II Mathematics of Machine Learning

Ishan Nath, Lent 2024

Based on Lectures by Prof. Rajen Shah

April 24, 2024

Contents

1 Introduction

Consider a pair of random variables (X, Y). Here X is the input, or features, or predictors, and Y is the response or output. We have $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where we assume $\mathcal{X} = \mathbb{R}^p$, with joint distribution P_0.

We wish to predict Y from X using a *hypothesis* $h : \mathcal{X} \to \mathcal{Y}$. We measure the quality of a prediction using a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. In a classification setting $\mathcal{Y} = \{-1, +1\}$. Typically, ℓ is the misclassification loss:

$$\ell(h(x), y) = \mathbb{1}_{\{h(x) \neq y\}}.$$

We refer to h as a *classifier*. In a regression setting, $\mathcal{Y} \in \mathbb{R}$. Typically, ℓ is the squared error loss:

$$\ell(h(x), y) = (h(x) - y)^2.$$

We aim to pick h with small *risk*: $R(h) = \mathbb{E}[\ell(h(X), Y)]$, when h is deterministic. However most of the time h is not deterministic.

The squared error risk is minimized by the *regression function*

$$x \mapsto \mathbb{E}[Y|X = x].$$

The classifier h_0 that minimises the 0-1 risk is known as the *Bayes classifier*. Its risk is known as the *Bayes risk*. In this context,

$$\eta(x) = \mathbb{P}(Y = 1|X = x)$$

is the *regression function.*

Proposition 1.1. *A Bayes classifier is given by*

$$h_0(x) = \begin{cases} 1 & \eta(x) > 1/2, \\ 0 & \textit{otherwise.} \end{cases}$$

Suppose we have iid copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y). These copies is what is known as our *training data*, D. Our task is to use this data to construct $\hat{h}$ such that the risk

$$R(\hat{h}) = \mathbb{E}[\ell(h(X), Y)|D]$$

is small. Note that $R(\hat{h})$ is now a random variable!

The classical statistics approach is as follows: model P_0 using some parametric family, and try to estimate the unknown parameters. The machine learning approach begins with a given class $\mathcal{H}$ from which we pick $\hat{h}$. Hence our task, in this course, is to find an algorithm to pick $\hat{h}$.

Example 1.1.

Linear regression is an example of this. We have

$$\mathcal{H} = \{x \mapsto \mu + \beta^T x, \mu \in \mathbb{R}, \beta \in \mathbb{R}^p\}.$$

We can generalize this a bit, and take

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{d} \phi_j(x)\beta_j \mid \beta \in \mathbb{R}^d \right\},$$

where $\phi_1, \dots, \phi_d$ are fixed *basis functions* $\phi_j : \mathbb{R}^p \to \mathbb{R}$. Or even more generally,

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{d} \phi_j(x)\beta_j \mid \beta \in \mathbb{R}^d, \phi_j \in \mathcal{B} \right\}$$

for a class $\mathcal{B}$ of functions $\phi : \mathbb{R}^p \to \mathbb{R}$.

For classification, we can take a linear model and compose it with the sign function, where $\operatorname{sgn}(0) = -1$.

1.1 Review of Conditional Expectation

Suppose $Z \in \mathbb{R}$, $W \in \mathbb{R}^d$ are random variables. There are two ways of thinking about $\mathbb{E}[Z|W]$.

- We can think of it as

  $$\mathbb{E}[Z|W = w] = \int z f_{z|w}(z|w) \, \mathrm{d}z = g(w),$$

  hence $\mathbb{E}[Z|W] = g(W)$.

- Or, $\mathbb{E}[Z|W]$ is our best guess of Z knowing only W.

Let's look at a couple of specific cases.

(i) If $Z \perp\!\!\!\perp W$ (meaning Z independent of W), then

$$\mathbb{E}[Z|W] = \mathbb{E}[Z].$$

More generally, if $U \perp\!\!\!\perp (Z, W)$ then

$$\mathbb{E}[Z|W, U] = \mathbb{E}[Z|W].$$

(ii) The tower property: if $f : \mathbb{R}^d \to \mathbb{R}^m$ then

$$\mathbb{E}[\mathbb{E}[Z|W]|f(W)] = \mathbb{E}[Z|f(W)].$$

In particular, if f is constant, then

$$\mathbb{E}[\mathbb{E}[Z|W]] = \mathbb{E}[Z].$$

(iii) We can fix what we know.

$$\begin{aligned}&\mathbb{E}[f(W_1, \dots, W_d)|W_1 = w_1, \dots, W_r = w_r]\\ &\qquad = \mathbb{E}[f(w_1, \dots, w_r, W_{r+1}, \dots, W_d)|W_1 = w_1, \dots, W_r = w_r].\end{aligned}$$

This is often used in a specific way. If $g : \mathbb{R}^d \to \mathbb{R}$, then

$$\mathbb{E}[g(W)Z|W] = g(W)\mathbb{E}[Z|W].$$

(iv) The best least squares predictor is

$$\mathbb{E}[(Z - g(W))^2] = \mathbb{E}[(Z - \mathbb{E}[Z|W])^2] + \mathbb{E}[(\mathbb{E}[Z|W] - g(W))^2].$$

We can prove this. Indeed,

$$\begin{aligned}\mathbb{E}[(Z - g(W))^2] &= \mathbb{E}[(Z - \mathbb{E}[Z|W] + \mathbb{E}[Z|W] - g(W))^2]\\ &= \mathbb{E}[(Z - \mathbb{E}[Z|W])^2] + \mathbb{E}[(\mathbb{E}[Z|W] - g(W))^2]\\ &\qquad + 2\mathbb{E}[(Z - \mathbb{E}[Z|W])(\mathbb{E}[Z|W] - g(W))],\end{aligned}$$

but now we find that

$$\begin{aligned}&\mathbb{E}[(Z - \mathbb{E}[Z|W])(\mathbb{E}[Z|W] - g(W))]\\ &\qquad = \mathbb{E}[\mathbb{E}[(Z - \mathbb{E}[Z|W])(\mathbb{E}[Z|W] - g(W))]|W]\\ &\qquad = \mathbb{E}[(\mathbb{E}[Z|W] - g(W))]\mathbb{E}[Z - \mathbb{E}[Z|W]|W] = 0,\end{aligned}$$

as this last term becomes 0.

The idea of the tower law is the same as the law of total probability: we can break our probability space into a disjoint union of spaces where we know a given random variable, and weighting by their respective probabilities gives us the original expectation.

Probabilistic results can be applied conditionally. In particular, we have conditional Jensen.

Recall $f : \mathbb{R} \to \mathbb{R}$ is convex if

$$tf(x) + (1 - t)f(y) \geq f(tx + (1 - t)y),$$

for all $f \in [0, 1]$. If f is convex, then Jensen gives

$$\mathbb{E}[f(Z)|W] \geq f(\mathbb{E}[Z|W]).$$

1.2 Bayes Risk

We will prove that h_0 given in the first proposition is indeed a Bayes classifier. Note we have that

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

So $h_0(x)$ must minimiser over $h(x)$, the value

$$\begin{aligned}\mathbb{P}(Y \neq h(x)|X = x) &= \mathbb{P}(Y = 1, h(x) = -1|X = x) + \mathbb{P}(Y = -1, h(x) = 1|X = x)\\ &= \mathbb{P}(Y = 1|X = x)\mathbb{1}_{\{h(x)=-1\}} + \mathbb{P}(Y = -1|X = x)\mathbb{1}_{\{h(x)=1\}}\\ &= \eta(x)\mathbb{1}_{\{h(x)=-1\}} + (1-\eta(x))\mathbb{1}_{\{h(x)=1\}}.\end{aligned}$$

Now if $\eta(x) > 1/2$, then $h_0(x) = 1$, and if it is less than $1/2$, then $h_0(x) = -1$. If $\eta(x) = 1/2$, then either choice is constant, so any $h(x)$ minimises it.

1.3 Empirical Risk Minimisation

The risk $R(h)$ is the expectation of $\ell(h(X), Y)$, where $(X, Y) \sim P_0$.

The *empirical risk* or *training error* is the expectation of $\ell(h(X), Y)$ with respect to the empirical measure, i.e.

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

Given $\mathcal{H}$, then

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

is the (a) *empirical risk minimiser*.

Here $\hat{h}$ is a random variable that is dependent on the training data D.

Example 1.2.

In the regression setting with

$$\mathcal{H} = \{x \mapsto \mu + \beta^T x \mid \mu \in \mathbb{R}, \beta \in \mathbb{R}^p\},$$

the empirical risk minimiser is the ordinary least squares solution: $\hat{}(x) = \hat{\mu} + \hat{\beta}^T x$, where

$$(\hat{\mu}, \hat{\beta}) = \operatorname*{argmin}_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu - X_i^T\beta)^2.$$

More generally, if we consider

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{d} \phi_j(x)\beta_j \mid \beta \in \mathbb{R}^d \right\},$$

where each $\phi_j : \mathbb{R}^p \to \mathbb{R}$, then suppose we have $\Phi \in \mathbb{R}^{n \times d}$ with $\Phi_{ij} = \phi_j(x_i)$. Then writing $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))$, the empirical risk minimiser is

$$\hat{h}(x) = \phi(x)^T (\Phi^T \Phi)^{-1} \Phi^T Y_{1:n},$$

where $Y_{1:n} = (Y_1, \ldots, Y_n)$.

1.4 Bias-variance Tradeoff

Consider $\hat{h} = \hat{h}_D$ trained on data $D = (X_i, Y_i)_{i=1}^n$. Let $(X, Y) \perp\!\!\!\perp D$. Then we define

$$\bar{h}(x) : x \mapsto \mathbb{E}[\hat{h}_D(x)].$$

Recall we can split the squared error into two terms. Using this,

$$\begin{aligned}
\mathbb{E}[R(\hat{h}_D)] &= \mathbb{E}[(Y - \hat{h}_D(X))^2] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y|D, X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - \hat{h}_D(X))^2] \\
&= \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2]|X] + \mathbb{E}[(\hat{h}_D(X) - \mathbb{E}[\hat{h}_D(X)|X])^2] \\
&\qquad + \mathbb{E}[(\bar{h}(X) - \mathbb{E}[Y|X])^2] \\
&= \mathbb{E}[(\mathbb{E}[Y|X] - \bar{h}(X))^2] + \mathbb{E}[\mathrm{Var}(\hat{h}_D(X)|X)] + \mathbb{E}[\mathrm{Var}(Y|X)].
\end{aligned}$$

Here we have three terms, the *squared bias*, the *variance of* $\hat{h}_D$, and the *irreducible variance.*

Consider $\hat{h}_D$, and empirical risk minimizer over the class $\mathcal{H}$. We expect that large $\mathcal{H}$ will lead to low bias, but high variance, and a small $\mathcal{H}$ will lead to high bias, but low variance.

It is instructive to consider a decomposition involving

$$\tilde{h}_{X_{1:n}} : x \mapsto \mathbb{E}[\hat{h}_D(x)|X_{1:n}].$$

Then we can show that

$$\begin{aligned}
&\mathbb{E}[(Y - \hat{h}_D(X))^2|X = x] = \mathbb{E}[(\mathbb{E}[Y|X = x] - \tilde{h}_{X_{1:n}}(x))^2] \\
&+\mathbb{E}[(\hat{h}_D(x) - \tilde{h}_{X_{1:n}}(x))^2] + \mathrm{Var}(Y|X = x).
\end{aligned}$$

Consider the ERM $\hat{h}_D$, in where

$$\hat{h}_D(x) = \phi(x)^T(\Phi^T\Phi)^{-1}\Phi^T Y_{1:n}.$$

We further assume that $\mathrm{Var}(Y|X = x) = \sigma^2$, i.e. constant in x. Then

$$\begin{aligned}\mathbb{E}[(\tilde{h}_D(x) - \tilde{h}_{X_{1:n}}(x))^2|X_{1:n}]\\ = \mathbb{E}[(\phi(x)^T(\Phi^T\Phi)^{-1}\Phi^T(Y_{1:n} - \mathbb{E}[Y_{1:n}|X_{1:n})]^2 \mid X_{1:n}]\\ = \phi(x)^T(\Phi^T\Phi)^{-1}\Phi^T\mathbb{E}[(Y_{1:n} - \mathbb{E}[Y_{1:n} \mid X_{1:n}])(Y_{1:n} - \mathbb{E}[Y_{1:n} \mid X_{1:n}])^T \mid X_{1:n}]\\ \times\Phi(\Phi^T\Phi)^{-1}\phi(x).\end{aligned}$$

Note that $\mathbb{E}[Y_j \mid X_{1:n}] = \mathbb{E}[Y_j \mid X_j]$, so if A is the matrix in the middle,

$$\begin{aligned}A_{jk} &= \mathbb{E}[(Y_j - \mathbb{E}[Y_j \mid X_j])(Y_k - \mathbb{E}[Y_k \mid X_k]) \mid X_j, X_k]\\ &= \mathbb{E}[Y_jY_k \mid X_j, X_k] - \mathbb{E}[Y_j \mid X_j]\mathbb{E}[Y_k \mid X_k].\end{aligned}$$

If $j \neq k$, then

$$\begin{aligned}\mathbb{E}[Y_jY_k \mid X_j, X_k] &= \mathbb{E}[Y_j\mathbb{E}[Y_k \mid X_j, Y_j] \mid X_j, X_k]\\ &= \mathbb{E}[Y_j\mathbb{E}[Y_k \mid X_k] \mid X_j, X_k]\\ &= \mathbb{E}[Y_j \mid X_j]\mathbb{E}[Y_k \mid X_k].\end{aligned}$$

Thus $A_{jk} = 0$ if $j \neq k$. Moreover, $A_{jj} = \sigma^2$, by our assumption, hence $A = \sigma^2 I$. Therefore we get

$$\mathbb{E}[(\hat{h}_D(x) - \tilde{h}_{X_{1:n}}(x))^2 \mid X_{1:n}] = \sigma^2\phi(x)^T(\Phi^T\Phi)^{-1}\phi(x).$$

Consider averaging this quantity over $x = X_1, \dots, X_n$. This is

$$\begin{aligned}&\frac{1}{n}\operatorname{tr}(\sum_{i=1}^n \sigma^2\phi(X_i)^2(\Phi^T\Phi)^{-1}\phi(X_i))\\ =&\frac{\sigma^2}{n}\operatorname{tr}\left(\sum_{i=1}^n \phi(X_i)\phi(X_i)^T(\Phi^T\Phi)^{-1}\right) = \frac{\sigma^2 d}{n}.\end{aligned}$$

So interestingly this quantity is on average linear in d, the dimension of basis functions.

In the above, we have the following classifiers:

- $\hat{h}_D$. This is the empirical risk minimizer, and is a function of the data D.
- $\bar{h}$. This is the average value of $\hat{h}_D$, over training data D. For a large class $\mathcal{H}$ we expect that the variance of $\hat{h}_D$ will be large.

- $\tilde{h}_{X_{1:n}}$. This is the average value of $\hat{h}_D$, given the values $X_1, \ldots, X_n$. We use it as a proxy for $\bar{h}$ in our calculation for the variance of $\hat{h}_D$.

In the above, it is important to know that $X_{1:n} \perp\!\!\!\perp X$, as X is part of (X, Y), the independent sample we are trying to predict.

1.5 Cross Validation

Let $\tilde{h}^{(1)}, \ldots, \tilde{h}^{(m)}$ be m "machine learning" methods, that take in training data $D = (X_i, Y_i)_{i=1}^n$, all iid, and output a hypothesis $\hat{h}_D^{(j)} : \mathcal{X} \to \mathcal{Y}$.

Given a loss ℓ and associated risk R, we may wish to pick $\hat{h}^{(j)}$ to minimize the risk $R(\hat{h}_D^{(j)}) = \mathbb{E}[\ell(\hat{h}_D^{(j)}(X), Y) \mid D]$, where $(X, Y) \perp\!\!\!\perp D$, and the variable (X, Y) is equal in distribution to (X_1, Y_1).

It is easier to attempt to minimize the expected risk

$$\mathbb{E}[R(\hat{h}_D^{(j)})] = \mathbb{E}[\mathbb{E}[\ell(\hat{h}_D^{(j)}, Y)|D]].$$

The idea of v-fold cross validation estimates this by splitting the data into v folds (i.e. groups) of roughly equal size.

We first split the data into v folds, say $A_1, \ldots, A_v$. Let D_{-v} denote all the data, except that in the k'th fold. For each j, we apply $\hat{h}^{(j)}$ to data D_{-k}, to get model $\hat{h}_{-k}^{(j)} = \hat{h}_{D_{-k}}^{(j)}$. Then we choose j to minimize

$$\mathrm{CV}(j) = \frac{1}{n} \sum_{k=1}^{v} \sum_{i \in A_k} \ell(\hat{h}_{-k}^{(j)}(X_i), Y_i).$$

Writing $\hat{j}$ for the minimizer, we select the hypothesis $\hat{h}_D^{(j)}$.

Note that for each $i \in A_k$,

$$\mathbb{E}[\ell(\hat{h}_{-k}^{(j)}(X_i), Y_i)] = \mathbb{E}[R(\hat{h}_{D_{-k}}^{(j)})].$$

This is similar to $\mathbb{E}[R(\hat{h}^{(j)})]$, except with a dataset of size n replaced by one with size $n - |A_k|$. Thus $\mathrm{CV}(j)$ gives a biased estimate of this quantity, with this bias typically decreasing as v increases.

The choice $v = n$ can give the least bias, but computationally may be expensive, and the variance may be large as the summands tend to be positively correlated. Typical choices are $v = 5$ or 10.

One thing we can do is to cross validate with permuted data. This reduces the variance slightly.

2 Popular Machine Learning Methods

2.1 Decision Trees

Decision trees (also known as regression trees in the context we shall be interested in) construct hypotheses of the form

$$T : x \mapsto \sum_{j=1}^{J} \gamma_j \mathbb{1}_{R_j}(x),$$

where R_j are rectangular regions that partition all of $\mathbb{R}^p$, and $\gamma_j \in \mathbb{R}$ are coefficients.

Decision trees are built to both choose the coefficients and the regions, hence we are sort of letting the data decide the basis functions to use. What decision trees don't do is empirical risk minimisation over the entire class of functions, because it is computationally difficult and due to overfitting.

One point is that given a region R_j, it is easy to find the coefficient γ_j that minimizes the risk: it will be the average value of the output in that region. The idea behind decision trees is that the $J = 2$ case is an interesting and solvable problem, and using this case we are able to recurse and generate more trees.

The recursive binary partitioning scheme is as follows:

1. Choose a maximum number of regions J, and initialise our regions $\mathcal{R} = \{\mathbb{R}^p\}$.
2. For each region $R \in \mathcal{R}$ such that $I = \{i \mid X_i \in R\}$ has $|I| \geq 2$, we do the following: for each $j = 1, \dots, p$, let S_j be the midpoints between adjacent $\{X_{ij}\}_{i \in I}$. Find the predictor $\hat{j}_R$ and split point $\hat{s}_R$ to minimize over $j = 1, \dots, p$ and $s \in S_j$

$$\min_{\gamma_l \in \mathbb{R}} \sum_{\substack{i \in I \\ X_{ij} \leq s}} (Y_i - \gamma_l)^2 + \min_{\gamma_r \in \mathbb{R}} \sum_{\substack{i \in I \\ X_{ij} > s}} (Y_i - \gamma_R)^2 - \min_{\gamma \in \mathbb{R}} \sum_{i \in I} (Y_i - \gamma)^2.$$

3. Let $\hat{R}$ be the region yielding the lowest value, and set

$$\hat{R}_l = \{x \in \hat{R} \mid X_{\hat{j}_R} \leq \hat{s}_r\}, \qquad \hat{R}_r = \hat{R} \setminus \hat{R}_l.$$

4. Update $\mathcal{R} = (\mathcal{R} \setminus \{\hat{R}\}) \cup \{\hat{R}_l, \hat{R}_r\}$.
5. Repeat these steps until $|\mathcal{R}| = J$.
6. Writing $\mathcal{R} = \{\hat{R}_1, \dots, \hat{R}_J\}$, let $\hat{I}_j = \{i \mid X_i \in \hat{R}_j\}$, and set

$$\hat{\gamma}_j = \frac{1}{|\hat{I}_j|} \sum_{i \in I_j} Y_i.$$

7. Then our final tree is
$$\hat{T} : x \mapsto \sum_{i=1}^{J} \hat{\gamma}_j \mathbb{1}_{\hat{R}_j}(x).$$

Note while the computations seem expensive, they can be organized in an efficient manner as follows: consider for simplicity we have the first split, so $I = \{1, \dots, n\}$, and suppose $X_1 < X_2 < \cdots < X_n$.

Then minimization is equivalent to finding the index m to minimise the sum of

$$\begin{aligned}\sum_{i \le m} \left(Y_i - \frac{1}{m} \sum_{j \le m} Y_j \right)^2 + \sum_{i > m} \left(Y_i - \frac{1}{n-m} \sum_{j > m} Y_j \right)^2 \\ = \sum_{i=1}^{n} Y_i^2 - \frac{1}{m} A_m^2 - \frac{1}{n-m} B_m^2,\end{aligned}$$

where $A_m = Y_1 + \cdots + Y_m$, and $B_m = \sum Y_i - A_m$. But then $A_{m+1} = A_m + Y_{m+1}$, and $B_{m+1} = B_m - Y_{m+1}$. Thus we can compute $A_1, \dots, A_n$ and $B_1, \dots, B_n$ in $\mathcal{O}(n)$ times, so the minimization may be performed in $\mathcal{O}(n)$ time.

The reason they are called decision trees is that, at each point, we can view the output of the model as a tree. Each time it forms a tree, we make a binary decision as to which region to put our data in, given a binary classification of our data.

One thing to note is our decision to use rectangular regions R_j. This is a form of bias-variance tradeoff: we could allow more types of regions if we wanted to, but the search would become computationally harder and we risk overfitting. In the end, it depends on the situation we are using this model in. If we, for example, expect rotational invariance then allowing a broader set of models makes sense, whereas if we expect x_1 and x_2 to correspond to variables which would have no linear relation, then it does not.

Another way to use this is to do the same decision tree, after rotating the data ourself, and then averaging over these rotations.

2.2 Random Forests

While trees are fast to compute, they can be unstable, because the initial split can be sensitive to the data, and the piecewise constant fits may not be a good approximation to $x \mapsto \mathbb{E}[Y|X = x]$. Recall our bias-variance decomposition:

$$\mathbb{E}[R(\hat{T}_D)] = \mathbb{E}[(\mathbb{E}[Y|X] - \bar{T}(X))^2] + \mathbb{E}[\mathrm{Var}(\hat{T}_D(X) \mid X)] + \mathbb{E}[\mathrm{Var}(Y|X)],$$

where $\bar{T}(X) = \mathbb{E}[\hat{T}_D(X)]$. A *random forest* attempts to estimate $\bar{T}$. If we have multiple datasets $D_1, \dots, D_B$, we could do this by finding

$$\frac{1}{B}\sum_{b=1}^{B} \hat{T}_{D_b}.$$

Random forest replicates this by sampling D with replacement to give datasets $D_1^*, \dots, D_B^*$ and performs the following:

1. For each $b = 1, \dots, B$, fit a decision tree $\tilde{T}^{(b)} = \hat{T}_{D_b^*}$, but when searching for the best predictor to split on, randomly sample m_{try} of the p predictors and choose the best split among these.
2. Output the average
$$f_{\text{rf}} = \frac{1}{B}\sum_{b=1}^{B} \tilde{T}^{(b)}.$$

The rationale for sampling predictors is to make $\tilde{T}^{(b)}$ more independent. Indeed, consider $b_1 \neq b_2$, and $x \in \mathbb{R}^p$ with

$$\operatorname{Corr}(\hat{T}^{(b_1)}(x), \hat{T}^{(b_2)}(x)) = \rho > 0.$$

Then we find

$$\begin{aligned}
\operatorname{Var}(f_{\text{rf}}(x)) &= \operatorname{Var}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{T}^{(b)}(x)\right) \\
&= \frac{1}{B^2}\sum_{b=1}^{B} \operatorname{Var}(\hat{T}^{(b)}(x)) + \frac{1}{B^2}\sum_{b_1 \neq b_2} \operatorname{Cov}(\hat{T}^{(b_1)}(x), \hat{T}^{(b_2)}(x)) \\
&= \frac{1}{B}\operatorname{Var}(\hat{T}^{(1)}(x)) + \frac{B(B-1)}{B^2}\rho \operatorname{Var}(\hat{T}^{(1)}(x)) \\
&= \left(\frac{1}{B}(1-\rho) + \rho\right)\operatorname{Var}(\hat{T}^{(1)}(x)).
\end{aligned}$$

So we expect that increasing B will decrease the variance. This is true, up to a certain point, due to this factor of ρ. Hence while we may want to minimize the bias term by choosing the optimal split point for each tree, injecting randomness allows us to decrease ρ, and also tries to solve the greedy nature of the trees.

A small m_{try} reduces the variance, but increases the bias. We can choose a best m_{try} through cross validation.

3 Statistical Learning Theory

Recall that, in the regression setting, using ordinary least squares on a set of d basis functions yields

$$\mathbb{E}(\hat{h}_D(X) - \tilde{h}_{X_{1:n}}(X))^2 = \mathbb{E}R(\hat{h}_D) - \mathbb{E}R(\tilde{h}_{X_{1:n}}(X)) \approx \frac{\sigma^2 d}{n},$$

when $\mathrm{Var}(Y|X=x) = \sigma^2$, a constant in x.

Consider $\hat{h}$ an ERM over a general class $\mathcal{H}$. We will study the *excess risk*

$$R(\hat{h}) - R(h^*), \text{ where } h^* = \mathrm{argmin}_{h\in\mathcal{H}} R(h).$$

We don't really have much to go off of, except that

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}(\hat{h}) + \underbrace{\hat{R}(\hat{h}) - \hat{R}(h^*)}_{\leq 0} + \hat{R}(h^*) - R(h^*) \\ &\leq \sup_{h\in\mathcal{H}}\{R(h) - \hat{R}(h)\} + \hat{R}(h^*) - R(h^*). \end{aligned}$$

Consider $|\mathcal{H}|$ finite. Then

$$\begin{aligned} \mathbb{P}\left(\max_{h\in\mathcal{H}}\{R(h) - \hat{R}(h)\} > t\right) &= \mathbb{P}\left(\bigcup_{h\in\mathcal{H}}\{R(h) - \hat{R}(h) > t\}\right) \\ &\leq \sum_{h\in\mathcal{H}} \mathbb{P}(R(h) - \hat{R}(h) > t). \end{aligned}$$

3.1 Sub-Gaussiannity and Hoeffding's Inequality

Recall:

Proposition 3.1 (Markov's Inequality)**.** *Let W be a non-negative random variable. Then $t\mathbb{1}_{\{W\geq t\}} \leq W$. Taking expectation and dividing by t,*

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}[W]}{t}.$$

Applying this with a strictly increasing transformation $\phi : R \to (0, \infty)$ and any random variable W,

$$\mathbb{P}(W \geq t) = \mathbb{P}(\phi(W) \geq \phi(t)) \leq \frac{\mathbb{E}\phi(W)}{\phi(t)}.$$

Taking $\phi(t) = e^{\alpha t}$ for $\alpha > 0$, yields the so-called *Chernoff bound*

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha>0} e^{-\alpha t}\mathbb{E}[e^{\alpha W}].$$

Example 3.1.

Suppose $W \sim N(0, \sigma^2)$. Then

$$\mathbb{E}e^{\alpha W} = e^{\alpha^2\sigma^2/2},$$

which gives, from the Chernoff bound,

$$\begin{aligned}\mathbb{P}(W \geq t) &\leq \exp\left(\inf_{\alpha>0}\left[-\alpha t + \frac{\alpha^2\sigma^2}{2}\right]\right)\\ &= e^{-t^2/(2\sigma^2)}.\end{aligned}$$

Definition 3.1. Say a random variable W is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}e^{\alpha(W-\mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2},$$

for all $\alpha \in \mathbb{R}$.

Then we have the following tail bound:

Proposition 3.2. *If W is σ-sub-Gaussian, then*

$$\mathbb{P}(W - \mathbb{E}W \geq t) \leq e^{-t^2/(2\sigma^2)},$$

for all $t \geq 0$.

Note that if W is σ-sub-Gaussian, then it is also σ'-sub-Gaussian for $\sigma' \geq \sigma$. Moreover, $-W$ is also σ-sub-Gaussian, so

$$\mathbb{P}(|W - \mathbb{E}W| \geq t) \leq \mathbb{P}(W - \mathbb{E}W \geq t) + \mathbb{P}(-(W - \mathbb{E}W) \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$

Note we do not require W to be non-negative, as $\phi(x) = e^{\alpha x}$ automatically makes our random variables positive.

Example 3.2.

A *Rademacher random variable* ϵ takes values $+1$ or -1, each with equal probability $1/2$. It is 1-sub-Gaussian, as

$$\mathbb{E}e^{\alpha\epsilon} = \frac{1}{2}(e^{\alpha} + e^{-\alpha}) = \sum_{k=0}^{\infty}\frac{(\alpha^2)^k}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{(\alpha^2/2)^k}{k!} = e^{\alpha^2/2}.$$

Lemma 3.1 (Hoeffding's Lemma)**.** *If W is a random variable taking values in $[a, b]$, then W is σ-sub-Gaussian with parameter $(b - a)/2$.*

Proof: We prove a weaker version with $\sigma = b - a$. First, we claim if random variable Z is symmetric with $|Z| \leq m$, then

$$\mathbb{E}e^{\alpha Z} \leq e^{\alpha^2 m^2/2}.$$

Indeed, $Z \overset{d}{=} \epsilon Z$, where $\epsilon \perp\!\!\!\perp Z$ is a Rademacher random variable:

$$\begin{aligned}\mathbb{E}e^{\alpha Z} &= \mathbb{E}\mathbb{E}[e^{\alpha \epsilon Z} \mid Z] \\ &\leq \mathbb{E}e^{\alpha^2 Z^2/2} \leq e^{\alpha^2 m^2/2}.\end{aligned}$$

Let W' be an independent copy of W. Then

$$\begin{aligned}\mathbb{E}e^{\alpha(W - \mathbb{E}W')} &= \mathbb{E}e^{\alpha \mathbb{E}[W - W' | W]} \\ &\leq \mathbb{E}\mathbb{E}[e^{\alpha(W-W')} \mid W] \\ &= \mathbb{E}e^{\alpha(W-W')}.\end{aligned}$$

But $W - W'$ is symmetric, and $|W - W'| \leq b - a$.

Proposition 3.3. *Suppose $W_1, \ldots, W_n$ are independent with each W_i σ_i-sub-Gaussian. Then for $\gamma \in \mathbb{R}^n$, $\gamma^T W$ is σ-sub-Gaussian with*

$$\sigma = \left(\sum_{i=1}^n \sigma_i^2 \gamma_i^2\right)^{1/2}.$$

Proof: Let's look at

$$\begin{aligned}\mathbb{E}\exp\left(\alpha \sum_{i=1}^n \gamma_i (W_i - \mathbb{E}W_i)\right) &= \mathbb{E}\prod_{i=1}^n e^{(\alpha(W_i - \mathbb{E}W_i)\gamma_i} \\ &= \prod_{i=1}^n \mathbb{E}e^{\alpha(W_i - \mathbb{E}W_i)\gamma_i} \\ &\leq \prod_{i=1}^n e^{\alpha^2 \gamma_i^2 \sigma_i^2/2} \\ &= \exp\left(\frac{\alpha^2 \sum \gamma_i^2 \sigma_i^2}{2}\right).\end{aligned}$$

From the above, if $W_1, \ldots, W_n$ are independent random variables taking values in

$[a_i, b_i]$ for all i, then for $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \mathbb{E}W_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\frac{1}{n^2}\sum(\frac{b_i-a_i}{2})^2}\right) = \exp\left(-\frac{2n^2t^2}{\sum(b_i-a_i)^2}\right).$$

This is known as *Hoeffding's inequality.*

Proposition 3.4. *Suppose $W_1, \ldots, W_d$ are mean zero and σ-sub-Gaussian. Then*

$$\mathbb{E}\max_{j=1,\ldots,d} W_j \leq \sigma\sqrt{2\log d}.$$

Proof: Let $\alpha > 0$. We have

$$\mathbb{E}\max_{j=1,\ldots,d} e^{\alpha W_j} \leq \sum_{j=1}^{d}\mathbb{E}e^{\alpha W_j} \leq de^{\alpha^2\sigma^2/2}.$$

Now we know

$$\mathbb{E}\exp(\alpha\max_j W_j) \geq \exp(\alpha\mathbb{E}\max_j W_j),$$

by Jensen's inequality applied to $u \mapsto e^{\alpha u}$. Hence we have that

$$\mathbb{E}\max_j W_j \leq \inf_{\alpha>0}\left(\frac{\log d}{\alpha} + \frac{\alpha\sigma^2}{2}\right).$$

Take α such that

$$\frac{\log d}{\alpha^2} = \frac{\sigma^2}{2} \implies \alpha = \frac{\sqrt{2\log d}}{\sigma}.$$

This gives

$$\mathbb{E}\max_j W_j \leq \sigma\sqrt{2\log d}.$$

Now we are done with the theory of concentration inequalities. We continue to applications.

3.2 Finite Hypothesis Class

Theorem 3.1. *Consider the classification setting (with 0-1 loss), and suppose $|\mathcal{H}| < \infty$. With probability at least $1-\delta$, and ERM $\hat{h}$ satisfies*

$$R(\hat{h}) - R(h^*) \leq \sqrt{\frac{2(\log|\mathcal{H}| + \log(1/\delta))}{n}}.$$

If we take $\delta = 0.01$ and $|\mathcal{H}| = 10^n$, then the right hand side is about $\sqrt{2\log 10}\sqrt{\frac{r+2}{n}}$.

Proof: Let $t > 0$. Then

$$\begin{aligned}\mathbb{P}(R(\hat{h}) - R(h^*) > t) &= \mathbb{P}(R(\hat{h}) - R(h^*) > t, \hat{h} \neq h^*) \\ &\geq \mathbb{P}(R(\hat{h}) - \hat{R}(\hat{h})) > t/2, \hat{h} \neq h^*) \qquad (1) \\ &\quad + \mathbb{P}(\hat{R}(h^*) - R(h^*)) > t/2). \qquad (2)\end{aligned}$$

We bound (2) first. By Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}[\ell(h^*(X_i), Y_i) - \mathbb{E}\ell(h^*(X_i), Y_i)] > t/2\right) \leq e^{-nt^2/2}.$$

For (1), let $\mathcal{H}_- = \mathcal{H} \setminus \{h^*\}$. Then

$$\{R(\hat{h}) - \hat{R}(\hat{h}) > t/2, \hat{h} \neq h^*\} \subseteq \bigcup_{h \in \mathcal{H}_-} \{R(h) - \hat{R}(h) > t/2\}.$$

Then from a union bound, we can bound (1) by

$$\sum_{h \in \mathcal{H}_-} \mathbb{P}(R(h) - \hat{R}(h) > t/2) \leq (|\mathcal{H}| - 1)e^{-nt^2/2},$$

again by Hoeffding's inequality. Hence

$$\mathbb{P}(R(\hat{h}) - R(h^*) > t) \leq |\mathcal{H}|e^{-nt^2/2} = \delta,$$

if we set

$$t = \sqrt{-\frac{2}{n}\log\frac{\delta}{|\mathcal{H}|}} = \sqrt{\frac{2(\log|\mathcal{H}| + \log(1/\delta))}{n}}.$$

Example 3.3.

Consider the classification setting with $X \in [0,1)^2$. Partition $[0,1)^2$ into m^2 disjoint squares $R_1, \ldots, R_{m^2} \subset [0,1)^2$ of the form $[\frac{r}{m}, \frac{r+1}{m}) \times [\frac{s}{m}, \frac{s+1}{m})$, for $r, s = 0, \ldots, m-1$. Let

$$\bar{Y}_j = \operatorname{sgn}\left(\sum_{i | X_i \in R_j} Y_i\right),$$

and the final classifier (the histogram classifier) will be

$$j\hat{h}^{\text{hist}}(x) = \sum_{i=1}^{m^2} \bar{Y}_j \mathbb{1}_{R_j}(x).$$

Then $\hat{h}^{\text{hist}}$ is an ERM over $\mathcal{H}$ consisting of the 2^{m^2} classifiers, each corresponding to a way of assigning labels $+1$ or -1 to the regions $R_1, \dots, R_{m^2}$.

Then our result gives us, with probability at least $1-\delta$,

$$R(\hat{h}^{\text{hist}}) - R(h^*) \le \sqrt{\frac{2m^2 \log 2 + 2\log(1/\delta)}{n}}.$$

3.3 Rademacher Complexity

Recall that we are minimizing

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h) - \hat{R}(h^*) + \hat{R}(h^*) - R(h^*) \\ &\le \sup_{h\in\mathcal{H}}(R(h) - \hat{R}(h)) + \hat{R}(h^*) - R(h^*). \end{aligned}$$

Taking expected values, we get

$$\mathbb{E}[R(\hat{h}) - R(h^*)] \le \mathbb{E}G = \mathbb{E}\sup_{h\in\mathcal{H}}(R(h) - \hat{R}(h)).$$

Let $Z_i = (X_i, Y_i)$ for all i, and define

$$\mathcal{F} = \{(x,y) \mapsto -\ell(h(x), y) \mid h \in \mathcal{H}\}.$$

Then we can see

$$G = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[f(Z_i) - \mathbb{E}f(Z_i)].$$

To motivate the following definitions, consider the special case when $f(Z_i) \overset{d}{=} -f(Z_i)$. Then $\mathbb{E}f(Z_i) = 0$, and if $\varepsilon_{1:n}$ are iid Rademacher random variables with $\varepsilon_{1:n} \perp\!\!\!\perp Z_{1:n}$, then $(f(Z_i))_{i=1}^n \overset{d}{=} (\varepsilon_i f(Z_i))_{i=1}^n$. Thus

$$\mathbb{E}G = \mathbb{E}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) \mid Z_{1:n}\right].$$

Definition 3.2. Let $\mathcal{F}$ be any class of functions $f : \mathcal{Z} \to \mathbb{R}$, and let $z_1, \dots, z_n \in \mathcal{Z}$.

- Let $\mathcal{F}(z_{1:n}) = \{(f(z_i), \ldots, f(z_n)) \mid f \in \mathcal{F}\} \subseteq \mathbb{R}^n$.
- Define the *empirical Rademacher complexity*

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(z_i),$$

where $\varepsilon_{1:n}$ are iid Rademacher. Given iid $Z_1, \ldots, Z_n \perp\!\!\!\perp \varepsilon_{1:n}$, we can may also view the empirical Rademacher complexity as a random variable

$$\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i)\varepsilon_i \mid Z_{1:n}\right].$$

- Define the *Rademacher complexity* of $\mathcal{F}$ as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})).$$

Theorem 3.2. *Let $\mathcal{F}$ and random variables $Z_{1:n}$ be as in the definition. Then*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - \mathbb{E}f(Z_i)]\right] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof: Introduce an independent copy $(Z_i')_{i=1}^n$ of $(Z_i)_{i=1}^n$. Then,

$$\begin{aligned}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - \mathbb{E}f(Z_i')]\right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - f(Z_i')] \mid Z_{1:n}\right)\right] \\ &\leq \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - f(Z_i')]\right).\end{aligned}$$

Indeed for any collection of random variables V_t, we have

$$\mathbb{E}V_t \leq \mathbb{E}\sup_{t'} V_{t'} \implies \sup_t \mathbb{E}V_t \leq \mathbb{E}\sup_t V_t.$$

Take $\varepsilon_{1:n}$ Rademacher random variables with $\varepsilon_{1:n} \perp\!\!\!\perp (Z_{1:n}, Z'_{1:n})$. Then,

$$\begin{aligned}\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - f(Z_i')] &\overset{d}{=} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(Z_i) - f(Z_i')] \\ &\leq \sup_{f \in \mathcal{F}} \frac{1}{n} \varepsilon_i f(Z_i) + \sup_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (-\varepsilon_i) g(Z_i).\end{aligned}$$

Thus we get

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}[f(Z_i)-\mathbb{E}f(Z_i)]\right)\leq 2\mathcal{R}_n(\mathcal{F}).$$

Theorem 3.3. *Let $\mathcal{F}=\{(x,y)\mapsto\ell(h(x),y)\mid h\in\mathcal{H}\}$ (here loss can be arbitrary). Then*

$$\mathbb{E}[R(\hat{h})-R(h^*)]\leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof: Recall that

$$\mathbb{E}[R(\hat{h})-R(h^*)]\leq\mathbb{E}\sup_{h\in\mathcal{H}}[R(h)-\hat{R}(h)]\leq 2\mathcal{R}_n(-\mathcal{F}),$$

by our previous theorem, where $-\mathcal{F}=\{-f\mid f\in\mathcal{F}\}$. But if $\varepsilon_{1:n}$ are iid Rademacher random variables, then $\varepsilon_{1:n}\overset{d}{=}-\varepsilon_{1:n}$, so $\mathcal{R}_n(\mathcal{F})=\mathcal{R}_n(-\mathcal{F})$.

3.4 VC Dimension

Consider a classification setting (with 0-1 loss). Our aim is to bound $\mathcal{R}_n(\mathcal{F})$ where

$$\mathcal{F}=\{(x,y)\mapsto\ell(h(x),y)\mid h\in\mathcal{H}\}.$$

Note that if $x_{1:n}\in\mathcal{X}^n$, and $y_{1:n}\in\{-1,1\}^n$ with $z_i=(x_i,y_i)$, then

$$|\mathcal{F}(z_{1:n})|=|\mathcal{H}(x_{1:n})|,$$

as we can match $(\ell(h(x_i),y_i))_{i=1}^n$ to $(h(x_i))_{i=1}^n$.

Lemma 3.2. *The empirical Rademacher complexity*

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))\leq\sqrt{\frac{2\log|\mathcal{H}(x_{1:n})|}{n}}.$$

Proof: Let $|\mathcal{F}(z_{1:n})|=d$, and $\mathcal{F}'=\{f_1,\dots,f_d\}\subseteq\mathcal{F}$ such that $\mathcal{F}'(z_{1:n})=\mathcal{F}(z_{1:n})$. Given $\varepsilon_{1:n}$ iid Rademacher random variables, let

$$W_j=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_j(z_i).$$

Then

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E} \max_{j=1,\dots,d} W_j.$$

Recall that each ε_i is 1-sub-Gaussian, so each W_j is $n^{-1/2}$-sub-Gaussian. Hence

$$\mathbb{E} \max_{j=1,\dots,d} W_j \le \sqrt{\frac{2\log d}{n}}.$$

Definition 3.3. Let $\mathcal{H}$ be a class of functions with $h : \mathcal{X} \to \{a, b\}$, with $|\mathcal{H}| \ge 2$. Say $\mathcal{H}$ *shatters* $x_{1:n} \in \mathcal{X}^n$ if $|\mathcal{H}(x_{1:n})| = 2^n$, i.e. our hypothesis class contains any function on this set.

Define the *shattering coefficient*

$$s(\mathcal{H}, n) = \max_{x_{1:n} \in \mathcal{X}^n} |\mathcal{H}(x_{1:n})|.$$

Define the *VC dimension* $\mathrm{VC}(\mathcal{H})$ to be the largest n such that $x_{1:n}$ is shattered. If no such n exists, say $\mathrm{VC}(\mathcal{H}) = \infty$, i.e.

$$\mathrm{VC}(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid s(\mathcal{H}, n) = 2^n\}.$$

Remark. Note that if $|\mathcal{H}(x_{1:n})| = 2^n$, then $|\mathcal{H}(x_{1:m})| = 2^m$ for all $m \le n$, so $\mathrm{VC}(\mathcal{H}) = n$ if and only if it shatters $x_{1:n}$, and it does not shatter any $x_{1:n+1}$.

Example 3.4.

Take $\mathcal{H} = \{\mathbb{1}_{[a,b)} \mid b > a\}$. This is an infinite class. For n points, we find that $s(\mathcal{H}, n) \le (n+1)^2$ since $(\mathbb{1}_{[a,b)}(x_i)) = (\mathbb{1}_{[a',b')}(x_i))$ if and only if a, a' are in the same interval $[x_i, x_{i+1})$, and so are b, b' (except maybe if they are degenerate).

For $n = 2$, any set can be shattered, and no $x_{1:3}$ can be shattered, as for $x_1 < x_2 < x_3$, we cannot have $x_1, x_3 \mapsto 1$ and $x_2 \to 0$. So $\mathrm{VC}(\mathcal{H}) = 2$.

Lemma 3.3 (Sauer-Shelah Lemma)**.** *Let $\mathcal{H}$ be as in the definition with $\mathrm{VC}(\mathcal{H}) < \infty$. Then $s(\mathcal{H}, n) \le (n+1)^{\mathrm{VC}(\mathcal{H})}$.*

An important consequence is that

$$\mathcal{R}_n(\mathcal{F}) \le \sqrt{\frac{2\mathrm{VC}(\mathcal{H}) \log(n+1)}{n}}.$$

Example 3.5.

Let $\mathcal{X} = \mathbb{R}^p$, and consider

$$\mathcal{A} = \left\{ \prod_{j=1}^{p} (-\infty, a_j] \mid a_1, \dots, a_p \in \mathbb{R} \right\},$$

and $\mathcal{H} = \{\mathbb{1}_A \mid A \in \mathcal{A}\}$. We claim that $\mathrm{VC}(\mathcal{H}) = p$. Indeed, the set of basis vectors $e_1, \dots, e_p$ can be shattered. For $I \subseteq \{1, \dots, p\}$, set $a_j = 1$ if $j \in I$, and 0 else. Then

$$e_j \in \prod_{j=1}^{p} (-\infty, a_j] \iff j \in I.$$

Now take $x_1, \dots, x_{p+1} \in \mathbb{R}^p$. For each coordinate $j = 1, \dots, p$, let $I_j = \{k \mid x_{kj} \geq x_{lj} \text{ for all } l\}$. There there must be some k^* not a unique element of any $I_1, \dots, I_p$, but then no $h \in \mathcal{H}$ has $h(x_{k^*}) = 0$ and $h(x_k) = 1$ for $k \neq k^*$, so $x_1, \dots, x_{p+1}$ cannot be shattered and $\mathrm{VC}(\mathcal{H}) = p$.

Let $\mathcal{F}$ be a vector space of function $f : \mathcal{X} \to \mathbb{R}$, e.g. affine maps. We can form classifiers $\mathcal{H} = \{\operatorname{sgn} f \mid f \in \mathcal{F}\}$.

Proposition 3.5. *Let $\mathcal{H}$ be as above, for some vector space $\mathcal{F}$. Then $\mathrm{VC}(\mathcal{H}) \leq \dim(\mathcal{F})$.*

Proof: Let $d = \dim(\mathcal{H}) + 1$, and take $x_1, \dots, x_d \in \mathcal{X}$. We need to show these cannot be shattered. Consider the linear map $L : \mathcal{F} \to \mathbb{R}^d$ by $f \mapsto (f(x_1), \dots, f(x_d))^T$. Then since $\dim(\mathcal{L}(\mathcal{F})) \leq \dim \mathcal{F} = d - 1$, there exists $\gamma \in \mathbb{R}^d$ orthogonal to everything in $L(\mathcal{F})$ i.e.

$$\sum_{i:\gamma_i > 0} \gamma_i f(x_i) + \sum_{i:\gamma_i \leq 0} \gamma_i f(x_i) = 0,$$

for all $f \in \mathcal{F}$. Without loss of generality there is i with $\gamma_i > 0$. But then we can't have $h(x_i) = 1$ for $f(x_i) > 0$, and $h(x_i) = -1$ for $f(x_i) < 0$, otherwise the left hand side of the above expression would be strictly positive.

Thus $x_1, \dots, x_d$ cannot be shattered, so $\mathrm{VC}(\mathcal{H}) \leq d - 1 = \dim \mathcal{F}$.

Example 3.6.

Let $\mathcal{X} = [0,1)^2$, and $\mathcal{F}$ to be the polynomials of degree at most d. Taking $\mathcal{H} = \{\operatorname{sgn} \circ f \mid f \in \mathcal{F}\}$, we get

$$\dim(\mathcal{F}) \leq (d+1) + (d) + (d-1) + \cdots + 1 = \frac{(d+1)(d+2)}{2}.$$

Hence for $d = 5$, $\dim(\mathcal{F}) = 21$. For an ERM $\hat{h}$, we have

$$\mathbb{E}[R(\hat{h}) - R(h^*)] \leq 2\sqrt{\frac{2 \times 21 \log(n+1)}{n}}.$$

Comparing this to $\hat{h}^{\text{hist}}$,

$$\mathbb{E}[R(\hat{h}^{\text{hist}} - R(h^*)] \leq \frac{1}{2\sqrt{n}}\sqrt{2m^2 \log 2}.$$

4 Computation for Empirical Risk Minimisation

Computation of the ERM with 0-1 can be computationally very hard. We thus aim to approximate the ERM via a convex optimisation problem, i.e. minimizing a convex function over a convex set.

4.1 Convexity

$C \subseteq \mathbb{R}^d$ is *convex* if

$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } t \in (0,1).$$

The intersections of convex sets are convex.

Definition 4.1.

- For $S \subseteq \mathbb{R}^d$, the *convex hull* $\operatorname{conv} S$ of S is the intersection of all convex sets containing S.
- $v \in \mathbb{R}^d$ is a *convex combination* of $v_1, \dots, v_n \in \mathbb{R}^d$ if

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m$$

for some $\alpha_1, \dots, \alpha_m \geq 0$, and $\alpha_1 + \cdots + \alpha_m = 1$.

Lemma 4.1. *For $S \subseteq \mathbb{R}^d$, $v \in \operatorname{conv} S$ if and only if v is a convex combination of some points in S.*

Proof: Let D be the set of all convex combinations of points in S.

We show $D \subseteq \operatorname{conv} S$. Intersections of convex sets are convex, so we know $\operatorname{conv} S$ is a convex set containing S. Thus any convex combinations of two points from S is in $\operatorname{conv} S$.

Let's induct. Suppose for $m \geq 2$, that any convex combination of m points is in $\operatorname{conv} S$. Take $v_1, \dots, v_{m+1} \in S$, and $\alpha_1, \dots, \alpha_{m+1} \geq 0$ with $\alpha_1 + \cdots + \alpha_{m+1} = 1$. Now consider $v = \alpha_1 v_1 + \cdots + \alpha_{m+1} + v_{m+1}$.

If $a_{m+1} = 1$, then $v = v_{m+1} \in S$. If not, call $t = a_1 + \cdots + a_m > 0$. Then

$$v = t\left(\frac{a_1}{t} v_1 + \cdots + \frac{a_m}{t} v_m\right) + (1-t) v_{m+1}.$$

By our induction hypothesis, the first part is in $\operatorname{conv}(S)$, by our induction hypothesis. Hence the entire thing is in $\operatorname{conv} S$.

Lemma 4.2. *Let $S \subseteq \mathbb{R}^d$. For any linear map $L : \mathbb{R}^d \to \mathbb{R}^m$,* $\operatorname{conv} L(S) =$ $L(\operatorname{conv} S)$*.*

Proof: $v \in \operatorname{conv} L(S)$ if and only if there exists $m \in \mathbb{N}$ and $v_1, \dots, v_m \in S$, $\alpha_i \geq 0$ with $\alpha_1 + \cdots + \alpha_m = 1$ with

$$v = \sum_{j=1}^{m} \alpha_j L(v_j) = L\left(\sum_{i=1}^{m} \alpha_j v_j\right) \in L(\operatorname{conv} S).$$

Moreover, $v \in L(\operatorname{conv} S)$ if and only if v takes the form $L(\sum_{i=1}^{m} \alpha_j v_j)$.

4.2 Convex Functions

Let $C \subseteq \mathbb{R}^d$ be convex. Then $f : C \to \mathbb{R}$ is *convex* if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y),$$

for $t \in (0,1)$ and $x, y \in C$. It is *strictly convex* if this inequality is strict.

By the triangle inequality, every norm is a convex function. Moreover, any local minimum of a convex function is a global minimum: if $x \in C$ is such that for all $y \in C$, there exists some $0 < t < 1$ with $f(x) \leq f((1-t)x + ty)$, then

$$f(x) \leq (1-t)f(x) + tf(y) \implies f(x) \leq f(y).$$

Convex functions satisfy a lot of properties. See notes.

4.3 Convex Surrogates

Consider $\mathcal{H} = \{x \mapsto \operatorname{sgn}(\beta^T x) \mid \beta \in \mathbb{R}^p\}$. Then ERM involves minimizing over $\beta \in \mathbb{R}^p$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq \operatorname{sgn}(X_i^T \beta)} \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,0]}(Y_i X_i^T \beta),$$

by ignoring when $X_i^T \beta = 0$. This is nasty to work with as $\mathbb{1}_{(-\infty,0]}$ is a non-convex, discontinuous function. If we replace it with a convex function, the resulting objective function will be convex. We therefore make the following changes to our setup:

- Take $\mathcal{H}$ to be a family of functions $h : X \to \mathbb{R}$. We obtain the corresponding classifier by composition with sgn.

- We consider losses of the form

  $$\ell(h(x), y) = \phi(yh(x)),$$

  where $\phi : \mathbb{R} \to (0, \infty)$ is convex.

The ϕ-risk is then

$$R_\phi(h) = \mathbb{E}\phi(Yh(X)).$$

We want to minimize the empirical ϕ-risk

$$\hat{R}_\phi(h) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i h(X_i)).$$

Classical choices of ϕ include the:

- Hinge loss: $\phi(u) = \max(1 - u, 0)$.
- Logistic loss: $\phi(u) = \log_2(1 + e^{-u}) = \log(1 + e^{-u})/2$.
- Exponential loss $\phi(u) = e^{-u}$.

We would like the minimizer $h_{\phi,0}$ of $R_\phi(h)$ over all functions h (assuming it exists) to be such that $x \mapsto \operatorname{sgn}(h_{\phi,0}(x))$ is equivalent to a Bayes classifier $x \mapsto \operatorname{sgn}(\eta(x) - 1/2)$, where $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$.

To study this, we examine the conditional ϕ-risk

$$\mathbb{E}[\phi(Yh(X)) \mid X = x] = \phi(h(x))\eta(x) + \phi(-h(x))(1 - \eta(x)).$$

Let $C_\eta(\alpha) = \phi(\alpha)\eta + \phi(-\alpha)(1 - \eta)$, for generic $\eta \in [0, 1]$.

Definition 4.2. We say $\phi : \mathbb{R} \to [0, \infty)$ is *classification calibrated* if, for any $\eta \in [0, 1]$ with $\eta \neq 1/2$,

$$\inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) \leq \inf_{\alpha(2\eta - 1) \leq 0} C_\eta(\alpha).$$

Theorem 4.1. *Let $\phi : \mathbb{R} \to [0, \infty)$ be convex. Then ϕ is classification calibrated if it is differentiable at* 0*, with $\phi'(0) < 0$.*

Proof: C_η is convex and differentiable at 0 with

$$C'_\eta(0) = \phi'(0)\eta - (1 - \eta)\phi'(0) = (2\eta - 1)\phi'(0).$$

Suppose $\eta > 1/2$, so $C'_\eta(0) < 0$. By convexity,

$$C_\eta(\alpha) \geq C_\eta(0) + \alpha C'_n(0) \geq C_n(0)$$

for $\alpha \leq 0$, and also

$$\lim_{\alpha \downarrow 0} \frac{C_\eta(\alpha) - C_\eta(0)}{\alpha} = C'_\eta(0) < 0,$$

so there exists $\alpha^* > 0$ with $C_\eta(\alpha^*) < C_\eta(0) = \inf_{\alpha \leq 0} C_\eta(\alpha)$.

Note that $C_{1/2+\theta}(\alpha) = C_{1/2-\theta}(-\alpha)$ for $\theta \in [0, 1/2]$, so when $\eta < 1/2$, there is $\alpha^* < 0$ with

$$C_\eta(\alpha^*) < C_\eta(0) = \inf_{\alpha \geq 0} C_\eta(\alpha).$$

Being classification calibrated means an optimal classifier will be a Bayes classifier.

4.4 Rademacher Complexity

Recall theorem 8. It says when $\mathcal{F} = \{(x, y) \mapsto \phi(yh(x)) \mid h \in \mathcal{H}\}$, then

$$\mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) \leq 2\mathcal{R}_n(\mathcal{F}).$$

Lemma 4.3 (Contraction Lemma). *Let $r = \sup |h(x)|$, over $h \in \mathcal{H}$ and $x \in \mathcal{X}$. Suppose there exists $L \geq 0$ with*

$$|\phi(u) - \phi(u')| \leq L|u - u'|,$$

for all $u, u' \in [-r, r]$, so f is L-Lipschitz on $[-r, r]$. Then for any $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$, writing $z_i = (x_i, y_i)$, we have

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq L\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})),$$

so $\mathcal{R}_n(\mathcal{F}) \leq L\mathcal{R}_n(\mathcal{H})$.

In the simple case $\mathcal{H} = \{x \mapsto x^T\beta \mid \beta \in \mathbb{R}^p\}$,

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}\left[\sup_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \epsilon_i X_i^T \beta\right]$$

may not be finite.

4.5 Constraints

Suppose $\xi = \{x \in \mathbb{R}^p \mid \|x\|_2 \leq C\}$, and consider

$$\mathcal{H} = \{x \mapsto x^T\beta \mid \|\beta\|_2 \leq \lambda\},$$

for some $\lambda > 0$. Then

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) &= \frac{1}{n}\mathbb{E}\left[\sup_{\|\beta\|_2 \le \lambda} \sum_{i=1}^n \epsilon_i x_i^T \beta\right] \\ &= \frac{\lambda}{n}\mathbb{E}\left(\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_2\right) \\ &\le \frac{\lambda}{n}\left(\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_2^2\right)^{1/2}.\end{aligned}$$

This thing can be found:

$$\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_2^2 = \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^n \epsilon_i \epsilon_j x_i^T x_j\right) = \sum_{i=1}^n \|x_i\|_2^2 \le nC^2,$$

so we finally get

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \le \frac{\lambda C}{\sqrt{n}}.$$

This also extends to variables in which $\mathbb{E}\|X_i\|_2^2 < \infty$.

Example 4.1.

Take ϕ to be the hinge loss, which is Lipschitz with constant 1, and $\mathcal{H}$ as above. Then $\operatorname{sgn} \circ \hat{h}$ is known as the *support vector classifier*.

Then the above results give

$$\mathbb{E}R_\phi(\hat{h}) - R(h^*) \le \frac{2\lambda C}{\sqrt{n}}.$$

4.6 ℓ_1 constraint

Recall the ℓ_1 norm of $u \in \mathbb{R}^p$ is

$$\|u\|_1 = \sum_{j=1}^p |u_j|.$$

Now consider

$$\mathcal{H} = \{x \mapsto x^T\beta \mid \|\beta\|_1 \le 1\}.$$

This looks like a diamond in the x, y plane. To bound $\hat{\mathcal{R}}(\mathcal{H}(x_{1:n}))$, we may use the following fact:

Lemma 4.4. *For any $A \subseteq \mathbb{R}^n$,*

$$\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}(\operatorname{conv} A).$$

Note that $\{\beta \mid \|\beta\|_1 < 1\} = \operatorname{conv}(S)$, where

$$S = \bigcup_{j=1}^{p} \{-e_j, e_j\}.$$

Indeed, if $\|\beta\|_1 = 1$, then

$$\beta = \sum_{j=1}^{p} \beta_j e_j = \sum_{j=1}^{p} |\beta_j| \operatorname{sgn}(\beta_j) e_j \in \operatorname{conv} S.$$

Next, if $\|\beta\|_1 < 1$, then

$$\beta = t \frac{\beta}{\|\beta\|_1} + (1-t)\left(\frac{-\beta}{\|\beta\|_1}\right) = (2t-1)\frac{\beta}{\|\beta\|_1},$$

where $t = (\|\beta\|_1 + 1)/2$. Now recall that

$$\mathcal{H}(x_{1:n}) = \{\begin{pmatrix} x_1^T\beta & \cdots & x_n^T\beta \end{pmatrix}^T \mid \|\beta\|_1 \leq \lambda\},$$

so letting $L(\beta) = \begin{pmatrix} x_1^T\beta & \cdots & x_n^T\beta \end{pmatrix}^T$, then

$$\mathcal{H}(x_{1:n}) = L(\operatorname{conv} S) = \operatorname{conv} L(S).$$

Thus,

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) = \hat{\mathcal{R}}(L(S)) = \frac{\lambda}{n}\mathbb{E}\left(\max_{j=1,\ldots,p}\left|\sum_{i=1}^{n} \epsilon_i x_{ij}\right|\right),$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid Rademacher. Now, note

$$\pm \sum_{i=1}^{n} \epsilon_i x_{ij}$$

is σ-sub-G, with

$$\sigma = \left(\sum_{i=1}^{n} x_{ij}^2\right)^{1/2}.$$

Therefore,

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \frac{\lambda}{n} \max_{j=1,\dots,p} \left(\sum_{i=1}^{n} x_{ij}^2 \right)^{1/2} \sqrt{2\log(2p)}.$$

If $\mathcal{X} = [-1,1]^p$, then

$$\hat{\mathcal{R}}(\mathcal{H}(x_{1:n})) \leq \lambda \sqrt{\frac{2\log(2p)}{n}}.$$

Example 4.2.

Take ϕ to be the hinge loss, and $\mathcal{H}_1 = \{x^T\beta \mid \|\beta\|_1 \leq \lambda_1\}$. Suppose $\mathcal{X} = [-1,1]^p$. Then the ERM $\hat{h}_1$ and $h_1^* = \operatorname{argmin} R_\phi(h)$ satisfy

$$\mathbb{E}R_\phi(\hat{h}_1) - R_\phi(h_1^*) \leq 2\lambda_1 \sqrt{\frac{2\log(2p)}{n}}.$$

Recall if $\mathcal{H}_2 = \{x \mapsto x^T\beta \mid \|\beta\|_2 \leq \lambda\}$, then the ERM $\hat{h}_2$ and $h_2^* = \operatorname{argmin} R_\phi(h)$ satisfy

$$\mathbb{E}R_\phi(\hat{h}_2) - R_\phi(h_2^*) \leq 2\lambda_2 \sqrt{\frac{p}{n}}.$$

Suppose there is $\beta^0 \in \mathbb{R}^p$ such that $h_0 : x \mapsto x^T\beta^0$ minimises $R_\phi(h)$ over all functions $h \in \{x \mapsto x^T\beta\}$.

- If $\beta^0 = \sqrt{p}^{-1} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T$, then taking $\lambda_1 = \sqrt{p}$ gives $h_1^* = h_0$, so
$$\mathbb{E}R_\phi(\hat{h}_1) - R_\phi(h_0) = \mathcal{O}\left(\sqrt{\frac{p \log p}{n}} \right).$$
- If $\beta^0 = \sqrt{s}^{-1} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}$, then taking $\lambda_1 = \sqrt{s}$ gives $h_1^* = h_0$, so
$$\mathbb{E}R_\phi(\hat{h}_1) - R_\phi(h_0) = \mathcal{O}\left(\sqrt{\frac{s \log p}{n}} \right).$$

In both cases, taking $\lambda_2 = 1$ gives $h_2^* = h_0$, so

$$\mathbb{E}R_\phi(\hat{h}_2) - R(h_0) \leq 2\sqrt{\frac{p}{n}}.$$

Conclusion: if every predictor is equally important, we expect ERM with $\mathcal{H}_2$ to outperform $\mathcal{H}_1$. However, if only $s \ll p$ predictors are important, then we expect ERM with $\mathcal{H}_1$ to do much better than $\mathcal{H}_2$.

4.7 Projections onto Convex Sets

Suppose we wish to minimize a convex function $f : C \to \mathbb{R}$, where $C \subseteq \mathbb{R}^d$ is a convex set. Given a candidate minimizer $\beta \in C$, consider the Taylor expansion

$$f(z) \approx f(\beta) + \nabla f(\beta)^T (z - \beta).$$

Taking $z = \beta + \eta\delta$ some some $\eta > 0$ and direction δ, we have

$$\min_{\delta | \|\delta\|_2 \leq 1} f(\beta + \eta\delta) \approx f(\beta) + \eta \min_{\delta | \|\delta\|_2 = 1} \nabla f(\beta)^T \delta.$$

This suggest updating β by moving in the direction $-\nabla f(\beta)$; however, this may result in leaving C, so we need a sensible way of getting back into C.

Proposition 4.1. *Let $C \subseteq \mathbb{R}^d$ be a closed convex set. Then for each $x \in \mathbb{R}^d$, the minimizer $\pi_C(x)$ of $z \mapsto \|z - x\|_2$ over $z \in C$ exists, and is unique.*

We call $\pi_C(x)$ the projection *of x on C. Also,*

- $(x - \pi_C(x))^T (z - \pi_C(x)) \leq 0$, *for all $z \in C$, $x \in \mathbb{R}^d$.*
- $\|\pi_C(x) - \pi_C(z)\|_2 \leq \|x - z\|_2$, *for all $x, z \in \mathbb{R}^d$.*

Proof: For existence, let $\mu = \inf_{z \in C} \|x - z\|_2$, and $B = \{z \mid \|x - z\|_2 \leq \mu + 1\}$. Then we have

$$\inf_{z \in C} \|x - z\|_2 = \inf_{z \in C \cap B} \|x - z\|_2.$$

But the right hand side is a minimum of a continuous function $z \mapsto \|z - x\|_2$ over a compact set, so the infimum is achieved.

For uniqueness, notice that $z \mapsto \|z - x\|_2^2$ is strictly convex, as the Hessian is positive definite. Hence its minimum on a convex set is unique, if it exists. Now we prove the two properties.

- For arbitrary x, let $\pi_C(x) = \pi$. If $z \in C$, then $(1 - t)\pi + tz \in C$ for all $t \in [0, 1]$ by convexity, so

$$\begin{aligned} \|x - \pi\|_2^2 \leq \|x - (1 - t)\pi - tz\|_2^2 &= \|x - \pi + t(\pi - z)\|_2^2 \\ &= \|x - \pi\|_2^2 + 2t(x - \pi)^T (\pi - z) + t^2 \|\pi - z\|_2^2, \end{aligned}$$

hence we get

$$(x - \pi)^T(z - \pi) \leq \frac{1}{2}t\|\pi - z\|_2^2 \leq 0,$$

as this holds for all t.

- Let $z \in \mathbb{R}^d$ be arbitrary. Then from the previous,

  $$(x - \pi_C(x))^T(\pi_C - \pi_C(x)) \leq 0,$$

  and moreover by interchanging x and z,

  $$(\pi_C(z) - z)^T(\pi_C(z) - \pi_C(x)) \leq 0.$$

  Adding these, we get

  $$\|\pi_C(z) - \pi_C(x)\|_2^2 \leq |(x-z)^T(\pi_C(z) - \pi_C(x)| \leq \|x-z\|_2\|\pi_C(z) - \pi_C(x)\|_2.$$

  Then we get our inequality by dividing, or if not the projections are equal and we get our result anyway.

4.8 Subgradients

Recall that for a convex function $f : \mathbb{R}^d \to \mathbb{R}$ differentiable at x, we have

$$f(z) \geq f(x) + \nabla f(x)^T(z - x).$$

Definition 4.3. A vector $g \in \mathbb{R}^d$ is a *subgradient* of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathbb{R}^d$ if

$$f(z) \geq f(x) + g^T(z - x),$$

for all $z \in \mathbb{R}^d$. The set of subgradients of f at x is called the *subdifferential* of f at x, and is denoted by $\partial f(x)$.

Example 4.3.

If $\phi(u) = \max(1 - u, 0)$, then

$$\partial\phi(1) = [-1, 0].$$

These satisfy the following.

Proposition 4.2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then $\partial f(x)$ is non-empty, for all $x \in \mathbb{R}^d$.*

Proposition 4.3. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.*

Proposition 4.4. *Let $f, f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex.*

(i) *For $\alpha > 0$, $\partial(\alpha f)(x) = \{\alpha g \mid g \in \partial f(x)\}$.*

(ii) *$\partial(f_1 + f_2)(x) = \{g_1 + g_2 \mid g_1 \in \partial f_1(x), g_2 \in \partial f_2(x)\}$.*

(iii) *If $h : \mathbb{R}^m \to \mathbb{R}$ is $x \mapsto f(Ax + b)$, where $A \in \mathbb{R}^{d\times m}$, $b \in \mathbb{R}^d$, then*

$$\partial h(x) = \{A^T g \mid g \in \partial f(Ax + b)\}.$$

Example 4.4.

Consider

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i x_i^T \beta),$$

where $\phi(u) = \max(1 - u, 0)$. Let $h_i(\beta) = \phi(y_i x_i^T \beta)$. Then

$$\partial h_i(\beta) = \{y_i x_i t \mid t \in [-1, 0]\}$$

when $y_i x_i^T \beta = 1$. By these previous properties, the subgradients of f and β take the form

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i t_i,$$

where $t_i \in [-1, 0]$, and

$$t_i = \begin{cases} -1 & y_i x_i^T \beta < 1, \\ 0 & y_i x_i^T \beta > 1. \end{cases}$$

4.9 Gradient Descent

Our goal is to minimize $f(\beta)$ subject to $\beta \in C$, where C is closed and convex, and f is convex. Our input is a guess $\beta_1 \in C$, and we have fixed $k \in \mathbb{N}$, the number of step, and step sizes $(\eta_s)_{s=1}^{k-1}$.

Then for $s = 1$ to $k - 1$, we find $g_s \in \partial f(\beta_s)$, and pick

$$z_{s+1} = \beta_s - \eta_s g_s, \qquad \beta_{s+1} = \pi_C(z_{s+1}).$$

We could return the last element, but instead we will return the average

$$\bar{\beta} = \frac{1}{k}\sum_{s=1}^{k} \beta_s.$$

Theorem 4.2. *Suppose $\hat{\beta}$ minimises the convex function $f : C \to \mathbb{R}$, where C is a closed, convex set. Suppose:*

- $\sup_{\beta \in C} \|\beta\|_2 \leq R < \infty$.
- $\sup_{\beta \in C} \sup_{g \in \partial f(\beta)} \|g\|_2 \leq L < \infty$.

If $\eta_s = \eta = 2R/(L\sqrt{k})$, then

$$f(\hat{\beta}) - f(\bar{\beta}) \leq \frac{2LR}{\sqrt{k}}.$$

Proof: We have that

$$f(\hat{\beta}) \geq f(\beta_s) + g_s^T(\hat{\beta} - \beta_s).$$

Thus, rearranging,

$$\begin{aligned} f(\hat{\beta}_s) - f(\hat{\beta}) &\leq g_s^T(\beta_s - \hat{\beta}) \\ &= \frac{1}{\eta}(\beta_s - z_{s+1})^T(\beta_s - \hat{\beta}). \end{aligned}$$

We now use the relation

$$u^T v = \frac{1}{2}(\|u\|_2^2 + \|v\|_2^2 - \|u - v\|_2^2).$$

Passing this along, we get

$$\frac{1}{2\eta}(\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2).$$

Now note that

$$\|z_{s+1} - \hat{\beta}\|_2^2 \geq \|\pi_C(z_{s+1}) - \pi(\hat{\beta})\|_2^2 = \|\beta_{s+1} - \hat{\beta}\|_2^2,$$

so we get

$$f(\beta_s) - f(\hat{\beta}) \leq \frac{1}{2\eta}(\eta^2\|g\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2),$$

hence taking the sum, by using Jensen's inequality,

$$
\begin{aligned}
f(\bar{\beta}) - f(\hat{\beta}) &\leq \frac{1}{k}\sum_{s=1}^{k} f(\beta_s) - f(\hat{\beta}) \\
&\leq \frac{\eta L^2}{2} + \frac{\|\beta_s - \hat{\beta}\|_2^2}{2\eta k} \leq \frac{\eta L^2}{2} + \frac{2R^2}{2\eta k} = \frac{2RL}{\sqrt{k}}.
\end{aligned}
$$

Example 4.5.

Doing empirical risk minimization with hinge loss $\phi(u) = \max(1-u, 0)$, where

$$
\begin{aligned}
\mathcal{H} &= \{x \mapsto x^T\beta \mid \|\beta\|_2 \leq \lambda\}, \\
\chi &= \{x \in \mathbb{R}^p \mid \|x\|_2 \leq C\},
\end{aligned}
$$

then given data (x_i, y_i), recall that the subgradient of the objective function of the ERM problem

$$
f(\beta) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i x_i^T \beta)
$$

takes the form

$$
g = \frac{1}{n}\sum_{i=1}^{n} y_i x_i t_i,
$$

where $t_i \in [-1, 0]$. By the triangle inequality, $\|g\|_2 \leq C$. Gradient descent with $\eta_s = 2\lambda/(C\sqrt{k})$ gives

$$
f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{2C\lambda}{\sqrt{k}}.
$$

4.10 Stochastic Gradient Descent

An issue with gradient descent is that computation of the gradient can involve a sweep over the entire dataset. *Stochastic gradient descent* (SGD) instead works with unbiased estimates of subgradients as are available when minimizing convex functions of the form

$$
f(\beta) = \mathbb{E}\tilde{f}(\beta; u),
$$

where $\tilde{f} : \mathbb{R}^p \times U \to \mathbb{R}$ is such that $\beta \to \tilde{f}(\beta; u)$ is convex for all $u \in U$, and u is a random variable taking values in U.

Example 4.6. (Important)

Taking $u \sim \text{Unif}\{1, 2, \dots, n\}$, then the ERM objective with $\mathcal{H} = \{h_\beta \mid \beta \in C\}$ may be written as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(h_\beta(x_i), y_i) = \mathbb{E}\ell(h_\beta(x_u), y_u),$$

when $\beta \mapsto \ell(h_\beta(x), y)$ is convex for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

The stochastic gradient descent algorithm is as follows: we are given a guess $\beta_1 \in C$, and $k \in \mathbb{N}$, the number of steps, and step sizes $(\eta_s)_{s=1}^{k-1}$, as well as iid copies of $u_1, u_2, \dots, u_{k-1}$ of u. For $s = 1$ to $k-1$, we pick $\tilde{g}_s \in \partial\tilde{f}(\beta_s; u_s)$ and let

$$z_{s+1} = \beta_s - \eta\tilde{g}_s, \qquad \beta_{s+1} = \pi_C(z_{s+1}).$$

Then we again return the average

$$\bar{\beta} = \frac{1}{k}\sum_{s=1}^{k} \beta_s.$$

The key point is that in this ERM example, computing $\tilde{g}_s$ involves only a single data point (x_{u_s}, y_{u_s}).

Theorem 4.3. *Suppose $\hat{\beta}$ minimizes $f : \mathbb{R}^p \to \mathbb{R}$ over a closed convex set $C \subseteq \mathbb{R}^p$. Suppose*

- $\sup_{\beta\in C} \|\beta\|_2 \le R < \infty$, *and*
- $\sup_{\beta\in C} \mathbb{E}\sup_{g\in\partial f(\beta,u)} \|g\|_2^2 \le L^2$.

Then taking $\eta_s = \eta = 2R/(L\sqrt{k})$, then

$$\mathbb{E}f(\bar{\beta}) - f(\hat{\beta}) \le \frac{2LR}{\sqrt{k}}.$$

Proof: Let $g_s = \mathbb{E}[\tilde{g}_s \mid \beta_s]$. Then $g_s \in \partial f(\beta_s)$. Indeed,

$$\tilde{f}(\beta; u_s) \ge \tilde{f}(\beta_s; u_s) + \tilde{g}_s^T(\beta - \beta_s),$$

so taking the expectation and conditioning on β_s, and that $u_s \perp\!\!\!\perp \beta_s$,

$$\mathbb{E}\tilde{f}(\beta; u_s) = f(\beta) \ge f(\beta_s) + g_s^T(\beta - \beta_s),$$

for all β. Thus

$$\begin{aligned} f(\beta_s) - f(\hat\beta) &\le g_s^T(\beta_s - \hat\beta) \\ &= \mathbb{E}[\tilde g_s^T(\beta_s - \hat\beta) \mid \beta_s] = \cdots \\ &\le \frac{1}{2\eta}\mathbb{E}[\eta^2\|\tilde g_s\|_2^2 + \|\beta_s - \hat\beta\|_2^2 - \|\beta_{s+1} - \hat\beta\|_2^2 \mid \beta_s], \end{aligned}$$

arguing the same way as in the previous proof. Taking expectations and averaging,

$$\begin{aligned} \mathbb{E}f(\bar\beta) - f(\hat\beta) &\le \mathbb{E}\left[\frac{1}{k}\sum_{s=1}^{k} f(\beta_s)\right] - f(\hat\beta) \\ &\le \frac{\eta L^2}{2} + \frac{\|\beta_1 - \hat\beta\|_2^2}{2\eta k} = \frac{2RL}{\sqrt{k}}, \end{aligned}$$

as $\|\beta_1 - \hat\beta\|_2^2 \le 4R^2$.

5 Popular Machine Learning Methods II

5.1 Adaboost

Given a set $\mathcal{B}$ of base classifiers $h : \mathcal{X} \to \{-1, 1\}$ such that $h \in \mathcal{B} \implies -h \in \mathcal{B}$, consider

$$\mathcal{H} = \left\{ \sum_{m=1}^{M} \beta_m h_m \mid \beta_m \geq 0, h_m \in \mathcal{B} \text{ for } m = 1, \dots, M \right\}.$$

Empirical risk minimization over this class is not easy. *Adaboost* can be motivated as a greedy ERM over $\mathcal{H}$ using the exponential loss $\phi(u) = e^{-u}$, and it works as follows:

Set $\hat{f}_0$ to be $x \mapsto 0$. Then for $m = 1, \dots, M$, we pick

$$(\beta_m, \hat{h}_m) = \text{argmin}_{\beta \geq 0, h \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i(\hat{f}_{m-1}(x_i) + \beta h(x_i))),$$

and then let $\hat{f}_m = \hat{f}_{m-1} + \hat{\beta}_m \hat{h}_m$. We then output the classifier $\text{sgn} \circ \hat{f}_m$. The key property of Adaboost is that this minimization is easy to compute. Define weights

$$w_i^{(m)} = \frac{1}{n} \exp(-y_i \hat{f}_{m-1}(x_i)),$$

then note

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i(\hat{f}_{m-1}(x_i) + \beta h(x_i))) &= \sum_{i=1}^{n} w_i^{(m)} \exp(-\beta y_i h(x_i)) \\ &= e^{\beta} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{y_i \neq h(x_i)\}} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)} (1 - \mathbb{1}_{\{y_i \neq h(x_i)\}}) \\ &= (e^{\beta} - e^{-\beta}) \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{y_i \neq h(x_i)\}} + e^{-\beta} \sum_{i=1}^{n} w_i^{(m)}. \end{aligned}$$

We are sort of minimizing a weighted misclassification loss. Let

$$\text{err}_m(h) = \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{y_i \neq h(x_i)\}} / \sum_{i=1}^{n} w_i^{(m)}.$$

Then $\hat{h}_m = \text{argmin}_{h \in \mathcal{B}} \text{err}_m(h)$, so $\hat{h}_m$ is a weighted ERM over the class $\mathcal{B}$. If $\text{err}_m(\hat{h}_m) > 0$, then

$$(e^{\hat{\beta}_m} + e^{-\hat{\beta}_m}) \text{err}_m(\hat{h}_m) = e^{-\hat{\beta}_m}.$$

Let $x = e^{\hat{\beta}_m}$, and $a = \mathrm{err}_m(\hat{h}_m)$. We have

$$(x^2+1)a = 1 \implies x = \sqrt{a^{-1}-1},$$

which we can solve to get

$$\hat{\beta}_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}_m(\hat{h}_m)}{\mathrm{err}_m(\hat{h}_m)}\right).$$

To use Adaboost effectively, we need the weighted ERM step to be computationally fast.

Example 5.1.

Let $\mathcal{X} \in \mathbb{R}^p$, and

$$\beta = \{x \mapsto \mathrm{sgn}(x_j - a), x \mapsto -\mathrm{sgn}(x_j - a) \mid a \in \mathbb{R}, j = 1, \dots, p\},$$

the class of *decision stumps*.

5.2 Gradient Boosting

Imagine applying gradient descent directly to minimise $R(h) = \mathbb{E}\ell(h(X), Y)$ over functions h.

We start with an initial guess $f_0 : \mathcal{X} \to \mathbb{R}$. Then for $m = 1, \dots, M$, we compute

$$\begin{aligned} g_m(x) &= \frac{\partial}{\partial \alpha}\mathbb{E}[\ell(\alpha, Y) \mid X = x]\Bigg|_{f_{m-1}(x)} \\ &= \mathbb{E}\left[\frac{\partial}{\partial \alpha}\ell(\alpha, Y)\Bigg|_{f_{m-1}(x)} \mid X = x\right], \end{aligned}$$

and then let $f_m = f_{m-1} - \eta g_m$ (for some $\eta > 0$). The problem is that we don't have access to this conditional expectation. Recall that

$$x \mapsto \mathbb{E}\left[\frac{\partial}{\partial \alpha}\ell(\alpha, Y)\Bigg|_{f_{m-1}(x)} \mid X = x\right]$$

minimizes

$$\mathbb{E}\left[\left(\frac{\partial}{\partial \alpha}\ell(\alpha, Y)\Bigg|_{f_{m-1}(x)} - g(X)\right)^2\right],$$

over all functions $g : \mathcal{X} \to \mathbb{R}$. *Gradient boosting* estimates this conditional expectation function using some bass regression procedure $\hat{h}$ that takes in training data $\mathcal{D}$ and outputs a hypothesis $\hat{h}_{\mathcal{D}} : \mathcal{X} \to \mathbb{R}$, which works as follows:

Our input data is $Y_{1:n}$, $X_{1:n}$, and we set $\eta > 0$. The base regression method is $\hat{h}$, and the number of iterations is M. We compute

$$\hat{\mu} = \operatorname{argmin}_{\mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mu, Y_i),$$

and set $\hat{f}_0(x) = \hat{\mu}$. For $m = 1, \ldots, M$, we find

$$W_i = \left. \frac{\partial \ell(\alpha, Y_i)}{\partial \alpha} \right|_{\hat{f}_{m-1}(X_i)},$$

and then we set $\hat{g}_m = \hat{h}(X_{1:n}, W_{1:n})$, i.e. we regress $W_{1:n}$ onto $X_{1:n}$, and then set

$$\hat{f}_m = \hat{f}_{m-1} - \eta \hat{g}_m.$$

We then return $\hat{f}_M$, or $\operatorname{sgn} \circ \hat{f}_M$.

Example 5.2.

For least squares, $\ell(\alpha, Y) = (\alpha - Y)^2$, and so

$$\left. \frac{\partial \ell(\alpha, Y_i)}{\partial \alpha} \right|_{f_{m-1}(X_i)} = -2(Y_i - f_{m-1}(X_i)).$$

For least-squares, we can consider $\hat{h}$ as the ERM of $\{x \mapsto \mu + x_j \beta \mid \mu \in \mathbb{R}, \beta \in \mathbb{R}, j = 1, \ldots, p\}$.

5.3 Feedforward Neural Networks

Feedforward neural networks are based on a class of hypotheses $h : \mathbb{R}^p \to \mathbb{R}$ of the form

$$h(x) = A^{(d)} \circ g \circ A^{(d-1)} \circ g \circ \cdots \circ g \circ A^{(2)} \circ g \circ A^{(1)}(x),$$

where we have

- d, the depth.
- $A^{(k)}(v) = \beta^{(k)} v + \mu^{(k)}$, where $v \in \mathbb{R}^{m_k}$, $\beta^{(k)} \in \mathbb{R}^{m_{k+1} \times m_k}$, $\mu^{(k)} \in \mathbb{R}^{m_{k+1}}$.

- $g : \mathbb{R}^m \to \mathbb{R}^m$ (for arbitrary m) is

  $$g(v) = \begin{pmatrix} \psi(v_1) \\ \cdots \\ \psi(v_m) \end{pmatrix},$$

  for a non-linear *activation function* $\psi : \mathbb{R} \to \mathbb{R}$. Usually, ψ is the rectified linear unit (ReLU): $u \mapsto \max(0, u)$, but historically people used the sigmoid $u \mapsto 1/(1 + e^{-u})$.

The parameters $(\beta^{(k)}, \mu^{(k)})$ are learned from the data $(x_1, y_1), \ldots, (x_n, y_n)$ by "attempting" to minimise the empirical ϕ risk using stochastic gradient descent. Suppose that ϕ and ψ are differentiable.

To compute the gradient of $z = \phi(yh(x))$ with respect to the parameters, we use the chain rule as follows. Let $h^{(0)} = x$, the input layer. Define inductively for $k = 1, \ldots, d - 1$

$$x^{(k)} = A^{(k)}(h^{(k-1)}), \qquad h^{(k)} = g(x^{(k)}).$$

Hence $h^{(k)}$ is the k'th hidden layer. Then $x^{(d)}$ is the output layer. Now,

$$\frac{\partial z}{\partial x^{(d)}} = y\phi'(yx(d)),$$

$$\frac{\partial z}{\partial \mu^{(d)}} = \frac{\partial z}{\partial x^{(d)}}, \qquad \frac{\partial z}{\partial \beta^{(d)}_{1k}} = \frac{\partial z}{\partial x^{(d)}} h^{(d-1)}_k,$$

$$\frac{\partial z}{\partial h^{(d-1)}_j} = \frac{\partial z}{\partial x^{(d)}} \beta^{(d)}_{1j}, \qquad \frac{\partial z}{\partial x^{(d-1)}_j} = \frac{\partial z}{\partial h^{(d-1)}_j} \psi'(x^{(d-1)}_j),$$

$$\frac{\partial z}{\partial \mu^{(d-1)}_j} = \frac{\partial z}{\partial x^{(d-1)}_j}, \qquad \frac{\partial z}{\partial \beta^{(d-1)}_{jk}} = \frac{\partial z}{\partial x^{(d-1)}_j} h^{(d-2)}_k.$$

Index