Learning Theory

Ingo Steinwart
University of Stuttgart

September 4, 2013

Informal Description of Supervised Learning

- X space of input samples
  Y space of labels, usually $Y \subset \mathbb{R}$.
- Already observed samples

$$D = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$$

Informal Description of Supervised Learning

- X space of input samples
  Y space of labels, usually $Y \subset \mathbb{R}$.
- Already observed samples

$$D = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (X \times Y)^n$$

- **Goal:**
  With the help of D find a function $f_D : X \to \mathbb{R}$ such that $f_D(x)$ is a good prediction of the label y for new, unseen x.
- **Learning method:**
  Assigns to every training set D a predictor $f_D : X \to \mathbb{R}$.

Illustration: Binary Classification

Problem:
The labels are ± 1.

Goal:
Make few mistakes on future data.

Illustration: Binary Classification

Problem:
The labels are ± 1.

Goal:
Make few mistakes on future data.

Example:

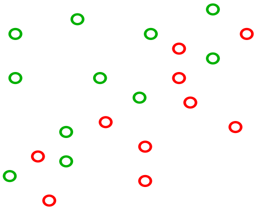


Illustration: Binary Classification

Problem:
The labels are ± 1.

Goal:
Make few mistakes on future data.

Example:

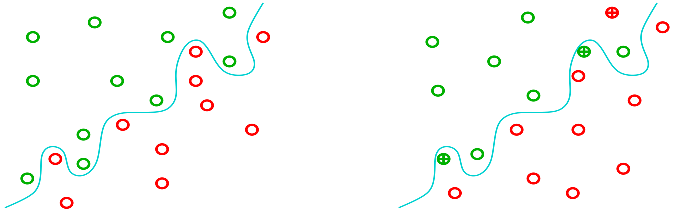


Illustration: Regression

Problem:
The labels are $\mathbb{R}$-valued.

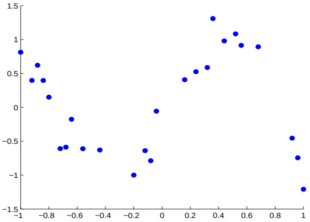
Goal:
Estimate label y for new data x as accurate as possible.

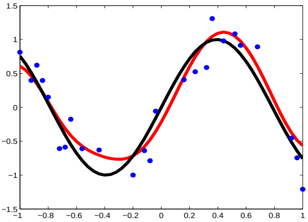
Illustration: Regression

Problem:
The labels are $\mathbb{R}$-valued.

Goal:
Estimate label y for new data x as accurate as possible.

Example:





Data Generation

Assumptions

- P is an unknown probability measure on $X \times Y$.
- $D = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (X \times Y)^n$ is sampled from P^n.
- Future samples (x, y) will also be sampled from P.

Data Generation

Assumptions

- P is an unknown probability measure on $X \times Y$.
- $D = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (X \times Y)^n$ is sampled from P^n.
- Future samples (x, y) will also be sampled from P.

Consequences

- The label y for a given x is, in general, not deterministic.
- The past and the future "look the same".
- We seek algorithms that "work well" for many (or even all) P.

Performance Evaluation I

Loss Function

$L : X \times Y \times \mathbb{R} \to [0, \infty)$ measures cost or loss $L(x, y, t)$ of predicting label y by value t at point x.

Interpretation

- As the name suggests, we prefer predictions with small loss.
- L is chosen by us.
- Since future (x, y) are random, it makes sense to consider the average loss of a predictor.

Performance Evaluation II

Risk

The risk of a predictor $f : X \to \mathbb{R}$ is the average loss

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L\big(x, y, f(x)\big)\, dP(x, y) \ .$$

For $D = ((x_1, y_1), \ldots, (x_n, y_n))$ the empirical risk is

$$\mathcal{R}_{L,D}(f) := \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) \, .$$

Interpretation

By the law of large numbers, we have P^∞-almost surely:

$$\mathcal{R}_{L,P}(f) = \lim_{|D| \to \infty} \mathcal{R}_{L,D}(f)$$

Thus, $\mathcal{R}_{L,P}(f)$ is the long-term average future loss when using f.

Performance Evaluation III

Bayes Risk and Bayes Predictor

The Bayes risk is the smallest possible risk

$$\mathcal{R}^*_{L,P} := \inf\{\, \mathcal{R}_{L,P}(f) \mid f : X \to \mathbb{R} \text{ (measurable) } \} \,.$$

A Bayes predictor is any function $f^*_{L,P} : X \to \mathbb{R}$ that satisfies

$$\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P} \,.$$

Performance Evaluation III

Bayes Risk and Bayes Predictor

The Bayes risk is the smallest possible risk

$$\mathcal{R}^*_{L,P} := \inf\{\, \mathcal{R}_{L,P}(f) \mid f : X \to \mathbb{R} \text{ (measurable) } \} .$$

A Bayes predictor is any function $f^*_{L,P} : X \to \mathbb{R}$ that satisfies

$$\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P} .$$

Interpretation

- We will never find a predictor whose risk is smaller than $\mathcal{R}^*_{L,P}$.
- We seek a predictor $f : X \to \mathbb{R}$ whose excess risk

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$$

is close to 0.

Performance Evaluation IV

Best Naïve Risk

The best naïve risk is the smallest risk one obtains by ignoring X:

$$\mathcal{R}^{\dagger}_{L,P} := \inf\{\, \mathcal{R}_{L,P}(c\mathbf{1}_X) \mid c \in \mathbb{R}\} \, .$$

Remarks

- The best naïve risk (and its minimizer) is usually easy to estimate.
- Using fancy learning algorithms only makes sense, if $\mathcal{R}^{*}_{L,P} < \mathcal{R}^{\dagger}_{L,P}$.

Performance Evaluation IV

Best Naïve Risk

The best naïve risk is the smallest risk one obtains by ignoring X:

$$\mathcal{R}^{\dagger}_{L,P} := \inf\{\, \mathcal{R}_{L,P}(c\mathbf{1}_X) \mid c \in \mathbb{R}\} \ .$$

Remarks

- The best naïve risk (and its minimizer) is usually easy to estimate.
- Using fancy learning algorithms only makes sense, if $\mathcal{R}^{*}_{L,P} < \mathcal{R}^{\dagger}_{L,P}$.

Equality

- Typically: $\mathcal{R}^{\dagger}_{L,P} = \mathcal{R}^{*}_{L,P}$ iff there is a constant Bayes predictor.
- If $P = P_X \otimes P_Y$, then $\mathcal{R}^{\dagger}_{L,P} = \mathcal{R}^{*}_{L,P}$, but the converse is false.

Learning Goals I

Binary Classification: $Y = \{-1, 1\}$

- $L(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)$ penalizes predictions t with $\operatorname{sign} t \neq y$.
- $\mathcal{R}_{L,P}(f) = P(\{(x,y) : \operatorname{sign} f(x) \neq y\})$.

Learning Goals I

Binary Classification: $Y = \{-1, 1\}$

- $L(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)$ penalizes predictions t with $\operatorname{sign} t \neq y$.
- $\mathcal{R}_{L,P}(f) = P(\{(x,y) : \operatorname{sign} f(x) \neq y\})$.

Optimal Risk

Let $\eta(x) := P(Y = 1|x)$ be the probability of a positive label at $x \in X$.

- Bayes risk: $\mathcal{R}^*_{L,P} = \mathbb{E}_{P_X} \min\{\eta, 1-\eta\}$.
- f is Bayes predictor iff $(2\eta - 1)\operatorname{sign} f \geq 0$.

Learning Goals I

Binary Classification: $Y = \{-1, 1\}$

- $L(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)$ penalizes predictions t with $\operatorname{sign} t \neq y$.
- $\mathcal{R}_{L,P}(f) = P(\{(x,y) : \operatorname{sign} f(x) \neq y\})$.

Optimal Risk

Let $\eta(x) := P(Y = 1|x)$ be the probability of a positive label at $x \in X$.

- Bayes risk: $\mathcal{R}^*_{L,P} = \mathbb{E}_{P_X} \min\{\eta, 1-\eta\}$.
- f is Bayes predictor iff $(2\eta - 1) \operatorname{sign} f \geq 0$.

Naïve Risk

- Naïve risk: $\mathcal{R}^\dagger_{L,P} = \min\{P(Y=1), 1 - P(Y=1)\}$
- $\mathcal{R}^\dagger_{L,P} = \mathcal{R}^*_{L,P}$ iff $\eta \geq 1/2$ or $\eta \leq 1/2$

Learning Goals II

Least Squares Regression: $Y \subset \mathbb{R}$

- $L(y, t) := (y - t)^2$
- Conditional expectation: $\mu_P(x) := \mathbb{E}_P(Y|x)$.
- Conditional variance: $\sigma_P^2(x) := \mathbb{E}_P(Y^2|x) - \mu^2(x)$.

Learning Goals II

Least Squares Regression: $Y \subset \mathbb{R}$

- $L(y, t) := (y - t)^2$
- Conditional expectation: $\mu_P(x) := \mathbb{E}_P(Y|x)$.
- Conditional variance: $\sigma_P^2(x) := \mathbb{E}_P(Y^2|x) - \mu^2(x)$.

Optimal Risk

- μ_P is the only Bayes predictor and $\mathcal{R}^*_{L,P} = \mathbb{E}_{P_X}\sigma_P^2$.
- Excess risk: $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \|f - \mu_P\|^2_{L_2(P_X)}$.

Least squares regression aims at estimating the conditional mean.

Learning Goals II

Least Squares Regression: $Y \subset \mathbb{R}$

- $L(y, t) := (y - t)^2$
- Conditional expectation: $\mu_P(x) := \mathbb{E}_P(Y|x)$.
- Conditional variance: $\sigma_P^2(x) := \mathbb{E}_P(Y^2|x) - \mu^2(x)$.

Optimal Risk

- μ_P is the only Bayes predictor and $\mathcal{R}^*_{L,P} = \mathbb{E}_{P_X}\sigma_P^2$.
- Excess risk: $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \|f - \mu_P\|^2_{L_2(P_X)}$.

Least squares regression aims at estimating the conditional mean.

Naïve Risk

- Naïve risk: $\mathcal{R}^\dagger_{L,P} = \operatorname{var} P_Y$.

Learning Goals III

Absolute Value Regression: $Y \subset \mathbb{R}$

- $L(y, t) := |y - t|$
- Conditional medians: $m_P(x) := \text{median}_P(Y|x)$.

Learning Goals III

Absolute Value Regression: $Y \subset \mathbb{R}$

- $L(y, t) := |y - t|$
- Conditional medians: $m_P(x) := \text{median}_P(Y|x)$.

Optimal Risk

- The medians m_P are the only Bayes predictors.
- Excess risk: $\mathcal{R}_{L,P}(f_n) - \mathcal{R}^*_{L,P} \to 0$ implies $f_n \to m_P$ in probability P_X.

Absolute value regression aims at estimating the conditional median.

Learning Goals III

Absolute Value Regression: $Y \subset \mathbb{R}$

- $L(y, t) := |y - t|$
- Conditional medians: $m_P(x) := \text{median}_P(Y|x)$.

Optimal Risk

- The medians m_P are the only Bayes predictors.
- Excess risk: $\mathcal{R}_{L,P}(f_n) - \mathcal{R}^*_{L,P} \to 0$ implies $f_n \to m_P$ in probability P_X.

Absolute value regression aims at estimating the conditional median.

Naïve Risk

- Naïve risk: $\mathcal{R}^{\dagger}_{L,P} = \text{median}\, P_Y$.

Questions in Statistical Learning I

Asymptotic Learning

A learning method is called universally consistent if

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_D) = \mathcal{R}^*_{L,P} \qquad \text{in probability } P^\infty \tag{1}$$

for every probability measure P on $X \times Y$.

Questions in Statistical Learning I

Asymptotic Learning

A learning method is called universally consistent if

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_D) = \mathcal{R}^*_{L,P} \qquad \text{in probability } P^\infty \tag{1}$$

for every probability measure P on $X \times Y$.

Good News

Many learning methods are universally consistent.

First result: Stone (1977), AoS

Questions in Statistical Learning II

Learning Rates

A learning method learns for a distribution P with rate $a_n \searrow 0$, if

$$\mathbb{E}_{D\sim P^n}\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + C_P a_n \,, \qquad\qquad n \geq 1.$$

Similar: learning rates in probability.

Questions in Statistical Learning II

Learning Rates

A learning method learns for a distribution P with rate $a_n \searrow 0$, if

$$\mathbb{E}_{D\sim P^n}\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + C_P a_n \,, \qquad\qquad n \geq 1.$$

Similar: learning rates in probability.

Bad News (Devroye, 1982, IEEE TPAMI)

If $|X| = \infty$, $|Y| \geq 2$, and L “non-trivial”, then it is impossible to obtain a learning rate that is independent of P.

Questions in Statistical Learning II

Learning Rates

A learning method learns for a distribution P with rate $a_n \searrow 0$, if

$$\mathbb{E}_{D\sim P^n}\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + C_P a_n\,, \qquad\qquad n \geq 1.$$

Similar: learning rates in probability.

Bad News (Devroye, 1982, IEEE TPAMI)

If $|X| = \infty$, $|Y| \geq 2$, and L "non-trivial", then it is impossible to obtain a learning rate that is independent of P.

Remark

If $|X| < \infty$, then it is usually easy to obtain a uniform learning rate for which C_P depends on $|X|$.

Questions in Statistical Learning III

Relative Learning Rates

- Let $\mathcal{P}$ be a set of distributions on $X \times Y$.
- A learning method learns $\mathcal{P}$ with rate $a_n \searrow 0$, if, for all $P \in \mathcal{P}$,

$$\mathbb{E}_{D\sim P^n}\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + C_P a_n \,, \qquad\qquad n \geq 1.$$

- The rate optimal (a_n) is minmax optimal, if, in addition, there is no learning method that learns $\mathcal{P}$ with a rate (b_n) such that $b_n/a_n \to 0$.

Questions in Statistical Learning III

Relative Learning Rates

- Let $\mathcal{P}$ be a set of distributions on $X \times Y$.
- A learning method learns $\mathcal{P}$ with rate $a_n \searrow 0$, if, for all $P \in \mathcal{P}$,

$$\mathbb{E}_{D\sim P^n}\mathcal{R}_{L,P}(f_D) \leq \mathcal{R}^*_{L,P} + C_P a_n\,, \qquad n \geq 1.$$

- The rate optimal (a_n) is minmax optimal, if, in addition, there is no learning method that learns $\mathcal{P}$ with a rate (b_n) such that $b_n/a_n \to 0$.

Tasks

- Identify interesting ("realistic") classes $\mathcal{P}$ with good optimal rates.
- Find learning algorithms that achieve these rates.

Example of Optimal Rates

Classical Least Squares Example

- $X = [0,1]^d$, $Y = [-1,1]$, L is least squares.
- W^m Sobolev space on X with order of smoothness $m > d/2$.
- $\mathcal{P}$ the set of P such that $f^*_{L,P} \in W^m$ with norm bounded by K.
- Optimal rate is $n^{-\frac{2m}{2m+d}}$.

Example of Optimal Rates

Classical Least Squares Example

- $X = [0,1]^d$, $Y = [-1,1]$, L is least squares.
- W^m Sobolev space on X with order of smoothness $m > d/2$.
- $\mathcal{P}$ the set of P such that $f^*_{L,P} \in W^m$ with norm bounded by K.
- Optimal rate is $n^{-\frac{2m}{2m+d}}$.

Remarks

- The smoother target $\mu = f^*_{L,P}$ is, the better it can be learned.
- The larger the input dimension is, the harder learning becomes.
- There exists various learning algorithms achieving the optimal rate.
- They usually require us to know m in advance.

Questions in Statistical Learning IV

Assumptions for Adaptivity

- Usually one has a familiy $(\mathcal{P}_\theta)_{\theta\in\Theta}$ of large sets $\mathcal{P}_\theta$ of distributions.
- Each set $\mathcal{P}_\theta$ has its own optimal rate.
- We don't know whether $P \in \mathcal{P}_\theta$ for some θ, but we hope so.
- If $P \in \mathcal{P}_\theta$, we don't know θ and we have no mean to estimate it.

Questions in Statistical Learning IV

Assumptions for Adaptivity

- Usually one has a familiy $(\mathcal{P}_\theta)_{\theta\in\Theta}$ of large sets $\mathcal{P}_\theta$ of distributions.
- Each set $\mathcal{P}_\theta$ has its own optimal rate.
- We don't know whether $P \in \mathcal{P}_\theta$ for some θ, but we hope so.
- If $P \in \mathcal{P}_\theta$, we don't know θ and we have no mean to estimate it.

Task

We seek learning algorithms that are

- universally consistent.
- learn all $\mathcal{P}_\theta$ with the optimal rate without knowing θ.

Such learning algorithms are adaptive to the unknown θ.

Questions in Statistical Learning V

Finite Sample Estimates

- Assume that our algorithm has some hyper-parameters $\lambda \in \Lambda$.
- For each P, λ, $\delta \in (0,1)$ and $n \geq 1$ we seek an $\varepsilon(P, \lambda, \delta, n)$ such that

$$\mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \leq \varepsilon(P, \lambda, \delta, n)$$

with probability P^n not smaller than $1 - \delta$.

Questions in Statistical Learning V

Finite Sample Estimates

- Assume that our algorithm has some hyper-parameters $\lambda \in \Lambda$.
- For each P, λ, $\delta \in (0,1)$ and $n \geq 1$ we seek an $\varepsilon(P,\lambda,\delta,n)$ such that

$$\mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \leq \varepsilon(P,\lambda,\delta,n)$$

with probability P^n not smaller than $1-\delta$.

Remarks

- If there exists a sequence (λ_n) with

$$\lim_{n\to\infty} \varepsilon(P,\lambda_n,\delta,n) = 0$$

for all P and δ, then the algorithm can be made universally consistent.
- We automatically obtain learning rates for such sequences.
- If $|X| = \infty$ and ..., then such $\varepsilon(P,\lambda,\delta,n)$ must depend on P.

Questions in Statistical Learning VI

Generalization Error Bounds

- Goal: Estimate risk $\mathcal{R}_{L,P}(f_{D,\lambda})$ by the performance of $f_{D,\lambda}$ on D.
- Find $\varepsilon(\lambda, \delta, n)$ such that with probability P^n not smaller than $1-\delta$:

$$\mathcal{R}_{L,P}(f_{D,\lambda}) \leq \mathcal{R}_{L,D}(f_{D,\lambda}) + \varepsilon(\lambda, \delta, n)\,.$$

Questions in Statistical Learning VI

Generalization Error Bounds

- Goal: Estimate risk $\mathcal{R}_{L,P}(f_{D,\lambda})$ by the performance of $f_{D,\lambda}$ on D.
- Find $\varepsilon(\lambda, \delta, n)$ such that with probability P^n not smaller than $1 - \delta$:

$$\mathcal{R}_{L,P}(f_{D,\lambda}) \leq \mathcal{R}_{L,D}(f_{D,\lambda}) + \varepsilon(\lambda, \delta, n) .$$

Remarks

- $\varepsilon(\lambda, \delta, n)$ must not depend on P since we do not know P.
- $\varepsilon(\lambda, \delta, n)$ can be used to derive parameter selection strategies such as structural risk minimization.
- Alternative: Use second data set D' and $\mathcal{R}_{L,D'}(f_{D,\lambda})$ as an estimate.

Summary

A "good" learning algorithm:

- Is universally consistent.
- Is adaptive for *realistic* classes of distributions.

Summary

A "good" learning algorithm:

- Is universally consistent.
- Is adaptive for *realistic* classes of distributions.
- Can be modified to new problems that have a different loss.
- Has a good record on real-world problems.
- Runs efficiently on a computer.
- ...

Empirical Risk Minimization

Definition

Let $\mathcal{F}$ be a set of functions $X \to \mathbb{R}$. A learning method whose predictors satisfy $f_D \in \mathcal{F}$ and

$$\mathcal{R}_{L,D}(f_D) = \min_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f)$$

is called empirical risk minimization (ERM).

Empirical Risk Minimization

Definition

Let $\mathcal{F}$ be a set of functions $X \to \mathbb{R}$. A learning method whose predictors satisfy $f_D \in \mathcal{F}$ and

$$\mathcal{R}_{L,D}(f_D) = \min_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f)$$

is called empirical risk minimization (ERM).

Remarks

- Not every $\mathcal{F}$ makes ERM possible.
- ERM is, in general, not unique.
- ERM may not be computationally feasible.

Empirical Risk Minimization

Danger of underfitting

- ERM can never produce predictors with risk better than

$$\mathcal{R}^*_{L,P,\mathcal{F}} := \inf\{\mathcal{R}_{L,P}(f) : f \in \mathcal{F}\}\,.$$

- Example: L least squares, $X = [0,1]$, P_X uniform distribution, $f^*_{L,P}$ not linear, and $\mathcal{F}$ set of linear functions, then

$$\mathcal{R}^*_{L,P,\mathcal{F}} > \mathcal{R}^*_{L,P}\,,$$

and thus ERM cannot be consistent.

Empirical Risk Minimization

Danger of overfitting

- If $\mathcal{F}$ is too large, ERM may overfit.
- Example: L least squares, $X = [0,1]$, P_X uniform distribution, $f^*_{L,P} = \mathbf{1}_X$, $\mathcal{R}^*_{L,P} = 0$, and $\mathcal{F}$ set of all functions. Then

$$f_D(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i \\ 0 & \text{otherwise.} \end{cases}$$

satisfies $\mathcal{R}_{L,D}(f_D) = 0$ but $\mathcal{R}_{L,P}(f_D) = 1$.

Summary of Last Session

- Risk of a predictor $f : X \to \mathbb{R}$ is

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L\big(x, y, f(x)\big)\, dP(x, y) \ .$$

- Bayes risk $\mathcal{R}^*_{L,P}$ is the smallest possible risk. A Bayes predictor $f^*_{L,P}$ achieves this minimal risk.

Summary of Last Session

- Risk of a predictor $f : X \to \mathbb{R}$ is

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(x, y, f(x))\, dP(x, y) .$$

- Bayes risk $\mathcal{R}^*_{L,P}$ is the smallest possible risk. A Bayes predictor $f^*_{L,P}$ achieves this minimal risk.
- Learning is

$$\mathcal{R}_{L,P}(f_D) \to \mathcal{R}^*_{L,P}$$

- Asymptotically, this is possible, but no uniform rates are possible.
- We seek adaptive learning algorithms. Ideally, these are fully automated.

Regularized ERM

Definition

Let $\mathcal{F}$ be a non-empty set of functions $X \to \mathbb{R}$ and $\Upsilon : \mathcal{F} \to [0, \infty)$ be a map. A learning method whose predictors satisfy $f_D \in \mathcal{F}$ and

$$\Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) = \inf_{f \in \mathcal{F}} \Bigl(\Upsilon(f) + \mathcal{R}_{L,D}(f) \Bigr)$$

is called regularized empirical risk minimization (RERM).

Regularized ERM

Definition

Let $\mathcal{F}$ be a non-empty set of functions $X \to \mathbb{R}$ and $\Upsilon : \mathcal{F} \to [0, \infty)$ be a map. A learning method whose predictors satisfy $f_D \in \mathcal{F}$ and

$$\Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) = \inf_{f \in \mathcal{F}} \Big(\Upsilon(f) + \mathcal{R}_{L,D}(f) \Big)$$

is called regularized empirical risk minimization (RERM).

Remarks

- $\Upsilon = 0$ yields ERM.
- All remarks about ERM apply to RERM, too.

Examples of Regularized ERM I

General Dictionary Methods

For bounded $h_1, \dots, h_m : X \to \mathbb{R}$ consider

$$\mathcal{F} := \left\{ f_c := \sum_{i=1}^{m} c_i h_i : (c_1, \dots, c_m) \in \mathbb{R}^m \right\},$$

Examples of Regularized ERM I

General Dictionary Methods

For bounded $h_1, \dots, h_m : X \to \mathbb{R}$ consider

$$\mathcal{F} := \left\{ f_c := \sum_{i=1}^{m} c_i h_i : (c_1, \dots, c_m) \in \mathbb{R}^m \right\},$$

Examples of Regularizers

- ℓ_1-regularization: $\Upsilon(f_c) = \lambda \|c\|_1 = \lambda \sum_{i=1}^m |c_i|$,
- ℓ_2-regularization: $\Upsilon(f_c) = \lambda \|c\|_2 = \lambda \sum_{i=1}^m |c_i|^2$,
- ℓ_∞-regularization: $\Upsilon(f_c) = \lambda \|c\|_\infty = \lambda \max_i |c_i|$,

or, in case of dependent h_i, we take the infimum over all representations.

Examples of Regularized ERM II

Further Examples

- Support Vector Machines
- Regularized Decision Trees
- ...

Regularized ERM: Norm Regularizers

Conventions

- Whenever we consider regularizers they will be of the form

$$\Upsilon(f) = \lambda \|f\|_E^\alpha , \qquad\qquad f \in \mathcal{F},$$

where $\alpha \geq 1$ and $E := \mathcal{F}$ is a vector space of functions $X \to \mathbb{R}$.

Regularized ERM: Norm Regularizers

Conventions

- Whenever we consider regularizers they will be of the form

$$\Upsilon(f) = \lambda \|f\|_E^\alpha\,, \qquad\qquad f \in \mathcal{F},$$

where $\alpha \geq 1$ and $E := \mathcal{F}$ is a vector space of functions $X \to \mathbb{R}$.

- In this case, we additionally assume that

$$\|f\|_\infty \leq \|f\|_E\,, \qquad\qquad f \in E.$$

Regularized ERM: Norm Regularizers

Conventions

- Whenever we consider regularizers they will be of the form

$$\Upsilon(f) = \lambda \|f\|_E^\alpha\,, \qquad f \in \mathcal{F},$$

where $\alpha \geq 1$ and $E := \mathcal{F}$ is a vector space of functions $X \to \mathbb{R}$.

- In this case, we additionally assume that

$$\|f\|_\infty \leq \|f\|_E\,, \qquad f \in E.$$

- In the following, we assume that the optimization problem also has a solution f_P, when we replace D by P:

$$f_P \in \arg\min_{f \in \mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,P}(f)$$

The Classical Argument I

Ansatz

- Assume that we have a data set D and an $\varepsilon > 0$ such that

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f) \right| \leq \varepsilon$$

The Classical Argument I

Ansatz

- Assume that we have a data set D and an $\varepsilon > 0$ such that

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f) \right| \leq \varepsilon$$

- Then we obtain

$$\begin{aligned} & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) \\ = \ & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D) + \mathcal{R}_{L,D}(f_D) \end{aligned}$$

The Classical Argument I

Ansatz

- Assume that we have a data set D and an $\varepsilon > 0$ such that

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f) \right| \leq \varepsilon$$

- Then we obtain

$$\begin{aligned} & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) \\ = \; & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D) + \mathcal{R}_{L,D}(f_D) \\ \leq \; & \Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) + \varepsilon \end{aligned}$$

The Classical Argument I

Ansatz

- Assume that we have a data set D and an $\varepsilon > 0$ such that

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f) \right| \leq \varepsilon$$

- Then we obtain

$$\begin{aligned} & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) \\ = \; & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D) + \mathcal{R}_{L,D}(f_D) \\ \leq \; & \Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) + \varepsilon \\ \leq \; & \Upsilon(f_P) + \mathcal{R}_{L,D}(f_P) + \varepsilon \end{aligned}$$

The Classical Argument I

Ansatz

- Assume that we have a data set D and an $\varepsilon > 0$ such that

$$\sup_{f \in \mathcal{F}} \left| \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f) \right| \leq \varepsilon$$

- Then we obtain

$$\begin{aligned}
& \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) \\
= \; & \Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D) + \mathcal{R}_{L,D}(f_D) \\
\leq \; & \Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) + \varepsilon \\
\leq \; & \Upsilon(f_P) + \mathcal{R}_{L,D}(f_P) + \varepsilon \\
\leq \; & \Upsilon(f_P) + \mathcal{R}_{L,P}(f_P) + 2\varepsilon
\end{aligned}$$

The Classical Argument II

Discussion

- The uniform bound

$$\sup_{f\in\mathcal{F}}\bigl|\mathcal{R}_{L,P}(f)-\mathcal{R}_{L,D}(f)\bigr|\leq\varepsilon \tag{2}$$

led to the inequality

$$\Upsilon(f_D)+\mathcal{R}_{L,P}(f_D)-\mathcal{R}^*_{L,P}\leq\Upsilon(f_P)+\mathcal{R}_{L,P}(f_P)-\mathcal{R}^*_{L,P}+2\varepsilon\,.$$

The Classical Argument II

Discussion

- The uniform bound

$$\sup_{f\in\mathcal{F}} \bigl|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\bigr| \leq \varepsilon \qquad (2)$$

led to the inequality

$$\Upsilon(f_D) + \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leq \Upsilon(f_P) + \mathcal{R}_{L,P}(f_P) - \mathcal{R}^*_{L,P} + 2\varepsilon\,.$$

- Since $\Upsilon(f_D) \geq 0$, all what remains to be done, is to estimate
  - the probability of (2)
  - the regularization error $\Upsilon(f_P) + \mathcal{R}_{L,P}(f_P) - \mathcal{R}^*_{L,P}$.

The Classical Argument III

Union Bound

- Assume that $\mathcal{F}$ is finite.
- The union bound gives

$$
\begin{aligned}
& P\Big(D : \sup_{f\in\mathcal{F}} \big|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\big| \leq \varepsilon\Big) \\
&= 1 - P\Big(D : \sup_{f\in\mathcal{F}} \big|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\big| > \varepsilon\Big) \\
&\geq 1 - \sum_{f\in\mathcal{F}} P\Big(D : \big|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\big| > \varepsilon\Big)
\end{aligned}
$$

The Classical Argument III

Union Bound

- Assume that $\mathcal{F}$ is finite.
- The union bound gives

$$\begin{aligned} & P\bigl(D : \sup_{f\in\mathcal{F}} \bigl|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\bigr| \leq \varepsilon\bigr) \\ &= 1 - P\bigl(D : \sup_{f\in\mathcal{F}} \bigl|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\bigr| > \varepsilon\bigr) \\ &\geq 1 - \sum_{f\in\mathcal{F}} P\bigl(D : \bigl|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\bigr| > \varepsilon\bigr) \end{aligned}$$

Consequences

- It suffices to bound $P\bigl(D : \bigl|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\bigr| > \varepsilon\bigr)$ for all f.
- No assumptions on P are made so far. In particular, so far data D does not need to be i.i.d. nor even random.

The Classical Argument IV

Hoeffding's Inequality

Let $(\Omega, \mathcal{A}, Q)$ be a probability space and $\xi_1, \ldots, \xi_n : \Omega \to [a, b]$ be independent random variables. Then, for all $\tau > 0$, $n \geq 1$, we have

$$Q\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}_Q \xi_i) \right| \geq (b-a) \sqrt{\frac{\tau}{2n}} \right) \leq 2e^{-\tau}\,.$$

The Classical Argument IV

Hoeffding's Inequality

Let $(\Omega, \mathcal{A}, Q)$ be a probability space and $\xi_1, \dots, \xi_n : \Omega \to [a, b]$ be independent random variables. Then, for all $\tau > 0$, $n \geq 1$, we have

$$Q\left(\left| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}_Q \xi_i) \right| \geq (b-a) \sqrt{\frac{\tau}{2n}} \right) \leq 2e^{-\tau} .$$

Application

- Consider $\Omega := (X \times Y)^n$ and $Q := P^n$.
- For $\xi_i(D) := L(x_i, y_i, f(x_i))$ we have $a = 0$ and

$$\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}_{P^n} \xi_i) = \mathcal{R}_{L,D}(f) - \mathcal{R}_{L,P}(f) .$$

- Assuming $L(x, y, f(x)) \leq B$ makes application of Hoeffding possible.

The Classical Argument V

Theorem for ERM

Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, $\mathcal{F}$ be a non-empty finite set of functions $f : X \to \mathbb{R}$, and $B > 0$ be a constant such that

$$L(x, y, f(x)) \leq B\,, \qquad\qquad (x, y) \in X \times Y,\ f \in \mathcal{F}.$$

The Classical Argument V

Theorem for ERM

Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, $\mathcal{F}$ be a non-empty finite set of functions $f : X \to \mathbb{R}$, and $B > 0$ be a constant such that

$$L(x, y, f(x)) \leq B\,, \qquad (x, y) \in X \times Y,\ f \in \mathcal{F}.$$

Then we have

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{F}|)}{n}}\,\right) \geq 1 - e^{-\tau}\,.$$

Remarks

- Does not specify approximation error $\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}$.
- If $|\mathcal{F}| = \infty$, the bound becomes meaningless.
- What happens, if we consider RERM with non-trivial regularizer?

ERM for Infinite $\mathcal{F}$: The General Approach

So far . . .

The union bound was the "trick" to make a conclusion from an estimate of

$$\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \geq \varepsilon$$

for a single f to all $f \in \mathcal{F}$. For infinite $\mathcal{F}$, this does not work!

ERM for Infinite $\mathcal{F}$: The General Approach

So far ...

The union bound was the “trick” to make a conclusion from an estimate of

$$\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \geq \varepsilon$$

for a single f to all $f \in \mathcal{F}$. For infinite $\mathcal{F}$, this does not work!

General Approach

Given some $\delta > 0$, find a finite $\mathcal{N}_\delta$ set of functions such that

$$\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \leq \sup_{f\in\mathcal{N}_\delta}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| + \delta$$

Then apply the union bound for $\mathcal{N}_\delta$. The rest remains unchanged.

ERM for Infinite $\mathcal{F}$: The General Approach

The old inequality

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{F}|)}{n}}\right) \geq 1 - e^{-\tau}\,.$$

The new inequality

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{N}_\delta|)}{n}} + \delta\right) \geq 1 - e^{-\tau}\,.$$

ERM for Infinite $\mathcal{F}$: The General Approach

The old inequality

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{F}|)}{n}}\right) \geq 1 - e^{-\tau}\,.$$

The new inequality

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{N}_\delta|)}{n}} + \delta\right) \geq 1 - e^{-\tau}\,.$$

Tasks

- For each $\delta > 0$, find a small set $\mathcal{N}_\delta$.
- Optimize the right-hand side wrt. δ.

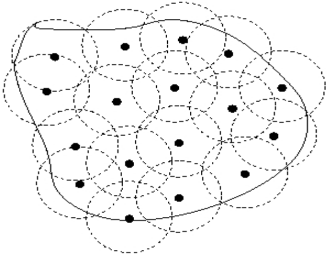
Covering Numbers

Definition

Let (M, d) be a metric space, $A \subset M$, and $\varepsilon > 0$. The ε-covering number of A is defined by

$$\mathcal{N}(A, d, \varepsilon) := \inf\left\{ n \geq 1 : \exists\, x_1, \ldots, x_n \in M \text{ such that } A \subset \bigcup_{i=1}^{n} B_d(x_i, \varepsilon) \right\}$$

where $\inf \emptyset := \infty$, and $B_d(x_i, \varepsilon)$ is the ball with radius ε and center x_i.



Covering Numbers

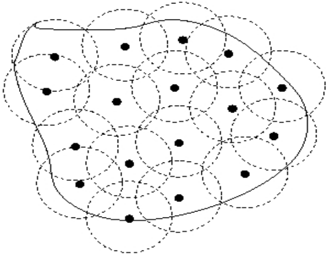
Definition

Let (M, d) be a metric space, $A \subset M$, and $\varepsilon > 0$. The ε-covering number of A is defined by

$$\mathcal{N}(A, d, \varepsilon) := \inf\left\{ n \geq 1 : \exists\, x_1, \ldots, x_n \in M \text{ such that } A \subset \bigcup_{i=1}^{n} B_d(x_i, \varepsilon) \right\}$$

where $\inf \emptyset := \infty$, and $B_d(x_i, \varepsilon)$ is the ball with radius ε and center x_i.

- $x_1, \ldots, x_n$ is called an ε-net.
- $\mathcal{N}(A, d, \varepsilon)$ is the size of the smallest ε-net.



Covering Numbers II

- Every bounded $A \subset \mathbb{R}^d$ satisfies

$$\mathcal{N}(A, \|\cdot\|, \varepsilon) \leq c\varepsilon^{-d}, \qquad \varepsilon > 0$$

where $c > 0$ is a constant and the norm $\|\cdot\|$ does only influence c.

Covering Numbers II

- Every bounded $A \subset \mathbb{R}^d$ satisfies

$$\mathcal{N}(A, \|\cdot\|, \varepsilon) \leq c\varepsilon^{-d}, \qquad \varepsilon > 0$$

  where $c > 0$ is a constant and the norm $\|\cdot\|$ does only influence c.
- For sets $\mathcal{F}$ of functions $f : X \to \mathbb{R}$, the behavior of $\mathcal{N}(\mathcal{F}, \|\cdot\|, \varepsilon)$ may be very different!
- The literature is full of estimates of $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|, \varepsilon)$.
- A typical estimate looks like

$$\ln \mathcal{N}(B_E, \|\cdot\|_F, \varepsilon) \leq c\varepsilon^{-2p}, \qquad \varepsilon > 0$$

  Here p may depend on the input dimension and the smoothness of the functions in E.

ERM with Infinite Sets

Theorem

- Let L be Lipschitz in its third argument, Lipschitz constant $= 1$.
- Assume that $\|L \circ f\|_\infty \leq B$ for all $f \in \mathcal{F}$.
- Let $\mathcal{N}_\varepsilon$ be a minimal ε-net of $\mathcal{F}$, i.e. $|\mathcal{N}_\varepsilon| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$.

ERM with Infinite Sets

Theorem

- Let L be Lipschitz in its third argument, Lipschitz constant $= 1$.
- Assume that $\|L \circ f\|_\infty \leq B$ for all $f \in \mathcal{F}$.
- Let $\mathcal{N}_\varepsilon$ be a minimal ε-net of $\mathcal{F}$, i.e. $|\mathcal{N}_\varepsilon| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$.

Then we have

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{N}_\varepsilon|)}{n}} + 2\varepsilon\right) \geq 1 - e^{-\tau}\,.$$

Using Covering Numbers VII

Example

- Let L satisfy assumptions on previous theorem.
- Let $\mathcal{F}$ set of functions with $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq c\varepsilon^{-2p}$.

Using Covering Numbers VII

Example

- Let L satisfy assumptions on previous theorem.
- Let $\mathcal{F}$ set of functions with $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq c\varepsilon^{-2p}$.
- Then we have

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 4c\varepsilon^{-2p}}{n}} + 2\varepsilon\right) \geq 1 - e^{-\tau}.$$

Using Covering Numbers VII

Example

- Let L satisfy assumptions on previous theorem.
- Let $\mathcal{F}$ set of functions with $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq c\varepsilon^{-2p}$.
- Then we have

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 4c\varepsilon^{-2p}}{n}} + 2\varepsilon \right) \geq 1 - e^{-\tau} .$$

- Optimizing wrt. ε gives a constant K_p such that

$$P^n\left(D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + K_p c^{\frac{1}{2+2p}} B\sqrt{\tau} n^{-\frac{1}{2+2p}} \right) \geq 1 - e^{-\tau} .$$

- For ERM over finite $\mathcal{F}$, we had "$p = 0$".

Standard Analysis for RERM

Difficulties when Analyzing RERM

- We are interested in RERMs, where $\mathcal{F}$ is a vector space E.
- Vector spaces E are never compact, thus $\ln \mathcal{N}(E, \|\cdot\|_\infty, \varepsilon) = \infty$.
- It seems that our approach does not work in this case.

Standard Analysis for RERM

Difficulties when Analyzing RERM

- We are interested in RERMs, where $\mathcal{F}$ is a vector space E.
- Vector spaces E are never compact, thus $\ln \mathcal{N}(E, \|\cdot\|_\infty, \varepsilon) = \infty$.
- It seems that our approach does not work in this case.

Solution

RERM actually solves its optimization problem

$$\Upsilon(f_D) + \mathcal{R}_{L,D}(f_D) = \inf_{f \in E} \Big(\Upsilon(f) + \mathcal{R}_{L,D}(f) \Big)$$

over a set, which is significantly smaller than E.

Norm Bound for RERM

Lemma

Assume that $L(x, y, 0) \leq 1$. Then, for any RERM predictor $f_{D,\lambda} \in E$ we have

$$\|f_{D,\lambda}\|_E \leq \lambda^{-1/\alpha}\,.$$

Norm Bound for RERM

Lemma

Assume that $L(x, y, 0) \leq 1$. Then, for any RERM predictor $f_{D,\lambda} \in E$ we have

$$\|f_{D,\lambda}\|_E \leq \lambda^{-1/\alpha}.$$

Consequence

RERM optimization problem is actually solved over the ball with radius

$$\lambda^{-1/\alpha}.$$

Norm Bound for RERM II

Proof

Our assumptions $L(x, y, t) \geq 0$ and $L(x, y, 0) \leq 1$ yield

Norm Bound for RERM II

Proof

Our assumptions $L(x, y, t) \geq 0$ and $L(x, y, 0) \leq 1$ yield

$$\lambda \|f_{D,\lambda}\|_E^\alpha \leq \lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,D}(f_{D,\lambda})$$

Norm Bound for RERM II

Proof

Our assumptions $L(x, y, t) \geq 0$ and $L(x, y, 0) \leq 1$ yield

$$\begin{aligned} \lambda \|f_{D,\lambda}\|_E^\alpha &\leq \lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,D}(f_{D,\lambda}) \\ &= \inf_{f \in E} \Bigl(\lambda \|f\|_E^\alpha + \mathcal{R}_{L,D}(f) \Bigr) \end{aligned}$$

Norm Bound for RERM II

Proof

Our assumptions $L(x, y, t) \geq 0$ and $L(x, y, 0) \leq 1$ yield

$$\begin{aligned} \lambda \|f_{D,\lambda}\|_E^\alpha &\leq \lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,D}(f_{D,\lambda}) \\ &= \inf_{f \in E} \Big(\lambda \|f\|_E^\alpha + \mathcal{R}_{L,D}(f) \Big) \\ &\leq \lambda \|0\|_E^\alpha + \mathcal{R}_{L,D}(0) \end{aligned}$$

Norm Bound for RERM II

Proof

Our assumptions $L(x, y, t) \geq 0$ and $L(x, y, 0) \leq 1$ yield

$$\begin{aligned} \lambda \|f_{D,\lambda}\|_E^\alpha &\leq \lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,D}(f_{D,\lambda}) \\ &= \inf_{f \in E} \Big(\lambda \|f\|_E^\alpha + \mathcal{R}_{L,D}(f) \Big) \\ &\leq \lambda \|0\|_E^\alpha + \mathcal{R}_{L,D}(0) \\ &\leq 1 \,. \end{aligned}$$

An Oracle Inequality

Theorem (Example)

- L Lipschitz continuous with $|L|_1 \leq 1$ and $L(x, y, 0) \leq 1$.
- E vector space with norm $\|\cdot\|_E$ satisfying $\|\cdot\|_\infty \leq \|\cdot\|_E$.
- $\Upsilon(f) = \lambda \|f\|_E^\alpha$.
- We have $\ln \mathcal{N}(B_E, \|\cdot\|_\infty, \varepsilon) \leq c\varepsilon^{-2p}$

An Oracle Inequality

Theorem (Example)

- L Lipschitz continuous with $|L|_1 \leq 1$ and $L(x, y, 0) \leq 1$.
- E vector space with norm $\|\cdot\|_E$ satisfying $\|\cdot\|_\infty \leq \|\cdot\|_E$.
- $\Upsilon(f) = \lambda \|f\|_E^\alpha$.
- We have $\ln \mathcal{N}(B_E, \|\cdot\|_\infty, \varepsilon) \leq c\varepsilon^{-2p}$

Then, for all $n \geq 1$, $\lambda \in (0, 1]$, $\tau \geq 1$, we have

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) < \lambda \|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda}) + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

with probability P^n not less than $1 - e^{-\tau}$.

Consequences of the Oracle Inequality

Oracle inequality

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) < \lambda \|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda})$$

$$+ K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

Consequences of the Oracle Inequality

Oracle inequality

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,P}^* \quad < \quad \lambda \|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P}^*$$

$$+ K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

Consequences of the Oracle Inequality

Oracle inequality

$$\begin{aligned}\lambda\|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,P}^* \quad < \quad & \lambda\|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,E}^* \\ & + \mathcal{R}_{L,P,E}^* - \mathcal{R}_{L,P}^* \\ & + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}\end{aligned}$$

Consequences of the Oracle Inequality

Oracle inequality

$$\begin{aligned}\lambda\|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,P}^* \quad < \quad & \lambda\|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,E}^* \\ & + \mathcal{R}_{L,P,E}^* - \mathcal{R}_{L,P}^* \\ & + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}\end{aligned}$$

- Regularization error: $A(\lambda) := \lambda\|f_{P,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}_{L,P,E}^*$
- Approximation error: $\mathcal{R}_{L,P,E}^* - \mathcal{R}_{L,P}^*$.
- Statistical error: $K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$.

Bounding the Remaining Errors

Lemma 1

If E is dense in $L_1(P_X)$, then $\mathcal{R}^*_{L,P,E} - \mathcal{R}^*_{L,P} = 0$.

Lemma 2

We have $\lim_{\lambda \to 0} A(\lambda) = 0$, and if there is an $f^* \in E$ with $\mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P,E}$, then

$$A(\lambda) \leq \lambda \|f^*\|_E^\alpha .$$

Bounding the Remaining Errors

Lemma 1
If E is dense in $L_1(P_X)$, then $\mathcal{R}^*_{L,P,E} - \mathcal{R}^*_{L,P} = 0$.

Lemma 2
We have $\lim_{\lambda\to 0} A(\lambda) = 0$, and if there is an $f^* \in E$ with $\mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P,E}$, then

$$A(\lambda) \leq \lambda \|f^*\|^\alpha_E \,.$$

Remarks

- A linear behaviour of A often requires such an f^*.
- A typical behavior is, for some $\beta \in (0,1]$, of the form

$$A(\lambda) \leq c\lambda^\beta$$

- A sufficient condition for such a behaviour can be described with the help of so-called "interpolation spaces of the real method".

Main Results for RERM

Oracle inequality

We assume $\mathcal{R}^*_{L,P,E} - \mathcal{R}^*_{L,P} = 0$.

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \quad < \quad A(\lambda) + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

Main Results for RERM

Oracle inequality
We assume $\mathcal{R}^*_{L,P,E} - \mathcal{R}^*_{L,P} = 0$.

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \quad < \quad A(\lambda) + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

Consequences

- Consistent, if $\lambda_n \to 0$ with $\lambda_n n^{\frac{\alpha}{2+2p}} \to \infty$.

Main Results for RERM

Oracle inequality

We assume $\mathcal{R}^*_{L,P,E} - \mathcal{R}^*_{L,P} = 0$.

$$\lambda \|f_{D,\lambda}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \quad < \quad A(\lambda) + K_p c^{\frac{1}{2+2p}} \sqrt{\tau} \lambda^{-\frac{1}{\alpha}} n^{-\frac{1}{2+2p}}$$

Consequences

- Consistent, if $\lambda_n \to 0$ with $\lambda_n n^{\frac{\alpha}{2+2p}} \to \infty$.
- If $A(\lambda) \leq c\lambda^\beta$, then

$$\lambda_n \sim n^{-\frac{\alpha}{(\alpha\beta+1)(2+2p)}}$$

achieves "best" rate

$$n^{-\frac{\alpha\beta}{(\alpha\beta+1)(2+2p)}}$$

Main Results for ERM II

Discussion

- Assumptions for consistency on E are minimal.
- More sophisticated algorithms can be devised from oracle inequality. For example, E could change with sample size, too.
- To achieve best learning rates, we need to know β.

Learning Rates: Hyper-Parameters III

Training-Validation Approach

Assume that L is clippable.

- Split data into equally sized parts D_1 and D_2. We write $m := n/2$.
- Fix a finite set $\Lambda \subset (0, 1]$ of candidate values for λ.
- For each $\lambda \in \Lambda$ compute $f_{D_1,\lambda}$.
- Pick the $\lambda_{D_2} \in \Lambda$ such that $\bar{f}_{D_1,\lambda_{D_2}}$ minimizes empirical risk $\mathcal{R}_{L,D_2}$.

Learning Rates: Hyper-Parameters III

Training-Validation Approach

Assume that L is clippable.

- Split data into equally sized parts D_1 and D_2. We write $m := n/2$.
- Fix a finite set $\Lambda \subset (0, 1]$ of candidate values for λ.
- For each $\lambda \in \Lambda$ compute $f_{D_1,\lambda}$.
- Pick the $\lambda_{D_2} \in \Lambda$ such that $\bar{f}_{D_1,\lambda_{D_2}}$ minimizes empirical risk $\mathcal{R}_{L,D_2}$.

Observation

Approach performs RERM on D_1 and ERM over $\mathcal{F} := \{\bar{f}_{D_1,\lambda} : \lambda \in \Lambda\}$ on D_2.

Learning Rates: Hyper-Parameters VI

Theorem

If Λ_n is a polynomially growing $n^{-\alpha/2}$-net of $(0, 1]$, our TV-RERM is consistent and enjoys the same best rates as RERM without knowing β.

Summary

Positive Aspects

- Finite sample estimates in forms of oracle inequalities.
- Consistency and learning rates.
- Adaptivity to best learning rates the analysis can provide.
- Framework applies to a variety of algorithms, e.g. SVMs with Gaussian kernels.
- Analysis is very robust to changes in the scenario.

Summary

Positive Aspects

- Finite sample estimates in forms of oracle inequalities.
- Consistency and learning rates.
- Adaptivity to best learning rates the analysis can provide.
- Framework applies to a variety of algorithms, e.g. SVMs with Gaussian kernels.
- Analysis is very robust to changes in the scenario.

Negative Aspect

- For RERM, the rates are never optimal!
- This analysis is out-dated.

Learning Rates: Non-Optimality I

- For RERM, with probability P^n not less than $1 - e^{-\tau}$ we have

$$\lambda_n \|f_{D,\lambda_n}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}^*_{L,P} \leq C\sqrt{\tau}\, n^{-\frac{\alpha\beta}{2(\alpha\beta+1)(1+p)}} \,. \tag{3}$$

Learning Rates: Non-Optimality I

- For RERM, with probability P^n not less than $1 - e^{-\tau}$ we have

$$\lambda_n \|f_{D,\lambda_n}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}_{L,P}^* \leq C\sqrt{\tau}\, n^{-\frac{\alpha\beta}{2(\alpha\beta+1)(1+p)}} . \tag{3}$$

- In the proof of this result we used $\lambda_n \|f_{D,\lambda_n}\|_E^\alpha \leq 1$, but (3) shows

$$\lambda_n \|f_{D,\lambda_n}\|_E^2 \leq C\sqrt{\tau}\, n^{-\frac{\alpha\beta}{2(\alpha\beta+1)(1+p)}} .$$

For large n this estimate is sharper!

Learning Rates: Non-Optimality I

- For RERM, with probability P^n not less than $1 - e^{-\tau}$ we have

$$\lambda_n \|f_{D,\lambda_n}\|_E^\alpha + \mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}_{L,P}^* \leq C\sqrt{\tau}\, n^{-\frac{\alpha\beta}{2(\alpha\beta+1)(1+p)}} . \tag{3}$$

- In the proof of this result we used $\lambda_n \|f_{D,\lambda_n}\|_E^\alpha \leq 1$, but (3) shows

$$\lambda_n \|f_{D,\lambda_n}\|_E^2 \leq C\sqrt{\tau}\, n^{-\frac{\alpha\beta}{2(\alpha\beta+1)(1+p)}} .$$

  For large n this estimate is sharper!

- Using the sharper estimate in the proof, we obtain a better learning rate.
- Argument can be iterated ...

Learning Rates: Non-Optimality II

Bernstein's Inequality
Let $(\Omega, \mathcal{A}, Q)$ be a probability space and $\xi_1, \dots, \xi_n : \Omega \to [-B, B]$ be independent random variables satisfying

- $\mathbb{E}_Q \xi_i = 0$
- $\mathbb{E}_Q \xi_i^2 \leq \sigma^2$

Then, for all $\tau > 0$, $n \geq 1$, we have

$$Q\left(\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| \geq \sqrt{\frac{2\sigma^2 \tau}{n}} + \frac{2B\tau}{3n} \right) \leq 2e^{-\tau} .$$

Learning Rates: Non-Optimality III

- Some loss functions or distributions allow a variance bound

$$\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^2 \leq V\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big)^\vartheta .$$

Learning Rates: Non-Optimality III

- Some loss functions or distributions allow a variance bound

$$\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^2 \leq V\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big)^\vartheta .$$

- Use Bernstein's inequality rather than Hoeffding's inequality leads to an oracle inequality with a
  - variance term, which is $O(n^{-1/2})$
  - supremum term, which is $O(n^{-1})$

Learning Rates: Non-Optimality III

- Some loss functions or distributions allow a variance bound

$$\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^2 \leq V\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big)^{\vartheta} .$$

- Use Bernstein's inequality rather than Hoeffding's inequality leads to an oracle inequality with a
  - variance term, which is $O(n^{-1/2})$
  - supremum term, which is $O(n^{-1})$
- Iteration in the proof:
  - Initial analysis provides small excess risk with high probability
  - Variance bound converts small excess risk into small variance
  - Variance term in oracle inequality becomes smaller, leading to a faster rate
  - ...

Learning Rates: Non-Optimality III

- Some loss functions or distributions allow a variance bound

$$\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^2 \leq V\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big)^\vartheta .$$

- Use Bernstein's inequality rather than Hoeffding's inequality leads to an oracle inequality with a
  - variance term, which is $O(n^{-1/2})$
  - supremum term, which is $O(n^{-1})$
- Iteration in the proof:
  - Initial analysis provides small excess risk with high probability
  - Variance bound converts small excess risk into small variance
  - Variance term in oracle inequality becomes smaller, leading to a faster rate
  - ...
- Rates up to $O(n^{-1})$ become possible. Iteration can be avoided!

Learning Rates: Non-Optimality IV

Further Reasons

- The fact that L is clippable, should be used to obtain a smaller supremum term.
- $\|\cdot\|_\infty$-covering numbers provide a worst-case tool.

Adaptivity of Standard SVMs

Theorem (Eberts & S. 2011)

- Consider an SVM with least squares loss and Gaussian kernel k_σ.
- Pick λ and σ by a suitable training/validation approach.

Then, for $m \in (d/2, \infty)$, the SVM learns every $f^*_{L,P} \in W^m(X)$ with the (essentially) optimal rate $n^{-\frac{2m}{2m+d}+\varepsilon}$ without knowing m.

Towards a Better Analysis for ERM I

Basic Setup

- We consider ERM over finite $\mathcal{F}$.
- We assume that a Bayes predictor $f^*_{L,P}$ exists.
- We consider excess losses

$$h_f := L \circ f - L \circ f^*_{L,P} .$$

Thus $\mathbb{E}_P h_f = \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$.

Towards a Better Analysis for ERM I

Basic Setup

- We consider ERM over finite $\mathcal{F}$.
- We assume that a Bayes predictor $f^*_{L,P}$ exists.
- We consider excess losses

$$h_f := L \circ f - L \circ f^*_{L,P} .$$

  Thus $\mathbb{E}_P h_f = \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$.
- Variance bound: $\mathbb{E}_P h_f^2 \leq V(\mathbb{E}_P h_f)^\vartheta$
- Supremum bound: $\|h_f\|_\infty \leq B$

Towards a Better Analysis for ERM II

Decomposition

- Let $f_P \in \mathcal{F}$ satisfy $\mathcal{R}_{L,P}(f_P) = \mathcal{R}^*_{L,P,\mathcal{F}}$.
- $\mathcal{R}_{L,D}(f_D) \leq \mathcal{R}_{L,D}(f_P)$ implies $\mathbb{E}_D h_{f_D} \leq \mathbb{E}_D h_{f_P}$.

Towards a Better Analysis for ERM II

Decomposition

- Let $f_P \in \mathcal{F}$ satisfy $\mathcal{R}_{L,P}(f_P) = \mathcal{R}^*_{L,P,\mathcal{F}}$.
- $\mathcal{R}_{L,D}(f_D) \le \mathcal{R}_{L,D}(f_P)$ implies $\mathbb{E}_D h_{f_D} \le \mathbb{E}_D h_{f_P}$.

This yields

$$\begin{aligned} \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,P}(f_P) &= \mathbb{E}_P h_{f_D} - \mathbb{E}_P h_{f_P} \\ &\le \mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D} + \mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} \end{aligned}$$

We will estimate the two differences separately.

Towards a Better Analysis for ERM III

Second Difference

We have $\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} = \mathbb{E}_D(h_{f_P} - \mathbb{E}_P h_{f_P})$.

Towards a Better Analysis for ERM III

Second Difference

We have $\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} = \mathbb{E}_D(h_{f_P} - \mathbb{E}_P h_{f_P})$.

- Centered: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P}) = 0$.
- Variance bound: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P})^2 \leq \mathbb{E}_P h_{f_P}^2 \leq V(\mathbb{E}_P h_{f_P})^\vartheta$
- Supremum bound: $\|h_{f_P} - \mathbb{E}_P h_{f_P}\|_\infty \leq 2B$

Towards a Better Analysis for ERM III

Second Difference

We have $\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} = \mathbb{E}_D(h_{f_P} - \mathbb{E}_P h_{f_P})$.

- Centered: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P}) = 0$.
- Variance bound: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P})^2 \leq \mathbb{E}_P h_{f_P}^2 \leq V(\mathbb{E}_P h_{f_P})^\vartheta$
- Supremum bound: $\|h_{f_P} - \mathbb{E}_P h_{f_P}\|_\infty \leq 2B$

Bernstein yields

$$\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} \quad \leq \quad \sqrt{\frac{2\tau V(\mathbb{E}_P h_{f_P})^\vartheta}{n}} + \frac{4B\tau}{3n}$$

Towards a Better Analysis for ERM III

Second Difference

We have $\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} = \mathbb{E}_D(h_{f_P} - \mathbb{E}_P h_{f_P})$.

- Centered: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P}) = 0$.
- Variance bound: $\mathbb{E}_P(h_{f_P} - \mathbb{E}_P h_{f_P})^2 \leq \mathbb{E}_P h_{f_P}^2 \leq V(\mathbb{E}_P h_{f_P})^\vartheta$
- Supremum bound: $\|h_{f_P} - \mathbb{E}_P h_{f_P}\|_\infty \leq 2B$

Bernstein yields

$$\begin{aligned}
\mathbb{E}_D h_{f_P} - \mathbb{E}_P h_{f_P} &\leq \sqrt{\frac{2\tau V(\mathbb{E}_P h_{f_P})^\vartheta}{n}} + \frac{4B\tau}{3n} \\
&\leq \mathbb{E}_P h_{f_P} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n}\,.
\end{aligned}$$

Towards a Better Analysis for ERM IV

First Difference

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}, \qquad f \in \mathcal{F},\ r > 0.$$

Towards a Better Analysis for ERM IV

First Difference

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}\,, \qquad\qquad f \in \mathcal{F},\ r > 0\,.$$

Bernstein Conditions

- Centered: $\mathbb{E}_P g_{f,r} = 0$.

Towards a Better Analysis for ERM IV

First Difference

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}, \qquad f \in \mathcal{F},\ r > 0 .$$

Bernstein Conditions

- Centered: $\mathbb{E}_P g_{f,r} = 0$.
- Variance bound:

$$\mathbb{E}_P g_{f,r}^2 \leq \frac{\mathbb{E}_P h_f^2}{(\mathbb{E}_P h_f + r)^2} \leq \frac{\mathbb{E}_P h_f^2}{r^{2-\vartheta}(\mathbb{E}_P h_f)^\vartheta} \leq V r^{\vartheta-2} .$$

Towards a Better Analysis for ERM IV

First Difference

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}, \qquad f \in \mathcal{F},\ r > 0\,.$$

Bernstein Conditions

- Centered: $\mathbb{E}_P g_{f,r} = 0$.
- Variance bound:

$$\mathbb{E}_P g_{f,r}^2 \le \frac{\mathbb{E}_P h_f^2}{(\mathbb{E}_P h_f + r)^2} \le \frac{\mathbb{E}_P h_f^2}{r^{2-\vartheta}(\mathbb{E}_P h_f)^\vartheta} \le V r^{\vartheta-2}\,.$$

- Supremum bound: $\|g_{f,r}\|_\infty \le \|\mathbb{E}_P h_f - h_f\|_\infty r^{-1} \le 2Br^{-1}$.

Towards a Better Analysis for ERM V

Application of Bernstein

With probability P^n not smaller than $1 - |\mathcal{F}|e^{-\tau}$ we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}$$

Towards a Better Analysis for ERM V

Application of Bernstein

With probability P^n not smaller than $1 - |\mathcal{F}|e^{-\tau}$ we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}$$

Transformation

The definition of $g_{f_D,r}$ and $f_D \in \mathcal{F}$ imply

$$\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D} < \mathbb{E}_P h_{f_D} \left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr} \right) + \sqrt{\frac{2V\tau r^{\vartheta}}{n}} + \frac{4B\tau}{3n} \, .$$

Towards a Better Analysis for ERM VI

Combination of the three Estimates

$$\begin{aligned}\mathbb{E}_P h_{f_D} - \mathbb{E}_P h_{f_P} \quad &< \quad \mathbb{E}_P h_{f_P} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{8B\tau}{3n} \\ &\quad + \mathbb{E}_P h_{f_D}\left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^{\vartheta}}{n}}\end{aligned}$$

Towards a Better Analysis for ERM VI

Combination of the three Estimates

$$
\begin{aligned}
\mathbb{E}_P h_{f_D} - \mathbb{E}_P h_{f_P} \quad < \quad & \mathbb{E}_P h_{f_P} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{8B\tau}{3n} \\
& + \mathbb{E}_P h_{f_D} \left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr} \right) + \sqrt{\frac{2V\tau r^{\vartheta}}{n}}
\end{aligned}
$$

Transformation

$$
\left(1 - \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} - \frac{4B\tau}{3nr}\right)\mathbb{E}_P h_{f_D} < 2\mathbb{E}_P h_{f_P} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{8B\tau}{3n} + \sqrt{\frac{2V\tau r^{\vartheta}}{n}}
$$

Final Step

For $r := \left(\frac{8V\tau}{n}\right)^{1/(2-\vartheta)}$, the factor on the lhs. is not smaller than $1/3$.

A Better Oracle Inequality for ERM

Theorem

Assume that there are $\vartheta \in [0,1]$, and $V \geq B^{2-\vartheta}$ such that

- $\mathcal{F}$ finite set of functions.
- Variance bound: $\mathbb{E}_P\big(L \circ f - L \circ f^*_{L,P}\big)^2 \leq V \cdot \big(\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})\big)^\vartheta$
- Supremum bound: $\|L \circ f - L \circ f^*_{L,P}\|_\infty \leq B$

A Better Oracle Inequality for ERM

Theorem

Assume that there are $\vartheta \in [0,1]$, and $V \geq B^{2-\vartheta}$ such that

- $\mathcal{F}$ finite set of functions.
- Variance bound: $\mathbb{E}_P\big(L \circ f - L \circ f^*_{L,P}\big)^2 \leq V \cdot \big(\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})\big)^\vartheta$
- Supremum bound: $\|L \circ f - L \circ f^*_{L,P}\|_\infty \leq B$

Then, for $\tau > 0$ and $n \geq 1$, we have with probability P^n not less than $1 - e^{-\tau}$:

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} < 6\big(\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}\big) + 4\left(\frac{8V\big(\tau + \ln(1 + |\mathcal{F}|)\big)}{n}\right)^{\frac{1}{2-\vartheta}} .$$