Permutation test for classification
&
Risk bounds for mixture of densities

Sayan Mukherjee
Center for Biological and Computational Learning
Cancer Genomics Group
MIT

The learning problem

Given a dataset $S = \{z_1 = (x_1, y_1), ..., z_n = (x_n, y_n)\}$ drawn i.i.d. from a distribution $P(x, y)$.

The learning problem

Given a dataset $S = \{z_1 = (x_1, y_1), ..., z_n = (x_n, y_n)\}$ drawn i.i.d. from a distribution $P(x, y)$.

An algorithm is the map $\mathcal{A} : S \to f_S$.

The learning problem

Given a dataset $S = \{z_1 = (x_1, y_1), ..., z_n = (x_n, y_n)\}$ drawn i.i.d. from a distribution $P(x, y)$.

An algorithm is the map $\mathcal{A} : S \to f_S$.

Empirical risk minimization

$$\mathcal{A} : \quad f_S \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i).$$

Definitions

- Empirical error: $R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)$

Definitions

- Empirical error: $R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$
- Expected error: $R[f] = \int_{X \times Y} V(f(x), y) dP(x, y)$

Definitions

- Empirical error: $R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$
- Expected error: $R[f] = \int_{X \times Y} V(f(x), y) dP(x, y)$
- Target function: $f_0 = \arg\min R[f] \qquad f_0 \in \mathcal{G}$

Definitions

- Empirical error: $R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$
- Expected error: $R[f] = \int_{X \times Y} V(f(x), y) dP(x, y)$
- Target function: $f_0 = \arg\min R[f] \qquad f_0 \in \mathcal{G}$
- Best function in the class: $f_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} R[f]$

Definitions

- Empirical error: $R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$
- Expected error: $R[f] = \int_{X\times Y} V(f(x), y) dP(x, y)$
- Target function: $f_0 = \arg\min R[f] \qquad f_0 \in \mathcal{G}$
- Best function in the class: $f_{\mathcal{H}} \in \arg\min_{f\in\mathcal{H}} R[f]$
- Function output: $f_S \in \arg\min_{f\in\mathcal{H}} R_{emp}[f]$

Approximation and estimation errors

$$R[f_S] - R[f_0] = R[f_S] - R[f_{\mathcal{H}}] \quad + \quad R[f_{\mathcal{H}}] - R[f_0]$$

$$\text{generalization error} = \text{estimation error} \quad + \quad \text{approximation error}$$

Approximation and estimation errors

$$R[f_S] - R[f_0] = R[f_S] - R[f_{\mathcal{H}}] \quad + \quad R[f_{\mathcal{H}}] - R[f_0]$$

generalization error = estimation error + approximation error

As $|\mathcal{H}|$ increases: approximation error decreases

estimation error increases.

Analysis of the tradeoff

1. Niyogi and Girosi: Approximation-estimation analysis for Radial Basis Functions

Analysis of the tradeoff

1. Niyogi and Girosi: Approximation-estimation analysis for Radial Basis Functions
2. Barron, Li and Barron: Approximation estimation analysis for neural networks and mixtures of densities

Analysis of the tradeoff

1. Niyogi and Girosi: Approximation-estimation analysis for Radial Basis Functions

2. Barron, Li and Barron: Approximation estimation analysis for neural networks and mixtures of densities

3. Cucker and Smale: Best choices for regularization parameters for RKHS

Analysis of the tradeoff

1. Niyogi and Girosi: Approximation-estimation analysis for Radial Basis Functions

2. Barron, Li and Barron: Approximation estimation analysis for neural networks and mixtures of densities

3. Cucker and Smale: Best choices for regularization parameters for RKHS

4. Smale and Zhou: Estimating the approximation error for RKHS

Two problems

1. Permutation tests for classification: uses label permutations to compute a bias variance tradeoff for classification.
   S. Mukherjee, P. Golland and D. Panchenko.

Two problems

1. Permutation tests for classification: uses label permutations to compute a bias variance tradeoff for classification.
   S. Mukherjee, P. Golland and D. Panchenko.

2. Risk bounds for mixture of densities: approximation and estimation bounds for mixture of densities models.
   A. Rakhlin, D. Panchenko, and S. Mukherjee

Permutation tests for classification

The permutation procedure described here is used extensively in gene expression analysis and image based clinical studies.

Permutation tests for classification

The permutation procedure described here is used extensively in gene expression analysis and image based clinical studies.

- Image-based clinical studies: detect neuroanatomical chances induced by diseases and predict disease development.

Permutation tests for classification

The permutation procedure described here is used extensively in gene expression analysis and image based clinical studies.

- Image-based clinical studies: detect neuroanatomical chances induced by diseases and predict disease development.
- Gene expression analysis: classify tissue morphology, lineage, treatment outcome, or drug sensitivity using DNA microarray data.

The practical problem

$x \in \mathbb{R}^{16,000}$, $\quad y \in \{-1, 1\}$ and $|S| \approx 50$.

The practical problem

$x \in \mathbb{R}^{16,000}, \quad y \in \{-1, 1\}$ and $|S| \approx 50$.

We compute a statistic $\mathcal{T}[S]$:

$$\text{Training error} \qquad \frac{1}{n}\sum_{i=1}^{n} V(f_S(x_i), y_i)$$

$$\text{Leave-one-out error} \qquad \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}(x_i), y_i).$$

The practical problem

$x \in \mathbb{R}^{16,000}, \quad y \in \{-1, 1\}$ and $|S| \approx 50$.

We compute a statistic $\mathcal{T}[S]$:

$$\text{Training error} \qquad \frac{1}{n}\sum_{i=1}^{n} V(f_S(x_i), y_i)$$

$$\text{Leave-one-out error} \qquad \frac{1}{n}\sum_{i=1}^{n} V(f_{S^i}(x_i), y_i).$$

Can we trust $\mathcal{T}[S]$?

Approximation estimation breakdown

$\mathcal{T}[S]$ is the proxy for approximation error.

Approximation estimation breakdown

$\mathcal{T}[S]$ is the proxy for approximation error.

$\text{Var}\{\mathcal{T}[\pi_1(S)], \ldots, \mathcal{T}[\pi_M(S)]\}$ is the proxy for estimation error:

Approximation estimation breakdown

$\mathcal{T}[S]$ is the proxy for approximation error.

$\text{Var}\{\mathcal{T}[\pi_1(S)], ..., \mathcal{T}[\pi_M(S)]\}$ is the proxy for estimation error:

- Repeat $m = 1, ..., M$ times
  - permute the labels: $\pi_m(S)$,
  - $t_m = \mathcal{T}[\pi_m(S)]$

Approximation estimation breakdown

$\mathcal{T}[S]$ is the proxy for approximation error.

$\text{Var}\{\mathcal{T}[\pi_1(S)], ..., \mathcal{T}[\pi_M(S)]\}$ is the proxy for estimation error:

- Repeat $m = 1, ..., M$ times
  - ⋆ permute the labels: $\pi_m(S)$,
  - ⋆ $t_m = \mathcal{T}[\pi_m(S)]$
- construct an empirical cummulative distribution

$$\hat{\mathbb{P}}(T \leq t) = \frac{1}{M} \sum_{m=1}^{M} \Theta(t - t_m),$$

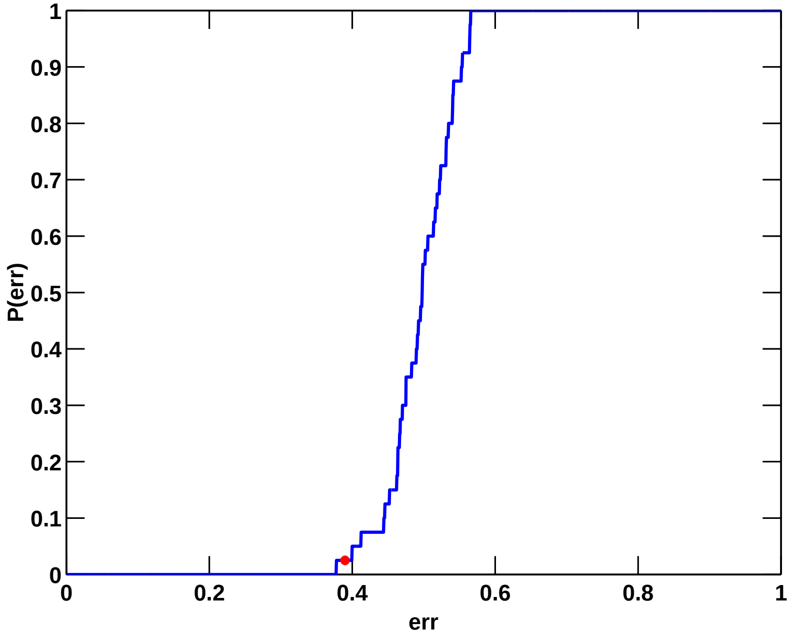
- the p-value of $\mathcal{T}[S]$ is $\hat{\mathbb{P}}(T \leq \mathcal{T}[S])$.

Toy example

$\mathcal{T}[S] = .39, .27, .25, .2$ for $\mathcal{H}_4 \subset \mathcal{H}_3 \subset \mathcal{H}_2 \subset \mathcal{H}_1$.

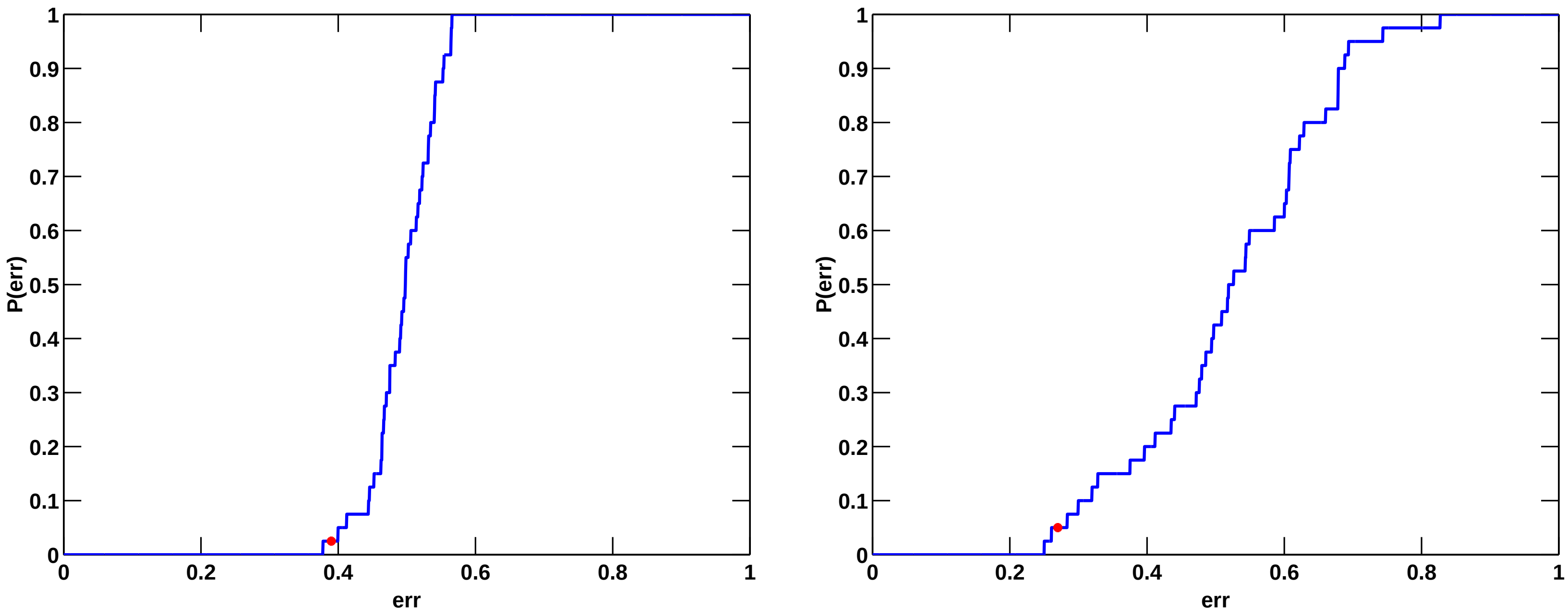
Toy example

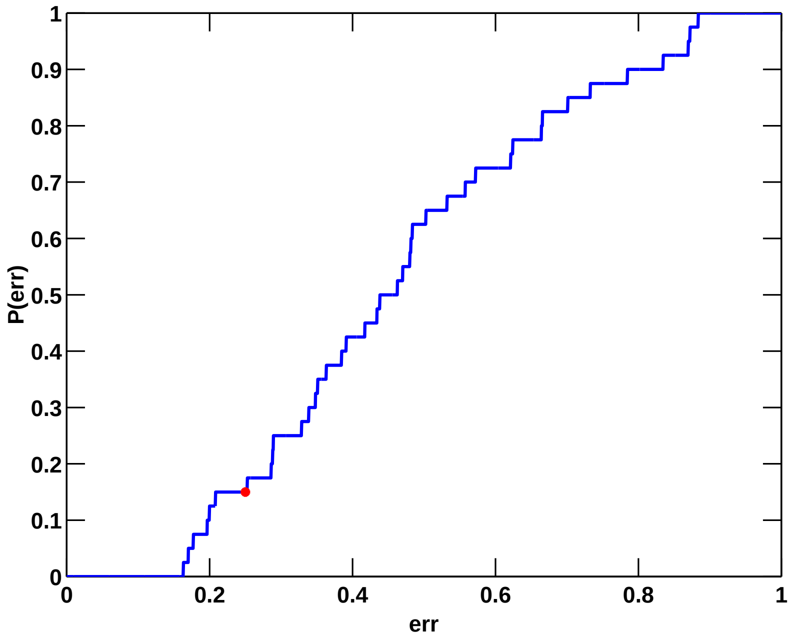
$$\mathcal{T}[S] = .39, .27, .25, .2 \text{ for } \mathcal{H}_4 \subset \mathcal{H}_3 \subset \mathcal{H}_2 \subset \mathcal{H}_1.$$

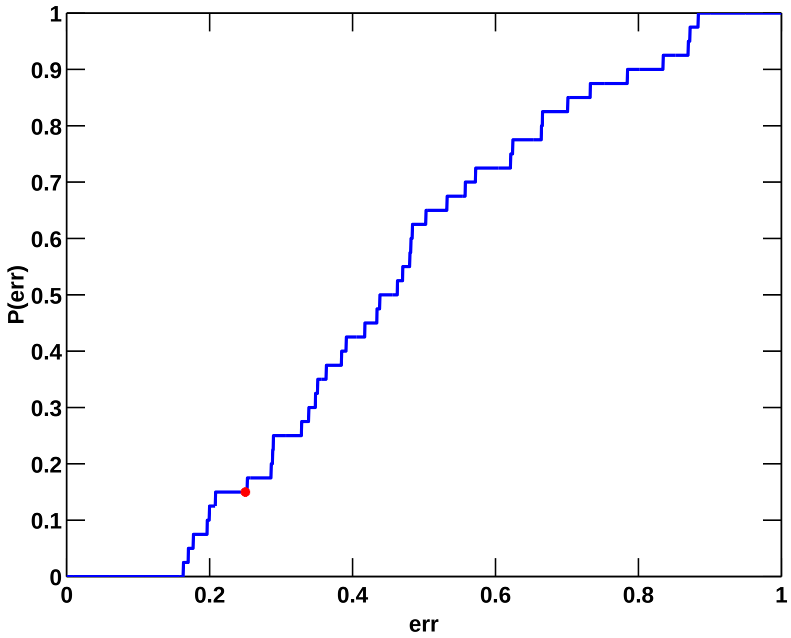


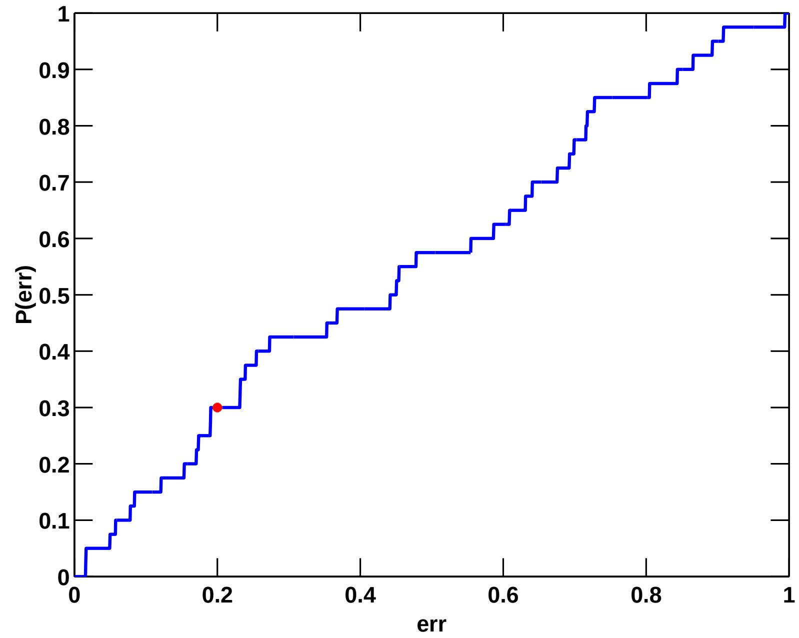
Toy example

$$\mathcal{T}[S] = .39, .27, .25, .2 \text{ for } \mathcal{H}_4 \subset \mathcal{H}_3 \subset \mathcal{H}_2 \subset \mathcal{H}_1.$$




P(err)
err
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1


P(err)
err
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8

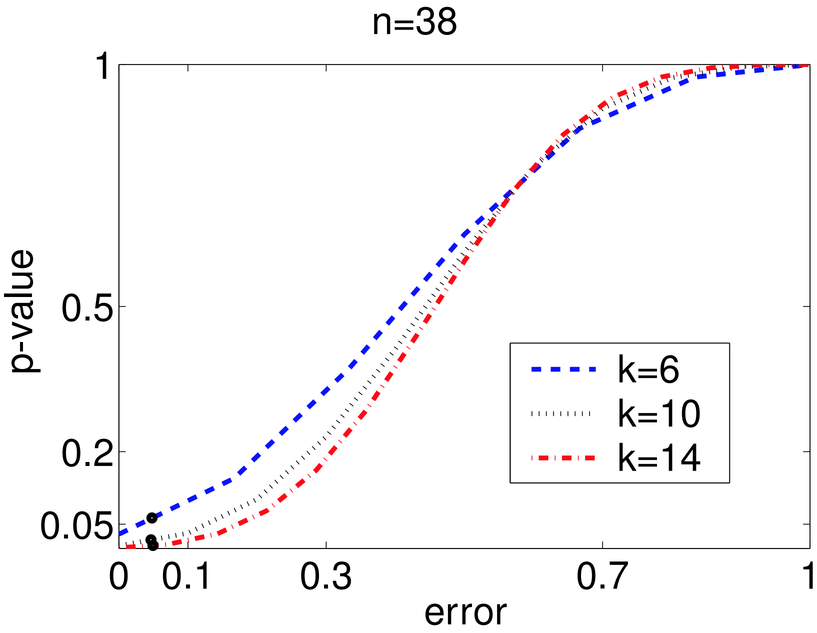

P(err)
err
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8

Leukemia

38 samples from 2 types of leukemia, picking k in leave-k-out.

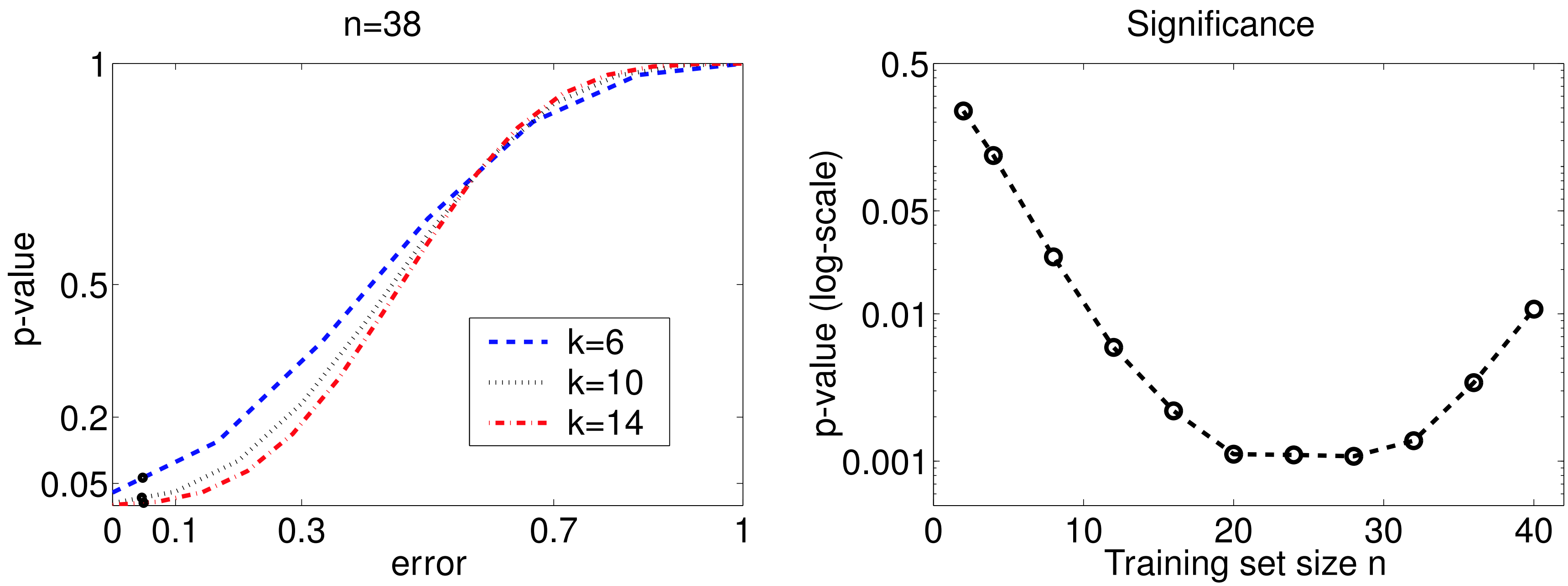
Leukemia

38 samples from 2 types of leukemia, picking k in leave-k-out.



Leukemia

38 samples from 2 types of leukemia, picking k in leave-k-out.



Generalization of the permutation process

Given $\mathcal{H}$ with target f_0. For a permutation $\pi(S)$ the smallest training error is

$$\begin{aligned} e_n(\pi(S)) &= \min_{f \in \mathcal{H}} P_n(f \triangle f_0) \\ &= \min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f, z_i^{\pi} \notin f_0) + I(z_i \notin f, z_i^{\pi} \in f_0) \right], \end{aligned}$$

where z_i is the ith sample z_i^{π} is the ith sample after permutation.

Generalization of the permutation process

Given $\mathcal{H}$ with target f_0. For a permutation $\pi(S)$ the smallest training error is

$$\begin{aligned} e_n(\pi(S)) &= \min_{f \in \mathcal{H}} P_n(f \triangle f_0) \\ &= \min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f, z_i^{\pi} \notin f_0) + I(z_i \notin f, z_i^{\pi} \in f_0) \right], \end{aligned}$$

where z_i is the ith sample z_i^{π} is the ith sample after permutation.

For a fixed $f \in \mathcal{H}$ the expected error (over permutations) is

$$\mathbb{E}_{\pi} P_n(f \triangle f_0) = P(z \in f)(1 - P(z \in f_0)) + (1 - P(z \in f))P(z \in f_0) = P(z \in f_0) \equiv P(f_0).$$

Generalization of the permutation process

Given $\mathcal{H}$ with target f_0. For a permutation $\pi(S)$ the smallest training error is

$$\begin{aligned} e_n(\pi(S)) &= \min_{f\in\mathcal{H}} P_n(f \triangle f_0) \\ &= \min_{f\in\mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f, z_i^\pi \notin f_0) + I(z_i \notin f, z_i^\pi \in f_0) \right], \end{aligned}$$

where z_i is the ith sample z_i^π is the ith sample after permutation.

For a fixed $f \in \mathcal{H}$ the expected error (over permutations) is

$$\mathbb{E}_\pi P_n(f \triangle f_0) = P(z \in f)(1 - P(z \in f_0)) + (1 - P(z \in f))P(z \in f_0) = P(z \in f_0) \equiv P(f_0).$$

For appropriate complexity assumptions on $\mathcal{H}$ prove that $e_n(\pi(S))$ is close to $P(f_0)$.

Concentration of the permutation process

The following maximization problem is equivalent to minimizing the empirical error on permuted data

$$e_n(\pi(S)) = P_n(z \in f_0) - \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2I(z_i^\pi \in f_0) - 1) \right].$$

Concentration of the permutation process

The following maximization problem is equivalent to minimizing the empirical error on permuted data

$$e_n(\pi(S)) = P_n(z \in f_0) - \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2I(z_i^{\pi} \in f_0) - 1) \right].$$

By Chernoff's inequality $P_n(z \in f_0)$ is close to $P(z \in f_0)$.

Concentration of the permutation process

The following maximization problem is equivalent to minimizing the empirical error on permuted data

$$e_n(\pi(S)) = P_n(z \in f_0) - \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2I(z_i^\pi \in f_0) - 1) \right].$$

By Chernoff's inequality $P_n(z \in f_0)$ is close to $P(z \in f_0)$.

So we need only bound the following process

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2I(z_i^\pi \in f_0) - 1) \right].$$

Bound on $G_n(\pi(S))$

Theorem 1. *If the* $\mathcal{H}$ *has VC dimension* V *then with probability* $1 - Ke^{-t/K}$

$$G_n(\pi(S)) \leq K \min\left(\sqrt{\frac{V \log n}{n}}, \frac{V \log n}{n(1 - 2P(f_0))^2}\right) + \sqrt{\frac{Kt}{n}}.$$

Bound on $G_n(\pi(S))$

Theorem 1. *If the* $\mathcal{H}$ *has VC dimension* V *then with probability* $1 - Ke^{-t/K}$

$$G_n(\pi(S)) \leq K \min\left(\sqrt{\frac{V \log n}{n}}, \frac{V \log n}{n(1 - 2P(f_0))^2}\right) + \sqrt{\frac{Kt}{n}}.$$

Therefore with probability $1 - Ke^{-t/K}$

$$P(z \in f_0) \leq P_n(z \in f_0) + K \min\left(\sqrt{\frac{V \log n}{n}}, \frac{V \log n}{n(1 - 2P(f_0))^2}\right) + \sqrt{\frac{Kt}{n}}.$$

Proof sketch (Part 1)

The process can be rewritten

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)\varepsilon_i \right],$$

where $\varepsilon_i = 2I(z_i^\pi \in f_0) - 1 = \pm 1$ are Bernoulli random variables with $P(\varepsilon_i = 1) = P(f_0)$ and (ε_i) depend on (z_i) only through the cardinality of $\{z_i \in f_0\}$.

Proof sketch (Part 1)

The process can be rewritten

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)\varepsilon_i \right],$$

where $\varepsilon_i = 2I(z_i^\pi \in f_0) - 1 = \pm 1$ are Bernoulli random variables with $P(\varepsilon_i = 1) = P(f_0)$ and (ε_i) depend on (z_i) only through the cardinality of $\{z_i \in f_0\}$.

A Bernoulli sequence (ε_i') independent of z can be constructed if a term is added

Proof sketch (Part 1)

The process can be rewritten

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i \right],$$

where $\varepsilon_i = 2I(z_i^\pi \in f_0) - 1 = \pm 1$ are Bernoulli random variables with $P(\varepsilon_i = 1) = P(f_0)$ and (ε_i) depend on (z_i) only through the cardinality of $\{z_i \in f_0\}$.

A Bernoulli sequence (ε_i') independent of z can be constructed if a term is added

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i' \right] + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \right|.$$

Proof sketch (Part 1)

The process can be rewritten

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i \right],$$

where $\varepsilon_i = 2I(z_i^{\pi} \in f_0) - 1 = \pm 1$ are Bernoulli random variables with $P(\varepsilon_i = 1) = P(f_0)$ and (ε_i) depend on (z_i) only through the cardinality of $\{z_i \in f_0\}$.

A Bernoulli sequence (ε_i') independent of z can be constructed if a term is added

$$G_n(\pi(S)) = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i' \right] + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \right|.$$

The second term can be bounded by applying Chernoff's inequality twice.

Proof sketch (Part 2)

We need to bound the following process

$$\sup_{f \in \mathcal{H}} R[f, \varepsilon'] = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i' \right],$$

where $P(\varepsilon_i' = 1) = P(f_0)$.

Proof sketch (Part 2)

We need to bound the following process

$$\sup_{f\in\mathcal{H}} R[f,\varepsilon'] = \sup_{f\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} I(z_i \in f)\varepsilon_i'\right],$$

where $P(\varepsilon_i' = 1) = P(f_0)$.

By Talagrand's inequality on the cube with probability $1 - e^{-Kt}$

$$\sup_{f\in\mathcal{H}} R[f,\varepsilon'] \le \mathbb{E}_{\varepsilon'} \sup_{f\in\mathcal{H}} R[f,\varepsilon'] + \sqrt{\frac{t}{n}}.$$

Proof sketch (Part 2)

We need to bound the following process

$$\sup_{f \in \mathcal{H}} R[f, \varepsilon'] = \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I(z_i \in f) \varepsilon_i' \right],$$

where $P(\varepsilon_i' = 1) = P(f_0)$.

By Talagrand's inequality on the cube with probability $1 - e^{-Kt}$

$$\sup_{f \in \mathcal{H}} R[f, \varepsilon'] \leq \mathbb{E}_{\varepsilon'} \sup_{f \in \mathcal{H}} R[f, \varepsilon'] + \sqrt{\frac{t}{n}}.$$

If $P(\varepsilon_i' = 1) = 1/2$ this is a Rademacher process and Dudley's entropy integral can be used to control $\mathbb{E}_{\varepsilon'} \sup_{f \in \mathcal{H}} R[f, \varepsilon']$.

Proof sketch (Part 3)

We transform the problem into such a form by adding and subtracting an independent sequence (ε_i'') such that $\mathbb{E}\varepsilon_i' = \mathbb{E}\varepsilon_i'' = (2P(f_0) - 1)$

Proof sketch (Part 3)

We transform the problem into such a form by adding and subtracting an independent sequence (ε_i'') such that $\mathbb{E}\varepsilon_i' = \mathbb{E}\varepsilon_i'' = (2P(f_0) - 1)$

$$\mathbb{E}_{\varepsilon'} \sup_{f \in \mathcal{H}} R[f, \varepsilon'] \leq \mathbb{E}_{\varepsilon', \varepsilon''} \sup_{f \in \mathcal{H}} \left[R[f, \varepsilon'] - R[f, \varepsilon''] + \frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2P(f_0) - 1) \right]$$

Proof sketch (Part 3)

We transform the problem into such a form by adding and subtracting an independent sequence (ε_i'') such that $\mathbb{E}\varepsilon_i' = \mathbb{E}\varepsilon_i'' = (2P(f_0) - 1)$

$$\mathbb{E}_{\varepsilon'} \sup_{f \in \mathcal{H}} R[f, \varepsilon'] \leq \mathbb{E}_{\varepsilon', \varepsilon''} \sup_{f \in \mathcal{H}} \left[R[f, \varepsilon'] - R[f, \varepsilon''] + \frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2P(f_0) - 1) \right]$$

$$\leq \quad \mathbb{E}_{\eta} \sup_{f \in \mathcal{H}} \left[R[f, \eta'] + \frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2P(f_0) - 1) \right]$$

where $\eta_i = (\varepsilon_i' - \varepsilon_i'')/2$ takes values $\{-1, 0, 1\}$ and $P(\eta_i = 1) = P(\eta_i = -1)$.

Proof sketch (Part 3)

We transform the problem into such a form by adding and subtracting an independent sequence (ε_i'') such that $\mathbb{E}\varepsilon_i' = \mathbb{E}\varepsilon_i'' = (2P(f_0) - 1)$

$$\mathbb{E}_{\varepsilon'} \sup_{f \in \mathcal{H}} R[f, \varepsilon'] \leq \mathbb{E}_{\varepsilon', \varepsilon''} \sup_{f \in \mathcal{H}} \left[R[f, \varepsilon'] - R[f, \varepsilon''] + \frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2P(f_0) - 1) \right]$$

$$\leq \quad \mathbb{E}_{\eta} \sup_{f \in \mathcal{H}} \left[R[f, \eta'] + \frac{1}{n} \sum_{i=1}^{n} I(z_i \in f)(2P(f_0) - 1) \right]$$

where $\eta_i = (\varepsilon_i' - \varepsilon_i'')/2$ takes values $\{-1, 0, 1\}$ and $P(\eta_i = 1) = P(\eta_i = -1)$.

Since η_i satisfy

$$\mathbb{P}\left(\sum_{i=1}^{n} \eta_i a_i > t \right) \leq e^{-\frac{t^2}{2 \sum_{i=1}^{n} a_i^2}}$$

we can use the entropy integral.

Proof sketch (Part 4)

By the entropy integral bound

$$\mathbb{E}_{\eta_i} \sup_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} I[z_i \in f]\eta_i \right] \leq K \frac{1}{\sqrt{n}} \int_0^{\sqrt{\mu}} \sqrt{\log \mathcal{N}(u, \mathcal{H})} du$$

where $\mu = \frac{1}{n} \sum_{i=1}^{n} I[z_i \in f]$.

The result of the theorem is obtained by computing the entropy integral and optimizing. □

Moral

- If $P(f_0) < 1/2$ then ignoring the "one dimensional terms" the rate of convergence is

$$O\left(\frac{V \log n}{n}\right).$$

- The weak dependency between (z_i) and a sequence (ε_i) can be broken with very little cost.

Risk bounds for mixture of densities

Given a dataset $S = \{x_1, ..., x_n\}$ drawn i.i.d. from an unknown bounded (from above and below) density f_0 estimate this density using k-component mixtures f_k where

$$f_k \in \mathcal{C}_k = \mathrm{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \theta_i \in \Theta \right\},$$

where $\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$.

Risk bounds for mixture of densities

Given a dataset $S = \{x_1, ..., x_n\}$ drawn i.i.d. from an unknown bounded (from above and below) density f_0 estimate this density using k-component mixtures f_k where

$$f_k \in \mathcal{C}_k = \text{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \theta_i \in \Theta \right\},$$

where $\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$.

We are given an algorithm

$$\mathcal{A} : S \to \hat{f}_k$$

where $\hat{f}_k \in \mathcal{C}_k$.

Risk bounds for mixture of densities

Given a dataset $S = \{x_1, ..., x_n\}$ drawn i.i.d. from an unknown bounded (from above and below) density f_0 estimate this density using k-component mixtures f_k where

$$f_k \in \mathcal{C}_k = \text{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \theta_i \in \Theta \right\},$$

where $\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$.

We are given an algorithm

$$\mathcal{A} : S \to \hat{f}_k$$

where $\hat{f}_k \in \mathcal{C}_k$.

We want to bound

$$\mathbb{E}_S[D(f_0 \| \hat{f}_k)] \leq \text{Approx}(\mathcal{C}_k) + \text{Est}(\mathcal{C}_k, n),$$

Risk bounds for mixture of densities

Given a dataset $S = \{x_1, ..., x_n\}$ drawn i.i.d. from an unknown bounded (from above and below) density f_0 estimate this density using k-component mixtures f_k where

$$f_k \in \mathcal{C}_k = \text{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \theta_i \in \Theta \right\},$$

where $\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$.

We are given an algorithm

$$\mathcal{A} : S \to \hat{f}_k$$

where $\hat{f}_k \in \mathcal{C}_k$.

We want to bound

$$\mathbb{E}_S[D(f_0 \| \hat{f}_k)] \leq \text{Approx}(\mathcal{C}_k) + \text{Est}(\mathcal{C}_k, n),$$

where $D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)}$.

The algorithms and some definitions

The following algorithms will be used

$$\mathcal{A}_{\mathrm{MLE}} : \widehat{f}_k = \arg\max_{\lambda,\theta} \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \lambda_j \phi_{\theta_j}(z_i) \right]$$

The algorithms and some definitions

The following algorithms will be used

$$\mathcal{A}_{\mathrm{MLE}} : \widehat{f}_k = \arg\max_{\lambda,\theta} \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} \lambda_j \phi_{\theta_j}(z_i)\right]$$

$$\mathcal{A}_{\mathrm{Greedy}} : \widehat{f}_k = \arg\max_{\theta,\lambda_k} \sum_{i=1}^{n} \log\left[(1-\lambda_k)\widehat{f}_{k-1}(z_i) + \lambda_k \phi_\theta(z_i)\right] .$$

The algorithms and some definitions

The following algorithms will be used

$$\mathcal{A}_{\mathrm{MLE}} : \hat{f}_k = \arg\max_{\lambda,\theta} \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} \lambda_j \phi_{\theta_j}(z_i)\right]$$

$$\mathcal{A}_{\mathrm{Greedy}} : \hat{f}_k = \arg\max_{\theta,\lambda_k} \sum_{i=1}^{n} \log\left[(1-\lambda_k)\hat{f}_{k-1}(z_i) + \lambda_k \phi_\theta(z_i)\right] .$$

We define the class

$$\mathcal{C} = \mathrm{conv}(\mathcal{H}) = \left\{ f : f(x) = \int_\Theta \phi_\theta(x) P(d\theta) \right\}$$

and

$$D(f_0 \| \mathcal{C}) = \inf_{g \in \mathcal{C}} D(f_0 \| g).$$

Approximation estimation tradeoff

Li and Barron proved the following:

Theorem 2. *Assume that* Θ *bounded and Lipschitz*

$$\sup_{x \in \mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \sum_{j=1}^{d} |\theta_j - \theta_j'|$$

for any $\theta, \theta' \in \Theta$*.*

Approximation estimation tradeoff

Li and Barron proved the following:

Theorem 2. *Assume that* Θ *bounded and Lipschitz*

$$\sup_{x \in \mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \sum_{j=1}^{d} |\theta_j - \theta'_j|$$

for any $\theta, \theta' \in \Theta$.

For either $\mathcal{A}_{\text{MLE}}$ *or* $\mathcal{A}_{\text{Greedy}}$

$$\mathbb{E}_S \left[D(f_0 \| \hat{f}_k) \right] - D(f_0 \| \mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2 k}{n} \log(n c_3).$$

Approximation estimation tradeoff

Li and Barron proved the following:

Theorem 2. *Assume that* Θ *bounded and Lipschitz*

$$\sup_{x \in \mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \sum_{j=1}^{d} |\theta_j - \theta'_j|$$

for any $\theta, \theta' \in \Theta$*.*

For either $\mathcal{A}_{\mathrm{MLE}}$ *or* $\mathcal{A}_{\mathrm{Greedy}}$

$$\mathbb{E}_S \left[D(f_0 \| \hat{f}_k) \right] - D(f_0 \| \mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2 k}{n} \log(n c_3).$$

The rate of convergence for optimal k *is* $O\left(\sqrt{\frac{\log n}{n}}\right)$*.*

There is no tradeoff in this problem

Alexander Rakhlin proved the following

Theorem 3. *For any bounded* f_0 *(*$a \leq f_0 \leq b$*) then for either* $\mathcal{A}_{\mathrm{MLE}}$ *or* $\mathcal{A}_{\mathrm{Greedy}}$

$$\mathbb{E}_S\left[D(f_0\|\hat{f}_k)\right] - D(f_0\|\mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}_S\left[\frac{c_2}{\sqrt{n}}\int_0^b \sqrt{\log \mathcal{N}(\mathcal{H}, u, d_x)}du\right],$$

where $\mathcal{N}(\mathcal{H}, u, d_x)$ *is the covering number of* $\mathcal{H}$ *with respect to the empirical distance.*

There is no tradeoff in this problem

Alexander Rakhlin proved the following

Theorem 3. *For any bounded* f_0 *(*$a \le f_0 \le b$*) then for either* $\mathcal{A}_{\text{MLE}}$ *or* $\mathcal{A}_{\text{Greedy}}$

$$\mathbb{E}_S\left[D(f_0\|\hat{f}_k)\right] - D(f_0\|\mathcal{C}) \le \frac{c_1}{k} + \mathbb{E}_S\left[\frac{c_2}{\sqrt{n}}\int_0^b \sqrt{\log \mathcal{N}(\mathcal{H}, u, d_x)}du\right],$$

where $\mathcal{N}(\mathcal{H}, u, d_x)$ *is the covering number of* $\mathcal{H}$ *with respect to the empirical distance.*

Under the conditions Li and Barron examined

$$\mathbb{E}_S\left[D(f_0\|\hat{f}_k)\right] - D(f_0\|\mathcal{C}) \le \frac{c_1}{k} + \frac{c_2}{\sqrt{n}}.$$

There is no tradeoff in this problem

Alexander Rakhlin proved the following

Theorem 3. *For any bounded* f_0 *(*$a \leq f_0 \leq b$*) then for either* $\mathcal{A}_{\mathrm{MLE}}$ *or* $\mathcal{A}_{\mathrm{Greedy}}$

$$\mathbb{E}_S\left[D(f_0 \| \hat{f}_k)\right] - D(f_0 \| \mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}_S\left[\frac{c_2}{\sqrt{n}} \int_0^b \sqrt{\log \mathcal{N}(\mathcal{H}, u, d_x)}\, du\right],$$

where $\mathcal{N}(\mathcal{H}, u, d_x)$ *is the covering number of* $\mathcal{H}$ *with respect to the empirical distance.*

Under the conditions Li and Barron examined

$$\mathbb{E}_S\left[D(f_0 \| \hat{f}_k)\right] - D(f_0 \| \mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}}.$$

There is no optimal k and only the complexity of $\mathcal{H}$ is involved.

Why only $\mathcal{H}$ (Part 1)

By McDiarmid's inequality with probability $1 - e^{-t}$

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| \leq \mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| + C\sqrt{t/n}.$$

Why only $\mathcal{H}$ (Part 1)

By McDiarmid's inequality with probability $1 - e^{-t}$

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| \leq \mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| + C\sqrt{t/n}.$$

By symmetrization

$$\mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| \leq 2\mathbb{E}_{S,\varepsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \log \frac{h(x_i)}{f_0(x_i)} \right|,$$

where $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$.

Why only $\mathcal{H}$ (Part 1)

By McDiarmid's inequality with probability $1 - e^{-t}$

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| \leq \mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| + C\sqrt{t/n}.$$

By symmetrization

$$\mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f_0(x_i)} - \mathbb{E} \log \frac{h}{f_0} \right| \leq 2\mathbb{E}_{S,\varepsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \log \frac{h(x_i)}{f_0(x_i)} \right|,$$

where $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$.

We will see that the Rademacher average above can be controlled only using $\mathcal{H}$.

Why only $\mathcal{H}$ (Part 2)

Lemma 1. *Comparison inequality for Rademacher processes*
If $G : \mathbb{R} \to \mathbb{R}$ *convex and non-decreasing and* $\phi_i : \mathbb{R} \to \mathbb{R}$

($i = 1, .., n$*) contractions (*$\psi_i(0) = 0$ *and* $|\psi_i(s) - \psi_i(t)| \leq |s - t|$*), then*

$$\mathbb{E}_\varepsilon G\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i \psi_i(f(x_i))\right] \leq \mathbb{E}_\varepsilon G\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i)\right].$$

Why only $\mathcal{H}$ (Part 2)

Lemma 1. *Comparison inequality for Rademacher processes*
If $G : \mathbb{R} \to \mathbb{R}$ *convex and non-decreasing and* $\phi_i : \mathbb{R} \to \mathbb{R}$

($i = 1, .., n$*) contractions (*$\psi_i(0) = 0$ *and* $|\psi_i(s) - \psi_i(t)| \leq |s - t|$*), then*

$$\mathbb{E}_\varepsilon G \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i \psi_i(f(x_i)) \right] \leq \mathbb{E}_\varepsilon G \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right] .$$

Applying the above lemma multiple times gives us the following bound

$$\mathbb{E}_{S,\varepsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \log \frac{h(x_i)}{f_0(x_i)} \right| \leq K_1 \ \mathbb{E}_{S,\varepsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i) \right| .$$

Why only $\mathcal{H}$ (Part 3)

The Rademacher averages of a class are equal to those of the convex hull, since a linear functional of convex combinations achieves its maximum value at the vertices. Therefore,

$$\mathbb{E}_\varepsilon \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i) \right| = \mathbb{E}_\varepsilon \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i) \right| .$$

Why only $\mathcal{H}$ (Part 3)

The Rademacher averages of a class are equal to those of the convex hull, since a linear functional of convex combinations achieves its maximum value at the vertices. Therefore,

$$\mathbb{E}_\varepsilon \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i) \right| = \mathbb{E}_\varepsilon \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i) \right| .$$

Using the Dudley integral bound

$$\mathbb{E}_\varepsilon \sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i) \right| \leq \frac{c_1}{\sqrt{n}} \int_0^b \log \sqrt{\mathcal{N}(\mathcal{H}, u, d_x)} du.$$

□

Moral

- For estimates that are convex combinations of Lipschitz functionals the estimation error bound should be a function of the complexity of the base class $\mathcal{H}$ and not the convex combination $\mathcal{C}$.

Moral

- For estimates that are convex combinations of Lipschitz functionals the estimation error bound should be a function of the complexity of the base class $\mathcal{H}$ and not the convex combination $\mathcal{C}$.

- When using mixture models control the complexity of the base class and use as many combinations as you want.