A 60 minute tour of Statistical Learning Theory

Theja Tulabandhula
Massachusetts Institute of Technology

Sourced from Shawe-Taylor, Kakade, Bousquet, Tewari, Mendelson, Rossi and various course pages.

WHAT IS THE BIG QUESTION?

- One of the important questions in science:
  What is learning?
  - Can a machine read a chapter from [Your favorite textbook] and answer questions at the back of the chapter?
- There are partial answers to "what is learning?"
- Such answers give algorithms.
- Given algorithms, theory
  - helps in measuring their quality.
  - helps design even better ones.
- Theory also more importantly,
  - helps formalize what is meant by learning.
  - helps understand what assumptions are made.

SCOPE OF THIS TALK

- Tools and techniques of Statistical Learning Theory (SLT)
- Building on probability, linear algebra, calculus.
- A whirlwind guided tour. No proofs or proof sketches.
- Avoid details of machine learning (is this allowed?)

GOAL

Understand assumptions we make while learning and know what guarantees we get via a statistical point of view.

INTRODUCTION

WHY A THEORY

- Model real phenomena so we can understand its properties.
- Then we can make predictions and know when they work
- SLT is one of many such theories
  - Others: Bayesian inference, Statistical physics, Mainstream statistics, Game theory, Valiant's PAC learning theory
- Every theory makes assumptions.
- Every theory has strengths and weaknesses.
- The better it matches real world, the better it is for our use.

LEARNING PHENOMENON

- Log Data $\rightarrow$ Build model $\rightarrow$ Predict future
  - Example: Supervised learning.
  - feature $x \in \mathcal{X}$, example: $\subset \mathbb{R}^p$ for regression.
  - label $y \in \mathcal{Y}$, example $\subset \{0,1\}$ for classification.
- Given realization S of $\mathsf{S} = \{\mathsf{x}_i, \mathsf{y}_i\}_{i=1}^n \overset{iid}{\sim} \mu^n_{\mathcal{X}\times\mathcal{Y}}$ and search set $\mathcal{F}$, "Learn"
  - Feed it to an Algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{F}$.
  - Outputs an element $f_{\mathcal{A}(S)} \in \mathcal{F}$
- Check if it is good. Use it for new realizations of (x, y).

STATISTICAL PERSPECTIVE

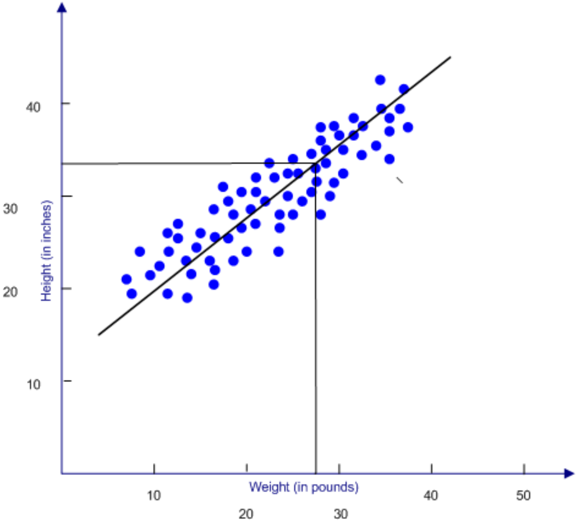
- Build probability based models.
  - Look at tails rather than mean.
  - Give answers as probabilistic guarantees (example: w.h.p)
- No free lunch
  - if there is no assumptions on how past relates to the future, learning is not possible.
  - if there is no restriction on the possible phenomena, generalization is impossible
- Assumptions:
  - Assume an unknown distribution $\mu_{\mathcal{X}\times\mathcal{Y}}$ exists, is stationary, and samples are drawn i.i.d.
  - Assume all explanations (models) belong to a set $\mathcal{F}$.

THEORY IS WHAT YOU MAKE OF IT.

$$\text{StatisticalLearning Theory} = \\ \text{theory} + \text{learning phenomenon} + \text{statistical perspective}$$

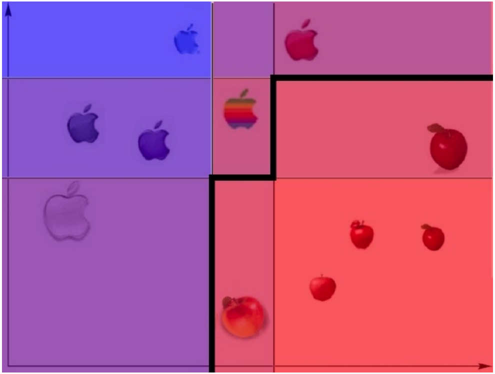
MACHINE LEARNING IN 3 PICTURES

1. Regression: $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \mathbb{R}$. Given new x find $\mathbb{E}[\mathsf{y}|\mathsf{x} = x]$.



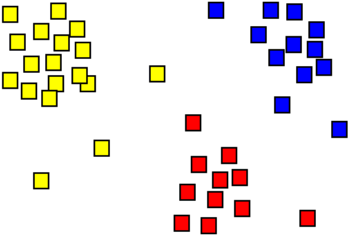
MACHINE LEARNING IN 3 PICTURES

2. Classification: $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \{0,1\}$. Given new x find $\Pr[\mathsf{y}|\mathsf{x} = x]$.



MACHINE LEARNING IN 3 PICTURES

3. Clustering: $\mathcal{X} \subset \mathbb{R}^p$. Given new x find $\Pr[\mathsf{x} = x]$.



Caution: clustering problems are bit different than classification or regression.

KEY OBJECTS I

- Data S is a realization of random variable S.
- Search set $\mathcal{F}$
  - example: bounded linear functions, set of polynomials etc.
  - $f : \mathcal{X} \to \mathcal{Y}$.
- Algorithm $\mathcal{A}$
  - input: S
  - output: Picks a function $f_{\mathcal{A}(S)} \in \mathcal{F}$
  - Can think of $f_{\mathcal{A}(\mathsf{S})}$ as a random variable.
- Loss function l
  - to assess any $f \in \mathcal{F}$
  - lots of variations depending on the learning task.
  - example: least squares $(f(x_i) - y_i)^2$ for regression.

KEY OBJECTS II

- Generalization error $R^{\text{exp.}}_{\mu_{X \times \mathcal{Y}}}(f)$
  - Given loss function l,

$$R^{\text{exp.}}_{\mu_{X \times \mathcal{Y}}}(f) := E_{\mu_{\mathcal{X} \times \mathcal{Y}}}[l(f(\mathsf{x}), \mathsf{y})]$$

  - Also known as Expected risk.
  - Special Case: Substitute $f = f_{\mathcal{A}(\mathsf{S})}$,
    - $R^{\text{exp.}}_{\mu_{X \times \mathcal{Y}}}(f_{\mathcal{A}(\mathsf{S})})$ is a random variable itself (depends on S).
- Empirical Risk
  - For any f, can't measure Generalization error.
  - So come up with an estimator for Generalization error
  - A re-substitution estimate using S

$$R^{\text{emp.}}_{S}(f) = \frac{1}{|S|} \sum_{(x_i, y_i) \in S} l(f(x_i), y_i)$$

  - Again, $R^{\text{emp.}}_{\mathsf{S}}(f_{\mathcal{A}(\mathsf{S})})$ is a random variable.

WHY THE NAME GENERALIZATION?

- Recall our assumptions:
  - $\mu_{\mathcal{X}\times\mathcal{Y}}$ is unknown, stationary and independent identically distributed samples.
  - Set $\mathcal{F}$ is our knowledge of what the model would come from.
- Come up with $\mathcal{F}$ by
  - preferring certain functions f over others.
  - restricting ourselves to some functions.
- If the set of models $\mathcal{F}$ is everything, not possible to generalize.
- Q: How to trade off knowledge and data?
  - Data can mislead us: overfitting vs underfitting
  - Knowledge can mislead us: if Bayes optimal model is not in $\mathcal{F}$
- A: this is formally known as approximation/ estimation tradeoff.

ALGORITHMS

- Empirical Risk Minimization:

$$f_{ERM(S)} \in \arg\min_{f \in \mathcal{F}} R_S^{\text{emp.}}(f)$$

  - Caution: always possible to get $f(x_i) \approx y_i$ by looking at richer search set $\mathcal{F}$.
- Structural Risk Minimization:

$$f_{SRM(S)} \in \arg\min_{j \in \mathcal{J}} \min_{f \in \mathcal{F}_j} R_S^{\text{emp.}}(f) + \text{pen}(\mathcal{F}_j, n)$$

  - prefers small empirical error
  - also takes into account capacity of set $\mathcal{F}_j$
- Regularized learning:

$$f_{reg.(S)} \in \arg\min_{f \in \mathcal{F}} R_S^{\text{emp.}}(f) + \text{reg}(f)$$

  - represents incorporating the knowledge.
  - example: ℓ_1-norm ball and linear function set $\mathcal{F}$

BOUNDS

TWO APPROACHES

- FIRST: statistical theory based on uniform convergence of empirical processes
  - Algorithm independent.
  - Assumes algorithm searches over entire search set.
  - worst case analysis.
- SECOND: statistical theory based on sensitivity analysis (or perturbation analysis)
  - Algorithm dependent analysis.
  - Is also known as stability analysis.
- We will focus on the former .
- The difference between empirical risk and expected risk will be the random variable of interest.
- Important component of both: Concentration.
  - A random variable that depends smoothly on the influence of many independent variables is essentially constant (Talagrand).

TOOLS WE KNOW

- Facts.
  - Union: $\Pr[A \text{ or } B] \leq \Pr[A] + \Pr[B]$
  - Inclusion: If $A \Rightarrow B$, then $\Pr[A] \leq \Pr[B]$
  - Inversion: If $\Pr[X \geq t] \leq F(t)$, then with probability $1 - \delta$, $X \leq F^{-1}(\delta)$
- Inequalities.
  - Jensen: for a convex f, $f(\mathbb{E}[\mathsf{x}]) \leq \mathbb{E}(f(\mathsf{x}))$
  - Markov: For $t > 0$, if $\mathsf{x} \geq 0$, $\Pr[\mathsf{x} \geq t] \leq E[\mathsf{x}]/t$
  - Chebyshev: For $t > 0$, $\Pr[|\mathsf{x} - \mathbb{E}[\mathsf{x}]| \geq t] \leq \mathrm{var}(\mathsf{x})/t^2$
  - Chernoff: For any t,$\Pr[\mathsf{x} \geq t] \leq \inf_{\lambda \geq 0} \mathbb{E}[e^{\lambda(\mathsf{x}-t)}]$

LAW OF LARGE NUMBERS (LLN)

- Law of large numbers (strong)

$$\Pr[\lim_{n\to\infty}\{\frac{1}{n}\sum_{i=1}^{n} g(\mathsf{z_i}) - \mathbb{E}[g(\mathsf{z})]\} = 0] = 1.$$

- Hoeffding: Quantitative version of LLN

$$\Pr[|\frac{1}{n}\sum_{i=1}^{n} g(\mathsf{z}_i) - \mathbb{E}[g(\mathsf{z})]| > \epsilon] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

- By inversion, can also write with probability at least $1-\delta$,

$$|\frac{1}{n}\sum_{i=1}^{n} g(\mathsf{z}_i) - \mathbb{E}[g(\mathsf{z})]| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

APPLYING HOEFFDINGS

- If the function DOES NOT depend on data, then replace $\mathbb{E}[g(\mathsf{z})]$ by $R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f)$ and $\frac{1}{n}\sum g(\mathsf{z}_i)$ by $R^{\text{emp.}}_{\mathsf{S}}(f)$.
- For any $\delta > 0$, with probability at least $1-\delta$,

$$R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) \leq R^{\text{emp.}}_{\mathsf{S}}(f) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

- For each $f \in \mathcal{F}$, there is a sample S for which the inequality is not true.
- For a realization S, only some $f \in \mathcal{F}$ will satisfy inequality.

UNIFORM DEVIATIONS

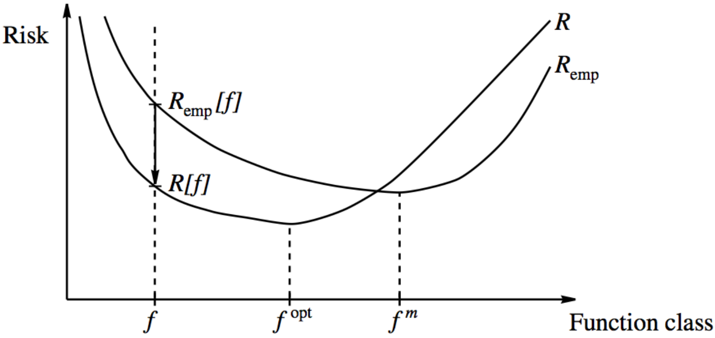
- Before seeing S, we don't know which model $\mathcal{A}$ picks. So consider uniform deviations:

$$|R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f_{\mathcal{A}(\mathsf{S})}) - R^{\text{emp.}}_{\mathsf{S}}(f_{\mathcal{A}(\mathsf{S})})| \leq \sup_{f\in\mathcal{F}} |R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) - R^{\text{emp.}}_{\mathsf{S}}(f)|$$

- Another reason for looking at uniform deviations:

$$R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f_{ERM(\mathsf{S})}) - \inf_{f\in\mathcal{F}} R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f)$$
$$= R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f_{ERM(\mathsf{S})}) - R^{\text{emp.}}_{\mathsf{S}}(f_{ERM(\mathsf{S})}) + R^{\text{emp.}}_{\mathsf{S}}(f_{ERM(\mathsf{S})}) - \inf_{f\in\mathcal{F}} I$$
$$\leq R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f_{ERM(\mathsf{S})}) - R^{\text{emp.}}_{\mathsf{S}}(f_{ERM(\mathsf{S})}) + \sup_{f\in\mathcal{F}} |R^{\text{emp.}}_{\mathsf{S}}(f) - R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}$$
$$\leq 2 \sup_{f\in\mathcal{F}} |R^{\text{emp.}}_{\mathsf{S}}(f) - R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f)|$$

- Leads to a bound which holds simultaneously for all functions in $\mathcal{F}$.
- Caution: This technique cannot be used to study $f_{\mathcal{A}(\mathsf{S})}$


Risk
R
R_emp
R_emp [f]
R[f]
f
f opt
f m
Function class

UNIFORM DEVIATION BOUND USING UNION BOUND.

- Let $|\mathcal{F}|$ be finite. And $0 \leq l(f(x), y) \leq 1$. Then, by union bound and one sided deviation inequality

$$\begin{aligned}\Pr[\exists f : R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) - R^{\text{emp.}}_{\mathsf{S}}(f) > \epsilon] &\leq \sum_{f\in\mathcal{F}} \Pr[R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) - R^{\text{emp.}}_{\mathsf{S}}(f) > \epsilon] \\ &\leq |\mathcal{F}| \exp(-2n\epsilon^2)\end{aligned}$$

- Or equivalently: With probability at least $1-\delta$,

$$\forall f \in \mathcal{F}, \;\; R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) \leq R^{\text{emp.}}_{\mathsf{S}}(f) + \sqrt{\frac{\log|\mathcal{F}| + \log\frac{1}{\delta}}{2n}}$$

HOW TO REASON WITH INFINITE $\mathcal{F}$: VC THEORY

- Simple search sets like the set of bounded linear functionals in some finite dimension are infinite.
- Growth function and VC dimension of $\{0,1\}$-valued search sets $\mathcal{F}$ help get around infinite sets.
- Uniform bound for infinite search set $\mathcal{F}$:

$$\Pr[\sup_{f\in\mathcal{F}} |R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) - R^{\text{emp.}}_{\mathsf{S}}(f)| \geq \epsilon] \leq 8\mathcal{S}_{\mathcal{F}}(n)\exp(-\frac{n\epsilon^2}{8})$$

- If $\mathcal{S}_{\mathcal{F}}(n)$ grows slower than the multiplicative exponential tail term, then uniform convergence happens.
- Sauer's lemma: $\mathcal{S}_{\mathcal{F}}(n) \leq (\frac{en}{VCdim(\mathcal{F})})^{VCdim(\mathcal{F})}$

COMPLEXITY MEASURES

VC DIMENSION

- Although $\mathcal{F}$ may be infinite, applying elements of $\mathcal{F}$ to S will give finite outcomes.
- Typically exponential. Example: 2^n for classification.
- The point at which the growth stops being exponential is when the complexity of $\mathcal{F}$ has been exhausted.
- Definition: $\Pi_{\mathcal{F}}(S) := \{f(x_1), f(x_2), ..., f(x_n) : f \in \mathcal{F}\}$
  - Behaviors on S realized by $\mathcal{F}$
  - If $\mathcal{F}$ makes a full realization, then $|\Pi_{\mathcal{F}}(S)| = 2^n$
  - Also looked at as a collection of subsets partitioning S.
- If $\Pi_{\mathcal{F}}(S) = 2^n$,then S is considered shattered by $\mathcal{F}$.
- $VCdim(\mathcal{F})$ is the size of the largest set S shattered by $\mathcal{F}$.
  - $VCdim(\mathcal{F}) = \max\{d : \exists |S| = d, \text{ and } \Pi_{\mathcal{F}}(S) = 2^d\}$
  - Growth function: $\mathcal{S}_{\mathcal{F}}(n) = \sup_{x_1,...,x_n} |\Pi_{\mathcal{F}}(S)|$
  - The $VCdim(\mathcal{F})$ for separating hyperplanes in $\mathbb{R}^d$ is $d+1$.

MORE COMPLEXITY MEASURES

- $VCdim(\mathcal{F})$ makes sense for $\{0, 1\}$-valued functions.
- What about real functions?
- There are various modifications to the definition. Leads to ϵ-fat shattering dimension and the idea of a margin.
- We will see two more:
  - Covering numbers.
  - Rademacher complexity
- Note: better generalization does not imply a better model.
  - Increasing $reg(f)$ or $pen(\mathcal{F}_j, n)$ increases the bias of the model and helps to reduce the variance
  - but any type of bias can either help or hurt the quality of modeling, depending on whether the knowledge associated with the bias is correct.

RADEMACHER COMPLEXITY

- Given S and $\mathcal{F}$, $\mathcal{F}_{|S}$ is the defined as the restriction of $\mathcal{F}$ with respect to S.
- The empirical Rademacher complexity of $\mathcal{F}_{|S}$ is:

$$\hat{\mathcal{R}}(\mathcal{F}_{|S}) = \mathbb{E}_{\sigma}[\sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i)]$$

  where $\{\sigma_i\}$ are Rademacher random variables ($\sigma_i = 1$ wp $1/2$ and -1 wp $1/2$).
- The Rademacher complexity is its expectation: $\mathcal{R}(\mathcal{F}) = \mathbb{E}_{S \sim (\mu_{\mathcal{X}})^n}[\hat{\mathcal{R}}(\mathcal{F}_{|S})]$.
- A uniform deviations statement very similar to VC uniform deviations statement can be proved.

RADEMACHER COMPLEXITY AND UNIFORM DEVIATIONS

- Using a deviation inequality called *McDiarmid's inequality*, write (using symmetrization) $R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f)$ in terms of the Rademacher complexity of a loss function set $l_{\mathcal{F}}$.
- Using the Ledoux-Talagrand's contraction lemma, relate this to Rademacher complexity of $\mathcal{F}$through a Lipschitz constant.

Theorem

For all $\delta > 0$, with probability at least $1-\delta, \forall f \in \mathcal{F}$,

$$R^{\text{exp.}}_{\mu_{X\times\mathcal{Y}}}(f) \leq R^{\text{emp.}}_{\mathsf{S}}(f) + \mathcal{L}\cdot\hat{\mathcal{R}}(\mathcal{F}_{|S}) + \frac{3}{\sqrt{2}}\sqrt{\frac{log\frac{1}{\delta}}{n}}.$$

RADEMACHER COMPLEXITY AND COVERING NUMBERS

- Discretization theorem (technique known as chaining)

Theorem
Let $\forall x \in \mathcal{X},\ f(x) \in [-b, b]$.

$$\frac{1}{b}\hat{\mathcal{R}}(\mathcal{F}_{|S}) \leq \inf_{\alpha>0}\left(\sqrt{\frac{2\log N(\alpha, \mathcal{F}_{|S}, \|\cdot\|_2)}{n}} + \alpha\right)$$

where $N(\alpha, \mathcal{F}_{|S}, \|\cdot\|_2)$ is the covering number of the set $\mathcal{F}_{|S}$.

COVERING NUMBERS

- Let $A \subseteq X$ be an arbitrary set and (X, dist) a (pseudo) metric space. Let $|\cdot|$ denote set size.
  - For any $\epsilon > 0$, an <u>ϵ-cover</u> for A is a finite set $U \subseteq X$ (not necessarily $\subseteq A$) s.t. $\forall x \in A, \exists u \in U$ with $\text{dist}(x, u) \leq \epsilon$.
  - A is totally bounded if A has a finite ϵ-cover for all $\epsilon > 0$. The <u>*covering number*</u> of A is then defined as $N(\epsilon, A, \text{dist}) := \inf_{U \in \mathcal{U}} |U|$ where $\mathcal{U}$ is the set of all ϵ-covers for A.
  - A set $R \subseteq X$ is ϵ-separated if $\forall x, y \in R, \text{dist}(x, y) > \epsilon$. The <u>*packing number*</u> $M(\epsilon, A, \text{dist}) := \sup_{R \in \mathcal{R}} |R|$, where $\mathcal{R}$ is the set of all ϵ-separated subsets of A.
- For every (pseudo) metric space (X, dist), $A \subseteq X$, and $\epsilon > 0$,
$$N(\epsilon, A, \text{dist}) \leq M(\epsilon, A, \text{dist}).$$
- Example: $N(\epsilon, \{x \in \mathbb{R}^d : ||x||_2 \leq 1\}, ||\cdot||_2) \leq (\frac{2}{\epsilon} + 1)^d$

COVERING NUMBERS AND UNIFORM DEVIATIONS

- Very similar to VC uniform deviation and Rademacher based uniform deviation results.

Theorem

Let $l_{\mathcal{F}}$ be a set of functions based on $\mathcal{F}$ with $0 \leq l(f(x), y) \leq M_{\text{bound}}, \quad \forall l \in l_{\mathcal{F}}$ and $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$. Then for any $\epsilon > 0$,

$$
\begin{aligned}
& P_{\mathsf{S}}(\exists l \in l_{\mathcal{F}} : |R^{\text{emp.}}_{\mathsf{S}}(f) - R^{\text{exp.}}_{\mu_{X \times \mathcal{Y}}}(f)| > \epsilon) \\
\leq \quad & 8\mathbb{E}\left[N\left(\epsilon/8, l_{\mathcal{F}}, \|\cdot\|_{L_1(\mu^m_{\mathcal{X} \times \mathcal{Y}})}\right)\right] \exp\left(\frac{-n\epsilon^2}{128 M^2_{\text{bound}}}\right).
\end{aligned}
$$

USE OF COMPLEXITY MEASURES

- Can also give us penalty term $pen(\mathcal{F}_j, n)$ for Structural Risk Minimization.
- Regularization looks very similar.
  - Thus, can justify it as some form of complexity control
  - There are other direct justifications: Sparsity or smoothness in regression.
- Adding regularization or penalty leads to algorithms, and the bounds we have seen before are independent of algorithms.
- Additional steps are required. For example, calibration.

SUMMARY

- Introduced SLT and its objects ($\mathsf{S}, \mathcal{F}, \mathcal{A}, R^{\text{exp.}}_{\mu_{\mathcal{X}\times\mathcal{Y}}}(f), R^{\text{emp.}}_{\mathsf{S}}(f)$)
- Looked at three algorithms: ERM, SRM and Regularization.
- In terms of generalization, we realize that data (S) cannot replace knowledge ($\mathcal{F}$)
- Only data + knowledge leads to low generalization error.
- Looked at Hoeffdings inequality and uniform deviations.
- Saw how to deal with infinite search sets: VC dimension, Rademacher complexity, Covering numbers.
- There were a lot of things we did not see.
- But we did look at the rationale behind the statistical aspects concerning learning phenomena.