Stochastic Convex Optimization	**Fall 2021**

Lecture 6: Uniform Convergence Bounds

Lecturer: Roi Livni *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

We are now going to utilize the uniform convergence setup in order to obtain generalization bounds for stochastic convex problems. We begin with the first technique that is going to be quite general, in fact will extend beyond convex problems. However, the bounds we will obtain are going to be *dimension dependent*. Will then discuss how, under further structual assumptions we can obtain improved uniform convergence bounds that are dimension independent.

6.1 Covering Numbers

We now begin to analyze the uniform convergence rate of convex problems. Though, we will, in fact only assume Lipschitness, thus the result of this section apply to a much larger family of functions.

Theorem 6.1. *Let $f(w,z)$ be L-Lipschitz, bounded by $-1 \leq f \leq 1$ over $\mathcal{W}$ and $\mathcal{W} = B(0,R)$. Then $f(w,z)$ has the u.c.p with sample complexity:*

$$m_{uc} = \tilde{O}\left(\frac{n}{\epsilon^2}\log\frac{LR}{\delta}\right).$$

Exercise 6.1. *Given an L Lipschitz convex function, C bounded $f(w,z)$, defined over $\mathcal{W} = B(0,R)$. Conclude that for any optimization procedure that, given a sample S of size m, returns a solution w_S such that*

$$\hat{F}_m(w_S) \leq \min_{w\in\mathcal{W}} \hat{F}_m(w) + \epsilon,$$

we have that for $m = \tilde{O}\left(C\frac{n}{\epsilon^2}\log\frac{LR}{\delta}\right)$ then

$$F(w_S) \leq \min_{w\in\mathcal{W}} F(w) + \epsilon.$$

For the proof, let us start with a simpler case, where we have a finite set of parameters and we want to estimate the performance of each and every one.

Lemma 6.2. *Let $\mathcal{W} = \{w_1, \ldots, w_N\}$ be a finite set of parameters. Then if $m \geq \frac{\log N/\delta}{\epsilon^2}$ then with probability $1 - \delta$ for every w_i (uniformly) we have that*

$$|F(w_i) - \hat{F}(w_i)| \leq \epsilon.$$

Proof. Consider the event E_j defined as

$$E_j = \{\{z_1, \ldots, z_m\} : \left| \frac{1}{m} \sum_{i=1}^{m} f(w_j, z_i) - \mathbb{E}_{z \sim D} \left[f(w_j, z) \right] \right| > \epsilon\}$$

Note that the single event E_j requires us to bound the generalization bound of a parameter w_j that is independent of the sample. Thus, for each E_j Hoeffding's inequality (theorem 5.3) states that the probability of E_j is at most $e^{-2m\epsilon^2}$.

However, we would like to upper bound the probability of the event E defined as follows:

$$E = \bigcup_{j=1}^{|\mathcal{W}|} E_j$$

We now apply the union bound inequality: the union bound states that for any event A and a set of events A_i: if $A = \cup_{i=1}^{T} A_i$ then

$$P(A) = P(\cup A_i) \leq \sum P(A_i).$$

Using the union bound and Hoeffding we obtain

$$\begin{aligned}
\mathbb{P}(E) &= \mathbb{P}\left(\bigcup_{j=1}^{|\mathcal{W}|} E_j \right) \\
&\leq \sum_{j=1}^{|\mathcal{W}|} \mathbb{P}(E_j) && \text{(Union bound)} \\
&\leq \sum_{j=1}^{|\mathcal{W}|} e^{-2m\epsilon^2} && \text{(Hoeffding)} \\
&= |\mathcal{W}| e^{-2m\epsilon^2}
\end{aligned}$$

Thus if $m = \frac{1}{2\epsilon^2} \log\left(\frac{|\mathcal{W}|}{\delta}\right)$ then with probability $1-\delta$ for any $w_j \in \mathcal{W}$ it holds that

$$\left| \frac{1}{m} \sum_{i=1}^{m} f(w_j, z_i) - \mathbb{E}_{z\sim D}\left[f(w_j, z)\right] \right| \leq \epsilon$$

□

The second ingredient of our proof is to *discretize*: Our space of parameters is a set in $\mathbb{R}^n$ which is infinte. Nevertheless we will show that it can be approximated by choosing a finite subset that approximates any parameter in our space:

Lemma 6.3. *Let $\mathcal{W} = B(0, R)$. There exists a finite set $\overline{W}$ of the size*

$$\left|\overline{W}\right| := N(\epsilon) = O\left(\left(\frac{LR}{\epsilon}\right)^n\right),$$

such that: for any $w \in \mathcal{W}$ there exists a $\bar{w} \in \overline{W}$ such that if $f(w)$ is L-Lipchitz and bounded (w.l.o.g $-1 \leq f \leq 1$) function, then:

$$|f(w) - f(\bar{w})| \leq \epsilon$$

proof of lemma 6.3. We would like to build $\overline{W}$ such that for any $w \in \mathcal{W}$ there exists $\bar{w} \in \overline{W}$ that satisfies

$$\|w - \bar{w}\| \leq \frac{\epsilon}{L} \tag{6.1}$$

This would imply that if f is L-Lipschitz then

$$|f(w) - f(\bar{w})| \leq L\|w - \bar{w}\| \leq \epsilon$$

It is easy to confirm that the following $\overline{W}$ will work:

$$\overline{W} = \left\{ \left(\pm k_1 \frac{\epsilon}{\sqrt{n}L}, \pm k_2 \frac{\epsilon}{\sqrt{n}L}, \ldots, \pm k_n \frac{\epsilon}{\sqrt{n}L}\right) : k_i = 1, 2, \ldots, \frac{LR}{\epsilon} \right\}$$

and also $\left|\overline{W}\right| = \left(2 \cdot \frac{LR}{\epsilon}\right)^n$. □

Exercise 6.2. .

1. *Show that for any $w \in B(0,1)$ there exists $\bar{w} \in \overline{W}$ such that $\|w - \bar{w}\| \leq \frac{\epsilon}{L}$ where*

$$\overline{W} = \left\{ \left(\pm k_1 \frac{\epsilon}{\sqrt{n}L}, \pm k_2 \frac{\epsilon}{\sqrt{n}L}, \ldots, \pm k_n \frac{\epsilon}{\sqrt{n}L} \right) : k_i = 1, 2, \ldots, \frac{LR}{\epsilon} \right\}$$

2. *Show that $|\overline{W}| \leq \left(2\frac{LR}{\epsilon}\right)^n$*

proof of theorem 6.1. Using the same construction of $\overline{W}$ as in the proof of lemma 6.3 together with lemma 6.2 we can conclude that if $m = \tilde{O}\left(\frac{1}{\epsilon^2} \log\left(\frac{|\overline{W}|}{\delta}\right)\right)$ then for any $\bar{w} \in \overline{W}$ with probability $1 - \delta$ we obtain

$$\left|F(\bar{w}) - \hat{F}_m(\bar{w})\right| \leq \epsilon$$

plugging $|\overline{W}| = O\left(\left(\frac{LR}{\epsilon}\right)^n\right)$ yields

$$\frac{1}{\epsilon^2} \log\left(\frac{|\overline{W}|}{\delta}\right) = \frac{n \log\left(\frac{LR}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

which is indeed $m = \tilde{O}\left(\frac{n}{\epsilon^2} \log \frac{LR}{\delta}\right)$. The only missing piece left to prove is that $|\hat{F}_m(w) - F(w)|$ is bounded by $O(\epsilon)$ for any $w \in \mathcal{W}$. Suppose that the event $\left|F(\bar{w}) - \hat{F}_m(\bar{w})\right| \leq \epsilon$ occurred for any $\bar{w} \in \overline{W}$. In addition, using lemma 6.3 we obtain that for any $w \in \mathcal{W}$ there exists $\bar{w} \in \overline{W}$ such that for all z

$$\left|f(w, z) - f(\bar{w}, z)\right| \leq \epsilon$$

and therefore

$$\left|F(w) - F(\bar{w})\right| \leq \Big| = \left|\mathbb{E}_{z\sim D} f(w, z) - \mathbb{E}_{z\sim D} f(\bar{w}, z)\right| \leq \max_z \left|f(w, z) - f(\bar{w}, z)\right|\Big| \leq \epsilon$$

for the same reason we obtain that

$$\left|\hat{F}_m(w) - \hat{F}_m(\bar{w})\right| \leq \epsilon$$

With this in hand we can conclude our proof

$$\begin{aligned}\left|\hat{F}_m(w) - F(w)\right| &= \left|\hat{F}_m(w) - \hat{F}_m(\bar{w}) + \hat{F}_m(\bar{w}) - F(\bar{w}) + F(\bar{w}) - F(w)\right| \\ &\leq \left|\hat{F}_m(w) - \hat{F}_m(\bar{w})\right| + \left|\hat{F}_m(\bar{w}) - F(\bar{w})\right| + \left|F(\bar{w}) - F(w)\right| \\ &\leq 3\epsilon\end{aligned}$$

□

6.2 General Linear Models

We now want to move towards dimension independent bounds. We will later see that, in general, uniform convergence techniques inherently fail to achieve such bounds. But, for now we want to show how, under further structual assumptions on the learning problem, we can exploit this general technique to achieve improved rates. Therefore we now look at a special class of convex optimization problems called *General Linear Model*. In a general linear model we consider an optimization problem of the form

$$\text{minimize,} \quad \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(w \cdot x_i, y_i)\right],$$

where $\ell(a, y)$ is a convex scalar function over a and , $x \in \mathbb{R}^n$, and $y \in \mathbb{R}$. GLMs capture a wide class of learning problem, such as SVMs, Linear regression and more.

6.2.1 Example – Support Vector Machine

Consider a prediction problem where $(x, y) \sim D$, $x \in \mathbb{R}^n$ and $y \in \{-1, 1\}$, are samples from some unknown distribution, we have that $\|x\| \leq 1$ and we assume that there exists a linear classifier with *margin*. That means that there is a classifier w, $\|w\| = 1$ and w.p 1:

$$y \cdot (w \cdot x) \geq \gamma.$$

Alternatively, there exists a classifier with $\|w\| = 1/\gamma$ such that $y \cdot (w \cdot x) \geq 1$. This is illustrated in fig. 6.1.

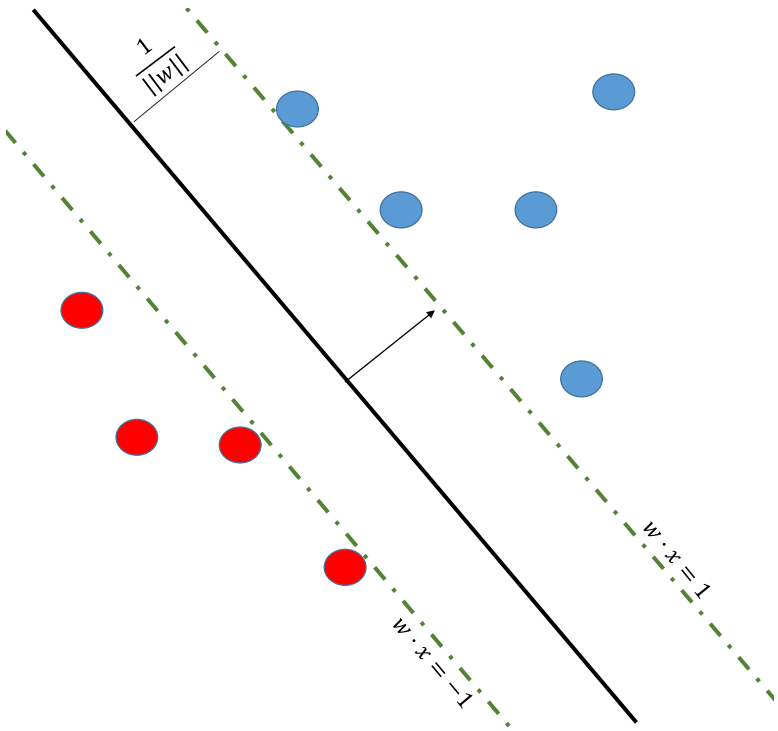

Figure 6.1: A linear separator with margin γ

That means that the sign of $\hat{w} \cdot x$ predicts y and is not sensitive to small perturbations. Geometrically, that means that the x's are distributed "away" from the decision boundary induced by w. One motivation to obtain dimension-independent bounds is to replace the dimension with the *margin*. Intuitively it makes sense that when the data is well-separated by a margin we expect improved rates of generalization. Note that we can find a linear classifier (over the empirical data) in the above case by considering the objective function:

$$\text{minimize } \frac{1}{m} \sum_{i=1}^{m} \ell(w \cdot x, y)$$
$$\text{s.t. } \|w\| \leq 1/\gamma$$

where $\ell(w \cdot x, y) = \max(0, 1 - yw \cdot x)$.

6.3 Rademacher Complexity

The Rademacher complexity is a tool to analyse the uniform convergence properties of a class of functions, by measuring how well we can fit an empirical sample.

The idea behind the Rademacher complexity is as follows: Suppose we are given a sample S and we divide it randomly to a train set of points S_{train}, and test set S_{test}: What could go wrong, if we train an algorithm by observing S_{train} and then testing it on S_{test}? Roughly, thing might go wrong, if for some function, the discrepency:

$$\sum_{\mathbf{z}\in S_{\text{train}}} h(w,\mathbf{z}) - \sum_{\mathbf{z}\in S_{\text{test}}} h(w,\mathbf{z})$$

is large.

This is exactly what the Rademacher complexity measures: We consider a random partition of a set into two distinct sets and we evaluate the discrepency w.r.t to the worst case function in $\mathcal{H}$.

For the discussion on Rademacher complexity it will be convenient to work with families of functions, instead of a parameter space $\mathcal{W}$: given a parameter space $\mathcal{W}$ we can think of it as inducing a family of function

$$\mathcal{H} = \{f : \mathcal{Z} \to \mathbb{R} | f = f(w,z), w \in \mathcal{W}\}.$$

Also, given a function h and a distribution D we denote

$$\mathcal{L}_D(h) = \mathop{\mathbb{E}}_{z\sim D}[h(z)],$$

and the given a sample S the empirical expected value:

$$\mathcal{L}_S(h) = \frac{1}{|S|}\sum_{i=1}^{m} h(z_i).$$

Finally, to define the Rademacher Complexity we begin with a simple definition of a Rademacher random variable:

Definition 6.4 (Rademacher random variables)**.** *Let σ be a vector whose elements are chosen independently and uniformly from $\{-1,+1\}$. That is, with probability $1/2$ a given element is either -1 or 1.*

Definition 6.5 (Empirical Rademacher complexity)**.** *Given a sample $S = \{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$ chosen from $\mathcal{D}^m$,*

define the empirical Rademacher complexity $\hat{\mathfrak{R}}_S(\mathcal{H})$ *as*

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(\mathbf{z}_i)\right]$$

Definition 6.6 (Rademacher complexity)**.** *For some* $m \geq 1$*, let the Rademacher complexity of* $\mathcal{H}$ *be the expectation of the empirical Rademacher over all samples* S *of size* m *drawn from some distribution* $\mathcal{D}$*.*

$$\mathfrak{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(\mathcal{H})\right]$$

Exercise 6.3. *Show that* $\hat{\mathfrak{R}}_S(\mathcal{H}) \geq 0$ *for every sample* S *and class of functions* $\mathcal{H}$*.*

Hint:

Use the fact that

$$\mathbb{E}[\sup_{h\in\mathcal{H}} h(X)] \geq \sup_{h\in\mathcal{H}} \mathbb{E}[h(X)].$$

Theorem 6.7. *Let* $\mathcal{H}$ *be a class of functions bounded by* c *and* $S = \{\mathbf{z}_i\}_{i=1}^m$ *a sample drawn IID then with probability* $1-\delta$

$$\sup_{h\in\mathcal{H}} \mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 2\mathfrak{R}_m(\mathcal{H}) + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right)$$

Proof. Let us write in shorthand $\Phi(S) = \sup_{h\in\mathcal{H}} \mathcal{L}_D(h) - \mathcal{L}_S(h)$.

For the proof we will utilize what is called a *ghost sample*: let us denote by S' a second sample that is drawn (i.i.d) from the distribution D of size m also (i.e. identical but independent from the sample S). Note that

$$\mathop{\mathbb{E}}_{S'\sim D^m} \frac{1}{m}\left[\sum_{i=1}^{m} h(z_i)\right] = \mathop{\mathbb{E}}_{z\sim D}\left[h(z)\right]. \tag{6.2}$$

We will first bound the expectation of $\Phi(S)$ then using a concentration inequality we will bound $\Phi(S)$ w.h.p:

$$\begin{aligned}
\mathbb{E}[\Phi(S)] &= \mathop{\mathbb{E}}_{S}[\sup_{h\in\mathcal{H}}\{\mathcal{L}_D(h) - \mathcal{L}_S(h)\}] && \text{by definition} \\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{h\in\mathcal{H}}\left\{\left[\mathop{\mathbb{E}}_{S'}\mathcal{L}_{S'}(h) - \mathcal{L}_S(h)\right]\right\}\right] && \text{eq. (6.2)} \\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{h\in\mathcal{H}}\{\mathcal{L}_{S'}(h) - \mathcal{L}_S(h)\}\right] \\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}[h(\mathbf{z}_i') - h(\mathbf{z}_i)]\right\}\right]
\end{aligned}$$

Now suppose we generate the sample S, S' as follows, after generating S, S' for every i with probability $1/2$ we set $\mathbf{z}_i' \in S'$ and $\mathbf{z}_i \in S$, but with probability $1/2$ we alternate and let $\mathbf{z}_i \in S'$ then clearly the sample S, S' are still generated by the same distribution. Therefore we have:

$$\begin{aligned}
&= \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[h(\mathbf{z}_i') - h(\mathbf{z}_i)\right]\right\}\right] && \sigma \text{ doesn't change } \mathbb{E} \\
&\leq \mathop{\mathbb{E}}_{S',\sigma}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[h(\mathbf{z}_i')\right]\right\}\right] + \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}-\sigma_i\left[h(\mathbf{z}_i)\right]\right\}\right] && \text{sub-additivity of supremum} \\
&= 2\mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[h(\mathbf{z}_i)\right]\right\}\right] && \sigma_i \text{ and } -\sigma_i \text{ distributed same way} \\
&= 2\mathfrak{R}_m(\mathcal{H}) \ \square
\end{aligned}$$

So far we've bounded the generalization error in expectation, to derive high probability bounds we rely on the following inequality

Lemma 6.8 (**Mcdiarmid Inequality**). *: Let X be some set and $f : X^m \to \mathbb{R}$ a function of m variables such that for some $a > 0$, for all $i \in [m]$ and for all $x_1, \ldots, x_m, x_i' \in X$ we have*

$$|h(x_1, \ldots, x_i, \ldots, x_m) - h(x_1, \ldots, x_i', \ldots, x_m)| \leq a.$$

Let $X_1, \ldots, X_m$ be IID random varaibles taking value at X then w.p at least $1 - \delta$:

$$|h(X_1, \ldots, X_m) - \mathbb{E}\left[h(X_1, \ldots, X_m)\right]| \leq a\sqrt{\ln(2/\delta)m/2}$$

To apply Mcdiarmind's inequality note that if $S, \bar{S}$ are two samples that differ on a single example, then $\Phi(S) - \Phi(\bar{S}) \leq \frac{2c}{m}$, hence we obtain the w.p at least $1 - \delta$

$$\sup_{h \in \mathcal{H}} \mathcal{L}_D(h) - L_S(h) \leq \mathbb{E}\left[\Phi(S)\right] + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right) \leq 2\mathfrak{R}_m(\mathcal{H}) + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right)$$

□

6.3.1 Rademacher Calculus

The following facts are easy to prove, and we leave them as an exercise:

Exercise 6.4.

1. *Let* $c \cdot \mathcal{H} + b = \{c \cdot h + b : h \in \mathcal{H}\}$. *Then*

$$\mathfrak{R}_m(\mathcal{H})(c \cdot \mathcal{H} + b) = |c|\mathfrak{R}_m(\mathcal{H})$$

2. *Let* $\text{conv}\mathcal{H} = \{\sum_{i=1}^k \alpha_i h_i : \ k \geq 1, \{h_i\} \subseteq \mathcal{H} \ \alpha_i \geq 0, \ \sum_{i=1}^k \alpha_i = 1\}$. *Then*

$$\mathfrak{R}_m(\text{conv}\mathcal{H}) = \mathfrak{R}_m(\mathcal{H})$$

The proof of the following Lemma can be found in the lecture notes:

Lemma 6.9. *Let* $\phi_{\mathbf{z}}$ *be* ρ*-Lipschitz functions for every* $\mathbf{z} \in \mathcal{X}$ *(i.e.* $|\phi_{\mathbf{z}}(a) - \phi_{\mathbf{z}}(b)| \leq \rho|a - b|$*). Denote*

$$\phi \circ \mathcal{H} = \{\phi_{\mathbf{z}}(h(\mathbf{z})) : h \in \mathcal{H}\}.$$

Then

$$\mathfrak{R}_m(\phi \circ \mathcal{H}) \leq \rho\mathfrak{R}_m(\mathcal{H})$$

Proof. w.l.o.g we may assume $\rho = 1$, the more general case will follow by setting $\phi'_{\mathbf{z}} = \frac{1}{\rho}\phi$ and applying

property 1. Given $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, Let us write

$$\hat{\mathfrak{R}}_S(\phi_{[t]} \circ \mathcal{H}) = \mathop{\mathbb{E}}_{\sigma}\left[\left|\sup_{h\in\mathcal{H}} \sum_{i\le t} \sigma_i \phi_{\mathbf{z}_i}(h(\mathbf{z}_i)) + \sum_{i>t} \sigma_i h(\mathbf{z}_i)\right|\right]$$

We will show by induction that $\hat{\mathfrak{R}}_S(\phi_{[t]} \circ \mathcal{H}) \le \hat{\mathfrak{R}}_S(\mathcal{H})$. Since the case for $t = 1$ is similar to the induction step, we will show the proof only for $t = 1$. Let us write $\pm\mathcal{H} = \{\pm h : h \in \mathcal{H}\}$ then:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma}\left[\left|\sup_{h\in\mathcal{H}} \sigma_1 \phi_{\mathbf{z}_1}(h(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h(\mathbf{z}_i)\right|\right] \\
=&\frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\left|\sup_{h\in\mathcal{H}} \phi_{\mathbf{z}_1}(h(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h(\mathbf{z}_i)\right|\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\left|\sup_{h'\in\mathcal{H}} -\phi_{\mathbf{z}_1}(h'(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right|\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\left|\sup_{h\in\mathcal{H}} \phi_{\mathbf{z}_1}(h(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h(\mathbf{z}_i)\right| + \left|\sup_{h'\in\mathcal{H}} -\phi_{\mathbf{z}_1}(h'(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right|\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\pm\mathcal{H}}\left[\phi_{\mathbf{z}_1}(h(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h(\mathbf{z}_i)\right] + \sup_{h'\in\pm\mathcal{H}}\left[-\phi_{\mathbf{z}_1}(h'(\mathbf{z}_1)) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right]\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h,h'\in\pm\mathcal{H}} (\phi_{\mathbf{z}_1}(h(\mathbf{z}_1)) - \phi_{\mathbf{z}_1}(h'(\mathbf{z}_1))) + \sum_{i>1} \sigma_i h(\mathbf{z}_i) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right] \\
&\le \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h,h'\in\pm\mathcal{H}} |h(\mathbf{z}_1) - h'(\mathbf{z}_1)| + \sum_{i>1} \sigma_i h(\mathbf{z}_i) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right]
\end{aligned}$$

Next, we claim that we can remove the absolute value from the term $|h(\mathbf{z}_1) - h'(\mathbf{z}_1)|$, since we are taking supremum over the terms, and the sum of the two other terms will not be effected by replacing h' with h, hence:

$$\begin{aligned}
&\frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h,h'\in\pm\mathcal{H}} |h(\mathbf{z}_1) - h'(\mathbf{z}_1)| + \sum_{i>1} \sigma_i h(\mathbf{z}_i) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h,h'\in\pm\mathcal{H}} h(\mathbf{z}_1) - h'(\mathbf{z}_1) + \sum_{i>1} \sigma_i h(\mathbf{z}_i) + \sum_{i>1} \sigma_i h'(\mathbf{z}_i)\right] = \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}} \sigma_1 h(\mathbf{z}_1) + \sum_{i>1} \sigma_i h(\mathbf{z}_i)\right] = \hat{\mathfrak{R}}_S(\mathcal{H})
\end{aligned}$$

$\square$

6.4 Norm-dependent Generalization Bounds

We now turn to obtain generalization bounds that depend on the *norm* of the classifier instead of the dimension.

Theorem 6.10. *Suppose D is a distribution over $Z = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ such that with probability 1 $\|x\| \leq R$. Let $\mathcal{W}_B = \{w : \|w\| \leq B\}$ and let $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a convex L-Lipschitz function bounded by c over $[-BR, BR] \times \mathcal{Y}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of an i.i.d sample of size m,*

$$\forall w \in \mathcal{W}_B, \quad \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [\ell(w \cdot x, y)] - \frac{1}{m} \sum_{i=1}^{m} \ell(w \cdot x_i, y_i) \right| \leq \frac{2LBR}{\sqrt{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

For the proof, we will need the following Lemma:

Lemma 6.11. *Let H be a linear space (for concreteness assume $H = \mathbb{R}^d$) and define $\mathcal{H} = \{h_w(x) : h_w(x) = w^\top x, \|w\| \leq B\}$ Then*

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\max_i \|x_i\|_2}{\sqrt{m}}$$

Proof. We will assume that $B = 1$, the general case then follows from item 1 in Exercise 6.4.

$$
\begin{aligned}
m\hat{\mathfrak{R}}_S(\mathcal{H})(\mathcal{H}) &= \mathbb{E}\left[\sup_{\|w\|\le 1} \sum_{i=1}^m \sigma_i w^\top x_i\right] \\
&= \mathbb{E}\left[\sup_{\|w\|\le 1} w^\top \left(\sum_{i=1}^m \sigma_i x_i\right)\right] \\
&= \mathbb{E}\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|\right] = \mathbb{E}\left[\sqrt{\left\|\sum_{i=1}^m \sigma_i x_i\right\|^2}\right] \\
&\le \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|^2\right]} && \text{Concavity of } \sqrt{} \\
&= \sqrt{\mathbb{E}\left[\sum_{i,j=1}^m \sigma_i \sigma_j x_i^\top (x_j)\right]} \\
&= \sqrt{\mathbb{E}\left[\sum_{i=1}^m \|x_i\|^2\right]} && \mathbb{E}(\sigma_i\sigma_j) = \begin{cases} 1 & i = j \\ 0 & \text{o.w.} \end{cases} \\
&\le \sqrt{m \max \|x_i\|^2}
\end{aligned}
$$

□

Proof of theorem 6.12. Let $\mathcal{H}$ be the class of functions over $\mathbf{z} = (x, y)$ of the form

$$h_w(\mathbf{z}) = \ell(w \cdot x, y).$$

Then note that theorem 6.12 follows directly from theorem 6.7 if we show that

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \le \frac{LBR}{\sqrt{m}}.$$

If $\mathcal{H}_B = \{h_w(\mathbf{z}) = w \cdot x : \|w\| \le B\}$, and denoting $\phi_{\mathbf{z}}(a) = \ell(a, y)$, we have that

$$\mathcal{H} = \phi \circ \mathcal{H}_B,$$

hence by lemmas 6.9 and 6.11 we have that

$$\mathfrak{R}_m(\mathcal{H}) = \mathfrak{R}_m(\phi \circ \mathcal{H}_B) \leq L\mathfrak{R}_m(\mathcal{H}_B) \leq \frac{RLB}{\sqrt{m}}.$$

□

Soft constraints We often do not want to restrict the class of predictors apriori, but to consider some tradeoff between regularization and fitting. Consider the following objective

$$\text{minimize, } \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1} \ell(w \cdot x_i, y_i) \tag{6.3}$$

Exercise 6.5. *Let ℓ be an 1-Lipschitz convex function. Assume also that for any y $\ell(0, y) \leq 1$. Show that the solution, $\hat{w}$, to the regularized objective in eq.* (6.3) *yields (w.p $1 - \delta$)*

$$F(\hat{w}) \leq F(w^\star) + \frac{\lambda}{2}B^2 + O\left(\frac{\log 1/\delta}{\sqrt{\lambda m}}\right).$$

Conclude that for choice $\lambda = O(1/B\sqrt[4]{m})$

$$F(\hat{w}) \leq \min_{\|w^\star\| \leq B} F(w^\star) + O\left(\frac{B \log 1/\delta}{\sqrt[4]{m}}\right).$$

Remark The above bound is far from optimal. In [2] the authors showed that the empirical risk minimizer of the regularized objectives in fact enjoys an error bound:

$$F(\hat{w}) \leq F(w^\star) + O\left(\frac{L^2B^2}{\lambda m}\right),$$

where $w^\star$ is the minimizer of the regularized objective. This leads to the optimal error rate of $O(1/\sqrt{m})$ (neglecting dependence in Lipschitness and radius of the ball) with a choice of $\lambda = \frac{1}{\sqrt{n}}$. The result relies on a technique called *localized Rademacher complexity.*

6.4.1 Algorithmic bias

Consider the unregularized objective

$$\text{minimize, } \frac{1}{m}\sum_{i=1}^{m}(w \cdot x_i - y_i)^2. \tag{6.4}$$

We will focus on the case that $m \ll d$, hence the above problem is ill-posed and there are more than one solution. It can also be seen that various solutions might minimize the empirical error and can fail to minimize the true expected risk.

Recall though that GD algorithm work by iteratively considering

$$w_t - \eta \nabla F_m(w_t) = w_t - \eta \frac{2}{m}\sum_{i=1}^{m}(w \cdot x_i - y_i)x_i.$$

Exercise 6.6. *Show that the solution to eq.* (6.4)*,* $\hat{w}$*, that lies in the span* $\text{span}(\{x_i\})$ *is the minimal norm solution. Namely, for any other solution* $\bar{w}$*, we have that* $\|\bar{w}\| \geq \|\hat{w}\|$*.*

As a result, by solving eq. (6.4) using GD we obtain a regularized solution even though we apriori considered unconstrained and unregularized objective. So if there exists a solution w^* with 0 error on eq. (6.4) with constant bounded norm (say 1 for simplicity), GD is guaranteed to converge to a generalizing solution using $O\left(\frac{\log 1/\delta}{\epsilon^2}\right)$ examples.

6.4.2 Other norm constraints

We have solely discussed ℓ_2-regularization, other norms can also induce generalization bounds. For example, in theorem 6.12 ℓ_2 can be replaced by ℓ_1 to yield the following guarantee:

Theorem 6.12. *[can be found in [1]] Suppose* D *is a distribution over* $Z = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^n \times \mathbb{R}$ *such that with probability* 1 $\|x\| \leq R$*. Let* $\mathcal{W}_B = \{w : \|w\|_1 \leq B\}$ *and let* $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ *be a convex L-Lipschitz function bounded by* c *over* $[-BR, BR] \times \mathcal{Y}$*. Then for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *over the choice on an i.i.d sample of size* m*,*

$$\forall w \in \mathcal{W}_B,\ |F(w) - \hat{F}_m(w)| \leq 2LBR\sqrt{\frac{\log n}{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}},$$

where

$$F(w) = \mathop{\mathbb{E}}_{z\sim D}[h(w,z)] = \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(w\cdot x, y)]\,,$$

and

$$\hat{F}_m(w) = \frac{1}{m}\sum_{i=1}^{m} h(w,z_i) = \frac{1}{m}\sum_{i=1}^{m} \ell(w\cdot x_i, y_i).$$

References

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

[2] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. *Advances in neural information processing systems*, 21:1545–1552, 2008.