum so we can narrow our search to ys such that $g^\top(x-y)$. In one dimension, this corresponds to the half line $\{y \in \mathbb{R} : y \leq x\}$ if $g > 0$ and the half line $\{y \in \mathbb{R} : y \geq x\}$ if $g < 0$. This concept leads to the idea of gradient descent.

2.2 Gradient Descent

$y \approx x$ and f differentiable the first order Taylor expansion of f at x yields $f(y) \approx f(x) + g^\top(y-x)$. This means that

$$\min_{|\hat{\mu}|_2=1} f(x+\varepsilon\hat{\mu}) \approx \min f(x) + g^\top(\varepsilon\hat{\mu})$$

which is minimized at $\hat{\mu} = -\frac{g}{|g|_2}$. Therefore to minimizes the linear approximation of f at x, one should move in direction opposite to the gradient.

Gradient descent is an algorithm that produces a sequence of points $\{x_j\}_{j\geq 1}$ such that (hopefully) $f(x_{j+1}) < f(x_j)$.

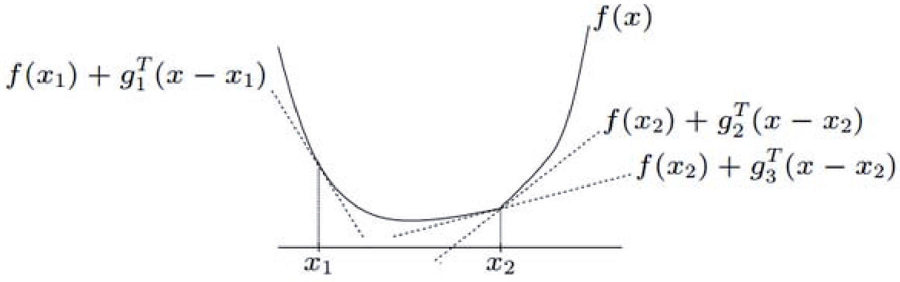

$f(x)$
$f(x_1) + g_1^T(x - x_1)$
$f(x_2) + g_2^T(x - x_2)$
$f(x_2) + g_3^T(x - x_2)$
x_1
x_2

Figure 2: Example where the subgradient of x_1 is a singleton and and the subgradient of x_2 contains multiple elements.
Source: https://optimization.mccormick.northwestern.edu/index.php/Subgradient_optimization

Algorithm 1 Gradient Descent algorithm

```
Input: $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$
for $s = 1$ to $k-1$ do
  $x_{s+1} = x_s - \eta_s g_s \,, \quad g_s \in \partial f(x_s)$
end for
return Either $\bar{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \underset{x\in\{x_1,\dots,x_k\}}{\operatorname{argmin}} f(x)$
```

Theorem: Let f be a convex L-Lipschitz function on $\mathbb{R}^d$ such that $x^* \in \operatorname{argmin}_{\mathbb{R}^d} f(x)$ exists. Assume that $|x_1 - x^*|_2 \leq R$. Then if $\eta_s = \eta = \frac{R}{L\sqrt{k}}$ for all $s \geq 1$, then

$$f(\frac{1}{k}\sum_{s=1}^{k} x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

and

$$\min_{1\leq s\leq k} f(x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Proof. Using the fact that $g_s = \frac{1}{\eta}(x_{s+1} - x_s)$ and the equality $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$,

$$\begin{aligned} f(x_s) - f(x^*) &\leq g_s^\top (x_s - x^*) = \frac{1}{\eta}(x_s - x_{s+1})^\top (x_s - x^*) \\ &= \frac{1}{2\eta}\Big[\|x_s - x_{s+1}\|^2 + \|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\Big] \\ &= \frac{\eta}{2}\|g_s\|^2 + \frac{1}{2\eta}(\delta_s^2 - \delta_{s+1}^2) \end{aligned}$$

where we have defined $\delta_s = \|x_s - x^*\|$. Using the Lipschitz condition

$$f(x_s) - f(x^*) \leq \frac{\eta}{2}L^2 + \frac{1}{2\eta}(\delta_s^2 - \delta_{s+1}^2)$$

Taking the average from 1, to k we get

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{\eta}{2}L^2 + \frac{1}{2k\eta}(\delta_1^2 - \delta_{s+1}^2) \leq \frac{\eta}{2}L^2 + \frac{1}{2k\eta}\delta_1^2 \leq \frac{\eta}{2}L^2 + \frac{R^2}{2k\eta}$$

Taking $\eta = \frac{R}{L\sqrt{k}}$ to minimize the expression, we obtain

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Noticing that the left-hand side of the inequality is larger than both $f(\sum_{s=1}^{k} x_s) - f(x^*)$ by Jensen's inequality and $\min_{1\leq s\leq k} f(x_s) - f(x^*)$ respectively, completes the proof. □

One flaw with this theorem is that the step size depends on k. We would rather have step sizes η_s that does not depend on k so the inequalities hold for all k. With the new step sizes,

$$\sum_{s=1}^{k} \eta_s[f(x_s) - f(x^*)] \leq \sum_{s=1}^{k} \frac{\eta_s^2}{2} L^2 + \frac{1}{2}\sum_{s=1}^{k}(\delta_s^2 - \delta_{s+1}^2) \leq \Big(\sum_{s=1}^{k} \eta_s^2\Big)\frac{L}{2} + \frac{R^2}{2}$$

After dividing by $\sum_{s=1}^{k} \eta_s$, we would like the right-hand side to approach 0. For this to happen we need $\frac{\sum \eta_s^2}{\sum \eta_s} \to 0$ and $\sum \eta_s \to \infty$. One candidate for the step size is $\eta_s = \frac{G}{\sqrt{s}}$ since then $\sum\limits_{s=1}^{k} \eta_s^2 \leq c_1 G^2 \log(k)$ and $\sum\limits_{s=1}^{k} \eta_s \geq c_2 G\sqrt{k}$. So we get

$$\Big(\sum_{s=1}^{k} \eta_s\Big)^{-1} \sum_{s=1}^{k} \eta_s[f(x_s) - f(x^*)] \leq \frac{c_1 GL \log k}{2c_2\sqrt{k}} + \frac{R^2}{2c_2 G\sqrt{k}}$$

Choosing G appropriately, the right-hand side approaches 0 at the rate of $LR\sqrt{\frac{\log k}{k}}$. Notice that we get an extra factor of $\sqrt{\log k}$. However, if we look at the sum from $k/2$ to k instead of 1 to k, $\sum\limits_{s=\frac{k}{2}}^{k} \eta_s^2 \leq c_1' G^2$ and $\sum\limits_{s=1}^{k} \eta_s \geq c_2' G\sqrt{k}$. Now we have

$$\min_{1 \leq s \leq k} f(x_s) - f(x^*) \leq \min_{\frac{k}{2} \leq s \leq k} f(x_s) - f(x^*) \leq \Big(\sum_{s=\frac{k}{2}}^{k} \eta_s\Big)^{-1} \sum_{s=\frac{k}{2}}^{k} \eta_s[f(x_s) - f(x^*)] \leq \frac{cLR}{\sqrt{k}}$$

which is the same rate as in the theorem and the step sizes are independent of k.

Important Remark: Note this rate only holds if we can ensure that $|x_{k/2} - x^*|_2 \leq R$ since we have replaced x_1 by $x_{k/2}$ in the telescoping sum. In general, this is not true for gradient descent, but it will be true for *projected gradient descent* in the next lecture.

One final remark is that the dimension d does not appear anywhere in the proof. However, the dimension does have an effect because for larger dimensions, the conditions f is L-Lipschitz and $|x_1 - x^*|_2 \leq R$ are stronger conditions in higher dimensions.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 12
Scribe: Michael Traub — Oct. 19, 2015

2.3 Projected Gradient Descent

In the original gradient descent formulation, we hope to optimize $\min_{x\in\mathcal{C}} f(x)$ where $\mathcal{C}$ and f are convex, but we did not constrain the intermediate x_k. Projected gradient descent will incorporate this condition.

2.3.1 Projection onto Closed Convex Set

First we must establish that it is possible to always be able to keep x_k in the convex set $\mathcal{C}$. One approach is to take the closest point $\pi(x_k) \in \mathcal{C}$.

Definition: Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^d$. Then $\forall\, x \in \mathbb{R}^d$, let $\pi(x) \in \mathcal{C}$ be the minimizer of

$$\|x - \pi(x)\| = \min_{z\in\mathcal{C}} \|x - z\|$$

where $\|\cdot\|$ denotes the Euclidean norm. Then $\pi(x)$ is unique and,

$$\langle \pi(x) - x, \pi(x) - z\rangle \le 0 \quad \forall\, z \in \mathcal{C} \tag{2.1}$$

Proof. From the definition of $\pi := \pi(x)$, we have $\|x-\pi\|^2 \le \|x - v\|^2$ for any $v \in \mathcal{C}$. Fix $w \in \mathcal{C}$ and define $v = (1-t)\pi + tw$ for $t \in (0,1]$. Observe that since $\mathcal{C}$ is convex we have $v \in \mathcal{C}$ so that

$$\|x-\pi\|^2 \le \|x-v\|^2 = \|x - \pi - t(w-\pi)\|^2$$

Expanding the right-hand side yields

$$\|x-\pi\|^2 \le \|x-\pi\|^2 - 2t\,\langle x-\pi, w-\pi\rangle + t^2\,\|w-\pi\|^2$$

This is equivalent to

$$\langle x-\pi, w-\pi\rangle \le t\,\|w-\pi\|^2$$

Since this is valid for all $t \in (0,1)$, letting $t \to 0$ yields (2.1).

Proof of Uniqueness. Assume π_1, $\pi_2 \in \mathcal{C}$ satisfy

$$\langle \pi_1 - x, \pi_1 - z\rangle \le 0 \quad \forall\, z \in C$$
$$\langle \pi_2 - x, \pi_2 - z\rangle \le 0 \quad \forall\, z \in C$$

Taking $z = \pi_2$ in the first inequality and $z = \pi_1$ in the second, we get

$$\langle \pi_1 - x, \pi_1 - \pi_2\rangle \le 0$$
$$\langle x - \pi_2, \pi_1 - \pi_2\rangle \le 0$$

Adding these two inequalities yields $\|\pi_1 - \pi_2\|^2 \le 0$ so that $\pi_1 = \pi_2$. □

2.3.2 Projected Gradient Descent

Algorithm 1 Projected Gradient Descent algorithm

```
  Input: $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$
  for $s = 1$ to $k-1$ do
    $y_{s+1} = x_s - \eta_s g_s\,, \quad g_s \in \partial f(x_s)$
    $x_{s+1} = \pi(y_{s+1})$
  end for
  return Either $\bar{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \underset{x\in\{x_1,\ldots,x_k\}}{\text{argmin}} f(x)$
```

Theorem: Let $\mathcal{C}$ be a closed, nonempty convex subset of $\mathbb{R}^d$ such that $\text{diam}(\mathcal{C}) \leq R$. Let f be a convex L-Lipschitz function on $\mathcal{C}$ such that $x^* \in \text{argmin}_{x\in\mathcal{C}} f(x)$ exists. Then if $\eta_s \equiv \eta = \frac{R}{L\sqrt{k}}$ then

$$f(\bar{x}) - f(x^*) \leq \frac{LR}{\sqrt{k}} \quad \text{ and } \quad f(\bar{x}^\circ) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Moreover, if $\eta_s = \frac{R}{L\sqrt{s}}$, then $\exists c > 0$ such that

$$f(\bar{x}) - f(x^*) \leq c\frac{LR}{\sqrt{k}} \quad \text{ and } \quad f(\bar{x}^\circ) - f(x^*) \leq c\frac{LR}{\sqrt{k}}$$

Proof. Again we will use the identity that $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$.

By convexity, we have

$$\begin{aligned} f(x_s) - f(x^*) &\leq g_s^\top(x_s - x^*) \\ &= \frac{1}{\eta}(x_s - y_{s+1})^\top(x_s - x^*) \\ &= \frac{1}{2\eta}\left[\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right] \end{aligned}$$

Next,

$$\begin{aligned} \|y_{s+1} - x^*\|^2 &= \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2\left\langle y_{s+1} - x_{s+1}, x_{s+1} - x^*\right\rangle \\ &= \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2\left\langle y_{s+1} - \pi(y_{s+1}), \pi(y_{s+1}) - x^*\right\rangle \\ &\geq \|x_{s+1} - x^*\|^2 \end{aligned}$$

where we used that $\langle x - \pi(x), \pi(x) - z\rangle \geq 0 \ \forall\, z \in \mathcal{C}$, and $x^* \in \mathcal{C}$. Also notice that $\|x_s - y_{s+1}\|^2 = \eta^2\|g_s\|^2 \leq \eta^2 L^2$ since f is L-Lipschitz with respect to $\|\cdot\|$. Using this we find

$$\begin{aligned} \frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) &\leq \frac{1}{k}\sum_{s=1}^{k}\frac{1}{2\eta}\left[\eta^2L^2 + \|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right] \\ &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k}\|x_1 - x^*\|^2 \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k} \end{aligned}$$

Minimizing over η we get $\frac{L^2}{2} = \frac{R^2}{2\eta^2 k} \implies \eta = \frac{R}{L\sqrt{k}}$, completing the proof

$$f(\bar{x}) - f(x^*) \leq \frac{RL}{\sqrt{k}}$$

Moreover, the proof of the bound for $f(\sum_{s=\frac{k}{2}}^{k} x_s) - f(x^*)$ is identical because $\left\| x_{\frac{k}{2}} - x^* \right\|^2 \leq R^2$ as well. □

2.3.3 Examples

Support Vector Machines

The SVM minimization as we have shown before is

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ \alpha^\top \mathbb{K} \alpha \leq C^2}} \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - Y_i f_\alpha(X_i)\right)$$

where $f_\alpha(X_i) = \alpha^\top \mathbb{K} e_i = \sum_{j=1}^{n} \alpha_j K(X_j, X_i)$. For convenience, call $g_i(\alpha) = \max\left(0, 1 - Y_i f_\alpha(X_i)\right)$. In this case executing the projection onto the ellipsoid $\{\alpha : \alpha^\top \mathbb{K} \alpha \leq C^2\}$ is not too hard, but we do not know about C, R, or L. We must determine these we can know that our bound is not exponential with respect to n. First we find L and start with the gradient of $g_i(\alpha)$:

$$\nabla g_i(\alpha) = \mathbb{I}(1 - Y_i f_\alpha(X_i) \geq 0) Y_i \mathbb{K} e_i$$

With this we bound the gradient of the φ-risk $\hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} g_i(\alpha)$.

$$\left\| \frac{\partial}{\partial \alpha} \hat{R}_{n,\varphi}(f_\alpha) \right\| = \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\alpha) \right\| \leq \frac{1}{n} \sum_{i=1}^{n} \|\mathbb{K} e_i\|_2$$

by the triangle inequality and the fact that that $\mathbb{I}(1 - Y_i f_\alpha(X_i) \geq 0) Y_i \leq 1$. We can now use the properties of our kernel K. Notice that $\|\mathbb{K} e_i\|$ is the ℓ_2 norm of the i^{th} column so $\|\mathbb{K} e_i\|_2 = \left(\sum_{j=1}^{n} K(X_j, X_i)^2 \right)^{\frac{1}{2}}$. We also know that

$$K(X_j, X_i)^2 = \langle K(X_j, \cdot), K(X_i, \cdot) \rangle \leq \|K(X_j, \cdot)\|_H \, \|K(X_i, \cdot)\|_H \leq k_{\max}^2$$

Combining all of these we get

$$\left\| \frac{\partial}{\partial \alpha} \hat{R}_{n,\varphi}(f_\alpha) \right\| \leq \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} k_{\max}^2 \right)^{\frac{1}{2}} = k_{\max} \sqrt{n} = L$$

To find R we try to evaluate $\operatorname{diam}\{\alpha^\top \mathbb{K} \alpha \leq C^2\} = 2 \max_{\alpha^\top \mathbb{K} \alpha \leq C^2} \sqrt{\alpha^\top \alpha}$. We can use the condition to put bounds on the diameter

$$C^2 \geq \alpha^\top \mathbb{K} \alpha \geq \lambda_{\min}(\mathbb{K}) \alpha^\top \alpha \implies \operatorname{diam}\{\alpha^\top \mathbb{K} \alpha \leq C^2\} \leq \frac{2C}{\sqrt{\lambda_{\min}(\mathbb{K})}}$$

We need to understand how small $\lambda_{\min}$ can get. While it is true that these exist random samples selected by an adversary that make $\lambda_{\min} = 0$, we will consider a random sample of

$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_d)$. This we can write these d-dimensional samples as a $d \times n$ matrix $\mathbb{X}$. We can rewrite the matrix $\mathbb{K}$ with entries $\mathbb{K}_{ij} = K(X_i, X_j) = \langle X_i, X_j \rangle_{\mathbb{R}^d}$ as a Wishart matrix $\mathbb{K} = \mathbb{X}^\top \mathbb{X}$ (in particular, $\frac{1}{d}\mathbb{X}^\top \mathbb{X}$ is Wishart). Using results from random matrix theory, if we take $n, d \to \infty$ but hold $\frac{n}{d}$ as a constant γ, then $\lambda_{\min}(\frac{\mathbb{K}}{d}) \to (1 - \sqrt{\gamma})^2$. Taking an approximation since we cannot take n, d to infinity, we get

$$\lambda_{\min}(\mathbb{K}) \simeq d\left(1 - 2\sqrt{\frac{n}{d}}\right) \geq \frac{d}{2}$$

using the fact that $d \gg n$. This means that $\lambda_{\min}$ becoming too small is not a problem when we model our samples as coming from multivariate Gaussians.

Now we turn our focus to the number of iterations k. Looking at our bound on the excess risk

$$\hat{R}_{n,\varphi}(f_{\alpha_R^\circ}) \leq \min_{\alpha^\top \mathbb{K}\alpha \leq C^2} \hat{R}_{n,\varphi}(f_\alpha) + C\sqrt{\frac{n}{k\lambda_{\min}(\mathbb{K})}} k_{\max}$$

we notice that our all of the constants in our stochastic term can be computed given the number of points and the kernel. Since statistical error is often $\frac{1}{\sqrt{n}}$, to be generous we want to have precision up to $\frac{1}{n}$ to allow for fast rates in special cases. This gives us

$$k \geq \frac{n^3 k_{\max}^2 C^2}{\lambda_{\min}(\mathbb{K})}$$

which is not bad since n is often not very big.

In [Bub15], the rates for many a wide rage of problems with various assumptions are available. For example, if we assume strong convexity and Lipschitz we can get an exponential rate so $k \sim \log n$. If gradient is Lipschitz, then we get get $\frac{1}{k}$ instead of $\frac{1}{\sqrt{k}}$ in the bound. However, often times we are not optimizing over functions with these nice properties.

Boosting

We already know that φ is L-Lipschitz for boosting because we required it before. Remember that our optimization problem is

$$\min_{\substack{\alpha \in \mathbb{R}^N \\ |\alpha|_1 \leq 1}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f_\alpha(X_i))$$

where $f_\alpha = \sum_{j=1}^{N} \alpha_j f_j$ and f_j is the j^{th} weak classifier. Remember before we had some rate like $c\sqrt{\frac{\log N}{n}}$ and we would hope to get some other rate that grows with $\log N$ since N can be very large. Taking the gradient of the φ-loss in this case we find

$$\nabla \hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{N} \varphi'(-Y_i f_\alpha(X_i))(-Y_i) F(X_i)$$

where $F(x)$ is the column vector $[f_1(x), \ldots, f_N(x)]^\top$. Since $|Y_i| \leq 1$ and $\varphi' \leq L$, we can bound the ℓ_2 norm of the gradient as

$$\begin{aligned}\left\|\nabla \hat{R}_{n,\varphi}(f_\alpha)\right\|_2 &\leq \frac{L}{n}\left\|\sum_{i=1}^{n} F(X_i)\right\| \\ &\leq \frac{L}{n}\sum_{i=1}^{n} \|F(X_i)\| \leq L\sqrt{N}\end{aligned}$$

using triangle inequality and the fact that $F(X_i)$ is a N-dimensional vector with each component bounded in absolute value by 1.

Using the fact that the diameter of the ℓ_1 ball is 2, $R = 2$ and the Lipschitz associated with our φ-risk is $L\sqrt{N}$ where L is the Lipschitz constant for φ. Our stochastic term $\frac{RL}{\sqrt{k}}$ becomes $2L\sqrt{\frac{N}{k}}$. Imposing the same $\frac{1}{n}$ error as before we find that $k \sim N^2 n$, which is very bad especially since we want $\log N$.

2.4 Mirror Descent

Boosting is an example of when we want to do gradient descent on a non-Euclidean space, in particular a ℓ_1 space. While the dual of the ℓ_2-norm is itself, the dual of the ℓ_1 norm is the ℓ_∞ or sup norm. We want this appear if we have an ℓ_1 constraint. The reason for this is not intuitive because we are taking about measures on the same space $\mathbb{R}^d$, but when we consider optimizations on other spaces we want a procedure that does is not indifferent to the measure we use. Mirror descent accomplishes this.

2.4.1 Bregman Projections

Definition: If $\|\cdot\|$ is some norm on $\mathbb{R}^d$, then $\|\cdot\|_*$ is its dual norm.

Example: If dual norm of the ℓ_p norm $\|\cdot\|_p$ is the ℓ_q norm $\|\cdot\|_q$, then $\frac{1}{p} + \frac{1}{q} = 1$. This is the limiting case of Hölder's inequality.

In general we can also refine our bounds on inner products in $\mathbb{R}^d$ to $x^\top y \leq \|x\| \, \|y\|_*$ if we consider x to be the primal and y to be the dual. Thinking like this, gradients live in the dual space, e.g. in $g_s^\top(x - x^*)$, $x - x^*$ is in the primal space, so g_s is in the dual. The transpose of the vectors suggest that these vectors come from spaces with different measure, even though all the vectors are in $\mathbb{R}^d$.

Definition: Convex function Φ on a convex set D is said to be
(i) L-Lipschitz with respect to $\|\cdot\|$ if $\|g\|_* \leq L \quad \forall g \in \partial\Phi(x) \quad \forall x \in D$
(ii) α-strongly convex with respect to $\|\cdot\|$ if

$$\Phi(y) \geq \Phi(x) + g^\top(y - x) + \frac{\alpha}{2}\|y - x\|^2$$

for all $x, y \in D$ and for $g \in \partial f(x)$

Example: If Φ is twice differentiable with Hessian H and $\|\cdot\|$ is the ℓ_2 norm, then all $\text{eig}(H) \geq \alpha$.

Definition (Bregman divergence): For a given convex function Φ on a convex set $\mathcal{D}$ with $x, y \in \mathcal{D}$, the Bregman divergence of y from x is defined as

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla\Phi(x)^\top(y - x)$$

This divergence is the error of the function $\Phi(y)$ from the linear approximation at x. Also note that this quantity is not symmetric with respect to x and y. If Φ is convex then $D_\Phi(y,x) \geq 0$ because the Hessian is positive semi-definite. If Φ is α-strongly convex then $D_\Phi(y,x) \geq \frac{\alpha}{2} \left\|y-x\right\|^2$ and if the quadratic approximation is good then this approximately holds in equality and this divergence behaves like Euclidean norm.

Proposition: Given convex function Φ on $\mathcal{D}$ with $x, y, z \in \mathcal{D}$

$$\left(\nabla\Phi(x) - \nabla\Phi(y)\right)^\top (x-z) = D_\Phi(x,y) + D_\Phi(z,x) - D_\Phi(z,y)$$

Proof. Looking at the right hand side

$$\begin{aligned}
&= \Phi(x) - \Phi(y) - \nabla\Phi(y)^\top(x-y) + \Phi(z) - \Phi(x) - \nabla\Phi(x)^\top(z-x) \\
&\qquad\qquad - \Big[\Phi(z) - \Phi(y) - \nabla\Phi(y)^\top(z-y)\Big] \\
&= \nabla\Phi(y)^\top(y-x+z-y) - \nabla\Phi(x)^\top(z-x) \\
&= \left(\nabla\Phi(x) - \nabla\Phi(y)\right)^\top (x-z)
\end{aligned}$$

□

Definition (Bregman projection): Given $x \in \mathbb{R}^d$, Φ a convex differentiable function on $\mathcal{D} \subset \mathbb{R}^d$ and convex $C \subset \bar{\mathcal{D}}$, the Bregman projection of x with respect to Φ is

$$\pi^\Phi(x) \in \underset{z \in C}{\operatorname{argmin}}\, D_\phi(x,z)$$

References

[Bub15] Sébastien Bubeck, *Convex optimization: algorithms and complexity*, Now Publishers Inc., 2015.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 13
Scribe: Mina Karzand Oct. 21, 2015

Previously, we analyzed the convergence of the projected gradient descent algorithm. We proved that optimizing the convex L-Lipschitz function f on a closed, convex set $\mathcal{C}$ with $\mathrm{diam}(\mathcal{C}) \leq R$ with step sizes $\eta_s = \frac{R}{L\sqrt{k}}$ would give us accuracy of $f(\overline{x}) \leq f(x^*) + \frac{LR}{\sqrt{k}}$ after k iterations.

Although it might seem that projected gradient descent algorithm provides dimension-free convergence rate, it is not always true. Reviewing the proof of convergence rate, we realize that dimension-free convergence is possible when the objective function f and the constraint set $\mathcal{C}$ are well-behaved in Euclidean norm (i.e., for all $x \in \mathcal{C}$ and $g \in \partial f(x)$, we have that $|x|_2$ and $|g|_2$ are independent of the ambient dimension). We provide an examples of the cases that these assumptions are not satisfied.

- Consider the differentiable, convex function f on the Euclidean ball $B_{2,n}$ such that $\|\nabla f(x)\|_\infty \leq 1,\ \forall x \in B_{2,n}$. This implies that $|\nabla f(x)|_2 \leq \sqrt{n}$ and the projected gradient descent converges to the minimum of f in $B_{2,n}$ at rate $\sqrt{\frac{n}{k}}$. Using the method of mirror descent we can get convergence rate of $\sqrt{\frac{\log(n)}{k}}$

To get better rates of convergence in the optimization problem, we can use the Mirror Descent algorithm. The idea is to change the Euclidean geometry to a more pertinent geometry to a problem at hand. We will define a new geometry by using a function which is sometimes called potential function $\Phi(x)$. We will use Bregman projection based on Bregman divergence to define this geometry.

The geometric intuition behind the mirror Descent algorithm is the following: The projected gradient described in previous lecture works in any arbitrary Hilbert space $\mathcal{H}$ so that the norm of vectors is associated with an inner product. Now, suppose we are interested in optimization in a Banach space $\mathcal{D}$. In other words, the norm (or the measure of distance) that we use does not derive from an inner product. In this case, the gradient descent does not even make sense since the gradient $\nabla f(x)$ are elements of dual space. Thus, the term $x - \eta \nabla f(x)$ cannot be performed. (Note that in Hilbert space used in projected gradient descent, the dual space of $\mathcal{H}$ is isometric to $\mathcal{H}$. Thus, we didn't have any such problems.)

The geometric insight of the Mirror Descent algorithm is that to perform the optimization in the primal space $\mathcal{D}$, one can first map the point $x \in \mathcal{D}$ in primal space to the dual space $\mathcal{D}^*$, then perform the gradient update in the dual space and finally map the optimal point back to the primal space. Note that at each update step, the new point in the primal space $\mathcal{D}$ might be outside of the constraint set $\mathcal{C} \subset \mathcal{D}$, in which case it should be projected into the constraint set $\mathcal{C}$. The projection associate with the Mirror Descent algorithm is Bergman Projection defined based on the notion of Bergman divergence.

Definition (Bregman Divergence): For given differentiable, α-strongly convex function $\Phi(x) : \mathcal{D} \to \mathbb{R}$, we define the Bregman divergence associated with Φ to be:

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x)$$

We will use the convex open set $\mathcal{D} \subset \mathbb{R}^n$ whose closure contains the constraint set $\mathcal{C} \subset \overline{\mathcal{D}}$. Bregman divergence is the error term of the first order Taylor expansion of the function Φ in $\mathcal{D}$.

Also, note that the function $\Phi(x)$ is said to be α-strongly convex w.r.t. a norm $\|.\|$ if

$$\Phi(y) - \Phi(x) - \nabla\Phi(x)^T(y-x) \geq \frac{\alpha}{2}\|y-x\|^2 .$$

We used the following property of the Euclidean norm:

$$2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$$

in the proof of convergence of projected gradient descent, where we chose $a = x_s - y_{s+1}$ and $b = x_s - x^*$.

To prove the convergence of the Mirror descent algorithm, we use the following property of the Bregman divergence in a similar fashion. This proposition shows that the Bregman divergence essentially behaves as the Euclidean norm squared in terms of projections:

Proposition: Given α-strongly differentiable convex function $\Phi : \mathcal{D} \to \mathbb{R}$, for all $x, y, z \in \mathcal{D}$,

$$[\nabla\Phi(x) - \nabla\Phi(y)]^\top (x-z) = D_\Phi(x,y) + D_\Phi(z,x) - D_\Phi(z,y) .$$

As described previously, the Bregman divergence is used in each step of the Mirror descent algorithm to project the updated value into the constraint set.

Definition (Bregman Projection): Given α-strongly differentiable convex function $\Phi : \mathcal{D} \to \mathbb{R}$ and for all $x \in \mathcal{D}$ and closed convex set $\mathcal{C} \subset \overline{\mathcal{D}}$

$$\Pi_{\mathcal{C}}^{\Phi}(x) = \underset{z \in \mathcal{C}\cap\mathcal{D}}{\operatorname{argmin}} D_\Phi(z,x)$$

2.4.2 Mirror Descent Algorithm

Algorithm 1 Mirror Descent algorithm

```
  Input: $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$, $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\zeta(x) = \nabla\Phi(x)$
  for $s = 1, \cdots, k$ do
    $\zeta(y_{s+1}) = \zeta(x_s) - \eta g_s$ for $g_s \in \partial f(x_s)$
    $x_{s+1} = \Pi_{\mathcal{C}}^{\Phi}(y_{s+1})$
  end for
  return  Either $\overline{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \operatorname{argmin}_{x\in\{x_1,\cdots,x_k\}} f(x)$
```

Proposition: Let $z \in \mathcal{C} \cap \mathcal{D}$, then $\forall y \in \mathcal{D}$,

$$(\nabla\Phi(\pi(y) - \nabla\Phi(y))^\top (\pi(y) - z) \leq 0$$

Moreover, $D_\Phi(z, \pi(y)) \le D_\Phi(z, y)$.

Proof. Define $\pi = \Pi_{\mathcal{C}}^{\Phi}(y)$ and $h(t) = D_\Phi(\pi + t(z - \pi), y)$. Since $h(t)$ is minimized at $t = 0$ (due to the definition of projection), we have

$$h'(0) = \nabla_x D_\Phi(x, y)|_{x=\pi}(z - \pi) \ge 0$$

where suing the definition of Bregman divergence,

$$\nabla_x D_\Phi(x, y) = \nabla\Phi(x) - \nabla\Phi(y)$$

Thus,

$$(\nabla\Phi(\pi) - \nabla\Phi(y))^\top (\pi - z) \le 0\,.$$

Using proposition 1, we know that

$$(\nabla\Phi(\pi) - \nabla\Phi(y))^\top (\pi - z) = D_\Phi(\pi, y) + D_\Phi(z, \pi) - D_\Phi(z, y) \le 0\,,$$

and since $D_\Phi(\pi, y) \ge 0$, we would have $D_\Phi(z, \pi) \le D_\Phi(z, y)$. □

Theorem: Assume that f is convex and L-Lipschitz w.r.t. $\|.\|$. Assume that Φ is α-strongly convex on $\mathcal{C} \cap\ \mathcal{D}$ w.r.t. $\|.\|$ and

$$R^2 = \sup_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x) - \min_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x)$$

take $x_1 = \operatorname{argmin}_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$ (assume that it exists). Then, Mirror Descent with $\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{R}}$ gives,

$$f(\overline{x}) - f(x^*) \le RL\sqrt{\frac{2}{\alpha k}} \quad \text{and} \quad f(\overline{x}^\circ) - f(x^*) \le RL\sqrt{\frac{2}{\alpha k}},$$

Proof. Take $x^\sharp \in \mathcal{C} \cap \mathcal{D}$. Similar to the proof of the projected gradient descent, we have:

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\overset{(i)}{\le} g_s^\top (x_s - x^\sharp) \\
&\overset{(ii)}{=} \frac{1}{\eta}\left(\zeta(x_s) - \zeta(y_{s+1})\right)^\top (x_s - x^\sharp) \\
&\overset{(iii)}{=} \frac{1}{\eta}\left(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right)^\top (x_s - x^\sharp) \\
&\overset{(iv)}{=} \frac{1}{\eta}\Big[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1})\Big] \\
&\overset{(v)}{\le} \frac{1}{\eta}\Big[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\Big] \\
&\overset{(vi)}{\le} \frac{\eta L^2}{2\alpha^2} + \frac{1}{\eta}\Big[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\Big]
\end{aligned}$$

Where (i) is due to convexity of the function f.

Equations (ii) and (iii) are direct results of Mirror descent algorithm.
Equation (iv) is the result of applying proposition 1.
Inequality (v) is a result of the fact that $x_{s+1} = \Pi_{\mathcal{C}}^{\Phi}(y_{s+1})$, thus for $x^\sharp \in \mathcal{C} \cap \mathcal{D}$, we have $D_\Phi(x^\sharp, y_{s+1}) \geq D_\Phi(x^\sharp, x_{s+1})$.
We will justify the following derivations to prove inequality (vi):

$$\begin{aligned}
D_\Phi(x_s, y_{s+1}) &\overset{(a)}{=} \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1}) \\
&\overset{(b)}{\leq} \left[\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right]^\top (x_s - y_{s+1}) - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\overset{(c)}{\leq} \eta\|g_s\|_*\|x_s - y_{s+1}\| - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\overset{(d)}{\leq} \frac{\eta^2 L^2}{2\alpha}\,.
\end{aligned}$$

Equation (a) is the definition of Bregman divergence.
To show inequality (b), we used the fact that Φ is α-strongly convex which implies that $\Phi(y_{s+1}) - \Phi(x_s) \geq \nabla\Phi(x_s)^T(y_{s+1} - x_s)\frac{\alpha}{2}\|y_{s+1} - x_s\|^2$.
According to the Mirror descent algorithm, $\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}) = \eta g_s$. We use Hölder's inequality to show that $g_s^\top(x_s - y_{s+1}) \leq \|g_s\|_*\|x_s - y_{s+1}\|$ and derive inequality (c).
Looking at the quadratic term $ax - bx^2$ for $a, b > 0$, it is not hard to show that $\max ax - bx^2 = \frac{a^2}{4b}$. We use this statement with $x = \|y_{s+1} - x_s\|$, $a = \eta\|g_s\|_* \leq L$ and $b = \frac{\alpha}{2}$ to derive inequality (d).

Again, we use telescopic sum to get

$$\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \leq \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x^\sharp, x_1)}{k\eta}\,. \tag{2.1}$$

We use the definition of Bregman divergence to get

$$\begin{aligned}
D_\Phi(x^\sharp, x_1) &= \Phi(x^\sharp) - \Phi(x_1) - \nabla\Phi(x_1)(x^\sharp - x_1) \\
&\leq \Phi(x^\sharp) - \Phi(x_1) \\
&\leq \sup_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x) - \min_{x \in \mathcal{C}\cap\mathcal{D}} \Phi(x) \\
&\leq R^2\,.
\end{aligned}$$

Where we used the fact $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$ in the description of the Mirror Descent algorithm to prove $\nabla\Phi(x_1)(x^\sharp - x_1) \geq 0$. We optimize the right hand side of equation (2.1) for η to get

$$\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \leq RL\sqrt{\frac{2}{\alpha k}}\,.$$

To conclude the proof, let $x^\sharp \to x^* \in \mathcal{C}$. □

Note that with the right geometry, we can get projected gradient descent as an instance the Mirror descent algorithm.

2.4.3 Remarks

The Mirror Descent is sometimes called Mirror Prox. We can write x_{s+1} as

$$\begin{aligned} x_{s+1} &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} D_\Phi(x, y_{s+1}) \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - \nabla\Phi^\top(y_{s+1})x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - [\nabla\Phi(x_s) - \eta g_s]^\top x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + \Phi(x) - \nabla\Phi^\top(x_s)x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + D_\Phi(x, x_s) \end{aligned}$$

Thus, we have

$$x_{s+1} = \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + D_\Phi(x, x_s)\,.$$

To get x_{s+1}, in the first term on the right hand side we look at linear approximations close to x_s in the direction determined by the subgradient g_s. If the function is linear, we would just look at the linear approximation term. But if the function is not linear, the linear approximation is only valid in a small neighborhood around x_s. Thus, we penalized by adding the term $D_\Phi(x, x_s)$. We can penalized by the square norm when we choose $D_\Phi(x, x_s) = \|x - x_s\|^2$. In this case we get back the projected gradient descent algorithm as an instance of Mirror descent algorithm.

But if we choose a different divergence $D_\Phi(x, x_s)$, we are changing the geometry and we can penalize differently in different directions depending on the geometry.

Thus, using the Mirror descent algorithm, we could replace the 2-norm in projected gradient descent algorithm by another norm, hoping to get less constraining Lipschitz constant. On the other hand, the norm is a lower bound on the strong convexity parameter. Thus, there is trade off in improvement of rate of convergence.

2.4.4 Examples

Euclidean Setup:

$\Phi(x) = \frac{1}{2}\|x\|^2$, $\quad \mathcal{D} = \mathbb{R}^d$, $\quad \nabla\Phi(x) = \zeta(x) = x$. Thus, the updates will be similar to the gradient descent.

$$\begin{aligned} D_\Phi(y, x) &= \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x\|^2 - x^\top y + \|x\|^2 \\ &= \frac{1}{2}\|x - y\|^2\,. \end{aligned}$$

Thus, Bregman projection with this potential function $\Phi(x)$ is the same as the usual Euclidean projection and the Mirror descent algorithm is exactly the same as the projected descent algorithm since it has the same update and same projection operator.

Note that $\alpha = 1$ since $D_\Phi(y, x) \geq \frac{1}{2}\|x - y\|^2$.

ℓ_1 **Setup:**

We look at $\mathcal{D} = \mathbb{R}_+^d \setminus \{0\}$.

Define $\Phi(x)$ to be the negative entropy so that:

$$\Phi(x) = \sum_{i=1}^{d} x_i \log(x_i), \qquad \zeta(x) = \nabla\Phi(x) = \{1 + \log(x_i)\}_{i=1}^{d}$$

Thus, looking at the update function $y^{(s+1)} = \nabla\Phi(x^{(s)}) - \eta g_s$, we get $\log(y_i^{(s+1)}) = \log(x_i^{(s)}) - \eta g_i^{(s)}$ and for all $i = 1, \cdots, d$, we have $y_i^{(s+1)} = x_i^{(s)} \exp(-\eta g_i^{(s)})$. Thus,

$$y^{(s)} = x^{(s)} \exp(-\eta g^{(s)})\,.$$

We call this setup exponential Gradient Descent or Mirror Descent with multiplicative weights.

The Bregman divergence of this mirror map is given by

$$\begin{aligned}
D_\Phi(y, x) &= \Phi(y) - \Phi(x) - \nabla\Phi^\top(x)(y - x) \\
&= \sum_{i=1}^{d} y_i \log(y_i) - \sum_{i=1}^{d} x_i \log(x_i) - \sum_{i=1}^{d}(1 + \log(x_i))(y_i - x_i) \\
&= \sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i}) + \sum_{i=1}^{d}(y_i - x_i)
\end{aligned}$$

Note that $\sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i})$ is call the Kullback-Leibler divergence (KL-div) between y and x.

We show that the projection with respect to this Bregman divergence on the simplex $\Delta_d = \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$ amounts to a simple renormalization $y \mapsto y/|y|_1$. To prove so, we provide the Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i}) + \sum_{i=1}^{d}(x_i - y_i) + \lambda(\sum_{i=1}^{d} x_i - 1)\,.$$

To find the Bregman projection, for all $i = 1, \cdots, d$ we write

$$\frac{\partial}{\partial x_i}\mathcal{L} = -\frac{y_i}{x_i} + 1 + \lambda = 0$$

Thus, for all i, we have $x_i = \gamma y_i$. We know that $\sum_{i=1}^{d} x_i = 1$. Thus, $\gamma = \frac{1}{\sum y_i}$.

Thus, we have $\Pi_{\Delta_d}^{\Phi}(y) = \frac{y}{|y|_1}$. The Mirror Descent algorithm with this update and projection would be:

$$\begin{aligned}
y_{s+1} &= x_s \exp(-\eta g_s) \\
x_{s+1} &= \frac{y}{|y|_1}\,.
\end{aligned}$$

To analyze the rate of convergence, we want to study the ℓ_1 norm on Δ_d. Thus, we have to show that for some α, Φ is α-strongly convex w.r.t $|\cdot|_1$ on Δ_d.

$$\begin{aligned} D_\Phi(y,x) &= KL(y,x) + \sum_i (x_i - y_i) \\ &= KL(y,x) \\ &\geq \frac{1}{2}|x-y|_1^2 \end{aligned}$$

Where we used the fact that $x, y \in \Delta_d$ to show $\sum_i (x_i - y_i) = 0$ and used Pinsker inequality show the result. Thus, Φ is 1-strongly convex w.r.t. $|\cdot|_1$ on Δ_d.

Remembering that $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$ was defined to be negative entropy, we know that $-\log(d) \leq \Phi(x) \leq 0$ for $x \in \Delta_d$. Thus,

$$R^2 = \max_{x \in \Delta_d} \Phi(x) - \min_{x \in \Delta_d} \Phi(x) = \log(d)\,.$$

Corollary: Let f be a convex function on Δ_d such that

$$\|g\|_\infty \leq L, \qquad \forall g \in \partial f(x), \quad \forall x \in \Delta_d\,.$$

Then, Mirror descent with $\eta = \frac{1}{L}\sqrt{\frac{2\log(d)}{k}}$ gives

$$f(\overline{x}_k) - f(x^*) \leq L\sqrt{\frac{2\log(d)}{k}}, \qquad f(x_k^\circ) - f(x^*) \leq L\sqrt{\frac{2\log(d)}{k}}$$

Boosting: For weak classifiers $f_1(x), \cdots, f_N(x)$ and $\alpha \in \Delta_n$, we define

$$f_\alpha = \sum_{j=1}^N \alpha_j f_j \qquad \text{and} \qquad F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_N(x) \end{pmatrix}$$

so that $f_\alpha(x)$ is the weighted majority vote classifier. Note that $|F|_\infty \leq 1$.

As shown before, in boosting, we have:

$$g = \nabla \widehat{R}_{n,\phi}(f_\alpha) = \frac{1}{n}\sum_{i=1}^n \phi'(-y_i f_\alpha(x_i))(-y_i)F(x_i)\,,$$

Since $|F|_\infty \leq 1$ and $|y|_\infty \leq 1$, then $|g|_\infty \leq L$ where L is the Lipschitz constant of ϕ (e.g., a constant like e or 2).

$$\widehat{R}_{n,\phi}(f_{\alpha_k^\circ}) - \min_{\alpha \in \Delta_n} \widehat{R}_{n,\phi}(f_\alpha) \leq L\sqrt{\frac{2\log(N)}{k}}$$

We need the number of iterations $k \approx n^2 \log(N)$.

The functions f_j's could hit all the vertices. Thus, if we want to fit them in a ball, the ball has to be radius $\sqrt{N}$. This is why the projected gradient descent would give the rate of $\sqrt{\frac{N}{k}}$. But by looking at the gradient we can determine the right geometry. In this case, the gradient is bounded by sup-norm which is usually the most constraining norm in projected gradient descent. Thus, using Mirror descent would be most beneficial.

Other Potential Functions:

There are other potential functions which are strongly convex w.r.t ℓ_1 norm. In particular, for

$$\Phi(x) = \frac{1}{p}|x|_p^p, \qquad p = 1 + \frac{1}{\log(d)}$$

then Φ is $c\sqrt{\log(d)}$-strongly convex w.r.t ℓ_1 norm.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Sylvain Carpentier

Lecture 14
Oct. 26, 2015

In this lecture we will wrap up the study of optimization techniques with stochastic optimization. The tools that we are going to develop will turn out to be very efficient in minimizing the φ-risk when we can bound the noise on the gradient.

3. STOCHASTIC OPTIMIZATION

3.1 Stochastic convex optimization

We are considering random functions $x \mapsto \ell(x, Z)$ where x is the optimization parameter and Z a random variable. Let P_Z be the distribution of Z and let us assume that $x \mapsto \ell(x, Z)$ is convex P_Z a.s. In particular, $\mathbb{E}[\ell(x, Z)]$ will also be convex. The goal of stochastic convex optimization is to approach $\min_{x \in \mathcal{C}} \mathbb{E}[\ell(x, Z)]$ when $\mathcal{C}$ is convex. For our purposes, $\mathcal{C}$ will be a deterministic convex set. However, stochastic convex optimization can be defined more broadly. The constraint can be itself stochastic :

$$\mathcal{C} = \{x, \mathbb{E}[g(x, Z)] \leq 0\}, \quad g \text{ convex } P_Z \text{ a.s.}$$

$$\mathcal{C} = \{x, \mathbb{P}[g(x, Z) \leq 0] \geq 1 - \varepsilon\}, \quad \text{“chance constraint”}$$

The second constraint is not convex a priori but remedies are possible (see [NS06, Nem12]). In the following, we will stick to the case where X is deterministic. A few optimization problems we tackled can be interpreted in this new framework.

3.1.1 Examples

Boosting. Recall that the goal in Boosting is to minimize the φ-risk:

$$\min_{\alpha \in \Lambda} \mathbb{E}[\varphi(-Y f_\alpha(X))] \,,$$

where Λ is the simplex of $\mathbb{R}^d$. Define $Z = (X, Y)$ and the random function $\ell(\alpha, Z) = \varphi(-Y f_\alpha(X))$, convex P_Z a.s.

Linear regression. Here the goal is the minimize the ℓ_2 risk:

$$\min_{\alpha \in \mathbb{R}^d} \mathbb{E}[(Y - f_\alpha(X))^2] \,.$$

Define $Z = (X, Y)$ and the random function $\ell(\alpha, Z) = (Y - f_\alpha(X))^2$, convex P_Z a.s.

Maximum likelihood. We consider samples $Z_1, \ldots, Z_n$ iid with density p_θ, $\theta \in \Theta$. For instance, $Z \sim \mathcal{N}(\theta, 1)$. The likelihood functions associated to this set of samples is $\theta \mapsto \prod_{i=1}^n p_\theta(Z_i)$. Let $p^*(Z)$ denote the true density of Z (it does not have to be of the form p_θ for some $\theta \in \Theta$. Then

$$\frac{1}{n}\mathbb{E}[\log \prod_{i=1}^n p_\theta(Z_i)] = -\int \log(\frac{p^*(z)}{p_\theta(z)}) p^*(z) dz + C = -\mathsf{KL}(p^*, p_\theta) + C$$

where C is a constant in θ. Hence maximizing the expected log-likelihood is equivalent to minimizing the expected Kullback-Leibler divergence:

$$\max_{\theta} \mathbb{E}[\log \prod_{i=1}^{n} p_\theta(Z_i)] \iff \mathsf{KL}(p^*, p_\theta)$$

External randomization. Assume that we want to minimize a function of the form

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)\,,$$

where the functions $f_1, \dots, f_n$ are convex. As we have seen, this arises a lot in empirical risk minimization. In this case, we treat this problem as deterministic problem but inject artificial randomness as follows. Let I be a random variable uniformly distributed on $[n] =: \{1, \dots, n\}$. We have the representation $f(x) = \mathbb{E}[f_I(x)]$, which falls into the context of stochastic convex optimization with $Z = I$ and $\ell(x, I) = f_I(x)$.

Important Remark: There is a key difference between the case where we assume that we are given independent random variables and the case where we generate artificial randomness. Let us illustrate this difference for Boosting. We are given $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d from some unknown distribution. In the first example, our aim is to minimize $\mathbb{E}[\varphi(-Yf_\alpha(X))]$ based on these n observations and we will that the stochastic gradient allows to do that by take one pair (X_i, Y_i) in each iteration. In particular, we can use each pair at most once. We say that we do *one pass* on the data.

We could also leverage our statistical analysis of the empirical risk minimizer from previous lectures and try to minimize the empirical φ-risk

$$\hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f_\alpha(X_i))$$

by generating k independent random variables $I_1, \dots, I_k$ uniform over $[n]$ and run the stochastic gradient descent to us one random variable I_j in each iteration. The difference here is that k can be arbitrary large, regardless of the number n of observations (we make *multiple passes* on the data). However, minimizing $\mathbb{E}_I[\varphi(-Y_I f_\alpha(X_I))|X_1, Y_1, \dots, X_n, Y_n]$ will perform no better than the empirical risk minimizer whose statistical performance is limited by the number n of observations.

3.2 Stochastic gradient descent

If the distribution of Z was known, then the function $x \mapsto \mathbb{E}[\ell(x, Z)]$ would be known and we could apply gradient descent, projected gradient descent or any other optimization tool seen before in the deterministic setup. However this is not the case in reality where the true distribution P_Z is unknown and we are only given the samples $Z_1, \dots, Z_n$ and the random function $\ell(x, Z)$. In what follows, we denote by $\partial \ell(x, Z)$ the set of subgradients of the function $y \mapsto \ell(y, Z)$ at point x.

Algorithm 1 Stochastic Gradient Descent algorithm

```
    **Input:** $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$, independent random variables $Z_1, \dots, Z_k$ with distribution $P_Z$.
    **for** $s = 1$ to $k-1$ **do**
        $y_{s+1} = x_s - \eta_s \tilde{g}_s\,, \quad \tilde{g}_s \in \partial \ell(x_s, Z_s)$
        $x_{s+1} = \pi_{\mathcal{C}}(y_{s+1})$
    **end for**
    **return** $\bar{x}_k = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

Note the difference here with the deterministic gradient descent which returns either $\bar{x}_k$ or $x_k^\circ = \underset{x_1,\dots,x_n}{\operatorname{argmin}} f(x)$. In the stochastic framework, the function $f(x) = \mathbb{E}[\ell(x, \xi)]$ is typically unknown and $\mathring{x}_k$ cannot be computed.

Theorem: Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^d$ such that $\operatorname{diam}(\mathcal{C}) \leq R$. Assume that he convex function $f(x) = \mathbb{E}[\ell(x, Z)]$ attains its minimum on $\mathcal{C}$ at $x^* \in \mathbb{R}^d$. Assume that $\ell(x, Z)$ is convex P_Z a.s. and that $\mathbb{E}\|\tilde{g}\|^2 \leq L^2$ for all $\tilde{g} \in \partial \ell(x, Z)$ for all x. Then if $\eta_s \equiv \eta = \frac{R}{L\sqrt{k}}$,

$$\mathbb{E}[f(\bar{x}_k)] - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Proof.

$$\begin{aligned}
f(x_s) - f(x^*) &\leq {g_s}^\top (x_s - x^*) \\
&= \mathbb{E}[\tilde{g}_s^\top (x_s - x^*) | x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(y_{s+1} - x_s)^\top (x_s - x^*) | x_s] \\
&= \frac{1}{2\eta}\mathbb{E}[\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2 | x_s] \\
&\leq \frac{1}{2\eta}(\eta^2 \mathbb{E}[\|\tilde{g}_s\|^2 | x_s] + \mathbb{E}[\|x_s - x^*\|^2 | x_s] - \mathbb{E}[\|x_{s+1} - x^*\|^2 | x_s]
\end{aligned}$$

Taking expectations and summing over s we get

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k}\,.$$

Using Jensen's inequality and chosing $\eta = \frac{R}{L\sqrt{k}}$, we get

$$\mathbb{E}[f(\bar{x_k})] - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

□

3.3 Stochastic Mirror Descent

We can also extend the Mirror Descent to a stochastic version as follows.

Algorithm 2 Mirror Descent algorithm

```
**Input:** $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$, $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\zeta(x) = \nabla\Phi(x)$, independent random variables $Z_1, \dots, Z_k$ with distribution $P_Z$.
**for** $s = 1, \cdots, k$ **do**
    $\zeta(y_{s+1}) = \zeta(x_s) - \eta\tilde{g}_s$ for $\tilde{g}_s \in \partial\ell(x_s, Z_s)$
    $x_{s+1} = \Pi_{\mathcal{C}}^{\Phi}(y_{s+1})$
**end for**
**return** $\overline{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

Theorem: Assume that Φ is α-strongly convex on $\mathcal{C} \cap\ \mathcal{D}$ w.r.t. $\|\cdot\|$ and

$$R^2 = \sup_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x) - \min_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$$

take $x_1 = \operatorname{argmin}_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$ (assume that it exists). Then, Stochastic Mirror Descent with $\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{R}}$ outputs $\bar{x}_k$, such that

$$\mathbb{E}[f(\bar{x}_k)] - f(x^*) \le RL\sqrt{\frac{2}{\alpha k}}\,.$$

Proof. We essentially reproduce the proof for the Mirror Descent algorithm.

Take $x^\sharp \in \mathcal{C} \cap \mathcal{D}$. We have

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\le g_s^\top(x_s - x^\sharp) \\
&\mathbb{E}[\tilde{g}_s^\top(x_s - x^*)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\zeta(x_s) - \zeta(y_{s+1}))^\top (x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top (x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1})\middle|x_s\right] \\
&\le \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\middle|x_s\right] \\
&\le \frac{\eta}{2\alpha^2}\mathbb{E}[\|\tilde{g}_s\|_*^2|x_s] + \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\middle|x_s\right]
\end{aligned}$$

where the last inequality comes from

$$\begin{aligned}
D_\Phi(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1}) \\
&\le \left[\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right]^\top (x_s - y_{s+1}) - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\le \eta\|\tilde{g}_s\|_*\|x_s - y_{s+1}\| - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\le \frac{\eta^2\|\tilde{g}_s\|_*^2}{2\alpha}\,.
\end{aligned}$$

Summing and taking expectations, we get

$$\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \le \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x^\sharp, x_1)}{k\eta}\,. \tag{3.1}$$

We conclude as in the previous lecture.

□

3.4 Stochastic coordinate descent

Let f be a convex L-Lipschitz and differentiable function on $\mathbb{R}^d$. Let us denote by $\nabla_i f$ the partial derivative of f in the direction e_i. One drawback of the Gradient Descent Algorithm is that at each step one has to update every coordinate $\nabla_i f$ of the gradient. The idea of the stochastic coordinate descent is to pick at each step a direction e_j uniformly and to choose that e_j to be the direction of the descent at that step. More precisely, of I is drawn uniformly on $[d]$, then $\mathbb{E}[d\nabla_I f(x)e_I] = \nabla f(x)$. Therefore, the vector $d\nabla_I f(x)e_I$ that has only one nonzero coordinate is an unbiased estimate of the gradient $\nabla f(x)$. We can use this estimate to perform stochastic gradient descent.

Algorithm 3 Stochastic Coordinate Descent algorithm

```
Input: $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\ge1}$, independent random variables $I_1, \dots, I_k$ uniform over $[d]$.
for $s = 1$ to $k - 1$ do
    $y_{s+1} = x_s - \eta_s d\nabla_I f(x)e_I\,, \quad \tilde{g}_s \in \partial\ell(x_s, Z_s)$
    $x_{s+1} = \pi_{\mathcal{C}}(y_{s+1})$
end for
return $\bar{x}_k = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

If we apply Stochastic Gradient Descent to this problem for $\eta = \frac{R}{L}\sqrt{\frac{2}{dk}}$, we directly obtain

$$\mathbb{E}[f(\bar{x_k})] - f(x^*) \le RL\sqrt{\frac{2d}{k}}$$

We are in a trade-off situation where the updates are much easier to implement but where we need more steps to reach the same precision as the gradient descent alogrithm.

References

[Nem12] Arkadi Nemirovski, *On safe tractable approximations of chance constraints*, European J. Oper. Res. **219** (2012), no. 3, 707–718. MR 2898951 (2012m:90133)

[NS06] Arkadi Nemirovski and Alexander Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim. **17** (2006), no. 4, 969–996. MR 2274500 (2007k:90077)

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 15

Scribe: Zach Izzo — Oct. 27, 2015

Part III

Online Learning

It is often the case that we will be asked to make a sequence of predictions, rather than just one prediction given a large number of data points. In particular, this situation will arise whenever we need to perform **online classification**: at time t, we have $(X_1, Y_1), \dots, (X_{t-1}, Y_{t-1})$ iid random variables, and given X_t, we are asked to predict $Y_t \in \{0, 1\}$. Consider the following examples.

Online Shortest Path: We have a graph $G = (V, E)$ with two distinguished vertices s and t, and we wish to find the shortest path from s to t. However, the edge weights $E_1, \dots, E_t$ change with time t. Our observations after time t may be all of the edge weights $E_1, \dots, E_t$; or our observations may only be the weights of edges through which our path traverses; or our observation may only be the sum of the weights of the edges we've traversed.

Dynamic Pricing: We have a sequence of customers, each of which places a value v_t on some product. Our goal is to set a price p_t for the tth customer, and our reward for doing so is p_t if $p_t \le v_t$ (in which case the customer buys the product at our price) or 0 otherwise (in which case the customer chooses not to buy the product). Our observations after time t may be $v_1, \dots, v_t$; or, perhaps more realistically, our observations may only be $\mathbb{1}(p_1 < v_1), \dots, \mathbb{1}(p_t < v_t)$. (In this case, we only know whether or not the customer bought the product.)

Sequential Investment: Given N assets, a portfolio is $\omega \in \Delta^N = \{x \in \mathbb{R}^n : x_i \ge 0, \sum_{i=1}^N x_i = 1\}$. ($\omega$ tells what percentage of our funds to invest in each stock. We could also allow for negative weights, which would correspond to shorting a stock.) At each time t, we wish to create a portfolio $\omega_t \in \Delta^N$ to maximize $\omega_t^T z_t$, where $z_t \in \mathbb{R}^N$ is a random variable which specifies the return of each asset at time t.

There are two general modelling approaches we can take: statistical or adversarial. Statistical methods typically require that the observations are iid, and that we can learn something about future points from past data. For example, in the dynamic pricing example, we could assume $v_t \sim N(v, 1)$. Another example is the Markowitz model for the sequential investment example, in which we assume that $\log(z_t) \sim \mathcal{N}(\mu, \Sigma)$.

In this lecture, we will focus on adversarial models. We assume that z_t can be any bounded sequence of numbers, and we will compare our predictions to the performance of some benchmark. In these types of models, one can imagine that we are playing a game against an opponent, and we are trying to minimize our losses regardless of the moves he plays. In this setting, we will frequently use optimization techniques such as mirror descent, as well as approaches from game theory and information theory.

1. PREDICTION WITH EXPERT ADVICE

1.1 Cumulative Regret

Let $\mathcal{A}$ be a convex set of actions we can take. For example, in the sequential investment example, $\mathcal{A} = \Delta^N$. If our options are discrete–for instance, choosing edges in a graph–then think of $\mathcal{A}$ as the convex hull of these options, and we can play one of the choices randomly according to some distribution. We will denote our adversary's moves by $\mathcal{Z}$. At time t, we simultaneously reveal $a_t \in \mathcal{A}$ and $z_t \in \mathcal{Z}$. Denote by $\ell(a_t, z_t)$ the loss associated to the player/decision maker taking action a_t and his adversary playing z_t.

In the general case, $\sum_{t=1}^n \ell(a_t, z_t)$ can be arbitrarily large. Therefore, rather than looking at the absolute loss for a series of n steps, we will compare our loss to the loss of a benchmark called an **expert**. An expert is simply some vector $b \in \mathcal{A}^n$, $b = (b_1, \dots, b_t, \dots, b_n)^T$. If we choose K experts $b^{(1)}, \dots, b^{(K)}$, then our benchmark value will be the minimum cumulative loss amongst of all the experts:

$$\text{benchmark} = \min_{1 \le j \le K} \sum_{t=1}^n \ell(b_t^{(j)}, z_t).$$

The **cumulative regret** is then defined as

$$R_n = \sum_{t=1}^n \ell(a_t, z_t) - \min_{1 \le j \le K} \sum_{t=1}^n \ell(b_t^{(j)}, z_t).$$

At time t, we have access to the following information:

1. All of our previous moves, i.e. $a_1, \dots, a_{t-1}$,
2. all of our adversary's previous moves, i.e. $z_1, \dots, z_{t-1}$, and
3. All of the experts' strategies, i.e. $b^{(1)}, \dots, b^{(K)}$.

Naively, one might try a strategy which chooses $a_t = b_t^*$, where b^* is the expert which has incurred minimal total loss for times $1, \dots, t-1$. Unfortunately, this strategy is easily exploitable by the adversary: he can simply choose an action which maximizes the loss for that move at each step. To modify our approach, we will instead take a convex combination of the experts' suggested moves, weighting each according to the performance of that expert thus far. To that end, we will replace $\ell(a_t, z_t)$ by $\ell(p, (b_t, z_t))$, where $p \in \Delta^K$ denotes a convex combination, $b_t = (b_t^{(1)}, \dots, b_t^{(K)})^T \in \mathcal{A}^K$ is the vector of the experts' moves at time t, and $z_t \in \mathcal{Z}$ is our adversary's move. Then

$$R_n = \sum_{t=1}^n \ell(p_t, z_t) - \min_{1 \le j \le K} \sum_{t=1}^n \ell(e_j, z_t)$$

where e_j is the vector whose jth entry is 1 and the rest of the entries are 0. Since we are restricting ourselves to convex combinations of the experts' moves, we can write $\mathcal{A} = \Delta^K$. We can now reduce our goal to an optimization problem:

$$\min_{\theta \in \Delta^K} \sum_{j=1}^K \theta_j \sum_{t=1}^n \ell(e_j, z_t).$$

From here, one option would be to use a projected gradient descent type algorithm: we define

$$q_{t+1} = p_t - \eta(\ell(e_1, z_t), \dots, \ell(e_K, z_T))^T$$

and then $p_{t+1} = \pi^{\Delta^K}(p_t)$ to be the projection of q_{t+1} onto the simplex.

1.2 Exponential Weights

Suppose we instead use stochastic mirror descent with $\Phi =$ negative entropy. Then

$$q_{t+1,j} = p_{t+1,j} \exp(-\eta \ell(e_j, z_t)), \quad p_{t+1,j} = \frac{q_t}{\sum_{l=1}^K q_{t+1,l}},$$

where we have defined

$$p_t = \sum_{j=1}^K \left(\frac{w_{t,j}}{\sum_{l=1}^K w_{t,l}} e_j \right), \quad w_{t,j} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_j, z_s) \right).$$

This process looks at the loss from each expert and downweights it exponentially according to the fraction of total loss incurred. For this reason, this method is called an **exponential weighting (EW) strategy**.

Recall the definition of the cumulative regret R_n:

$$R_n = \sum_{t=1}^n \ell(p_t, z_t) - \min_{1 \le j \le K} \sum_{t=1}^n \ell(e_j, z_t).$$

Then we have the following theorem.

Theorem: Assume $\ell(\cdot, z)$ is convex for all $z \in \mathcal{Z}$ and that $\ell(p, z) \in [0, 1]$ for all $p \in \Delta^K, z \in \mathcal{Z}$. Then the EW strategy has regret

$$R_n \le \frac{\log K}{\eta} + \frac{\eta n}{2}.$$

In particular, for $\eta = \sqrt{\frac{2 \log K}{n}}$,

$$R_n \le \sqrt{2n \log K}.$$

Proof. We will recycle much of the mirror descent proof. Define

$$f_t(p) = \sum_{j=1}^K p_j \ell(e_j, z_t).$$

Denote $\|\cdot\| := |\cdot|_1$. Then

$$\frac{1}{n} \sum_{t=1}^n f_t(p_t) - f_t(p^*) \le \frac{\eta \frac{1}{n} \sum_{t=1}^n \|g_t\|_*^2}{2} + \frac{\log K}{\eta n},$$

where $g_t \in \partial f_t(p_t)$ and $\|\cdot\|_*$ is the dual norm (in this case $\|\cdot\|_* = |\cdot|_\infty$). The 2 in the denominator of the first term of this sum comes from setting $\alpha = 1$ in the mirror descent proof. Now,

$$g_t \in \partial f_t(p_t) \Rightarrow g_t = (\ell(e_1, z_t), \dots, \ell(e_K, z_t))^T.$$

Furthermore, since $\ell(p, z) \in [0, 1]$, we have $\|g_t\|_* = |g_t|_\infty \leq 1$ for all t. Thus

$$\frac{\eta \frac{1}{n}\sum_{t=1}^n \|g_t\|_*^2}{2} + \frac{\log K}{n\eta} \leq \frac{\eta}{2} + \frac{\log K}{\eta n}.$$

Substituting for f_t yields

$$\sum_{t=1}^n \sum_{j=1}^K p_{t,j}\ell(e_j, z_t) - \boxed{\min_{p \in \Delta^K} \sum_{j=1}^K \sum_{t=1}^n p_j \ell(e_j, z_t)} \leq \frac{\eta n}{2} + \frac{\log K}{\eta}.$$

Note that the boxed term is actually $\min_{1 \leq j \leq K} \sum_{t=1}^n \ell(e_j, z_t)$. Furthermore, applying Jensen's to the unboxed term gives

$$\sum_{t=1}^n \sum_{j=1}^K p_{t,j}\ell(e_j, z_t) \geq \sum_{t=1}^n \ell(p_t, z_t).$$

Substituting these expressions then yields

$$R_n \leq \frac{\eta n}{2} + \frac{\log K}{\eta}.$$

We optimize over η to reach the desired conclusion. □

We now offer a different proof of the same theorem which will give us the optimal constant in the error bound. Define

$$w_{t,j} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_j, z_s)\right), \quad W_t = \sum_{j=1}^K w_{t,j}, \quad p_t = \frac{\sum_{j=1}^K w_{t,j} e_j}{W_t}.$$

For $t = 1$, we initialize $w_{1,j} = 1$, so $W_1 = K$. It should be noted that the starting values for $w_{1,j}$ are uniform, so we're starting at the correct point (i.e. maximal entropy) for mirrored descent. Now we have

$$\begin{aligned}
\log\left(\frac{W_{t+1}}{W_t}\right) &= \log\left(\frac{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_j, z_s)\right) \exp(-\eta \ell(e_j, z_t))}{\sum_{l=1}^K \exp\left(-\eta \sum_{j=1}^{t-1} \ell(e_l, z_s)\right)}.\right) \\
&= \log\left(\mathbb{E}_{J \sim p_t}\left[\exp(-\eta \ell(e_J, z_t))\right]\right) \\
\text{Hoeffding's lemma } \Rightarrow &\leq \log\left(e^{\frac{1}{8}\eta^2} e^{-\eta \mathbb{E}_J \ell(e_J, z_t)}\right) \\
&= \frac{\eta^2}{8} - \eta \mathbb{E}_J \ell(e_J, z_t) \\
\text{Jensen's } \Rightarrow &\leq \frac{\eta^2}{8} - \eta \ell(\mathbb{E}_J e_J, z_t) = \frac{\eta^2}{8} - \eta \ell(p_t, z_t)
\end{aligned}$$

since $\mathbb{E}_J e_j = \sum_{j=1}^K p_{t,j} e_j$. If we sum over t, the sum telescopes. Since $W_1 = K$, we are left with

$$\log(W_{n+1}) - \log(K) \le \frac{n\eta^2}{8} - \eta \sum_{t=1}^n \ell(p_t, z_t).$$

We have

$$\log(W_{n+1}) = \log\left(\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^n \ell(e_j, z_s)\right)\right),$$

so setting $j^* = \mathrm{argmin}_{1\le j\le K} \sum_{t=1}^n \ell(e_j, z_t)$, we obtain

$$\log(W_{n+1}) \ge \log\left(\exp\left(-\eta \sum_{s=1}^n \ell(e_{j*}, z_s)\right)\right) = -\eta \sum_{t=1}^n \ell(e_{j^*}, z_t).$$

Rearranging, we have

$$\sum_{t=1}^n \ell(p_t, z_t) - \sum_{t=1}^n \ell(e_{j^*}, z_t) \le \frac{\eta n}{8} + \frac{\log K}{\eta}.$$

Finally, we optimize over η to arrive at

$$\eta = \sqrt{\frac{8\log K}{n}} \Rightarrow R_n \le \sqrt{\frac{n\log K}{2}}.$$

The improved constant comes from the assumption that our loss lies in an interval of size 1 (namely $[0,1]$) rather than in an interval of size 2 (namely $[-1,1]$).

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Haihao (Sean) Lu

Lecture 16
Nov. 2, 2015

Recall that in last lecture, we talked about prediction with expert advice. Remember that $l(e_j, z_t)$ means the loss of expert j at time t, where z_t is one adversary's move. In this lecture, for simplexity we replace the notation z_t and denote by z_t the loss associated to all experts at time t:

$$z_t = \begin{pmatrix} \ell(e_1, z_t) \\ \vdots \\ \ell(e_K, z_t) \end{pmatrix} ,$$

whereby for $p \in \Delta^K$, $p^\top z_t = \sum_{j=1}^K p_j \ell(e_j, z_t)$. This gives an alternative definition of $f_t(p)$ in last lecture. Actually it is easy to check $f_t(p) = p^\top z_t$, thus we can rewrite the theorem for exponential weighting(EW) strategy as

$$R_n \le \sum_{t=1}^n p_t^\top z_t - \min_{p \in \Delta^k} \sum_{t=1}^n p^\top z_t \le \sqrt{2n \log K},$$

where the first inequality is Jensen inequality:

$$\sum_{t=1}^n p_z^\top z_t \ge \sum_{t=1}^n \ell(p_z, z_t) \ .$$

We consider EW strategy for bounded convex losses. Without loss of generality, we assume $\ell(p, z) \in [0, 1]$, for all $(p, z) \in \Delta^K \times \mathcal{Z}$, thus in notation here, we expect $p_t \in \Delta^K$ and $z_t \in [0, 1]^K$. Indeed if $\ell(p, z) \in [m, M]$ then one can work with a rescaled loss $\bar{\ell}(a, z) = \frac{\ell(a,z)-m}{M-m}$. Note that now we have bounded gradient on p_t, since z_t is bounded.

2. FOLLOW THE PERTURBED LEADER (FPL)

In this section, we consider a different strategy, called Follow the Perturbed Leader.

At first, we introduce Follow the Leader strategy, and give an example to show that Follow the Leader can be hazardous sometimes. At time t, assume that choose

$$p_t = \underset{p \in \Delta^K}{\operatorname{argmin}} \sum_{s=1}^{t-1} p^\top z_s.$$

Note that the function to be optimized is linear in p, whereby the optimal solution should be a vertex of the simplex. This method can be viewed as a greedy algorithm, however, it might not be a good strategy.

Consider the following example. Let $K = 2$, $z_1 = (0, \varepsilon)^\top$, $z_2 = (0, 1)^\top$, $z_3 = (1, 0)^\top$, $z_4 = (0, 1)^\top$ and so on (alternatively having $(0, 1)^\top$ and $(1, 0)^\top$ when $t \ge 2$), where ε is small enough. Then with Following the Leader Strategy, we have that p_1 is arbitrary and in the best case $p_1 = (1, 0)^\top$, and $p_2 = (1, 0)^\top$, $p_3 = (0, 1)^\top$, $p_4 = (1, 0)^\top$ and so on (alternatively having $(0, 1)^\top$ and $(1, 0)^\top$ when $t \ge 2$).

In the above example, we have

$$\sum_{t=1}^{n} p_t^\top z_t - \min_{p\in\Delta^k} \sum_{t=1}^{n} p^\top z_t \leq n - 1 - \frac{n}{2} \leq \frac{n}{2} - 1 \ ,$$

which gives raise to linear regret.

Now let's consider FPL. FPL regularizes FL by adding a small amount of noise, which can guarantee square root regret under oblivious adversary situation.

Algorithm 1 Follow the Perturbed Leader (FPL)

```
    **Input:** Let $\xi$ be a random variables uniformly drawn on $[0, \frac{1}{\eta}]^K$.
    **for** $t = 1$ to $n$ **do**
        $$p_t = \operatorname{argmin}_{p\in\Delta^K} \sum_{s=1}^{t-1} \big(p^\top z_s + \xi\big).$$
    **end for**
```

We analyze this strategy in oblivious adversaries, which means the sequence z_t is chosen ahead of time, rather than adaptively given. The following theorem gives a bound for regret of FPL:

Theorem: FPL with $\eta = \frac{1}{\sqrt{kn}}$ yields expected regret:

$$\mathbb{E}_\xi[R_n] \leq 2\sqrt{2nK} \ .$$

Before proving the theorem, we introduce the so-called Be-The-Leader Lemma at first.

Lemma: (Be-The-Leader)
For all loss function $\ell(p, z)$, let

$$p_t^* = \arg \min_{p\in\Delta^K} \sum_{s=1}^{t} \ell(p, z_s) \ ,$$

then we have

$$\sum_{t=1}^{n} \ell(p_t^*, z_t) \leq \sum_{t=1}^{n} \ell(p_n^*, z_t)$$

Proof. The proof goes by induction on n. For $n = 1$, it is clearly true. From n to $n+1$, it follows from:

$$\begin{aligned}\sum_{t=1}^{n+1} \ell(p_t^*, z_t) &= \sum_{i=1}^{n} \ell(p_t^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \\ &\leq \sum_{i=1}^{n} \ell(p_n^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \\ &\leq \sum_{i=1}^{n} \ell(p_{n+1}^*, z_t) + \ell(p_{n+1}^*, z_{n+1}) \ ,\end{aligned}$$

where the first inequality uses induction and the second inequality follows from the definition of p_n^*. □

Proof of Theorem. Define

$$q_t = \operatorname*{argmin}_{p \in \Delta^K} p^\top (\xi + \sum_{s=1}^{t} z_s) \ .$$

Using the Be-The-Leader Lemma with

$$\ell(p, z_t) = \begin{cases} p^T(\xi + z_1) & if \ t = 1 \\ p^T z_t & if \ t > 1 \ , \end{cases}$$

we have

$$q_1^\top \xi + \sum_{t=1}^{n} q_t^\top z_t \leq \min_{q \in \Delta^K} q^\top (\xi + \sum_{t=1}^{n} z_t) \ ,$$

whereby for any $q \in \Delta^K$,

$$\sum_{i=1}^{n} \left(q_t^\top z_t - q^\top z_t \right) \leq \left(q^\top - q_1^\top \right) \xi \leq \|q - q_1\|_1 \|\xi\|_\infty \leq \frac{2}{\eta} \ ,$$

where the second inequality uses Hölder's inequality and the third inequality is from the fact that q and q_1 are on the simplex and ξ is in the box.

Now let

$$q_t = \arg \min_{p \in \Delta^K} p^\top \left(\xi + z_t + \sum_{s=1}^{t} z_s \right)$$

and

$$p_t = \arg \min_{p \in \Delta^K} p^\top \left(\xi + 0 + \sum_{s=1}^{t} z_s \right) \ .$$

Therefore,

$$\begin{aligned}\mathbb{E}[R_n] &\leq \sum_{i=1}^{n} p_t^\top z_t - \min_{p \in \Delta^k} \sum_{i=1}^{n} p^\top z_t \\ &\leq \sum_{i=1}^{n} \left(q_t^\top z_t - p^{*T} z_t \right) + \sum_{i=1}^{n} \mathbb{E}[(p_t - q_t)^\top z_t] \\ &\leq \frac{2}{\eta} + \sum_{i=1}^{n} \mathbb{E}[(p_t - q_t)^\top z_t] \ , \qquad (2.1)\end{aligned}$$

where $p^* = \arg\min_{p\in\Delta^K} \sum_{t=1}^n p^\top z_t$.

Now let

$$h(\xi) = z_t^\top \left(\arg\min_{p\in\Delta^K} p^\top [\xi + \sum_{s=1}^{t-1} z_s] \right) ,$$

then we have a easy observation that

$$\mathbb{E}[z_t^\top (p_t - q_t)] = \mathbb{E}[h(\xi)] - \mathbb{E}[h(\xi + z_t)] .$$

Hence,

$$\begin{aligned}
\mathbb{E}[z_t^\top (p_t - q_t)] &= \eta^K \int_{\xi\in[0,\frac{1}{\eta}]^K} h(\xi) d\xi - \eta^K \int_{\xi\in z_t+[0,\frac{1}{\eta}]^K} h(\xi) d\xi \\
&\leq \eta^K \int_{\xi\in[0,\frac{1}{\eta}]^K \setminus \left\{ z_t + [0,\frac{1}{\eta}]^K \right\}} h(\xi) d\xi \\
&\leq \eta^K \int_{\xi\in[0,\frac{1}{\eta}]^K \setminus \left\{ z_t + [0,\frac{1}{\eta}]^K \right\}} 1 d\xi \\
&= \mathbb{P}\left(\exists i \in [K], \xi(i) \leq z_t(i)\right) \\
&\leq \sum_{i=1}^K \mathbb{P}\left(\mathbf{Unif}\left([0, \frac{1}{\eta}] \right) \leq z_t(i) \right) \\
&\leq \eta K z_t(i) \leq \eta K \ , \qquad (2.2)
\end{aligned}$$

where the first inequality is from the fact that $h(\xi) \geq 0$, the second inequality uses $h(\xi) \leq 1$, the second equation is just geometry and the last inequality is due to $z_t(i) \leq 1$.

Combining (2.1) and (2.2) together, we have

$$\mathbb{E}[R_n] \leq \frac{2}{\eta} + \eta K n .$$

In particular, with $\eta = \sqrt{\frac{2}{Kn}}$, we have

$$\mathbb{E}[R_n] \leq 2\sqrt{2Kn} ,$$

which completes the proof. □

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 18
Scribe: Haihao (Sean) Lu — Nov. 2, 2015

3. STOCHASTIC BANDITS

3.1 Setup

The stochastic multi-armed bandit is a classical model for decision making and is defined as follows:

There are K arms(different actions). Iteratively, a decision maker chooses an arm $k \in \{1, \dots, K\}$, yielding a sequence $X_{K,1}, \dots, X_{K,t}, \dots$, which are i.i.d random variables with mean μ_k. Define $\mu_* = \max_j \mu_j$ or $* \in \arg\max$. A policy π is a sequence $\{\pi_t\}_{t\geq 1}$, which indicates which arm to be pulled at time t. $\pi_t \in \{1, \dots, K\}$ and it depends only on the observations strictly interior to t. The regret is then defined as:

$$\begin{aligned}
R_n &= \max_k \mathbb{E}[\sum_{t=1}^n X_{K,t}] - \mathbb{E}[\sum_{t=1}^n X_{\pi_t,t}] \\
&= n\mu_* - \mathbb{E}[\sum_{t=1}^n X_{\pi_t,t}] \\
&= n\mu_* - \mathbb{E}[\mathbb{E}[\sum_{t=1}^n X_{\pi_t,t} \mid \pi_t]] \\
&= \sum_{k=1}^K \Delta_k \mathbb{E}[T_k(n)] \ ,
\end{aligned}$$

where $\Delta_k = \mu_* - \mu_k$ and $T_k(n) = \sum_{t=1}^n \mathbb{I}(\pi_t = k)$ is the number of time when arm k was pulled.

3.2 Warm Up: Full Info Case

Assume in this subsection that $K = 2$ and we observe the full information $\begin{pmatrix} X_{1,t} \\ \vdots \\ X_{K,t} \end{pmatrix}$ at time t after choosing π_t. So in each iteration, a normal idea is to choose the arm with highest average return so far. That is

$$\pi_t = \underset{k=1,2}{\operatorname{argmax}} \bar{X}_{k,t}$$

where

$$\bar{X}_{k,t} = \frac{1}{t}\sum_{s=1}^t X_{k,s}$$

Assume from now on that all random variable $X_{k,t}$ are subGaussian with variance proxy σ^2, which means $\mathbb{E}[e^{ux}] \leq e^{\frac{u^2\sigma^2}{2}}$ for all $u \in \mathbb{R}$. For example, $N(0, \sigma^2)$ is subGaussian with variance proxy σ^2 and any bounded random variable $X \in [a, b]$ is subGaussian with variance proxy $(b-a)^2/4$ by Hoeffding's Lemma.

Therefore,

$$R_n = \Delta \mathbb{E}[T_2(n)] \ , \tag{3.1}$$

where $\Delta = \mu_1 - \mu_2$. Besides,

$$\begin{aligned} T_2(n) &= 1 + \sum_{t=2}^{n} \mathbb{I}(\bar{X}_{2,t} > \bar{X}_{1,t}) \\ &= 1 + \sum_{t=2}^{n} \mathbb{I}(\bar{X}_{2,t} - \bar{X}_{1,t} - (\mu_2 - \mu_1) \geq \Delta) \ . \end{aligned}$$

It is easy to check that $(\bar{X}_{2,t} - \bar{X}_{1,t}) - (\mu_2 - \mu_1)$ is centered subGaussian with variance proxy $2\sigma^2$, whereby

$$\mathbb{E}[\mathbb{I}(\bar{X}_{2,t} > \bar{X}_{1,t})] \leq e^{-\frac{t\Delta^2}{4\sigma^2}}$$

by a simple Chernoff Bound. Therefore,

$$R_n \leq \Delta(1 + \sum_{t=0}^{\infty} e^{-\frac{t\Delta^2}{4\sigma^2}}) \leq \Delta + \frac{4\sigma^2}{\Delta} \ , \tag{3.2}$$

whereby the benchmark is

$$R_n \leq \Delta + \frac{4\sigma^2}{\Delta} \ .$$

3.3 Upper Confidence Bound (UCB)

Without loss of generality, from now on we assume $\sigma = 1$. A trivial idea is that after s pulls on arm k, we use $\hat{\mu}_{k,s} = \frac{1}{s}\sum_{j \in \{\text{pulls of } k\}} X_{K,j}$ and choose the one with largest $\hat{\mu}_{k,s}$. The problem of this trivial policy is that for some arm, we might try it for only limited times, which give a bad average and then we never try it again. In order to overcome this limitation, a good idea is to choose the arm with highest upper bound estimate on the mean of each arm at some probability lever. Note that the arm with less tries would have a large deviations from its mean. This is called Upper Confidence Bound policy.

Algorithm 1 Upper Confidence Bound (UCB)

```
for $t = 1$ to $K$ do
    $\pi_t = t$
end for
for $t = K + 1$ to $n$ do
    $$T_k(t) = \sum_{s=1}^{t-1} \mathbb{1}(\pi_t = k)$$
    (number of time we have pull arm $k$ before time $t$)
    $$\hat{\mu}_{k,t} = \frac{1}{T_k(t)} \sum_{s=1}^{t-1} X_{K,t\wedge s}$$
    $$\pi_t \in \operatorname*{argmax}_{k\in[K]} \left\{ \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log(t)}{T_k(t)}} \right\} ,$$
end for
```

Theorem: The UCB policy has regret

$$R_n \leq 8 \sum_{k,\Delta_k>0} \frac{\log n}{\Delta_k} + (1 + \frac{\pi^2}{3}) \sum_{k=1}^{K} \Delta_k$$

Proof. From now on we fix k such that $\Delta_k > 0$. Then

$$\mathbb{E}[T_k(n)] = 1 + \sum_{t=K+1}^{n} \mathbb{P}(\pi_t = k) \ .$$

Note that for $t > K$,

$$\begin{aligned} \{\pi_t = k\} &\subseteq \{\hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}} \leq \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}\} \\ &\subseteq \left\{ \{\mu_k \geq \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}}\} \bigcup \{\mu_* \geq \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}\} \bigcup \{\mu_* \leq \mu_k + 2\sqrt{\frac{2\log t}{T_k(t)}}, \pi_t = k\} \right\} \end{aligned}$$

And from a union bound, we have

$$\begin{aligned} \mathbb{P}(\hat{\mu}_{k,t} - \mu_k < -2\sqrt{\frac{2\log t}{T_k(t)}}) &= \mathbb{P}(\hat{\mu}_{k,t} - \mu_k < 2\sqrt{\frac{2\log t}{T_k(t)}}) \\ &\leq \sum_{s=1}^{t} \exp(\frac{-s\frac{8\log t}{s}}{2}) \\ &= \frac{1}{t^3} \end{aligned}$$

Thus $\mathbb{P}(\mu_k > \hat{\mu}_{k,t} + 2\sqrt{\frac{2\log t}{T_k(t)}}) \leq \frac{1}{t^3}$ and similarly we have $\mathbb{P}(\mu_* > \hat{\mu}_{*,t} + 2\sqrt{\frac{2\log t}{T_*(t)}}) \leq \frac{1}{t^3}$, whereby

$$\begin{aligned}
\sum_{t=K+1}^{n} \mathbb{P}(\pi_t = k) &\leq 2\sum_{t=1}^{n} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(\mu_* \leq \mu_k + 2\sqrt{\frac{2\log t}{T_k(t)}}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(T_k(t) \leq \frac{8\log t}{\Delta_k^2}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{t=1}^{n} \mathbb{P}(T_k(t) \leq \frac{8\log n}{\Delta_k^2}, \pi_t = k) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^3} + \sum_{s=1}^{\infty} \mathbb{P}(s \leq \frac{8\log n}{\Delta_k^2}) \\
&\leq 2\sum_{t=1}^{\infty} \frac{1}{t^2} + \frac{8\log n}{\Delta_k^2} \\
&= \frac{\pi^2}{3} + \frac{8\log n}{\Delta_k^2} \ ,
\end{aligned}$$

where s is the counter of pulling arm k. Therefore we have

$$R_n = \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)] \leq \sum_{k,\Delta_k>0} \Delta_k (1 + \frac{\pi^2}{3} + \frac{8\log n}{\Delta_k^2}) \ ,$$

which furnishes the proof. □

Consider the case $K = 2$ at first, then from the theorem above we know $R_n \sim \frac{\log n}{\Delta}$, which is consistent with intuition that when the difference of two arm is small, it is hard to distinguish which to choose. On the other hand, it always hold that $R_n \leq n\Delta$. Combining these two results, we have $R_n \leq \frac{\log n}{\Delta} \wedge n\Delta$, whereby $R_n \leq \frac{\log(n\Delta^2)}{\Delta}$ up to a constant. Actually it turns out to be the optimal bound. When $K \geq 3$, we can similarly get the result that $R_n \leq \sum_k \frac{\log(n\Delta_k^2)}{\Delta_k}$. This, however, is not the optimal bound. The optimal bound should be $\sum_k \frac{\log(n/H)}{\Delta_k}$, which includes the harmonic sum and $H = \sum_k \frac{1}{\Delta_k^2}$. See [Lat15].

3.4 Bounded Regret

From above we know UCB policy can give regret that increases with at most rate $\log n$ with n. In this section we would consider whether it is possible to have bounded regret. Actually it turns out that if there is a known separator between the expected reward of optimal arm and other arms, there is a bounded regret policy.

We would only consider the case when $K = 2$ here. Without loss of generality, we assume $\mu_1 = \frac{\Delta}{2}$ and $\mu_2 = -\frac{\Delta}{2}$, then there is a natural separator 0.

Algorithm 2 Bounded Regret Policy (BRP)

```
$\pi_1 = 1$ and $\pi_2 = 2$
for $t = 3$ to $n$ do
  if $\max_k \hat{\mu}_{k,t} > 0$ then
    then $\pi_t = \text{argmax}_k \hat{\mu}_{k,t}$
  else
    $\pi_t = 1$, $\pi_{t+1} = 2$
  end if
end for
```

Theorem: BRP has regret

$$R_n \leq \Delta + \frac{16}{\Delta} \ .$$

Proof.

$$\mathbb{P}(\pi_t = 2) = \mathbb{P}(\hat{\mu}_{2,t} > 0, \pi_t = 2) + \mathbb{P}(\hat{\mu}_{2,t} \leq 0, \pi_t = 2)$$

Note that

$$\begin{aligned} \sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{2,t} > 0, \pi_t = 2) &\leq \mathbb{E} \sum_{t=3}^{n} \mathbb{1}(\hat{\mu}_{2,t} > 0, \pi_t = 2) \\ &\leq \mathbb{E} \sum_{t=3}^{n} \mathbb{1}(\hat{\mu}_{2,t} - \mu_2 > 0, \pi_t = 2) \\ &\leq \sum_{s=1}^{\infty} e^{-\frac{s\Delta^2}{8}} \\ &= \frac{8}{\Delta^2} \ , \end{aligned}$$

where s is the counter of pulling arm 2 and the third inequality is a Chernoff bound. Similarly,

$$\begin{aligned} \sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{2,t} \leq 0, \pi_t = 2) &= \sum_{t=3}^{n} \mathbb{P}(\hat{\mu}_{1,t} \leq 0, \pi_{t-1} = 1) \\ &\leq \frac{8}{\Delta^2} \ , \end{aligned}$$

Combining these two inequality, we have

$$R_n \leq \Delta(1 + \frac{16}{\Delta^2}) \ ,$$

□

References

[Lat15] Tor Lattimore, *Optimally confident UCB : Improved regret for finite-armed bandits*, Arxiv:1507.07880, 2015.

18.657: Mathematics of Machine Learning

Lecturer: Alexander Rakhlin
Scribe: Kevin Li

Lecture 19
Nov. 16, 2015

4. PREDICTION OF INDIVIDUAL SEQUENCES

In this lecture, we will try to predict the next bit given the previous bits in the sequence. Given completely random bits, it would be impossible to correctly predict more than half of the bits. However, certain cases including predicting bits generated by a human can be correct greater than half the time due to the inability of humans to produce truly random bits. We will show that the existence of a prediction algorithm that can predict better than a given threshold exists if and only if the threshold satifies certain probabilistic inequalities. For more information on this topic, you can look at the lecture notes at http://stat.wharton.upenn.edu/~rakhlin/courses/stat928/stat928_notes.pdf

4.1 The Problem

To state the problem formally, given a sequence $y_1, \dots, y_n, \dots \in \{-1, +1\}$, we want to find a prediction algorithm $\hat{y}_t = \hat{y}_t(y_1, \dots, y_{t-1})$ that correctly predicts y_t as much as possible.

In order to get a grasp of the problem, we will consider the case where $y_1, \dots, y_n \overset{iid}{\sim} Ber(p)$. It is easy to see that we can get

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}\right] \to \min\{p, 1-p\}$$

by letting $\hat{y}_t$ equal majority vote of the first $t-1$ bits. Eventually, the bit that occurs with higher probability will always have occurred more times. So the central limit theorem shows that our loss will approach $\min\{p, 1-p\}$ at the rate of $O(\frac{1}{\sqrt{n}})$.

Knowing that the distribution of the bits are iid Bernoulli random variables made the prediction problem fairly easy. More surprisingly is the fact that we can achieve the same for any individual sequence.

Claim: There is an algorithm such that the following holds for any sequence $y_1, \dots, y_n, \dots$.

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\} - \min\{\bar{y}_n, 1-\bar{y}_n\} \leq 0 \text{ a.s.}$$

It is clear that no deterministic strategy can achieve this bound. For any deterministic strategy, we can just choose $y_t = -\hat{y}_t$ and the predictions would be wrong every time. So we need a non-deterministic algorithm that chooses $\hat{q}_t = \mathbb{E}[\hat{y}_t] \in [-1, 1]$.

To prove this claim, we will look at a more general problem. Take a fixed horizon $n \geq 1$, and function $\phi : \{\pm 1\}^n \to \mathbb{R}$. Does there exist a randomized prediction strategy such that for any $y_1, \dots, y_n$

$$\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \dots, y_n) \ ?$$

For certain ϕ such as $\phi \equiv 0$, it is clear that no randomized strategy exists. However for $\phi \equiv \frac{1}{2}$, the strategy of randomly predicting the next bit ($\hat{q}_t = 0$) satisfies the inequality.

Lemma: For a stable ϕ, the following are equivalent

a) $\exists (\hat{q}_t)_{t=1,\ldots,n} \forall y_1, \ldots, y_n \quad \mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \ldots, y_n)$

b) $\mathbb{E}[\phi(\epsilon_1, \ldots, \epsilon_n)] \geq \frac{1}{2}$ where $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables

where stable is defined as follows

Definition (Stable Function): A function $\phi : \{\pm 1\}^n \to \mathbb{R}$ is stable if

$$|\phi(\ldots, y_i, \ldots) - \phi(\ldots, -y_i, \ldots)| \leq \frac{1}{n}$$

Proof. ($a \implies b$) Suppose $\mathbb{E}\phi < \frac{1}{2}$. Take $(y_1, \ldots, y_n) = (\epsilon_1, \ldots, \epsilon_n)$. Then $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq \epsilon_t\}] = \frac{1}{2} > \mathbb{E}[\phi]$ so there must exist a sequence $(\epsilon_1, \ldots, \epsilon_n)$ such that $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq \epsilon_t\}] > \phi(\epsilon_1, \ldots, \epsilon_n)$.

($b \implies a$) Recursively define $V(y_1, \ldots, y_t)$ such that $\forall y_1, \ldots, y_n$

$$V(y_1, \ldots, y_{t-1}) = \min_{q_t \in [-1,1]} \max_{y_t \in \pm 1} \left(\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{y}_t \neq y_t\}] + V(y_1, \ldots y_n)\right)$$

Looking at the definition, we can see that $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] = V(\emptyset) - V(y_1, \ldots, y_n)$. Now we note that $V(y_1, \ldots, y_t) = -\frac{t}{2n} - \mathbb{E}[\phi(y_1, \ldots, y_t, \epsilon_{t+1}, \ldots, \epsilon_n)]$ satisfies the recursive definition since

$$\begin{aligned}
&\min_{\hat{q}_t} \max_{y_t} \frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{y}_t \neq y_t\}] - \mathbb{E}[\phi(y_1, \ldots, y_t, \epsilon_{t+1}, \ldots, \epsilon_n)] - \frac{t}{2n} \\
=& \min_{\hat{q}_t} \max_{y_t} \frac{-\hat{q}_t y_t}{2n} - \mathbb{E}[\phi(y_1, \ldots, y_t, \epsilon_{t+1}, \ldots, \epsilon_n)] - \frac{t-1}{2n} \\
=& \min_{\hat{q}_t} \max\{-\frac{\hat{q}_t}{2n} - \mathbb{E}[\phi(y_1, \ldots, y_{t-1}, 1, \epsilon_{t+1}, \ldots, \epsilon_n)] - \frac{t-1}{2n}, \frac{\hat{q}_t}{2n} - \mathbb{E}[\phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n)] - \frac{t-1}{2n}\} \\
=& - \mathbb{E}[\phi(y_1, \ldots, y_{t-1}, \epsilon_t, \epsilon_{t+1}, \ldots, \epsilon_n)] - \frac{t-1}{2n} \\
=& V(y_1, \ldots, y_{t-1})
\end{aligned}$$

The first equality uses the fact that for $a, b \in \{\pm 1\}$, $\mathbb{1}\{a \neq b\} = \frac{1-ab}{2}$, the second uses the fact that $y_t \in \{\pm 1\}$, the third minimizes the entire expression by choosing $\hat{q}_t$ so that the two expressions in the max are equal. Here the fact that ϕ is stable means $\hat{q}_t \in [-1, 1]$ and is the only place where we need ϕ to be stable.

Therefore we have

$$\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] = V(\emptyset) - V(y_1, \ldots, y_n) = -\mathbb{E}[\phi(\epsilon_1, \ldots, \epsilon_n)] + \frac{1}{2} + \phi(y_1, \ldots, y_n) \leq \phi(y_1, \ldots, y_n)$$

by b). □

By choosing $\phi = \min\{\bar{y}, 1-\bar{y}\} + \frac{c}{\sqrt{n}}$, this shows there is an algorithm that satisfies our original claim.

4.2 Extensions

4.2.1 Supervised Learning

We can extend the problem to a regression type problem by observing x_t and trying to predict y_t. In this case, the objective we are trying to minimize would be

$$\frac{1}{n}\sum l(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum l(f(x_t), y_t)$$

It turns out that the best achievable performance in such problems is governed by martingale (or, sequential) analogues of Rademacher averages, covering numbers, combinatorial dimensions, and so on. Much of Statistical Learning techniques extend to this setting of online learning. In addition, the minimax/relaxation framework gives a systematic way of developing new prediction algorithms (in a way similar to the bit prediction problem).

4.2.2 Equivalence to Tail Bounds

We can also obtain probabilistic tail bound on functions ϕ on hypercube by using part a) of the earlier lemma. Rearranging part a) of the lemma we get $1 - 2\phi(y_1, \ldots, y_n) \le \frac{1}{n}\sum \hat{q}_t y_t$. This implies

$$\mathbb{P}\big(\phi(\epsilon_1, \ldots, \epsilon_n) < \frac{1-\mu}{2}\big) = \mathbb{P}\big(1 - 2\phi(\epsilon_1, \ldots, \epsilon_n) > \mu\big) \le \mathbb{P}\big(\frac{1}{n}\sum \hat{q}_t\epsilon_t > \mu\big) \le e^{-\frac{\mu^2}{2n}}$$

So $\mathbb{E}\phi \ge \frac{1}{2} \implies$ existence of a strategy $\implies$ tail bound for $\phi < \frac{1}{2}$.

We can extend the results to higher dimensions. Consider $z_1, \ldots, z_n \in B_2$ where B_2 is a ball in a Hilbert space. We can define recursively $\hat{y}_0 = 0$ and $\hat{y}_{t+1} = \mathrm{Proj}_{B_2}(\hat{y}_t - \frac{1}{\sqrt{n}} z_t)$. Based on the properties of projections, for every $y^* \in B_2$, we have $\frac{1}{n}\sum\langle \hat{y}_t - y^*, z_t\rangle \le \frac{1}{\sqrt{n}}$. Taking $y^* = \dfrac{\sum z_t}{\|\sum z_t\|}$,

$$\forall z_1, \ldots, z_n, \quad \|\sum_{t=1}^n z_t\| - \sqrt{n} \le \sum_{t=1}^n \langle \hat{y}_t, -z_t\rangle$$

Take a martingale difference sequence $Z_1, \ldots, Z_n$ with values in B_2. Then

$$\mathbb{P}\big(\|\sum_{t=1}^n Z_t\| - \sqrt{n} > \mu\big) \le \mathbb{P}(\sum_{t=1}^n \langle \hat{y}_t, -Z_t\rangle > \mu) \le e^{-\frac{n\mu^2}{2}}$$

Integrating out the tail,

$$\mathbb{E}\|\sum_{t=1}^n Z_t\| \le c\sqrt{n}$$

It can be shown using Von Neumann minimax theorem that

$$\exists (\hat{y}_t) \forall z_1, \ldots, z_n, y^* \in B_2 \quad \sum_{t=1}^{n} \langle \hat{y}_t - y^*, z_t \rangle \ \le \sup_{\text{MDS}W_1,\ldots,W_n} E\| \sum_{t=1}^{n} W_t \| \le c\sqrt{n}$$

where the supremum is over all martingale difference sequences (MDS) with values in B_2. By the previous part, this upper bound is $c\sqrt{n}$. We conclude an interesting equivalence of (a) deterministic statements that hold for all sequences, (b) tail bounds on the size of a martingale, and (c) in-expectation bound on this size.

In fact, this connection between probabilistic bounds and existence of prediction strategies for individual sequences is more general and requires further investigation.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet　　Lecture 20

Scribe: Vira Semenova　　Nov. 23, 2015

In this lecture, we talk about the adversarial bandits under limited feedback. Adversarial bandit is a setup in which the loss function $l(a,z) : \mathcal{A}x\mathcal{Z}$ is determinitic. Limited feedback means that the information available to the DM after the step t is $\mathcal{I}_t = \{l(a_1, z_1), ..., l(a_{t-1}, z_t)\}$, namely consits of the realised losses of the past steps only.

5. ADVERSARIAL BANDITS

Consider the problem of prediction with expert advice. Let the set of adversary moves be $\mathcal{Z}$ and the set of actions of a decision maker $\mathcal{A} = \{e_1, ..., e_K\}$. At time t, $a_t \in \mathcal{A}$ and $z_t \in \mathcal{Z}$ are simultaneously revealed.Denote the loss associated to the decision $a_t \in \mathcal{A}$ and his adversary playing z_t by $l(a_t, z_t)$. We compare the total loss after n steps to the minimum expert loss, namely:

$$\min_{1 \le j \le K} \sum_{t=1}^{n} l_t(e_j, z_t),$$

where e_j is the choice of expert $j \in \{1, 2, .., K\}$.

The cumulative regret is then defined as

$$R_n = \sum_{t=1}^{n} l_t(a_t, z_t) - \min_{1 \le j \le K} \sum_{t=1}^{n} l_t(e_j, z_t)$$

The feedback at step t can be either full or limited. The full feedback setup means that the vector $f = (l(e_1, z_t), ..., l(e_K, z_t))^\top$ of losses incurred at a pair of adversary's choice z_t and each bandit $e_j \in \{e_1, ..., e_K\}$ is observed after each step t. Hence, the information available to the DM after the step t is $\mathcal{I}_t = \cup_{t'=1}^{t} \{l(a_1, z_t'), ..., l(a_K, z_{t'})\}$. The limited feedback means that the time $-t$ feedback consists of the realised loss $l(a_t, z_t)$ only. Namely, the information available to the DM is $\mathcal{I}_t = \{l(a_1, z_1), ..., l(a_t, z_t)\}$. An example of the first setup is portfolio optimization problems, where the loss of all possible portfolios is observed at time t. An example of the second setup is a path planning problem and dynamic pricing, where the loss of the chosen decision only is observed. This lecture has limited feedback setup.

The two strategies, defined in the past lectures, were exponential weights, which yield the regret of order $R_n \le c\sqrt{n \log K}$ and Follow the Perturbed Leader. We would like to play exponential weights, defined as:

$$p_{t,j} = \frac{\exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))}{\sum_{l=1}^{k} \exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))}$$

This decision rule is not feasible, since the loss $l(e_j, z_t)$ are not part of the feedback if $e_j \neq a_t$. We will estimate it by

$$\hat{l}(e_j, z_t) = \frac{l(e_j, z_t) \mathbb{1}(a_t = e_j)}{P(a_t = e_j)}$$

Lemma: $\hat{l}(e_j, z_t)$ is an unbiased estimator of $l(e_j, z_t)$

Proof. $E_{a_t}\hat{l}(e_j, z_t) = \sum_{k=1}^{K} \frac{l(e_k,z_t)\mathbb{1}(e_k=e_t)}{P(a_t=e_j)} P(a_t = e_k) = l(e_j, z_t)$ □

Definition (Exp 3 Algorithm): Let $\eta > 0$ be fixed. Define the exponential weights as

$$p_{t+1,j}^{(j)} = \frac{\exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}{\sum_{l=1}^{k} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}$$

($Exp3$ stands for Exponential weights for Exploration and Exploitation.)

We will show that the regret of Exp3 is bounded by $\sqrt{2nK \log K}$. This bound is $\sqrt{K}$ times bigger than the bound on the regret under the full feedback. The $\sqrt{K}$ multiplier is the price of have smaller information set at the time t. The are methods that allow to get rid of $\log K$ term in this expression. On the other hand, it can be shown that $\sqrt{2nK}$ is the optimal regret.

Proof. Let $W_{t,j} = \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))$, $W_t = \sum_{j=1}^{k} W_{t,j}$, and $p_t = \frac{\sum_{j=1}^{K} W_{t,j} e_j}{W_t}$.

$$\log(\frac{W_{t+1}}{W_t}) = \log(\frac{\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s)) \exp(-\eta \hat{l}(e_j, z_t))}{\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}) \tag{5.1}$$

$$= \log(\mathbb{E}_{\mathcal{J} \sim p_t} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_{\mathcal{J}}, z_s))) \tag{5.2}$$

$$\leq^* \log(1 - \eta \mathbb{E}_{\mathcal{J} \sim p_t} \hat{l}(e_{\mathcal{J}}, z_s) + \frac{\eta^2}{2} \mathbb{E}_{\mathcal{J} \sim p_t} \hat{l}^2(e_{\mathcal{J}}, z_s) \tag{5.3}$$

where $*$ inequality is obtained by plugging in $\mathbb{E}_{\mathcal{J} \sim p_t} \hat{l}(e_{\mathcal{J}}, z_t)$ into the inequality

$$\exp x \geq 1 - \eta x + \frac{\eta^2 x^2}{2}$$

.

$$\mathbb{E}_{\mathcal{J} \sim p_t} \hat{l}(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \hat{l}(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \frac{l(e_j, z_t) \mathbb{1}(a_t = e_j)}{P(a_t = e_j)} = l(a_t, z_t) \tag{5.4}$$

$$\mathbb{E}_{\mathcal{J} \sim p_t} \hat{l}^2(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \hat{l}^2(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \frac{l^2(e_j, z_t) \mathbb{1}(a_t = e_j)}{P^2(a_t = e_j)} \tag{5.5}$$

$$= \frac{l^2(e_j, z_t)}{P_{a_t,t}} \leq \frac{1}{P_{a_t,t}} \tag{5.6}$$

Summing from 1 through n, we get
$\log(W_{t+1}) \leq \log(W_1) - \eta \sum_{t=1}^{n} l(a_t, z_t) + \frac{\eta^2}{2} \sum \frac{1}{P_{a_t,t}}$.

For $t = 1$, we initialize $w_{1,j} = 1$, so $W_1 = K$.
Since $\mathbb{E}_{\mathcal{J}} \frac{1}{P_{a_t,t}} = \sum_{j=1}^{K} \frac{p_{j,t}}{p_{j,t}} = K$, the expression above becomes
$\mathbb{E} \log(W_{n+1}) - \log K \leq -\eta \sum_{t=1}^{n} l(a_t, z_t) + \frac{\eta^2 Kn}{2}$
Noting that $\log(W_{n+1}) = \log(\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))$
and defining $j^* = \operatorname{argmin}_{1 \leq j \leq K} \sum_{t=1}^{n} l(e_j, z_t)$, we obtain:

$$\log(W_{n+1}) \geq \log(\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))) = -\eta \sum_{s=1}^{t-1} l(e_j, z_s)$$

Together:

$$\sum_{t=1}^{n} l(a_t, z_t) - \min_{1 \leq j \leq K} \sum_{t=1}^{n} l(e_j, z_t) \leq \frac{\log K}{\eta} + \frac{\eta K n}{2}$$

The choice of $\eta := \sqrt{2 \log KnK}$ yields the bound $R_n \leq \sqrt{2K \log Kn}$. □

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 21
Scribe: Ali Makhdoumi Nov. 25, 2015

6. LINEAR BANDITS

Recall form last lectures that in prediction with expert advise, at each time t, the player plays $a_t \in \{e_1, \dots, e_k\}$ and the adversary plays z_t such that $l(a_t, z_t) \leq 1$ for some loss function. One example of such loss function is linear function $l(a_t, z_t) = a_t^T z_t$ where $|z_t|_\infty \leq 1$. Linear bandits are a more general setting where the player selects an action $a_t \in \mathcal{A} \subset \mathbb{R}^k$, where $\mathcal{A}$ is a convex set and the adversary selects $z_t \in \mathcal{Z}$ such that $|z_t^T a_t| \leq 1$. Similar to the prediction with expert advise, the regret is defined as

$$R_n = \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a \in \mathcal{K}} \sum_{t=1}^n a^T z_t,$$

where A_t is a random variable in $\mathcal{A}$. Note that in the prediction with expert advise, the set $\mathcal{A}$ was essentially a polyhedron and we had $\min_{a\in\mathcal{K}} \sum_{t=1}^n a^T z_t = \min_{1\leq j\leq k} e_j^T z_t$. However, in the linear bandit setting the minimizer of $a^T z_t$ can be any point of the set $\mathcal{A}$ and essentially the umber experts that the player tries to "compete" with are infinity. Similar, to the prediction with expert advise we have two settings:

1 **Full feedback:** after time t, the player observes z_t.

2 **Bandit feedback:** after time t, the player observes $A_t^T z_t$, where A_t is the action that player has chosen at time t.

We next, see if we can use the bounds we have developed in the prediction with expert advise in this setting. In particular, we have shown the following bounds for prediction with expert advise:

1 **Prediction with k expert advise, full feedback:** $R_n \leq \sqrt{2n \log k}$.

2 **Prediction with k expert advise, bandit feedback:** $R_n \leq \sqrt{2nk \log k}$.

The idea to deal with linear bandits is to discretize the set $\mathcal{A}$. Suppose that $\mathcal{A}$ is bounded (e.g., $\mathcal{A} \subset B_2$, where B_2 is the l_2 ball in $\mathbb{R}^k$). We can use a $\frac{1}{n}$-covering of $\mathcal{A}$ which we have shown to be of size (smaller than)$O(n^k)$. This means there exist $y_1, \dots, y_{|\mathcal{N}|}$ such that for any $a \in \mathcal{A}$, there exist y_i such that $||y_i - a|| \leq \frac{1}{n}$. We now can bound the regret for general case, where the experts can be any point in $\mathcal{A}$, based on the regret on the discrete set, $\mathcal{N} = \{y_1, \dots, y_{|\mathcal{N}|}\}$,as follows.

$$\begin{aligned}
R_n &= \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a\in\mathcal{A}} \sum_{t=1}^n a^T z_t \\
&= \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a\in\mathcal{N}} \sum_{t=1}^n a^T z_t + o(1).
\end{aligned}$$

Therefore, we restrict actions A_t to a combination of the actions that belong to $\{y_1, \dots, y_{|\mathcal{N}|}\}$ (we can always do this), then using the bounds for the prediction with expert advise, we obtain the following bounds:

1 **Linear bandit, full feedback:** $R_n \leq \sqrt{2n\log(n^k)} = O(\sqrt{kn\log n})$, which in terms of dependency to n is of order $O(\sqrt{n})$ that is what we expect to have.

2 **Linear bandit, bandit feedback:** $R_n \leq \sqrt{2nn^k\log(n^k)} = \Omega(n)$, which is useless in terms of dependency of n as we expect to obtain $O(\sqrt{n})$ behavior.

The topic of this lecture is to provide bounds for the linear bandit in the bandit feedback.
Problem Setup: Let us recap the problem formulation:

- at time t, player chooses action $a_t \in \mathcal{A} \subset [-1,1]^k$.
- at time t, adversary chooses $z_t \in \mathcal{Z} \subset \mathbb{R}^k$, where $a_t^T z_t = \langle a_t, z_t\rangle \in [0,1]$.
- Bandit feedback: player observes $\langle a_t, z_t\rangle$(rather than z_t in the full feedback setup).

Literature: $O(n^{3/4})$ regret bound has been shown in [BB04]. Later on this bound has been improved to $O(n^{2/3})$ in [BK04] and [VH06] with "Geometric Hedge algorithm", which we will describe and analyze below. We need the following assumption to show the results:
Assumption: There exist δ such that $\delta e_1, \ldots, \delta e_k \in \mathcal{A}$. This assumption guarantees that $\mathcal{A}$ has full-dimension around zero.
We also discretize $\mathcal{A}$ with a $\frac{1}{n}$-net of size Cn^k and only consider the resulting discrete set and denote it by $\mathcal{A}$, where $|\mathcal{A}| \leq (3n)^k$. All we need to do is to bound

$$R_n = \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a\in\mathcal{A}} \sum_{t=1}^n a^T z_t.$$

For any t and a, we define

$$p_t(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{z}_s^T a\right)}{\sum_{a\in\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{z}_s^T a\right)},$$

where η is a parameter (that we will choose later) and $\hat{z}_t$ is defined to incorporate the idea of exploration versus exploitation. The algorithm which is termed *Geometric Hedge Algorithm* is as follows:
At time t we have

- **Exploitation:** with probability $1-\gamma$ draw a_t according to p_t and let $\hat{z}_t = 0$.
- **Exploration:** with probability $\frac{\gamma}{k}$ let $a_t = \delta e_j$ for some $1 \leq j \leq k$ and $\hat{z}_t = \frac{k}{\delta^2\gamma}\langle a_t, z_t\rangle a_t = \frac{k}{\gamma} z_t^{(j)} e_j$.

Note that δ is the the parameter that we have by assumption on the set $\mathcal{A}$, and η and γ are the parameters of the algorithm that we shall choose later.

Theorem: Using Geometric Hedge algorithm for linear bandit with bandit feedback, with $\gamma = \frac{1}{n^{1/3}}$ and $\eta = \sqrt{\frac{\log n}{kn^{4/3}}}$, we have

$$\mathbb{E}[R_n] \leq Cn^{2/3}\sqrt{\log n}\; k^{3/2}.$$

Proof. Let the overall distribution of a_t be q_t defined as $q_t = (1-\gamma)p_t + \gamma U$, where U is a uniform distribution over the set $\{\delta e_1, \ldots, \delta e_k\}$. Under this distribution, $\hat{z}_t$ is an unbiased estimator of z_t, i.e.,

$$\mathbb{E}_{a_t \sim q_t}[\hat{z}_t] = 0(1-\gamma) + \sum_{j=1}^{k} \frac{\gamma}{k} \frac{k}{\gamma} z_t^{(j)} e_j = z_t.$$

following the same lines of the proof that we had for analyzing exponential weight algorithm, we will define

$$w_t = \sum_{a \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} a^T \hat{z}_s\right).$$

We then have

$$\begin{aligned}
\log\left(\frac{w_{t+1}}{w_t}\right) &= \log\left(\sum_{a \in \mathcal{A}} p_t(a) \exp\left(-\eta a^T \hat{z}_t\right)\right) \\
&\overset{e^{-x} \leq 1 - x + \frac{x^2}{2}}{\leq} \log\left(\sum_{a \in \mathcal{A}} p_t(a) \left(1 - \eta a^T \hat{z}_t + \frac{1}{2}\eta^2 (a^T \hat{z}_s)^2\right)\right) \\
&= \log\left(1 + \sum_{a \in \mathcal{A}} p_t(a) \left(-\eta a^T \hat{z}_t + \frac{1}{2}\eta^2 (a^T \hat{z}_t)^2\right)\right) \\
&\overset{\log(1+x) \leq x}{\leq} \sum_{a \in \mathcal{A}} p_t(a) \left(-\eta a^T \hat{z}_t + \frac{1}{2}\eta^2 (a^T \hat{z}_t)^2\right).
\end{aligned}$$

Taking expectation from both sides leads to

$$\begin{aligned}
&\mathbb{E}_{a_t \sim q_t}\left[\log\left(\frac{w_{t+1}}{w_t}\right)\right] \leq -\eta \mathbb{E}_{a_t \sim q_t}\left[\sum_{a \in \mathcal{A}} p_t(a) a^T \hat{z}_t\right] + \frac{\eta^2}{2} \mathbb{E}_{a_t \sim q_t}\left[\sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right] \\
&= -\eta \mathbb{E}_{a_t \sim p_t}\left[a_t^T \hat{z}_t\right] + \frac{\eta^2}{2} \mathbb{E}_{a_t \sim q_t}\left[\sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right] \\
&\overset{q_t = (1-\gamma)p_t + \gamma U}{=} \frac{-\eta}{1-\gamma} \mathbb{E}_{a_t \sim q_t}\left[a_t^T \hat{z}_t\right] + \eta \frac{\gamma}{1-\gamma} \mathbb{E}_{a_t \sim U}\left[a_t^T \hat{z}_t\right] + \frac{\eta^2}{2} \mathbb{E}_{a_t \sim q_t}\left[\sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right] \\
&\overset{a_t^T z_t \leq 1}{\leq} \frac{-\eta}{1-\gamma} \mathbb{E}_{a_t \sim q_t}\left[a_t^T \hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma} + \frac{\eta^2}{2} \mathbb{E}_{a_t \sim q_t}\left[\sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right].
\end{aligned}$$

We next, take summation of the previous relation for $t = 1$ up to n and use a telescopic cancellation to obtain

$$\begin{aligned}
\mathbb{E}\left[\log w_{n+1}\right] &\leq \mathbb{E}\left[\log w_1\right] - \frac{\eta}{1-\gamma} \mathbb{E}\left[\sum_{t=1}^{n} a_t^T \hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma} n + \frac{\eta^2}{2} \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right] \\
&\leq \mathbb{E}\left[\log w_1\right] - \eta \mathbb{E}\left[\sum_{t=1}^{n} a_t^T \hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma} n + \frac{\eta^2}{2} \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2\right]. \qquad (6.1)
\end{aligned}$$

Note that for all $a^* \in \mathcal{A}$ we have

$$\log(w_{n+1}) = \log\left(\sum_{a\in\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{n} a^T \hat{z}_s\right)\right) \geq -\eta \sum_{s=1}^{n} \langle a^*, \hat{z}_s\rangle.$$

Using $\mathbb{E}[\hat{z}_s] = z_s$, leads to

$$\mathbb{E}\left[\log(w_{n+1})\right] \geq -\eta \sum_{s=1}^{n} \langle a^*, z_s\rangle. \tag{6.2}$$

We also have that

$$\log(w_1) = \log|\mathcal{A}| \leq 2k\log n. \tag{6.3}$$

Plugging (6.2) and (6.3) into (6.1), leads to

$$\mathbb{E}[R_n] \leq \frac{\gamma}{1-\gamma} n + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] + \frac{2k\log n}{\eta}. \tag{6.4}$$

It remains to control the quadratic term $\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right]$. We use the fact that $|z_t^{(j)}|, |a_t^{(j)}| \leq 1$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] &= \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\mathbb{E}_{q_t}[(a^T\hat{z}_t)^2] \\
&= \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\left((1-\gamma)0 + \sum_{j=1}^{k}\frac{\gamma}{k}\left(\frac{k}{\gamma}\right)^2 [a^j z_t^{(j)}]^2\right) \\
&\overset{|a^j z_t^{(j)}|\leq 1}{\leq} \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\left(\frac{k^2}{\gamma}\right) = n\frac{k^2}{\gamma}.
\end{aligned}$$

Plugging this bound into (6.4), we have

$$\mathbb{E}[R_n] \leq \gamma n + \frac{\eta}{2} n\frac{k^2}{\gamma} + \frac{2k\log n}{\eta}.$$

Letting $\gamma = \frac{1}{n^{1/3}}$ and $\eta = \sqrt{\frac{\log n}{kn^{4/3}}}$ leads to

$$\mathbb{E}[R_n] \leq Ck^{3/2}n^{2/3}\sqrt{\log n}.$$

□

Literature: The bound we just proved has been improved in [VKH07] where they show $O(d^{3/2}\sqrt{n\log n})$ bound with a better exploration in the algorithm. The exploration that we used in the algorithm was coordinate-wise. The key is that we have a linear problem and we can use better tools from linear regression such as least square estimation. However, we will describe a slightly different approach in which we never explore and the exploration is completely done with the exponential weighting. This approach also gives a better performance in terms of the dependency on k. In particular, we obtain the bound $O(d\sqrt{n\log n})$ which coincides with the bound recently shown in [BCK 12] using a John's ellipsoid.

Theorem: Let $C_t = \mathbb{E}_{a_t \sim q_t}[a_t a_t^T]$, $\hat{z}_t = (a_t^T z_t) C_t^{-1} a_t$, and $\gamma = 0$ (so that $p_t = q_t$). Using Geometric Hedge algorithm with $\eta = 2\sqrt{\frac{\log n}{n}}$ for linear bandit with bandit feedback leads to

$$\mathbb{E}[R_n] \leq CK\sqrt{n \log n}.$$

Proof. We follow the same lines of the proof as the previous theorem to obtain (6.4). Note that the only fact that we used in order to obtain (6.4) is unbiasedness, i.e., $\mathbb{E}[\hat{z}_t] = z_t$, which holds here as well since

$$\mathbb{E}[\hat{z}_t] = \mathbb{E}[C_t^{-1} a_t a_t^T z_t] = C_t^{-1}\mathbb{E}[a_t a_t^T] z_t = z_t.$$

Note that we can use pseudo-inverse instead of inverse so that invertibility is not an issue. Therefore, rewriting (6.4) with $\gamma = 0$, we obtain

$$\mathbb{E}[R_n] \leq \frac{\eta}{2}\mathbb{E}_{a_t \sim p_t}\left[\sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a)(a^T \hat{z}_t)^2\right] + \frac{2k \log n}{\eta}.$$

We now bound the quadratic term as follows

$$\begin{aligned}
&\mathbb{E}_{a_t \sim p_t}\left[\sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a)(a^T \hat{z}_t)^2\right] = \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a)\mathbb{E}_{a_t \sim p_t}\left[(a^T \hat{z}_t)^2\right] \\
&\overset{C_t^T = C_t,\ \hat{z}_t = (a_t^T z_t) C_t^{-1} a_t}{=} \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a) a^T \mathbb{E}\left[\hat{z}_t \hat{z}_t^T\right] a = \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a) a^T \mathbb{E}\left[(a_t^T z_t)^2 C_t^{-1} a_t a_t^T C_t^{-1}\right] a \\
&\overset{|a_t^T z_t| \leq 1}{\leq} \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a) a^T C_t^{-1}\mathbb{E}\left[a_t a_t^T\right] C_t^{-1} a \overset{\mathbb{E}[a_t a_t^T] = C_t}{=} \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a) a^T C_t^{-1} a \\
&= \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a)\mathrm{tr}(a^T C_t^{-1} a) \overset{\mathrm{tr}(AB) = \mathrm{tr}(BA)}{=} \sum_{t=1}^{n}\sum_{a \in \mathcal{A}} p_t(a)\mathrm{tr}(C_t^{-1} a a^T) \\
&= \sum_{t=1}^{n}\mathrm{tr}(C_t^{-1}\mathbb{E}_{a \sim p_t}[a a^T]) = \sum_{t=1}^{n}\mathrm{tr}(C_t^{-1} C_t) = \sum_{t=1}^{n}\mathrm{tr}(I_k) = kn.
\end{aligned}$$

Plugging this bound into previous bound yields

$$\mathbb{E}[R_n] \leq \frac{\eta}{2} nk + \frac{2k \log n}{\eta}.$$

Letting $\eta = 2\sqrt{\frac{\log n}{n}}$, leads to $\mathbb{E}[R_n] \leq Ck\sqrt{n \log n}$. □

References

[BCK 12] Bubeck, Sbastien, Nicolo Cesa-Bianchi, and Sham M. Kakade. *Towards minimax policies for online linear optimization with bandit feedback.* arXiv preprint arXiv:1202.3079 (2012). APA

[BB04] McMahan, H. Brendan, and Avrim Blum. *Online geometric optimization in the bandit setting against an adaptive adversary.* Conference on Learning theory (COLT) 2004.

[VH06] Dani, Varsha, and Thomas P. Hayes. *Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary.* Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm. Society for Industrial and Applied Mathematics, 2006.

[BK04] Awerbuch, Baruch, and Robert D. Kleinberg. *Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches.*Proceedings of the thirty-sixth annual ACM symposium on Theory of computing. ACM, 2004.

[VKH07] Dani, Varsha, Sham M. Kakade, and Thomas P. Hayes, *The price of bandit information for online optimization*, Advances in Neural Information Processing Systems. 2007.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 22
Scribe: Aden Forrow — Nov. 30, 2015

7. BLACKWELL'S APPROACHABILITY

7.1 Vector Losses

David Blackwell introduced approachability in 1956 as a generalization of zero sum game theory to vector payoffs. Born in 1919, Blackwell was the first black tenured professor at UC Berkeley and the seventh black PhD in math in the US.

Recall our setup for online linear optimization. At time t, we choose an action $a_t \in \Delta_K$ and the adversary chooses $z_t \in B_\infty(1)$. We then get a loss $\ell(a_t, z_t) = \langle a_t, z_t \rangle$. In the full information case, where we observe z_t and not just $\ell(a_t, z_t)$, this is the same as prediction with expert advice. Exponential weights leads to a regret bound

$$R_n \leq \sqrt{\frac{n}{2}\log(K)}.$$

The setup of a zero sum game is nearly identical:

- Player 1 plays a mixed strategy $p \in \Delta_n$.
- Player 2 plays $q \in \Delta_m$.
- Player 1's payoff is $p^\top M q$.

Here M is the game's payoff matrix.

Theorem: Von Neumann Minimax Theorem

$$\max_{p\in\Delta_n}\min_{q\in\Delta_m} p^\top M q = \min_{q\in\Delta_m}\max_{p\in\Delta_n} p^\top M q.$$

The minimax is called the value of the game. Each player can prevent the other from doing any better than this. The minimax theorem implies that if there is a good response p_q to any individual q, then there is a silver bullet strategy p that works for any q.

Corollary: If $\forall q \in \Delta_n$, $\exists p$ such that $p^\top M q \geq c$, then $\exists p$ such that $\forall q$, $p^\top M q \geq c$.

Von Neumann's minimax theorem can be extended to more general sets. The following theorem is due to Sion (1958).

Theorem: Sion's Minimax Theorem Let A and Z be convex, compact spaces, and $f : A \times Z \to \mathbb{R}$. If $f(a, \cdot)$ is upper semicontinuous and quasiconcave on Z $\forall a \in A$ and

$f(\cdot, z)$ is lower semicontinuous and quasiconvex on A $\forall z \in Z$, then

$$\inf_{a \in A} \sup_{z \in Z} f(a, z) = \sup_{z \in Z} \inf_{a \in A} f(a, z).$$

(Note - this wasn't given explicitly in lecture, but we do use it later.) Quasiconvex and quasiconcave are weaker conditions than convex and concave respectively.

Blackwell looked at the case with vector losses. We have the following setup:

- Player 1 plays $a \in A$
- Player 2 plays $z \in Z$
- Player 1's payoff is $\ell(a, z) \in \mathbb{R}^d$

We suppose A and Z are both compact and convex, that $\ell(a, z)$ is bilinear, and that $\|\ell(a, z)\| \leq R$ $\forall a \in A$, $z \in Z$. All norms in this section are Euclidean norms. Can we translate the minimax theorem directly to this new setting? That is, if we fix a set $S \subset \mathbb{R}^d$, and if $\forall z$ $\exists a$ such that $\ell(a, z) \in S$, does there exist an a such that $\forall z$ $\ell(a, z) \in S$?

No. We'll construct a counterexample. Let $A = Z = [0, 1]$, $\ell(a, z) = (a, z)$, and $S = \{(a, z) \in [0, 1]^2 : a = z\}$. Clearly, for any $z \in Z$ there is an $a \in A$ such that $a = z$ and $\ell(a, z) \in S$, but there is no $a \in A$ such that $\forall z$, $a = z$.

Instead of looking for a single best strategy, we'll play a repeated game. At time t, player 1 plays $a_t = a_t(a_1, z_1, \ldots, a_{t-1}, z_{t-1})$ and player 2 plays $z_t = z_t(a_1, z_1, \ldots, a_{t-1}, z_{t-1})$. Player 1's average loss after n iterations is

$$\bar{\ell}_n = \frac{1}{n} \sum_{t=1}^{n} \ell(a_t, z_t)$$

Let $d(x, S)$ be the distance between a point $x \in \mathbb{R}^d$ and the set S, i.e.

$$d(x, S) = \inf_{s \in S} \|x - s\|.$$

If S is convex, the infimum is a minimum attained only at the projection of x in S.

Definition: A set S is *approachable* if there exists a strategy $a_t = a_t(a_1, z_1, \ldots, a_{t-1}, z_{t-1})$ such that $\lim_{n\to\infty} d(\bar{\ell}_n, S) = 0$.

Whether a set is approachable depends on the loss function $\ell(a, z)$. In our example, we can choose $a_0 = 0$ and $a_t = z_{t-1}$ to get

$$\lim_{n\to\infty} \bar{\ell}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} (z_{t-1}, z_t) = (\bar{z}, \bar{z}) \in S.$$

So this S is approachable.

7.2 Blackwell's Theorem

We have the same conditions on A, Z, and $\ell(a, z)$ as before.

Theorem: Blackwell's Theorem Let S be a closed convex set of $\mathbb{R}^2$ with $\|x\| \leq R$ $\forall x \in S$. If $\forall z$, $\exists a$ such that $\ell(a, z) \in S$, then S is approachable.

Moreover, there exists a strategy such that

$$d(\bar{\ell}_n, S) \leq \frac{2R}{\sqrt{n}}$$

Proof. We'll prove the rate; approachability of S follows immediately. The idea here is to transform the problem to a scalar one where Sion's theorem applies by using half spaces.

Suppose we have a half space $H = \{x \in \mathbb{R}^d : \langle w, x\rangle \leq c\}$ with $S \subset H$. By assumption, $\forall z$ $\exists a$ such that $\ell(a, z) \in H$. That is, $\forall z$ $\exists a$ such that $\langle w, \ell(a, z)\rangle \leq c$, or

$$\max_{z\in Z} \min_{a\in A} \langle w, \ell(a, z)\rangle \leq c.$$

By Sion's theorem,

$$\min_{a\in A} \max_{z\in Z} \langle w, \ell(a, z)\rangle \leq c.$$

So $\exists a^*_H$ such that $\forall z$ $\ell(a, z) \in H$.

This works for any H containing S. We want to choose H_t so that $\ell(a_t, z_t)$ brings the average $\bar{\ell}_t$ closer to S than $\bar{\ell}_{t-1}$. An intuitive choice is to have the hyperplane W bounding H_t be the separating hyperplane between S and $\bar{\ell}_{t-1}$ closest to S. This is Blackwell's strategy: let W be the hyperplane through $\pi_t \in \operatorname{argmin}_{\mu\in S} \|\bar{\ell}_{t-1} - \mu\|$ with normal vector $\bar{\ell}_{t-1} - \pi_t$. Then

$$H = \{x \in \mathbb{R}^d : \langle x - \pi_t, \bar{\ell}_{t-1} - \pi_t\rangle \leq 0\}.$$

Find a^*_H and play it.

We need one more equality before proving convergence. The average loss can be expanded:

$$\begin{aligned}\bar{\ell}_t &= \frac{t-1}{t}\bar{\ell}_{t-1} + \frac{1}{t}\ell_t \\ &= \frac{t-1}{t}(\bar{\ell}_{t-1} - \pi_t) + \frac{t-1}{t}\pi_t + \frac{1}{t}\ell_t\end{aligned}$$

Now we look at the distance of the average from S, using the above equation and the definition of π_{t+1}:

$$\begin{aligned}d(\bar{\ell}_t, S)^2 &= \|\bar{\ell}_t - \pi_{t+1}\|^2 \\ &\leq \|\bar{\ell}_t - \pi_t\|^2 \\ &= \left\|\frac{t-1}{t}(\bar{\ell}_{t-1} - \pi_t) + \frac{1}{t}(\ell_t - \pi_t)\right\|^2 \\ &= \left(\frac{t-1}{t}\right)^2 d(\bar{\ell}_{t-1}, S)^2 + \frac{\|\ell_t - \pi_t\|^2}{t^2} + 2\frac{t-1}{t^2}\langle \ell_t - \pi_t, \bar{\ell}_{t-1} - \pi_t\rangle\end{aligned}$$

Since $\ell_t \in H$, the last term is negative; since ℓ_t and π_t are both bounded by R, the middle term is bounded by $\frac{4R^2}{t^2}$. Letting $\mu_t^2 = t^2 d(\bar{\ell}_t, S)^2$, we have a recurrence relation

$$\mu_t^2 \leq \mu_{t-1}^2 + 4R^2,$$

implying

$$\mu_n^2 \leq 4nR^2.$$

Rewriting in terms of the distance gives the desired bound,

$$d(\bar{\ell}_t, S) \leq \frac{2R}{\sqrt{n}}$$

□

Note that this proof fails for nonconvex S.

7.3 Regret Minimization via Approachability

Consider the case $A = \Delta_K$, $Z = B_\infty^K(1)$. As we showed before, exponential weights $R_n \leq c\sqrt{n\log(K)}$. We can get the same dependence on n with an approachability-based strategy. First recall that

$$\begin{aligned}\frac{1}{n}R_n &= \frac{1}{n}\sum_{t=1}^{n}\ell(a_t, z_t) - \min_j \frac{1}{n}\sum_{t=1}^{n}\ell(e_j, z_t)\\ &= \max_j \left[\frac{1}{n}\sum_{t=1}^{n}\ell(a_t, z_t) - \frac{1}{n}\sum_{t=1}^{n}\ell(e_j, z_t)\right]\end{aligned}$$

If we define a vector average loss

$$\bar{\ell}_n = \frac{1}{n}\sum_{t=1}^{n}\left(\ell(a_t, z_t) - \ell(e_1, z_t), \ldots, \ell(a_t, z_t) - \ell(e_K, z_t)\right) \in \mathbb{R}^K,$$

$\frac{R_n}{n} \to 0$ if and only if all components of $\bar{\ell}_n$ are nonpositive. That is, we need $d(\bar{\ell}_n, O_K^-) \to 0$, where $O_K^- = \{x \in \mathbb{R}^K : -1 \leq x_i \leq 0, \forall i\}$ is the nonpositive orthant. Using Blackwell's approachability strategy, we get

$$\frac{R_n}{n} \leq d(\bar{\ell}_n, O_K^-) \leq c\sqrt{\frac{K}{n}}.$$

The K dependence is worse than exponential weights, $\sqrt{K}$ instead of $\sqrt{\log(K)}$.

How do we find a_H^*? As a concrete example, let $K = 2$. We need a_H^* tp satisfy

$$\langle w, \ell(a_H^*, z)\rangle = \langle w, \langle a_H^*, z\rangle y - z\rangle \leq c$$

for all z. Here y is the vector of all ones. Note that $c \geq 0$ since 0 is in S and therefore in H. Rearranging,

$$\langle a_H^*, z\rangle\langle w, y\rangle \leq \langle w, z\rangle + c,$$

Choosing $a_H^* = \frac{w}{\langle w, y\rangle}$ will work; the inequality reduces to

$$\langle w, z\rangle \leq \langle w, z\rangle + c.$$

Approachability in the bandit setting with only partial feedback is still an open problem.

References

[Bla56] D. Blackwell, *An analog of the minimax theorem for vector payoffs*, Pacific J. Math. 6 (1956), no. 1, 1–8

[Sio58] M. Sion, *On general minimax theorems.* Pacific J. Math. 8 (1958), no. 1, 171–176.

18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Jonathan Weed

Lecture 23
Dec. 2, 2015

1. POTENTIAL BASED APPROACHABILITY

Last lecture, we saw Blackwell's celebrated Approachability Theorem, which establishes a procedure by which a player can ensure that the average (vector) payoff in a repeated game approaches a convex set. The central idea was to construct a hyperplane separating the convex set from the point $\bar{\ell}_{t-1}$, the average loss so far. By projecting perpendicular to this hyperplane, we obtained a scalar-valued problem to which von Neumann's minimax theorem could be applied. The set S is approachable as long as we can always find a "silver bullet," a choice of action a_t for which the loss vector ℓ_t lies on the side of the hyperplane containing S. (See Figure 1.)

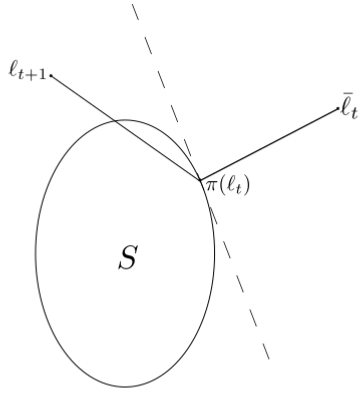


Figure 1: Blackwell approachability

Concretely, Blackwell's Theorem also implied the existence of a regret-minimizing algorithm for expert advice. Indeed, if we define the vector loss ℓ_t by $(\ell_t)_i = \ell(a_t, z_t) - \ell(e_i, z_t)$, then the average regret at time t is equivalent to the sup-norm distance between the average loss $\bar{\ell}_t$ and the negative orthant. Approaching the negative orthant therefore corresponds to achieving sublinear regret.

However, this reduction yielded suboptimal rates. To bound average regret, we replaced the sup-norm distance by the Euclidean distance, which led to an extra factor of $\sqrt{k}$ appearing in our bound. In the sequel, we develop a more sophisticated version of approachability that allows us to adapt to the geometry of our problem. (Much of what follows resembles out development of the mirror descent algorithm, though the two approaches differ in crucial details.)

1.1 Potential functions

We recall the setup of mirror descent, first described in Lecture 13. Mirror descent achieved accelerated rates by employing a potential function which was strongly convex with respect to the given norm. In this case, we seek what is in some sense the opposite: a function whose gradient does not change too quickly. In particular, we make the following definition.

Definition: A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a *potential* for $S \in \mathbb{R}$ if it satisfies the following properties:

- Φ is convex.
- $\Phi(x) \leq 0$ for $x \in S$.
- $\Phi(y) = 0$ for $y \in \partial S$.
- $\Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle \leq \frac{h}{2}\|x - y\|^2$, where by abuse of notation we use $\nabla\Phi(x)$ to denote a subgradient of Φ at x.

Given such a function, we recall two associated notions from the mirror descent algorithm. The Bregman divergence associated to Φ is given by

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle \,.$$

Likewise, the associated Bregman projection is

$$\pi(x) = \operatorname*{argmin}_{y \in S} D_\Phi(y, x)\,.$$

We aim to use the function Φ as a stand-in for the Euclidean distance that we employed in our proof of Blackwell's theorem. To that end, the following lemma establishes several properties that will allow us to generalize the notion of a separating hyperplane.

Lemma: For any convex, closed set S and $z \in S$, $x \in S^C$, the following properties hold.

- $\langle z - \pi(x), \nabla\Phi(x)\rangle \leq 0$,
- $\langle x - \pi(x), \nabla\Phi(x)\rangle \geq \Phi(x)$.

In particular, if Φ is positive on S^C, then $H := \{y \mid \langle y - \Phi(x), \nabla\Phi(x)\rangle = 0\}$ is a separating hyperplane.

Our proof requires the following proposition, whose proof appears in our analysis of the mirror descent algorithm and is omitted here.

Proposition: For all $z \in S$, it holds

$$\langle \nabla\Phi(\pi(x)) - \nabla\Phi(x), \pi(x) - z\rangle \leq 0\,.$$

Proof of Lemma. Denote by π the projection $\pi(x)$. The first claim follows upon expanding the expression on the left-hand side as follows

$$\langle z - \pi, \nabla\Phi(x)\rangle = \langle z - \pi, \nabla\Phi(x) - \nabla\Phi(\pi)\rangle + \langle z - \pi, \nabla\Phi(\pi)\rangle.$$

The above Proposition implies that the first term is nonpositive. Since the function Φ is convex, we obtain

$$0 \geq \Phi(z) \geq \Phi(\pi) + \langle z - \pi, \nabla\Phi(\pi)\rangle \,.$$

Since π lies on the boundary of S, by assumption $\Phi(\pi) = 0$ and the claim follows.

For the second claim, we again use convexity:

$$\Phi(\pi) \geq \Phi(x) + \langle \pi - x, \nabla\Phi(x)\rangle \,.$$

Since $\Phi(\pi) = 0$, the claim follows. □

1.2 Potential based approachability

With the definitions in place, the algorithm for approachability is essentially the same as it before we introduced the potential function. As before, we will use a projection defined by the hyperplane $H = \{y \mid \langle y - \pi(\bar{\ell}_{t-1}), \nabla\Phi(\bar{\ell}_{t-1}\rangle = 0\}$ and von Neumann's minmax theorem to find a "silver bullet" a_t^* such that $\ell_t = \ell(a_t^*, z_t)$ satisfies

$$\langle \ell_t - \pi_t, \nabla\Phi(\bar{\ell}_{t-1})\rangle \leq 0 \,.$$

All that remains to do is to analyze this procedure's performance. We have the following theorem.

Theorem: If $\|\ell(a,z)\| \leq R$ holds for all $z \in \mathcal{A}, z \in \mathcal{Z}$ and all assumptions above are satisfied, then

$$\Phi(\bar{\ell}_n) \leq \frac{4R^2 h \log n}{n} \,.$$

Proof. The definition of the potential Φ required that Φ be upper bounded by a quadratic function. The proof below is a simple application of that bound.

As before, we note the identity

$$\bar{\ell}_t = \bar{\ell}_{t-1} + \frac{\ell_t - \bar{\ell}_{t-1}}{t} \,.$$

This expression and the definition of Φ imply.

$$\Phi(\bar{\ell}_t) \leq \Phi(\bar{\ell}_{t-1}) + \frac{1}{t}\langle \ell_t - \bar{\ell}_{t-1}, \nabla\Phi(\bar{\ell}_{t-1})\rangle + \frac{h}{2t^2}\|\ell_t - \bar{\ell}_{t-1}\|^2 \,.$$

The last term is the easiest to control. By assumption, ℓ_t and $\bar{\ell}_{t-1}$ are contained in a ball of radius R, so $\|\ell_t - \bar{\ell}_{t-1}\|^2 \leq 4R^2$.

To bound the second term, write

$$\frac{1}{t}\langle \ell_t - \bar{\ell}_{t-1}, \nabla\Phi(\bar{\ell}_{t-1})\rangle = \frac{1}{t}\langle \ell_t - \pi_t, \nabla\Phi(\bar{\ell}_{t-1})\rangle + \frac{1}{t}\langle \pi_t - \bar{\ell}_{t-1}, \nabla\Phi(\bar{\ell}_{t-1})\rangle \,.$$

The first term is nonpositive by assumption, since this is how the algorithm constructs the silver bullet. By the above Lemma, the inner product in the second term is at most $-\Phi(\bar{\ell}_{t-1})$.

We obtain

$$\Phi(\bar{\ell}_t) \leq \left(\frac{t-1}{t}\right)\Phi(\bar{\ell}_{t-1}) + \frac{2hR^2}{t^2} \,.$$

Defining $u_t = t\Phi(\bar{\ell}_t)$ and rearranging, we obtain the recurrence

$$u_t \leq u_{t-1} + \frac{2hR^2}{t}\,,$$

So

$$u_n = \sum_{t=1}^{n} u_t - u_{t-1} \leq 2hR^2 \sum_{t=1}^{n} \frac{1}{t} \leq 4hR^2 \log n\,.$$

Applying the definition of u_n proves the claim. □

1.3 Application to regret minimization

We now show that potential based approachability provides an improved bound on regret minimization. Our ultimate goal is to replace the bound $\sqrt{nk}$ (which we proved last lecture) by $\sqrt{n \log k}$ (which we know to be the optimal bound for prediction with expert advice). We will be able to achieve this goal up to logarithmic terms in n. (A more careful analysis of the potential defined below does actually yields an optimal rate.)

Recall that $\frac{R_n}{n} = d_\infty(\bar{\ell}_n, O_K^-)$, where R_n is the cumulative regret after n rounds and O_K^- is the negative orthant. It is not hard to see that $d_\infty = \|x_+\|_\infty$, where x_+ is the positive part of the vector x.

We define the following potential function:

$$\Phi(x) = \frac{1}{\eta} \log \left(\frac{1}{K} \sum_{j=1}^{K} e^{\eta (x_j)_+} \right)\,.$$

The function Φ is a kind of "soft max" of the positive entries of x. (Note that this definition does not agree with the use of the term soft max in the literature—the difference is the presence of the factor $\frac{1}{K}$.) The terminology soft max is justified by noting that

$$\|x_+\|_\infty = \max_j (x_j)_+ \leq \max_j \frac{1}{\eta} \log \frac{1}{K} e^{\eta (x_j)_+} + \frac{\log K}{\eta} \leq \Phi(x) + \frac{\log K}{\eta}\,.$$

The potential function therefore serves as an upper bound on the sup distance, up to an additive logarithmic factor.

The function Φ defined in this way is clearly convex and zero on the negative orthant. To verify that it is a potential, it remains to show that Φ can be bounded by a quadratic.

Away from the negative orthant, Φ is twice differentiable and we can compute the Hessian explicitly:

$$\nabla^2 \Phi(x) = \eta \operatorname{diag}(\nabla \Phi(x)) - \eta \nabla \Phi \nabla \Phi^\top\,.$$

For any vector u such that $\|u\|_2 = 1$, we therefore have

$$u^\top \nabla^2 \Phi(x) u = \eta \sum_{j=1}^{K} u_j^2 (\nabla \Phi(x))_j - \eta (u^\top \nabla \Phi(x))^2 \leq \eta \sum_{j=1}^{K} u_j^2 (\nabla \Phi(x))_j \leq \eta\,,$$

since $\|u\|_2 = 1$ and $\|\nabla \Phi(x)\|_1 \leq 1$.

We conclude that $\nabla^2 \Phi(x) \preceq \eta I$, which for nonnegative x and y implies the bound

$$\Phi(y) - \Phi(x) - \langle \nabla \Phi(x), y - x \rangle \leq \frac{\eta}{2} \|y - x\|^2\,.$$

In fact, this bound holds everywhere. Therefore Φ is a valid potential function for the negative orthant, with $h = \eta$.

The above theorem then implies that we can ensure

$$\frac{R_n}{n} \leq \Phi(\bar{\ell}_n) + \frac{\log K}{\eta} \leq \frac{4R^2 \eta \log n}{n} + \frac{\log K}{\eta}\,.$$

To optimize this bound, we pick $\eta = \frac{1}{2R}\sqrt{\frac{n \log K}{\log n}}$ and obtain the bound

$$R_n \leq 4R\sqrt{n \log n \log K}\,.$$

As alluded to earlier, a more careful analysis can remove the $\log n$ term. Indeed, for this particular choice of Φ, we can modify the above Lemma to obtain the sharper bound

$$\langle x - \pi(x), \nabla \Phi(x) \rangle \geq 2\Phi(x)\,.$$

When we substitute this expression into the above proof, we obtain the recurrence relation

$$\Phi(\bar{\ell}_t) \leq \frac{t-2}{t}\Phi(\bar{\ell}_{t-1}) + \frac{c}{t^2}\,.$$

This small change is enough to prevent the appearance of $\log n$ in the final bound.