STAT615 SP26 Lecture Notes: Statistical Learning

Yaqin Chen

Contents

0 Lecture 0: Introduction & Overview

0.1 Syllabus and Logistics

- Discuss syllabus.
- Exam Dates.
- Homeworks.
- Grading.

0.2 Brief Overview

We will be interested in the problem of **prediction** from **observed data** in the context of supervised learning.

- **Training Data:** $(X_1, Y_1), ..., (X_n, Y_n) \in (\mathcal{X} \times \mathcal{Y})$
- **Test Data:** (X, Y) where X is the input (Observed) and Y is the output (Not Known / Not Observed).

0.2.1 Goal

Use the test data to build a "predictor" that we can use to predict the output when we input X.

Note: To say something interesting about X from previously observed data, we have to imagine there is a connection between the training data and the test data.

Assumption: We assume that $(X_1, Y_1), ..., (X_n, Y_n), (X, Y)$ are all random variables obtained by sampling a joint distribution ρ (independent).

0.3 Classification and Regression

0.3.1 Example 1: Linear Regression

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \mathbb{R}$

Question: What do you do in linear regression?
We minimize the squared error:

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\langle \beta, X_i \rangle + \alpha))^2$$

Solution: Assuming invertibility, the solution is:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = (X^T X)^{-1} X^T \vec{Y}$$

Where the design matrix X typically includes a column of 1s for the intercept.

$$X \in \mathbb{R}^{n \times (d+1)}$$

In general, we solve the linear system:

$$X^T\vec{y} = (X^TX)\begin{pmatrix}\hat{\beta} \\ \hat{\alpha}\end{pmatrix}$$

After this, we can predict our output for X:

$$\hat{Y} = \langle\hat{\beta}, X\rangle + \hat{\alpha}$$

Question: Did we succeed? There is no way to really know for a single test, but we can try to see if we succeed in a **probabilistic sense**:

$$\mathbb{E}[\ell(\langle\beta, X\rangle + \alpha, Y)]$$

For example, consider the squared loss: $\ell(y, y') = |y - y'|^2$.

0.3.2 Theoretical Assumptions

1. Why did we restrict to linear predictors? Theoretical assumptions or practical reasons. But if we want more generality, shouldn't we try to find "universal" methods?

2. Why did we try to solve the empirical minimization?

$$\min_{\alpha,\beta} \frac{1}{n}\sum(Y_i - (\langle\beta, X_i\rangle + \alpha))^2$$

Later we will say that in fact, a reasonable thing to do is to solve:

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} \mathbb{E}_{(X,Y)\sim\rho}[(Y - (\langle\beta, X\rangle + \alpha))^2]$$

The problem: Typically we don't know what ρ is! How do we know that by solving the empirical version, we'll be close to solving the true expectation?

Later we will recognize this question as the **Statistical analysis of Empirical Risk Minimization (ERM)**.

0.3.3 Example 2: Quantile Regression

For a given X, linear regression outputs a single y. In quantile regression, we don't want to fully commit to a single output for the test point X.

Rather, we want to give a **confidence set**:

$$[L(X), U(X)]$$

For a given x, we output an interval $[L(x), U(x)]$, not just a point. This highlights the difference between point estimation and uncertainty quantification.

0.3.4 Example 3: Classification

$\mathcal{Y}$ is a finite set of labels or classes.

- $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{\text{cat}, \text{dog}\}$ (Binary)
- $\mathcal{Y} = \{0, 1, 2, ..., 9\}$ (Multiclass)

Goal: Construct a map $f : \mathcal{X} \to \mathcal{Y}$. Given a new observed X, it outputs a label in $\mathcal{Y}$ ($f(X)$).

Logistic Regression (Binary Case): For a given (α, β), consider:

$$p(x) = \frac{\exp(\langle \beta, x \rangle + \alpha)}{1 + \exp(\langle \beta, x \rangle + \alpha)}$$

Then define:

$$f(x) = \begin{cases} 1 & \text{if } p(x) \geq \frac{1}{2} \\ 0 & \text{if } p(x) < \frac{1}{2} \end{cases}$$

Questions:

- How do we obtain (α, β)?
- Is logistic regression the only thing we can do?
- Under what conditions is it the best thing?
- If we tune (α, β) using our training data, how do we know that we have succeeded?

0.3.5 Example 4: Uncertainty Quantification (UQ) in Classification

Approach 1: Suppose that for a given x, we output probabilities $p(x) \in [0,1]^{|\mathcal{Y}|}$.

$$p(x) = (p_1(x), ..., p_k(x))$$

This tells us how confident I am about X belonging to class k. (0: Not confident at all, 1: As confident as we can be).

Approach 2: Suppose that for a given X, we output a subset $C(X) \subseteq \mathcal{Y}$. $C(X)$ does not have to be a singleton and it contains the classes x is believed to belong to.

0.4 Course Roadmap

Exam 1:

- Classification, Regression, Quantile Regression
- Proximity classifiers
- Risk minimization
- Statistical properties of classifiers
- Focus on binary classification

Exam 2:

- Support vector machines (SVM) and other linear classifiers
- Reproducing Kernel Hilbert Spaces (RKHS)

Final Exam:

- All the above
- Neural Networks vs Kernel machines
- Conformal Prediction

1 Lecture 1: Classification & The Bayes Classifier

1.1 Goal and Setup

- Classification problem.
- Bayes classifier.

Notation:

- $\mathcal{X}$ = input space (e.g., $\mathbb{R}^d$)
- $\mathcal{Y}$ = output space
- x = feature vector, input, data point
- y = label

Supervised Learning: We seek a function $f : \mathcal{X} \mapsto \mathcal{Y}$. We learn from a given dataset $(x_1, y_1), ..., (x_n, y_n)$.

Examples:

- **MNIST:** x_i is an image of a handwritten digit. $y_i \in \{0, 1, ..., 9\}$.
- **CIFAR 10:** x_i is an image (airplane, automobile, bird, cat, etc.). $y_i \in \{1, ..., k\}$.
- **Traffic Signs:** x_i is an image of a traffic sign. y_i could be "Stop" or "50mph".

Distinctions:

- If $\mathcal{Y} = \{1, ..., k\}$ (discrete), it is **Classification**.
- If $\mathcal{Y} = \mathbb{R}$ or $\mathbb{R}^d$, it is **Regression**.

1.2 Distributional Assumptions

The data $(x_1, y_1), ..., (x_n, y_n)$ come from a distribution of input-output pairs.

- ρ is a probability distribution over $\mathcal{X} \times \mathcal{Y}$.
- (X, Y) is a random variable.

Example: $\mathcal{Y} = \{0, 1\}$. It could be that $\mathbb{P}(Y = 1|X = x) > 0$ and $\mathbb{P}(Y = 0|X = x) > 0$. The mapping $x \mapsto Y(x) = y$ is the "Ground Truth".

1.3 Minimizing Error

Our goal is to build the "best" possible classifier. Given $f : \mathcal{X} \to \mathcal{Y}$, compute the average misclassification error:

$$\text{Error} = \mathbb{P}(f(X) \neq Y)$$

Goal: Find f that makes the average misclassification error as small as possible.

Questions:

1. Does there exist a "best" classifier?
2. Can we construct this "best" classifier? (If we only observe the data $(X_1, Y_1), ..., (X_n, Y_n)$).
3. If we cannot build this classifier from the observed data, then what can we do in that case?

1.4 Loss Functions and Risk

Definition: The 0-1 Loss

$$\ell : \{1, ..., k\} \times \{1, ..., k\} \to \mathbb{R}$$

$$\ell(y, y') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{if } y = y' \end{cases}$$

Relative to this loss function and relative to ρ, we can define the **risk** of a given classifier $f : \mathcal{X} \to \mathcal{Y} = \{1, ..., k\}$:

$$R(f) := \mathbb{E}_{(X,Y)\sim\rho}[\ell(f(X), Y)] = \mathbb{P}_{(X,Y)\sim\rho}[f(X) \neq Y]$$

We want to solve: $\min_f R(f)$.

Question: Can you think of settings where other types of loss functions are more appropriate than the 0-1 loss?

General Risk:

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

Example (Cost Sensitivity): Suppose y_1 = Stop Sign, y_2 = 50mph, y_3 = 40mph. If the true y = Stop Sign, predicting 50mph seems worse than predicting 40mph. However, 0-1 loss penalizes them equally (loss = 1). A different loss function might be needed here.

1.5 The Bayes Classifier

Theorem (Bayes Classifier): Let ℓ be the 0-1 loss and $\mathcal{Y} = \{0, 1\}$ (binary problem). Let:

$$f_B(x) := \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{otherwise} \end{cases}$$

Then f_B (the Bayes Classifier) minimizes the average misclassification error.

Observations:

1. f_B depends on ρ (the true distribution).
2. We expect f_B constructed properly to do well at classifying (X, Y).
3. **Bayes Rate:** The risk associated with the Bayes classifier.

Alternative Formulation: Using Bayes' rule, the condition $\mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x)$ is equivalent to:

$$\rho_{X|Y=1}(x) \cdot \mathbb{P}(Y = 1) \geq \rho_{X|Y=0}(x) \cdot \mathbb{P}(Y = 0)$$

Or:

$$\mathbb{P}(X = x, Y = 1) \geq \mathbb{P}(X = x, Y = 0)$$

(This is useful when the input space $\mathcal{X}$ is discrete).

2 Lecture 2: Optimality of the Bayes Classifier

2.1 Goal Revisited

We seek a function $f : \mathcal{X} \to \{0, 1\}$ to minimize the risk:

$$\min_f \mathbb{P}(f(X) \neq Y)$$

Implicitly, we are working with the 0-1 loss.

2.2 The Bayes Classifier

Theorem: Given a distribution ρ over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{0, 1\}$, let:

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

The **Bayes Classifier** is defined as:

$$f_B(x) := \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{if } \mathbb{P}(Y = 0|X = x) > \mathbb{P}(Y = 1|X = x) \end{cases}$$

Then f_B minimizes the average misclassification error.

2.2.1 Proof of Optimality

Proof: Let $f : \mathcal{X} \to \{0, 1\}$ be any classifier. We want to show that:

$$\mathbb{P}(f(X) \neq Y) \geq \mathbb{P}(f_B(X) \neq Y)$$

The key idea is to use conditional expectations and the **Tower Property** of conditional expectation:

$$\mathbb{P}(f(X) \neq Y) = \mathbb{E}[\mathbf{1}_{f(X)\neq Y}] = \mathbb{E}[\mathbb{E}[\mathbf{1}_{f(X)\neq Y}|X]]$$

Let us focus on how to choose f at a specific x to make the conditional risk $\mathbb{E}[\mathbf{1}_{f(X)\neq Y}|X = x]$ smallest. There are two options in this binary case:

1. $f(x) = 1$: In this case, the error probability is:

$$\mathbb{E}[\mathbf{1}_{1\neq Y}|X = x] = \mathbb{P}(Y \neq 1|X = x) = \mathbb{P}(Y = 0|X = x)$$

2. $f(x) = 0$: In this case, the error probability is:

$$\mathbb{E}[\mathbf{1}_{0\neq Y}|X = x] = \mathbb{P}(Y \neq 0|X = x) = \mathbb{P}(Y = 1|X = x)$$

The best choice is the one that results in the smaller error.

- If $\mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x)$, then $\mathbb{P}(Y = 0|X = x)$ is smaller, so we should choose $f(x) = 1$.
- If $\mathbb{P}(Y = 0|X = x) > \mathbb{P}(Y = 1|X = x)$, then $\mathbb{P}(Y = 1|X = x)$ is smaller, so we should choose $f(x) = 0$.

This matches exactly the definition of $f_B(x)$. Thus, for any f:

$$\mathbb{E}[\mathbf{1}_{f(X)\neq Y}|X = x] \geq \mathbb{E}[\mathbf{1}_{f_B(X)\neq Y}|X = x]$$

Since this is true for every x, we can take the expectation over X and get:

$$\mathbb{P}(f(X) \neq Y) \geq \mathbb{P}(f_B(X) \neq Y)$$

□

2.3 Bayes Risk

The minimal achievable risk is called the **Bayes Risk**, denoted R_B^*:

$$R_B^* := \min_f \mathbb{P}(f(X) \neq Y) = \mathbb{P}(f_B(X) \neq Y)$$

Observations:

1. Regardless of the distribution, $R_B^* \leq 1/2$ (since we can always achieve at least 50% accuracy by predicting the most common class or random guessing in the worst case).
2. Both f_B and R_B^* depend on ρ.
3. If we change the loss function (e.g., cost-sensitive classification), the formula for f_B would change.

2.4 Examples

2.4.1 Example 1: Discrete Case

Let $\mathcal{X} = \{a, b, c, d, e, f\}$ and $\mathcal{Y} = \{0, 1\}$. Consider the unnormalized scores (or joint probabilities scaled) $q(x, y)$ given in the table below:

x	a	b	c	d	e	f
$q(x,0)$	0	0.1	0.2	0.5	0.8	0.9
$q(x,1)$	0.9	0.3	0.1	0.3	0.1	0.3

The Bayes classifier $f_B(x)$ selects the class with the higher score:

- $f_B(a) = 1$ (since $0.9 > 0$)
- $f_B(b) = 1$ (since $0.3 > 0.1$)
- $f_B(c) = 0$ (since $0.1 < 0.2$)
- $f_B(d) = 0$ (since $0.3 < 0.5$)
- $f_B(e) = 0$ (since $0.1 < 0.8$)
- $f_B(f) = 0$ (since $0.3 < 0.9$)

2.4.2 Example 2: Gaussian Mixture Models (GMM)

Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. Let $\mathbb{P}(Y = 1) = w_1$ and $\mathbb{P}(Y = 0) = w_0$. The conditional distributions are Gaussian:

$$X|(Y = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$X|(Y = 0) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

The Bayes classifier is:

$$f_B(x) = 1 \iff w_1\rho_1(x) \geq w_0\rho_0(x)$$

where $\rho_k(x)$ is the PDF of class k. Substituting the Gaussian density:

$$w_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \geq w_0 \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

Taking logs leads to the decision boundary.

- If $\sigma_1 \neq \sigma_0$, we get a quadratic term in x, leading to **QDA (Quadratic Discriminant Analysis)**.
- If $\sigma_1 = \sigma_0$, the quadratic terms cancel, leaving a linear inequality in x, leading to **LDA (Linear Discriminant Analysis)**.

2.5 Plug-in Classifiers

In practice, we do not know ρ (and thus $\eta(x)$). We have training data $(X_1, Y_1), ..., (X_n, Y_n)$. Recall that $f_B(x) = 1$ if $\eta(x) \geq 1/2$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$. Also note that $\mathbb{P}(Y = 0|X = x) = 1 - \eta(x)$, so the condition is equivalent to $\eta(x) \geq 1 - \eta(x) \implies \eta(x) \geq 1/2$.

Idea: Replace the unknown $\eta(x)$ with an estimate $\hat{\eta}_n(x)$ constructed from the training data.

$$\hat{\eta}_n(x) = \hat{\eta}_n(x; (X_1, Y_1), ..., (X_n, Y_n))$$

The resulting classifier is:

$$\hat{f}_n(x) = \begin{cases} 1 & \text{if } \hat{\eta}_n(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

This is called a **Plug-in Classifier**.

3 Lecture 3: Consistency of Plug-in Classifiers

3.1 Risk Bound for Plug-in Classifiers

To understand how good $\hat{f}_n$ is, we need to relate the classification risk to the quality of the estimation of $\eta(x)$.

Lemma: Let f be any classifier of the form:

$$f(x) = \begin{cases} 1 & \text{if } \alpha(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

for some function $\alpha(x)$. Then:

$$0 \leq R(f) - R_B^* \leq 2\mathbb{E}_X[|\eta(X) - \alpha(X)|]$$

Or more precisely:

$$R(f) - R_B^* = \mathbb{E}_X[|2\eta(X) - 1| \cdot \mathbf{1}_{f(X) \neq f_B(X)}]$$

3.1.1 Proof of Lemma

We use the tower property to write the risks.

$$R(f) - R_B^* = \mathbb{E}_X[\mathbb{E}[\mathbf{1}_{f(X) \neq Y} - \mathbf{1}_{f_B(X) \neq Y}|X]]$$

Let's analyze the term inside the expectation for a fixed x:

$$\mathbb{E}[\mathbf{1}_{f(x) \neq Y} - \mathbf{1}_{f_B(x) \neq Y}|X = x] = \mathbb{P}(f(x) \neq Y|X = x) - \mathbb{P}(f_B(x) \neq Y|X = x)$$

This difference is 0 if $f(x) = f_B(x)$. If $f(x) \neq f_B(x)$:

- Case A: $f(x) = 1, f_B(x) = 0$. Since $f_B(x) = 0 \implies \eta(x) < 1/2$. Difference $= (1 - \eta(x)) - \eta(x) = 1 - 2\eta(x) = |2\eta(x) - 1|$.
- Case B: $f(x) = 0, f_B(x) = 1$. Since $f_B(x) = 1 \implies \eta(x) \geq 1/2$. Difference $= \eta(x) - (1 - \eta(x)) = 2\eta(x) - 1 = |2\eta(x) - 1|$.

Thus, $R(f) - R_B^* = \mathbb{E}_X[|2\eta(X) - 1| \cdot \mathbf{1}_{f(X) \neq f_B(X)}]$.

Now, we show that $|2\eta(x) - 1| \cdot \mathbf{1}_{f(x) \neq f_B(x)} \leq 2|\eta(x) - \alpha(x)|$.

- **Case 1:** $f(x) \neq f_B(x)$ and $\eta(x) < 1/2$. Here $f_B(x) = 0$, so $f(x) = 1$, which implies $\alpha(x) \geq 1/2$. We have $|2\eta(x) - 1| = 1 - 2\eta(x)$. Also $2|\eta(x) - \alpha(x)| = 2(\alpha(x) - \eta(x))$ (since $\alpha \geq 1/2 > \eta$). Is $1 - 2\eta(x) \leq 2\alpha(x) - 2\eta(x)$? Yes, because $1 \leq 2\alpha(x) \iff 1/2 \leq \alpha(x)$.
- **Case 2:** $f(x) \neq f_B(x)$ and $\eta(x) \geq 1/2$. Here $f_B(x) = 1$, so $f(x) = 0$, which implies $\alpha(x) < 1/2$. We have $|2\eta(x) - 1| = 2\eta(x) - 1$. Also $2|\eta(x) - \alpha(x)| = 2(\eta(x) - \alpha(x))$. Is $2\eta(x) - 1 \leq 2\eta(x) - 2\alpha(x)$? Yes, because $-1 \leq -2\alpha(x) \iff 1/2 \geq \alpha(x)$.

Therefore, substituting this back into the expectation:

$$R(f) - R_B^* \leq 2\mathbb{E}_X[|\eta(X) - \alpha(X)|]$$

$\square$

3.2 Corollary for Learned Classifiers

Let $(X_1, Y_1), ..., (X_n, Y_n) \sim \rho$ and let $\hat{f}_n$ be the plug-in classifier using $\hat{\eta}_n$. Then:

$$\mathbb{E}_{\text{train}}[R(\hat{f}_n)] - R_B^* \leq 2\mathbb{E}_{\text{train},X}[|\eta(X) - \hat{\eta}_n(X)|]$$

This means that if our estimator $\hat{\eta}_n$ converges to the true η in L_1 norm (average absolute error), then the classifier $\hat{f}_n$ converges to the optimal Bayes classifier in terms of risk.

3.3 Proximity-Based Estimators

How do we construct $\hat{\eta}_n(x)$? A common approach is using a weighted average of training labels:

$$\hat{\eta}_n(x) = \sum_{i=1}^{n} Y_i w_{in}(x)$$

where weights $w_{in}(x) \geq 0$ and $\sum w_{in}(x) = 1$. The intuition is that $w_{in}(x)$ should be large if X_i is "close" to x.

Example (Kernel Regression): Let $\mathcal{X} = \mathbb{R}^d$.

$$w_{in}(x) = \frac{K\left(\frac{\|x-X_i\|}{\gamma_n}\right)}{\sum_{j=1}^{n} K\left(\frac{\|x-X_j\|}{\gamma_n}\right)}$$

Common choice is the Gaussian Kernel: $K(u) = \exp(-u^2)$. Here γ_n is a bandwidth parameter.

3.4 Consistency

Our goal is to study conditions on the weights that guarantee:

1. **Weak Consistency (in expectation):**

$$\lim_{n\to\infty} \mathbb{E}[R(\hat{f}_n)] - R_B^* = 0$$

   This can be reduced to showing $\lim_{n\to\infty} \mathbb{E}[|\eta(X) - \hat{\eta}_n(X)|] = 0$.

2. **Strong Consistency (almost surely):**

$$\lim_{n\to\infty} R(\hat{f}_n) - R_B^* = 0 \quad \text{a.s.}$$

   (Note that $R(\hat{f}_n)$ is a random variable depending on the training data).

4 Lecture 4: Consistency of Classifiers

4.1 Recap and Motivation

We have discussed two main approaches:

1. **Plug-in Classifiers (Similarity Classifiers):** Approximating the Bayes classifier $f_B(x)$ which uses the 0-1 loss.

$$f_B(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 0|X = x) \\ 0 & \text{else} \end{cases}$$

2. **Empirical Risk Minimization (ERM):** Finding $\hat{f}$ that minimizes the empirical risk, motivating $R(f) = \mathbb{P}(f(X) \neq Y)$.

Goal: To build $\hat{f}_n$ in such a way that $R(\hat{f}_n) - R_B^*$ goes to zero.

4.2 Notions of Consistency

Let $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ be the training data.

Definition 1 (Consistency)**.** *We say that a sequence of classifiers $\{\hat{f}_n\}$ is **consistent** for the distribution ρ if:*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n}[R_\rho(\hat{f}_n)] - R_{B,\rho}^* = 0$$

Here $R_{B,\rho}^* = \min_f \mathbb{P}(f(X) \neq Y)$ is the Bayes risk (a number), and $R_\rho(\hat{f}_n) = \mathbb{P}(\hat{f}_n(X) \neq Y|\mathcal{D}_n)$ is a random variable depending on the training data.

Definition 2 (Strong Consistency)**.** *We say $\{\hat{f}_n\}$ is **strongly consistent** for ρ if:*

$$\lim_{n\to\infty} (R_\rho(\hat{f}_n) - R_{B,\rho}^*) = 0 \quad \textit{almost surely.}$$

Remark: Since the variables $R_\rho(\hat{f}_n) - R_{B,\rho}^*$ are uniformly bounded (by a constant, e.g., 1), strong consistency implies consistency (by the Dominated Convergence Theorem).

4.3 Universal Consistency

Definition 3 (Universal Consistency)**.** *We say $\{\hat{f}_n\}$ is **universally consistent** if it is consistent for **any** distribution ρ on $\mathcal{X} \times \mathcal{Y}$.*

$$\lim_{n\to\infty} \mathbb{E}[R_\rho(\hat{f}_n)] - R_{B,\rho}^* = 0 \quad \forall \rho$$

Definition 4 (Universal Strong Consistency)**.** *We say that a family $\{\hat{f}_n\}$ is **universally strongly consistent** if it is strongly consistent for all ρ.*

4.3.1 Statistical Inference Analogy

To understand the difference between consistency and universal consistency, consider a simpler setting: Let μ be a distribution over $\mathbb{R}$ with finite first moment. Let $\theta_\mu := \mathbb{E}_{Z\sim\mu}[Z]$. Goal: Build an estimator $\hat{\theta}_n(Z_1, ..., Z_n)$ that is consistent for θ_μ.

- **Sample Mean:** $\hat{\theta}_n = \frac{1}{n}\sum Z_i$. By SLLN, $\hat{\theta}_n \to \theta_\mu$ almost surely regardless of what μ is (as long as finite mean). This is **universally strongly consistent**.
- **Sample Median:** Is **NOT** universally consistent for the mean. It is consistent only if the distribution μ has a mean and median that coincide (e.g., symmetric distributions). If μ does not satisfy this property, it fails.

4.4 Consistency of Plug-in Classifiers

Recall the plug-in classifier:

$$\hat{f}_n(x) = \begin{cases} 1 & \text{if } \hat{\eta}_n(x) \geq 1/2 \\ 0 & \text{else} \end{cases}$$

where $\hat{\eta}_n(x)$ approximates $\eta(x) = \mathbb{P}(Y = 1|X = x)$. From previous lectures, we know:

$$R(\hat{f}_n) - R_B^* \leq 2\mathbb{E}_X[|\eta(X) - \hat{\eta}_n(X)|]$$

Our goal is to construct $\hat{\eta}_n$ such that:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n}\mathbb{E}_X[|\eta(X) - \hat{\eta}_n(X)|] = 0$$

(This implies consistency). Or almost surely for strong consistency.

4.4.1 Weight-based Estimators

We focus on proximity-based estimators of the form:

$$\hat{\eta}_n(x) = \sum_{i=1}^{n} Y_i w_{in}(x; X_1, ..., X_n)$$

where $\sum_{i=1}^n w_{in}(x) = 1$ and $w_{in}(x) \geq 0$.

Example 1: Kernel-based (Nadaraya-Watson) Select $\gamma_n > 0$.

$$w_{in}(x) = \frac{\exp(-\frac{\|X_i - x\|^2}{\gamma_n^2})}{\sum_{j=1}^n \exp(-\frac{\|X_j - x\|^2}{\gamma_n^2})}$$

Example 2: Histograms $\mathcal{X} = \mathbb{R}^d$. Introduce $h_n > 0$. Partition the space into bins of size h_n. Define pre-weights $\tilde{w}_{in}(x) = 1$ if X_i belongs to the same bin as x, and 0 otherwise. Normalization:

- If there is at least one X_j in the bin:

$$w_{in}(x) = \frac{\tilde{w}_{in}(x)}{\sum_{j=1}^n \tilde{w}_{jn}(x)}$$

- If the bin is empty: $w_{in}(x) = 1/n$.

Example 3: k-Nearest Neighbors (k-NN) $\mathcal{X} = \mathbb{R}^d$. Let $k_n \in \mathbb{N}$ be between 1 and n. Pre-weights: $\tilde{w}_{in}(x) = 1$ if X_i is among the k-closest points to x, else 0. Formally, order the data by distance to x: $\|x - X_{(1)}\| \leq \|x - X_{(2)}\| \leq ....$ The weights are uniform over the k neighbors:

$$w_{in}(x) = \frac{\tilde{w}_{in}(x)}{\sum_{j=1}^{n} \tilde{w}_{jn}(x)} = \begin{cases} 1/k & \text{if neighbor} \\ 0 & \text{else} \end{cases}$$

4.5 Stone's Theorem (1977)

Let $\mathcal{X} = \mathbb{R}^d$. Let $w_{in}(x)$ be a collection of weights satisfying the following three conditions:

1. **Stability:** $\exists C > 0$ such that for every non-negative function $g : \mathcal{X} \to (0, \infty)$ with $\mathbb{E}[g(X)] < \infty$:
$$\mathbb{E}_{X,\mathcal{D}_n}\left[\sum_{i=1}^{n} w_{in}(x) g(X_i)\right] \leq C\mathbb{E}[g(X)]$$

2. **Localization:** For all $a > 0$:
$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} w_{in}(X)\mathbf{1}_{\|X - X_i\| > a}\right] = 0$$
(Weights are only large for points X_i that are close enough to X).

3. **vanishing Weights:**
$$\lim_{n\to\infty} \mathbb{E}\left[\max_{i=1...n} w_{in}(X)\right] = 0$$
(We want to give weights to enough points, not just a few).

Conclusion: If these conditions hold, then:

$$\lim_{n\to\infty} \mathbb{E}[|\eta(X) - \hat{\eta}_n(X)|] = 0$$

Corollary: The induced family of classifiers is universally consistent.

4.5.1 Proof Sketch

Define the "smoothed" true function $\bar{\eta}_n(x) := \sum_{i=1}^{n} \eta(X_i) w_{in}(x)$. We decompose the error:

$$\mathbb{E}[|\eta(X) - \hat{\eta}_n(X)|] \leq \underbrace{\mathbb{E}[|\eta(X) - \bar{\eta}_n(X)|]}_{\text{Bias}} + \underbrace{\mathbb{E}[|\bar{\eta}_n(X) - \hat{\eta}_n(X)|]}_{\text{Variance}}$$

Recall $\hat{\eta}_n(x) = \sum Y_i w_{in}(x)$ and $\mathbb{E}[Y_i|X_i] = \eta(X_i)$.

5 Lecture 5: Universal Consistency & Concentration

5.1 Proof of Bias Convergence in Stone's Theorem

We want to show the Bias term goes to 0:

$$\text{Bias} = \mathbb{E}[|\eta(X) - \bar{\eta}_n(X)|] = \mathbb{E}\left[\left|\sum_{i=1}^{n}(\eta(X) - \eta(X_i))w_{in}(X)\right|\right]$$

Using triangle inequality and the fact that $\sum w_{in} = 1$:

$$\leq \mathbb{E}\left[\sum_{i=1}^{n} |\eta(X) - \eta(X_i)| w_{in}(X)\right]$$

Let $a > 0$. Split the sum based on distance $\|X - X_i\| > a$ and $\leq a$.

$$\leq \mathbb{E}\left[\sum_{i=1}^{n} |\eta(X) - \eta(X_i)| w_{in}(X)\mathbf{1}_{\|X-X_i\|>a}\right] + \mathbb{E}\left[\sum_{i=1}^{n} |\eta(X) - \eta(X_i)| w_{in}(X)\mathbf{1}_{\|X-X_i\|\leq a}\right]$$

Since $|\eta(x) - \eta(x')| \leq 1$:

$$\leq 2\mathbb{E}\left[\sum_{i=1}^{n} w_{in}(X)\mathbf{1}_{\|X-X_i\|>a}\right] + \mathbb{E}\left[\sum_{i=1}^{n} |\eta(X) - \eta(X_i)| w_{in}(X)\mathbf{1}_{\|X-X_i\|\leq a}\right]$$

Case 1: η is Lipschitz. Suppose $|\eta(x) - \eta(x')| \leq \text{Lip}(\eta)\|x - x'\|$. The second term becomes bounded by $a \cdot \text{Lip}(\eta)$. Taking limits:

$$\limsup_{n\to\infty} \text{Bias} \leq 2 \lim_{n\to\infty} \mathbb{E}\left[\sum w_{in}\mathbf{1}_{>a}\right] + a \cdot \text{Lip}(\eta)$$

The first term is 0 by Stone's Condition 2. So $\limsup \leq a \cdot \text{Lip}(\eta)$. Since a is arbitrary, the Bias goes to 0.

Case 2: General η. If η is not Lipschitz, we use density arguments. We can approximate any measurable function with a sequence of Lipschitz functions $\{h_s\}$ in $L^1(\rho_X)$.

$$\lim_{s\to\infty} \mathbb{E}[|h_s(X) - \eta(X)|] = 0$$

We define $\bar{h}_{s,n}(x) = \sum h_s(X_i)w_{in}(x)$. Using triangle inequality:

$$\mathbb{E}|\eta - \bar{\eta}_n| \leq \mathbb{E}|\eta - h_s| + \mathbb{E}|h_s - \bar{h}_{s,n}| + \mathbb{E}|\bar{h}_{s,n} - \bar{\eta}_n|$$

1. $\mathbb{E}|\eta - h_s|$ is small for large s. 2. $\mathbb{E}|h_s - \bar{h}_{s,n}|$ goes to 0 by the Lipschitz argument (Condition 2). 3. $\mathbb{E}|\bar{h}_{s,n} - \bar{\eta}_n|$ involves $\sum |h_s(X_i) - \eta(X_i)|w_{in}(X)$. By Stone's Condition 1 (Stability), this is bounded by $C\mathbb{E}|h_s - \eta|$, which is small.

Summary of Stone's Proof Requirements:

- To prove **Bias** $\to 0$: Need Conditions 1 and 2.
- To prove **Variance** $\to 0$: Need Condition 3 (Max weight $\to 0$). (Homework exercise).

5.2 Applying Stone's Theorem

5.2.1 Universal Consistency of Histograms

Suppose $\mathcal{X} = \mathbb{R}^d$. Weights defined by histograms with bin width h_n. The family is universally consistent if:

1. $h_n \to 0$ as $n \to \infty$ (Ensures localization).
2. $nh_n^d \to \infty$ as $n \to \infty$ (Ensures max weight $\to 0$, roughly $1/(n\text{Vol}) \approx 1/(nh_n^d)$).

5.2.2 Universal Consistency of k-NN

Suppose $\mathcal{X} = \mathbb{R}^d$. Weights defined by k_n-NN. The family is universally consistent if:

1. $k_n/n \to 0$ as $n \to \infty$ (Localization).
2. $k_n \to \infty$ as $n \to \infty$ (Vanishing weights, since $w_{in} \approx 1/k_n$).

5.3 Strong Consistency and Concentration Inequalities

Stone's theorem gives us L^1 convergence (consistency). How do we study **strong consistency**? We need tools to bound probabilities of large deviations.

5.3.1 Probability Reviews

Let $Z_1, ..., Z_n$ be i.i.d. real random variables with $\mathbb{E}[Z_i] = m$.
Strong Law of Large Numbers (SLLN): $\frac{1}{n}\sum(Z_i - m) \to 0$ almost surely.
Central Limit Theorem (CLT): Suppose $\text{Var}(Z_i) = \sigma^2 < \infty$.

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum(Z_i - m)\right) \xrightarrow{d} \mathcal{N}(0,1)$$

For fixed t:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum(Z_i - m)\right) \geq t\right) = \int_t^\infty \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds$$

Berry-Esseen Theorem: A quantitative version of CLT. If $\mathbb{E}|Z_i|^3 < \infty$ (let $\gamma = \mathbb{E}|Z_i|^3$), then:

$$\left|\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\text{Sum} \geq t\right) - \text{GaussianTail}(t)\right| \leq \frac{C\gamma}{\sigma\sqrt{n}}$$

Limitation: When n is large, the deviation probability might be very small, and the error term $O(1/\sqrt{n})$ might dominate the probability itself. This motivates Concentration Inequalities.

5.3.2 Concentration Inequalities

Goal: Quantify precisely the behavior of tails for finite n. If the sum were truly Gaussian $Z \sim \mathcal{N}(0,1)$, we have the bound $\mathbb{P}(Z \geq t) \leq e^{-t^2/2}$. We want similar bounds for sums of random variables.

Markov's Inequality: Let W be a real-valued non-negative random variable.

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}[W]}{t} \quad \forall t > 0$$

Proof: $\mathbb{P}(W \geq t) = \mathbb{E}[\mathbf{1}_{W \geq t}] \leq \mathbb{E}[\frac{W}{t}\mathbf{1}_{W \geq t}] \leq \frac{\mathbb{E}[W]}{t}$.

Generalized Markov: Let $\phi : \mathbb{R} \to (0, \infty)$ be non-decreasing.

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

Chernoff's Bound (Motivation): Choose $\phi(t) = \exp(st)$ for some $s > 0$.

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}[e^{sW}]}{e^{st}} \quad \forall s > 0$$

To get the tightest bound, we minimize over $s > 0$:

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

Next time, we will apply this to the sample mean W_n.

6 Lecture 6: Concentration Inequalities

6.1 Recap: Chernoff Bound

Recall Markov's inequality: $\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}[W]}{t}$ for $W \geq 0$. Applying this to e^{sW} yields Chernoff's bound: For any real-valued random variable W:

$$\mathbb{P}(W \geq t) \leq \inf_{s>0} \frac{\mathbb{E}[\exp(sW)]}{\exp(st)}$$

We will use this to prove Hoeffding's and Bernstein's inequalities.

6.2 Sub-Gaussian Random Variables

Definition 5. *A random variable W with $\mathbb{E}[W] = 0$ is called* ***Sub-Gaussian*** *with variance parameter $\sigma^2 > 0$ if:*

$$\mathbb{E}[\exp(sW)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right) \quad \forall s \in \mathbb{R}$$

This condition means the moment generating function of W is bounded by that of a Gaussian $\mathcal{N}(0, \sigma^2)$.

6.3 Hoeffding's Inequality

Theorem: Let $Z_1, ..., Z_n$ be independent real-valued random variables such that:

1. $\mathbb{E}[Z_i] = 0$
2. Z_i is sub-Gaussian with parameter σ_i^2.

Then for all $t > 0$:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right) \leq \exp\left(-\frac{n^2 t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right)$$

In particular, if $\sigma_i = \sigma$ for all i, this simplifies to:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

6.3.1 Proof of Hoeffding's Inequality

Let $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$. By Chernoff's bound:

$$\begin{aligned}
\mathbb{P}(W \geq t) &\leq \inf_{s>0} \frac{\mathbb{E}[\exp(s \cdot \frac{1}{n}\sum Z_i)]}{\exp(st)} \\
&= \inf_{s>0} \frac{\prod_{i=1}^{n} \mathbb{E}[\exp(\frac{s}{n} Z_i)]}{\exp(st)} \quad \text{(Independence)} \\
&\leq \inf_{s>0} \frac{\prod_{i=1}^{n} \exp(\frac{\sigma_i^2 (s/n)^2}{2})}{\exp(st)} \quad \text{(Sub-Gaussianity)} \\
&= \inf_{s>0} \exp\left(\frac{s^2}{2n^2}\sum_{i=1}^{n}\sigma_i^2 - st\right)
\end{aligned}$$

To minimize the exponent, we optimize over s. Taking the derivative with respect to s and setting to 0:

$$\frac{s}{n^2}\sum \sigma_i^2 - t = 0 \implies s^* = \frac{tn^2}{\sum \sigma_i^2}$$

Substituting s^* back yields the result. $\square$

Remarks:

1. The variables Z_i must be independent, but need not be identically distributed.
2. If $\mathbb{E}[Z_i] \neq 0$, we apply the bound to $Z_i - \mathbb{E}[Z_i]$.
3. Concentration for absolute values:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum(Z_i - \mathbb{E}[Z_i])\right| \geq t\right) \leq 2\exp\left(-\frac{n^2t^2}{2\sum \sigma_i^2}\right)$$

6.4 Hoeffding's Lemma for Bounded Random Variables

Lemma: If Z is a random variable with $\mathbb{E}[Z] = 0$ and $a \leq Z \leq b$, then Z is sub-Gaussian with $\sigma^2 = \frac{(b-a)^2}{4}$.

$$\mathbb{E}[\exp(sZ)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Corollary: Let $Z_1, ..., Z_n$ be independent with $a_i \leq Z_i \leq b_i$. Then:

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

Proof Sketch of Lemma: uses convexity of the exponential function. Since $x \mapsto e^{sx}$ is convex, the chord connecting (a, e^{sa}) and (b, e^{sb}) lies above the function. This allows bounding the moment generating function.

6.5 Bernstein's Inequality

Hoeffding's inequality relies only on the range $(b-a)$. It does not account for the variance being potentially small. **Theorem:** Let $Z_1, ..., Z_n$ be independent with $\mathbb{E}[Z_i] = 0$, $|Z_i| \leq M$, and $\mathrm{Var}(Z_i) = \sigma_i^2$. Then:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}n^2t^2}{\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{3}Mnt}\right)$$

Or equivalently for the sum $S_n = \sum Z_i$:

$$\mathbb{P}(S_n \geq t) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum \sigma_i^2 + \frac{1}{3}Mt}\right)$$

Comparison: When σ^2 is much smaller than the range M^2, Bernstein's inequality is sharper than Hoeffding's. Specifically, if $\sigma^2 \leq \frac{2}{3}M^2$ (approx), Bernstein wins.

7 Lecture 7: Martingales and McDiarmid's Inequality

7.1 Motivation

Hoeffding and Bernstein require sums of **independent** random variables. Recall our plug-in classifier analysis involved $\hat{\eta}_n(x)$, which is a function of independent variables (X_i, Y_i), but the structure is not a simple sum. We need tools for functions $g(Z_1, ..., Z_n)$ where Z_i are independent.

7.2 Martingales

Definition 6 (Martingale Difference Sequence)**.** *Let $\{Z_1, ..., Z_n\}$ and $\{V_1, ..., V_n\}$ be sequences of random variables. We say $\{Z_k\}$ is a **Martingale Difference Sequence (MDS)** w.r.t. $\{V_k\}$ if:*

1. *Z_k is a deterministic function of $V_1, ..., V_k$.*

2. *$\mathbb{E}[Z_k | V_1, ..., V_{k-1}] = 0$ for all k.*

Definition 7 (Martingale)**.** *A sequence $\{M_1, ..., M_n\}$ is a **Martingale** w.r.t. $\{V_k\}$ if:*

1. *M_k is a function of $V_1, ..., V_k$.*

2. *$\mathbb{E}[M_k | V_1, ..., V_{k-1}] = M_{k-1}$.*

Observation: If $\{M_k\}$ is a martingale, then the differences $Z_k = M_k - M_{k-1}$ (with $M_0 = \mathbb{E}[M_1]$) form a martingale difference sequence.

Example (Doob Martingale): Let $Z_1, ..., Z_n$ be independent random variables with $\mathbb{E}[Z_i] = 0$. Let $M_k = \sum_{i=1}^{k} Z_i$. This is a martingale w.r.t $\{Z_i\}$. Proof:

$$\mathbb{E}[M_k | Z_1, ..., Z_{k-1}] = \mathbb{E}[Z_k + M_{k-1} | Z_{1:k-1}] = \mathbb{E}[Z_k] + M_{k-1} = 0 + M_{k-1} = M_{k-1}$$

7.3 Hoeffding-Azuma Inequality

This is Hoeffding's inequality generalized to Martingale Difference Sequences.

Theorem: Let $Z_1, ..., Z_n$ be a martingale difference sequence w.r.t some $\{V_i\}$. Suppose that $U_k \le Z_k \le U_k + C_k$, where C_k is a fixed constant and U_k is a function of $V_1, ..., V_{k-1}$. Then:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \ge t\right) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n} C_k^2}\right)$$

This looks just like Hoeffding's, but allows for dependence structure via the martingale property!

7.4 McDiarmid's Inequality (Bounded Differences Inequality)

We want a concentration bound for a general function $g(Z_1, ..., Z_n)$ of independent random variables $Z_i \in \mathcal{S}_i$.

Bounded Difference Condition: Assume that for all $i = 1, ..., n$, changing the i-th coordinate changes the function value by at most c_i:

$$\sup_{z_1,...,z_n,z_i'} |g(z_1, ..., z_i, ..., z_n) - g(z_1, ..., z_i', ..., z_n)| \le c_i$$

Theorem (McDiarmid): Under the bounded difference condition:

$$\mathbb{P}(g(Z_1, ..., Z_n) - \mathbb{E}[g] \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

7.4.1 Relation to Hoeffding

Let $g(Z_1, ..., Z_n) = \frac{1}{n}\sum Z_i$ where $Z_i \in [a_i, b_i]$. Changing Z_i to Z_i' changes the sum by at most $|Z_i - Z_i'|/n \le (b_i - a_i)/n$. So $c_i = (b_i - a_i)/n$. McDiarmid gives:

$$\exp\left(-\frac{2t^2}{\sum((b_i - a_i)/n)^2}\right) = \exp\left(-\frac{2n^2t^2}{\sum(b_i - a_i)^2}\right)$$

This recovers Hoeffding's inequality exactly.

7.4.2 Sketch of Proof

We construct a Doob Martingale: Let $M_k = \mathbb{E}[g(Z_1, ..., Z_n)|Z_1, ..., Z_k]$.

- $M_0 = \mathbb{E}[g]$
- $M_n = g(Z_1, ..., Z_n)$

Define differences $D_k = M_k - M_{k-1}$. Then $g - \mathbb{E}[g] = \sum_{k=1}^n D_k$. $\{D_k\}$ is a martingale difference sequence. One can show that the bounded difference condition on g implies boundedness of the martingale differences:

$$\sup |D_k| \le c_k$$

(Technically, the range is bounded). Applying the Hoeffding-Azuma inequality to this sequence yields the result.

8 Lecture 8: Concentration Applications and Strong Consistency

8.1 Examples of Concentration Inequalities

Example 1: Let $N \sim \text{Bin}(n, p)$. We are interested in bounding $\mathbb{P}(N - \mathbb{E}[N] \ge t)$. Note that $N = \sum_{i=1}^n Z_i$ where $Z_i \sim \text{Ber}(p)$, and the Z_i are independent. In particular, $Z_i \in \{0, 1\} \subseteq [0, 1]$, which means they are bounded random variables. By applying Hoeffding's inequality:

$$\begin{aligned}
\mathbb{P}(N - \mathbb{E}[N] \ge t) &= \mathbb{P}\left(\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \ge t\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \ge \frac{t}{n}\right) \\
&\le \exp\left(\frac{-2n^2(t/n)^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) = \exp\left(-\frac{2t^2}{n}\right)
\end{aligned}$$

Example 2: Let $\xi_1, \dots, \xi_n \sim$ Cauchy(γ). The PDF of Cauchy is given by $p(x) = \frac{1}{\pi\gamma(1+(x/\gamma)^2)}$. Whereas the Gaussian tail is approximately $e^{-t^2/2}$, the Cauchy distribution is approximately $\frac{\gamma}{t^2}$. The Cauchy distribution has very heavy tails. Therefore, they are NOT sub-Gaussian random variables. Not even the expectation $\mathbb{E}[X]$ is well defined. Because of this, we CANNOT use Hoeffding's or Bernstein's inequality in this case.

Example 3: Consider the function $g(u_1, \dots, u_n) = \cos\left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)$. Let $\xi_1, \dots, \xi_n$ be i.i.d. Cauchy distributed random variables. **Question:** Can we use McDiarmid's inequality to bound $\mathbb{P}(g(\xi_1, \dots, \xi_n) - \mathbb{E}[g(\xi_1, \dots, \xi_n)] \geq t)$? **YES!**: The only assumption on the $\xi_1, \dots, \xi_n$ is that they are independent. Now, does g satisfy the bounded increment property? In this case,

$$|g(u_1, \dots, u_i, \dots, u_n) - g(u_1, \dots, u_i', \dots, u_n)| \leq |g(u_1, \dots, u_i, \dots, u_n)| + |g(u_1, \dots, u_i', \dots, u_n)| \leq 2$$

This is always true, but not always useful. A direct application of McDiarmid's leads to:

$$\mathbb{P}(g(\xi_1, \dots, \xi_n) - \mathbb{E}[g(\xi_1, \dots, \xi_n)] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(\frac{-2t^2}{4n}\right) = \exp\left(\frac{-t^2}{2n}\right)$$

Example 4: Suppose $\xi_1, \dots, \xi_n$ are i.i.d. random variables with $\xi_i \in [a_i, b_i]$. **Question:** For these random variables, do we have a better bound for $\mathbb{P}(g(\xi_1, \dots, \xi_n) - \mathbb{E}[g(\xi_1, \dots, \xi_n)] \geq t)$? **YES!**, because we can use a better bound:

$$|g(u_1, \dots, u_i, \dots, u_n) - g(u_1, \dots, u_i', \dots, u_n)| = \left|\cos\left(\frac{1}{n}\sum_{j\neq i} u_j + \frac{u_i}{n}\right) - \cos\left(\frac{1}{n}\sum_{j\neq i} u_j + \frac{u_i'}{n}\right)\right|$$

By using the Fundamental Theorem of Calculus that is relevant: $\cos(x) = \cos(y) + \int_y^x (-\sin(r))dr$. But the domain of g for this problem is $[a_1, b_1] \times \cdots \times [a_n, b_n]$ (because ξ_i is assumed to fall within $[a_i, b_i]$). In this domain, $\frac{|u_i - u_i'|}{n} \leq \frac{b_i - a_i}{n} =: c_i$. Hence:

$$\mathbb{P}(g(\xi_1, \dots, \xi_n) - \mathbb{E}[g(\xi_1, \dots, \xi_n)] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(\frac{-2t^2}{\frac{1}{n^2}\sum_{i=1}^n (b_i - a_i)^2}\right)$$

8.2 High Probability Statements

Statement 1: $\mathbb{P}(|W_n - \mathbb{E}[W_n]| \geq t) \leq 2\exp(-cnt^2)$ For example in Hoeffding's inequality.

Statement 2: With probability greater than or equal to $1-\delta$ we have: $|W_n - \mathbb{E}[W_n]| \leq f(\delta, n)$.

What are we fixing in each statement?

- Statement 1: threshold t.
- Statement 2: probability δ.

In the above example: $\delta = 2\exp(-cnt^2)$.

$$\log\left(\frac{\delta}{2}\right) = -cnt^2 \implies \sqrt{\frac{1}{cn}\log\left(\frac{2}{\delta}\right)} = t$$

This implies:

$$\mathbb{P}\left(|W_n - \mathbb{E}[W_n]| \geq \sqrt{\frac{1}{cn}\log\left(\frac{2}{\delta}\right)}\right) \leq \delta$$

$$\mathbb{P}\left(|W_n - \mathbb{E}[W_n]| < \sqrt{\frac{1}{cn}\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta$$

8.3 Borel-Cantelli Lemma and Almost Sure Convergence

To apply concentration inequalities in the study of strong consistency of plug-in classifiers, we need the following theorem:

Theorem (Borel-Cantelli): Let $E_1, E_2, E_3, \dots$ be events in a probability space. Suppose that: $\sum_{i=1}^{\infty} \mathbb{P}(E_i) < \infty$. (This is a stronger condition than $\lim \mathbb{P}(E_i) = 0$). Then the event $E = \{\exists N \text{ s.t. } \forall n \geq N, E_n \text{ DOES NOT OCCUR}\}$ is such that $\mathbb{P}(E) = 1$. In words, this means that almost surely E_n does not occur for all large enough n.

Example 1: Suppose $Z_1, Z_2, \dots$ are Bernoulli(p_i) such that $\sum_{i=1}^{\infty} p_i < \infty$, where $p_i = \mathbb{P}(Z_i = 1)$. Let $E = \{\text{eventually, we don't see heads any more}\}$. By Borel-Cantelli $\implies \mathbb{P}(E) = 1$.

Example 2: $Z_1, Z_2, \dots$ are i.i.d. random variables with $|Z_i| \leq M$ for all i, and $\mathbb{E}[Z_i] = \mu$. By Hoeffding's inequality, $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2M^2}\right)$. **Goal:** Prove that $\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| = 0$ almost surely. (In other words, proving the SLLN in this setting).

Let $E_n := \left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| > t_n\right\}$, where t_n is to be chosen. By Hoeffding's inequality: $\mathbb{P}(E_n) \leq 2\exp\left(\frac{-nt_n^2}{2M^2}\right)$. We need the sum to be finite: $\sum_{n=1}^{\infty} \mathbb{P}(E_n) \leq \sum_{n=1}^{\infty} 2\exp\left(\frac{-nt_n^2}{2M^2}\right) < \infty$. We need to choose t_n such that by Borel-Cantelli, with probability one, we will be able to find N such that E_n does not occur. In other words, $\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \leq t_n$ for all large enough $n \geq N$. If in addition t_n was chosen so that $\lim_{n\to\infty} t_n = 0$, then $\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| = 0$.

We need t_n such that:

1. $\sum_{n=1}^{\infty} \exp\left(-\frac{nt_n^2}{2M^2}\right) < \infty$
2. $\lim_{n\to\infty} t_n = 0$

Let $t_n := \sqrt{4M^2\frac{\log(n)}{n}}$. Then for condition 1:

$$\sum_{n=1}^{\infty} \exp\left(-\frac{n}{2M^2}4M^2\frac{\log(n)}{n}\right) = \sum_{n=1}^{\infty} \exp(-2\log(n)) = \sum_{n=1}^{\infty} \exp\left(\log\left(\frac{1}{n^2}\right)\right) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

Note: We have just proved the SLLN for bounded random variables. Recall that the SLLN is much more general.

8.4 Strong Consistency of Histogram Classifiers

Finally returning to strong consistency of similarity classifiers.

Theorem: Let $\hat{f}_n$ be a histogram classifier built from data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$. Let h_n be the length scale of the histogram. If:

1. $h_n \to 0$ as $n \to \infty$
2. $nh_n^d \to \infty$ as $n \to \infty$

Then $\{\hat{f}_n\}$ is universally strongly consistent.

Proof: Step 1: The empirical classifier is $\hat{f}_n(x) = 1$ if $\hat{\eta}_n(x) \geq \frac{1}{2}$, else 0. Here $\hat{\eta}_n(x) = \sum_{i=1}^{n} Y_i w_{in}(x)$. Let $\tilde{w}_{in}(x) = 1$ if X_i belongs to the same cell as x, else 0. The weights are normalized: $w_{in}(x) = \frac{\tilde{w}_{in}(x)}{\sum_{j=1}^{n} \tilde{w}_{jn}(x)}$ if the denominator $\neq 0$, else $\frac{1}{n}$.

Notation: For a given x, let $A_n(x)$ be the cell at level n that x belongs to, which has side length h_n. Let $N_n(x)$ be the number of X_j such that $X_j \in A_n(x)$. Then we can write $w_{in}(x) = \frac{1}{N_n(x)}\mathbf{1}_{A_n(x)}(X_i)$ if $N_n(x) > 0$, else $\frac{1}{n}$.

Therefore, the condition $\hat{\eta}_n(x) \geq \frac{1}{2}$ is equivalent to $\frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{N_n(x)} \geq \frac{1}{2}$.

$$\begin{aligned}
2\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i) &\geq \sum_{i=1}^n \mathbf{1}_{A_n(x)}(X_i) \\
&= \sum_{i=1}^n (Y_i + (1 - Y_i))\mathbf{1}_{A_n(x)}(X_i) \\
\iff \sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i) &\geq \sum_{i=1}^n (1 - Y_i)\mathbf{1}_{A_n(x)}(X_i)
\end{aligned}$$

By dividing both sides by $n\rho_X(A_n(x))$, we can define empirical quantities:

- $\hat{\alpha}_n^1(x) := \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))} \approx \eta(x)$
- $\hat{\alpha}_n^0(x) := \frac{\sum_{i=1}^n (1-Y_i) \mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))} \approx 1 - \eta(x)$

Then the decision rule becomes: $\hat{f}_n(x) = 1$ if $\hat{\alpha}_n^1(x) \geq \hat{\alpha}_n^0(x)$, else 0.

Step 2: Before, we established $R_\rho(\hat{f}_n) - R^*_{B,\rho} \leq 2\mathbb{E}_{X\sim\rho_X}[|\hat{\eta}_n(X) - \eta(X)|]$. We need to extend this a bit:

$$R_\rho(\hat{f}_n) - R^*_{B,\rho} \leq \mathbb{E}_{X\sim\rho_X}[|\eta(X) - \hat{\alpha}_n^1(X)|] + \mathbb{E}_{X\sim\rho_X}[|(1 - \eta(X)) - \hat{\alpha}_n^0(X)|]$$

(See HW2 for derivation details). If we can show that the RHS goes to zero as $n \to \infty$ (almost surely), then the LHS will also go to zero as $n \to \infty$ (almost surely). This implies Strong Consistency.

We focus on the first term: $W_n := \mathbb{E}_{X\sim\rho_X}[|\eta(X) - \hat{\alpha}_n^1(X)|]$. We want to show $\lim_{n\to\infty} W_n = 0$ almost surely. Note that W_n is a function of the data sequence:

$$W_n = g((X_1, Y_1), \dots, (X_n, Y_n)) = \mathbb{E}_{X\sim\rho_X}\left[\left|\eta(X) - \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(X)}(X_i)}{n\rho_X(A_n(X))}\right|\right]$$

To show that $W_n \to 0$ as $n \to \infty$ almost surely, we will split it:

$$W_n = (W_n - \mathbb{E}[W_n]) + \mathbb{E}[W_n]$$

And we will show both parts converge to 0 as $n \to \infty$ almost surely.

9 Lecture 9: Universal Strong Consistency and Empirical Risk Minimization

9.1 Universal Strong Consistency of Histogram Classifiers

Theorem: (Universal Strong Consistency of histogram classifiers).Let $\mathcal{X} = \mathbb{R}^d$ and suppose $\hat{f}_n$ is the plug-in classifier built with histogram-based weights with length-scale h_n.If:

1. $h_n \to 0$ as $n \to \infty$
2. $nh_n^d \to \infty$ as $n \to \infty$

Then $\{\hat{f}_n\}_{n \in \mathbb{N}}$ is universally strongly consistent.

Proof:

Step 1: To prove this, it is easier to introduce some notation and rewrite some expressions.

Notation: For a given $x \in \mathbb{R}^d$:

- $A_n(x) :=$ cell at level n that x belongs to (with side length h_n).
- $\mu_n(x) :=$ number of X_j such that $X_j \in A_n(x)$ (i.e., the number of training points in the same cell as x).

The weights are defined as:

$$w_{in}(x) = \begin{cases} \frac{\mathbf{1}_{A_n(x)}(X_i)}{\mu_n(x)} & \text{if } \mu_n(x) > 0 \\ \frac{1}{n} & \text{else} \end{cases}$$

For simplicity, imagine we have the plug-in classifier:

$$\hat{f}_n(x) = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{\mu_n(x)} \geq \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

Let us rewrite the condition $\frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{\mu_n(x)} \geq \frac{1}{2}$:

$$\begin{aligned} \iff\ & 2\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i) \geq \sum i = 1^n \mathbf{1}_{A_n(x)}(X_i) \\ \iff\ & 2\sum i = 1^n Y_i \mathbf{1}_{A_n(x)}(X_i) \geq \sum i = 1^n (Y_i + (1 - Y_i))\mathbf{1}_{A_n(x)}(X_i) \\ \iff\ & \sum i = 1^n Y_i \mathbf{1}_{A_n(x)}(X_i) \geq \sum i = 1^n (1 - Y_i)\mathbf{1}_{A_n(x)}(X_i) \end{aligned}$$

Dividing both sides by $n\rho_X(A_n(x))$ gives:

$$\underbrace{\frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))}}_{=: \hat{\alpha}_n^1(x) \approx \eta(x)} \geq \underbrace{\frac{\sum_{i=1}^n (1 - Y_i)\mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))}}_{=: \hat{\alpha}_n^0(x) \approx 1 - \eta(x)}$$

Then the decision rule becomes:

$$\hat{f}_n(x) = \begin{cases} 1 & \text{if } \hat{\alpha}_n^1(x) \geq \hat{\alpha}_n^0(x) \\ 0 & \text{else} \end{cases}$$

Step 2: Extend the risk bound.Before, we established:

$$R_\rho(\hat{f}_n) - R_{B,\rho}^* \leq 2\mathbb{E}_{X \sim \rho_X}[|\hat{\eta}_n(X) - \eta(X)|]$$

We need to extend this a bit (see HW2 for details). If $\hat{\alpha}_n^1$ and $\hat{\alpha}_n^0$ do not have to add to one, the extended inequality is:

$$R_\rho(\hat{f}_n) - R_{B,\rho}^* \leq \mathbb{E}_{X \sim \rho_X}[|\eta(X) - \hat{\alpha}_n^1(X)|] + \mathbb{E}_{X \sim \rho_X}[|(1 - \eta(X)) - \hat{\alpha}_n^0(X)|]$$

If we can show that the RHS goes to zero as $n \to \infty$ (almost surely), then the LHS will also go to zero as $n \to \infty$ (almost surely).Let us consider the terms:

$$\hat{\alpha}_n^1(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))}$$
$$\hat{\alpha}_n^0(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \frac{\sum_{i=1}^n (1 - Y_i) \mathbf{1}_{A_n(x)}(X_i)}{n\rho_X(A_n(x))}$$

(Note that the number of points in $A_n(x)$ is exactly $\mu_n(x)$).

Step 3: Prove convergence.Goal: Prove that, almost surely,

$$\lim_{n\to\infty} \underbrace{\mathbb{E}_{X\sim\rho_X}\left[\left|\eta(X) - \frac{\sum_{i=1}^n Y_i \mathbf{1}_{A_n(X)}(X_i)}{n\rho_X(A_n(X))}\right|\right]}_{=:\, g_n((X_1,Y_1),\dots,(X_n,Y_n))} = 0$$

We decompose g_n into a variance-like term and a bias-like term:

$$\begin{aligned} g_n((X_1, Y_1), \dots, (X_n, Y_n)) = (g_n((X_1, Y_1), \dots, (X_n, Y_n)) - \mathbb{E}_{\text{Data}}[g_n]) \quad &(\text{Variance}) \\ + \mathbb{E}_{\text{Data}}[g_n] \quad (\text{Bias}) \end{aligned}$$

Bias term:

$$\text{Bias} = \mathbb{E}_{\text{Data}}\mathbb{E}_{X\sim\rho_X}[|\eta(X) - \hat{\alpha}_n^1(X)|]$$

By a small adaptation of what was done for HW2 (essentially adapting Stone's theorem), since we assume $h_n \to 0$ and $nh_n^d \to \infty$, then:

$$\lim_{n\to\infty} \mathbb{E}_{\text{Data},X}[|\eta(X) - \hat{\alpha}_n^1(X)|] = 0$$

Variance term: We use McDiarmid's inequality.

$$g_n((X_1, Y_1), \dots, (X_n, Y_n)) = \mathbb{E}_{X\sim\rho_X}\left[\left|\eta(X) - \frac{\sum_{j=1}^n Y_j \mathbf{1}_{A_n(X)}(X_j)}{n\rho_X(A_n(X))}\right|\right]$$

Change the i-th variable $(X_i, Y_i) \to (\tilde{X}_i, \tilde{Y}_i)$. We estimate the difference:

$$\begin{aligned} &\left|g_n((X_1, Y_1), \dots, (X_i, Y_i), \dots) - g_n((X_1, Y_1), \dots, (\tilde{X}i, \tilde{Y}i), \dots)\right| \\ &\leq \mathbb{E}X \sim \rho_X \left[\left|\frac{Y_i \mathbf{1}_{A_n(X)}(X_i)}{n\rho_X(A_n(X))} - \frac{\tilde{Y}i \mathbf{1}_{A_n(X)}(\tilde{X}i)}{n\rho_X(A_n(X))}\right|\right] \\ &\leq \mathbb{E}X \sim \rho_X \left[\frac{Y_i \mathbf{1}_{A_n(X)}(X_i)}{n\rho_X(A_n(X))}\right] + \mathbb{E}X \sim \rho_X \left[\frac{\tilde{Y}i \mathbf{1}_{A_n(X)}(\tilde{X}_i)}{n\rho_X(A_n(X))}\right] \quad (\text{Triangle Inequality}) \end{aligned}$$

To understand the bound on the expectation:

$$\mathbb{E}_{X\sim\rho_X}\left[\frac{Y_i \mathbf{1}_{A_n(X)}(X_i)}{n\rho_X(A_n(X))}\right]$$

Note that $X_i \in A_n(X) \implies A_n(X) = A_n(X_i)$. Therefore:

$$= \mathbb{E}_{X\sim\rho_X}\left[\frac{Y_i \mathbf{1}_{A_n(X_i)}(X)}{n\rho_X(A_n(X_i))}\right] = \frac{Y_i \rho_X(A_n(X_i))}{n\rho_X(A_n(X_i))} = \frac{Y_i}{n} \leq \frac{1}{n}$$

So the difference is bounded by $\frac{1}{n} + \frac{1}{n} = \frac{2}{n} =: c_i$.By McDiarmid's inequality:

$$\mathbb{P}\left(|g_n - \mathbb{E}[g_n]| \geq t_n\right) \leq 2\exp\left(-\frac{2t_n^2}{\sum_{i=1}^n c_i^2}\right)$$
$$= 2\exp\left(-\frac{2t_n^2}{\sum_{i=1}^n \frac{4}{n^2}}\right) = 2\exp\left(-\frac{nt_n^2}{2}\right)$$

We need:$t_n \to 0$, $\sum_{n=1}^{\infty} \exp\left(-\frac{nt_n^2}{2}\right) < \infty$.Let $t_n = \sqrt{\frac{C\log(n)}{n}}$ for $C > 2$.

By Borel-Cantelli, we get:

$$\lim_{n\to\infty} |g_n - \mathbb{E}[g_n]| = 0 \quad \text{almost surely.}$$

9.2 Empirical Risk Minimization

The Bayes classifier is defined as:

$$f_B(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

We know that $f_B \in \operatorname{argmin}_f R_\rho(f)$. Until now, we tried to build Similarity Classifiers to approximate f_B. Another approach is **Empirical Risk Minimization**.**Definition:** Given training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we define the **empirical risk** R_n:

$$R_n(f) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{f(X_i)\neq Y_i}$$

where f is a classifier $f : \mathcal{X} \to \mathcal{Y}$.Compare this to the true risk (e.g., $\mathcal{Y} = \{0, 1\}$):

$$R_\rho(f) = \mathbb{E}_{(X,Y)\sim\rho}[\mathbf{1}_{f(X)\neq Y}]$$

If $f_B \in \operatorname{argmin}_f R_\rho(f)$, how about we consider $f_n^* \in \operatorname{argmin}_f R_n(f)$?**Not a good idea directly.****Example:** Assume ρ_X has a density.Choose $\delta > 0$ such that for a fixed $\epsilon > 0$, we have:

$$\rho_X\left(\bigcup_{i=1}^n B(X_i, \delta)\right) \leq \epsilon$$

Define the classifier:

$$f_n(x) = \begin{cases} Y_i & \text{if } x \in B(X_i, \delta) \\ 1 - f_B(x) & \text{if } x \notin \bigcup_{i=1}^n B(X_i, \delta) \end{cases}$$

By construction: $R_n(f_n) = 0$. (It perfectly memorizes the training data).On the other hand, the true risk is terrible: $R_\rho(f_n) \geq \frac{1}{2} - \epsilon$.

How to adjust this idea?Idea: Only work with a **subfamily** of classifiers. Let $\mathcal{F}$ be this family.

Example: $\mathcal{X} = \mathbb{R}^d$; $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{F} = \{\text{all linear classifiers}\}$.For a hyperplane $H_{a,b} = \{x \text{ s.t. } \langle a, x\rangle + b = 0\}$, the positive half-space is $H_{a,b}^+ = \{x \text{ s.t. } \langle a, x\rangle + b \geq 0\}$.

Empirical Risk Minimization (ERM):

$$\min_{f\in\mathcal{F}} R_n(f)$$

We define the following minimizers:

- $f_n^* \in \text{argmin}_{f \in \mathcal{F}} R_n(f)$ (The Empirical Risk Minimizer)
- $f_{\mathcal{F}}^* \in \text{argmin}_{f \in \mathcal{F}} R_\rho(f)$ (The best classifier within the class $\mathcal{F}$)
- $f_B \in \text{argmin}_{\text{all classifiers}} R_\rho(f)$ (The overall optimal Bayes classifier)

10 Lecture 10: Empirical Risk Minimization and Uniform Convergence

10.1 The Pitfall of Direct Empirical Risk Minimization

We know that the Bayes classifier f_B minimizes the true risk:

$$f_B \in \arg\min_f R_\rho(f)$$

How about we consider minimizing the empirical risk over all possible functions?

$$\arg\min_f R_n(f)$$

This is not a good idea directly! Example (Overfitting): Suppose the distribution ρ has a density. Consider the training data $X_1, X_2, ..., X_n \in \mathcal{X}$. Choose a radius $\delta > 0$ small enough such that:

$$\mathbb{P}_X\left(\bigcup_{i=1}^n B(X_i, \delta)\right) \leq \epsilon$$

Construct a classifier $f_n(x)$ as follows:

$$f_n(x) = \begin{cases} Y_i & \text{if } x \in B(X_i, \delta) \\ 1 - f_B(x) & \text{if } x \notin \bigcup_{i=1}^n B(X_i, \delta) \end{cases}$$

By construction, the empirical risk is exactly zero:

$$R_n(f_n) = 0$$

On the other hand, the true risk is extremely high (since it completely disagrees with the optimal Bayes classifier outside the small balls around the training points):

$$R_\rho(f_n) \geq \frac{1}{2} - \epsilon$$

Minimizing R_n over *all* functions does not imply we will "generalize well". It simply leads to overfitting the data.

10.2 Hypothesis Classes

Idea: Only work with a subfamily of classifiers. Let $\mathcal{F}$ be this family.

Example: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{F} = \{\text{all linear classifiers}\}$. For a given vector a and scalar b:

$$\begin{aligned} \mathcal{H}^+_{a,b} &= x \text{ s.t. } \langle a, x\rangle + b \geq 0 \\ \mathcal{H}_{a,b} &= x \text{ s.t. } \langle a, x\rangle + b = 0 \\ \mathcal{H}^-_{a,b} &= x \text{ s.t. } \langle a, x\rangle + b \leq 0 \end{aligned}$$

The classifier is defined as:

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathcal{H}^+_{a,b} \\ 0 & \text{else} \end{cases}$$

10.3 Empirical Risk Minimization (ERM)

We define three classifiers of interest:

- **Empirical Risk Minimizer over $\mathcal{F}$:** $f_n^* \in \arg\min_{f \in \mathcal{F}} R_n(f)$, where $R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq Y_i}$.
- **True Risk Minimizer over $\mathcal{F}$:** $f^* \in \arg\min_{f \in \mathcal{F}} R_\rho(f)$.
- **Bayes Classifier (over all functions):** $f_B \in \arg\min_f R_\rho(f)$.

Goal: Understand how large the gap $R_\rho(f_n^*) - R_{B,\rho}^*$ can be.
Compare to Similarity Classifiers: For plug-in classifiers, we studied consistency and strong consistency via Stone's Theorem and McDiarmid's inequality, analyzing terms like $\mathbb{E}_X[|\eta(X) - \hat{\eta}_n(X)|]$. How do we study consistency and strong consistency for classifiers obtained through ERM?

10.4 Generalization Bounds and Uniform Convergence

We want to establish statements of the following forms:

1. **Empirical Risk vs True Risk:** With probability at least $1-\delta$:
$$R_\rho(f_n^*) \leq R_n(f_n^*) + \epsilon(n, \mathcal{F}, \delta)$$
2. **True Risk Comparison:** With probability greater than $1-\delta$:
$$R_\rho(f_n^*) \leq R_\rho(f^*) + \epsilon(n, \mathcal{F}, \delta)$$
3. **Overall Risk Bound:** With probability at least $1-\delta$:
$$R_\rho(f_n^*) \leq R_{B,\rho}^* + \epsilon(n, \mathcal{F}, \delta) + \text{Bias}(\mathcal{F})$$

To achieve all this, we need the following **key statement**: With probability greater than $1-\delta$:

$$\sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \leq \epsilon(n, \mathcal{F}, \delta)$$

10.4.1 Deriving the Statements

Observations:

- Since $f_n^* \in \mathcal{F}$ and f^* minimizes R_ρ over $\mathcal{F}$: $R_\rho(f^*) \leq R_\rho(f_n^*)$.
- Since f^* is in $\mathcal{F}$ and f_n^* minimizes R_n over $\mathcal{F}$: $R_n(f_n^*) \leq R_n(f^*)$.

Statement 1:

$$|R_n(f_n^*) - R_\rho(f_n^*)| \leq \sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \leq \epsilon(n, \mathcal{F}, \delta)$$

Statement 2: We decompose the excess risk over the class:

$$\begin{aligned}
0 \leq R_\rho(f_n^*) - R_\rho(f^*) &= R_\rho(f_n^*) - R_n(f_n^*) + R_n(f_n^*) - R_n(f^*) + R_n(f^*) - R_\rho(f^*) \\
&\leq (R_\rho(f_n^*) - R_n(f_n^*)) + 0 + (R_n(f^*) - R_\rho(f^*)) \\
&\leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \\
&\leq \tilde{\epsilon}(n, \mathcal{F}, \delta) \quad (\text{where } \tilde{\epsilon} = 2\epsilon)
\end{aligned}$$

Statement 3:

$$\begin{aligned}
0 &\leq R_\rho(f_n^*) - R_{B,\rho}^* \\
&= \underbrace{R_\rho(f_n^*) - R_\rho(f^*)}_{\text{Random, bounded by } \epsilon(n, \mathcal{F}, \delta)} + \underbrace{R * \rho(f^*) - R_{B,\rho}^*}_{\text{Bias}(\mathcal{F})} \\
&\leq \epsilon(n, \mathcal{F}, \delta) + \text{Bias}(\mathcal{F})
\end{aligned}$$

The Bias term depends on the choice of the family $\mathcal{F}$ (how well $\mathcal{F}$ can approximate the Bayes classifier), not on the training data.

10.5 Vapnik-Chervonenkis (VC) Theorem

Theorem (Vapnik-Chervonenkis):

$$\sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \to 0 \text{ in probability}$$

if and only if:

$$|R_\rho(f_n^*) - R_\rho(f^*)| \to 0 \text{ in probability}$$

(References: Von Luxburg, Schölkopf; Lugosi).

10.6 Bounding the Supremum

How do we obtain concentration bounds for $\mathbb{P}(\sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \geq t)$?

Case 1: $\mathcal{F} = \{f\}$ (A single classifier) Using Hoeffding's inequality for bounded variables (since the 0-1 loss is bounded by 1):

$$\begin{aligned}
\mathbb{P}(|R_n(f) - R_\rho(f)| \geq t) &= \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \left(\mathbf{1}_{f(X_i) \neq Y_i} - \mathbb{E}[\mathbf{1}_{f(X) \neq Y}] \right) \right| \geq t \right) \\
&\leq 2 \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n 1^2} \right) = 2 \exp(-2nt^2)
\end{aligned}$$

Case 2: $\mathcal{F} = \{f^1, ..., f^M\}$ (Finite collection) Using the Union Bound:

$$\begin{aligned}
\mathbb{P}\left(\sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)| \geq t \right) &= \mathbb{P}\left(\max_{j=1...M} |R_n(f^j) - R_\rho(f^j)| \geq t \right) \\
&\leq \sum_{j=1}^M \mathbb{P}(|R_n(f^j) - R_\rho(f^j)| \geq t) \\
&\leq \sum_{j=1}^M 2 \exp(-2nt^2) = 2M \exp(-2nt^2)
\end{aligned}$$

To find the required t for a given confidence level δ:

$$2M\exp(-2nt^2) = \delta \implies 2nt^2 = \log\left(\frac{2M}{\delta}\right) \implies t = \sqrt{\frac{1}{2n}\log\left(\frac{2M}{\delta}\right)}$$

Thus, if $|\mathcal{F}| = M$, with probability greater than $1-\delta$:

$$\sup_{f\in\mathcal{F}} |R_n(f) - R_\rho(f)| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2M}{\delta}\right)} := \epsilon(n, \mathcal{F}, \delta)$$

Case 3: What do we do when $\mathcal{F}$ is NOT finite? Key Insight: It could happen that two different functions $f, \tilde{f} \in \mathcal{F}$ produce the exact same values on the training data $X_1, ..., X_n$. If $f(X_i) = \tilde{f}(X_i)$ for all $i = 1...n$, then $R_n(f) = R_n(\tilde{f})$. We effectively only care about the number of distinct classifications the class $\mathcal{F}$ can realize on the dataset.

10.7 The Shattering Number (Growth Function)

Definition 8. *Let $\mathcal{F}$ be a set of binary classifiers. The set of behaviors of $\mathcal{F}$ on a specific sample $X_1, ..., X_n$ is:*

$$\mathcal{F}_{X_1,\dots,X_n} = \{(f(X_1), ..., f(X_n)) : f \in \mathcal{F}\}$$

*We define the **Shattering Number** of $\mathcal{F}$ as:*

$$S(n, \mathcal{F}) := \sup_{X_1,\dots,X_n} |\mathcal{F}_{X_1,\dots,X_n}|$$

(The supremum is taken over all possible sets of n points in $\mathcal{X}$).

Notes:

- The maximum possible value is $S(n, \mathcal{F}) \leq 2^n$ (every point can be labeled 0 or 1).
- Later we will show that for "nice" families $\mathcal{F}$, the shattering number $S(n, \mathcal{F})$ scales *polynomially* in n, rather than exponentially.

Example: Linear classifiers in $\mathbb{R}^2$ Let $\mathcal{F} = \{\text{all linear classifiers in } \mathbb{R}^2\}$.

- $n = 1$: A single point can be classified as 0 or 1. $S(1, \mathcal{F}) = 2 = 2^1$.
- $n = 2$: Two points can be separated in all 4 possible ways. $S(2, \mathcal{F}) = 4 = 2^2$.
- $n = 3$: For three points (not on a single line), a linear classifier can achieve all 8 possible labelings. $S(3, \mathcal{F}) = 8 = 2^3$.
- $n = 4$: No matter how you arrange 4 points in $\mathbb{R}^2$, you can never achieve all 16 possible labelings (e.g., if arranged in a convex quadrilateral, you cannot separate the diagonals with a single line). Thus, $S(4, \mathcal{F}) < 16 = 2^4$.

11 Lecture 11: Shattering, VC Dimension, and Rademacher Complexity

11.1 Recap: The ERM Problem

We are interested in studying Empirical Risk Minimization (ERM):

$$f_n^* \in \arg\min_{f \in \mathcal{F}} R_n(f) \quad \text{where} \quad R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq Y_i}$$

We want to bound:

- $R_\rho(f_n^*) - R_n(f_n^*) \leq ?$
- $R_\rho(f_n^*) - R_\rho(f^*) \leq ?$

Recall that $f^* \in \arg\min_{f \in \mathcal{F}} R_\rho(f)$. We saw that it suffices to obtain high-probability bounds for the uniform deviation:

$$\sup_{f \in \mathcal{F}} |R_n(f) - R_\rho(f)|$$

11.2 Shattering Number (Growth Function)

We introduced the shattering number to obtain concentration bounds for infinite families $\mathcal{F}$.

Definition 9. *Given a family of classifiers $\mathcal{F}$, the shattering number is:*

$$S(n, \mathcal{F}) := \sup_{X_1,\ldots,X_n \text{ distinct points}} |\mathcal{F}_{X_1,\ldots,X_n}|$$

where $\mathcal{F}_{X_1,\ldots,X_n} = \{(f(X_1), \ldots, f(X_n)) \text{ s.t. } f \in \mathcal{F}\}$.

11.2.1 Example: Linear Classifiers in $\mathbb{R}^2$

Let $\mathcal{F} = \{\text{all linear classifiers in } \mathbb{R}^2\}$.

1. $S(1, \mathcal{F})$**:** Take an arbitrary point $X_1 \in \mathbb{R}^2$. How many different values can $\mathcal{F}$ induce on X_1?

- Can we induce 0? Yes, by placing the hyperplane such that $X_1 \notin \mathcal{H}_{a,b}^+$.
- Can we induce 1? Yes, by placing the hyperplane such that $X_1 \in \mathcal{H}_{a,b}^+$.

So $|\mathcal{F}_{X_1}| = 2$, hence $S(1, \mathcal{F}) = 2$.

2. $S(2, \mathcal{F})$**:** Let $X_1, X_2 \in \mathbb{R}^2$. Can we find functions to induce $(0,0), (1,0), (0,1), (1,1)$? Yes, by drawing lines to separate the points in all 4 possible ways. Hence $S(2, \mathcal{F}) = 4 = 2^2$.

3. $S(3, \mathcal{F})$**:** If we take X_1, X_2, X_3 as collinear points, we cannot induce the tuple $(1,0,1)$. If X_1 and X_3 are in the positive half-space, the convexity of the half-space forces X_2 to also be in the positive half-space. Thus $|\mathcal{F}_{X_1,X_2,X_3}| < 8$ for collinear points. Does this mean $S(3, \mathcal{F}) < 8$? No! The supremum is taken over *all* arrangements. If we place

X_1, X_2, X_3 in a triangle (non-collinear), we can shatter them (induce all 8 labelings). Thus, $S(3, \mathcal{F}) = 8 = 2^3$.

4. $S(4, \mathcal{F})$**:** For $n = 4$, it turns out that no matter how you arrange 4 points in $\mathbb{R}^2$, you can never achieve all 16 labelings. Thus, $S(4, \mathcal{F}) < 16 = 2^4$.

11.3 Vapnik-Chervonenkis (VC) Bounds

Shattering numbers can be used to obtain bounds for the uniform deviation.

Theorem (Vapnik-Chervonenkis): Provided $t \geq \sqrt{\frac{2}{n}}$, then:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R_n(f) - R_\rho(f)| \geq t\right) \leq 4S(2n, \mathcal{F}) \exp\left(-\frac{nt^2}{8}\right)$$

Setting $\delta = 4S(2n, \mathcal{F}) \exp(-nt^2/8)$ gives:

$$t = \sqrt{\frac{8}{n} \log\left(\frac{4S(2n, \mathcal{F})}{\delta}\right)}$$

As long as $\sqrt{8 \log(4S(2n, \mathcal{F}))} \geq \sqrt{2}$, with probability at least $1 - \delta$:

$$\sup_{f\in\mathcal{F}} |R_n(f) - R_\rho(f)| \leq \sqrt{\frac{8}{n} \log\left(\frac{4S(2n, \mathcal{F})}{\delta}\right)} := \epsilon(n, \mathcal{F}, \delta)$$

Observation: Recall $S(2n, \mathcal{F}) \leq 2^{2n} = 4^n$. If $\mathcal{F}$ is too rich and $S(2n, \mathcal{F}) = 4^n$, the bound becomes:

$$\approx \sqrt{\frac{8}{n} \log(4^n)} = \sqrt{\frac{8n \log(4)}{n}} = \sqrt{8 \log(4)} = \text{Constant}$$

Therefore, if $S(2n, \mathcal{F})$ grows exponentially in n, the concentration bound is NOT useful (it doesn't go to 0).

11.4 VC Dimension

When does $S(n, \mathcal{F})$ grow exponentially?

Definition 10. *The **VC-dimension** of $\mathcal{F}$ is:*

$$VC(\mathcal{F}) := \max\{n \in \mathbb{N} \ s.t. \ S(n, \mathcal{F}) = 2^n\}$$

This is the size of the largest set of points that can be shattered by $\mathcal{F}$. If $VC(\mathcal{F}) = N$, then $S(N, \mathcal{F}) = 2^N$ but $S(N + 1, \mathcal{F}) < 2^{N+1}$. Note: $VC(\mathcal{F})$ can be ∞ if $S(n, \mathcal{F}) = 2^n$ for all n.

Sauer-Shelah Lemma (Vapnik, Chervonenkis, Sauer, Shelah): If $VC(\mathcal{F}) < \infty$, then for all n:

$$S(n, \mathcal{F}) \leq \left(\frac{en}{VC(\mathcal{F})}\right)^{VC(\mathcal{F})}$$

Here e is the Euler's number. In other words, if the VC dimension is finite, the shattering number grows **polynomially** in n! The degree of the polynomial is determined by $VC(\mathcal{F})$.

Corollary: If $VC(\mathcal{F}) < \infty$, then with probability greater than $1-\delta$:

$$\sup_{f\in\mathcal{F}} |R_n(f) - R_\rho(f)| \leq \sqrt{\frac{8}{n}\log(\frac{4S(2n,\mathcal{F})}{\delta})}$$
$$\leq \sqrt{\frac{8}{n}\log(\frac{4(en)^{VC(\mathcal{F})}}{\delta VC(\mathcal{F})^{VC(\mathcal{F})}})}$$
$$\leq \sqrt{\frac{8}{n}VC(\mathcal{F})\log\left(\frac{4en}{VC(\mathcal{F})}\right) + \frac{8}{n}\log\left(\frac{1}{\delta}\right)}$$

Note: As $n \to \infty$, this bound goes to 0 (since $\log(n)/n \to 0$). Also note: This bound does NOT depend on the distribution ρ, only on the class $\mathcal{F}$.

Examples:

- $\mathcal{F} = \{\text{all linear classifiers in } \mathbb{R}^d\}$: $VC(\mathcal{F}) = d+1$.
- $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$ (Thresholds on the real line): $VC(\mathcal{F}) = 1 < \infty$. This implies the Glivenko-Cantelli theorem:

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(-\infty,t]}(X_i) - F(t)\right| = 0 \quad \text{a.s.}$$

  where $F(t)$ is the CDF.

11.5 Rademacher Complexity

VC dimension is theoretically useful but difficult to compute in practice. Rademacher complexity is a data-dependent measure of the richness of a function class.

Convention change: Assume classifiers map to $\{-1,1\}$ instead of $\{0,1\}$, so $f : \mathcal{X} \to \{-1,1\}$ and $Y_i \in \{-1,1\}$.

Definition 11. *Given a family $\mathcal{F}$ of functions mapping to $\{-1,1\}$, we define:*

- ***Conditional Rademacher Average:***

$$\tilde{Rad}_n(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} R_n^\sigma(f)\right]$$

- ***Rademacher Average (Complexity):***

$$Rad_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n^\sigma(f)\right] = \mathbb{E}_X[\tilde{Rad}_n(\mathcal{F})]$$

Here $R_n^\sigma(f) = \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$. $X_1, ..., X_n$ are the data, and $\sigma_1, ..., \sigma_n$ are independent ***Rademacher variables*** *($\mathbb{P}(\sigma_i = -1) = 1/2, \mathbb{P}(\sigma_i = 1) = 1/2$). $\mathbb{E}_\sigma$ is the expectation over σ only, given the data.*

Notice that $\tilde{\text{Rad}}_n(\mathcal{F})$ is a random variable depending on the sample $X_1, ..., X_n$. It measures how well the function class can correlate with random noise (the Rademacher variables) on the specific dataset.

Generalization Bounds via Rademacher Complexity:

Theorem 1: For all $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} (R(f) - R_n(f)) \leq 2\text{Rad}_n(\mathcal{F}) + \sqrt{\frac{c \log(1/\delta)}{n}}$$

Corollary: If $\text{Rad}_n(\mathcal{F}) \to 0$ as $n \to \infty$, then ERM with $\mathcal{F}$ is universally strongly consistent. Unfortunately, $\text{Rad}_n(\mathcal{F})$ is still distribution-dependent.

Theorem 2 (Data-dependent bound): For all $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} (R(f) - R_n(f)) \leq 2\tilde{\text{Rad}}_n(\mathcal{F}) + \sqrt{\frac{\tilde{c} \log(1/\delta)}{n}}$$

Note: This uses the *conditional* Rademacher average, which can be estimated directly from the training data without knowing the true distribution!

12 Lecture 12: Rademacher Complexity and Generalization Bounds

12.1 Definitions and Intuition

Recall new convention: $\mathcal{Y} = \{-1, 1\}$.

We define two notions of complexity for a function class $\mathcal{F}$ mapping to $\{-1, 1\}$:

$$\text{Rad}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} R_n^\sigma(f)\right]$$

$$\tilde{\text{Rad}}_n(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} R_n^\sigma(f)\right]$$

where:

$$R_n^\sigma(f) := \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$$

- $\sigma_1, ..., \sigma_n$ are independent Rademacher random variables ($\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$).
- They are independent of the data $(X_1, Y_1), ..., (X_n, Y_n)$.
- $\mathbb{E}_\sigma[\cdot]$ denotes the expectation over $\sigma_1, ..., \sigma_n$ only.

Question: What is $\text{Rad}_n(\mathcal{F})$ (or $\tilde{\text{Rad}}_n(\mathcal{F})$) trying to measure?

Let's look at two extreme cases:

1. **Case 1 (Simplest family):** $\mathcal{F} = \{\tilde{f}\}$, where $\tilde{f} \equiv 1$ (a constant function).

   $$\sup_{f \in \mathcal{F}} R_n^\sigma(f) = R_n^\sigma(\tilde{f}) = \frac{1}{n}\sum_{i=1}^n \sigma_i$$

   This is the average of zero-mean independent random variables. Taking the expectation, we get exactly:

   $$\text{Rad}_n(\mathcal{F}) = \tilde{\text{Rad}}_n(\mathcal{F}) = 0$$

   $\mathcal{F}$ is the simplest family, and $\text{Rad}_n(\mathcal{F}) = 0$ reflects that.

2. **Case 2 (Most complex family):** $\mathcal{F} = \{\text{all binary classifiers}\}$.

   $$\sup_{f \in \mathcal{F}} R_n^\sigma(f) = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$$

   Since $\mathcal{F}$ contains all functions, for any realization of σ_i, we can simply choose a function f such that $f(X_i) = \sigma_i$.

   $$\geq \frac{1}{n}\sum_{i=1}^n \sigma_i^2 = 1$$

   Since we also naturally have $R_n^\sigma(f) \leq 1$, in this case we get:

   $$\text{Rad}_n(\mathcal{F}) = \tilde{\text{Rad}}_n(\mathcal{F}) = 1$$

   This $\mathcal{F}$ is the most complex family, and the Rademacher complexity is maximized at 1.

Remark: We will see that for generalization to occur, we need $\text{Rad}_n(\mathcal{F}), \tilde{\text{Rad}}_n(\mathcal{F}) \to 0$ as $n \to \infty$.

12.2 Generalization Bounds based on Rademacher Complexity

Theorem 1. *For all $\delta \in (0,1)$, with probability at least $1-\delta$ we have:*

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq 2Rad_n(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

Theorem 2. *For all $\delta \in (0,1)$, with probability at least $1-\delta$ we have:*

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq 2\tilde{Rad}_n(\mathcal{F}) + 3\sqrt{\frac{2\log(2/\delta)}{n}}$$

(Note: The empirical/conditional Rademacher complexity incurs an extra penalty term when substituting the true Rademacher average).

Remark: The above theorems, especially Theorem 2, are more useful than the concentration bound using VC-dimension:

1. VC-dimension can be difficult to compute or estimate for many families $\mathcal{F}$.
2. The inequality is distribution-dependent, which can yield tighter bounds.

In contrast, as we will soon see, $\tilde{\text{Rad}}_n(\mathcal{F})$ can be estimated in practice directly from the training data.

12.3 Proof of Theorem 1

We are interested in bounding $\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$. We will do it in two steps:

$$\begin{aligned}\sup_{f \in \mathcal{F}} |R_n(f) - R(f)| = &\left(\sup_{f \in \mathcal{F}} |R_n(f) - R(f)| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|\right]\right) \\ &+ \mathbb{E}\left[\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|\right]\end{aligned}$$

12.3.1 Step 1: Concentration via McDiarmid's Inequality

Given $(x_1, y_1), ..., (x_n, y_n)$ arbitrary points, define:

$$g((x_1, y_1), ..., (x_n, y_n)) := \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$

We can check that the function g satisfies the conditions for using McDiarmid's inequality. Consider changing one data point from (x_i, y_i) to (x_i', y_i'):

$$\begin{aligned}&g((x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)) - g((x_1, y_1), ..., (x_i', y_i'), ..., (x_n, y_n)) \\ &= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j \neq i} \mathbf{1}_{f(x_j) \neq y_j} + \frac{1}{n} \mathbf{1}_{f(x_i) \neq y_i} - R(f) \right| - \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j \neq i} \mathbf{1}_{f(x_j) \neq y_j} + \frac{1}{n} \mathbf{1}_{f(x_i') \neq y_i'} - R(f) \right|\end{aligned}$$

Using the property that $|\sup|A| - \sup|B|| \leq \sup|A-B|$, and letting f^* be the maximizer for the first supremum:

$$\begin{aligned}&\leq \left| \frac{1}{n} \mathbf{1}_{f^*(x_i) \neq y_i} - \frac{1}{n} \mathbf{1}_{f^*(x_i') \neq y_i'} \right| \\ &\leq \frac{1}{n} \left| \mathbf{1}_{f^*(x_i) \neq y_i} - \mathbf{1}_{f^*(x_i') \neq y_i'} \right| \leq \frac{1}{n}\end{aligned}$$

(Note: Depending on the loss formulation, bounds often use $c_i = 1/n$ or $c_i = 2/n$. To align with the standard theorem statement yielding $\sqrt{2\log(2/\delta)/n}$, we use $c_i = 2/n$.) Let's assign $c_i := \frac{2}{n}$.

Using McDiarmid's inequality, we get:

$$\begin{aligned}
\mathbb{P}\Big(g(\dots) - \mathbb{E}[g(\dots)] \geq t\Big) &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \\
&= \exp\left(-\frac{2t^2}{n(4/n^2)}\right) = \exp\left(-\frac{nt^2}{2}\right) =: \frac{\delta}{2}
\end{aligned}$$

(Note: we apply it two-sided to get the absolute value, hence $\delta/2$ or $2\exp(-nt^2/2) = \delta$.)

In terms of δ, t can be written as:

$$t = \sqrt{\frac{2\log(2/\delta)}{n}}$$

12.3.2 Step 2: Bounding the Expectation

It remains to bound the expected value of the supremum by the Rademacher complexity:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|\right] \leq 2\mathrm{Rad}_n(\mathcal{F})$$

(This completes the first part of the proof; the detailed symmetrization argument for Step 2 typically follows in the next lecture).

13 Lecture 13: Symmetrization and Linear Classifiers

13.1 Proof of Theorem 1 (Continued)

It remains to prove:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|R_n(f)-R(f)|\right]\leq 2\mathrm{Rad}_n(\mathcal{F})$$

13.1.1 Symmetrization via Ghost Samples

Consider an independent copy of the data $(X_1,Y_1),...,(X_n,Y_n)$ that we denote:

$$(X_1',Y_1'),...,(X_n',Y_n')$$

These are also i.i.d. from ρ. These are often called "ghost samples". In particular, for any fixed f:

$$\mathbb{E}\left[\mathbf{1}_{f(X_i')\neq Y_i'}\right]=R_\rho(f)$$

Recall the empirical risks:

$$R_n(f)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{f(X_i)\neq Y_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{1-f(X_i)Y_i}{2}$$
$$R_n'(f)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{f(X_i')\neq Y_i'}=\frac{1}{n}\sum_{i=1}^{n}\frac{1-f(X_i')Y_i'}{2}$$

Now, we rewrite the true risk using the ghost sample:

$$\mathbb{E}_{(X_i,Y_i)_{i=1\ldots n}\sim\rho}\left[\sup_{f\in\mathcal{F}}|R_n(f)-R_\rho(f)|\right]=\mathbb{E}_{(X_i,Y_i)_{i=1\ldots n}}\left[\sup_{f\in\mathcal{F}}\left|R_n(f)-\mathbb{E}_{(X_i',Y_i')_{i=1\ldots n}}[R_n'(f)]\right|\right]$$

Because $R_n(f)$ does not depend on the ghost sample $(X_i',Y_i')_{i=1\ldots n}$, we can bring it inside the inner expectation:

$$=\mathbb{E}_{(X_i,Y_i)_{i=1\ldots n}}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{(X_i',Y_i')_{i=1\ldots n}}\left[R_n(f)-R_n'(f)\right]\right|\right]$$

Because the supremum of an expectation is less than or equal to the expectation of the supremum ($\sup|\mathbb{E}[\cdot]|\leq\mathbb{E}[\sup|\cdot|]$):

$$\leq\mathbb{E}_{(X_i,Y_i)_{i=1\ldots n}}\left[\mathbb{E}_{(X_i',Y_i')_{i=1\ldots n}}\left[\sup_{f\in\mathcal{F}}|R_n(f)-R_n'(f)|\right]\right]$$

Combining the expectations:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|R_n(f)-R(f)|\right]\leq\mathbb{E}_{(X_i,Y_i),(X_i',Y_i')}\left[\sup_{f\in\mathcal{F}}|R_n(f)-R_n'(f)|\right]$$

13.1.2 Introducing Rademacher Variables

Observe that for any realization of the data and ghost data:

$$R_n(f) - R'_n(f) = \frac{1}{2n}\sum_{i=1}^{n}\left(f(X'_i)Y'_i - f(X_i)Y_i\right)$$

Because (X_i, Y_i) and (X'_i, Y'_i) are drawn i.i.d. from the same distribution, swapping them does not change the distribution of the sum. This means $f(X'_i)Y'_i - f(X_i)Y_i$ has the exact same distribution as:

$$\sigma_i\left(f(X'_i)Y'_i - f(X_i)Y_i\right)$$

where $\sigma_i \in \{-1, 1\}$ are independent Rademacher random variables. To understand this, suppose $n = 2$ and $\sigma_1 = +1, \sigma_2 = -1$. Original sum:

$$(f(X'_1)Y'_1 - f(X_1)Y_1) + (f(X'_2)Y'_2 - f(X_2)Y_2)$$

With σ:

$$(f(X'_1)Y'_1 - f(X_1)Y_1) - (f(X'_2)Y'_2 - f(X_2)Y_2) = (f(X'_1)Y'_1 - f(X_1)Y_1) + (f(X_2)Y_2 - f(X'_2)Y'_2)$$

It is as if the two sets (X_2, Y_2) and (X'_2, Y'_2) were swapped. The distribution remains identical.

Hence, we can introduce σ_i:

$$\begin{aligned}
&\mathbb{E}_{(X_i,Y_i),(X'_i,Y'_i)}\left[\sup_{f\in\mathcal{F}}|R_n(f) - R'_n(f)|\right] \\
&= \mathbb{E}_{\sigma,(X_i,Y_i),(X'_i,Y'_i)}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{2n}\sum_{i=1}^{n}\sigma_i(f(X'_i)Y'_i - f(X_i)Y_i)\right|\right]
\end{aligned}$$

In turn, using the triangle inequality $(\sup|A+B| \le \sup|A| + \sup|B|)$:

$$\begin{aligned}
&\le \mathbb{E}_{\sigma,X,X',Y,Y'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{2n}\sum_{i=1}^{n}\sigma_i f(X'_i)Y'_i\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{2n}\sum_{i=1}^{n}\sigma_i f(X_i)Y_i\right|\right] \\
&= 2\mathbb{E}_{\sigma,X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{2n}\sum_{i=1}^{n}\sigma_i Y_i f(X_i)\right|\right]
\end{aligned}$$

13.1.3 Dropping the Labels Y_i

To understand this final step: note that the random variable $\sigma_i Y_i | X_i$ has the same distribution as $\sigma_i | X_i$. Indeed, since $Y_i \in \{-1, 1\}$ and σ_i is a symmetric coin flip independent of Y_i:

$$\begin{aligned}
\mathbb{P}(Y_i\sigma_i = 1|X_i) &= \mathbb{P}(Y_i = 1, \sigma_i = 1|X_i) + \mathbb{P}(Y_i = -1, \sigma_i = -1|X_i) \\
&= \mathbb{P}(\sigma_i = 1|X_i, Y_i = 1)\mathbb{P}(Y_i = 1|X_i) + \mathbb{P}(\sigma_i = -1|X_i, Y_i = -1)\mathbb{P}(Y_i = -1|X_i) \\
&= \frac{1}{2}\mathbb{P}(Y_i = 1|X_i) + \frac{1}{2}\mathbb{P}(Y_i = -1|X_i) \\
&= \frac{1}{2}(\mathbb{P}(Y_i = 1|X_i) + \mathbb{P}(Y_i = -1|X_i)) = \frac{1}{2} \\
&= \mathbb{P}(\sigma_i = 1|X_i)
\end{aligned}$$

Since the distribution is the same, we can drop the Y_i:

$$= 2\mathbb{E}_{\sigma,X}\left[\sup_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^{n}\sigma_i Y_i f(X_i)\right] = \mathbb{E}_{\sigma,X}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right] = \mathrm{Rad}_n(\mathcal{F})$$

(Note: the absolute value can be dropped because if σ_i is symmetric, $-\sigma_i$ has the same distribution). This completes the proof.

13.2 Estimating Rademacher Complexity in Practice

Recall the conditional Rademacher complexity:

$$\tilde{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right]$$

This depends on the fixed training data $X_1, ..., X_n$. To estimate it:

1. Choose K large enough.
2. For $k = 1, ..., K$:
   - Generate n independent Rademacher variables $\sigma_1^{(k)}, ..., \sigma_n^{(k)}$.
   - Solve the optimization problem (which is the same computational problem as ERM!):
     $$m_k = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i^{(k)} f(X_i)$$
3. Approximate:
   $$\hat{\mathrm{Rad}}_n(\mathcal{F}) := \frac{1}{K}\sum_{k=1}^{K} m_k \approx \tilde{\mathrm{Rad}}_n(\mathcal{F})$$

Remark: You can set K as large as you want because generating Rademacher variables is computationally cheap (unlike getting more real training samples).

13.3 Summary of Part 1

1. Introduced the notion of the Bayes classifier.
2. Studied two types of data-driven classifiers:
   - **Plug-in Classifiers:** (e.g., Histograms, k-NN). Studied via Stone's Theorem and Concentration Bounds.
   - **ERM-based Classifiers:** Studied via VC-theory and Rademacher Complexity.
3. This was done to answer the question: When can data-driven classifiers generalize?

13.4 Part 2: Linear and Kernel Classifiers

13.4.1 Linear Classifiers

What choices of $\mathcal{F}$ are used in practice?

- **Linear Classifiers** on $\mathcal{X} = \mathbb{R}^d$: $f_{w,b}(x) = 1$ if $\langle w, x \rangle + b \geq 0$, else 0 (or -1).
- **Kernel-based Classifiers**: $f_h(x) = 1$ if $h(x) \geq 0$, where h belongs to a Reproducing Kernel Hilbert Space (RKHS).
- **Neural Networks**: $f_{w,b}(x) = 1$ if $\sum_{j=1}^{W} v_j \sigma(\langle w_j, x \rangle + b_j) \geq 0$ (e.g., Feed-forward NN with width W and one hidden layer).

In Part 2, we will investigate linear classifiers and kernel-based classifiers more carefully.

13.4.2 Linear Classifiers in $\mathbb{R}^d$

Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$. Let $\beta \in \mathbb{R}^d$ (with $\beta \neq \vec{0}$) and $\beta_0 \in \mathbb{R}$. Define the decision boundary (a hyperplane):

$$\mathcal{H}_{\beta,\beta_0} := \{x \in \mathbb{R}^d : \langle \beta, x \rangle + \beta_0 = 0\}$$

Let x_0 be a point on the hyperplane. Then $\langle \beta, x_0 \rangle = -\beta_0$. For any $x \in \mathcal{H}_{\beta,\beta_0}$:

$$\langle \beta, x - x_0 \rangle = \langle \beta, x \rangle - \langle \beta, x_0 \rangle = -\beta_0 - (-\beta_0) = 0$$

So β is the normal vector to the hyperplane.

Define the half-spaces:

$$\mathcal{H}^+_{\beta,\beta_0} := \{x \in \mathbb{R}^d : \langle \beta, x \rangle + \beta_0 \geq 0\}$$
$$\mathcal{H}^-_{\beta,\beta_0} := \{x \in \mathbb{R}^d : \langle \beta, x \rangle + \beta_0 \leq 0\}$$

The linear classifier is:

$$f_{\beta,\beta_0}(x) = \begin{cases} 1 & \text{if } \langle \beta, x \rangle + \beta_0 \geq 0 \\ -1 & \text{if } \langle \beta, x \rangle + \beta_0 < 0 \end{cases}$$

Question: Given data $(X_i, Y_i)_{i=1\dots n}$, how do we construct a linear classifier? Three different ways to obtain β, β_0 from data:

1. **LDA and QDA** (Linear / Quadratic Discriminant Analysis) - Generative, Bayes classifier inspired.
2. **Logistic Regression** - ERM based.
3. **Perceptrons and Support Vector Machines (SVM)** - ERM + Margin criterion based.

13.4.3 1. LDA and QDA

Let $\mathcal{Y} = \{1, 2, ..., K\}$ (not just binary), $\mathcal{X} \in \mathbb{R}^d$. Assume a **Gaussian Mixture Model** generative process for $(X, Y) \sim \rho$:

$$\begin{aligned}\mathbb{P}(Y = j) &= w_j \\ X|Y = j &\sim \mathcal{N}(\mu_j, \Sigma_j)\end{aligned}$$

The Bayes classifier minimizes the risk by picking the most probable class:

$$\begin{aligned}f_B(x) &= \arg\max_{j=1...K} \mathbb{P}(Y = j|X = x) \\ &= \arg\max_{j=1...K} \frac{\rho_{X|Y=j}(x) \cdot w_j}{\rho_X(x)} \\ &= \arg\max_{j=1...K} \rho_{X|Y=j}(x) \cdot w_j \\ &= \arg\max_{j=1...K} \log\left(\rho_{X|Y=j}(x) \cdot w_j\right)\end{aligned}$$

Plugging in the multivariate Gaussian density:

$$\begin{aligned}&= \arg\max_{j=1...K} \left\{\log\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma_j)^{1/2}}\right) - \frac{1}{2}\langle\Sigma_j^{-1}(x-\mu_j), (x-\mu_j)\rangle + \log(w_j)\right\} \\ &= \arg\min_{j=1...K} \left\{\frac{1}{2}\log(\det(\Sigma_j)) + \frac{1}{2}\langle\Sigma_j^{-1}(x-\mu_j), (x-\mu_j)\rangle - \log(w_j)\right\}\end{aligned}$$

Let the term inside the arg min be $\delta_j(x)$. The optimal rule is $f_B(x) = \arg\min_j \delta_j(x)$. Because $\delta_j(x)$ contains a quadratic term $\langle\Sigma_j^{-1}x, x\rangle$ that depends on j, the decision boundaries $\delta_i(x) = \delta_j(x)$ are quadratic. This is **Quadratic Discriminant Analysis (QDA)**.

Estimating Parameters from Data (Plug-in principle): Given $(X_1, Y_1), ..., (X_n, Y_n)$:

$$\begin{aligned}\hat{\mu}_j &= \frac{1}{N_j}\sum_{i:Y_i=j} X_i \quad \text{(where } N_j = \#\{i : Y_i = j\}) \\ (\hat{\Sigma}_j)_{ml} &= \frac{1}{N_j}\sum_{i:Y_i=j}(X_{i,l} - \hat{\mu}_{j,l})(X_{i,m} - \hat{\mu}_{j,m}) \\ \hat{w}_j &= \frac{N_j}{n}\end{aligned}$$

We then define the data-driven classifier $\hat{f}_n(x) = \arg\min_j \hat{\delta}_j(x)$.

13.4.4 Linear Discriminant Analysis (LDA)

Assume all classes share the same covariance matrix: $\Sigma_1 = ... = \Sigma_K = \Sigma$. Then:

$$\begin{aligned}\delta_j(x) &= \frac{1}{2}\log(\det(\Sigma)) + \frac{1}{2}\langle\Sigma^{-1}(x-\mu_j), (x-\mu_j)\rangle - \log(w_j) \\ &= (\text{terms independent of } j) - \langle\Sigma^{-1}x, \mu_j\rangle + \frac{1}{2}\langle\Sigma^{-1}\mu_j, \mu_j\rangle - \log(w_j)\end{aligned}$$

The quadratic term $\langle\Sigma^{-1}x, x\rangle$ is now the same for all j, so it does not affect the arg min and can be ignored. The Bayes classifier becomes:

$$f_B(x) = \arg\min_{j=1...K} \left\{-\langle\Sigma^{-1}\mu_j, x\rangle + \frac{1}{2}\langle\Sigma^{-1}\mu_j, \mu_j\rangle - \log(w_j)\right\}$$

Notice that the decision function is now **linear** in x. The decision boundaries are hyperplanes. We estimate the shared Σ by pooling the empirical covariances, and plug the estimates into the linear formula to get $\hat{f}_n(x)$.

14 Lecture 14: Linear Classifiers (Part 2)

14.1 Setup and Hyperplanes

Consider the family of all linear classifiers in $\mathbb{R}^d$:

$$\mathcal{F} = \{\text{all linear classifiers in } \mathbb{R}^d\}$$

We use the convention $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. Let $\beta \in \mathbb{R}^d$ (with $\beta \neq 0$) and $\beta_0 \in \mathbb{R}$. We define the decision boundary as the hyperplane:

$$\mathcal{H}_{\beta,\beta_0} := \{x \in \mathbb{R}^d : \langle \beta, x \rangle + \beta_0 = 0\}$$

Note: If $x_0 \in \mathcal{H}_{\beta,\beta_0}$ and $x \in \mathcal{H}_{\beta,\beta_0}$:

$$\langle x - x_0, \beta \rangle = \langle x, \beta \rangle - \langle x_0, \beta \rangle = -\beta_0 + \beta_0 = 0$$

That is, the vector $x - x_0$ on the hyperplane is perpendicular to β. Thus, β is the normal vector to the hyperplane.

We define the two half-spaces:

$$\begin{aligned} \mathcal{H}^+_{\beta,\beta_0} &:= \{x \in \mathbb{R}^d \text{ s.t. } \langle \beta, x \rangle + \beta_0 \geq 0\} \\ \mathcal{H}^-_{\beta,\beta_0} &:= \{x \in \mathbb{R}^d \text{ s.t. } \langle \beta, x \rangle + \beta_0 \leq 0\} \end{aligned}$$

The classifier is defined as:

$$f_{\beta,\beta_0}(x) = \begin{cases} 1 & \text{if } \langle \beta, x \rangle + \beta_0 \geq 0 \\ -1 & \text{if } \langle \beta, x \rangle + \beta_0 < 0 \end{cases}$$

Question: Given observations $(X_i, Y_i)_{i=1}^n$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1, 1\}$, how do we construct a linear classifier from the observations?

We will discuss at least 3 different ways to tune $\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}$:

1. **LDA and QDA** (Linear / Quadratic Discriminant Analysis) - *Bayes Classifier inspired.*

2. **Logistic Regression** - *ERM based.*

3. **Perceptrons and SVMs** (Support Vector Machines) - *ERM + Margin Criterion.*

14.2 1. LDA and QDA

Let $\mathcal{Y} = \{1, 2, ..., K\}$ (not just binary classification). Assume $(X, Y) \sim \rho$ where ρ is a **Gaussian Mixture Model**:

$$\begin{aligned} w_j &= \mathbb{P}(Y = j) \quad \text{for } j = 1, ..., K \\ X|Y = j &\sim \mathcal{N}(\mu_j, \Sigma_j) \end{aligned}$$

Here w_j, μ_j, Σ_j are the parameters. For this model, the Bayes classifier is:

$$\begin{aligned}
f_{Bayes}(x) &= \arg\max_{j=1...K} \mathbb{P}(Y = j|X = x) \\
&= \arg\max_{j=1...K} \frac{\rho_{X|Y=j}(x) \cdot w_j}{\rho_X(x)} \\
&= \arg\max_{j=1...K} \rho_{X|Y=j}(x) \cdot w_j \\
&= \arg\max_{j=1...K} \log\left(\rho_{X|Y=j}(x) \cdot w_j\right) \\
&= \arg\max_{j=1...K} \left\{\log\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma_j)^{1/2}}\right) - \frac{1}{2}\langle\Sigma_j^{-1}(x-\mu_j), (x-\mu_j)\rangle + \log(w_j)\right\} \\
&= \arg\min_{j=1...K} \left\{-\log\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma_j)^{1/2}}\right) + \frac{1}{2}\langle\Sigma_j^{-1}(x-\mu_j), (x-\mu_j)\rangle - \log(w_j)\right\} \\
&= \arg\min_{j=1...K} \delta_j(x)
\end{aligned}$$

Why QDA? Because $\delta_j(x)$ is a quadratic function of x.

Estimating from Data: What do we do when we only have data $(X_1, Y_1), ..., (X_n, Y_n)$? We estimate the parameters:

$$\begin{aligned}
\hat{\mu}_j &:= \frac{\sum_{Y_i=j} X_i}{\#\{i : Y_i = j\}} \approx \mu_j \\
(\hat{\Sigma}_j)_{ml} &:= \frac{1}{N_j}\sum_{Y_i=j}(X_{i,l} - \hat{\mu}_{j,l})(X_{i,m} - \hat{\mu}_{j,m}) \approx (\Sigma_j)_{ml} \\
\hat{w}_j &:= \frac{N_j}{n}
\end{aligned}$$

Then we define our empirical classifier:

$$\hat{f}_n(x) := \arg\min_{j=1...K} \hat{\delta}_j(x)$$

Note: We don't have to strictly assume that (X_i, Y_i) actually come from this distribution to define and use the above classifier.

LDA (Linear Discriminant Analysis): This is a special case of QDA when we assume $\Sigma_1 = ... = \Sigma_K = \Sigma$. In this case:

$$\begin{aligned}
\delta_j(x) &= -\log\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\right) + \frac{1}{2}\langle\Sigma^{-1}(x-\mu_j), (x-\mu_j)\rangle - \log(w_j) \\
&= (\text{Something that doesn't depend on } j) - \langle\Sigma^{-1}x, \mu_j\rangle + \frac{1}{2}\langle\Sigma^{-1}\mu_j, \mu_j\rangle - \log(w_j)
\end{aligned}$$

Hence:

$$\begin{aligned}
f_{Bayes}(x) &= \arg\min_{j=1...K} \delta_j(x) \\
&= \arg\min_{j=1...K} \left\{-\langle\Sigma^{-1}\mu_j, x\rangle + \frac{1}{2}\langle\Sigma^{-1}\mu_j, \mu_j\rangle - \log(w_j)\right\} \\
&= \arg\min_{j=1...K} \ell_j(x)
\end{aligned}$$

where $\ell_j(x)$ is a linear function of x. For the data-driven version, $\hat{f}_n(x) = \arg\min_j \hat{\ell}_j(x)$, replacing μ, Σ, w with empirical estimates $\hat{\mu}, \hat{\Sigma}, \hat{w}$.

14.3 2. Logistic Regression

Let $\mathcal{Y} = \{1, ..., K\}$. The parameters are $(\vec{\beta}_j, \beta_{0j}) \in \mathbb{R}^d \times \mathbb{R}$ for $j = 1, ..., K-1$. Now we imagine $(X, Y) \sim \rho$ where:

$$\mathbb{P}(Y = j | X = x) = \frac{\exp(\langle x, \vec{\beta}_j \rangle + \beta_{0j})}{1 + \sum_{l=1}^{K-1} \exp(\langle x, \vec{\beta}_l \rangle + \beta_{0l})} \quad \text{if } j = 1, ..., K-1$$
$$\mathbb{P}(Y = K | X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\langle x, \vec{\beta}_l \rangle + \beta_{0l})}$$

The Bayes classifier for this model is:

$$\begin{aligned} f_{Bayes}(x) &= \arg \max_{j=1...K} \mathbb{P}(Y = j | X = x) \\ &= \arg \max_{j=1...K} \varphi_j(x) \end{aligned}$$

where $\varphi_j(x) = \exp(\langle x, \vec{\beta}_j \rangle + \beta_{0j})$ for $j = 1...K-1$, and $\varphi_K(x) = 1$. Taking the log:

$$\begin{aligned} f_{Bayes}(x) &= \arg \max_{j=1...K} \log(\varphi_j(x)) \\ \log(\varphi_j(x)) &= \langle x, \vec{\beta}_j \rangle + \beta_{0j} =: \ell_j(x) \quad \text{(linear function of } x) \\ \log(\varphi_K(x)) &= 0 \end{aligned}$$

In summary: $f_{Bayes}(x) = \arg\max_j \ell_j(x)$.

Estimating from Data: We estimate parameters from data using Maximum Likelihood Estimation (MLE):

$$\max_{\{\vec{\beta}_j, \beta_{0j}\}_{j=1}^{K-1}} \sum_{i=1}^{n} \log(\mathbb{P}(Y = Y_i | X = X_i))$$

Optimize over the parameter choices to get $\{\hat{\beta}_j, \hat{\beta}_{0j}\}_{j=1}^{K-1}$. Then define:

$$\hat{f}_n(x) := \arg \max_{j=1...K} \hat{\ell}_j(x)$$

14.4 3. Perceptrons and SVMs

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. Parameters are $(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R}$.

14.4.1 Perceptron

The objective function for the Perceptron is:

$$\mathcal{O}(\beta, \beta_0) = \sum_{i \in \mathcal{M}_{\beta,\beta_0}} \text{dist}(X_i, \mathcal{H}_{\beta,\beta_0})$$

where $\mathcal{M}_{\beta,\beta_0} := \{i \text{ s.t. } Y_i \neq f_{\beta,\beta_0}(X_i)\}$ is the set of misclassified points. **Remark:** This objective function is not differentiable. Moreover, it may have infinitely many solutions, not all of them desirable (multiple valid separating hyperplanes).

14.4.2 Support Vector Machines (SVMs)

For motivation, let's start by assuming that (X_i, Y_i) are linearly separable. That is, $\exists \beta, \beta_0$ such that:

$$f_{\beta,\beta_0}(X_i) = Y_i \quad \forall i = 1, ..., n$$

(which implies $\mathcal{M}_{\beta,\beta_0} = \emptyset$).

In SVMs, we want to maximize the margin:

$$\begin{aligned} &\max_{\beta,\beta_0} \text{Margin}(\beta, \beta_0) \\ &\text{s.t. } \mathcal{M}_{\beta,\beta_0} = \emptyset \end{aligned}$$

where $\text{Margin}(\beta, \beta_0) = \min_{i=1...n} \text{dist}(X_i, \mathcal{H}_{\beta,\beta_0})$. A larger margin is intuitively good: we want classifiers to be less sensitive to perturbations of the data.

Our goal is to rewrite the margin maximization problem in a tractable way. As a first step, consider (β, β_0) such that $\mathcal{M}_{\beta,\beta_0} = \emptyset$ (with $\beta \neq 0$). Note that:

$$\begin{aligned} \mathcal{H}_{\beta,\beta_0} &= \{x \text{ s.t. } \langle x, \beta\rangle + \beta_0 = 0\} \\ &= \left\{x \text{ s.t. } \left\langle x, \frac{\beta}{\|\beta\|}\right\rangle + \frac{\beta_0}{\|\beta\|} = 0\right\} \\ &= \mathcal{H}_{\tilde{\beta},\tilde{\beta}_0} \end{aligned}$$

Thus, the above problem is equivalent to:

$$\begin{aligned} &\max_{\tilde{\beta},\tilde{\beta}_0} \text{Margin}(\tilde{\beta}, \tilde{\beta}_0) \\ &\text{s.t. } \mathcal{M}_{\tilde{\beta},\tilde{\beta}_0} = \emptyset \\ &\|\tilde{\beta}\| = 1 \end{aligned}$$

15 Lecture 15: Support Vector Machines (Part 2) and Duality

15.1 Geometric Version of SVMs

Recall our dataset: $(X_1, Y_1), ..., (X_n, Y_n)$ with $Y_i \in \{-1, 1\}$. Recall: A data set is **linearly separable** if:

$$\exists(\beta, \beta_0) \text{ s.t. } \mathcal{M}_{\beta,\beta_0} = \emptyset$$

where $\mathcal{M}_{\beta,\beta_0} := \{i \text{ s.t. } Y_i \neq f_{\beta,\beta_0}(X_i)\}$ is the set of misclassified points by f_{β,β_0}.

Definition 12. *Given (β, β_0) such that $\mathcal{M}_{\beta,\beta_0} = \emptyset$, the margin is defined as:*

$$\mathit{margin}(\beta, \beta_0) := \min_{i=1,\ldots,n} \mathit{dist}(X_i, \mathcal{H}_{\beta,\beta_0})$$

SVMs (Linearly Separable Case):

$$\begin{aligned} \max_{\beta,\beta_0} \quad & \text{margin}(\beta, \beta_0) \\ \text{s.t.} \quad & \mathcal{M}_{\beta,\beta_0} = \emptyset \end{aligned}$$

This is the geometric version of SVMs. **Goal:** To show that this problem is equivalent to a **convex optimization** problem.

15.2 Rewriting the Margin

First, note that for any $x \in \mathbb{R}^d$, the distance to the hyperplane is:

$$\text{dist}(x, \mathcal{H}_{\beta,\beta_0}) = \frac{|\langle x, \beta\rangle + \beta_0|}{\|\beta\|}$$

Proof: Let $x_0 \in \mathcal{H}_{\beta,\beta_0}$ be arbitrary. Then the vector $x - x_0$ projected onto the normal vector $\frac{\beta}{\|\beta\|}$ gives the distance:

$$\begin{aligned} \text{dist}(x, \mathcal{H}_{\beta,\beta_0}) &= \left\|\text{Proj}_\beta(x - x_0)\right\| \\ &= \left|\left\langle x - x_0, \frac{\beta}{\|\beta\|}\right\rangle\right| \\ &= \frac{|\langle \beta, x\rangle - \langle \beta, x_0\rangle|}{\|\beta\|} = \frac{|\langle \beta, x\rangle + \beta_0|}{\|\beta\|} \end{aligned}$$

Next, note that $\mathcal{M}_{\beta,\beta_0} = \emptyset$ if and only if:

$$Y_i(\langle X_i, \beta\rangle + \beta_0) > 0 \quad \forall i = 1, ..., n$$

Note: If $\alpha > 0$, then $\mathcal{H}_{\alpha\beta,\alpha\beta_0} = \mathcal{H}_{\beta,\beta_0}$ (indeed $\langle x, \beta\rangle + \beta_0 = 0 \iff \alpha(\langle x, \beta\rangle + \beta_0) = 0$). Hence, if linear separability holds, we can **rescale** (β, β_0) so that:

$$\min_{i=1,\ldots,n} Y_i(\langle X_i, \beta\rangle + \beta_0) = 1$$

which implies $\min_{i=1,\ldots,n} |\langle X_i, \beta\rangle + \beta_0| = 1$.

For pairs (β, β_0) satisfying this scaling, we thus have:

$$\min_{i=1,\ldots,n} \operatorname{dist}(X_i, \mathcal{H}_{\beta,\beta_0}) = \min_{i=1,\ldots,n} \frac{|\langle X_i, \beta\rangle + \beta_0|}{\|\beta\|} = \frac{1}{\|\beta\|}$$

In conclusion: The SVM problem is equivalent to:

$$\begin{aligned} \max_{\beta,\beta_0} \quad & \frac{1}{\|\beta\|} \\ \text{s.t.} \quad & Y_i(\langle X_i, \beta\rangle + \beta_0) \geq 1 \quad \forall i = 1, ..., n \end{aligned}$$

In turn, the above is equivalent to:

$$\begin{aligned} \min_{\beta,\beta_0} \quad & \frac{1}{2}\|\beta\|^2 \quad \text{(Convex Objective)} \\ \text{s.t.} \quad & Y_i(\langle X_i, \beta\rangle + \beta_0) \geq 1 \quad \forall i = 1, ..., n \quad \text{(Linear inequality constraints)} \end{aligned}$$

This is a **Convex Optimization Problem**.

Observations: The above optimization problem is feasible (i.e., there are β, β_0 satisfying the constraints) **if and only if** the data set is linearly separable. (We will later discuss how to define SVMs in the non-linearly separable case).

15.3 Duality for Convex Optimization Problems

Consider the following Primal problem (P):

$$\begin{aligned} \min_{z\in\mathbb{R}^s} \quad & f(z) \\ \text{s.t.} \quad & h_i(z) \leq 0 \quad \forall i = 1, ..., n \quad (\text{\# of constraints}) \end{aligned}$$

where $f : \mathbb{R}^s \to \mathbb{R}$ and $h_i : \mathbb{R}^s \to \mathbb{R}$ are differentiable functions. We will later assume that $f, h_1, ..., h_n$ are all convex functions.

Dual Formulation: First, we introduce the notion of the **Lagrangian**:

$$L(z, \lambda) := f(z) + \sum_{i=1}^{n} \lambda_i h_i(z)$$

where $z \in \mathbb{R}^s$ and $\lambda \in \mathbb{R}^n$ with $\lambda \geq 0$.

Dual Objective Function:

$$g(\lambda) := \min_{z\in\mathbb{R}^s} L(z, \lambda) \quad (\text{for a fixed } \lambda)$$

Dual Maximization Problem (D):

$$\begin{aligned} \max_{\lambda\in\mathbb{R}^n} \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda_i \geq 0 \quad \forall i = 1, ..., n \end{aligned}$$

Question: How are (P) and (D) related? Suppose $f, h_1, ..., h_n$ are differentiable and convex. Here is some intuition:

$$\max_{\lambda\geq 0} \min_{z\in\mathbb{R}^s} L(z, \lambda) \leq \min_{z\in\mathbb{R}^s} \max_{\lambda\geq 0} L(z, \lambda)$$

Suppose we can justify swapping max and min. Notice that:

$$\max_{\lambda \geq 0} \left\{ f(z) + \sum_{i=1}^{n} \lambda_i h_i(z) \right\} = \begin{cases} f(z) & \text{if } h_i(z) \leq 0 \quad \text{(Case 1)} \\ \infty & \text{if } h_i(z) > 0 \quad \text{(Case 2)} \end{cases}$$

So $\min_{z \in \mathbb{R}^s} \max_{\lambda \geq 0} L(z, \lambda)$ is exactly equivalent to solving the primal problem $\min_z f(z)$ subject to $h_i(z) \leq 0$.

15.4 Karush-Kuhn-Tucker (KKT) Theorem

Slater's Condition: Suppose there exists $\tilde{z} \in \mathbb{R}^s$ such that $h_i(\tilde{z}) < 0$ for all $i = 1, ..., n$.

Theorem 3 (KKT)**.** *Suppose $f, h_1, ..., h_n$ are convex and differentiable. Then, for every solution z^* of (P), there exists λ^* of (D) such that:*

1. $0 = \nabla f(z^*) + \sum_{i=1}^{n} \lambda_i^* \nabla h_i(z^*)$ ***(Stationarity Condition)***
2. $h_i(z^*) \leq 0 \quad \forall i = 1, ..., n$ ***(Primal Feasibility)***
3. $\lambda_i^* \geq 0 \quad \forall i = 1, ..., n$ ***(Dual Feasibility)***
4. $\lambda_i^* h_i(z^*) = 0 \quad \forall i = 1, ..., n$ ***(Complementary Slackness)***

Conversely, if (z^*, λ^*) satisfy conditions 1-4, then z^* is a solution of (P) and λ^* is a solution of (D).

15.5 Dual of SVM (Linearly Separable)

Primal problem:

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{\|\beta\|^2}{2} \\ \text{s.t.} \quad & 1 - Y_i(\langle X_i, \beta \rangle + \beta_0) \leq 0 \quad \forall i = 1, ..., n \end{aligned}$$

Here $Z = (\beta, \beta_0) \in \mathbb{R}^{d+1}$, $f(\beta, \beta_0) = \frac{\|\beta\|^2}{2}$, and $h_i(\beta, \beta_0) = 1 - Y_i(\langle X_i, \beta \rangle + \beta_0)$.

The Lagrangian is:

$$\begin{aligned} L((\beta, \beta_0), \lambda) &= \frac{\|\beta\|^2}{2} + \sum_{i=1}^{n} \lambda_i \left(1 - Y_i(\langle X_i, \beta \rangle + \beta_0)\right) \\ &= \frac{\|\beta\|^2}{2} + \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i Y_i \langle X_i, \beta \rangle - \left(\sum_{i=1}^{n} \lambda_i Y_i \right) \beta_0 \end{aligned}$$

The dual objective is $g(\lambda) := \min_{(\beta, \beta_0)} L((\beta, \beta_0), \lambda)$. We analyze this based on λ:
Case 1: If $\sum_{i=1}^{n} \lambda_i Y_i = 0$, then the β_0 term vanishes:

$$g(\lambda) = \min_{\beta} \left\{ \frac{\|\beta\|^2}{2} + \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i Y_i \langle X_i, \beta \rangle \right\}$$

Optimality condition for β (taking gradient and setting to 0):

$$0 = \beta - \sum_{i=1}^{n} \lambda_i Y_i X_i \implies \beta = \sum_{i=1}^{n} \lambda_i Y_i X_i$$

Substituting this optimal β back into $g(\lambda)$:

$$
\begin{aligned}
g(\lambda) &= \frac{1}{2}\left\|\sum_{i=1}^{n} \lambda_i Y_i X_i\right\|^2 + \sum_{i=1}^{n} \lambda_i - \left\langle \sum_{i=1}^{n} \lambda_i Y_i X_i, \sum_{j=1}^{n} \lambda_j Y_j X_j \right\rangle \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \langle X_i, X_j\rangle + \sum_{i=1}^{n} \lambda_i
\end{aligned}
$$

Case 2: If $\sum_{i=1}^{n} \lambda_i Y_i \neq 0$, then:

$$
g(\lambda) = \inf_{\beta, \beta_0} L((\beta, \beta_0), \lambda) = -\infty
$$

(because we can set β_0 to be arbitrarily large, positive or negative, to drive the term $-(\sum \lambda_i Y_i)\beta_0$ to $-\infty$).

In conclusion: The Dual SVM problem is:

$$
\begin{aligned}
\max_{\lambda \in \mathbb{R}^n} \quad & \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Y_i Y_j \langle X_i, X_j\rangle \\
\text{s.t.} \quad & \lambda_i \geq 0 \quad \forall i = 1, ..., n \\
\text{and} \quad & \sum_{i=1}^{n} \lambda_i Y_i = 0
\end{aligned}
$$

15.6 KKT Conditions for SVMs

From the Stationarity condition of KKT:

$$
0 = \nabla_\beta L = \beta^* - \sum_{i=1}^{n} \lambda_i^* Y_i X_i
$$

$$
\implies \beta^* = \sum_{i=1}^{n} \lambda_i^* Y_i X_i
$$

Once we have solved the dual SVM problem to find the optimal λ_i^*, the optimal weight vector β^* is determined by the above formula.

What about β_0^*? We will discuss this in the next lecture.

16 Lecture 16: SVM Dual, KKT Conditions, and Consequences (Part 2)

16.1 Recall: SVM Primal and Dual Problems

Primal SVM Problem: Given $(X_1, Y_1), ..., (X_n, Y_n)$ with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. Assume the dataset is linearly separable.

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{\|\beta\|^2}{2} \\ \text{s.t.} \quad & Y_i(\langle X_i, \beta \rangle + \beta_0) \geq 1 \quad \forall i = 1, ..., n \end{aligned}$$

where $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$.

Dual SVM Problem:

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^n} \quad & -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n Y_i Y_j \langle X_i, X_j \rangle \lambda_i \lambda_j + \sum_{i=1}^n \lambda_i \\ \text{s.t.} \quad & \lambda_i \geq 0 \quad \forall i = 1, ..., n \\ \text{and} \quad & \sum_{i=1}^n \lambda_i Y_i = 0 \end{aligned}$$

Here, $\lambda \in \mathbb{R}^n$ represents the Lagrange multipliers. Notice that the objective is concave in λ. We can rewrite the quadratic term as $-\frac{1}{2}\langle A\lambda, \lambda \rangle$, where $A \in \mathbb{R}^{n \times n}$ is given by:

$$A_{ij} = Y_i Y_j \langle X_i, X_j \rangle = \langle Y_i X_i, Y_j X_j \rangle$$

A is a **Gram matrix**, which implies that A is Positive Semi-Definite (PSD).

16.2 KKT Conditions for SVMs

Recall the general KKT conditions for a primal problem with parameters z and dual multipliers λ:

1. **Stationarity:** $0 = \nabla_z f(z^*) + \sum_{i=1}^n \lambda_i^* \nabla_z h_i(z^*)$
2. **Primal Feasibility:** $h_i(z^*) \leq 0 \quad \forall i = 1, ..., n$
3. **Dual Feasibility:** $\lambda_i^* \geq 0 \quad \forall i = 1, ..., n$
4. **Complementary Slackness:** $\lambda_i^* h_i(z^*) = 0 \quad \forall i = 1, ..., n$

Back to SVMs: Let $Z = (\beta, \beta_0)$. $f(\beta, \beta_0) = \frac{\|\beta\|^2}{2}$, and $h_i(\beta, \beta_0) = 1 - Y_i(\langle X_i, \beta \rangle + \beta_0)$.

From Stationarity:

$$\begin{aligned} 0 &= \nabla_\beta f(\beta^*, \beta_0^*) + \sum_{i=1}^n \lambda_i^* \nabla_\beta h_i(\beta^*, \beta_0^*) \\ &= \beta^* + \sum_{i=1}^n \lambda_i^* (-Y_i X_i) \\ \implies \beta^* &= \sum_{i=1}^n \lambda_i^* Y_i X_i \end{aligned}$$

This gives us the normal vector of our hyperplane in terms of the solution to the dual problem.

Question: How do we find β_0^* in terms of λ^*?
Answer: We use Complementary Slackness. Let i_0 be an index such that $\lambda_{i_0}^* > 0$ (non-zero). Then by complementary slackness:

$$\lambda_{i_0}^* (1 - Y_{i_0}(\langle X_{i_0}, \beta^* \rangle + \beta_0^*)) = 0$$

Since $\lambda_{i_0}^* > 0$, we must have:

$$1 - Y_{i_0}(\langle X_{i_0}, \beta^* \rangle + \beta_0^*) = 0 \implies 1 = Y_{i_0}(\langle X_{i_0}, \beta^* \rangle + \beta_0^*)$$

Multiplying both sides by Y_{i_0} (and noting $Y_{i_0}^2 = 1$):

$$Y_{i_0} = \langle X_{i_0}, \beta^* \rangle + \beta_0^* \implies \beta_0^* = Y_{i_0} - \langle X_{i_0}, \beta^* \rangle$$

Summary: We can use the solution to the dual problem to build the solution to the primal problem (P). **Question:** Can we always find such an i_0? **Answer:** Yes! As long as not all the Y_i are $+1$ (or all -1), there is always at least one i_0 such that $\lambda_{i_0}^* > 0$.

16.3 Some Consequences of Duality

16.3.1 Observation 1: Robustness

Definition 13 (Support Vector)**.** *A point X_i is said to be a **support vector** if it lies exactly on the margin boundary:*

$$Y_i(\langle X_i, \beta^* \rangle + \beta_0^*) = 1 \iff 1 - Y_i(\langle X_i, \beta^* \rangle + \beta_0^*) = 0$$

Suppose j is an index such that X_j is **not** a support vector. This means $1 - Y_j(\langle X_j, \beta^* \rangle + \beta_0^*) < 0$. Consider modifying the dataset by moving this non-support vector X_j to a new position $\tilde{X}_j$, such that it still satisfies:

$$1 - Y_j(\langle \tilde{X}_j, \beta^* \rangle + \beta_0^*) \leq 0$$

(The point remains outside the margin on the correct side). **Question:** What is the solution to the SVM for the new data set?
Answer: (β^*, β_0^*) from the original data set **continues to be optimal** for the new data set!

Proof Sketch using KKT: Step 1: Propose the candidate solution for the new problem: $\tilde{\beta} = \beta^*$, $\tilde{\beta}_0 = \beta_0^*$, and $\tilde{\lambda} = \lambda^*$. Step 2: Verify that this candidate satisfies the KKT conditions for the new problem. Condition 1, 2, and 3 naturally hold. Let's verify Condition 4 (Complementary Slackness): For $i \neq j$, $\tilde{\lambda}_i(1 - Y_i(\langle X_i, \tilde{\beta} \rangle + \tilde{\beta}_0)) = \lambda_i^*(1 - Y_i(\langle X_i, \beta^* \rangle + \beta_0^*)) = 0$. For the perturbed point j, note that since it was not a support vector originally, its corresponding dual variable $\lambda_j^* = 0$. Therefore, $\tilde{\lambda}_j = 0$, which immediately implies:

$$\tilde{\lambda}_j(1 - Y_j(\langle \tilde{X}_j, \tilde{\beta} \rangle + \tilde{\beta}_0)) = 0 \cdot (\ldots) = 0$$

So (β^*, β_0^*) continues to be the solution even after the perturbation!

16.3.2 Observation 2: Dependence on Inner Products

Recall our formulas for the optimal parameters:

$$\beta^* = \sum_{i=1}^{n} \lambda_i^* Y_i X_i$$

$$\beta_0^* = Y_{i_0} - \langle X_{i_0}, \beta^* \rangle = Y_{i_0} - \sum_{j=1}^{n} \lambda_j^* Y_j \langle X_{i_0}, X_j \rangle$$

The induced classifier for a new point x is:

$$f(x) = \text{sign}(\langle x, \beta^* \rangle + \beta_0^*)$$

$$\langle x, \beta^* \rangle + \beta_0^* = \sum_{i=1}^{n} \lambda_i^* Y_i \langle X_i, x \rangle + Y_{i_0} - \sum_{j=1}^{n} \lambda_j^* Y_j \langle X_{i_0}, X_j \rangle$$

Key Insight: The dependence on features is entirely through **inner products** ($\langle X_i, x \rangle$ and $\langle X_{i_0}, X_j \rangle$).

16.3.3 Observation 3: Dimensionality

- **Primal Problem:** $\min \frac{\|\beta\|^2}{2}$ s.t. constraints. The optimization variables are $\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}$. The dimension of the search space is $d + 1$.

- **Dual Problem:** $\max_{\lambda \in \mathbb{R}^n} -\frac{1}{2} \sum \sum \lambda_i \lambda_j Y_i Y_j \langle X_i, X_j \rangle + \sum \lambda_i$. The optimization variables are $\lambda \in \mathbb{R}^n$. The dimension of the search space is n (the number of data points).

In summary: We only need to look at the dual problem and inner products of features. If $d \gg n$ (e.g., infinite-dimensional feature space), solving the dual is much more efficient. This naturally leads to the Kernel Trick.

17 Lecture 17: Soft Margin SVMs and Kernels (Part 2)

17.1 Soft Margin SVMs

In contrast to hard margin SVMs, we introduce Soft Margin SVMs to handle non-linearly separable data. Given $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^d \times \{-1, 1\}$.

Primal Problem:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}, \xi_1, ..., \xi_n \in \mathbb{R}} \quad & \frac{\|\beta\|^2}{2} + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & Y_i(\langle X_i, \beta \rangle + \beta_0) \geq 1 - \xi_i \quad \forall i = 1, ..., n \\ & \xi_i \geq 0 \quad \forall i = 1, ..., n \end{aligned}$$

Remark: $C \in (0, \infty)$ is a parameter. **Note:** As $C \to \infty$, we recover the original (Hard margin) SVM problem.

Remark: How to choose the parameter C? **Cross-Validation.** We split the data into:

- **Training:** $(X_1, Y_1), ..., (X_{n/2}, Y_{n/2})$
- **Tuning:** $(X_{n/2+1}, Y_{n/2+1}), ..., (X_n, Y_n)$

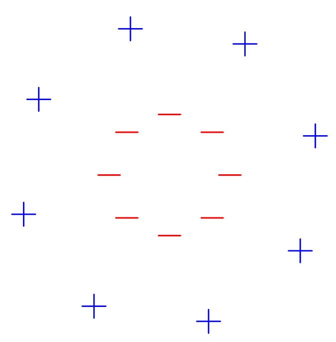
For each value of C, solve the Soft SVM problem for the training data to get $(\beta_C^*, \beta_{0C}^*)$ for different choices of parameter. Now compute the empirical error on the tuning set:

$$\ell(C) := \frac{1}{n/2} \sum_{j=1}^{n/2} \mathbf{1}_{f_{\beta_C^*, \beta_{0C}^*}(X_{n/2+j}) \neq Y_{n/2+j}}$$

Pick the C that makes the above the smallest. Soft margin SVMs are effective when the data set is close to being linearly separable. For C large...

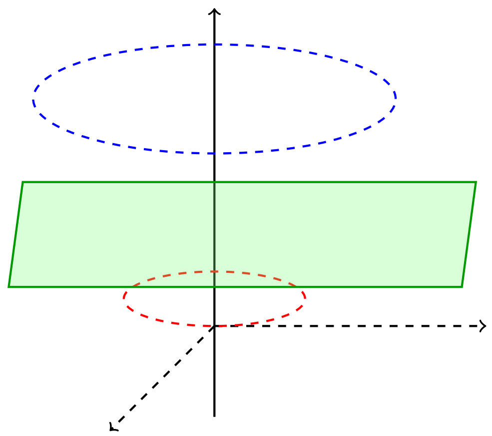
17.2 Non-linear Geometry and Data Embeddings

However, how to deal with a situation like this one:



Here the geometry of the data does not get along with linear classifiers. Original data set lives in $\mathcal{X} = \mathbb{R}^2$. Let us define an embedding Ψ:

$$\Psi : (x, z) \in \mathbb{R}^2 \mapsto (x, z, x^2 + z^2) \in \mathbb{R}^3$$



The mapped dataset $(\Psi(X_1), Y_1), ..., (\Psi(X_n), Y_n) \in \mathbb{R}^3 \times \{-1, 1\}$ is now linearly separable. In the space $\mathcal{H} = \mathbb{R}^3$ (the embedded space), we can now try to do SVMs:

(P) Primal Problem in $\mathcal{H}$:

$$\min_{\psi \in \mathbb{R}^3, \psi_0 \in \mathbb{R}} \quad \frac{\|\psi\|^2}{2}$$
$$\text{s.t.} \quad Y_i(\langle \Psi(X_i), \psi \rangle_{\mathcal{H}} + \psi_0) \geq 1 \quad \forall i = 1, ..., n$$

(D) Dual Problem in $\mathcal{H}$:

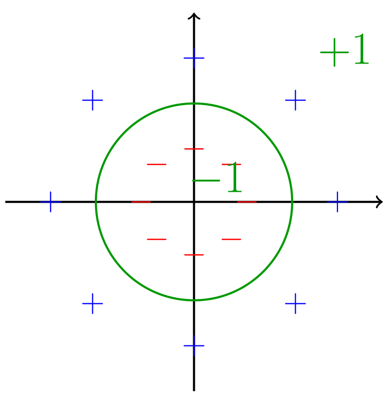
$$\max_{\lambda \in \mathbb{R}^n} \quad -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \langle \Psi(X_i), \Psi(X_j) \rangle_{\mathcal{H}} \lambda_i \lambda_j + \sum_{i=1}^{n} \lambda_i$$
$$\text{s.t.} \quad \sum_{i=1}^{n} Y_i \lambda_i = 0$$
$$\lambda_i \geq 0 \quad \forall i = 1, ..., n$$

I can now talk about two classifiers:

- **In $\mathcal{H}$:** let $z \in \mathcal{H}$.

$$f_{\psi^*, \psi_0^*}(z) = \begin{cases} 1 & \text{if } \langle \psi^*, z \rangle_{\mathcal{H}} + \psi_0^* \geq 0 \\ -1 & \text{else} \end{cases}$$

- **In $\mathbb{R}^2$:**



We define the classifier in the original space:

$$\ell_{\Psi}(x) := f_{\psi^*, \psi_0^*}(\Psi(x)) \quad \text{for } x \in \mathbb{R}^2$$

Note: $\ell_\Psi(x)$ is NOT a linear classifier on $\mathbb{R}^2$. It is the composition of a linear classifier in $\mathcal{H}$ and the embedding $\Psi : \mathbb{R}^2 \to \mathcal{H}$.

$$\ell_\Psi(x) = f_{\psi^*,\psi_0^*}(\Psi(x)) = \begin{cases} 1 & \text{if } \langle \Psi(x), \psi^* \rangle_{\mathcal{H}} + \psi_0^* \geq 0 \\ -1 & \text{else} \end{cases}$$

Recall from KKT conditions:

$$\psi^* = \sum_{i=1}^{n} \lambda_i^* Y_i \Psi(X_i)$$

$$\psi_0^* = Y_{i_0} - \sum_{j=1}^{n} \lambda_j^* Y_j \langle \Psi(X_j), \Psi(X_{i_0}) \rangle_{\mathcal{H}}$$

where i_0 is such that $\lambda_{i_0}^* > 0$. Moreover, λ^* solves the dual problem (D). Plugging ψ^* and ψ_0^* in the expression for $\ell_\Psi(x)$, we get:

$$\ell_\Psi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} \lambda_i^* Y_i \langle \Psi(X_i), \Psi(x) \rangle_{\mathcal{H}} + Y_{i_0} - \sum_{j=1}^{n} \lambda_j^* Y_j \langle \Psi(X_j), \Psi(X_{i_0}) \rangle_{\mathcal{H}} \geq 0 \\ -1 & \text{else} \end{cases}$$

The above motivates the following definition:

$$K(x, \tilde{x}) := \langle \Psi(x), \Psi(\tilde{x}) \rangle_{\mathcal{H}}$$

And defining:

$$\ell_K(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} \lambda_i^* Y_i K(X_i, x) + Y_{i_0} - \sum_{j=1}^{n} \lambda_j^* Y_j K(X_j, X_{i_0}) \geq 0 \\ -1 & \text{else} \end{cases}$$

where λ^* solves:

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^n} \quad & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j K(X_i, X_j) \lambda_i \lambda_j + \sum_{i=1}^{n} \lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} Y_i \lambda_i = 0 \\ & \lambda_i \geq 0 \quad \forall i = 1, ..., n \end{aligned}$$

Then we will define $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that it behaves like an inner product and use it to define classifiers.

17.3 Kernels

Let $\mathcal{X}$ be an arbitrary data space (for simplicity think of $\mathcal{X} = \mathbb{R}^d$). We say that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **Kernel** if the following conditions hold: $\forall m \in \mathbb{N}$, and every collection of points $x_1, ..., x_m \in \mathcal{X}$, the Gram matrix

$$A = \begin{pmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_m, x_1) & K(x_m, x_2) & \cdots & K(x_m, x_m) \end{pmatrix} \in \mathbb{R}^{m \times m}$$

is symmetric and Positive Semi-definite (PSD).

Recall: Let $A \in \mathbb{R}^{m \times m}$.

- **Symmetry:** $A_{ij} = A_{ji} \quad \forall i, j.$
- **PSD:** $\forall v \in \mathbb{R}^m$, $\langle Av, v\rangle_{\mathbb{R}^m} \geq 0$. (If A is symmetric, the above is equivalent to all eigenvalues of A being non-negative).

17.4 Properties of Kernels

1. Inner products define Kernels. $K(x, \tilde{x}) = \langle x, \tilde{x}\rangle_{\mathbb{R}^d}$. Let $m \in \mathbb{N}$ and let $x_1, ..., x_m \in \mathbb{R}^d$. Gram matrix: $A_{ij} = \langle X_i, X_j\rangle_{\mathbb{R}^d}$.

- *Symmetry:* $A_{ij} = \langle X_i, X_j\rangle = \langle X_j, X_i\rangle = A_{ji}$.
- *PSD:* Let $v \in \mathbb{R}^m$.

$$\begin{aligned}
\langle Av, v\rangle_{\mathbb{R}^m} &= \sum_{i=1}^{m}\sum_{j=1}^{m} A_{ij} v_i v_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \langle X_i, X_j\rangle_{\mathbb{R}^d} v_i v_j \\
&= \sum_{i=1}^{m} v_i \left\langle X_i, \sum_{j=1}^{m} v_j X_j \right\rangle_{\mathbb{R}^d} \\
&= \left\langle \sum_{i=1}^{m} v_i X_i, \sum_{j=1}^{m} v_j X_j \right\rangle_{\mathbb{R}^d} \\
&= \left\| \sum_{i=1}^{m} v_i X_i \right\|_{\mathbb{R}^d}^2 \geq 0
\end{aligned}$$

2. Positive linear combination of Kernels is a Kernel. Let K_1 and K_2 be kernels over $\mathcal{X}$. Then

$$\tilde{K}(x, \tilde{x}) = a_1 K_1(x, \tilde{x}) + a_2 K_2(x, \tilde{x})$$

for $a_1, a_2 \geq 0$. Then $\tilde{K}$ is also a kernel.

3. Scaling by a function. Let K be a kernel over $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}$ be any function. Then

$$\tilde{K}(x, \tilde{x}) = f(x) f(\tilde{x}) K(x, \tilde{x})$$

is a kernel.

4. Products of kernels are kernels. If K_1 and K_2 are kernels over $\mathcal{X}$, then

$$\tilde{K}(x, \tilde{x}) = K_1(x, \tilde{x}) K_2(x, \tilde{x})$$

is also a kernel.

5. Inner products composed with transformations are kernels. Let $\mathcal{H}$ be a Hilbert Space (think of it as $\mathbb{R}^S$). Let $\Psi : \mathcal{X} \to \mathcal{H}$. Then

$$K(x, \tilde{x}) = \langle \Psi(x), \Psi(\tilde{x})\rangle_{\mathcal{H}}$$

is a kernel over $\mathcal{X}$.

17.5 Examples of Kernels

1. **Polynomial Kernel:** $K(x, \tilde{x}) = (\langle x, \tilde{x} \rangle)^n$ is a kernel.

2. $\sum_{p=1}^{s} a_p(\langle x, \tilde{x} \rangle)^p$ is a kernel as long as $a_p \geq 0$.

3. $\sum_{p=0}^{s} \frac{1}{p!}(\langle x, \tilde{x} \rangle)^p$ is a kernel.

4. $\sum_{p=0}^{\infty} \frac{1}{p!}(\langle x, \tilde{x} \rangle)^p = \exp(\langle x, \tilde{x} \rangle)$ is a kernel.

5. **Gaussian (RBF) Kernel:** $K(x, \tilde{x}) = \exp\left(-\frac{\|x-\tilde{x}\|^2}{2}\right)$.

   Proof:

   $$\begin{aligned} \|x - \tilde{x}\|^2 &= \langle x - \tilde{x}, x - \tilde{x} \rangle \\ &= \|x\|^2 - 2\langle x, \tilde{x} \rangle + \|\tilde{x}\|^2 \end{aligned}$$

   Therefore:

   $$\begin{aligned} K(x, \tilde{x}) &= \exp\left(\langle x, \tilde{x} \rangle - \frac{\|x\|^2}{2} - \frac{\|\tilde{x}\|^2}{2}\right) \\ &= \exp\left(-\frac{\|x\|^2}{2}\right) \exp\left(-\frac{\|\tilde{x}\|^2}{2}\right) \exp(\langle x, \tilde{x} \rangle) \end{aligned}$$

   Letting $f(x) = \exp(-\|x\|^2/2)$ and $f(\tilde{x}) = \exp(-\|\tilde{x}\|^2/2)$, this matches Property 3 where the base kernel is $\exp(\langle x, \tilde{x} \rangle)$ from Example 4. Thus, it is a valid kernel.

18 Lecture 18: Hilbert Spaces and RKHS

18.1 Overview

- Definition of a Hilbert space. Some properties and some examples.
- RKHS (Reproducing Kernel Hilbert spaces).
- Data embeddings, Kernels, RKHS.

We clarify the way in which Kernels are related to data embeddings:

- **Kernels:** $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $(x, \tilde{x}) \mapsto K(x, \tilde{x})$
- **Embeddings:** $\Psi : \mathcal{X} \to \mathcal{H}$ (data space into $\mathcal{H}$)

18.2 Hilbert Spaces

Definition 14. *A **Hilbert space** $\mathcal{H}$ is a vector space together with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is a complete metric space.*

Properties:

1. If we have elements $v, w \in \mathcal{H}$, we can add them: $v + w \in \mathcal{H}$, and we can multiply by scalars $\alpha \in \mathbb{R}$: $\alpha v \in \mathcal{H}$.
2. $\langle u, v \rangle_{\mathcal{H}} \in \mathbb{R}$ and we have the following properties:
   - **Symmetry:** $\langle u, v \rangle_{\mathcal{H}} = \langle v, u \rangle_{\mathcal{H}}$
   - **Bi-linearity:** $\langle \alpha_1 u_1 + \alpha_2 u_2, v \rangle_{\mathcal{H}} = \alpha_1 \langle u_1, v \rangle_{\mathcal{H}} + \alpha_2 \langle u_2, v \rangle_{\mathcal{H}}$
   - **Positivity:** $\langle u, u \rangle_{\mathcal{H}} \geq 0 \quad \forall u \in \mathcal{H}$ and $\langle u, u \rangle_{\mathcal{H}} = 0 \iff u = 0$ (Zero element in $\mathcal{H}$).
3. Because of the positivity property, we can define:

   $$\|u\|_{\mathcal{H}} := \sqrt{\langle u, u \rangle_{\mathcal{H}}}$$

   This turns out to be a norm. In particular,

   $$d(u, v) = \|u - v\|_{\mathcal{H}}$$

   is a distance function.

An important inequality in Hilbert spaces is the so-called **Cauchy-Schwarz inequality**:

$$|\langle u, v \rangle_{\mathcal{H}}| \leq \|u\|_{\mathcal{H}} \cdot \|v\|_{\mathcal{H}}$$

Examples of Hilbert spaces:

1. $(\mathbb{R}^m, \langle \cdot, \cdot \rangle_{\mathbb{R}^m})$

$$u, v \in \mathbb{R}^m, \quad u = (u^1, ..., u^m), \quad v = (v^1, ..., v^m)$$

$$\langle u, v \rangle_{\mathbb{R}^m} = \sum_{i=1}^{m} u^i v^i$$

2. $(\ell^2(\mathbb{N}), \langle \cdot, \cdot \rangle_{\ell^2(\mathbb{N})})$ $u \in \ell^2(\mathbb{N})$ is a sequence $u = (u^1, u^2, u^3, ...)$ satisfying $\sum_{i=1}^{\infty} (u^i)^2 < \infty$. Given $u, v \in \mathcal{H}$:

$$\langle u, v \rangle_{\mathcal{H}} := \sum_{i=1}^{\infty} u^i v^i$$

3. $L^2(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} \text{ measurable and } \int_{-\infty}^{\infty} (f(x))^2 dx < \infty \right\}$

$$\langle f, g \rangle_{L^2(\mathbb{R})} := \int_{-\infty}^{\infty} f(x) g(x) dx$$

Note: Actually, the elements of $L^2(\mathbb{R})$ are equivalence classes of measurable functions. We won't discuss details, but keep in mind this definition needs adjustment.

18.3 Reproducing Kernel Hilbert Spaces (RKHS)

These are the spaces we will be interested in. They relate Kernels and Embeddings in a very nice way. Need some definitions first...

Definition 15. *Let $\mathcal{H}$ be a Hilbert space. A linear function $L : \mathcal{H} \to \mathbb{R}$ is continuous if there is a constant $C > 0$ s.t.*

$$|L(f)| \leq C \|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$

Example: Fix $g \in \mathcal{H}$ and define $L_g(f) := \langle f, g \rangle_{\mathcal{H}}$. Note L_g is linear (because of bi-linearity of inner product). Moreover, using Cauchy-Schwarz:

$$|L_g(f)| = |\langle f, g \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \underbrace{(\|g\|_{\mathcal{H}})}_{:=C} \quad \forall f \in \mathcal{H}$$

It turns out all linear continuous functions L on a Hilbert space are of the form above...

Theorem 4 (Riesz Representation Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space. For any $L : \mathcal{H} \to \mathbb{R}$ linear and continuous, there exists a unique $g \in \mathcal{H}$ s.t.*

$$L(f) = L_g(f) = \langle f, g \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$

Definition 16 (RKHS)**.** *Let $\mathcal{X}$ be a set. Let $\mathcal{H}$ be a Hilbert space of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ (i.e., elements in $\mathcal{H}$ are functions from $\mathcal{X}$ into $\mathbb{R}$). $\mathcal{H}$ is said to be a **Reproducing Kernel Hilbert Space (RKHS)** if for every $x \in \mathcal{X}$ the map (evaluation map):*

$$L_x : f \in \mathcal{H} \mapsto f(x)$$

is a continuous map.

In particular:

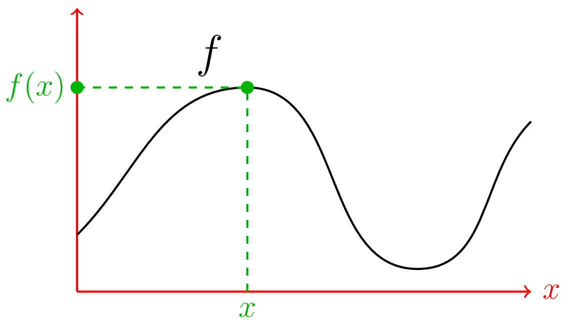
$$|f(x) - \tilde{f}(x)| \leq C_x \|f - \tilde{f}\|_{\mathcal{H}} \quad \forall f, \tilde{f}$$

Remark: L_x is linear.

$$L_x(\alpha f + \beta g) = \alpha f(x) + \beta g(x) = \alpha L_x(f) + \beta L_x(g)$$

Now, if L_x is continuous, by Riesz Representation Theorem, we have: $\exists! K_x \in \mathcal{H}$ s.t.

$$f(x) = L_x(f) = \langle K_x, f \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$



That is, to evaluate $f \in \mathcal{H}$ at x, we take the inner product between f and K_x. Now, let $\tilde{x} \in \mathcal{X}$ be another element in $\mathcal{X}$. Let us consider its associated $K_{\tilde{x}}$. We have:

$$f(x) = \langle K_x, f \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$
$$f(\tilde{x}) = \langle K_{\tilde{x}}, f \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$

So what if we take $f = K_{\tilde{x}}$ in the first equation? We obtain:

$$K_{\tilde{x}}(x) = \langle K_x, K_{\tilde{x}} \rangle_{\mathcal{H}}$$

This motivates defining the function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$K(x, \tilde{x}) := \langle K_x, K_{\tilde{x}} \rangle_{\mathcal{H}}$$

We can also define $\Psi : \mathcal{X} \to \mathcal{H}$ mapping $x \mapsto K_x$. Then $\tilde{K}(x, \tilde{x}) = \langle \Psi(x), \Psi(\tilde{x}) \rangle_{\mathcal{H}}$. K is thus a Kernel over $\mathcal{X}$!

So far: RKHS $\implies$ Kernel K. The Kernel has the property:

$$\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H}$$

"It reproduces elements in $\mathcal{H}$" (hence the name).

18.4 Moore-Aronszajn Theorem

What about the other way around? Kernel $\implies$ RKHS? Yes!

Theorem 5 (Moore-Aronszajn theorem)**.** *Defines:*

$$\mathcal{H} = \left\{ \sum_{i=1}^{\infty} a_i K(\cdot, x_i) \text{ for some collection } x_i \in \mathcal{X} \text{ and } a_i \in \mathbb{R} \text{ with } \sum_{i,j} a_i a_j K(x_i, x_j) < \infty \right\}$$

Then define:

$$\left\langle \sum_{i=1}^{\infty} a_i K(\cdot, x_i), \sum_{j=1}^{\infty} b_j K(\cdot, \tilde{x}_j) \right\rangle_{\mathcal{H}} := \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_i b_j K(x_i, \tilde{x}_j)$$

Think of this as $a^T K b$ where

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \end{pmatrix}, \quad K = \begin{pmatrix} K(x_1, \tilde{x}_1) & K(x_1, \tilde{x}_2) & \cdots \\ K(x_2, \tilde{x}_1) & K(x_2, \tilde{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

It can be shown that $\mathcal{H}$ with this inner product is a Hilbert space. Moreover, K is a reproducing Kernel: Let $f = \sum_{i=1}^{\infty} a_i K(\cdot, x_i)$. Then,

$$\begin{aligned}
\langle f, K(\cdot, x)\rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{\infty} a_i K(\cdot, x_i), K(\cdot, x) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\infty} a_i \langle K(\cdot, x_i), K(\cdot, x)\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\infty} a_i K(x, x_i) = f(x)
\end{aligned}$$

So: RKHS $\iff$ Kernels.

Next, we will study minimization problems of the form:

$$\min_{f \in \mathcal{H}} \underbrace{\sum_{i=1}^{n} \ell_i(f(x_i))}_{\text{Empirical Risk}} + \underbrace{C\|f\|_{\mathcal{H}}^2}_{\text{Regularizer}}$$

where $x_1, ..., x_n$ are data points and $C > 0$ is a parameter. We'll show, for example, that this type of problem induces a great variety of methods for classification and regression. In the next lecture we will give some examples and prove a fundamental result about solutions to the above optimization problem (Representer Theorem).

19 Lecture 19: Examples of RKHS and Representer Theorem (Part 2)

19.1 Examples of RKHS

Let us first consider examples of RKHS.

Example 1: Let $\mathcal{X} = \mathbb{R}^d$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be given by:

$$K(x, \tilde{x}) = \exp(-\|x - \tilde{x}\|^2)$$

Since K is a kernel, we can use the Moore-Aronszajn theorem to conclude that

$$\mathcal{H} = \left\{ \sum_{n=1}^{\infty} a_n \exp(-\| \cdot - x_n\|^2) \text{ s.t. } \sum_{n=1}^{\infty}\sum_{l=1}^{\infty} a_n a_l \exp(-\|x_l - x_n\|^2) < \infty \right\}$$

together with the inner product

$$\left\langle \sum_{n=1}^{\infty} a_n K(\cdot, x_n), \sum_{l=1}^{\infty} b_l K(\cdot, \tilde{x}_l) \right\rangle_{\mathcal{H}} = \sum_{n=1}^{\infty}\sum_{l=1}^{\infty} a_n b_l K(x_n, \tilde{x}_l)$$

is an RKHS over $\mathcal{X} = \mathbb{R}^d$.

Note, for example, that $K(\cdot, x) = \exp(-\| \cdot - x\|^2)$ belongs to $\mathcal{H}$. Note, also, that

$$\exp(-\| \cdot - x_1\|^2) + 2\exp(-\| \cdot - x_2\|^2)$$

also belongs to $\mathcal{H}$.

Example 2: Let us now give an example where the kernel is not given explicitly. Let $\mathcal{X} = [0, 1]$ (unit interval) in $\mathbb{R}$. Consider the following set:

$$\mathcal{H} = \left\{ f : [0, 1] \to \mathbb{R} \text{ s.t. } f(x) = \int_0^x f'(t)dt \quad \forall x \in [0, 1] \text{ and } \int_0^1 |f'(t)|^2 dt < \infty \right\}$$

$$\langle f, g\rangle_{\mathcal{H}} := \int_0^1 f'(t)g'(t)dt$$

Is $\mathcal{H}$ an RKHS? If so, what is its reproducing kernel?

Note: For all $f \in \mathcal{H}$ we have

$$\begin{aligned}
|f(x)| &= \left| \int_0^x f'(t)dt \right| \\
&\leq \int_0^x |f'(t)|dt \\
&\leq \int_0^1 |f'(t)|dt \quad \text{(Jensen's inequality)} \\
&\leq \left(\int_0^1 1^2 dt \right)^{1/2} \left(\int_0^1 |f'(t)|^2 dt \right)^{1/2} \\
&= \|f\|_{\mathcal{H}}
\end{aligned}$$

Hence, L_x (the evaluation map) is continuous. Since this is true for every $x \in \mathcal{X} = [0, 1]$, we conclude that $\mathcal{H}$ is indeed an RKHS.

Next, we need to obtain its reproducing kernel. (Which we know exists following abstract theory). By definition,

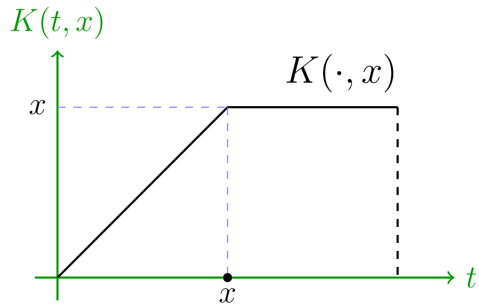
$$\begin{aligned} f(x) &= \int_0^x f'(t)dt \\ &= \int_0^1 f'(t)\mathbf{1}_{[0,x]}(t)dt \end{aligned}$$

If $\mathbf{1}_{[0,x]}(t) = \frac{d}{dt}K(t,x)$, then the above can be written as:

$$= \langle f, K(\cdot, x)\rangle_{\mathcal{H}}$$

and we would have found our reproducing kernel. Hence, we just need to integrate:

$$K(t,x) := \int_0^t \mathbf{1}_{[0,x]}(s)ds = \min\{x,t\}$$



19.2 Empirical Risk Minimization using Kernels

Let $\mathcal{X}$ be our data space and suppose $x_1, ..., x_n \in \mathcal{X}$ are our observations (features). For a given kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we will study the problem:

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{n} \ell_i(f(x_i)) + C\|f\|_{\mathcal{H}}^2 \tag{*}$$

Here:

- $\mathcal{H}$ is the RKHS associated to K (Moore-Aronszajn theorem).
- $\ell_i : \mathbb{R} \to \mathbb{R}$ are suitable loss functions (we'll give examples soon).
- $C > 0$ is a hyperparameter (like in Soft SVMs).
- Again, $x_1, ..., x_n$ are our observations.

Remark 1: Problem (*) is like Ridge Regression! (See HW 4) Suppose $\ell_i(t) = (t - y_i)^2$. Recall that $y_i \in \mathbb{R}$ is some label.

$$f(x_i) = \langle f, K(\cdot, x_i)\rangle_{\mathcal{H}}$$

Also, consider the canonical data embedding:

$$\Psi : x \mapsto K(\cdot, x)$$

So that we can write $f(x_i) = \langle f, \Psi(x_i) \rangle_{\mathcal{H}}$. The objective function can thus be written as:

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{n} (y_i - \langle \Psi(x_i), f \rangle_{\mathcal{H}})^2 + C\|f\|_{\mathcal{H}}^2$$

This is Ridge Regression in $\mathcal{H}$ for the data set $(\Psi(x_1), y_1), ..., (\Psi(x_n), y_n)$!

Remark 2: Problem (*) is a regularized version of ERM.

$$\min_{f \in \mathcal{H}} \quad \underbrace{\sum_{i=1}^{n} \ell_i(f(x_i))}_{\text{Empirical Risk}} + \underbrace{C\|f\|_{\mathcal{H}}^2}_{\text{Regularizer}}$$

where $\mathcal{H}$ is the class of functions over which we are minimizing.

19.3 Representer Theorem

Next, we discuss the structure of solutions to problem (*). The following is a fundamental result about RKHS.

Theorem 6 (Representer Theorem)**.** *Let K be a kernel over $\mathcal{X}$ and let $\mathcal{H}$ be its corresponding RKHS. Let $\ell_i : \mathbb{R} \to \mathbb{R}$ be arbitrary loss functions and let $C > 0$. Then all solutions to*

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{n} \ell_i(f(x_i)) + C\|f\|_{\mathcal{H}}^2$$

have the form:

$$f^*(\cdot) = \sum_{i=1}^{n} a_i K(\cdot, x_i)$$

(Note: Same data points appearing in the formulation of the problem!)

Before showing the proof, let us understand the consequences of this result. First of all, note that (*) is an **infinite dimensional** optimization problem (this should sound intimidating). However, the Representer theorem implies that it is enough to solve an **n-dimensional optimization problem**! Indeed, since we know solutions have the form $\sum_{i=1}^{n} a_i K(\cdot, x_i)$, it remains to find the coefficients $a_1, ..., a_n$. How do we find them? By solving the following n-dimensional problem:

$$\min_{\vec{a}=(a_1,...,a_n) \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \ell_i \left(\sum_{j=1}^{n} a_j K(x_i, x_j) \right) + C \left\| \sum_{j=1}^{n} a_j K(\cdot, x_j) \right\|_{\mathcal{H}}^2$$

We can actually express

$$\left\| \sum_{j=1}^{n} a_j K(\cdot, x_j) \right\|_{\mathcal{H}}^2 = \sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i K(x_i, x_j)$$

The bottom line is that the coefficients $a_1, ..., a_n$ are obtained by solving:

$$\min_{\vec{a}=(a_1,...,a_n) \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \ell_i \left(\sum_{j=1}^{n} a_j K(x_i, x_j) \right) + C \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j)$$

19.4 Example: Kernel Ridge Regression

Suppose $\ell_i(t) = (t - y_i)^2$, where $y_i \in \mathbb{R}$ (squared loss). This loss naturally arises in Regression problems. Problem (*) becomes:

$$\min_{a_1,\ldots,a_n} \quad \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_j K(x_i, x_j) - y_i \right)^2 + C \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j)$$

It will be convenient to introduce matrix notation:

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}$$

The above objective function can be written as:

$$\min_{\vec{a}} \quad \|\mathbf{K}\vec{a} - \vec{y}\|^2_{\mathbb{R}^n} + C\langle \mathbf{K}\vec{a}, \vec{a} \rangle_{\mathbb{R}^n}$$

This is a convex optimization problem so that considering the gradient in $\vec{a}$ of the objective and setting it to be 0 will give us the solution to the problem:

$$\begin{aligned} \vec{0} &= \nabla_{\vec{a}}(\|\mathbf{K}\vec{a} - \vec{y}\|^2 + C\langle \mathbf{K}\vec{a}, \vec{a} \rangle) \\ &= 2\mathbf{K}^T(\mathbf{K}\vec{a} - \vec{y}) + 2C\mathbf{K}\vec{a} \end{aligned}$$

(Since $\mathbf{K}$ is symmetric, $\mathbf{K} = \mathbf{K}^T$). Hence,

$$\mathbf{K}\vec{y} = (\mathbf{K}^2 + C\mathbf{K})\vec{a}$$

So, provided $\mathbf{K}$ is invertible, we can write:

$$\vec{a} = (\mathbf{K}^2 + C\mathbf{K})^{-1}\mathbf{K}\vec{y}$$

Back to $\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + C\|f\|^2_{\mathcal{H}}$, we conclude that

$$f^*(\cdot) = \sum_{i=1}^{n} a_i^* K(\cdot, x_i)$$

where $\vec{a}^* = (a_1^*, ..., a_n^*)^T = (\mathbf{K}^2 + C\mathbf{K})^{-1}\mathbf{K}\vec{y}$.

20 Lecture 20: Representer Theorem Proof and Kernel Applications

20.1 Proof of the Representer Theorem

Let us first present the proof of the Representer theorem. Let $x_1, ..., x_n \in \mathcal{X}$ be fixed. Define the subspace S:

$$S = \left\{ \sum_{i=1}^{n} a_i K(\cdot, x_i) : a_1, ..., a_n \in \mathbb{R} \right\}$$

(Another way of writing S is $S = \text{Span}\{K(\cdot, x_1), ..., K(\cdot, x_n)\}$, i.e., S is the set of linear combinations of $K(\cdot, x_1), ..., K(\cdot, x_n)$).

For a given $f \in \mathcal{H}$, let $\text{Proj}_S(f)$ be its projection onto S. $\text{Proj}_S(f) \in S$ and has the property that:

$$f - \text{Proj}_S(f) \perp S$$

(it is perpendicular to S). That means:

$$\langle f - \text{Proj}_S(f), g \rangle_{\mathcal{H}} = 0 \quad \forall g \in S$$

Let us denote $f_S := \text{Proj}_S(f)$.

In particular, for all $i = 1, ..., n$ we have $K(\cdot, x_i) \in S$, so:

$$\langle f - f_S, K(\cdot, x_i) \rangle_{\mathcal{H}} = 0$$

Hence, for all $i = 1, ..., n$ we have:

$$\langle f, K(\cdot, x_i) \rangle_{\mathcal{H}} = \langle f_S, K(\cdot, x_i) \rangle_{\mathcal{H}}$$

By the reproducing property, this means:

$$f(x_i) = f_S(x_i)$$

Thus, the empirical risk is exactly the same:

$$\sum_{i=1}^{n} \ell_i(f(x_i)) = \sum_{i=1}^{n} \ell_i(f_S(x_i))$$

On the other hand, by the Pythagorean theorem:

$$\|f_S\|_{\mathcal{H}}^2 + \|f - f_S\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{H}}^2$$

Since $\|f - f_S\|_{\mathcal{H}}^2 > 0$ unless $f \in S$, we have:

$$\|f_S\|_{\mathcal{H}}^2 < \|f\|_{\mathcal{H}}^2 \quad \text{(strict inequality unless } f \in S)$$

Hence, if $f \notin S$, we have:

$$\sum_{i=1}^{n} \ell_i(f_S(x_i)) + C\|f_S\|_{\mathcal{H}}^2 < \sum_{i=1}^{n} \ell_i(f(x_i)) + C\|f\|_{\mathcal{H}}^2$$

Therefore, no $f \notin S$ can be a minimizer of the problem. Thus, minimizers must belong to S. $\square$

20.2 Examples of Regularized ERM with Kernels

Let us consider more examples of problems of the form:

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{n} \ell_i(f(x_i)) + C\|f\|_{\mathcal{H}}^2$$

20.2.1 1. Quantile Regression with Kernels

Suppose $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$. For a given $\tau \in (0, 1)$, consider:

$$\ell_i(s) = \ell_\tau(y_i - s)$$

where

$$\ell_\tau(u) := \begin{cases} (\tau - 1)u & \text{if } u < 0 \\ \tau u & \text{if } u \geq 0 \end{cases}$$

Recall that the optimization $\min_f \mathbb{E}_{(X,Y)\sim\rho}[\ell_\tau(Y - f(X))]$ over all $f : \mathcal{X} \to \mathbb{R}$ is solved by $f = q_\tau$, where $q_\tau(x)$ is the τ-th quantile of the conditional distribution of Y given $X = x$. (Area under the curve $= \tau$). Thus,

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{n} \ell_\tau(y_i - f(x_i)) + C\|f\|_{\mathcal{H}}^2$$

is a regularized ERM problem to approximate q_τ from finite data.

20.2.2 2. Logistic Regression with Kernels

Next, we consider examples of loss functions useful in classification problems. Suppose $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ (setup for binary classification). Consider:

$$\ell_i(s) := -\log\left(\frac{\exp(y_i \cdot s)}{1 + \exp(y_i \cdot s)}\right)$$

For example, suppose $y_i = 1$, then $\ell_i(f(x_i))$ is small when the sign of $f(x_i)$ is $+1$ (in other words, when it is consistent with y_i). Conversely, if $y_i = -1$, it penalizes positive outputs.

When we consider a solution f^* of the regularized ERM, we will get $f^* : \mathcal{X} \to \mathbb{R}$ that doesn't necessarily output in the set $\{-1, 1\}$. How can we obtain a classifier from f^*?

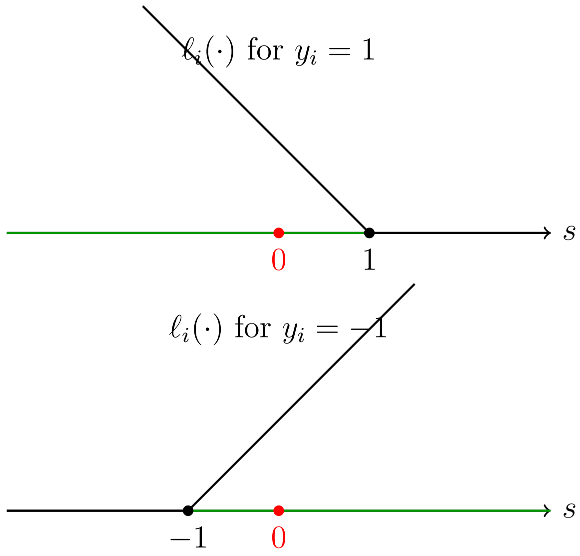
- **Hard classifier:** $C_f(x) = \begin{cases} 1 & \text{if } f^*(x) \geq 0 \\ -1 & \text{else} \end{cases}$
- **Soft classifier:** $C_f(x)$ outputs 1 with probability $\frac{\exp(f^*(x))}{1+\exp(f^*(x))}$, and outputs -1 with probability $\frac{1}{1+\exp(f^*(x))}$.

20.2.3 3. Hinge Loss with Kernels (SVMs)

Again in the setting of binary classification: $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$. Consider the loss function:

$$\ell_i(s) = \max\{0, 1 - sy_i\}$$

If $y_i = 1$, then $\ell_i(s) = \max\{0, 1 - s\}$. If $y_i = -1$, then $\ell_i(s) = \max\{0, 1 + s\}$.



Remark: You'll show that the optimization problem with $\ell_i(\cdot)$ being the Hinge loss gives rise to Soft Margin SVMs.

20.3 More Examples of Kernels and Bochner's Theorem

For this part we'll assume $\mathcal{X} = \mathbb{R}^d$.

Definition 17. *We say a Kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is **stationary** or **homogeneous** if it has the form:*

$$K(x, \tilde{x}) = h(x - \tilde{x})$$

for some function $h : \mathbb{R}^d \to \mathbb{R}$.

Example: Gaussian Kernel is stationary: $K(x, \tilde{x}) = \exp(-\|x - \tilde{x}\|^2/2) = h(x - \tilde{x})$ where $h(u) = \exp(-\|u\|^2/2)$.

Theorem 7 (Bochner's Theorem)**.** *Let Z be a random vector in $\mathbb{R}^d$. Let $i = \sqrt{-1}$. Let $\phi_Z(v) := \mathbb{E}[\exp(i\langle v, Z\rangle)]$ for $v \in \mathbb{R}^d$ be Z's characteristic function. If $\phi_Z(\cdot)$ is a real-valued function, then:*

$$K(x, \tilde{x}) := \phi_Z(x - \tilde{x})$$

is a Kernel.

Conversely, if $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a stationary/homogeneous Kernel (i.e., $K(x, \tilde{x}) = h(x - \tilde{x})$), then there is a constant $A > 0$ and a random vector $Z \in \mathbb{R}^d$ such that:

$$K(x, \tilde{x}) = A\phi_Z(x - \tilde{x}) \quad \forall x, \tilde{x} \in \mathbb{R}^d$$

The previous result can be used to obtain new examples of Kernels in $\mathbb{R}^d$. However, we just need to say what random vectors Z have real-valued characteristic functions.
Proposition: Z's characteristic function is real-valued if and only if Z and $-Z$ have the same distribution (symmetric).

This theorem allows us to provide more examples of Kernels:

1. **Rational Quadratic family:** $K(x, \tilde{x}) = h(x - \tilde{x})$ where

$$h(u) = \left(1 + \frac{\|u\|^2}{\alpha \ell^2}\right)^{-\alpha}$$

   where $\alpha > 0, \ell > 0$ (hyperparameters). **Note:** As $\alpha \to \infty$, we recover $\exp(-\|u\|^2/\ell^2)$.

2. **Exponential family:** $K(x, \tilde{x}) = h(x - \tilde{x})$ where

$$h(u) = \exp\left(-\frac{\|u\|^\gamma}{\ell^\gamma}\right), \quad 0 < \gamma \leq 2$$

   In particular, when $\gamma = 1$, $h(u) = \exp(-\|u\|/\ell)$ (Exponential). **Q:** What random variable has this as characteristic function? (Hint: Cauchy for $\gamma = 1$).

3. **Matérn class:** (See Chapters 2 & 3 in the book "Gaussian Processes for Machine Learning" by Rasmussen and Williams).

21 Lecture 21: Gaussian Processes and Mercer's Theorem

21.1 Gaussian Processes

Definition 18 (Gaussian Process)**.** *Let $\mathcal{X}$ be a set (data space). A collection of random variables $\{Z_x\}_{x\in\mathcal{X}}$ (a family of real-valued random variables indexed by $\mathcal{X}$) is said to be a **Gaussian process** with mean zero if for any finite collection $x_1, ..., x_s \in \mathcal{X}$, the random vector:*

$$(Z_{x_1}, ..., Z_{x_s})$$

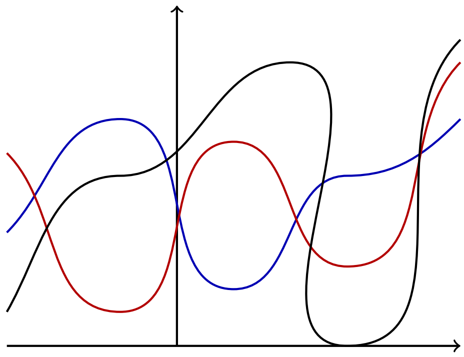
is Gaussian with mean zero and some covariance matrix.

Definition 19 (Covariance function)**.** *Let $\{Z_x\}_{x\in\mathcal{X}}$ be a Gaussian process over $\mathcal{X}$. The covariance function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the Gaussian process is given by:*

$$K(x, \tilde{x}) := Cov(Z_x, Z_{\tilde{x}})$$

Remark: If $\{Z_x\}$ is a Gaussian process, then $(Z_{x_1}, ..., Z_{x_s}) \sim \mathcal{N}_s(0, \Sigma)$ where $\Sigma_{ij} = K(x_i, x_j)$.

Q: What is the interpretation of a sample from a Gaussian process? **A:** A random function over $\mathcal{X}$.



(Qualitative Observation: How rough or irregular these samples look like depends on the covariance function of the process).

21.2 Connection between Covariance functions and Kernels

Proposition: If $\{Z_x\}_{x\in\mathcal{X}}$ is a Gaussian process over $\mathcal{X}$, then its covariance function K is a Kernel. *Proof:* Let $x_1, ..., x_s \in \mathcal{X}$. We consider:

$$A = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_s) \\ \vdots & \ddots & \vdots \\ K(x_s, x_1) & \cdots & K(x_s, x_s) \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(Z_{x_1}, Z_{x_1}) & \cdots & \mathrm{Cov}(Z_{x_1}, Z_{x_s}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(Z_{x_s}, Z_{x_1}) & \cdots & \mathrm{Cov}(Z_{x_s}, Z_{x_s}) \end{pmatrix}$$

A is the covariance matrix of $(Z_{x_1}, ..., Z_{x_s})$. Covariance matrices are always symmetric and PSD. Thus K is a kernel. $\square$

Q1: Given a Kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (potentially satisfying some additional properties), can we construct a Gaussian process $\{Z_x\}_{x\in\mathcal{X}}$ whose covariance function coincides

with K?
Q2: If so, how can we produce samples from such a Gaussian process?

To answer these questions, we'll focus on the setting $\mathcal{X} = \mathbb{R}^d$. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Kernel that is continuous and bounded (i.e., $|K(x, \tilde{x})| \leq C \quad \forall x, \tilde{x} \in \mathbb{R}^d$). Under these assumptions, we will discuss Mercer's theorem which generalizes the Spectral theorem for PSD matrices.

List of fundamental results about Kernels:

- **Moore-Aronszajn Theorem:** (RKHS and Kernels)
- **Representer Theorem:** (Solutions to certain optim problems in RKHS)
- **Bochner's Theorem:** (Characterization of Homogeneous Kernels in $\mathbb{R}^d$)
- **Mercer's Theorem:** (Connection between Kernels and Gaussian processes)

21.3 Mercer's Theorem

To state Mercer's Theorem, we need to introduce some additional definitions. Let $p : \mathbb{R}^d \to \mathbb{R}$ be a strictly positive function (chosen by convenience, think of it as a density). Relative to this p, we define $T : L^2(p) \to L^2(p)$ via convolution:

$$Tf(x) := \int_{\mathbb{R}^d} K(x, \tilde{x}) f(\tilde{x}) p(\tilde{x}) d\tilde{x}$$

Note:

$$\begin{aligned}
\|Tf\|^2_{L^2(p)} &= \int_{\mathbb{R}^d} (Tf(x))^2 p(x) dx \\
&= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} K(x, \tilde{x}) f(\tilde{x}) p(\tilde{x}) d\tilde{x} \right)^2 p(x) dx \\
&\leq \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} K(x, \tilde{x})^2 p(\tilde{x}) d\tilde{x} \right) \left(\int_{\mathbb{R}^d} f(\tilde{x})^2 p(\tilde{x}) d\tilde{x} \right) p(x) dx
\end{aligned}$$

Since K is bounded and PSD, $K(x, \tilde{x})^2 \leq K(x, x) K(\tilde{x}, \tilde{x}) \leq C^2$. So:

$$\leq \|f\|^2_{L^2(p)} C^2 \int_{\mathbb{R}^d} p(x) dx \implies \|Tf\|^2_{L^2(p)} \leq C^2 \|f\|^2_{L^2(p)}$$

Observation: Notice that $Tf(x) = \int K(x, \tilde{x}) f(\tilde{x}) p(\tilde{x}) d\tilde{x}$ is analogous to $(Av)_i = \sum_{j=1}^s A_{ij} v_j$ where $A \in \mathbb{R}^{s \times s}$. The operator T is analogous to matrix multiplication.

Recall: Spectral Theorem (For PSD matrices): Let $A \in \mathbb{R}^{s \times s}$ be a symmetric PSD matrix. Then there exists a collection $\{\phi_1, ..., \phi_s\} \subseteq \mathbb{R}^s$ such that:

1. $A\phi_k = \lambda_k \phi_k$ for some $\lambda_k \geq 0$. (ϕ_k are eigenvectors of A).
2. $\{\phi_1, ..., \phi_s\}$ form an orthonormal basis for $\mathbb{R}^s$: $\langle \phi_k, \phi_l \rangle = \delta_{kl}$. And $\forall v \in \mathbb{R}^s$: $v = \sum_{k=1}^s \langle v, \phi_k \rangle \phi_k$.
3. $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_s \geq 0$.
4. $A = \sum_{k=1}^s \lambda_k \phi_k \phi_k^T$.

Theorem 8 (Mercer's Theorem). *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded continuous Kernel. There exist a collection $\{\psi_k\}_{k\in\mathbb{N}} \subseteq L^2(p)$ such that:*

1. *$T\psi_k = \lambda_k \psi_k$ for $\lambda_k \geq 0$. (The ψ_k are eigenfunctions of T).*

2. *$\{\psi_k\}_{k\in\mathbb{N}}$ forms an orthonormal basis for $L^2(p)$:*

$$\langle \psi_k, \psi_l \rangle_{L^2(p)} = \int \psi_k(x)\psi_l(x)p(x)dx = \delta_{kl}$$

And $\forall f \in L^2(p)$:

$$f(\cdot) = \sum_{k=1}^{\infty} \langle f, \psi_k \rangle_{L^2(p)} \psi_k(\cdot)$$

3. *$\lambda_1 \geq \lambda_2 \geq ... \lambda_k \to 0$ as $k \to \infty$.*

4. *$K(x, \tilde{x}) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(\tilde{x})$.*

(This is a result from Functional Analysis).
Let us just prove (4): Let $x \in \mathcal{X}$. $K(\cdot, x) \in L^2(p)$.

$$\int_{\mathbb{R}^d} (K(\tilde{x}, x))^2 p(\tilde{x})d\tilde{x} \leq \int_{\mathbb{R}^d} K(\tilde{x}, \tilde{x})K(x, x)p(\tilde{x})d\tilde{x} \leq C^2 \int_{\mathbb{R}^d} p(\tilde{x})d\tilde{x} < \infty$$

Then $K(\cdot, x)$ admits the representation:

$$K(\cdot, x) = \sum_{k=1}^{\infty} \langle K(\cdot, x), \psi_k \rangle_{L^2(p)} \psi_k(\cdot)$$

(Notice this is NOT the RKHS inner product!). By definition:

$$\langle K(\cdot, x), \psi_k \rangle_{L^2(p)} = \int_{\mathbb{R}^d} K(\tilde{x}, x)\psi_k(\tilde{x})p(\tilde{x})d\tilde{x} = T\psi_k(x) = \lambda_k \psi_k(x)$$

Therefore:

$$K(\cdot, x) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(\cdot) \implies K(\tilde{x}, x) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(\tilde{x})$$

21.4 Karhunen-Loève Expansion

Q1: How do we generate a Gaussian process with covariance function $= K$? We use the **Karhunen-Loève expansion**.

Given K, suppose we obtain $\{(\lambda_k, \psi_k)\}_{k\in\mathbb{N}}$ associated to T. Let $\xi_1, \xi_2, ... \sim_{i.i.d.} \mathcal{N}(0, 1)$ (Standard Gaussian). We define the random function:

$$Z(\cdot) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k \psi_k(\cdot)$$

Claim: Z is a Gaussian process with covariance function K.

To see this, note:

$$\begin{aligned}\mathrm{Cov}(Z_x, Z_{\tilde{x}}) &= \mathrm{Cov}\left(\sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k\psi_k(x), \sum_{l=1}^{\infty} \sqrt{\lambda_l}\xi_l\psi_l(\tilde{x})\right) \\ &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \sqrt{\lambda_k}\sqrt{\lambda_l}\psi_k(x)\psi_l(\tilde{x})\mathrm{Cov}(\xi_k, \xi_l)\end{aligned}$$

Since $\mathrm{Cov}(\xi_k, \xi_l) = 1$ if $k = l$ and 0 if $k \neq l$:

$$= \sum_{k=1}^{\infty} \lambda_k\psi_k(x)\psi_k(\tilde{x}) = K(x, \tilde{x})$$

This beautifully closes the loop!

22 Lecture 22: Sampling from GPs and Limits of Kernels

22.1 Sampling from a Gaussian Process

Q1: How do we generate samples from a Gaussian process with covariance function K?

22.1.1 Method 1: Using the Karhunen-Loève expansion

Given K, suppose we obtain $\{(\lambda_k, \psi_k)\}_{k\in\mathbb{N}}$ associated to T. Let $\xi_1, \xi_2, \cdots \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ (Standard Gaussian). We define the random function:

$$Z(\cdot) = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k\psi_k(\cdot)$$

Claim: Z is a Gaussian process with covariance function K.
To see this, note:

$$\begin{aligned}\mathrm{Cov}(Z_x, Z_{\tilde{x}}) &= \mathrm{Cov}\left(\sum_{k=1}^{\infty}\sqrt{\lambda_k}\xi_k\psi_k(x), \sum_{l=1}^{\infty}\sqrt{\lambda_l}\xi_l\psi_l(\tilde{x})\right)\\ &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sqrt{\lambda_k}\sqrt{\lambda_l}\psi_k(x)\psi_l(\tilde{x})\mathrm{Cov}(\xi_k,\xi_l)\end{aligned}$$

Since $\mathrm{Cov}(\xi_k, \xi_l) = 1$ if $k = l$ and 0 if not:

$$\begin{aligned}&= \sum_{k=1}^{\infty}\lambda_k\psi_k(x)\psi_k(\tilde{x})\\ &= K(x,\tilde{x}).\end{aligned}$$

Need to know $\{(\lambda_l, \psi_l)\}_{l\in\mathbb{N}}$. If you do know this, and you have to produce 1 sample from the process, all you need to do is to produce independent samples from a standard Gaussian (e.g., `np.random.randn(N)` in Python). Your "sample" from this Gaussian process is the function:

$$x \in \mathcal{X} \mapsto \sum_{l=1}^{\infty}\xi_l\sqrt{\lambda_l}\psi_l(x)$$

Since in practice we cannot generate ∞ many i.i.d. $\mathcal{N}(0,1)$, we generate N i.i.d. $\xi_1, \dots, \xi_N$ and consider the function:

$$\sum_{l=1}^{N}\xi_l\sqrt{\lambda_l}\psi_l(x)$$

Remark: Here we "discretize" by truncating the infinite sum. However, note that our discretization allows us to evaluate the produced function at any given $x \in \mathcal{X}$.

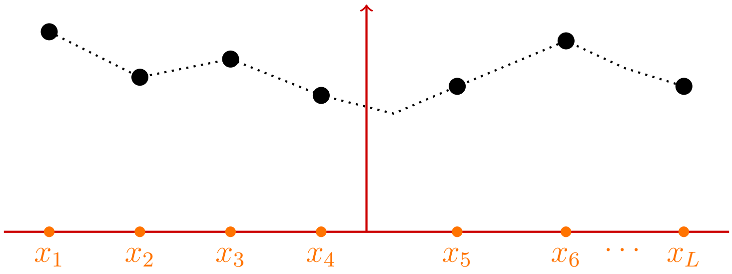
22.1.2 Method 2: Discretizing the data space $\mathcal{X}$

Suppose that $x_1, \dots, x_L$ are points in $\mathcal{X}$. For example, imagine $\mathcal{X} = [0,1]$ and $x_i = \frac{i}{L}$ for $i = 0, \dots, L$.

If Z is a Gaussian process over $\mathcal{X}$ with covariance function K, then $(Z(x_1), \ldots, Z(x_L))$ is multivariate Gaussian (L-dimensional) with mean vector zero and covariance matrix $\Sigma \in \mathbb{R}^{L \times L}$ satisfying:

$$\Sigma_{ij} = K(x_i, x_j) \quad \text{for } i, j = 1, \ldots, L.$$

Thus, to at least sample our random function at the points $x_1, \ldots, x_L$, we may as well sample from $\mathcal{N}(\vec{0}, \Sigma)$. The output could be represented as:



Naturally, if the discretization is finer and finer, we will get a function that looks as if it had been defined everywhere, not just at the points $x_1, \ldots, x_L$. The advantage of this approach is that we don't need to know the eigenpairs $\{(\lambda_k, \psi_k)\}$, but the disadvantage is that the function sampled is only defined at discrete points.

If the $\{(\lambda_k, \psi_k)\}$ are available, use Karhunen-Loève. Examples where eigenpairs are available:

1. Gaussian Kernel (optional)
2. $\mathcal{X} = [0, 1]$, $K(s, t) = \min\{s, t\}$

Among others...

22.2 Eigenpairs of $K(s, t) = \min\{s, t\}$

Let us briefly discuss the eigenpairs of $K(s, t) = \min\{s, t\}$. Take $p(x) \equiv 1$ as the density function. Then

$$Tf(t) = \int_0^1 K(s, t) f(s) ds.$$

The equation for the eigenpairs is:

$$\begin{aligned} \lambda f(t) &= Tf(t) \\ &= \int_0^1 (\min\{s, t\}) f(s) ds \\ &= \int_0^t s f(s) ds + t \int_t^1 f(s) ds \quad \forall t \in [0, 1] \end{aligned}$$

Let's differentiate on both sides of the above equality:

$$\lambda f'(t) = t f(t) + \int_t^1 f(s) ds - t f(t) = \int_t^1 f(s) ds.$$

Differentiating again:

$$\lambda f''(t) = -f(t) \quad \text{on } (0, 1).$$

On the other hand,

$$\lambda f(0) = \int_0^1 (\min\{0, s\}) f(s) ds = 0 \implies f(0) = 0.$$

And also

$$\lambda f'(1) = \int_1^1 f(s) ds = 0 \implies f'(1) = 0.$$

Thus f must satisfy:

$$\begin{aligned} \lambda f''(t) &= -f(t) \quad \text{on } (0,1) \\ f(0) &= 0 \\ f'(1) &= 0 \end{aligned}$$

For some $\lambda > 0$. This is a second order linear ODE with boundary conditions. The only solutions are given by

$$\lambda_k = \frac{1}{k^2 \pi^2}, \quad k = 1, 2, \ldots$$

and

$$\psi_k(t) = c_k \sin(k\pi t)$$

up to a constant that we pick so that $\int_0^1 (\psi_k(t))^2 dt = 1$.

22.3 Why go beyond Kernels?

We have discussed many different aspects of Kernel-based methods...

Some nice features:

1. Usually just need to solve convex optimization problems to complete learning tasks:

$$\min_{a_1, \ldots, a_n} \sum_{i=1}^n \ell_i \left(\sum_{j=1}^n a_j K(\cdot, x_j) \right) + C \sum_{i=1}^n \sum_{j=1}^n a_i a_j K(x_i, x_j)$$

   This is convex in a if ℓ_i are convex. This is the case for all loss functions we discussed in class.

2. **Universal Approximation:** Let $A \subseteq \mathbb{R}^d$ be a compact set and take K to be the Gaussian Kernel over A:

$$K(x, \tilde{x}) = \exp\left(\frac{-\|x - \tilde{x}\|^2}{\sigma^2} \right)$$

   for some fixed σ. Then, for all $g : A \to \mathbb{R}$ continuous and every $\epsilon > 0$, there exists $f \in \mathcal{H}$ (RKHS of K) such that:

$$\sup_{x \in A} |f(x) - g(x)| \leq \epsilon.$$

Why does one need to go beyond Kernels? While the universal approximation property is desirable, the question is how complex the approximating function f has to be. For example, in practice we will work with

$$f(\cdot) = \sum_{i=1}^n a_i K(\cdot, x_i)$$

(because these are the functions that arise when fitting on n data points, thanks to Representer theorem).

Q: How large does n have to be to approximate g? Even if we assume $g \in C^k(\mathbb{R}^d)$ (k-times differentiable), in general n will have to scale like:

$$n \sim \epsilon^{-\frac{d}{k}}$$

If $k = 2$ (for example, the function is twice differentiable) and $d = 400$ (e.g., 20×20 pixels):

$$n \sim \epsilon^{-200}$$

So, if $\epsilon = 1/2$ (we are willing to be loose in our approximations), then

$$n \sim 2^{200}$$

This is a huge number!

Remark: This phenomenon will happen with any space of functions (even neural networks) and has to do with something called **No Free Lunch theorems**.

The question is, can Kernels perform well in tasks that may be of interest (Here task can be thought of as a function we want to learn from data)? Not really! One specific setting where Kernels fare poorly is in approximating **compositional functions**: Suppose $d = 2^L$.

$$g(x_1, \ldots, x_d) = \ldots h(h(x_1, x_2), h(x_3, x_4)), h(h(x_5, x_6), h(x_7, x_8))$$

Example: A simple example of a compositional function is summation:

$$g(x_1, \ldots, x_d) = x_1 + \cdots + x_d$$

It has the form above with $h(a, b) = a + b$. $g : \mathbb{R}^d \to \mathbb{R}$ is very simple. Why would we have to pay such a high sample complexity price if g is fully determined by $h : \mathbb{R}^2 \to \mathbb{R}$?

Bengio and LeCun (2007) use this discussion to motivate deep learning. Compositional functions are the types of functions that deep neural networks can approximate quite well!

23 Lecture 23: Course Summary and Conformal Prediction

23.1 Stat 615: Summary

What have we talked about in this class?

- **Definition of the classification problem:**
  - Bayes classifier (and Bayes risk) or how to build the best possible classifier when the underlying data distribution is known.
  - Bayes classifier in binary and multi-label settings and under different loss functions.
  - Linear regression and quantile regression.
- **How to construct classifiers from finite data:**
  - Plug-in similarity classifiers.
  - Linear classifiers from classical statistics: classification using linear regression, LDA (and QDA), logistic regression.
  - Support vector machines (hard and soft versions) without change of geometry.
  - Generic empirical risk minimization.
  - (Regularized) Empirical risk minimization in RKHSs (using logistic loss, Hinge loss). With Hinge loss, we recover what we called kernelized SVMs.
- **Statistical properties of classifiers built from data:** assuming data are i.i.d. sampled from a distribution on pairs (x, y), what can we say about the statistical properties of:
  - Similarity classifiers (Stone's theorem, universal strong consistency of histograms and k-NN classifiers). Some mathematical tools introduced to answer this question: concentration inequalities (Hoeffding's, Bernstein's, McDiarmid's inequalities).
  - Classifiers using empirical risk minimization. Mathematical tools introduced to answer this question: VC theory and Rademacher complexity.
- **Regression functions built from data (x, y) when y can take any value in $\mathbb{R}$:**
  - Linear regression as in classical statistics.
  - Ridge regression (or regularized least squares).
  - Non-parametric regression using RKHS, i.e., regularized empirical risk minimization in RKHS using something like the squared-loss.
  - Non-parametric quantile regression using RKHS, i.e., regularized empirical risk minimization in RKHS using a suitable loss.
- **Kernels, reproducing kernel Hilbert spaces, data embeddings and relation to classification and regression problems:**

- What is a kernel?
- What is the relationship between changing the geometry of a data set and selecting a kernel?
- Kernels and RKHS (Moore-Aronszajn theorem, reproducing property of a kernel).
- RKHS and applications to classification and regression (regularized empirical risk minimization in RKHS and Representer Theorem).
- The Gaussian kernel in $\mathbb{R}^d$ and some of its properties (see HW 4).
- Characterization of stationary/homogeneous kernels in $\mathbb{R}^d$ (Bochner's theorem).

- **Relation between kernels and Gaussian processes:**
  - Gaussian processes in spaces of functions or how to define priors when unknown is a function.
  - Gaussian processes and their covariance functions.
  - How to build a Gaussian process with a covariance function that matches a pre-specified kernel: Karhunen-Loève expansion of a Gaussian process in terms of eigenfunctions of a certain operator involving the pre-specified kernel.

Some of the things that we did not talk about in this class...

- **Unsupervised learning:** we always assumed that our data sets consist of pairs (x_i, y_i) that we use to define empirical risks (for optimization-based approaches) to build predictors. But what if we only had access to x_i data, i.e., we didn't have any corresponding labels? An example of an unsupervised learning problem is clustering, where the goal is to find representative subgroups of the data.
- **Random forests** (other ways to build classifiers).
- **Reinforcement learning**, contrastive learning, learning with fairness constraints, adversarial learning, federated learning, etc. Many of these are frameworks for learning in either supervised or unsupervised learning settings.

Last time, we also briefly discussed why kernel-based ML is not sufficient for learning in high-dimensional settings. Bengio and LeCun (2007) then argue that the failure of kernel-based methods to approximate compositional functions warrants the use of deep learning methods.

Scaling Learning Algorithms towards AI
Yoshua Bengio (1) and Yann LeCun (2)
(1) Yoshua.Bengio@umontreal.ca, Département d'Informatique et Recherche Opérationnelle, Université de Montréal.
(2) yann@cs.nyu.edu, The Courant Institute of Mathematical Sciences, New York University, New York, NY.

23.2 Uncertainty Quantification through Conformal Prediction

Today, we will consider one last important and relatively recent idea in machine learning: uncertainty quantification through conformal prediction. We will discuss these ideas in the context of classification and regression.

23.2.1 1. Binary Classification

Suppose we have data:

- $(X_1, Y_1), ..., (X_n, Y_n) \sim \rho$ (Training Data)
- $(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_m, \tilde{Y}_m) \sim \rho$ (Held out Data / Test Data)

Step 1: The training data is used to obtain a predictor $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$. For example, $\hat{f}_n$ could be obtained by solving:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell_i(f(X_i)) + C\|f\|_{\mathcal{H}}^2$$

where $\mathcal{H}$ is the RKHS. $\hat{f}_n$ could have been obtained in any other way. For example:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f(X_i))$$

where $\mathcal{F}$ is a family of neural networks with parameters satisfying some conditions.

Step 2: Introduce a "score" measuring the agreement between a predicted label $\hat{f}_n(x)$ and an actual label y.

$$S(\hat{f}_n(x), y) \in \mathbb{R}$$

(The intuition is that the larger $S(\cdot, \cdot)$, the less agreement between the predicted and the actual label). This is a subjective measure of "confidence"! How is this score function defined? For example, in the case of binary classification $\mathcal{Y} = \{-1, 1\}$, we can define:

$$S(\hat{f}_n(x), y) := \exp(-\hat{f}_n(x)y)$$

Step 3: Compute scores at the held out data:

$$S(\hat{f}_n(\tilde{X}_1), \tilde{Y}_1), ..., S(\hat{f}_n(\tilde{X}_m), \tilde{Y}_m)$$

and arrange them in increasing order:

$$S(\hat{f}_n(\tilde{X}_{(1)}), \tilde{Y}_{(1)}) \leq \cdots \leq S(\hat{f}_n(\tilde{X}_{(m)}), \tilde{Y}_{(m)})$$

Moreover, for a given $\alpha \in (0, 1)$ (Fixed), find the $(1-\alpha)$-quantile of the above collection of numbers:

$$\hat{q}_{1-\alpha} = S(\hat{f}_n(\tilde{X}_{(k)}), \tilde{Y}_{(k)}) \quad \text{where } k = \lceil (1-\alpha)(m+1) \rceil$$

(i.e., we pick the point at index $\lceil (1-\alpha)(m+1) \rceil$).

Step 4: Define, for each $x \in \mathcal{X}$, the (confidence) set:

$$C(x) := \{y \in \mathcal{Y} \text{ s.t. } S(\hat{f}_n(x), y) \leq \hat{q}_{1-\alpha}\}$$

How can this be used? For a given input x, instead of outputting $\hat{f}_n(x)$ (or a hard label), we output the whole confidence set $C(x)$.

23.2.2 2. Multiclass Setting

Example in multiclass setting ($\mathcal{Y} = \{1, ..., K\}$): *(Figure 1: Prediction set examples on Imagenet. We show three progressively more difficult examples of the class for squirrel and the prediction sets, i.e., $C(X_{\text{test}})$, generated by conformal prediction. Easy examples yield a single class "squirrel", while ambiguous examples yield sets like {squirrel, fox, bucket, barrel}).* **Remark:** The intuition is that the bigger $C(x)$ is, the more "uncertain" our prediction is.

The approach is just as before, we just need to adjust a few details:

- **Step 1:** Suppose, for example, that $\hat{f}_n(x) \in \mathbb{R}^K$.
- **Step 2:** Suppose we define

$$S(\hat{f}_n(x), y) = \exp(-(\hat{f}_n(x))_y)$$

  where $(\hat{f}_n(x))_y$ is the y-th entry of $\hat{f}_n(x)$.

23.2.3 3. Regression Setting

As in the multiclass case, we just need to adjust a few things: **Step 1:** Suppose $\hat{f}_n(x) = (\hat{f}_n^-(x), \hat{f}_n^+(x)) \in \mathbb{R}^2$. For example, $\hat{f}_n(\cdot)$ could be obtained through two quartile regression procedures: $\hat{f}_n^-(\cdot)$ solves:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell_{0.25}(y_i - f(X_i)) + C\|f\|_{\mathcal{H}}^2$$

and $\hat{f}_n^+(\cdot)$ solves:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell_{0.75}(y_i - f(X_i)) + C\|f\|_{\mathcal{H}}^2$$

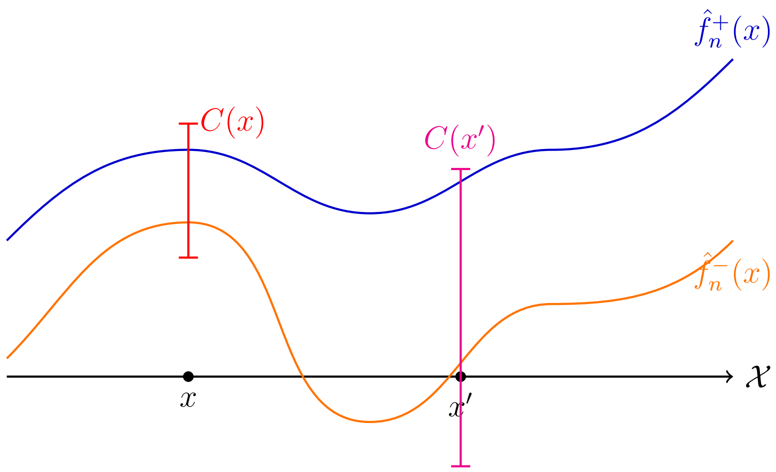
(Kernel-based quartile regression), where, recall:

$$\ell_\tau(s) = \begin{cases} (\tau - 1)s & \text{if } s < 0 \\ \tau s & \text{if } s \geq 0 \end{cases}$$

Step 2: The score could be set to be:

$$S(\hat{f}_n(x), y) = \max\{\hat{f}_n^-(x) - y, y - \hat{f}_n^+(x), 0\}$$

In that case we have the following geometry:



The larger the interval $C(x)$, the more "uncertain" we are about our prediction.

23.3 Theoretical Guarantee

Why is this reasonable?

Theorem 9. *Suppose* $(X_1, Y_1), ..., (X_n, Y_n)$ *and* $(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_m, \tilde{Y}_m) \sim_{i.i.d.} \rho$. *Let* $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ *(Not depending on the held-out data* $(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_m, \tilde{Y}_m)$*). Then:*

$$\mathbb{P}_{(X,Y)\sim\rho}(Y \in C(X)) \geq 1 - \alpha$$

Reference: See *"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"* by Anastasios N. Angelopoulos and Stephen Bates.