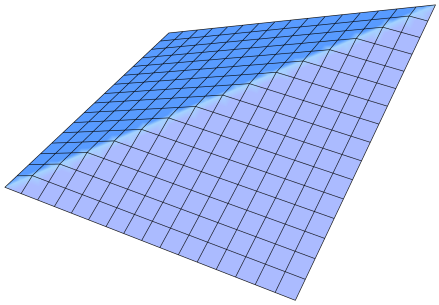

STA 3000Y: Advanced Theory of Statistics (I)

Professor Daniel Roy • Fall 2018 • University of Toronto

T 2:00-5:00PM in UC 53

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Tuesday, September 11, 2018

1. Complexity, Concentration, and Learnability

Risk and PAC Learnability

Parametric statistics begins with the assumption that the data arise from a specified parametric family. Here in **statistical learning theory**, by contrast, the distribution is not assumed to have a finite-dimensional parameter. We retain only the strong but useful assumption that repeated observations are independent and identically distributed. Under mild conditions, this assumption

makes population quantities estimable from empirical averages through laws of large numbers and central limit theorems. The two recurring phenomena will be **concentration of measure**, whereby an empirical quantity is close to its expectation, and **uniform convergence**, whereby this approximation holds simultaneously over a class of candidate predictors.

The same machinery underlies the consistency of M-estimators and many semiparametric procedures. Interpreting independent and identically distributed sampling as a promise that present observations are relevant to future observations also clarifies what is lost when that assumption is removed. Without it, absolute guarantees about future performance are generally unavailable. Online learning replaces them with relative guarantees: a strategy is successful when its cumulative performance is not much worse than that of an appropriate benchmark.

Let $Z_1, \dots, Z_m$ be independent observations with common but unknown distribution D, and write $Z_i = (X_i, Y_i)$. The variable $X_i \in \mathcal{X}$ is an **example** or **instance** and $Y_i \in \mathcal{Y}$ is its **label** or **class** or **response**. Initially, take $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, 1\}$. For example, X_i might be an image of tissue and Y_i the indicator that a physician diagnoses the tissue as diseased. The corresponding empirical measure is

$$D_m \coloneqq \frac{1}{m} \sum_{i=1}^{m} \delta_{Z_i},$$

where δ_{Z_i} denotes the **Dirac probability measure** at Z_i. Because its support points are random, D_m is a random probability measure. Although D is unknown, D_m can be used to estimate quantities defined in terms of D.

Definition 1.1 A **hypothesis class** $\mathcal{H}$ is a set of functions from $\mathcal{X}$ to $\mathcal{Y}$. Its elements are called **hypotheses**, or **predictors**.

A hypothesis $h \in \mathcal{H}$ predicts the label Y of an instance $X \in \mathcal{X}$ by $h(X)$. The goal of learning is to select a hypothesis with good predictive performance. For a hypothesis $h \in \mathcal{H}$, define the **zero-one loss** by

$$\ell_{0,1}(h(x), y) \coloneqq \begin{cases} 1, & h(x) \neq y, \\ 0, & h(x) = y. \end{cases}$$

The associated **risk**, or **error**, is

$$L_D(h) \coloneqq D\big(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : h(x) \neq y\}\big) = \int_{\mathcal{X} \times \mathcal{Y}} \ell_{0,1}(h(x), y) \, \mathrm{d}D(x, y).$$

We wish to find $h^* \in \mathcal{H}$ such that $L_D(h^*)$ is as small as possible — that is, $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_D(h)$. However, this goal is usually too ambitious because D is unknown.

Note that if

$$f_h(x, y) \coloneqq \ell_{0,1}(h(x), y) \qquad \text{and} \qquad \mathcal{F} \coloneqq \{f_h : h \in \mathcal{H}\},$$

then the same quantity can be written as

$$L_D(h) = \int f_h \, \mathrm{d}D = D(f_h).$$

Thus D may be viewed either as a probability measure on the underlying measurable space or as the expectation functional $f \mapsto D(f)$. The set $\mathcal{F}$ is the **loss class** induced by $\mathcal{H}$. The corresponding **empirical risk**, or **training error**, is

$$L_S(h) = D_m(f_h) = \frac{1}{m} \sum_{i=1}^{m} f_h(X_i, Y_i) = \frac{|\{i \in [m] : h(X_i) \neq Y_i\}|}{m},$$

where $[m] \coloneqq \{1, \ldots, m\}$. This follows from the identity $\int f \, \mathrm{d}\delta_z = f(z)$ and the linearity of integration.

Exercise 1.2 Show that

$$\mathbb{E}_{S \sim D^m}[D_m(f_h)] = D(f_h),$$

or equivalently that $\mathbb{E}_{S \sim D^m}[L_S(h)] = L_D(h)$. Thus empirical risk is an unbiased estimate of the risk of any fixed hypothesis h.

Solution. By linearity of expectation and the identical distribution of the observations,

$$\mathbb{E}_{S \sim D^m}[D_m(f_h)] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[f_h(Z_i)] = \frac{1}{m} \sum_{i=1}^{m} D(f_h) = D(f_h).$$

Since $D_m(f_h) = L_S(h)$ and $D(f_h) = L_D(h)$, the equivalent risk identity follows. □

For a fixed h, the strong law of large numbers and the central limit theorem give

$$D_m(f_h) \xrightarrow{\text{a.s.}} D(f_h) \qquad \text{and} \qquad \sqrt{m}\big(D_m(f_h) - D(f_h)\big) \xrightarrow{d} \mathcal{N}\big(0, D(f_h)(1 - D(f_h))\big).$$

The difficulty is that a learning algorithm does not evaluate a fixed h; it chooses h only *after* seeing the data. To see what can go wrong, let $X \sim \mathrm{Unif}[0, 1]$, set

$$Y = \mathbb{1}(X \geq 1/2),$$

and take $\mathcal{H}$ to be the class of all functions from $[0, 1]$ to $\{0, 1\}$. Thus the distribution has a perfect

classifier, but the feature distribution has no atoms. Define the **memorization rule**

$$h_S(x) = \mathbb{1}\big(\text{there is an } i \in [m] \text{ such that } X_i = x \text{ and } Y_i = 1\big).$$

For each realized sample, h_S is an element of $\mathcal{H}$, but h_S itself is random because it depends on S. More precisely, the **learning rule** maps samples to functions in $\mathcal{H}$. By construction,

$$L_S(h_S) = D_m(f_{h_S}) = 0 \quad \text{almost surely.}$$

Nevertheless, the rule merely memorizes the observed instances. An independent test feature is almost surely different from every training feature, so h_S predicts zero on it. Consequently,

$$\mathbb{E}_{S\sim D^m}[L_S(h_S)] = 0 \qquad \text{and} \qquad \mathbb{E}_{S\sim D^m}[L_D(h_S)] = \frac{1}{2}.$$

The empirical minimizer can therefore have chance-level predictive performance even when its training error is zero.

Exact minimization of L_D is also too ambitious as a statistical objective. The minimizer is a functional of the unknown distribution, and finite samples cannot reliably distinguish arbitrarily similar distributions. For example, one thousand tosses may not distinguish a coin with success probability $3/4$ from one with success probability $3/4 + 10^{-6}$. The usual relaxation is to require a predictor $\widehat{h}$ satisfying

$$L_D(\widehat{h}) \le \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon$$

with high probability.

Definition 1.3 A class $\mathcal{H}$ is **agnostically probably approximately correct learnable**, or **agnostically PAC learnable**, if there are a **sample complexity**

$$m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$$

and learning rules $A_m : \mathcal{Z}^m \times (0,1)^2 \to \mathcal{H}$ such that the following holds: for every distribution D, every $\varepsilon, \delta \in (0,1)$, and every $m \ge m_{\mathcal{H}}(\varepsilon, \delta)$,

$$\mathbb{P}_{S\sim D^m}\left(L_D(A_m(S,\varepsilon,\delta)) \le \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon\right) \ge 1 - \delta.$$

The value $m_{\mathcal{H}}(\varepsilon, \delta)$ is the number of observations sufficient, uniformly over D, to obtain risk within ε of the best risk in $\mathcal{H}$ with confidence at least $1 - \delta$. In particular, if $\mathcal{X}$ is uncountable, the class of all functions from $\mathcal{X}$ to $\mathcal{Y}$ is not PAC learnable (we will see this later).

Definition 1.4 A distribution D on $\mathcal{X} \times \mathcal{Y}$ is **realizable by $\mathcal{H}$** if there is a hypothesis $h^* \in \mathcal{H}$ such that

$$D\big(\{(x, y) : h^*(x) = y\}\big) = 1.$$

A class $\mathcal{H}$ is **probably approximately correct learnable**, or **PAC learnable**, if there are a sample complexity function

$$m_{\mathcal{H}}^{\mathrm{PAC}} \colon (0,1)^2 \to \mathbb{N}$$

and learning rules $A_m \colon \mathcal{Z}^m \times (0,1)^2 \to \mathcal{H}$ such that, for every distribution D realizable by $\mathcal{H}$, every $\varepsilon, \delta \in (0,1)$, and every $m \geq m_{\mathcal{H}}^{\mathrm{PAC}}(\varepsilon, \delta)$,

$$\mathbb{P}_{S \sim D^m}\left(L_D(A_m(S, \varepsilon, \delta)) \leq \varepsilon\right) \geq 1 - \delta.$$

Definition 1.4 is the realizable specialization of Definition 1.3, because the best risk in $\mathcal{H}$ is zero. Dependence of the learning rule on (ε, δ) can be removed, although that equivalence is not immediate. More deeply, for binary classification the classes that are PAC learnable are exactly those that are agnostically PAC learnable.

Theorem 1.5 Every finite hypothesis class $\mathcal{H}$ is PAC learnable in the realizable setting. In particular, empirical risk minimization succeeds whenever

$$m \geq \frac{\log|\mathcal{H}| + \log(1/\delta)}{\varepsilon}.$$

Proof. Because the problem is realizable, at least one hypothesis has empirical risk zero. Fix a bad hypothesis $h' \in \mathcal{H}$ with $L_D(h') > \varepsilon$. The probability that h' makes no error on m independent observations is

$$\mathbb{P}_{S \sim D^m}(L_S(h') = 0) = D^m\left(\{S : L_S(h') = 0\}\right) = (1 - L_D(h'))^m < (1 - \varepsilon)^m.$$

The bad event is that some hypothesis with risk greater than ε nevertheless has empirical risk zero. By the union bound,

$$\begin{aligned}
&\mathbb{P}_{S \sim D^m}\big(\text{there is an } h \in \mathcal{H} \text{ with } L_S(h) = 0 \text{ and } L_D(h) > \varepsilon\big) \\
&= D^m\left(\{S : \exists h \in \mathcal{H} \text{ with } L_S(h) = 0 \text{ and } L_D(h) > \varepsilon\}\right) \\
&= D^m\left(\bigcup_{h : L_D(h) > \varepsilon} \{S : L_S(h) = 0\}\right)
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{h:L_D(h)>\varepsilon} D^m\left(\{S : L_S(h) = 0\}\right) \\
&\leq |\mathcal{H}|(1-\varepsilon)^m \\
&\leq |\mathcal{H}|e^{-m\varepsilon}.
\end{aligned}$$

The final expression is at most δ when

$$m \geq \frac{\log|\mathcal{H}| + \log(1/\delta)}{\varepsilon}.$$

Consequently, with probability at least $1-\delta$, every hypothesis of empirical risk zero has risk at most ε. Any empirical risk minimizer therefore satisfies the required guarantee. □

An **empirical risk minimizer** is any learning rule satisfying

$$A(S) \in \operatorname*{argmin}_{h\in\mathcal{H}} D_m(f_h) = \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h).$$

The sample bound in Theorem 1.5 shows that m must be at least commensurate with $\log|\mathcal{H}|$. For an infinite class, cardinality is not the right measure of size; Vapnik–Chervonenkis dimension will eventually replace $\log|\mathcal{H}|$.

A useful sufficient condition for empirical risk minimization is **uniform convergence**. For $\varepsilon, \delta \in (0,1)$, suppose that

$$\mathbb{P}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}} |L_S(h) - L_D(h)| > \varepsilon\right) < \delta. \tag{1.1}$$

The centered family

$$\{L_S(h) - L_D(h) : h \in \mathcal{H}\}$$

is an **empirical process**, and (1.1) controls its supremum.

Proposition 1.6 Assume that the event

$$\sup_{h\in\mathcal{H}} |L_S(h) - L_D(h)| < \varepsilon$$

holds. If h_S^* minimizes L_S over $\mathcal{H}$ and h^* minimizes L_D over $\mathcal{H}$, then

$$L_D(h_S^*) \leq L_D(h^*) + 2\varepsilon.$$

Proof. Insert the empirical risks of h_S^* and h^* to obtain

$$L_D(h_S^*) - L_D(h^*) = \underbrace{L_D(h_S^*) - L_S(h_S^*)}_{<\varepsilon} + \underbrace{L_S(h_S^*) - L_S(h^*)}_{\leq 0} + \underbrace{L_S(h^*) - L_D(h^*)}_{<\varepsilon} \leq 2\varepsilon.$$

□

Thus uniform convergence yields uniform control of empirical risk. For binary classification, the fundamental equivalence will connect PAC learnability, uniform convergence, finite Vapnik–Chervonenkis dimension, and agnostic PAC learnability.

Lecture 2 — Tuesday, September 18, 2018

Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a general loss function, and let $\mathbb{P}$ be a distribution on $\mathcal{Z}$. The loss class induced by $\mathcal{H}$ is

$$\mathcal{F} := \ell \circ \mathcal{H} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}.$$

In binary classification, for instance, ℓ may be the zero-one loss. The risk of $f \in \mathcal{F}$ is $\mathbb{P}(f)$ and its empirical risk is $\mathbb{P}_m(f)$. If f_S minimizes $\mathbb{P}_m$ over $\mathcal{F}$ and f^* minimizes $\mathbb{P}$ over $\mathcal{F}$, then uniform convergence gives

$$\mathbb{P}(f_S) - \mathbb{P}(f^*) = \mathbb{P}(f_S) - \mathbb{P}_m(f_S) + \mathbb{P}_m(f_S) - \mathbb{P}_m(f^*) + \mathbb{P}_m(f^*) - \mathbb{P}(f^*) \leq 2 \sup_{f \in \mathcal{F}} |\mathbb{P}(f) - \mathbb{P}_m(f)|.$$

This is the same excess risk argument as Proposition 1.6. The remaining task is to control the supremum. Rademacher complexity provides one way to quantify its size.

Rademacher Complexity

Fix a class $\mathcal{G}$ of functions from $\mathcal{Z}$ to $[a, b]$ and a sample $S = (z_1, \ldots, z_m)$. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ have independent coordinates, each uniformly distributed on $\{-1, 1\}$ (i.e., each σ_i has a **Rademacher distribution**).

Definition 2.1 The **empirical Rademacher complexity** of $\mathcal{G}$ on S is

$$\widehat{R}_S(\mathcal{G}) \coloneqq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\right].$$

If $S \sim D^m$, the **Rademacher complexity** of $\mathcal{G}$ at sample size m is

$$R_m(\mathcal{G}) \coloneqq \mathbb{E}_{S\sim D^m}\left[\widehat{R}_S(\mathcal{G})\right].$$

For $\boldsymbol{g}(S) \coloneqq (g(z_1), \ldots, g(z_m))$, the random signed average can be written as

$$\frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i) = \frac{1}{m}\langle \boldsymbol{\sigma}, \boldsymbol{g}(S)\rangle.$$

The supremum measures how well the class can correlate with independent random signs. A very flexible class can fit this artificial noise, whereas a less flexible class cannot.

Define the **one-sided uniform deviation**

$$\Phi(S) \coloneqq \sup_{g\in\mathcal{G}} \Big(\mathbb{E}_{z\sim D}[g(z)] - \widehat{\mathbb{E}}_S[g]\Big) \qquad \text{and} \qquad \widehat{\mathbb{E}}_S[g] \coloneqq \frac{1}{m}\sum_{i=1}^{m} g(z_i).$$

For each fixed g, the difference inside the supremum is a mean-zero random variable. The supremum need not have mean zero, and its expectation is controlled by symmetrization.

Proposition 2.2 For the class $\mathcal{G}$ above,

$$\mathbb{E}_{S\sim D^m}[\Phi(S)] \le 2R_m(\mathcal{G}).$$

Proof. Let $S' = (z'_1, \ldots, z'_m)$ be an independent sample from D^m. Because

$$\mathbb{E}_{z\sim D}[g(z)] = \mathbb{E}_{S'}\left[\widehat{\mathbb{E}}_{S'}[g]\right],$$

Jensen's inequality and the convexity of the supremum give

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \mathbb{E}_{S'}\left[\widehat{\mathbb{E}}_{S'}[g] - \widehat{\mathbb{E}}_S[g]\right]\right] \le \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\big(g(z'_i) - g(z_i)\big)\right].$$

Each difference $g(z'_i) - g(z_i)$ is symmetric, so multiplying it by an independent random sign does

not change its distribution. Therefore,

$$\begin{aligned}\mathbb{E}_S[\Phi(S)] &\le \mathbb{E}_{S,S',\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(g(z_i')-g(z_i)\big)\right]\\ &\le \mathbb{E}_{S',\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i')\right]+\mathbb{E}_{S,\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}(-\sigma_i)g(z_i)\right]\\ &= 2R_m(\mathcal{G}),\end{aligned}$$

since $-\sigma$ has the same distribution as σ. □

Concentration of Measure

Lemma 2.3 (McDiarmid's inequality) Let $X_1,\dots,X_m$ be independent random variables taking values in a set $\mathcal{V}$, and let $f:\mathcal{V}^m\to\mathbb{R}$. Suppose that changing only coordinate i changes the value of f by at most c_i. Then, with probability at least $1-\delta$,

$$|f(X_1,\dots,X_m)-\mathbb{E}[f(X_1,\dots,X_m)]| \le \sqrt{\frac{\log(2/\delta)}{2}\sum_{i=1}^{m}c_i^2}.$$

In particular, if every c_i is equal to c, the right-hand side is

$$c\sqrt{\frac{m\log(2/\delta)}{2}}.$$

Changing one observation in S changes $\Phi(S)$ by at most $(b-a)/m$. Applying Lemma 2.3 with $c_i=(b-a)/m$, and then applying Proposition 2.2, shows that with probability at least $1-\delta$,

$$\Phi(S)\le\mathbb{E}[\Phi(S)]+(b-a)\sqrt{\frac{\log(2/\delta)}{2m}}\le 2R_m(\mathcal{G})+(b-a)\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{2.1}$$

In particular, if g_S is selected by empirical risk minimization, then

$$\mathbb{E}[g_S]-\widehat{\mathbb{E}}_S[g_S]\le\Phi(S).$$

Thus (2.1) bounds the risk of the selected function by its empirical risk plus a complexity penalty.

Corollary 2.4 Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $[a,b]$, let $S \sim D^m$, and let $\delta \in (0,1)$. The following statements hold:

1. With probability at least $1-\delta$, every $g \in \mathcal{G}$ satisfies
$$\mathbb{E}[g] \le \widehat{\mathbb{E}}_S[g] + 2R_m(\mathcal{G}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2m}}.$$
2. With probability at least $1-2\delta$, every $g \in \mathcal{G}$ satisfies
$$\mathbb{E}[g] \le \widehat{\mathbb{E}}_S[g] + 2\widehat{R}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Proof. (1) is exactly (2.1), applied simultaneously to the class.

For (2), apply Lemma 2.3 to the map $S \mapsto \widehat{R}_S(\mathcal{G})$. Replacing one observation changes this quantity by at most $(b-a)/m$, so with probability at least $1-\delta$,
$$R_m(\mathcal{G}) \le \widehat{R}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2m}}.$$
Combining this event with (1) and using the union bound gives (2). □

The empirical bound is especially useful because $\widehat{R}_S(\mathcal{G})$ depends only on the observed sample. A successful class should fit independent random signs poorly, and hence should have small empirical Rademacher complexity. For a $\{-1,1\}$-valued hypothesis class $\mathcal{H}$, with labels in $\{-1,1\}$ and zero-one loss $\ell_{0,1}(h,(x,y)) = (1-yh(x))/2$, write $S_{\mathcal{X}} = (x_1,\dots,x_m)$ for the input portion of the labeled sample S. Sign symmetry gives
$$\widehat{R}_S(\ell_{0,1} \circ \mathcal{H}) = \frac{1}{2}\widehat{R}_{S_{\mathcal{X}}}(\mathcal{H}).$$
Computing this quantity directly can still be difficult, so the next step is to control it through a combinatorial measure of the size of the class.

Definition 2.5 For a class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$, the **growth function** is
$$\Pi_{\mathcal{G}}(m) \coloneqq \max_{x_1,\dots,x_m \in \mathcal{X}} |\{(g(x_1),\dots,g(x_m)) : g \in \mathcal{G}\}|, \quad m \in \mathbb{N}.$$

For any set of m instances, the class produces a collection of vectors of function values. The growth function is the largest possible number of distinct such vectors. In binary classification these vectors are labelings. If $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$ and $\Pi_{\mathcal{H}}(m) = 2^m$, then some set of m instances realizes every binary labeling, and the empirical Rademacher complexity on that set is 1.

Lemma 2.6 (Massart's lemma) Let $A \subseteq \mathbb{R}^m$ be nonempty and finite, and let

$$r := \max_{\boldsymbol{x} \in A} \|\boldsymbol{x}\|_2.$$

Then

$$\mathbb{E}_\sigma \left[\sup_{\boldsymbol{x} \in A} \frac{1}{m} \langle \boldsymbol{\sigma}, \boldsymbol{x} \rangle \right] \leq \frac{r\sqrt{2 \log |A|}}{m}.$$

A detailed proof of Lemma 2.6 will be given later. For now, apply the result to

$$\mathcal{G}|_S := \{(g(x_1), \ldots, g(x_m)) : g \in \mathcal{G}\}.$$

If $\mathcal{G}$ is binary valued, every vector in $\mathcal{G}|_S$ has Euclidean norm $\sqrt{m}$.

Corollary 2.7 Let $\mathcal{G}$ be a nonempty class of functions satisfying $|g(x)| \leq 1$ for every $g \in \mathcal{G}$ and $x \in \mathcal{X}$, and suppose $\Pi_{\mathcal{G}}(m) < \infty$. Then

$$\widehat{R}_S(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}} \qquad \text{and} \qquad R_m(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}}.$$

Proof. By Definition 2.1,

$$\widehat{R}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{x} \in \mathcal{G}|_S} \frac{1}{m} \langle \boldsymbol{\sigma}, \boldsymbol{x} \rangle \right].$$

In Massart's lemma, take $A = \mathcal{G}|_S$. Then $r = \sqrt{m}$ and

$$|\mathcal{G}|_S| \leq \Pi_{\mathcal{G}}(m).$$

It follows that

$$\widehat{R}_S(\mathcal{G}) \leq \frac{\sqrt{m}\sqrt{2 \log \Pi_{\mathcal{G}}(m)}}{m} = \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}}.$$

Taking the expectation over S gives the same bound for $R_m(\mathcal{G})$. □

This estimate is purely combinatorial and does not depend on the data distribution. In the worst

case, $\Pi_{\mathcal{G}}(m) = 2^m$, so the upper bound in Corollary 2.7 is of constant order and does not vanish with m. Useful generalization guarantees therefore require subexponential growth. For example, combining Corollaries 2.4 and 2.7 shows that if $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$, then with probability at least $1-\delta$, every $g \in \mathcal{G}$ satisfies

$$\mathbb{E}[g] \leq \widehat{\mathbb{E}}_S[g] + 2\sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}} + \sqrt{\frac{\log(2/\delta)}{2m}}.$$

The remaining problem is to control the growth function.

Definition 2.8 Let $\mathcal{H}$ be a binary-valued hypothesis class, with labels represented by either $\{0,1\}$ or $\{-1,1\}$. A finite set $S \subseteq \mathcal{X}$ is **shattered** by $\mathcal{H}$ if every binary labeling of S is realized by some $h \in \mathcal{H}$. The **Vapnik–Chervonenkis dimension** (or **VC dimension**) of $\mathcal{H}$ is

$$\mathrm{VCdim}(\mathcal{H}) := \sup\{m \in \mathbb{N} : \Pi_{\mathcal{H}}(m) = 2^m\}.$$

Thus $\mathrm{VCdim}(\mathcal{H})$ is the largest sample size, if one exists, on which the class can realize every possible labeling. Lower bounds are commonly established by exhibiting a shattered set, whereas upper bounds require proving that no larger set can be shattered.

Example 2.9 Let

$$\mathcal{H}_{\mathrm{ray}} := \{x \mapsto \mathbb{1}(x \geq a) : a \in \mathbb{R}\}.$$

Determine $\mathrm{VCdim}(\mathcal{H}_{\mathrm{ray}})$. The geometry of this class is shown in Figure 1.

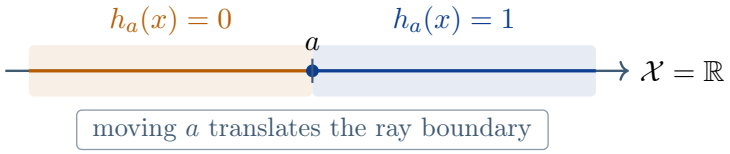

FIGURE 1

Solution. Any single point can be assigned either label by moving the threshold past that point. Two ordered points cannot be labeled 1 and then 0, so no two-point set is shattered. Therefore,

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{ray}}) = 1.$$

□

Theorem 2.10 (Sauer–Shelah lemma) If $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then for every $m \in \mathbb{N}$,

$$\Pi_{\mathcal{H}}(m) \le \sum_{i=0}^{\min(d,m)} \binom{m}{i}.$$

If $m \ge d \ge 1$, then

$$\Pi_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^{d}.$$

For $d = 0$, one has $\Pi_{\mathcal{H}}(m) \le 1$. In every case, $\Pi_{\mathcal{H}}(m) \in O(m^{d})$ as $m \to \infty$.

The logarithm of the largest possible growth function is linear in m until the Vapnik–Chervonenkis dimension is reached, after which Theorem 2.10 bounds it by a multiple of $d \log m$. This change is illustrated in Figure 2.

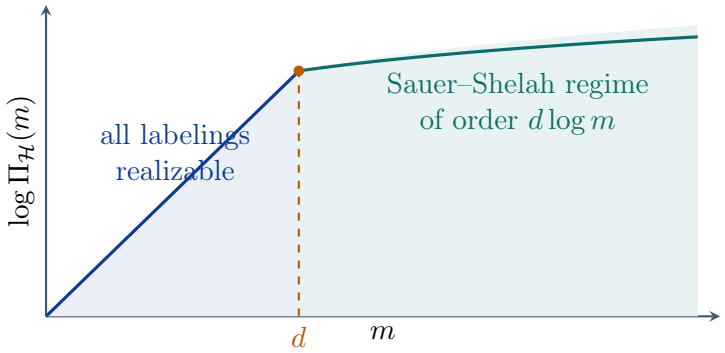

Figure 2

Combining the Sauer–Shelah lemma with the Rademacher bounds above gives, for a universal constant $k > 0$, a bound of the form

$$\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \le 2\sqrt{\frac{2d \log m}{m}} + k\sqrt{\frac{\log(2/\delta)}{m}}$$

with probability at least $1 - \delta$. Consequently, finite Vapnik–Chervonenkis dimension implies uniform convergence and hence agnostic PAC learnability by Proposition 1.6.

Lecture 3 — Tuesday, September 25, 2018

Theorem 3.1 (Fundamental theorem of statistical learning) For a binary-valued hypothesis class $\mathcal{H}$, the following statements are equivalent.

1. The class $\mathcal{H}$ is PAC learnable.
2. The class $\mathcal{H}$ is agnostically PAC learnable.
3. The class $\mathcal{H}$ has the uniform convergence property.
4. The dimension $\mathrm{VCdim}(\mathcal{H})$ is finite.

Proof. (4) *implies* (3). This implication follows from the Sauer–Shelah lemma and the Rademacher bound derived above.

(3) *implies* (2). This is the uniform excess risk argument in Proposition 1.6.

(2) *implies* (1). The realizable setting is a special case of the agnostic setting in which the optimal risk is zero.

(1) *implies* (4). We prove the contrapositive by establishing a **no-free-lunch** lower bound. Let A be an arbitrary learning algorithm, and suppose $\mathcal{H}$ shatters a set $B \subseteq \mathcal{X}$ of cardinality d. Let D be the uniform distribution on B, and let

$$S = ((X_1, Y_1), \dots, (X_m, Y_m))$$

be the labeled sample supplied to A. Denote $A(S)$ by h_S and define the set of unobserved points by

$$\overline{S} \coloneqq \{z \in B : z \neq X_i \text{ for every } i \in [m]\}.$$

Choose a labeling function $f : B \to \{0, 1\}$ uniformly at random, and generate the sample labels by $Y_i = f(X_i)$. Conditional on the observed sample, each label $f(x)$ at an unobserved point $x \in \overline{S}$ remains a fair random bit, independent of the prediction $h_S(x)$. Hence

$$\mathbb{E}_{f|S}[L_{D,f}(h_S)] \geq \mathbb{E}_{x \sim D}\left[\mathbb{1}(x \in \overline{S})\mathbb{E}_{f|S}\left[\mathbb{1}(h_S(x) \neq f(x))\right]\right]$$

$$= \frac{1}{2}\mathbb{E}_{x\sim D}\left[\mathbb{1}(x \in \overline{S})\right] = \frac{1}{2}D(\overline{S}).$$

Taking expectation over S and interchanging the order of expectation gives

$$\mathbb{E}_f \mathbb{E}_{S\sim D^m}[L_{D,f}(h_S)] \geq \frac{1}{2}\mathbb{E}_{S\sim D^m}[D(\overline{S})].$$

Because D is uniform on B, a fixed point of B is absent from the sample with probability $(1-1/d)^m$. Therefore,

$$\mathbb{E}_{S\sim D^m}[D(\overline{S})] = \left(1 - \frac{1}{d}\right)^m.$$

It follows that some fixed labeling f_0 satisfies

$$\mathbb{E}_{S\sim D^m}[L_{D,f_0}(h_S)] \geq \frac{1}{2}\left(1 - \frac{1}{d}\right)^m. \tag{3.1}$$

Since B is shattered, this labeling is realized by a hypothesis in $\mathcal{H}$ and the resulting problem is realizable. Set $X = L_{D,f_0}(h_S) \in [0,1]$. For any $\varepsilon \in (0,1)$,

$$\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}(X > \varepsilon)] + \mathbb{E}[X\mathbb{1}(X \leq \varepsilon)] \leq \mathbb{P}(X > \varepsilon) + \varepsilon\big(1 - \mathbb{P}(X > \varepsilon)\big).$$

Combining this inequality with (3.1) yields

$$\mathbb{P}_{S\sim D^m}\big(L_{D,f_0}(h_S) > \varepsilon\big) \geq \frac{\frac{1}{2}(1-1/d)^m - \varepsilon}{1-\varepsilon}. \tag{3.2}$$

Choose $\varepsilon < 1/2$ and

$$\delta < \frac{1/2 - \varepsilon}{1-\varepsilon} \qquad \text{and} \qquad \psi \coloneqq 2\big(\delta(1-\varepsilon) + \varepsilon\big) < 1.$$

The right-hand side of (3.2) is at least δ whenever

$$m \leq \frac{\log(1/\psi)}{-\log(1-1/d)}.$$

For $d \geq 2$, the inequality

$$-\log\left(1 - \frac{1}{d}\right) \leq \frac{1}{d-1}$$

shows that the last upper bound is at least

$$(d-1)\log(1/\psi).$$

Thus, writing $m^*_{\mathcal{H}}(\varepsilon,\delta)$ for the smallest possible sample complexity among all algorithms, we have

that

$$\text{if } \operatorname{VCdim}(\mathcal{H}) \geq d \qquad \text{then} \qquad m^*_{\mathcal{H}}(\varepsilon, \delta) \geq \lfloor (d-1)\log(1/\psi) \rfloor .$$

If $\operatorname{VCdim}(\mathcal{H}) = \infty$, this lower bound can be made arbitrarily large, so $\mathcal{H}$ is not PAC learnable. This proves the contrapositive and completes the cycle of implications. □

Historical Note () A pathological example of Shai Ben-David and collaborators showed that set-theoretic questions related to the continuum hypothesis can enter the study of learnability. The standard finite-dimensional theory above is unaffected, but the example illustrates how subtle unrestricted learning problems can become.

For completeness, we now supply the moment-generating-function proof of Massart's lemma.

Proof of Lemma 2.6. Let

$$X \coloneqq \sup_{\boldsymbol{y} \in A} \langle \boldsymbol{\sigma}, \boldsymbol{y} \rangle \qquad \text{and} \qquad r \coloneqq \max_{\boldsymbol{y} \in A} \|\boldsymbol{y}\|_2 .$$

The claim is immediate if $r = 0$ or $|A| = 1$, so assume otherwise. For $t > 0$, Jensen's inequality gives

$$\exp(t\mathbb{E}[X]) \leq \mathbb{E}\left[e^{tX}\right],$$

and therefore

$$\mathbb{E}[X] \leq \frac{1}{t} \log \mathbb{E}\left[e^{tX}\right]. \tag{3.3}$$

Because the exponential function is increasing,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\sup_{\boldsymbol{y} \in A} e^{t\langle \boldsymbol{\sigma}, \boldsymbol{y} \rangle}\right] \leq \sum_{\boldsymbol{y} \in A} \mathbb{E}\left[e^{t\langle \boldsymbol{\sigma}, \boldsymbol{y} \rangle}\right] = \sum_{\boldsymbol{y} \in A} \prod_{i=1}^{m} \mathbb{E}\left[e^{t\sigma_i y_i}\right].$$

For a Rademacher random variable σ_i,

$$\mathbb{E}\left[e^{t\sigma_i y_i}\right] = \frac{1}{2}\left(e^{ty_i} + e^{-ty_i}\right) = \cosh(ty_i) \leq e^{t^2 y_i^2/2}.$$

Consequently,

$$\mathbb{E}\left[e^{tX}\right] \leq \sum_{\boldsymbol{y} \in A} \exp\left(\frac{t^2}{2} \sum_{i=1}^{m} y_i^2\right) = \sum_{\boldsymbol{y} \in A} \exp\left(\frac{t^2}{2}\|\boldsymbol{y}\|_2^2\right) \leq |A| \exp\left(\frac{t^2 r^2}{2}\right).$$

Substitution into (3.3) yields

$$\mathbb{E}[X] \le \frac{\log |A|}{t} + \frac{tr^2}{2}.$$

The right-hand side is minimized at

$$t = \frac{\sqrt{2 \log |A|}}{r},$$

which gives

$$\mathbb{E}[X] \le r\sqrt{2 \log |A|}.$$

Dividing by m proves Lemma 2.6. □

The same sign symmetry argument also gives the scaling identity

$$\widehat{R}_S(c\mathcal{G}) = |c|\, \widehat{R}_S(\mathcal{G}), \quad c \in \mathbb{R}.$$

Lecture 4 — Tuesday, October 02, 2018

Sample compression provides a second route to learnability. The central idea is to retain a small part of the sample, reconstruct a hypothesis from that part, and use the remaining observations to validate the reconstructed hypothesis. This perspective also anticipates model selection, regularization, and PAC-Bayes bounds. Sample compression is a very active area of research!

To begin, split a sample S of size m into a training subsample T and a validation subsample V, as illustrated in Figure 3. The hypothesis h_T is constructed using only T. Conditional on T, the observations in V remain independent draws from D, and hence

$$\mathbb{E}_{V \sim D^{|V|}}\big[L_V(h_T)\big] = L_D(h_T).$$

Theorem 4.1 (Bernstein's inequality) Assume that the loss takes values in $[0, 1]$. For every distribution D and every $\delta \in (0, 1)$,

$$\mathbb{P}_{V \sim D^{|V|}} \left\{ L_D(h_T) - L_V(h_T) \ge \sqrt{\frac{2L_V(h_T)\log(1/\delta)}{|V|}} + \frac{4\log(1/\delta)}{|V|} \right\} \le \delta.$$

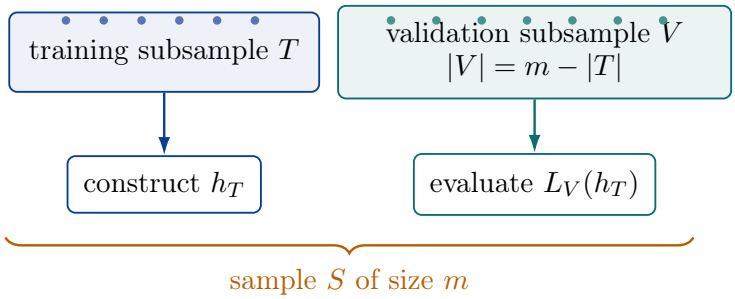

FIGURE 3

For comparison, Hoeffding's inequality gives

$$\mathbb{P}_{V\sim D^{|V|}}\left\{L_D(h_T) - L_V(h_T) > \sqrt{\frac{\log(1/\delta)}{2|V|}}\right\} \le \delta.$$

Unlike this additive bound, the leading term in Theorem 4.1 becomes smaller when the validation loss is small.

We now allow the training subsample to be selected *after* observing all of the data. For $j \ge 0$, let $[m]^j_{\neq}$ denote the ordered j-tuples of distinct indices, with $[m]^0_{\neq}$ containing only the empty tuple. For a fixed budget k, define

$$\mathcal{T}_{m,k} := \bigcup_{j=0}^{\min(k,m)} [m]^j_{\neq} \qquad \text{and} \qquad N_{m,k} := |\mathcal{T}_{m,k}|.$$

For $m \ge 1$,

$$N_{m,k} = \sum_{j=0}^{\min(k,m)} \frac{m!}{(m-j)!} \le (k+1)m^k.$$

Let $\mathcal{J}$ be a nonempty finite set of side-information values. For $I = (i_1, \dots, i_j) \in \mathcal{T}_{m,k}$, write

$$S_I = (z_{i_1}, \dots, z_{i_j}) \qquad \text{and} \qquad V_I = (z_\ell : \ell \in [m] \setminus \{i_1, \dots, i_j\}),$$

where the latter sample retains the original index order. Given reconstruction maps

$$B_j \colon \mathcal{Z}^j \times \mathcal{J} \to \mathcal{Y}^{\mathcal{X}}, \qquad 0 \le j \le k,$$

set $h_{I,s} = B_j(S_I, s)$ for $s \in \mathcal{J}$. For each fixed pair (I, s), the hypothesis $h_{I,s}$ depends only on S_I and is therefore independent of the $m - j$ observations in V_I.

Suppose $m > k$. Applying Theorem 4.1 with failure probability $\delta/(N_{m,k}|\mathcal{J}|)$ and taking a union bound shows that, with probability at least $1-\delta$,

$$L_D(h_{I,s}) - L_{V_I}(h_{I,s}) \le \sqrt{\frac{2L_{V_I}(h_{I,s})\log(N_{m,k}|\mathcal{J}|/\delta)}{m-j}} + \frac{4\log(N_{m,k}|\mathcal{J}|/\delta)}{m-j} \tag{4.1}$$

simultaneously for every $I \in \mathcal{T}_{m,k}$ and $s \in \mathcal{J}$. Consequently, the following three-stage procedure remains valid even when (I, s) is selected using the entire sample:

1. Select an index tuple I of size at most k and side information s using S.
2. Reconstruct the hypothesis $h_{I,s} = B_{|I|}(S_I, s)$.
3. Report the confidence bound for $L_D(h_{I,s})$ obtained from (4.1).

The logarithmic factor is the price paid for selecting the retained indices and the side information after seeing the data. This index-based formulation also remains unambiguous when the sample contains repeated observations.

Definition 4.2 Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. A **sample compression scheme** with kernel budget k and finite side-information set $\mathcal{J}$ consists, for every m, of a compression map

$$A_m \colon \mathcal{Z}^m \to \mathcal{T}_{m,k} \times \mathcal{J}$$

and reconstruction maps $B_j \colon \mathcal{Z}^j \times \mathcal{J} \to \mathcal{Y}^{\mathcal{X}}$, $0 \le j \le k$. If $A_m(S) = (I, s)$, the reconstructed hypothesis $B_{|I|}(S_I, s)$ must agree with every observation in S whenever S is realizable by $\mathcal{H}$. The size of the scheme is

$$k + \lceil \log_2 |\mathcal{J}| \rceil .$$

The scheme is **proper** if every reconstructed hypothesis belongs to $\mathcal{H}$.

The map A_m compresses the sample, whereas the appropriate B_j reconstructs a hypothesis from the retained examples and side information. Allowing at most k examples, rather than exactly k, also covers samples with $m < k$.

Proposition 4.3 Suppose that $\mathcal{H}$ has a realizable compression scheme with kernel budget k. An agnostic procedure with the same budget can be constructed as follows:

1. Find an empirical risk minimizer h^* over $\mathcal{H}$, and let S' consist of the examples in S that h^* classifies correctly.
2. Apply the realizable compression scheme to S'.
3. Return the hypothesis h reconstructed from the compressed subsample of S'.

Then $L_S(h) \le L_S(h^*)$. If the compression scheme is proper, equality holds.

Proof. The sample S' is realizable because h^* labels every one of its examples correctly. Thus the hypothesis reconstructed in (3) is also correct on all of S'. It can therefore make errors only on examples on which h^* errs, so $L_S(h) \le L_S(h^*)$. If the scheme is proper, then $h \in \mathcal{H}$; empirical optimality of h^* supplies the reverse inequality. □

Theorem 4.4 If $\mathcal{H}$ has a compression scheme of finite size, then $\mathcal{H}$ is improperly PAC learnable. If the scheme is proper, the resulting PAC learner is proper.

Proof. Let the scheme have kernel budget k and side-information set $\mathcal{J}$. In the realizable setting, if $A_m(S) = (I, s)$, then $h = B_{|I|}(S_I, s)$ is consistent with all of S. Its loss on V_I is therefore zero. For $m > k$, (4.1) and $m - |I| \ge m - k$ give, with probability at least $1 - \delta$,

$$L_D(h) \le \frac{4 \log(N_{m,k} |\mathcal{J}| / \delta)}{m - k}.$$

For fixed k and finite $\mathcal{J}$, the right-hand side converges to zero as m tends to infinity, which proves the claim. □

Theorem 4.5 (Moran–Yehudayoff compression theorem) Every binary hypothesis class of finite VC dimension d has a sample compression scheme of size $2^{O(d)}$.

Historical Note (2018) Littlestone and Warmuth introduced sample compression and conjectured that every class of VC dimension d admits a compression scheme whose size is proportional to d. Shay Moran and Amir Yehudayoff proved the weaker, but still dimension-dependent, bound in Theorem 4.5; see *Sample compression schemes for VC classes*. The stronger linear size conjecture remains open.

Lecture 5 — Tuesday, October 16, 2018

2. Nonuniform Learnability and PAC-Bayes

Nonuniform Learnability

Definition 5.1 A hypothesis class $\mathcal{H}$ is **nonuniformly learnable** if there are learning rules

$$A_m \colon \mathcal{Z}^m \times (0,1)^2 \to \mathcal{H}$$

and a sample complexity

$$m_{\mathcal{H}}^{\mathrm{NU}} \colon (0,1)^2 \times \mathcal{H} \to \mathbb{N}$$

such that the following holds: for every $\varepsilon, \delta \in (0,1)$, every $h \in \mathcal{H}$, every distribution D, and every $m \geq m_{\mathcal{H}}^{\mathrm{NU}}(\varepsilon, \delta, h)$,

$$\mathbb{P}_{S \sim D^m} \left(L_D(A_m(S, \varepsilon, \delta)) \leq L_D(h) + \varepsilon \right) \geq 1 - \delta.$$

The sample complexity in Definition 5.1 depends on the comparator h. Equivalently, for a subclass $\mathcal{H}' \subseteq \mathcal{H}$, the learner competes with $\inf_{h \in \mathcal{H}'} L_D(h)$ once the sample size is at least the supremum of the relevant comparator-dependent sample complexities. More precisely, using accuracy $\varepsilon/2$ in Definition 5.1, choosing an element of $\mathcal{H}'$ whose risk is within $\varepsilon/2$ of the infimum gives

$$\mathbb{P}_{S \sim D^m} \left(L_D(A_m(S, \varepsilon/2, \delta)) \leq \inf_{h \in \mathcal{H}'} L_D(h) + \varepsilon \right) \geq 1 - \delta$$

whenever

$$m \geq \sup_{h \in \mathcal{H}'} m_{\mathcal{H}}^{\mathrm{NU}}(\varepsilon/2, \delta, h).$$

If this supremum is finite for $\mathcal{H}' = \mathcal{H}$, then the class is agnostically PAC learnable.

Theorem 5.2 Let $\mathcal{H}$ be a binary hypothesis class. Then $\mathcal{H}$ is nonuniformly learnable if and only if

$$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n,$$

where every $\mathcal{H}_n$ is PAC learnable.

Proof. *Forward implication.* Assume that $\mathcal{H}$ is nonuniformly learnable, and define

$$\mathcal{H}_n \coloneqq \left\{h \in \mathcal{H} : m_{\mathcal{H}}^{\mathrm{NU}}(1/8, 1/7, h) \leq n\right\}.$$

Every $h \in \mathcal{H}$ belongs to some $\mathcal{H}_n$, so $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$. If some $\mathcal{H}_n$ had infinite VC dimension, the no-free-lunch argument in the proof of Theorem 3.1 could be applied with sample size n. It would produce a realizable distribution for which the learner has risk greater than $1/8$ with probability at least $1/7$. This contradicts the definition of $\mathcal{H}_n$. Therefore every $\mathcal{H}_n$ has finite VC dimension and is PAC learnable by Theorem 3.1.

Reverse implication. Suppose that $\mathcal{H} = \bigcup_{j\in\mathbb{N}} \mathcal{G}_j$, where every $\mathcal{G}_j$ is PAC learnable. Replacing the original classes by

$$\mathcal{H}_n \coloneqq \bigcup_{j=1}^{n} \mathcal{G}_j$$

makes the sequence increasing without changing its union. Each $\mathcal{H}_n$ has finite VC dimension and therefore has the uniform convergence property.

Indeed,

$$\Pi_{\mathcal{H}_n}(m) \leq \sum_{j=1}^{n} \Pi_{\mathcal{G}_j}(m).$$

Every term on the right grows at most polynomially by the Sauer–Shelah lemma, whereas a class of infinite VC dimension has growth function 2^m for every m. This proves the asserted finiteness of the VC dimension.

Let $(w_n)_{n\in\mathbb{N}}$ be a fixed sequence of positive weights satisfying $\sum_{n=1}^{\infty} w_n \leq 1$. For fixed n, m, and η, put

$$\mathcal{E}_{n,m,\eta} \coloneqq \left\{\varepsilon \in (0,1) : m_{\mathcal{H}_n}^{\mathrm{UC}}(\varepsilon, \eta) \leq m\right\}.$$

Define the confidence radius by

$$\varepsilon_n(m, \eta) \coloneqq \begin{cases} \min\{1, \inf \mathcal{E}_{n,m,\eta} + 1/m\}, & \mathcal{E}_{n,m,\eta} \neq \varnothing, \\ 1, & \mathcal{E}_{n,m,\eta} = \varnothing. \end{cases}$$

The added $1/m$ avoids assuming that the defining infimum is attained. If the radius equals 1, the desired deviation event is empty; if it is smaller than 1, some admissible element of $\mathcal{E}_{n,m,\eta}$ lies

below it. Uniform convergence and the union bound therefore give

$$\begin{aligned}
&\mathbb{P}\left(\exists n \in \mathbb{N}, h \in \mathcal{H}_n : |L_D(h) - L_S(h)| > \varepsilon_n(m, \delta w_n)\right) \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}\left(\exists h \in \mathcal{H}_n : |L_D(h) - L_S(h)| > \varepsilon_n(m, \delta w_n)\right) \\
&\leq \sum_{n=1}^{\infty} \delta w_n \\
&\leq \delta.
\end{aligned}$$

Thus, for every distribution D, with probability at least $1 - \delta$,

$$|L_D(h) - L_S(h)| \leq \varepsilon_n(m, \delta w_n) \quad \text{for every } n \in \mathbb{N} \text{ and every } h \in \mathcal{H}_n. \tag{5.1}$$

For $h \in \mathcal{H}$, let $n(h) = \inf\{n \in \mathbb{N} : h \in \mathcal{H}_n\}$ be the first index for which $h \in \mathcal{H}_{n(h)}$, and define

$$\mathrm{pen}_m(h, \delta) \coloneqq \varepsilon_{n(h)}(m, \delta w_{n(h)}).$$

The **structural risk minimization rule** returns

$$\widehat{h} \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \left\{ L_S(h) + \mathrm{pen}_m(h, \delta) \right\}.$$

If a minimizer does not exist, an approximate minimizer with arbitrarily small optimization error may be used. On the event in (5.1), every comparator $h \in \mathcal{H}$ satisfies

$$L_D(\widehat{h}) \leq L_S(\widehat{h}) + \mathrm{pen}_m(\widehat{h}, \delta) \leq L_S(h) + \mathrm{pen}_m(h, \delta) \leq L_D(h) + 2\,\mathrm{pen}_m(h, \delta).$$

For fixed n, η, and $\gamma > 0$, take $m \geq m^{\mathrm{UC}}_{\mathcal{H}_n}(\gamma/2, \eta)$ and $m \geq 2/\gamma$. Then the definition gives $\varepsilon_n(m, \eta) \leq \gamma$. Thus, for each fixed h, the penalty converges to zero as m tends to infinity. A comparator-dependent sample complexity therefore makes the final term at most ε, proving nonuniform learnability. □

The tradeoff behind structural risk minimization is depicted in Figure 4. Moving to a richer class may reduce empirical risk, but it generally enlarges the confidence penalty. The algorithm minimizes their sum.

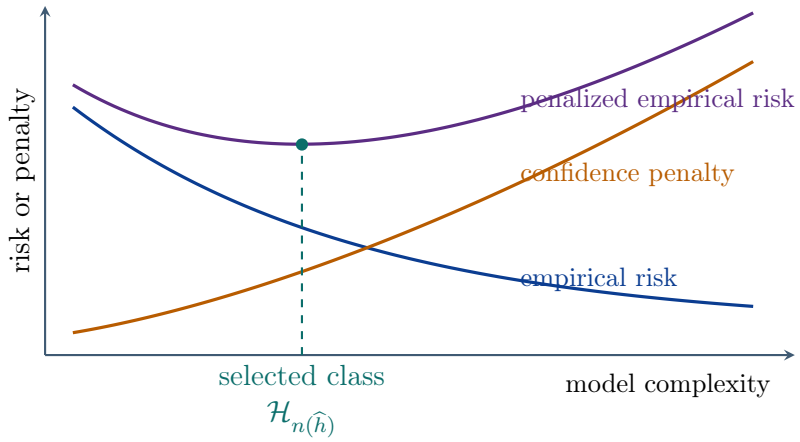

FIGURE 4

Lecture 6 — Tuesday, October 09, 2018

Structural Risk Minimization

The proof of Theorem 5.2 produces an explicit model selection method. Given an increasing sequence

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \qquad \text{with} \qquad \bigcup_{n\in\mathbb{N}} \mathcal{H}_n = \mathcal{H},$$

structural risk minimization balances two quantities. The empirical optimum

$$\widehat{L}_n \coloneqq \inf_{h\in\mathcal{H}_n} L_S(h)$$

cannot increase with n, whereas the confidence radius $\varepsilon_n(m, \delta w_n)$ typically grows with model complexity. These opposing trends are shown in Figure 5.

Proposition 6.1 Let the hypotheses be assigned penalties as in the proof of Theorem 5.2, and let

$$\widehat{h} \in \arg\min_{h\in\mathcal{H}} \left\{ L_S(h) + \mathrm{pen}_m(h, \delta) \right\}.$$

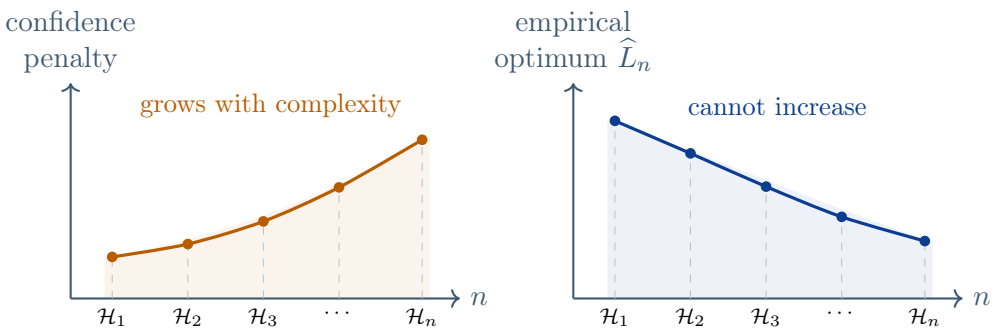

FIGURE 5

Then, with probability at least $1-\delta$,

$$L_D(\widehat{h}) \le \inf_{h\in\mathcal{H}} \{L_D(h) + 2\,\mathrm{pen}_m(h,\delta)\}\,. \tag{6.1}$$

Proof. On the simultaneous event in (5.1), empirical and population risk differ by at most the assigned penalty. For any $h \in \mathcal{H}$, optimality of $\widehat{h}$ gives

$$L_D(\widehat{h}) \le L_S(\widehat{h}) + \mathrm{pen}_m(\widehat{h},\delta) \le L_S(h) + \mathrm{pen}_m(h,\delta) \le L_D(h) + 2\,\mathrm{pen}_m(h,\delta).$$

Taking the infimum over h proves (6.1). □

The factor of two in (6.1) comes from using the uniform convergence bound once for the selected hypothesis and once for the comparator. These two deviations are depicted in Figure 6.

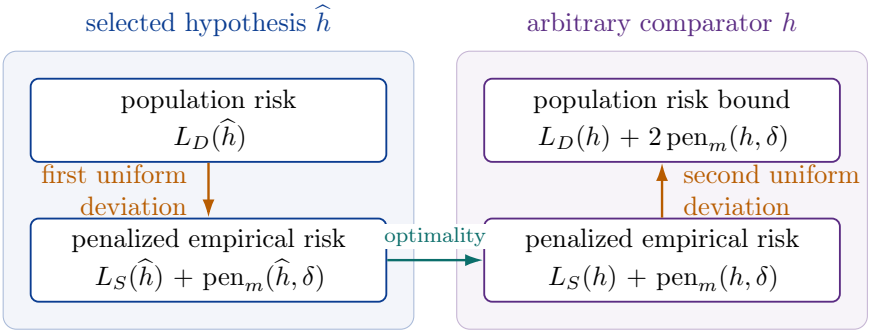

FIGURE 6

The same construction has an especially transparent form for a countable hypothesis class.

Theorem 6.2 Let $\mathcal{H} = \{h_n : n \in \mathbb{N}\}$, and let $(w_n)_{n\in\mathbb{N}}$ be positive weights, chosen independently of the sample, such that

$$\sum_{n=1}^{\infty} w_n \leq 1.$$

For every distribution D and every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|L_D(h_n) - L_S(h_n)| \leq \sqrt{\frac{\log(1/w_n) + \log(2/\delta)}{2m}} \quad \text{for every } n \in \mathbb{N}. \tag{6.2}$$

Proof. For a fixed n, Hoeffding's inequality gives

$$\mathbb{P}_{S\sim D^m}\left(|L_D(h_n) - L_S(h_n)| > \sqrt{\frac{\log(2/(\delta w_n))}{2m}}\right) \leq \delta w_n.$$

The union bound and the inequality $\sum_{n=1}^{\infty} w_n \leq 1$ show that all of these events hold simultaneously with probability at least $1-\delta$. Finally,

$$\log\left(\frac{2}{\delta w_n}\right) = \log\left(\frac{1}{w_n}\right) + \log\left(\frac{2}{\delta}\right),$$

which gives (6.2). □

For fixed m and δ, the penalty in (6.2) grows as w_n decreases. Since the weights of an infinite sequence must tend to zero along some subsequence, sufficiently low-weight hypotheses receive vacuous bounds. The weights must be fixed before observing the sample; otherwise the simultaneous probability guarantee does not follow. They can be interpreted as a prior allocation of confidence across hypotheses, even though the guarantee itself is entirely frequentist.

Occam's Razor

The penalty in Theorem 6.2 has an interpretation in terms of **description lengths**. Assign a positive weight $w(h)$ to every $h \in \mathcal{H}$, with

$$\sum_{h\in\mathcal{H}} w(h) \leq 1,$$

and define the complexity

$$c(h) := \log\left(\frac{1}{w(h)}\right).$$

A large weight corresponds to a short description and hence to a simple hypothesis. The summability requirement ensures that only a limited number of hypotheses can receive short descriptions. For example, a vocabulary of 50,000 words permits 50,000 one-word strings but $50{,}000^2$ two-word strings, so longer descriptions must be assigned progressively smaller individual weights.

Corollary 6.3 (Occam bound) With probability at least $1-\delta$,

$$L_D(h) \le L_S(h) + \sqrt{\frac{c(h) + \log(2/\delta)}{2m}} \quad \text{for every } h \in \mathcal{H}.$$

Thus, among hypotheses with the same empirical risk, the bound in Corollary 6.3 favors the one with the larger prior weight, or equivalently the shorter description. This principle is commonly called **Occam's razor**.

For example, consider a regression problem in which both linear and polynomial regression classes are available. The linear class has a simple description and can be assigned a relatively large weight, whereas the polynomial class is more complex and receives a smaller weight. The Occam's razor bound then penalizes the polynomial class more heavily, especially when the sample size is small. On a small sample, many flexible polynomial fits may have low empirical risk, even when a linear relation generated the data (Figure 7 illustrates several such fits). However, as the sample grows, a polynomial chosen merely to interpolate an earlier sample is unlikely to continue fitting new observations (Figure 8).

The appropriate notion of simplicity is therefore relative to the available sample size. A quadratic class may be adequately controlled by several hundred observations, whereas a degree-100 polynomial class carries a much larger complexity penalty. Declaring a complicated, data-selected hypothesis to be simple only after seeing its low empirical risk invalidates the bound, because the weights in Corollary 6.3 must be fixed independently of the sample.

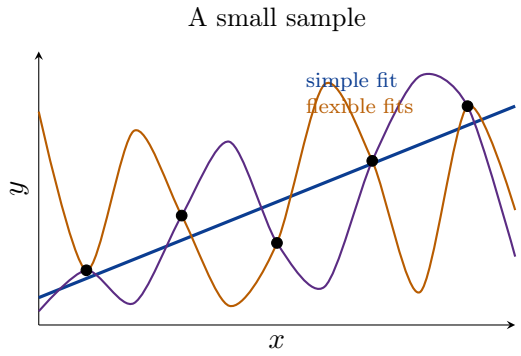

Figure 7

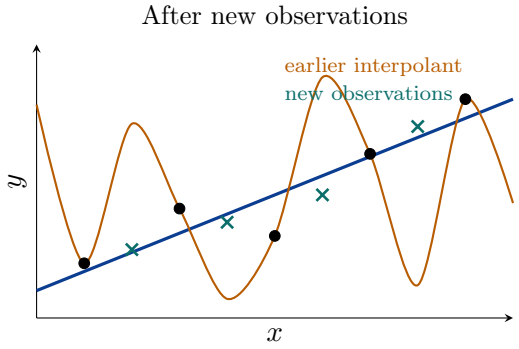

Figure 8

PAC-Bayes

The weights in Theorem 6.2 must be fixed before the sample is observed. PAC-Bayes analysis retains this fixed prior but permits the learner to choose a data-dependent distribution over hypotheses.

Definition 6.4 A **Gibbs classifier** is a probability distribution $\mathbb{Q}$ on $\mathcal{H}$. To predict at x, it draws $h \sim \mathbb{Q}$ and returns $h(x)$. Its population and empirical risks are

$$L_D(\mathbb{Q}) := \mathbb{E}_{h\sim\mathbb{Q}}[L_D(h)] \qquad \text{and} \qquad L_S(\mathbb{Q}) := \mathbb{E}_{h\sim\mathbb{Q}}[L_S(h)].$$

Equivalently, if $\widehat{D} = m^{-1}\sum_{i=1}^{m} \delta_{(X_i,Y_i)}$ is the empirical measure, then $L_S(\mathbb{Q}) = L_{\widehat{D}}(\mathbb{Q})$.

Definition 6.5 Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $\mathcal{H}$. If $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, their **Kullback–Leibler divergence** is

$$\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{h\sim\mathbb{Q}}\left[\log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(h)\right)\right].$$

If $\mathbb{Q}$ is not absolutely continuous with respect to $\mathbb{P}$, set $\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) = +\infty$.

Theorem 6.6 (PAC-Bayes binary relative entropy bound) Fix a prior distribution $\mathbb{P}$ on $\mathcal{H}$ independently of the sample. For every distribution D and every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathrm{kl}\big(L_S(\mathbb{Q})\|L_D(\mathbb{Q})\big) \le \frac{\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log(2\sqrt{m}/\delta)}{m} \quad \text{for every probability distribution } \mathbb{Q} \text{ on } \mathcal{H}, \tag{6.3}$$

where

$$\mathrm{kl}(q\|p) := q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

is the **binary relative entropy**, with its continuous extension at the endpoints.

Pinsker's inequality, $\mathrm{kl}(q\|p) \ge 2(q-p)^2$, immediately gives the more familiar consequence

$$L_D(\mathbb{Q}) \le L_S(\mathbb{Q}) + \sqrt{\frac{\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log(2\sqrt{m}/\delta)}{2m}} \tag{6.4}$$

simultaneously for every $\mathbb{Q}$.

If $\mathbb{Q} = \delta_h$ is a point mass and $\mathbb{P}(\{h\}) > 0$, then

$$\mathrm{KL}(\delta_h \| \mathbb{P}) = \log\left(\frac{1}{\mathbb{P}(\{h\})}\right).$$

Thus (6.4) recovers an Occam bound. The advantage of PAC-Bayes analysis is that $\mathbb{Q}$ may spread its mass over many hypotheses after the sample has been observed.

Example 6.7 Let $A \subseteq \mathcal{H}$ satisfy $\mathbb{P}(A) > 0$, and define the conditional distribution

$$\mathbb{Q}(B) \coloneqq \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

Compute $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$.

Solution. The Radon–Nikodym derivative is

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(h) = \frac{\mathbb{1}(h \in A)}{\mathbb{P}(A)} \quad \text{for } \mathbb{Q}\text{-almost every } h.$$

Consequently,

$$\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\log\left(\frac{1}{\mathbb{P}(A)}\right)\right] = \log\left(\frac{1}{\mathbb{P}(A)}\right).$$

For example, if $A = \{h, h'\}$ and the two hypotheses have the same empirical risk, conditioning on both pools their prior mass and may give a smaller complexity term than selecting either point mass. □

Lecture 7 — Tuesday, October 23, 2018

The terminology of PAC-Bayes is motivated by Bayesian prediction, although the resulting guarantee is frequentist. In a **Bayesian model**, a latent state Θ has prior distribution π, and conditional on $\Theta = \theta$ the observations are independent with distribution $\mathbb{P}_\theta$. The **posterior distribution** is

$$\Theta \mid S \sim \pi(\cdot \mid S).$$

If

$$\mathbb{P}(Y_i = 1 \mid X_i = x, \Theta = \theta) = \kappa(x, \theta),$$

then the **posterior predictive probability** is

$$\mathbb{P}(Y_{m+1} = 1 \mid X_{m+1} = x, S) = \int \kappa(x, \theta)\, \pi(\mathrm{d}\theta \mid S). \tag{7.1}$$

This averages predictive uncertainty over the posterior distribution of the latent state. The conditional independence structure is depicted in Figure 9.

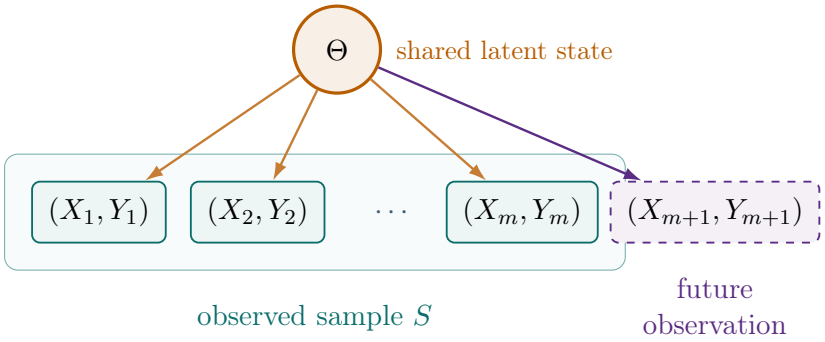

FIGURE 9

Thresholding (7.1) at $1/2$ gives the Bayesian classifier

$$A(S)(x) = \mathbb{1}\left(\int \kappa(x, \theta)\, \pi(\mathrm{d}\theta \mid S) \geq \frac{1}{2} \right).$$

The observations are conditionally independent given Θ, but they are generally dependent under the marginal Bayesian model because they share the same latent state.

PAC-Bayes analysis separates the probabilistic guarantee from this generative interpretation. It assumes only that $S \sim D^m$, fixes a prior $\mathbb{P}$ independently of S, and allows the learning rule to return any data-dependent distribution $\mathbb{Q}_S$ on $\mathcal{H}$. The analogy is summarized below.

Bayesian analysis	PAC-Bayes analysis
posterior predictive distribution	Gibbs distribution $\mathbb{Q}_S$
generative prior on a latent state	sample-independent reference $\mathbb{P}$
posterior probability statement	frequentist high-probability bound

For binary hypotheses with values in $\{0, 1\}$, define

$$q_{\mathbb{Q}}(x) := \mathbb{P}_{h \sim \mathbb{Q}}(h(x) = 1).$$

Tonelli's theorem gives

$$\begin{aligned} L_D(\mathbb{Q}) &= \mathbb{E}_{h\sim\mathbb{Q}}\mathbb{E}_{(X,Y)\sim D}\left[\mathbb{1}(h(X)\neq Y)\right] \\ &= \mathbb{E}_{(X,Y)\sim D}\left[\mathbb{P}_{h\sim\mathbb{Q}}(h(X)\neq Y)\right] \\ &= \mathbb{E}_{(X,Y)\sim D}\left[q_{\mathbb{Q}}(X)^{1-Y}\big(1-q_{\mathbb{Q}}(X)\big)^{Y}\right]. \end{aligned}$$

This quantity is called the **Gibbs risk**. Indeed, when $Y=0$ the randomized classifier errs with probability $q_{\mathbb{Q}}(X)$, and when $Y=1$ it errs with probability $1-q_{\mathbb{Q}}(X)$. The associated deterministic **majority vote classifier** is

$$B_{\mathbb{Q}}(x) \coloneqq \mathbb{1}\big(q_{\mathbb{Q}}(x)\geq 1/2\big).$$

If $\mathbb{Q}=\mathbb{Q}_S$ is data dependent, then $q_{\mathbb{Q}_S}\colon \mathcal{X}\to[0,1]$ plays the same formal role as a posterior predictive probability, without requiring a Bayesian generative model.

Proposition 7.1 Suppose that $\mathcal{Y}=\{-1,1\}$ and that every $h\in\mathcal{H}$ is $\{-1,1\}$-valued. Define

$$B_{\mathbb{Q}}(x) \coloneqq \operatorname{sign}\left(\mathbb{E}_{h\sim\mathbb{Q}}[h(x)]\right)$$

with an arbitrary convention when the expectation is zero. Then

$$L_D(B_{\mathbb{Q}}) \leq 2L_D(\mathbb{Q}).$$

Proof. The majority vote is incorrect only if

$$Y\mathbb{E}_{h\sim\mathbb{Q}}[h(X)]\leq 0.$$

Therefore, Markov's inequality yields

$$\begin{aligned} L_D(B_{\mathbb{Q}}) &\leq \mathbb{P}_{(X,Y)\sim D}\left(1-Y\mathbb{E}_{h\sim\mathbb{Q}}[h(X)]\geq 1\right) \\ &\leq \mathbb{E}_{(X,Y)\sim D}\left[1-Y\mathbb{E}_{h\sim\mathbb{Q}}[h(X)]\right] \\ &= \mathbb{E}_{(X,Y)\sim D}\mathbb{E}_{h\sim\mathbb{Q}}\left[1-Yh(X)\right]. \end{aligned}$$

Because $Yh(X)$ equals 1 for a correct prediction and -1 for an error,

$$1-Yh(X) = 2\mathbb{1}(h(X)\neq Y).$$

Substitution gives $L_D(B_{\mathbb{Q}})\leq 2L_D(\mathbb{Q})$. □

Thus a PAC-Bayes bound on the Gibbs risk also controls the corresponding deterministic majority

vote. This observation applies to many ensemble procedures, including boosting and collections of decision trees or neural networks.

Theorem 7.2 (Donsker–Varadhan change-of-measure inequality) Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $\mathcal{H}$, with $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}$. For every measurable $\phi\colon \mathcal{H} \to \mathbb{R}$ such that $\mathbb{E}_{h\sim\mathbb{P}}[e^{\phi(h)}]$ is finite,

$$\mathbb{E}_{h\sim\mathbb{Q}}[\phi(h)] \le \mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log \mathbb{E}_{h\sim\mathbb{P}}[e^{\phi(h)}].$$

Proof. Let $r = \mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P}$. Jensen's inequality for the concave logarithm gives

$$\mathbb{E}_{h\sim\mathbb{Q}}[\phi(h)] - \mathrm{KL}(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\log\left(\frac{e^{\phi(h)}}{r(h)}\right)\right] \le \log \mathbb{E}_{h\sim\mathbb{Q}}\left[\frac{e^{\phi(h)}}{r(h)}\right] = \log \mathbb{E}_{h\sim\mathbb{P}}[e^{\phi(h)}].$$

Rearranging proves the claim. □

Theorem 7.3 Assume binary classification with zero-one loss. Let $\Delta\colon [0,1]^2 \to [0,\infty)$ be jointly convex, let $\lambda > 0$, and define

$$\mathcal{I}_\Delta(m,\lambda) \coloneqq \sup_{p\in[0,1]} \mathbb{E}_{K\sim\mathrm{Bin}(m,p)}\left[\exp\left(\lambda\Delta(K/m,p)\right)\right].$$

Fix a prior $\mathbb{P}$ independently of the sample. With probability at least $1-\delta$,

$$\Delta\big(L_S(\mathbb{Q}), L_D(\mathbb{Q})\big) \le \frac{\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log\big(\mathcal{I}_\Delta(m,\lambda)/\delta\big)}{\lambda} \tag{7.2}$$

simultaneously for every probability distribution $\mathbb{Q}$ on $\mathcal{H}$.

Proof. For a fixed sample, joint convexity and Jensen's inequality give

$$\lambda\Delta\big(L_S(\mathbb{Q}), L_D(\mathbb{Q})\big) \le \mathbb{E}_{h\sim\mathbb{Q}}\left[\lambda\Delta\big(L_S(h), L_D(h)\big)\right].$$

If $\mathbb{Q}$ is not absolutely continuous with respect to $\mathbb{P}$, then $\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) = \infty$ and the claimed inequality is trivial. Otherwise, apply Theorem 7.2 with

$$\phi_S(h) \coloneqq \lambda\Delta\big(L_S(h), L_D(h)\big)$$

and define

$$X_{\mathbb{P}}(S) \coloneqq \mathbb{E}_{h\sim\mathbb{P}}\left[e^{\phi_S(h)}\right].$$

It follows deterministically, for every $\mathbb{Q}$, that

$$\lambda\Delta\big(L_S(\mathbb{Q}), L_D(\mathbb{Q})\big) \le \mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log X_{\mathbb{P}}(S). \tag{7.3}$$

Markov's inequality shows that, with probability at least $1-\delta$,

$$X_{\mathbb{P}}(S) \le \frac{\mathbb{E}_{S'\sim D^m}[X_{\mathbb{P}}(S')]}{\delta}.$$

For each fixed h, the random variable $mL_{S'}(h)$ has the binomial distribution with parameters m and $L_D(h)$. Therefore, Tonelli's theorem gives

$$\mathbb{E}_{S'\sim D^m}[X_{\mathbb{P}}(S')] = \mathbb{E}_{h\sim\mathbb{P}}\mathbb{E}_{S'\sim D^m}\left[e^{\lambda\Delta(L_{S'}(h),L_D(h))}\right] \le \sup_{p\in[0,1]} \mathbb{E}_{K\sim\mathrm{Bin}(m,p)}\left[e^{\lambda\Delta(K/m,p)}\right] = \mathcal{I}_\Delta(m,\lambda).$$

Substituting this bound into (7.3) proves (7.2). The same event works for every $\mathbb{Q}$, including distributions selected after observing S. □

Lemma 7.4 For every positive integer m,

$$\sup_{p\in[0,1]} \mathbb{E}_{K\sim\mathrm{Bin}(m,p)}\left[e^{m\,\mathrm{kl}(K/m\|p)}\right] \le 2\sqrt{m}.$$

Proof. For $p\in(0,1)$, direct substitution gives

$$\begin{aligned}
&\mathbb{E}_{K\sim\mathrm{Bin}(m,p)}\left[e^{m\,\mathrm{kl}(K/m\|p)}\right] \\
&\quad= \sum_{k=0}^{m}\binom{m}{k}p^k(1-p)^{m-k}\left(\frac{k/m}{p}\right)^k\left(\frac{1-k/m}{1-p}\right)^{m-k} \\
&\quad= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k}.
\end{aligned}$$

The expression is independent of p. Its two endpoint summands equal 1. For $1\le k\le m-1$, the standard two-sided Stirling bounds give

$$\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} \le e^{1/(12m)}\sqrt{\frac{m}{2\pi k(m-k)}}.$$

Moreover,

$$\sum_{k=1}^{m-1} \frac{1}{\sqrt{k(m-k)}} \le \int_0^m \frac{\mathrm{d}x}{\sqrt{x(m-x)}} = \pi.$$

To see the inequality, compare the summands on the left half with the integrals over the unit intervals immediately to their left and those on the right half with the intervals immediately to their right; the integrand decreases up to $m/2$ and increases thereafter. Hence the entire moment is at most

$$2 + e^{1/(12m)}\sqrt{\frac{\pi m}{2}} \le 2\sqrt{m}$$

for $m \ge 8$. Direct substitution verifies the remaining cases $m = 1, \dots, 7$. At $p = 0$ or $p = 1$, the binomial variable is constant and the moment equals 1, so the same upper bound holds. □

Proof of Theorem 6.6. Apply Theorem 7.3 with

$$\Delta(q, p) = \mathrm{kl}(q\|p) \qquad \text{and} \qquad \lambda = m.$$

Binary relative entropy is jointly convex, and Lemma 7.4 gives

$$\mathcal{I}_\Delta(m, m) \le 2\sqrt{m}.$$

Substitution into (7.2) yields (6.3). □

The distribution $\mathbb{P}$ is called the **prior** because it must be selected independently of the sample. The data-dependent distribution $\mathbb{Q}$ is called the **posterior**, even when it is not a Bayesian posterior. The term $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})$ discourages choosing a posterior that departs too sharply from the prior.

Remark 7.5 (Catoni-type bounds) Other choices of the convex deviation function avoid the painful square root in (6.4). For example, a **Catoni bound** yields, for $c > 0$, a bound of the form

$$L_D(\mathbb{Q}) \le \frac{1 - \exp\left(-cL_S(\mathbb{Q}) - \left(\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) + \log(1/\delta)\right)/m\right)}{1 - e^{-c}}$$

simultaneously for every $\mathbb{Q}$. Expanding the exponential shows the characteristic tradeoff between empirical risk and a complexity term proportional to $\mathrm{KL}(\mathbb{Q}\|\mathbb{P})/m$.

This tradeoff leads naturally to a **Gibbs posterior**. For a parameter $\beta > 0$, define

$$\frac{\mathrm{d}\mathbb{Q}_\beta}{\mathrm{d}\mathbb{P}}(h) \coloneqq \frac{\exp(-\beta m L_S(h))}{Z_\beta} \qquad \text{and} \qquad Z_\beta \coloneqq \mathbb{E}_{h\sim\mathbb{P}}\left[\exp(-\beta m L_S(h))\right].$$

For every $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}$,

$$mL_S(\mathbb{Q}) + \frac{1}{\beta}\,\mathrm{KL}(\mathbb{Q}\|\mathbb{P}) = \frac{1}{\beta}\left\{\mathrm{KL}(\mathbb{Q}\|\mathbb{Q}_\beta) - \log Z_\beta\right\}.$$

Consequently, $\mathbb{Q}_\beta$ minimizes the regularized empirical objective on the left. Hypotheses with high empirical risk receive exponentially less mass.

The connection with an ordinary Bayesian posterior becomes exact for the log loss. If

$$L_S(h) = -\frac{1}{m}\sum_{i=1}^{m} \log p_h(Y_i \mid X_i),$$

then, for $\beta = 1$,

$$\exp(-mL_S(h)) = \prod_{i=1}^{m} p_h(Y_i \mid X_i)$$

is the **likelihood**. Multiplying it by the prior and normalizing gives Bayes' rule.

Lecture 8 — Tuesday, November 06, 2018

Linear Classification and the Perceptron

We now return to deterministic hypothesis classes and examine linear classification, sparsity, and a basic algorithm for finding a separating hyperplane.

Definition 8.1 The class of **affine linear scores** on $\mathbb{R}^d$ is

$$\mathcal{L}_d := \left\{h_{\boldsymbol{w},b} : \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R}\right\} \qquad \text{and} \qquad h_{\boldsymbol{w},b}(\boldsymbol{x}) := \langle \boldsymbol{w}, \boldsymbol{x}\rangle + b = \sum_{j=1}^{d} w_j x_j + b.$$

An affine score can be homogenized by augmenting the input:

$$\widetilde{\boldsymbol{x}} := (\boldsymbol{x}, 1) \in \mathbb{R}^{d+1} \qquad \text{and} \qquad \widetilde{\boldsymbol{w}} := (\boldsymbol{w}, b) \in \mathbb{R}^{d+1}.$$

Then

$$\langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}}\rangle = \langle \boldsymbol{w}, \boldsymbol{x}\rangle + b.$$

Thus every affine problem in $\mathbb{R}^d$ can be treated as a homogeneous linear problem in $\mathbb{R}^{d+1}$. Very nice!

Definition 8.2 Define

$$\operatorname{sign}(u) := \begin{cases} +1, & u \geq 0, \\ -1, & u < 0. \end{cases}$$

The class of **affine halfspace classifiers** is

$$\mathcal{H}_d^{\mathrm{HS}} := \left\{ \boldsymbol{x} \mapsto \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) : \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}.$$

Example 8.3 Let $\mathcal{Y} = \{-1, 1\}$ and $d = 2$. What is the classifier represented by a separating line?

The line $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$ divides the plane into two halfspaces. The classifier assigns label 1 to one side and label -1 to the other, as illustrated in Figure 10:

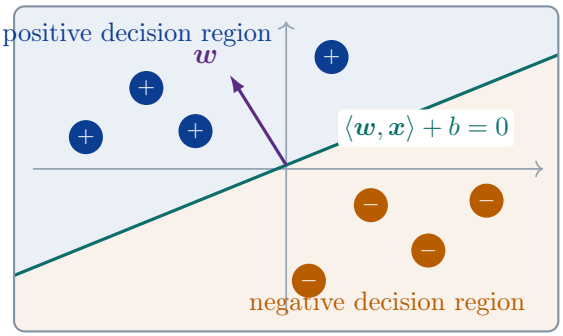

FIGURE 10

The positive decision region is

$$\left\{ \boldsymbol{x} \in \mathbb{R}^2 : \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \geq 0 \right\},$$

and the negative decision region is its complement.

The affine halfspace class satisfies

$$\operatorname{VCdim}(\mathcal{H}_d^{\mathrm{HS}}) = d + 1.$$

The extra dimension is the intercept. Equivalently, it is the dimension introduced by the homoge-

nization in Definition 8.2.

When d is much larger than m, structural restrictions become essential. For example,

$$\boldsymbol{w} = (0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0)$$

is sparse because only a small number k of its coordinates are nonzero. If the support is fixed in advance, the resulting affine halfspace class has VC dimension $k + 1$. If the support may also be selected from the data, the union over possible supports carries an additional model selection cost. Structural risk minimization can express this preference by assigning larger prior weights, and hence smaller penalties, to lower-dimensional sparse models. The tradeoff is deliberate: if the best classifier is not sparse, substantially more data may be required.

In the separable setting, finding an empirical risk minimizer reduces to a linear feasibility problem. A general linear program has the form

$$\max_{\boldsymbol{w} \in \mathbb{R}^d} \langle \boldsymbol{u}, \boldsymbol{w} \rangle \quad \text{subject to} \quad \mathbf{A}\boldsymbol{w} \geq \boldsymbol{v}.$$

For labels $y_i \in \{-1, 1\}$, a vector $\boldsymbol{w}$ separates the sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ exactly when

$$y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0 \quad \text{for every } i \in [m].$$

If such a $\boldsymbol{w}$ exists, define

$$\gamma \coloneqq \min_{i \in [m]} y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0 \qquad \text{and} \qquad \overline{\boldsymbol{w}} \coloneqq \frac{\boldsymbol{w}}{\gamma}.$$

Then $y_i \langle \overline{\boldsymbol{w}}, \boldsymbol{x}_i \rangle \geq 1$ for every i. Let

$$\mathbf{A} \coloneqq \begin{bmatrix} y_1 \boldsymbol{x}_1^{\mathsf{T}} \\ \vdots \\ y_m \boldsymbol{x}_m^{\mathsf{T}} \end{bmatrix}.$$

The margin constraints become

$$\mathbf{A}\boldsymbol{w} = \begin{bmatrix} y_1 \langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle \\ \vdots \\ y_m \langle \boldsymbol{w}, \boldsymbol{x}_m \rangle \end{bmatrix} \geq \mathbf{1}.$$

Thus separation is equivalent to feasibility of the linear program

$$\max_{\boldsymbol{w} \in \mathbb{R}^d} 0 \quad \text{subject to} \quad \mathbf{A}\boldsymbol{w} \geq \mathbf{1}.$$

The **batch Perceptron algorithm** searches for a feasible separator by repeatedly correcting a nonpositive signed margin:

1. Initialize $\boldsymbol{w}^{(0)} = \mathbf{0}$.
2. If there is an $i \in [m]$ such that
$$y_i \langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_i \rangle \le 0,$$
choose one such index and set
$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + y_i \boldsymbol{x}_i.$$
3. If every signed margin is positive, stop and return $\boldsymbol{w}^{(t)}$.

The weak inequality in (2) is essential: at the initialization in (1), every signed margin equals zero. An update improves the signed margin of the selected observation by

$$y_i \langle \boldsymbol{w}^{(t+1)}, \boldsymbol{x}_i \rangle = y_i \langle \boldsymbol{w}^{(t)} + y_i \boldsymbol{x}_i, \boldsymbol{x}_i \rangle = y_i \langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_i \rangle + \|\boldsymbol{x}_i\|_2^2.$$

Accordingly, each update adds or subtracts one observed feature vector. The procedure also admits a stochastic gradient interpretation developed later.

Theorem 8.4 (Perceptron convergence theorem) Assume that $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ is linearly separable, and define

$$B := \inf \left\{ \|\boldsymbol{u}\|_2 : \boldsymbol{u} \in \mathbb{R}^d \text{ and } y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle \ge 1 \text{ for every } i \in [m] \right\}$$

and

$$R := \max_{i \in [m]} \|\boldsymbol{x}_i\|_2.$$

Then the batch Perceptron makes at most $(RB)^2$ updates.

Proof. Fix any $\boldsymbol{u} \in \mathbb{R}^d$ satisfying

$$y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle \ge 1 \quad \text{for every } i \in [m].$$

Suppose the algorithm makes T updates, and denote the vector after those updates by $\boldsymbol{w}^{(T)}$. We use two estimates:

1. The iterates make at least unit progress in the direction $\boldsymbol{u}$ at every update:

$$\langle \boldsymbol{u}, \boldsymbol{w}^{(T)} \rangle \geq T.$$

   Indeed, if observation i causes update t, then

$$\langle \boldsymbol{u}, \boldsymbol{w}^{(t+1)} \rangle - \langle \boldsymbol{u}, \boldsymbol{w}^{(t)} \rangle = \langle \boldsymbol{u}, y_i \boldsymbol{x}_i \rangle = y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle \geq 1.$$

   Summing over the T updates verifies the claim.

2. The squared norm grows by at most R^2 at every update:

$$\left\| \boldsymbol{w}^{(T)} \right\|_2^2 \leq TR^2.$$

   At an update,

$$\left\| \boldsymbol{w}^{(t+1)} \right\|_2^2 = \left\| \boldsymbol{w}^{(t)} + y_i \boldsymbol{x}_i \right\|_2^2 = \left\| \boldsymbol{w}^{(t)} \right\|_2^2 + 2y_i \langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_i \rangle + y_i^2 \|\boldsymbol{x}_i\|_2^2 \leq \left\| \boldsymbol{w}^{(t)} \right\|_2^2 + R^2,$$

   because the selected signed margin is nonpositive and $y_i^2 = 1$. Summing again proves the claim.

Combining (1), the Cauchy–Schwarz inequality, and (2) yields

$$T \leq \langle \boldsymbol{u}, \boldsymbol{w}^{(T)} \rangle \leq \|\boldsymbol{u}\|_2 \left\| \boldsymbol{w}^{(T)} \right\|_2 \leq \|\boldsymbol{u}\|_2 R\sqrt{T}.$$

If $T > 0$, division by $\sqrt{T}$ gives

$$T \leq R^2 \|\boldsymbol{u}\|_2^2.$$

Taking the infimum over all unit margin separators $\boldsymbol{u}$ gives $T \leq (RB)^2$. Equivalently, the cosine between $\boldsymbol{u}$ and $\boldsymbol{w}^{(T)}$ is at least $\sqrt{T}/(R\|\boldsymbol{u}\|_2)$ and cannot exceed one. □

Remark 8.5 The proof does not require the examples to be sampled from a distribution. An adversary may present a new example whenever the current vector is wrong. As long as all presented examples obey the same radius and margin conditions, Theorem 8.4 limits the total number of mistakes to $(RB)^2$.

Regression in a Learning Framework

The definitions of population risk and empirical risk extend directly to arbitrary output spaces and loss functions. For linear regression, take

$$\mathcal{H}_{\text{reg}} \coloneqq \left\{ \boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b : \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}$$

and use squared loss,

$$\ell(h, (\boldsymbol{x}, y)) \coloneqq (h(\boldsymbol{x}) - y)^2.$$

The resulting risks are

$$L_{\mathbb{P}}(h) \coloneqq \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}}\left[(h(\boldsymbol{X}) - Y)^2\right] \qquad \text{and} \qquad L_S(h) \coloneqq \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}_i) - y_i)^2.$$

The latter is the **mean squared error**. Without bounds on the inputs, outputs, parameters, or loss, the unrestricted class need not be PAC learnable. Norm constraints and regularization provide the required control.

After homogenizing the intercept as above, write the empirical risk as

$$L_S(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}\big(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i\big)^2.$$

It is differentiable, with

$$\nabla L_S(\boldsymbol{w}) = \frac{2}{m}\sum_{i=1}^{m}\big(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i\big)\boldsymbol{x}_i.$$

Define

$$\mathbf{A} \coloneqq \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}} \qquad \text{and} \qquad \boldsymbol{b} \coloneqq \sum_{i=1}^{m} y_i \boldsymbol{x}_i.$$

Then the first-order condition $\nabla L_S(\boldsymbol{w}) = \mathbf{0}$ is precisely the normal equation

$$\mathbf{A}\boldsymbol{w} = \boldsymbol{b}.$$

If $\mathbf{A}$ is singular, there may be many empirical minimizers; regularization will make the solution unique.

Binary regression gives another convex example. Modeling the conditional probability $p(\boldsymbol{x}) =$

$\mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ through the log odds,

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b,$$

leads to logistic regression. For labels $y \in \{-1, 1\}$, its individual loss can be written as

$$\ell_{\log}(\boldsymbol{w}, (\boldsymbol{x}, y)) \coloneqq \log\left(1 + e^{-y\langle \boldsymbol{w}, \boldsymbol{x} \rangle}\right),$$

which is convex in $\boldsymbol{w}$.

Thus, in these problems, the empirical risk is a convex function of the parameters. Convexity guarantees that every local minimum is global. Appropriate first-order methods then converge under suitable choices of step size and regularity assumptions. Many finite-dimensional convex problems also admit polynomial-time algorithms under standard encoding or oracle assumptions. By contrast, nonconvex problems may have many local minima, and important classes of global nonconvex optimization problems are NP-hard.

Least squares regression is already convex. **Ridge regression** adds a squared norm penalty, making the objective strongly convex and favoring solutions of smaller norm. Convexity is therefore both an optimization property and a useful organizing principle for learning guarantees.

Definition 8.6 A subset C of a real vector space is **convex** if

$$\alpha \boldsymbol{u} + (1 - \alpha)\boldsymbol{v} \in C$$

for every $\boldsymbol{u}, \boldsymbol{v} \in C$ and every $\alpha \in [0, 1]$.

Euclidean space, open and closed balls, halfspaces, and arbitrary intersections of convex sets are convex.

Definition 8.7 Let C be convex. A function $f \colon C \to \mathbb{R}$ is **convex** if

$$f\big(\alpha \boldsymbol{u} + (1 - \alpha)\boldsymbol{v}\big) \leq \alpha f(\boldsymbol{u}) + (1 - \alpha) f(\boldsymbol{v})$$

for every $\boldsymbol{u}, \boldsymbol{v} \in C$ and every $\alpha \in [0, 1]$. Equivalently, its epigraph

$$\operatorname{epi}(f) \coloneqq \{(\boldsymbol{x}, t) \in C \times \mathbb{R} : t \geq f(\boldsymbol{x})\}$$

is convex.

Proposition 8.8 Every local minimizer of a convex function is a global minimizer.

Proof. Let $\boldsymbol{u}$ be a local minimizer, so that $f(\boldsymbol{u}) \leq f(\boldsymbol{x})$ whenever $\|\boldsymbol{x}-\boldsymbol{u}\|_2 < r$ for some $r > 0$. Fix any $\boldsymbol{v}$ in the domain. Choose $\alpha \in (0,1)$ small enough that

$$\boldsymbol{u} + \alpha(\boldsymbol{v}-\boldsymbol{u}) \in B(\boldsymbol{u}, r).$$

Local minimality and convexity give

$$f(\boldsymbol{u}) \leq f\big((1-\alpha)\boldsymbol{u} + \alpha\boldsymbol{v}\big) \leq (1-\alpha)f(\boldsymbol{u}) + \alpha f(\boldsymbol{v}).$$

Subtracting $(1-\alpha)f(\boldsymbol{u})$ and dividing by α yields $f(\boldsymbol{u}) \leq f(\boldsymbol{v})$. Since $\boldsymbol{v}$ was arbitrary, $\boldsymbol{u}$ is a global minimizer. □

Proposition 8.9 If f is differentiable on an open convex set, then f is convex if and only if

$$f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{u}-\boldsymbol{w}\rangle$$

for every $\boldsymbol{u}$ and $\boldsymbol{w}$ in its domain. If f is twice continuously differentiable, this is equivalent to the Hessian $\nabla^2 f(\boldsymbol{w})$ being positive semidefinite at every $\boldsymbol{w}$.

Proof. *Forward implication.* Assume that f is convex. Fix $\boldsymbol{u}$ and $\boldsymbol{w}$ in its domain. For $t \in (0,1)$, convexity gives

$$f\big(\boldsymbol{w} + t(\boldsymbol{u}-\boldsymbol{w})\big) \leq (1-t)f(\boldsymbol{w}) + tf(\boldsymbol{u}).$$

After subtracting $f(\boldsymbol{w})$ and dividing by t, let t decrease to zero. Differentiability yields

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{u}-\boldsymbol{w}\rangle \leq f(\boldsymbol{u}) - f(\boldsymbol{w}),$$

which is the first-order inequality.

Reverse implication. Assume that the first-order inequality holds. For $\boldsymbol{u}, \boldsymbol{v}$ in the domain and $\alpha \in [0,1]$, set $\boldsymbol{z} = \alpha\boldsymbol{u} + (1-\alpha)\boldsymbol{v}$. Applying the inequality at $\boldsymbol{z}$ first to $\boldsymbol{u}$ and then to $\boldsymbol{v}$ gives

$$f(\boldsymbol{u}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{u}-\boldsymbol{z}\rangle \qquad \text{and} \qquad f(\boldsymbol{v}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{v}-\boldsymbol{z}\rangle.$$

Multiply the first inequality by α, multiply the second by $1-\alpha$, and add. Since

$$\alpha(\boldsymbol{u}-\boldsymbol{z}) + (1-\alpha)(\boldsymbol{v}-\boldsymbol{z}) = 0,$$

the inner product terms cancel, leaving

$$f(\boldsymbol{z}) \le \alpha f(\boldsymbol{u}) + (1-\alpha) f(\boldsymbol{v}).$$

Thus f is convex.

Finally, suppose that f is twice continuously differentiable. If f is convex, then for every $\boldsymbol{w}$ and direction $\boldsymbol{a}$, the one-dimensional function $t \mapsto f(\boldsymbol{w} + t\boldsymbol{a})$ is convex near zero. Its second derivative at zero is therefore nonnegative:

$$\boldsymbol{a}^\top \nabla^2 f(\boldsymbol{w}) \boldsymbol{a} \ge 0.$$

Hence the Hessian is positive semidefinite. Conversely, if every Hessian is positive semidefinite, then the second derivative of $t \mapsto f((1-t)\boldsymbol{u} + t\boldsymbol{v})$ is nonnegative along every line segment in the domain. Each such restriction is convex, which is precisely the convexity of f. □

In one dimension, the conditions in Proposition 8.9 are equivalent to the derivative being nondecreasing and, when the second derivative exists, to $f'' \ge 0$.

Example 8.10 Verify the convexity of the squared loss and the **softplus function** $g(t) = \log\bigl(1 + e^t\bigr)$, and show that composing a convex scalar function with an affine score preserves convexity.

Solution. The second derivatives are

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} t^2 = 2 \qquad \text{and} \qquad g''(t) = \frac{e^t}{(1+e^t)^2},$$

both of which are nonnegative. Hence both scalar functions are convex. If $g \colon \mathbb{R} \to \mathbb{R}$ is convex and

$$f(\boldsymbol{w}) = g(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b),$$

then for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\alpha \in [0,1]$,

$$f\bigl(\alpha\boldsymbol{u} + (1-\alpha)\boldsymbol{v}\bigr) = g\bigl(\alpha(\langle \boldsymbol{u}, \boldsymbol{x} \rangle + b) + (1-\alpha)(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + b)\bigr) \le \alpha f(\boldsymbol{u}) + (1-\alpha) f(\boldsymbol{v}).$$

Taking $g(t) = t^2$ and replacing the intercept by $-y$ proves that

$$\boldsymbol{w} \mapsto (\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y)^2$$

is convex. □

Proposition 8.11 If $f_1, \dots, f_k$ are convex, then their pointwise maximum

$$x \mapsto \max_{i \in [k]} f_i(x)$$

is convex. If $a_1, \dots, a_k$ are nonnegative, then

$$x \mapsto \sum_{i=1}^{k} a_i f_i(x)$$

is also convex.

In particular, an empirical average of convex losses is convex. Thus the least squares empirical risk

$$L_S(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \big(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i \big)^2$$

is convex.

Definition 8.12 A learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ is **convex** if $\mathcal{H} \subseteq \mathbb{R}^d$ is convex and, for every $z \in \mathcal{Z}$, the map

$$\boldsymbol{w} \mapsto \ell(\boldsymbol{w}, \boldsymbol{z})$$

is convex on $\mathcal{H}$.

The domain requirement in Definition 8.12 is essential because convexity is defined along line segments that must remain in the domain. Linear regression and logistic regression are both convex learning problems.

Definition 8.13 Given a regularizer $R \colon \mathbb{R}^d \to [0, \infty)$ and $\lambda > 0$, a **regularized loss minimizer** is

$$\widehat{\boldsymbol{w}}_\lambda \in \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ L_S(\boldsymbol{w}) + \lambda R(\boldsymbol{w}) \right\}.$$

For squared loss with $R(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2$, it is convenient to use the normalization

$$\widehat{\boldsymbol{w}}_\lambda \coloneqq \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ \frac{1}{2m} \sum_{i=1}^{m} \big(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i \big)^2 + \lambda \|\boldsymbol{w}\|_2^2 \right\}. \tag{8.1}$$

This is **ridge regression** (or **Tikhonov regression**). The nested constraint sets

$$\mathcal{H}_r \coloneqq \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2^2 \le r \right\}$$

give the corresponding structural risk minimization viewpoint: the regularizer penalizes movement into larger norm balls.

Using the matrix $\mathbf{A}$ and the vector $\boldsymbol{b}$ defined above, the gradient of the objective in (8.1) is

$$\frac{1}{m}(\mathbf{A}\boldsymbol{w} - \boldsymbol{b}) + 2\lambda \boldsymbol{w}.$$

The unique minimizer therefore satisfies

$$(\mathbf{A} + 2\lambda m \mathbf{I})\widehat{\boldsymbol{w}}_\lambda = \boldsymbol{b}$$

and hence

$$\widehat{\boldsymbol{w}}_\lambda = (\mathbf{A} + 2\lambda m \mathbf{I})^{-1}\boldsymbol{b}. \tag{8.2}$$

For every nonzero $\boldsymbol{v} \in \mathbb{R}^d$,

$$\boldsymbol{v}^\mathsf{T}(\mathbf{A} + 2\lambda m \mathbf{I})\boldsymbol{v} = \sum_{i=1}^{m} \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle^2 + 2\lambda m \|\boldsymbol{v}\|_2^2 > 0.$$

Thus $\mathbf{A} + 2\lambda m \mathbf{I}$ is positive definite and invertible, even when the unregularized design matrix $\mathbf{A}$ is singular. Besides controlling statistical complexity, ridge regularization improves numerical stability.

The next objective is to turn this stability into a learning guarantee. Under boundedness assumptions such as $\|\boldsymbol{w}\|_2 \le B$ and $Y \in [-1, 1]$, the goal is an expected excess risk bound of the form

$$\mathbb{E}_{S \sim D^m}\left[L_D(A(S)) \right] \le \inf_{\boldsymbol{w} \in \mathcal{H}} L_D(\boldsymbol{w}) + \varepsilon,$$

with a sample size depending on the accuracy and norm bounds rather than directly on the ambient dimension.

Lecture 9 — Tuesday, November 13, 2018

Stable Rules Don't Overfit

Notation 9.1 Let $S = (z_1, \dots, z_m) \sim D^m$, let $z' \sim D$ be independent of S, and let I be uniformly distributed on $[m]$, independently of everything else. Write

$$S^{I \leftarrow z'} \coloneqq (z_1, \dots, z_{I-1}, z', z_{I+1}, \dots, z_m)$$

for the sample obtained by replacing z_I with z'.

Theorem 9.2 For every learning rule A for which all of the displayed expectations are finite,

$$\mathbb{E}_{S \sim D^m}\left[L_D(A(S)) - L_S(A(S))\right] = \mathbb{E}_{S,z',I}\left[\ell\big(A(S^{I \leftarrow z'}), z_I\big) - \ell(A(S), z_I)\right]. \tag{9.1}$$

Proof. Introducing an independent observation and a uniform index gives

$$\mathbb{E}_S\left[L_D(A(S)) - L_S(A(S))\right] = \mathbb{E}_{S,z',I}\left[\ell(A(S), z') - \ell(A(S), z_I)\right].$$

The pair (z_I, z') is exchangeable. Swapping these two observations transforms (S, z') into $(S^{I \leftarrow z'}, z_I)$ without changing their joint distribution. Consequently,

$$\mathbb{E}_{S,z',I}[\ell(A(S), z')] = \mathbb{E}_{S,z',I}\left[\ell\big(A(S^{I \leftarrow z'}), z_I\big)\right],$$

which proves (9.1). □

Definition 9.3 A learning rule A is **on-average replace-one stable** with rate $\varepsilon \colon \mathbb{N} \to [0, \infty)$ if

$$\mathbb{E}_{S,z',I}\left[\ell\big(A(S^{I \leftarrow z'}), z_I\big) - \ell(A(S), z_I)\right] \le \varepsilon(m)$$

for every positive integer m.

By Theorem 9.2, the expected generalization gap is exactly the replace-one stability quantity.

Hence

$$\underbrace{\mathbb{E}[L_D(A(S))]}_{\text{expected risk}} = \underbrace{\mathbb{E}[L_S(A(S))]}_{\text{expected empirical risk}} + \underbrace{\mathbb{E}[L_D(A(S)) - L_S(A(S))]}_{\text{stability term}}.$$

A constant learning rule has perfect stability but may have poor empirical risk. Successful learning requires both terms to be small.

We study the **Tikhonov regularization rule**

$$A_\lambda(S) \coloneqq \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ L_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2 \right\}, \quad \lambda > 0.$$

We will leave λ unspecified for now. The curvature introduced by the squared norm makes the output insensitive to replacing one observation.

Definition 9.4 Let $C \subseteq \mathbb{R}^d$ be convex. A function $f\colon C \to \mathbb{R}$ is **μ-strongly convex** if

$$\alpha f(\boldsymbol{u}) + (1-\alpha) f(\boldsymbol{w}) - f\big(\alpha\boldsymbol{u} + (1-\alpha)\boldsymbol{w}\big) \geq \frac{\mu}{2}\alpha(1-\alpha)\|\boldsymbol{u}-\boldsymbol{w}\|_2^2$$

for every $\boldsymbol{u}, \boldsymbol{w} \in C$ and every $\alpha \in [0,1]$.

Strong convexity quantifies the vertical gap between the chord joining $(\boldsymbol{u}, f(\boldsymbol{u}))$ and $(\boldsymbol{w}, f(\boldsymbol{w}))$ and the graph of f, as shown in Figure 11.

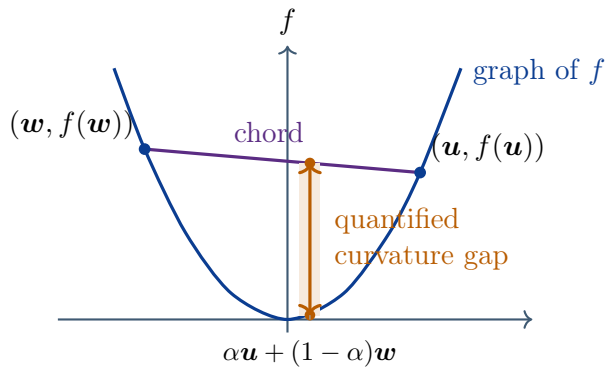

FIGURE 11

Proposition 9.5 The following statements hold:

1. The function

$$\boldsymbol{w} \mapsto \frac{\mu}{2}\|\boldsymbol{w}\|_2^2$$

is μ-strongly convex.

2. If f is μ-strongly convex and $c > 0$, then cf is $c\mu$-strongly convex.
3. If f is μ-strongly convex and g is convex, then $f + g$ is μ-strongly convex.

Proof. For (1), set $\overline{\alpha} = 1 - \alpha$. Expanding the squared norm gives

$$\begin{aligned}
&\alpha\frac{\mu}{2}\|\boldsymbol{u}\|_2^2 + \overline{\alpha}\frac{\mu}{2}\|\boldsymbol{w}\|_2^2 - \frac{\mu}{2}\|\alpha\boldsymbol{u} + \overline{\alpha}\boldsymbol{w}\|_2^2 \\
&\quad = \frac{\mu}{2}\alpha\overline{\alpha}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{w}\|_2^2 - 2\langle\boldsymbol{u}, \boldsymbol{w}\rangle\right) \\
&\quad = \frac{\mu}{2}\alpha(1-\alpha)\|\boldsymbol{u} - \boldsymbol{w}\|_2^2.
\end{aligned}$$

This is equality in Definition 9.4. Multiplying the defining inequality by c proves (2). Adding the convexity inequality for g to the strong convexity inequality for f proves (3). □

Proposition 9.6 If f is μ-strongly convex and $\boldsymbol{u}$ minimizes f, then

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \geq \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{u}\|_2^2$$

for every $\boldsymbol{w}$ in the domain.

Proof. For $\alpha \in (0, 1)$, minimality of $\boldsymbol{u}$ and strong convexity give

$$f(\boldsymbol{u}) \leq f\big((1-\alpha)\boldsymbol{u} + \alpha\boldsymbol{w}\big) \leq (1-\alpha)f(\boldsymbol{u}) + \alpha f(\boldsymbol{w}) - \frac{\mu}{2}\alpha(1-\alpha)\|\boldsymbol{w} - \boldsymbol{u}\|_2^2.$$

After rearranging and dividing by α,

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \geq \frac{\mu}{2}(1-\alpha)\|\boldsymbol{w} - \boldsymbol{u}\|_2^2.$$

Letting α decrease to zero proves the result. □

Define

$$F_S(\boldsymbol{w}) \coloneqq L_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2.$$

By (1) and (3), F_S is 2λ-strongly convex. Let

$$\boldsymbol{u} \coloneqq A_\lambda(S) \qquad \text{and} \qquad \boldsymbol{v} \coloneqq A_\lambda(S^{I \leftarrow z'}).$$

The quadratic growth bound in Proposition 9.6 gives

$$\lambda\|\boldsymbol{v}-\boldsymbol{u}\|_2^2 \le F_S(\boldsymbol{v}) - F_S(\boldsymbol{u}). \tag{9.2}$$

Only one summand of the empirical risk changes under replacement, so

$$F_S(\boldsymbol{v}) - F_S(\boldsymbol{u}) = F_{S^{I\leftarrow z'}}(\boldsymbol{v}) - F_{S^{I\leftarrow z'}}(\boldsymbol{u}) + \frac{\ell(\boldsymbol{v},z_I) - \ell(\boldsymbol{u},z_I) + \ell(\boldsymbol{u},z') - \ell(\boldsymbol{v},z')}{m}.$$

Since $\boldsymbol{v}$ minimizes $F_{S^{I\leftarrow z'}}$, the first difference is nonpositive. Combining this fact with (9.2) yields

$$\lambda\|\boldsymbol{v}-\boldsymbol{u}\|_2^2 \le \frac{\ell(\boldsymbol{v},z_I) - \ell(\boldsymbol{u},z_I) + \ell(\boldsymbol{u},z') - \ell(\boldsymbol{v},z')}{m}. \tag{9.3}$$

Definition 9.7 The loss ℓ is **ρ-Lipschitz** in its parameter if

$$|\ell(\boldsymbol{v},z) - \ell(\boldsymbol{u},z)| \le \rho\|\boldsymbol{v}-\boldsymbol{u}\|_2$$

for every $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d$ and every $z\in\mathcal{Z}$.

Geometrically, a one-dimensional ρ-Lipschitz function cannot leave the cone of slopes $\pm\rho$ based at any point of its graph. This is illustrated for $\rho = 1/2$ in Figure 12.

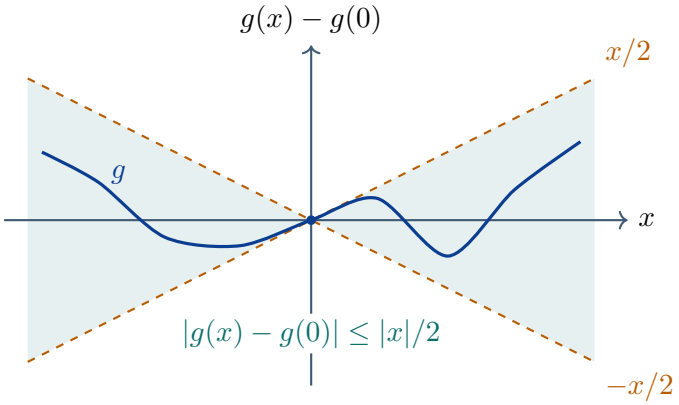

FIGURE 12

Theorem 9.8 Assume that $\boldsymbol{w}\mapsto\ell(\boldsymbol{w},z)$ is convex and ρ-Lipschitz for every $z\in\mathcal{Z}$. Then the rule

$$A_\lambda(S) = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d}\left\{L_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2\right\}$$

is on-average replace-one stable with rate

$$\varepsilon(m) = \frac{2\rho^2}{\lambda m}.$$

Equivalently,

$$\mathbb{E}_{S\sim D^m}\left[L_D(A_\lambda(S)) - L_S(A_\lambda(S))\right] \leq \frac{2\rho^2}{\lambda m}.$$

Proof. Apply the Lipschitz condition in Definition 9.7 to the two loss differences on the right-hand side of (9.3). This gives

$$\lambda\|\boldsymbol{v}-\boldsymbol{u}\|_2^2 \leq \frac{2\rho}{m}\|\boldsymbol{v}-\boldsymbol{u}\|_2.$$

If $\boldsymbol{u} = \boldsymbol{v}$, the desired parameter bound is immediate. Otherwise, division by $\lambda\|\boldsymbol{v}-\boldsymbol{u}\|_2$ gives

$$\|\boldsymbol{v}-\boldsymbol{u}\|_2 \leq \frac{2\rho}{\lambda m}. \tag{9.4}$$

Applying the Lipschitz condition once more at z_I yields

$$\ell(\boldsymbol{v}, z_I) - \ell(\boldsymbol{u}, z_I) \leq \rho\|\boldsymbol{v}-\boldsymbol{u}\|_2 \leq \frac{2\rho^2}{\lambda m}.$$

Taking expectations and using Theorem 9.2 proves both claims. □

3. Optimization Methods for Learning

Smoothness and Convex Learning

Definition 9.9 A differentiable function $g\colon \mathbb{R}^d \to \mathbb{R}$ is **β-smooth** if its gradient is β-Lipschitz:

$$\|\nabla g(\boldsymbol{u}) - \nabla g(\boldsymbol{w})\|_2 \leq \beta\|\boldsymbol{u}-\boldsymbol{w}\|_2$$

for every $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^d$.

Theorem 9.10 (Descent lemma) If g is β-smooth, then

$$g(\boldsymbol{v}) \leq g(\boldsymbol{w}) + \langle \nabla g(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle + \frac{\beta}{2} \|\boldsymbol{v} - \boldsymbol{w}\|_2^2$$

for every $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$. If g is also convex, then

$$g(\boldsymbol{w}) + \langle \nabla g(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle \leq g(\boldsymbol{v})$$

as well.

Proof. Let $\boldsymbol{\gamma}(t) = \boldsymbol{w} + t(\boldsymbol{v} - \boldsymbol{w})$. By the fundamental theorem of calculus,

$$g(\boldsymbol{v}) - g(\boldsymbol{w}) = \int_0^1 \langle \nabla g(\boldsymbol{\gamma}(t)), \boldsymbol{v} - \boldsymbol{w} \rangle \, \mathrm{d}t = \langle \nabla g(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle + \int_0^1 \langle \nabla g(\boldsymbol{\gamma}(t)) - \nabla g(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle \, \mathrm{d}t.$$

The Cauchy–Schwarz inequality and smoothness bound the integrand in absolute value by

$$\beta t \|\boldsymbol{v} - \boldsymbol{w}\|_2^2.$$

Integration gives the upper bound. The lower bound is the first-order characterization of convexity in Proposition 8.9. □

Proposition 9.11 If g is nonnegative and β-smooth for some $\beta > 0$, then

$$\|\nabla g(\boldsymbol{w})\|_2^2 \leq 2\beta g(\boldsymbol{w})$$

for every $\boldsymbol{w} \in \mathbb{R}^d$.

Proof. Apply Theorem 9.10 with

$$\boldsymbol{v} = \boldsymbol{w} - \frac{1}{\beta} \nabla g(\boldsymbol{w}).$$

Then

$$0 \leq g(\boldsymbol{v}) \leq g(\boldsymbol{w}) - \frac{1}{2\beta} \|\nabla g(\boldsymbol{w})\|_2^2.$$

Rearranging proves the claim. □

Smoothness can therefore replace a global Lipschitz assumption in refined stability arguments: the gradient, and hence the local Lipschitz behavior, is controlled by the loss value itself.

We now combine empirical optimality with the stability bound from Theorem 9.8.

Corollary 9.12 Assume that each loss is convex and ρ-Lipschitz. For every comparator $\boldsymbol{w}^* \in \mathbb{R}^d$,

$$\mathbb{E}_{S\sim D^m}\left[L_D(A_\lambda(S))\right] \le L_D(\boldsymbol{w}^*) + \lambda\|\boldsymbol{w}^*\|_2^2 + \frac{2\rho^2}{\lambda m}.$$

Proof. Optimality of $A_\lambda(S)$ gives

$$L_S(A_\lambda(S)) + \lambda\|A_\lambda(S)\|_2^2 \le L_S(\boldsymbol{w}^*) + \lambda\|\boldsymbol{w}^*\|_2^2.$$

Dropping the nonnegative regularization term on the left and taking expectations yields the **oracle inequality**

$$\mathbb{E}_S[L_S(A_\lambda(S))] \le L_D(\boldsymbol{w}^*) + \lambda\|\boldsymbol{w}^*\|_2^2.$$

Adding the expected generalization gap bound from Theorem 9.8 proves the result. □

Corollary 9.13 Suppose that the learning problem is convex, every loss is ρ-Lipschitz, and

$$\mathcal{H} \subseteq \left\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \le B\right\}.$$

Assume $B > 0$ and $\rho > 0$. (We say that the learning problem is **convex-Lipschitz-bounded**.) Choose

$$\lambda := \sqrt{\frac{2\rho^2}{B^2 m}}.$$

Then

$$\mathbb{E}_{S\sim D^m}\left[L_D(A_\lambda(S))\right] \le \inf_{\boldsymbol{w}\in\mathcal{H}} L_D(\boldsymbol{w}) + \rho B\sqrt{\frac{8}{m}}.$$

Proof. Apply Corollary 9.12 to any $\boldsymbol{w}^* \in \mathcal{H}$ and use $\|\boldsymbol{w}^*\|_2 \le B$. The two error terms are

$$\lambda B^2 \qquad \text{and} \qquad \frac{2\rho^2}{\lambda m}.$$

The displayed choice of λ balances them, and their sum is $\rho B\sqrt{8/m}$. Taking the infimum over $\boldsymbol{w}^*$ proves the result. □

Consequently, expected excess risk at most ε is guaranteed whenever

$$m \geq \frac{8\rho^2 B^2}{\varepsilon^2}.$$

This rate is independent of the ambient dimension. Smooth, nonnegative losses can be treated through the self-boundedness property in Proposition 9.11; the central requirement is enough control of how individual losses change with the parameter.

Lecture 10 — Tuesday, November 20, 2018

Gradient Descent

Gradient descent is *very* important. Essentially all of artificial intelligence rests on the shoulders of gradient descent. It is the workhorse of deep learning.

Definition 10.1 Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be differentiable. Given an initial point $\boldsymbol{w}^{(1)}$ and a step size $\eta > 0$, **gradient descent** generates

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla f(\boldsymbol{w}^{(t)}), \quad t = 1, 2, \ldots.$$

For a convex objective, the negative gradient is a direction of local decrease, and every local minimum is global by Proposition 8.8. Figure 13 depicts level sets of a convex bowl over a bounded feasible region.

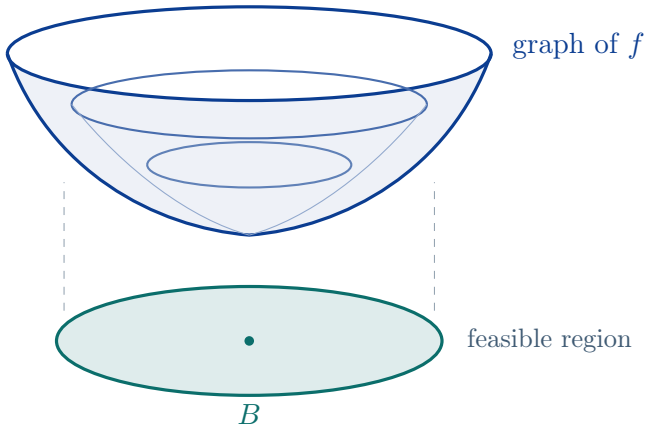

Figure 13

If the update rolls downhill, we can be sure that the objective decreases at each step. Recall that

$$f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{u} - \boldsymbol{w} \rangle$$

from a first-order Taylor expansion. However, we do not want to rely on the linear approximation too much, so we impose a penalty:

$$\boldsymbol{w}^{(t+1)} = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ \eta \left[f(\boldsymbol{w}^{(t)}) + \langle \nabla f(\boldsymbol{w}^{(t)}), \boldsymbol{w} - \boldsymbol{w}^{(t)} \rangle \right] + \frac{1}{2} \left\| \boldsymbol{w} - \boldsymbol{w}^{(t)} \right\|_2^2 \right\}.$$

The linear term encourages descent, whereas the squared distance term prevents the new iterate from relying too heavily on a local approximation. Differentiating this auxiliary objective and setting its gradient to zero recovers the update in Definition 10.1:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla f(\boldsymbol{w}^{(t)}).$$

Stochastic Gradient Descent

Stochastic gradient descent replaces the exact gradient by a random direction $\boldsymbol{V}_t$ whose conditional expectation is a gradient:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{V}_t \qquad \text{and} \qquad \mathbb{E}[\boldsymbol{V}_t \mid \boldsymbol{w}^{(t)}] = \nabla f(\boldsymbol{w}^{(t)}).$$

This makes each iteration inexpensive when f is itself an expectation or a large empirical average.

The method is also widely used for nonconvex objectives, such as the loss associated with a multilayer neural network

$$\boldsymbol{x} \mapsto \sigma\big(\mathbf{W}_3 \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \boldsymbol{x}))\big).$$

Nonconvex objectives may have several local minima and saddle points, as suggested by Figure 14. Stochasticity can help an algorithm move away from some unstable saddle points, but convex convergence guarantees do not automatically extend to this setting.

Convex objectives need not be differentiable. The absolute value penalty used in least absolute shrinkage and selection, for example, is not differentiable at zero. Subgradients extend the first-order inequality to this setting.

Definition 10.2 Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be convex. A vector $\boldsymbol{v} \in \mathbb{R}^d$ is a **subgradient** of f at $\boldsymbol{w}$

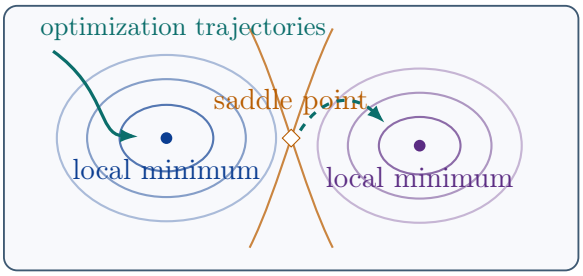

FIGURE 14

if

$$f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{w} \rangle$$

for every $\boldsymbol{u} \in \mathbb{R}^d$. The set of all subgradients at $\boldsymbol{w}$ is the **subdifferential**

$$\partial f(\boldsymbol{w}) \coloneqq \left\{ \boldsymbol{v} \in \mathbb{R}^d : f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{w} \rangle \text{ for every } \boldsymbol{u} \right\}.$$

At a differentiability point, $\partial f(\boldsymbol{w}) = \{\nabla f(\boldsymbol{w})\}$. At a corner, several supporting slopes may be valid. For instance,

$$\partial |x| \big|_{x=0} = [-1, 1],$$

as depicted in Figure 15.

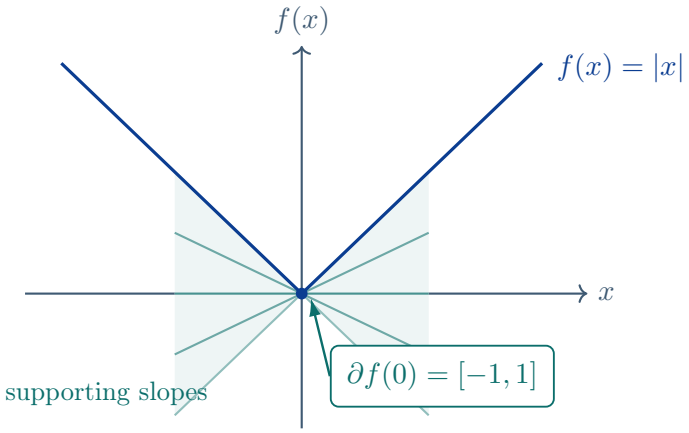

FIGURE 15

Proposition 10.3 Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then the following statements hold:

1. If f is differentiable, it is ρ-Lipschitz if and only if

$$\|\nabla f(\boldsymbol{w})\|_2 \leq \rho \quad \text{for every } \boldsymbol{w} \in \mathbb{R}^d.$$

2. In general, f is ρ-Lipschitz if and only if

$$\|\boldsymbol{v}\|_2 \leq \rho \quad \text{for every } \boldsymbol{w} \in \mathbb{R}^d \text{ and every } \boldsymbol{v} \in \partial f(\boldsymbol{w}).$$

Proof. It suffices to prove (2); statement (1) follows because the subdifferential of a differentiable convex function is the singleton containing its gradient.

Reverse implication. Suppose first that all subgradients have norm at most ρ. If $\boldsymbol{v} \in \partial f(\boldsymbol{u})$, then

$$f(\boldsymbol{u}) - f(\boldsymbol{w}) \leq \langle \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{w} \rangle \leq \rho \|\boldsymbol{u} - \boldsymbol{w}\|_2.$$

Interchanging $\boldsymbol{u}$ and $\boldsymbol{w}$ gives the corresponding lower bound, so f is ρ-Lipschitz.

Forward implication. Suppose that f is ρ-Lipschitz, and take a nonzero $\boldsymbol{v} \in \partial f(\boldsymbol{w})$. For $t > 0$, set

$$\boldsymbol{u} = \boldsymbol{w} + t\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}.$$

The subgradient inequality and the Lipschitz condition imply

$$t\|\boldsymbol{v}\|_2 \leq f(\boldsymbol{u}) - f(\boldsymbol{w}) \leq \rho t.$$

Thus $\|\boldsymbol{v}\|_2 \leq \rho$. The zero subgradient satisfies the same bound as well. □

Figure 16 illustrates that a Lipschitz constant restricts the slopes of all supporting lines.

Theorem 10.4 Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex and ρ-Lipschitz. Suppose $\boldsymbol{w}^*$ minimizes f and $\|\boldsymbol{w}^*\|_2 \leq B$. Initialize $\boldsymbol{w}^{(1)} = \boldsymbol{0}$, choose

$$\boldsymbol{v}_t \in \partial f(\boldsymbol{w}^{(t)}),$$

and update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{v}_t.$$

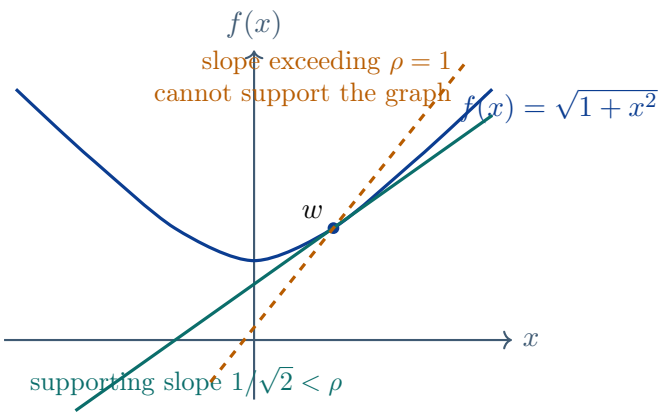

FIGURE 16

For the averaged iterate

$$\overline{\boldsymbol{w}}_T \coloneqq \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}^{(t)},$$

one has

$$f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*) \le \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}. \tag{10.1}$$

In particular, if $B > 0$ and $\rho > 0$, choosing $\eta = B/(\rho\sqrt{T})$ gives

$$f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*) \le \frac{B\rho}{\sqrt{T}}.$$

Proof. Jensen's inequality and the subgradient inequality give

$$f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*) \le \frac{1}{T}\sum_{t=1}^{T}\big(f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*)\big) \le \frac{1}{T}\sum_{t=1}^{T}\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t\rangle. \tag{10.2}$$

The update identity gives

$$\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\right\|_2^2 = \left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^* - \eta\boldsymbol{v}_t\right\|_2^2 = \left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\right\|_2^2 - 2\eta\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t\rangle + \eta^2\|\boldsymbol{v}_t\|_2^2.$$

Rearranging and summing over t yields the telescoping identity

$$\sum_{t=1}^{T}\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t\rangle = \frac{\left\|\boldsymbol{w}^{(1)} - \boldsymbol{w}^*\right\|_2^2 - \left\|\boldsymbol{w}^{(T+1)} - \boldsymbol{w}^*\right\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|_2^2.$$

By (2), $\|\boldsymbol{v}_t\|_2 \le \rho$. Moreover, $\boldsymbol{w}^{(1)} = \mathbf{0}$ and $\|\boldsymbol{w}^*\|_2 \le B$. Dropping the nonpositive terminal term and substituting into (10.2) proves (10.1).

The function of η on the right-hand side of (10.1) is minimized at

$$\eta = \frac{B}{\rho\sqrt{T}},$$

which gives the final bound. Thus accuracy ε is achieved after at most $(B\rho/\varepsilon)^2$ iterations. □

Theorem 10.5 Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex, let $\boldsymbol{w}^*$ minimize f, and suppose $\|\boldsymbol{w}^*\|_2 \le B$. Let $\mathcal{F}_{t-1}$ contain the randomness revealed before iteration t. Starting from $\boldsymbol{w}^{(1)} = \mathbf{0}$, suppose

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{V}_t,$$

where

$$\mathbb{E}[\boldsymbol{V}_t \mid \mathcal{F}_{t-1}] \in \partial f(\boldsymbol{w}^{(t)}) \qquad \text{and} \qquad \mathbb{E}[\|\boldsymbol{V}_t\|_2^2 \mid \mathcal{F}_{t-1}] \le \rho^2.$$

Then

$$\mathbb{E}\left[f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*)\right] \le \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}.$$

If $B > 0$ and $\rho > 0$, taking $\eta = B/(\rho\sqrt{T})$ makes the right-hand side $B\rho/\sqrt{T}$.

Proof. Let

$$\overline{\boldsymbol{V}}_t \coloneqq \mathbb{E}[\boldsymbol{V}_t \mid \mathcal{F}_{t-1}].$$

Since $\boldsymbol{w}^{(t)}$ is measurable with respect to $\mathcal{F}_{t-1}$ and $\overline{\boldsymbol{V}}_t \in \partial f(\boldsymbol{w}^{(t)})$,

$$f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*) \le \langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \overline{\boldsymbol{V}}_t \rangle.$$

The tower property gives

$$\mathbb{E}\left[\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{V}_t \rangle\right] = \mathbb{E}\left[\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \overline{\boldsymbol{V}}_t \rangle\right].$$

Taking expectations in the same potential identity used in the proof of Theorem 10.4, summing over t, and using the conditional second moment bound yields

$$\sum_{t=1}^{T} \mathbb{E}\left[f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*)\right] \le \frac{B^2}{2\eta} + \frac{\eta T \rho^2}{2}.$$

Jensen's inequality for the averaged iterate completes the proof. □

Corollary 10.6 Let

$$L_D(\boldsymbol{w}) = \mathbb{E}_{Z\sim D}[\ell(\boldsymbol{w}, Z)],$$

where $\boldsymbol{w} \mapsto \ell(\boldsymbol{w}, z)$ is convex and ρ-Lipschitz for every z. At iteration t, draw $Z_t \sim D$ independently, choose

$$\boldsymbol{V}_t \in \partial\ell(\boldsymbol{w}^{(t)}, Z_t),$$

and make the stochastic subgradient update. Assume $B > 0$ and $\rho > 0$. Then

$$\mathbb{E}[L_D(\overline{\boldsymbol{w}}_T)] \leq \inf_{\|\boldsymbol{w}\|_2 \leq B} L_D(\boldsymbol{w}) + \frac{B\rho}{\sqrt{T}}$$

when $\eta = B/(\rho\sqrt{T})$.

Proof. For every comparator $\boldsymbol{u}$, the pointwise subgradient inequality gives

$$\ell(\boldsymbol{u}, Z_t) - \ell(\boldsymbol{w}^{(t)}, Z_t) \geq \langle \boldsymbol{V}_t, \boldsymbol{u} - \boldsymbol{w}^{(t)} \rangle.$$

Taking the conditional expectation over Z_t shows that

$$\mathbb{E}[\boldsymbol{V}_t \mid \mathcal{F}_{t-1}] \in \partial L_D(\boldsymbol{w}^{(t)}).$$

Moreover, (2) gives $\|\boldsymbol{V}_t\|_2 \leq \rho$. The result follows from Theorem 10.5. □

Thus expected excess risk at most ε is obtained after

$$T \geq \left(\frac{B\rho}{\varepsilon}\right)^2$$

independent observations. If Lipschitz control is guaranteed only on a constrained region, the iterates should be projected back into that region.

Definition 10.7 For a nonempty closed convex set $\mathcal{H} \subseteq \mathbb{R}^d$, the **Euclidean projection** of $\boldsymbol{x}$ onto $\mathcal{H}$ is

$$\Pi_{\mathcal{H}}(\boldsymbol{x}) \coloneqq \arg\min_{\boldsymbol{w}\in\mathcal{H}} \|\boldsymbol{w} - \boldsymbol{x}\|_2.$$

Projected stochastic subgradient descent uses the two-stage update

$$\boldsymbol{w}^{(t+1/2)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{V}_t \qquad \text{and} \qquad \boldsymbol{w}^{(t+1)} = \Pi_{\mathcal{H}}(\boldsymbol{w}^{(t+1/2)}).$$

The geometry of the projection step is shown in Figure 17.

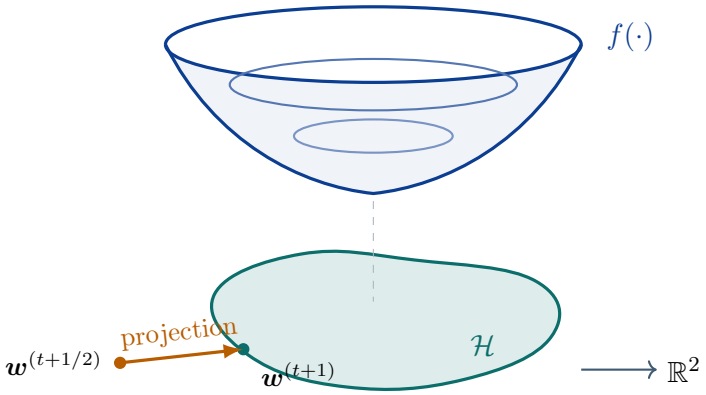

FIGURE 17

Proposition 10.8 For every $\boldsymbol{x} \in \mathbb{R}^d$ and every $\boldsymbol{u} \in \mathcal{H}$,

$$\|\Pi_{\mathcal{H}}(\boldsymbol{x}) - \boldsymbol{u}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{u}\|_2.$$

Proof. Let $\boldsymbol{p} = \Pi_{\mathcal{H}}(\boldsymbol{x})$. The first-order optimality condition for projection onto a closed convex set is

$$\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{u} - \boldsymbol{p} \rangle \leq 0 \quad \text{for every } \boldsymbol{u} \in \mathcal{H}.$$

Therefore,

$$\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 = \|\boldsymbol{x} - \boldsymbol{p}\|_2^2 + \|\boldsymbol{p} - \boldsymbol{u}\|_2^2 + 2\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{p} - \boldsymbol{u} \rangle \geq \|\boldsymbol{p} - \boldsymbol{u}\|_2^2.$$

Taking square roots proves the claim. □

By Proposition 10.8, projection can only improve the distance term used in the potential argument. The convergence bounds in Theorems 10.4 and 10.5 therefore remain valid for projected iterates with comparators in $\mathcal{H}$.

4. Margin and Instance-Based Learning

Support Vector Machines

For labels $y_i \in \{-1, 1\}$, an affine classifier predicts

$$h_{\boldsymbol{w},b}(\boldsymbol{x}) \coloneqq \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b).$$

It classifies a sample correctly exactly when

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0 \quad \text{for every } i \in [m].$$

Among all separating hyperplanes, a **support vector machine** selects one with a large distance from the closest observation.

Lecture 11 — Tuesday, November 27, 2018

Lemma 11.1 Let $\boldsymbol{w} \neq \boldsymbol{0}$ and

$$H_{\boldsymbol{w},b} \coloneqq \left\{ \boldsymbol{u} \in \mathbb{R}^d : \langle \boldsymbol{w}, \boldsymbol{u} \rangle + b = 0 \right\}.$$

Then

$$\operatorname{dist}(\boldsymbol{x}, H_{\boldsymbol{w},b}) = \frac{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b|}{\|\boldsymbol{w}\|_2}.$$

Proof. Set

$$\boldsymbol{u}^* \coloneqq \boldsymbol{x} - \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b}{\|\boldsymbol{w}\|_2^2} \boldsymbol{w}.$$

Then $\langle \boldsymbol{w}, \boldsymbol{u}^* \rangle + b = 0$, so $\boldsymbol{u}^* \in H_{\boldsymbol{w},b}$, and

$$\|\boldsymbol{x} - \boldsymbol{u}^*\|_2 = \frac{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b|}{\|\boldsymbol{w}\|_2}.$$

For any $\boldsymbol{u} \in H_{\boldsymbol{w},b}$, the Cauchy–Schwarz inequality gives

$$|\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b| = |\langle \boldsymbol{w}, \boldsymbol{x} - \boldsymbol{u} \rangle| \leq \|\boldsymbol{w}\|_2 \|\boldsymbol{x} - \boldsymbol{u}\|_2.$$

Thus no point of the hyperplane is closer than $\boldsymbol{u}^*$. □

Definition 11.2 The **functional margin** of $(\boldsymbol{w}, b)$ on observation $(\boldsymbol{x}_i, y_i)$ is

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b).$$

Its **geometric margin** is

$$\frac{y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)}{\|\boldsymbol{w}\|_2}.$$

For a correctly classified sample, both quantities are positive.

The functional margin changes under rescaling of $(\boldsymbol{w}, b)$, whereas the geometric margin does not. Maximizing the smallest geometric margin favors a separator that remains valid under moderate perturbations of the observations, as suggested by Figure 18.

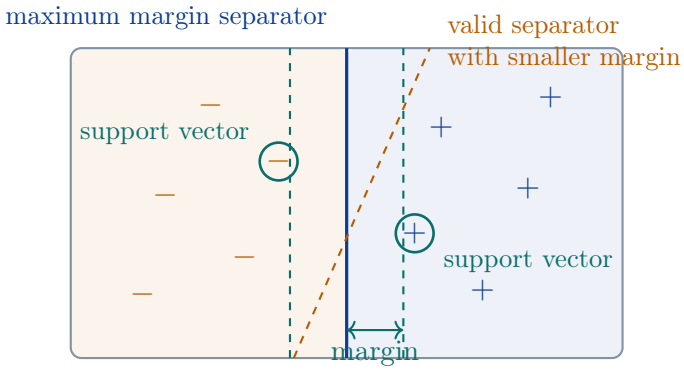

FIGURE 18

Hard Margin Support Vector Machines

The **hard margin setting** assumes that the sample is linearly separable. By Lemma 11.1, the maximum margin problem is

$$\underset{(\boldsymbol{w},b):\, \|\boldsymbol{w}\|_2=1}{\text{maximize}} \; \min_{i \in [m]} y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b), \tag{11.1}$$

where only separating pairs $(\boldsymbol{w}, b)$ are feasible.

Proposition 11.3 Assume that the sample is linearly separable and contains both labels. Problem (11.1) is equivalent, up to positive rescaling, to the quadratic program

$$\underset{\boldsymbol{w}\in\mathbb{R}^d,\, b\in\mathbb{R}}{\text{minimize}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \quad \text{subject to} \quad y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq 1 \text{ for every } i \in [m]. \tag{11.2}$$

If $(\boldsymbol{w}_0, b_0)$ solves (11.2), then

$$\widehat{\boldsymbol{w}} = \frac{\boldsymbol{w}_0}{\|\boldsymbol{w}_0\|_2} \qquad \text{and} \qquad \widehat{b} = \frac{b_0}{\|\boldsymbol{w}_0\|_2}$$

solves (11.1), and its geometric margin is $1/\|\boldsymbol{w}_0\|_2$.

Proof. *Forward implication.* Let $(\boldsymbol{w}^*, b^*)$ solve (11.1), and write

$$\gamma^* := \min_{i\in[m]} y_i(\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + b^*) > 0.$$

Because $\|\boldsymbol{w}^*\|_2 = 1$, the rescaled pair

$$\left(\frac{\boldsymbol{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*}\right)$$

is feasible for (11.2) and has norm $1/\gamma^*$. If $(\boldsymbol{w}_0, b_0)$ minimizes (11.2), then

$$\|\boldsymbol{w}_0\|_2 \leq \frac{1}{\gamma^*}. \tag{11.3}$$

On the other hand, normalizing the feasible pair $(\boldsymbol{w}_0, b_0)$ produces a unit normal with margin at least $1/\|\boldsymbol{w}_0\|_2$. Optimality of γ^* in (11.1) therefore gives

$$\frac{1}{\|\boldsymbol{w}_0\|_2} \leq \gamma^*. \tag{11.4}$$

Equations (11.3) and (11.4) are equalities. Hence the normalized quadratic program solution solves the unit normal problem.

Reverse implication. Let $(\boldsymbol{w}_0, b_0)$ solve (11.2). The preceding normalization gives a feasible unit normal separator of margin $1/\|\boldsymbol{w}_0\|_2$. If another unit normal pair had larger margin γ, scaling it by $1/\gamma$ would produce a feasible point of (11.2) with norm strictly smaller than $\|\boldsymbol{w}_0\|_2$, contradicting optimality. Thus the normalized pair solves (11.1). □

Definition 11.4 For $y \in \{-1, 1\}$, the **hinge loss** is

$$\ell_{\text{hinge}}((\boldsymbol{w}, b), (\boldsymbol{x}, y)) \coloneqq \big(1 - y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\big)_+,$$

where $(a)_+ \coloneqq \max\{a, 0\}$.

The constrained program (11.2) can be studied through Lagrange multipliers. When perfect separation is impossible, fix $C > 0$. Slack variables then lead to the **soft margin program**

$$\begin{aligned} \underset{\boldsymbol{w}, b, \xi_1, \ldots, \xi_m}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \quad i \in [m], \\ & \xi_i \geq 0, \quad i \in [m]. \end{aligned} \tag{11.5}$$

At the optimum,

$$\xi_i = \big(1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)\big)_+,$$

so (11.5) is equivalent to minimizing a squared norm penalty plus empirical hinge loss. This connects support vector machines to the convex regularization framework developed above.

The Rademacher bound in Corollary 2.4 controls a loss class. Applying it directly to zero-one halfspace loss does not make the bound margin sensitive. Indeed, positive rescaling does not change the classifier:

$$\operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) = \operatorname{sign}\left(\left\langle \frac{\boldsymbol{w}}{c}, \boldsymbol{x} \right\rangle + \frac{b}{c}\right), \quad c > 0.$$

This scale invariance is illustrated in Figure 19. More explicitly, the zero-one loss is

$$\ell_{0,1}((\boldsymbol{w}, b), (\boldsymbol{x}, y)) \coloneqq \begin{cases} 1, & y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \leq 0, \\ 0, & y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) > 0. \end{cases}$$

For any positive c,

$$\ell_{0,1}((\boldsymbol{w}, b), (\boldsymbol{x}, y)) = \ell_{0,1}((\boldsymbol{w}/c, b/c), (\boldsymbol{x}, y)).$$

Consequently, a norm constraint alone does not reduce the class of zero-one halfspace losses: every represented halfspace can be rescaled to have an arbitrarily small parameter norm. The hard margin constraints in (11.2) fix the scale, but a direct generalization analysis requires a loss that responds to the magnitude of the signed margin. The ramp loss supplies this scale sensitivity.

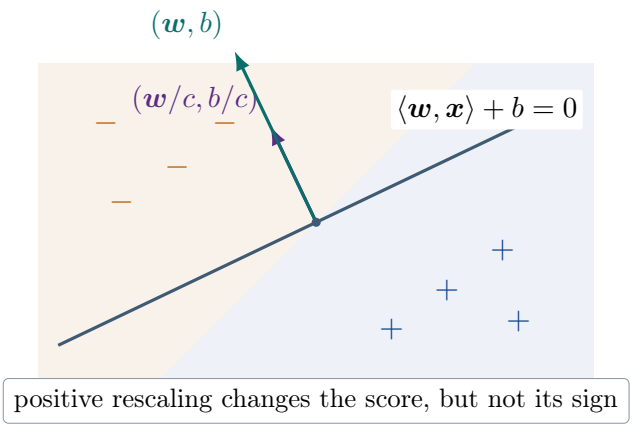

Figure 19

Ramp Loss

Definition 11.5 For a margin parameter $\gamma > 0$, define

$$\phi_\gamma(t) := \begin{cases} 1, & t \le 0, \\ 1 - t/\gamma, & 0 < t < \gamma, \\ 0, & t \ge \gamma. \end{cases}$$

The corresponding **ramp loss** is

$$\ell_\gamma((\boldsymbol{w}, b), (\boldsymbol{x}, y)) := \phi_\gamma\big(y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\big).$$

The ramp loss upper bounds zero-one loss but also distinguishes correctly classified observations according to their signed margin. It is $1/\gamma$-Lipschitz in that margin. The two losses are compared in Figure 20.

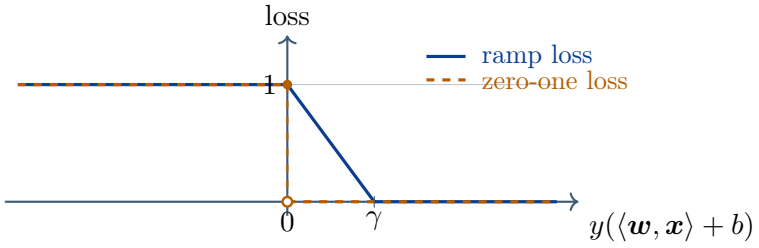

Figure 20

Lemma 11.6 (Rademacher contraction lemma) Let $\mathcal{F}$ be a real-valued function class and, for each $i \in [m]$, let $\psi_i \colon \mathbb{R} \to \mathbb{R}$ be L-Lipschitz with $\psi_i(0) = 0$. Then

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \psi_i(f(z_i)) \right] \leq L \widehat{R}_S(\mathcal{F}).$$

Proof. We contract one coordinate at a time. Condition on $\sigma_1, \ldots, \sigma_{m-1}$ and write

$$a_f \coloneqq \sum_{i=1}^{m-1} \sigma_i \psi_i(f(z_i)) \qquad \text{and} \qquad x_f \coloneqq f(z_m).$$

Averaging over σ_m gives

$$\begin{aligned} &\frac{1}{2} \sup_{f \in \mathcal{F}} \{a_f + \psi_m(x_f)\} + \frac{1}{2} \sup_{g \in \mathcal{F}} \{a_g - \psi_m(x_g)\} \\ &\quad = \frac{1}{2} \sup_{f,g \in \mathcal{F}} \{a_f + a_g + \psi_m(x_f) - \psi_m(x_g)\} \\ &\quad \leq \frac{1}{2} \sup_{f,g \in \mathcal{F}} \{a_f + a_g + L|x_f - x_g|\}. \end{aligned}$$

Because the supremum is symmetric in f and g, the last expression equals the corresponding average with Lx_f and Lx_g in place of $\psi_m(x_f)$ and $\psi_m(x_g)$. Repeating this argument for coordinates $m-1, \ldots, 1$ proves the result after division by m. □

For the ramp loss, take

$$\psi_i(t) \coloneqq \phi_\gamma(y_i t) - \phi_\gamma(0).$$

Each ψ_i is $1/\gamma$-Lipschitz and vanishes at zero. Subtracting the constant $\phi_\gamma(0)$ does not change empirical Rademacher complexity because the expected signed sum of a fixed constant is zero.

Proposition 11.7 Let

$$\mathcal{F}_B \coloneqq \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle : \|\boldsymbol{w}\|_2 \leq B\}.$$

For every sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$,

$$\widehat{R}_S(\mathcal{F}_B) \leq \frac{B}{m} \left(\sum_{i=1}^{m} \|\boldsymbol{x}_i\|_2^2 \right)^{1/2}.$$

In particular, if $\|\boldsymbol{x}_i\|_2 \le R$ for every i, then

$$\widehat{R}_S(\mathcal{F}_B) \le \frac{BR}{\sqrt{m}}.$$

Proof. By duality of the Euclidean norm,

$$\begin{aligned} m\widehat{R}_S(\mathcal{F}_B) &= \mathbb{E}_\sigma\left[\sup_{\|\boldsymbol{w}\|_2 \le B} \sum_{i=1}^m \sigma_i \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right] \\ &= B\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2\right] \\ &\le B\left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2^2\right]\right)^{1/2} \\ &= B\left(\sum_{i=1}^m \|\boldsymbol{x}_i\|_2^2\right)^{1/2}. \end{aligned}$$

The cross terms vanish because $\mathbb{E}[\sigma_i\sigma_j] = 0$ when $i \ne j$. The final claim follows from the radius bound. □

Corollary 11.8 Suppose $\|\boldsymbol{X}\|_2 \le R$ almost surely, and consider homogeneous linear scores with $\|\boldsymbol{w}\|_2 \le B$. With probability at least $1 - 2\delta$, every such $\boldsymbol{w}$ satisfies

$$L_D^{0,1}(\boldsymbol{w}) \le L_D^{\gamma}(\boldsymbol{w}) \le L_S^{\gamma}(\boldsymbol{w}) + \frac{2BR}{\gamma\sqrt{m}} + 3\sqrt{\frac{\log(2/\delta)}{2m}}, \tag{11.6}$$

where L^γ denotes ramp risk with margin parameter γ.

Proof. The first inequality follows pointwise from

$$\ell_{0,1}(\boldsymbol{w}, z) \le \ell_\gamma(\boldsymbol{w}, z).$$

By Lemma 11.6 and Proposition 11.7,

$$\widehat{R}_S(\ell_\gamma \circ \mathcal{F}_B) \le \frac{BR}{\gamma\sqrt{m}}.$$

Apply the empirical bound in (2) to the $[0,1]$-valued ramp loss class. □

An intercept can be included by augmenting $\boldsymbol{x}$ with a constant coordinate and constraining the augmented parameter norm. If γ is selected after observing the data, a simultaneous bound over a fixed grid of candidate margins, obtained by a union bound, accounts for that selection.

Theorem 11.9 Assume that $\|\boldsymbol{X}\|_2 \le R$ almost surely and that there is a vector $\boldsymbol{w}^* \in \mathbb{R}^d$ satisfying

$$\mathbb{P}_{(\boldsymbol{X},Y)\sim D}\left(Y\langle \boldsymbol{w}^*, \boldsymbol{X}\rangle \ge 1\right) = 1.$$

Let $\boldsymbol{w}_S$ minimize $\|\boldsymbol{w}\|_2$ subject to

$$y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle \ge 1 \quad \text{for every } i \in [m].$$

Then, with probability at least $1-2\delta$,

$$L_D^{0,1}(\boldsymbol{w}_S) \le \frac{2R\|\boldsymbol{w}^*\|_2}{\sqrt{m}} + 3\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Proof. The oracle vector $\boldsymbol{w}^*$ is feasible for the sample almost surely. Minimality of $\boldsymbol{w}_S$ therefore gives

$$\|\boldsymbol{w}_S\|_2 \le \|\boldsymbol{w}^*\|_2.$$

Moreover, every training observation has signed margin at least one, so

$$L_S^1(\boldsymbol{w}_S) = 0.$$

Apply Corollary 11.8 with $B = \|\boldsymbol{w}^*\|_2$ and $\gamma = 1$. □

The bound in Theorem 11.9 depends on the input radius and the norm of a unit margin separator, but not on the ambient dimension d. This is the principal benefit of a margin-based analysis.

Nearest Neighbor Methods

Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and let $\mathcal{Y} = \{0,1\}$. For a sample

$$S = ((x_1,y_1),\dots,(x_m,y_m)),$$

order the indices for each query $x \in \mathcal{X}$ so that

$$d_{\mathcal{X}}(x, x_{\pi_1(x)}) \le \cdots \le d_{\mathcal{X}}(x, x_{\pi_m(x)}),$$

using a fixed rule to break ties.

Definition 11.10 For an odd positive integer $k \le m$, the ***k*-nearest-neighbor classifier** is

$$h_{S,k}(x) := \mathbb{1}\left(\frac{1}{k}\sum_{j=1}^{k} y_{\pi_j(x)} \ge \frac{1}{2}\right).$$

The case $k = 1$ is the **one-nearest-neighbor classifier**.

Nearest-neighbor analysis does not begin with uniform convergence over a fixed hypothesis class. Instead, it compares the data-dependent rule with the **Bayes classifier**. Let $\mathbb{P}_X$ be the marginal distribution of X and define the **regression function**

$$\eta(x) := \mathbb{P}(Y = 1 \mid X = x).$$

The Bayes classifier is

$$h^*(x) = \mathbb{1}(\eta(x) \ge 1/2),$$

and its conditional error at x is $\min\{\eta(x), 1 - \eta(x)\}$.

Theorem 11.11 Assume that η is c-Lipschitz:

$$\left|\eta(x) - \eta(x')\right| \le c\, d_{\mathcal{X}}(x, x')$$

for every $x, x' \in \mathcal{X}$. Then

$$\mathbb{E}_{S\sim D^m}\left[L_D(h_{S,1})\right] \le 2L_D(h^*) + c\,\mathbb{E}_{S,X}\left[d_{\mathcal{X}}(X, x_{\pi_1(X)})\right].$$

Proof. Condition on the query $X = x$ and its nearest sample point $x_{\pi_1(x)} = x'$. The query label and nearest-neighbor label are conditionally independent Bernoulli variables with success probabilities $\eta(x)$ and $\eta(x')$. Hence the conditional error probability is

$$\eta(x)(1 - \eta(x')) + (1 - \eta(x))\eta(x') = 2\eta(x)(1 - \eta(x)) + \big(\eta(x) - \eta(x')\big)(2\eta(x) - 1).$$

The first term is at most

$$2\min\{\eta(x), 1 - \eta(x)\},$$

and the second is at most

$$\big|\eta(x) - \eta(x')\big| \le c\, d_{\mathcal{X}}(x, x').$$

Taking expectation over the query and the sample proves the result. □

The random distance appearing in Theorem 11.11 is illustrated in Figure 21.

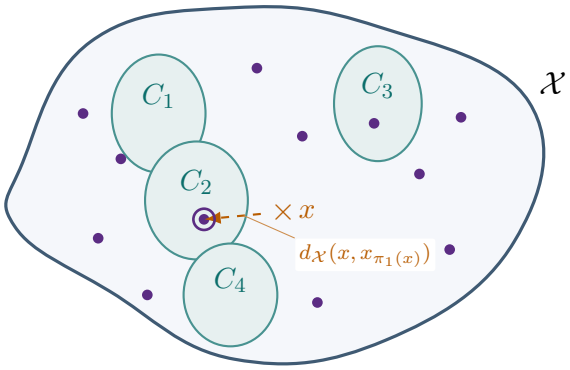

FIGURE 21

Lemma 11.12 Let $C_1, \dots, C_r$ be a measurable partition of $\mathcal{X}$. Let

$$S_X = (X_1, \dots, X_m) \sim \mathbb{P}_X^m.$$

Then

$$\mathbb{E}_{S_X}\left[\sum_{j=1}^{r} \mathbb{P}_X(C_j)\mathbb{1}(S_X \cap C_j = \varnothing)\right] \le \frac{r}{em}.$$

Proof. Write $p_j = \mathbb{P}_X(C_j)$. The probability that cell C_j receives no sample point is $(1 - p_j)^m$. Therefore the expected missing mass is

$$\sum_{j=1}^{r} p_j (1 - p_j)^m \le \sum_{j=1}^{r} p_j e^{-m p_j}.$$

The function $p \mapsto p e^{-mp}$ is maximized at $p = 1/m$, where its value is $1/(em)$. Summing over the r cells proves the result. □

Theorem 11.13 Let $\mathcal{X} = [0,1]^d$ with Euclidean distance. For every distribution $\mathbb{P}_X$ on $\mathcal{X}$,

$$\mathbb{E}_{S_X, \boldsymbol{X}} \left[\left\| \boldsymbol{X} - \boldsymbol{x}_{\pi_1(\boldsymbol{X})} \right\|_2 \right] \le \frac{4\sqrt{d}}{m^{1/(d+1)}}.$$

Proof. Partition the unit cube into n^d cubes of side length $1/n$, as in Figure 22. If the cell containing the query $\boldsymbol{X}$ also contains a sample point, then the nearest-neighbor distance is at most the cell diameter $\sqrt{d}/n$. If that cell is empty, the distance is at most the diameter $\sqrt{d}$ of the unit cube. Hence Lemma 11.12 gives

$$\mathbb{E}_{S_X, \boldsymbol{X}} \left[\left\| \boldsymbol{X} - \boldsymbol{x}_{\pi_1(\boldsymbol{X})} \right\|_2 \right] \le \sqrt{d} \left(\frac{n^d}{em} + \frac{1}{n} \right). \tag{11.7}$$

Choose

$$n = \max \left\{ 1, \left\lfloor m^{1/(d+1)} \right\rfloor \right\}.$$

When $m^{1/(d+1)} \ge 2$, one has

$$\frac{n^d}{m} \le m^{-1/(d+1)} \qquad \text{and} \qquad \frac{1}{n} \le 2m^{-1/(d+1)}.$$

Substitution into (11.7) gives the claimed bound, with room in the constant. When $m^{1/(d+1)} < 2$, the trivial diameter bound is already no larger than the displayed right-hand side. □

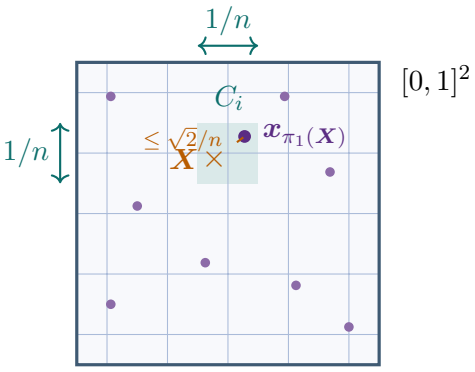

FIGURE 22

Corollary 11.14 Under the assumptions of Theorem 11.11, if $\mathcal{X} = [0,1]^d$, then

$$\mathbb{E}_{S\sim D^m}\left[L_D(h_{S,1})\right] \leq 2L_D(h^*) + \frac{4c\sqrt{d}}{m^{1/(d+1)}}.$$

The exponent $1/(d+1)$ deteriorates rapidly with dimension. This is a concrete manifestation of the **curse of dimensionality**: without additional structure, a prohibitively large sample may be required before the nearest observed point is close to a new query.

5. Online Learning

Realizable Online Classification

All preceding statistical guarantees assumed independent and identically distributed observations. That assumption connects past data to future performance. Online learning removes this probabilistic link and evaluates a learner on an arbitrary sequence of instances and labels.

Without a sampling model, absolute prediction guarantees are impossible. Performance is therefore measured relative to a benchmark, such as the best fixed hypothesis in a class. This viewpoint includes prediction with expert advice, universal coding, and game-theoretic formulations of probability.

Let $\mathcal{X}$ be an instance space and initially take $\mathcal{Y} = \{0,1\}$. At each round t, the interaction follows the protocol displayed below.

round	instance	prediction	revealed label	loss
1	x_1	$p_1 \in \mathcal{Y}$	y_1	$\ell(p_1, y_1)$
2	x_2	$p_2 \in \mathcal{Y}$	y_2	$\ell(p_2, y_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	x_T	$p_T \in \mathcal{Y}$	y_T	$\ell(p_T, y_T)$

For zero-one loss,

$$\ell_{0,1}(p_t, y_t) \coloneqq \begin{cases} 1, & p_t \neq y_t, \\ 0, & p_t = y_t. \end{cases}$$

The total number of mistakes through round T is

$$M_T \coloneqq \sum_{t=1}^{T} \ell_{0,1}(p_t, y_t).$$

Definition 11.15 A labeled sequence $((x_t, y_t))_{t=1}^{T}$ is **realizable** by $\mathcal{H}$ if there is a single $h^* \in \mathcal{H}$ such that

$$h^*(x_t) = y_t \quad \text{for every } t \in [T].$$

The environment may choose the sequence adaptively, but realizability requires one hypothesis to remain consistent with the entire history. Given the observations through round $t-1$, define the **version space**

$$V_t \coloneqq \{h \in \mathcal{H} : h(x_s) = y_s \text{ for every } s < t\}.$$

A simple consistent learner selects any $h_t \in V_t$ and predicts $h_t(x_t)$. When $\mathcal{H}$ is finite, each mistake eliminates the selected hypothesis while h^* remains. Therefore this learner makes at most $|\mathcal{H}| - 1$ mistakes. The halving algorithm improves this linear dependence to a logarithmic one.

Lecture 12 — Tuesday, December 04, 2018

The Halving Algorithm

Definition 12.1 Let $\mathcal{H}$ be finite. The **halving algorithm** proceeds as follows.

1. Initialize $V_1 = \mathcal{H}$.
2. On round t, receive x_t and predict
$$p_t \in \arg\max_{r \in \{0,1\}} |\{h \in V_t : h(x_t) = r\}|,$$
breaking ties arbitrarily.
3. After observing y_t, update
$$V_{t+1} = \{h \in V_t : h(x_t) = y_t\}.$$

Theorem 12.2 On every sequence realizable by a finite class $\mathcal{H}$, the halving algorithm makes at most
$$\log_2 |\mathcal{H}|$$
mistakes.

Proof. The realizing hypothesis h^* belongs to every version space, so $|V_t| \geq 1$. Whenever the prediction in (2) is wrong, the update in (3) retains the minority side of the split. Thus a mistake implies
$$|V_{t+1}| \leq \frac{|V_t|}{2}.$$
If M mistakes have occurred, then
$$1 \leq |V_{T+1}| \leq |\mathcal{H}|\, 2^{-M}.$$
Taking base-two logarithms gives $M \leq \log_2 |\mathcal{H}|$. □

Cardinality is not always the most informative potential. Consider the indicator class
$$\mathcal{H}_{\text{ind}} \coloneqq \{h_j : i \mapsto \mathbb{1}(i = j),\ j \in \mathcal{X}\},$$
where $\mathcal{X} = [m]$ or $\mathbb{N}$. Querying an index j separates the single hypothesis h_j from all remaining hypotheses. This motivates a sequential dimension that measures the depth of adaptive binary distinctions rather than the number of hypotheses.

Definition 12.3 A **complete binary instance tree of depth d** has an internal node labeled by an instance $x_s \in \mathcal{X}$ for every binary string $s \in \{0,1\}^{<d}$. The root corresponds to the empty string ε, and the two edges leaving each internal node are labeled 0 and 1. The tree is **shattered by $\mathcal{H}$** if, for every root-to-leaf sequence $(y_1, \dots, y_d) \in \{0,1\}^d$, there is a hypothesis $h \in \mathcal{H}$ such that

$$h\big(x_{y_1 \cdots y_{t-1}}\big) = y_t \quad \text{for every } t \in [d].$$

Thus, the class realizes every adaptive sequence of edge labels. A depth-two instance tree is shown in Figure 23.

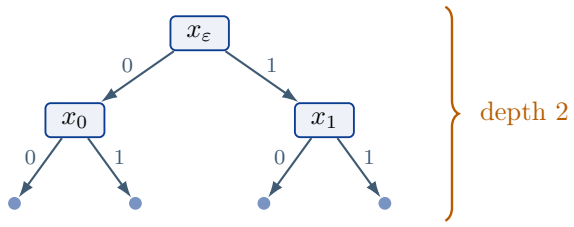

FIGURE 23

A shattered tree serves as a strategy tree for the environment. After observing the learner's prediction, the environment may reveal the opposite label and follow the corresponding edge, while Definition 12.3 guarantees that the labels along the chosen path remain realizable.

Definition 12.4 The **Littlestone dimension** of $\mathcal{H}$, denoted by $\mathrm{Ldim}(\mathcal{H})$, is the largest depth of a complete binary instance tree shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters trees of every finite depth, then $\mathrm{Ldim}(\mathcal{H}) = \infty$. We adopt the convention $\mathrm{Ldim}(\varnothing) = -1$.

Proposition 12.5 If $\mathcal{H}$ is finite, then

$$\mathrm{Ldim}(\mathcal{H}) \le \log_2 |\mathcal{H}|.$$

Proof. If $\mathcal{H}$ shatters a tree of depth d, then each of its 2^d root-to-leaf label sequences must be realized by a distinct hypothesis: two different paths disagree at their first divergent edge. Hence $|\mathcal{H}| \ge 2^d$, which gives the claimed inequality. □

Example 12.6 Suppose that $\mathcal{X}$ contains at least two points. For the indicator class $\mathcal{H}_{\text{ind}}$ defined above,

$$\text{Ldim}(\mathcal{H}_{\text{ind}}) = 1.$$

The class shatters a depth-one tree: label its root by any $j \in \mathcal{X}$, use h_j to realize the label 1, and use any h_k with $k \neq j$ to realize the label 0. At a root labeled by j, however, the branch carrying label 1 contains only h_j. A singleton class cannot shatter a tree of depth one, so no depth-two tree is shattered.

Example 12.7 Let

$$\mathcal{H}_{\text{th}} \coloneqq \{x \mapsto \mathbb{1}(x \le a) : a \in \mathbb{R}\}.$$

Although $\text{VCdim}(\mathcal{H}_{\text{th}}) = 1$, its Littlestone dimension is infinite.

It is enough to restrict the instance space to $[0, 1]$. The first instance in the recursive construction is the midpoint $1/2$; the resulting split is shown in Figure 24.

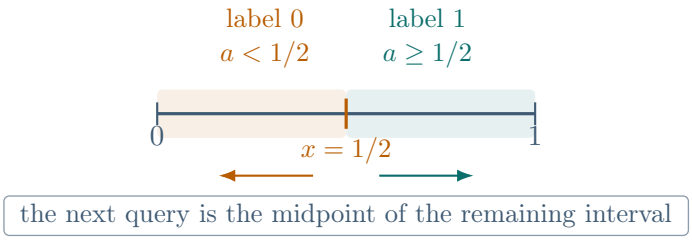

FIGURE 24

At every subsequent node, choose the midpoint of the interval of threshold parameters consistent with the preceding edge labels. The first two levels of this construction are shown in Figure 25.

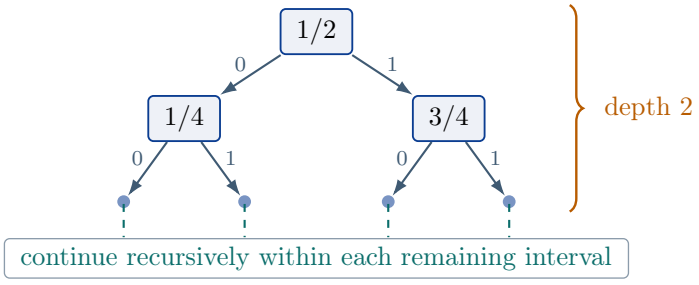

FIGURE 25

Begin with the interval $(0, 1)$ of admissible threshold parameters. At an internal node associated with an interval (a, b), label the node by $x = (a + b)/2$. Following the edge labeled 0 restricts the threshold parameter to (a, x), while following the edge labeled 1 restricts it to (x, b). Both intervals are nonempty. Consequently, every finite root-to-leaf path leaves a nonempty interval of threshold parameters, and any parameter in that interval realizes the entire path. The class therefore shatters a tree of every finite depth.

Moreover, Littlestone dimension is monotone under inclusion: if $\mathcal{H} \subseteq \mathcal{H}'$, then every tree shattered by $\mathcal{H}$ is also shattered by $\mathcal{H}'$. Since every class of affine halfspaces in positive dimension contains a copy of the one-dimensional threshold class, its Littlestone dimension is also infinite.

Proposition 12.5 resembles the mistake bound for the halving algorithm, but Littlestone dimension leads to an optimal strategy even when cardinality is uninformative. For example, querying j in the indicator class splits the version space as shown in Figure 26.

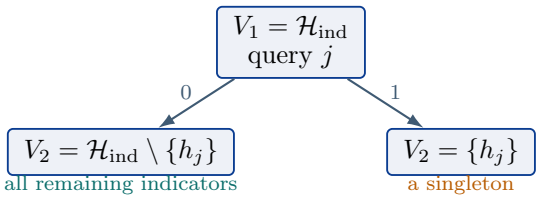

FIGURE 26

More generally, an instance x_t partitions a version space V_t into

$$V_t^r := \{h \in V_t : h(x_t) = r\}, \quad r \in \{0, 1\},$$

as depicted in Figure 27.

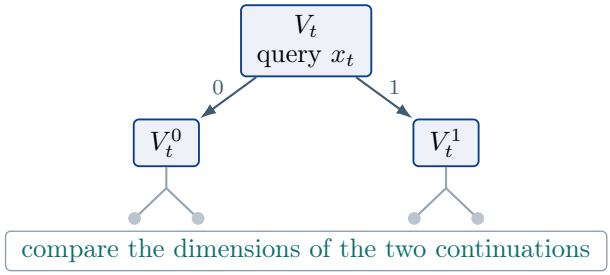

FIGURE 27

Definition 12.8 On round t, the **standard optimal algorithm** forms the two restricted version spaces V_t^0 and V_t^1 and predicts

$$p_t \in \operatorname*{argmax}_{r \in \{0,1\}} \operatorname{Ldim}(V_t^r).$$

After observing y_t, it updates $V_{t+1} = V_t^{y_t}$.

The following elementary splitting property drives the analysis.

Lemma 12.9 Let V be a hypothesis class with finite Littlestone dimension d, and let $x \in \mathcal{X}$. It is impossible for both

$$V^0 = \{h \in V : h(x) = 0\} \qquad \text{and} \qquad V^1 = \{h \in V : h(x) = 1\}$$

to have Littlestone dimension d.

Proof. If both subclasses shattered depth-d trees, place those two trees below a new root labeled by x, putting the tree for V^0 below the edge labeled 0 and the tree for V^1 below the edge labeled 1. The resulting depth-$(d+1)$ tree would be shattered by V, contrary to $\operatorname{Ldim}(V) = d$. □

Definition 12.10 The **optimal realizable mistake bound** $M^*(\mathcal{H})$ is the smallest worst-case number of mistakes among all deterministic online learners, where the worst case ranges over all finite sequences realizable by $\mathcal{H}$.

Theorem 12.11 (Littlestone's mistake bound theorem) For every nonempty binary hypothesis class $\mathcal{H}$,

$$M^*(\mathcal{H}) = \operatorname{Ldim}(\mathcal{H}).$$

When the Littlestone dimension is finite, the standard optimal algorithm attains this bound.

Proof. For the lower bound, suppose that $\mathcal{H}$ shatters a tree of depth d. The environment begins at the root. On every round, it presents the instance at the current node, reveals the label opposite to the learner's prediction, and follows the edge carrying that label. Definition 12.3 ensures that the resulting length-d sequence is realizable by $\mathcal{H}$, while the learner makes a mistake on every round. Hence $M^*(\mathcal{H}) \geq d$. Taking the largest possible d, or arbitrarily large finite d when the Littlestone dimension is infinite, proves the lower bound.

For the upper bound, assume that $d = \mathrm{Ldim}(\mathcal{H}) < \infty$ and run the standard optimal algorithm. If it makes a mistake at time t, then the actual version space $V_t^{y_t}$ is the branch it did not predict. By the prediction rule,

$$\mathrm{Ldim}(V_t^{y_t}) \le \mathrm{Ldim}(V_t^{p_t}).$$

If $\mathrm{Ldim}(V_t^{y_t})$ equalled $\mathrm{Ldim}(V_t)$, monotonicity under inclusion would force both branches to have the same dimension as V_t, contradicting Lemma 12.9. Thus every mistake decreases the Littlestone dimension of the version space by at least one. Starting from $V_1 = \mathcal{H}$, the algorithm can therefore make at most d mistakes. □

Theorem 12.12 Every binary hypothesis class satisfies

$$\mathrm{VCdim}(\mathcal{H}) \le \mathrm{Ldim}(\mathcal{H}).$$

Proof. Suppose that $\{x_1, \dots, x_d\}$ is shattered in the VC sense. Form a complete binary tree of depth d by labeling every node at level t with x_t, independently of the preceding edge labels. A schematic view appears in Figure 28.

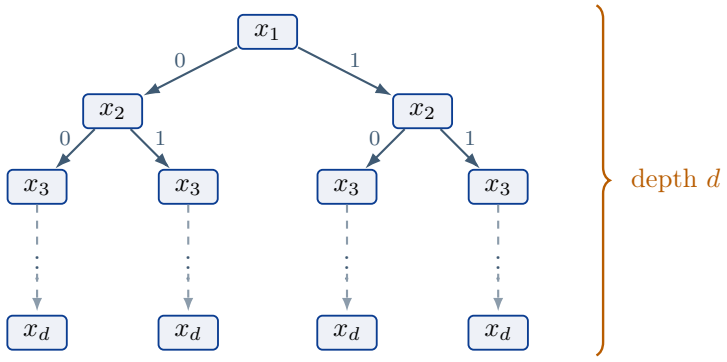

FIGURE 28

Every root-to-leaf path specifies a binary labeling of $\{x_1, \dots, x_d\}$. Since that set is VC shattered, some hypothesis in $\mathcal{H}$ realizes the path. The displayed tree is therefore shattered in the sense of Definition 12.3, which shows that $\mathrm{Ldim}(\mathcal{H}) \ge d$. □

The preceding theory gives exact worst-case mistake bounds for realizable sequences. Without a realizability assumption, no hypothesis need predict every label correctly. An absolute mistake bound is therefore generally impossible. The appropriate objective is relative: compare the learner's cumulative loss with that of the best fixed decision selected in hindsight.

Online Convex Optimization

Definition 12.13 In **online convex optimization**, the learner uses a nonempty closed convex decision set $\mathcal{H} \subseteq \mathbb{R}^d$, the environment uses an arbitrary outcome space $\mathcal{Z}$, and their interaction is evaluated by a loss $\ell\colon \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ such that $\ell(\,\cdot\,, z)$ is convex for every $z \in \mathcal{Z}$. On each round $t = 1, 2, \ldots$, the following events occur:

1. The learner selects a decision $\boldsymbol{w}^{(t)} \in \mathcal{H}$.
2. The environment reveals an outcome $z_t \in \mathcal{Z}$.
3. The learner incurs the loss $\ell\big(\boldsymbol{w}^{(t)}, z_t\big)$.

Definition 12.14 Write $f_t = \ell(\,\cdot\,, z_t)$. The learner's **regret against a fixed comparator** $\boldsymbol{w}^* \in \mathcal{H}$ after T rounds is

$$\operatorname{Regret}_T(\boldsymbol{w}^*) \coloneqq \sum_{t=1}^{T} f_t\big(\boldsymbol{w}^{(t)}\big) - \sum_{t=1}^{T} f_t(\boldsymbol{w}^*).$$

Its regret on the outcome sequence is

$$\operatorname{Regret}_T(\mathcal{H}) \coloneqq \sup_{\boldsymbol{w}^* \in \mathcal{H}} \operatorname{Regret}_T(\boldsymbol{w}^*) = \sum_{t=1}^{T} f_t\big(\boldsymbol{w}^{(t)}\big) - \inf_{\boldsymbol{w}^* \in \mathcal{H}} \sum_{t=1}^{T} f_t(\boldsymbol{w}^*).$$

An algorithm is **no regret** if $\operatorname{Regret}_T(\mathcal{H})/T \to 0$ for every admissible outcome sequence.

Definition 12.15 Assume that $\mathbf{0} \in \mathcal{H}$, choose a step size $\eta > 0$, and initialize $\boldsymbol{w}^{(1)} = \mathbf{0}$. On round t, **online gradient descent** performs Steps (1), (2), and (3), chooses a subgradient

$$\boldsymbol{v}_t \in \partial f_t\big(\boldsymbol{w}^{(t)}\big),$$

and updates

$$\boldsymbol{w}^{(t+1)} = \Pi_{\mathcal{H}}\big(\boldsymbol{w}^{(t)} - \eta \boldsymbol{v}_t\big),$$

where $\Pi_{\mathcal{H}}$ denotes Euclidean projection onto $\mathcal{H}$. When $\mathcal{H} = \mathbb{R}^d$, the projection is the identity map.

Unlike stochastic subgradient descent, online gradient descent does not assume that the outcomes

are independent and identically distributed. The environment may choose them adaptively. Nevertheless, the same one-step potential argument yields a deterministic regret bound.

Theorem 12.16 Suppose that every f_t is convex and that online gradient descent chooses subgradients satisfying $\|\boldsymbol{v}_t\|_2 \le \rho$. Then, for every comparator $\boldsymbol{w}^* \in \mathcal{H}$,

$$\operatorname{Regret}_T(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta\rho^2 T}{2}. \tag{12.1}$$

The subgradient bound follows, for example, when each f_t is the restriction to $\mathcal{H}$ of a convex ρ-Lipschitz function on $\mathbb{R}^d$. Merely assuming relative Lipschitz continuity on $\mathcal{H}$ would not bound every possible choice of a boundary subgradient.

Proof. Euclidean projection onto a closed convex set cannot increase the distance to any point in that set. Hence

$$\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\right\|^2 \le \left\|\boldsymbol{w}^{(t)} - \eta\boldsymbol{v}_t - \boldsymbol{w}^*\right\|^2 = \left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\right\|^2 - 2\eta\left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t \right\rangle + \eta^2\|\boldsymbol{v}_t\|^2.$$

By the subgradient inequality in Definition 10.2,

$$\left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t \right\rangle \ge f_t\big(\boldsymbol{w}^{(t)}\big) - f_t(\boldsymbol{w}^*).$$

Rearranging the preceding display gives

$$f_t\big(\boldsymbol{w}^{(t)}\big) - f_t(\boldsymbol{w}^*) \le \frac{\left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\right\|^2 - \left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\right\|^2}{2\eta} + \frac{\eta}{2}\|\boldsymbol{v}_t\|^2.$$

Summing over $t \in [T]$ telescopes the squared distances. Since $\boldsymbol{w}^{(1)} = \boldsymbol{0}$ and the final squared distance is nonnegative,

$$\operatorname{Regret}_T(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2.$$

The assumed bound $\|\boldsymbol{v}_t\| \le \rho$ proves (12.1). □

The generic choice $\eta = 1/\sqrt{T}$ gives

$$\operatorname{Regret}_T(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{w}^*\|^2 + \rho^2}{2}\sqrt{T}.$$

If the comparison class additionally satisfies $\|\boldsymbol{w}^*\| \le B$ for some $B > 0$, and $\rho > 0$, choosing

$\eta = B/(\rho\sqrt{T})$ yields the sharper uniform bound

$$\mathrm{Regret}_T(\mathcal{H}) \le B\rho\sqrt{T}.$$

Both bounds are sublinear in T, and the second is dimension free.

We now return to online binary classification and analyze the Perceptron through the online convex optimization framework. Consider the homogeneous halfspace class

$$\mathcal{H}_{\mathrm{hom},d} \coloneqq \left\{ \boldsymbol{x} \mapsto \mathrm{sign}\big(\langle \boldsymbol{w}, \boldsymbol{x}\rangle\big) : \boldsymbol{w} \in \mathbb{R}^d \right\}, \qquad \mathcal{X} = \mathbb{R}^d, \qquad \text{and} \qquad \mathcal{Y} = \{-1, 1\}.$$

As Example 12.7 demonstrates for a subclass of affine halfspaces, realizability alone need not provide a finite online mistake bound. An environment may exploit unbounded feature norms or present points with margins approaching zero. The Perceptron bound below makes these two obstructions explicit.

For a labeled example $z = (\boldsymbol{x}, y)$, the **zero-one margin loss** is

$$\ell_{01}(\boldsymbol{w}, z) \coloneqq \mathbb{1}\big(y\langle \boldsymbol{w}, \boldsymbol{x}\rangle \le 0\big).$$

The associated hinge loss

$$\ell_{\mathrm{hinge}}(\boldsymbol{w}, z) \coloneqq \big(1 - y\langle \boldsymbol{w}, \boldsymbol{x}\rangle\big)_+$$

has two important properties:

1. It is convex as a function of $\boldsymbol{w}$.
2. It satisfies $\ell_{\mathrm{hinge}}(\boldsymbol{w}, z) \ge \ell_{01}(\boldsymbol{w}, z)$ for every $\boldsymbol{w}$, so it is a convex surrogate for the zero-one loss.

Online Perceptron as Online Gradient Descent

Definition 12.17 Fix a step size $\eta > 0$ and initialize $\boldsymbol{w}^{(1)} = \boldsymbol{0}$. At time t, the **online Perceptron** predicts

$$p_t = \mathrm{sign}\big(\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_t\rangle\big).$$

After receiving y_t, it updates according to

$$\boldsymbol{w}^{(t+1)} = \begin{cases} \boldsymbol{w}^{(t)}, & \text{if } y_t\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_t\rangle > 0, \\ \boldsymbol{w}^{(t)} + \eta y_t \boldsymbol{x}_t, & \text{if } y_t\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_t\rangle \le 0. \end{cases} \tag{12.2}$$

A zero signed margin is counted as an error and triggers an update.

Let

$$\mathcal{M} \coloneqq \Big\{ t \in [T] : y_t\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_t\rangle \le 0 \Big\}$$

be the set of update rounds. To express (12.2) as online gradient descent, define the adaptive convex loss

$$f_t(\boldsymbol{w}) \coloneqq \mathbb{1}(t \in \mathcal{M})\big(1 - y_t\langle \boldsymbol{w}, \boldsymbol{x}_t\rangle\big)_+. \tag{12.3}$$

If $t \notin \mathcal{M}$, then f_t is identically zero and we take $\boldsymbol{v}_t = \boldsymbol{0}$. If $t \in \mathcal{M}$, then

$$\boldsymbol{v}_t = -y_t \boldsymbol{x}_t \in \partial f_t\big(\boldsymbol{w}^{(t)}\big).$$

Thus $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{v}_t$ is precisely the Perceptron update. Iterating this identity gives

$$\boldsymbol{w}^{(t)} = \eta \sum_{\substack{i \in \mathcal{M} \\ i < t}} y_i \boldsymbol{x}_i. \tag{12.4}$$

Since $\eta > 0$, its value does not affect the signs of the predictions:

$$p_t = \operatorname{sign}\left(\left\langle \sum_{\substack{i \in \mathcal{M} \\ i < t}} y_i \boldsymbol{x}_i, \boldsymbol{x}_t \right\rangle \right).$$

Theorem 12.18 Run the online Perceptron on $((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T))$, and suppose that

$$R \coloneqq \max_{t \in [T]} \|\boldsymbol{x}_t\|_2 < \infty.$$

For a comparator $\boldsymbol{w}^* \in \mathbb{R}^d$, define its hinge loss on the update rounds by

$$L_{\mathcal{M}}(\boldsymbol{w}^*) \coloneqq \sum_{t \in \mathcal{M}} \big(1 - y_t\langle \boldsymbol{w}^*, \boldsymbol{x}_t\rangle\big)_+.$$

Then

$$|\mathcal{M}| \le \frac{1}{4}\left(R\|\boldsymbol{w}^*\| + \sqrt{R^2\|\boldsymbol{w}^*\|^2 + 4L_{\mathcal{M}}(\boldsymbol{w}^*)}\right)^2 \tag{12.5}$$

$$\le \left(R\|\boldsymbol{w}^*\| + \sqrt{L_{\mathcal{M}}(\boldsymbol{w}^*)}\right)^2. \tag{12.6}$$

Proof. On every update round, the current signed margin is nonpositive, so $f_t(\boldsymbol{w}^{(t)}) \ge 1$. On all other rounds, f_t is zero. Consequently,

$$\sum_{t=1}^{T} f_t(\boldsymbol{w}^{(t)}) \ge |\mathcal{M}|.$$

The potential calculation in the proof of Theorem 12.16, applied without projection, therefore yields

$$|\mathcal{M}| - L_{\mathcal{M}}(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2 \le \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta R^2|\mathcal{M}|}{2}.$$

Optimizing the right-hand side over $\eta > 0$ gives

$$|\mathcal{M}| - L_{\mathcal{M}}(\boldsymbol{w}^*) \le R\|\boldsymbol{w}^*\|\sqrt{|\mathcal{M}|}.$$

The case $R\|\boldsymbol{w}^*\| = 0$ follows by a limiting argument. Otherwise, the optimizing step size is $\eta = \|\boldsymbol{w}^*\|/(R\sqrt{|\mathcal{M}|})$.

Set $u = \sqrt{|\mathcal{M}|}$ and $a = R\|\boldsymbol{w}^*\|$. The last inequality is $u^2 - au - L_{\mathcal{M}}(\boldsymbol{w}^*) \le 0$, so the quadratic formula gives

$$u \le \frac{a + \sqrt{a^2 + 4L_{\mathcal{M}}(\boldsymbol{w}^*)}}{2}.$$

Squaring proves (12.5). Finally, $\sqrt{a^2 + 4L} \le a + 2\sqrt{L}$ for $a, L \ge 0$, which proves (12.6). □

Corollary 12.19 (Perceptron margin bound) Suppose that there is a vector $\boldsymbol{w}^*$ satisfying

$$y_t\langle \boldsymbol{w}^*, \boldsymbol{x}_t\rangle \ge 1 \quad \text{for every } t \in [T].$$

Then

$$|\mathcal{M}| \le R^2\|\boldsymbol{w}^*\|^2.$$

Equivalently, if there is a unit vector $\boldsymbol{u}$ for which $y_t\langle \boldsymbol{u}, \boldsymbol{x}_t\rangle \geq \gamma > 0$ on every round, then

$$|\mathcal{M}| \leq \frac{R^2}{\gamma^2}.$$

Proof. The first conclusion follows from (12.5) because $L_{\mathcal{M}}(\boldsymbol{w}^*) = 0$. For the second, take $\boldsymbol{w}^* = \boldsymbol{u}/\gamma$; the resulting mistake bound decreases as the margin γ grows. □

Appendix: Lecture and Tutorial Index