Overfitting and Structural Risk Minimization

Dániel Csaba

Topics in Computational Economics
New York University

April 15, 2016

Introduction

One of the central issues in finite sample statistical inference is overfitting

`scikit-learn` package offers remedies

- regularization
- penalty term
- tuning parameter ...

Objective: try and look at these in a common framework

Statistical Learning Problem

Objective: prediction

- learn functional dependence from finite observations

Stable environment – probabilistic relationship

$$(y, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^d$$
$$P(y, \mathbf{x}) = P(y \mid \mathbf{x}) \cdot P(\mathbf{x})$$

IID observations

$$\mathcal{D}_n := \{(y_i, \mathbf{x}_i)\}_{i=1}^n$$

- independent – each observation yields maximum information
- identically distributed – learning is possible

Loss and Target

Provide a function, $f : X \mapsto Y$, which predicts y "well" as a function of $\mathbf{x}$

Define what we mean by "well"

- some form of discrepancy—loss: $L(y, f(\mathbf{x}))$—in expectation

Risk functional: $$R(f) := \int_{Y,X} L(y, f(\mathbf{x}))dP(y, \mathbf{x})$$

These define the target

$$f_0 := \arg \inf_{f \in \mathcal{F}} R(f)$$

Empirical Risk Minimization Principle

Issue: the true distribution P is unknown

Analogue estimation

- use empirical distribution and minimize empirical risk

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \; R_{emp}(f; n)$$

$$= \arg\min_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$$

Minimizing over all functions in $\mathcal{F}$ would not make sense

Instead, choose a hypothesis space $\mathcal{H} \subseteq \mathcal{F}$

Target and Hypothesis Space

The form of the loss function defines a feature of the distribution

- regression – conditional mean (squared), median (absolute)
- classification – logistic (cross entropy)
- density – MLE (-log(density))

Choice of hypothesis space, $\mathcal{H} \subseteq \mathcal{F}$, the class within which one approximates the target

- linear
- polynomial
- parametric derived from theoretical model

Estimation- and Approximation Error

Tension while choosing $\mathcal{H}$

Decomposing the risk – denote $f_{\mathcal{H}} := \arg\min_{f \in \mathcal{H}} R(f)$

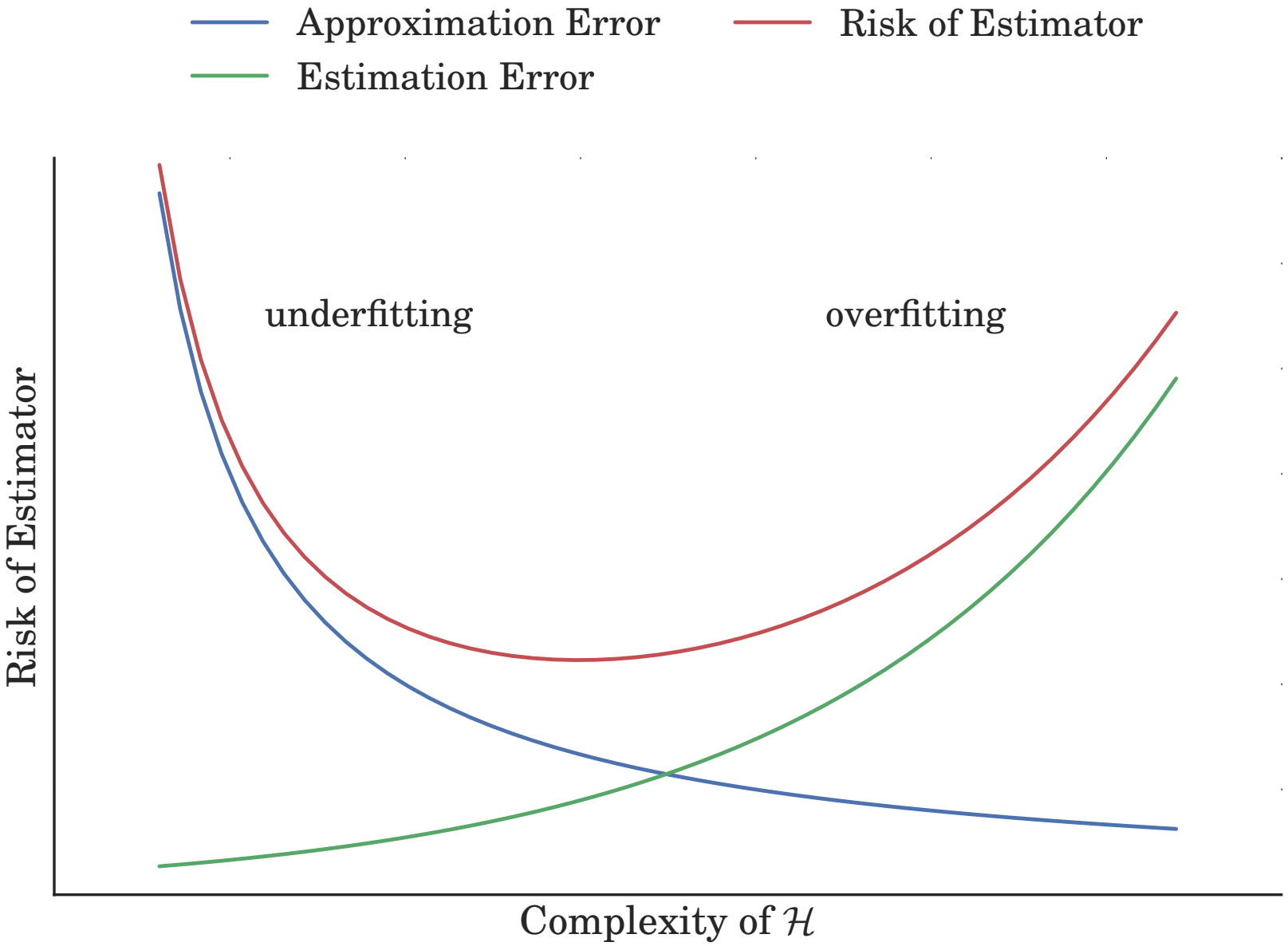
$$R(\hat{f}_n) - R(f_0) = \underbrace{R(\hat{f}_n) - R(f_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{H}}) - R(f_{\mathcal{F}})}_{\text{approximation error}}$$

Estimation error

- random quantity
- noise the estimator picks up

Approximation error

- deterministic quantity
- distance between $\mathcal{H}$ and target


Approximation Error
Risk of Estimator
Estimation Error
underfitting
overfitting
Risk of Estimator
Complexity of $\mathcal{H}$

Overfitting and Noise

Overfitting: pick the hypothesis with lower epirical risk and ultimately get higher true risk

Too much attention payed to a given realization of the sample

Estimation error $R(\hat{f}_n) - R(f_{\mathcal{H}})$ is a random quantity

Stochastic noise

- observations from the target are coming with noise
- higher level implies that the esimator is picking up more noise

Deterministic noise

- difference between $f_{\mathcal{H}}$ and $f_{\mathcal{F}}$ acts like noise
- unfortunately increasing $\mathcal{H}$ does not only affect the determinstc noise

Consistency and No Free Lunch

We can estimate $\hat{f}_n := \arg\min_{f \in \mathcal{H}} R_{emp}(f; n)$

We are interested in $R_{emp}(\hat{f}_n; n) \simeq R(\hat{f}_n)$

Conditions for two-sided uniform convergence (VC 1968, 1971)

$$\lim_{n \to \infty} P\left\{ \sup_{f \in \mathcal{H}} \left| R(f) - R_{emp}(f; n) \right| > \epsilon \right\} = 0 \quad \forall \epsilon > 0.$$

No free lunch (Devroye et al 1996)

- Any algorithm, in any finite sample can be arbitrarily far from the true risk for some distributions.

Capacity and Non-asymptotic Bounds

Capacity measure of the set $\{L(y, f(\mathbf{x})), f \in \mathcal{H}\}$ plays key role – $C_{\mathcal{H}}$

Bounds on estimation error

$\forall\ f \in \mathcal{H}$ with probability at least $1 - \delta$ we have that

$$\left|R(f) - R_{emp}(f; n)\right| \leq \Omega(C_{\mathcal{H}}, n, \delta)$$

Bound gets

- tighter – as n increases, δ decreases
- looser – as capacity $C_{\mathcal{H}}$ increases

Structural Risk Minimization

The bound on the risk consists of two terms

$$R(\hat{f}_n^{\mathcal{H}}) = R_{emp}(\hat{f}_n^{\mathcal{H}}) + (R(\hat{f}_n^{\mathcal{H}}) - R_{emp}(\hat{f}_n^{\mathcal{H}}))$$

$$R(\hat{f}_n) \leq R_{emp}(\hat{f}_n) + \Omega(C_{\mathcal{H}}, n)$$

Empirical risk – monotone decreasing in $\mathcal{H}$

Confidence interval – increasing in the capacity of $\mathcal{H}$

Objective is to optimally trade-off in-sample error and reliability of that error

Structural Risk Minimization

Capacity has to be a control variable

Define a structure on $\{L(y, f(\mathbf{x})), f \in \mathcal{F}\}$

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_k \subseteq \cdots \qquad (\cup_i \mathcal{H}_i = \mathcal{F})$$

such that the corresponding capacities are finite and satisfy

$$C_{\mathcal{H}_1} \leq C_{\mathcal{H}_2} \leq \cdots \leq C_{\mathcal{H}_k} \leq \cdots$$

THe SRM principle chooses $\mathcal{H}_k$ and corresponding $\hat{f}_n^k$ according to

$$\min_k \left\{ R_{emp}(\hat{f}_n^k) + \Omega(C_{\mathcal{H}_k}, n) \right\}$$

Model Selection in Practice

In practice the bounds are rarely tight and other methods are used to select the model

Heuristically, the sturcture often takes the form

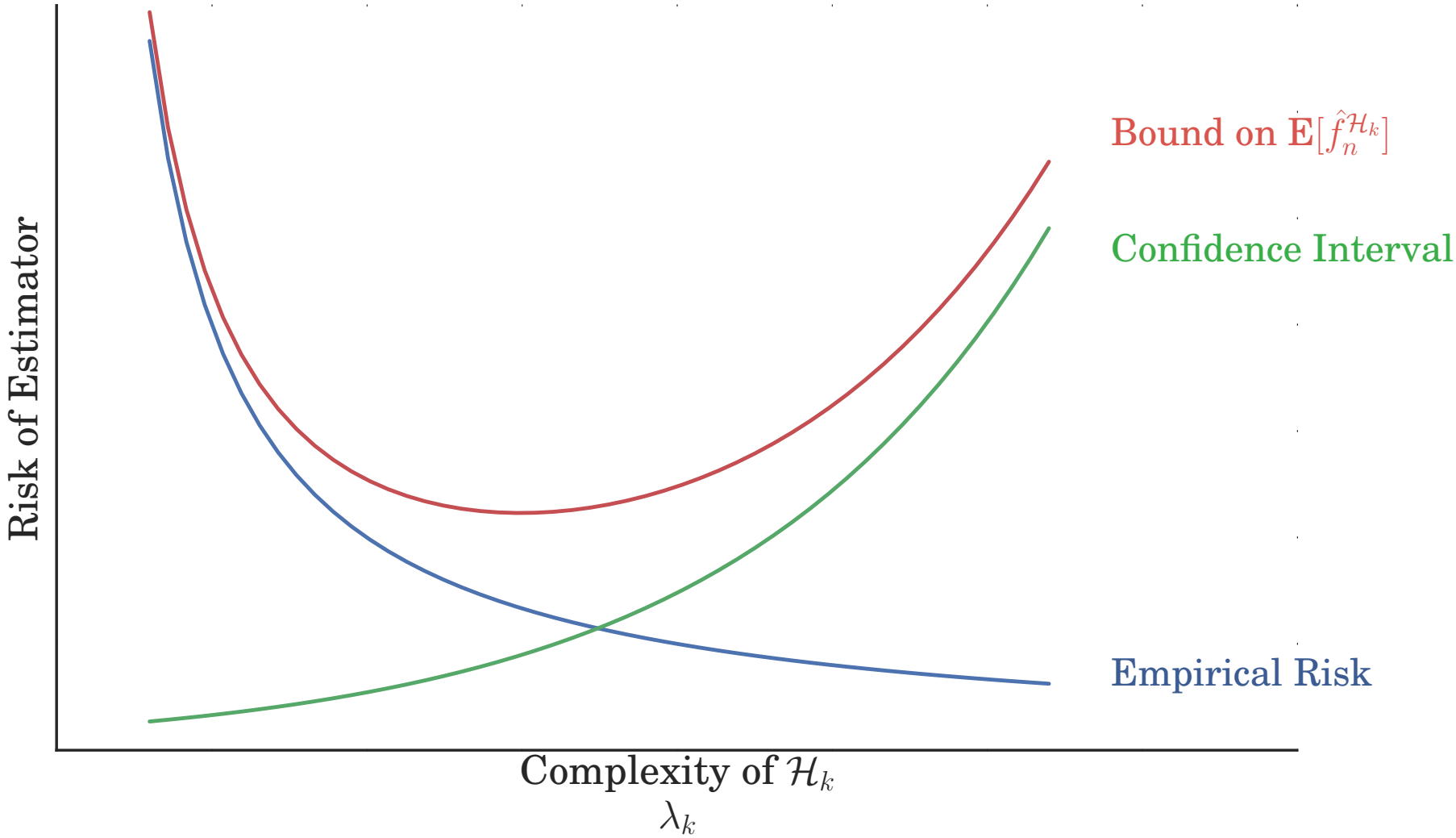
$$(\mathcal{F}, \lambda_1) \subseteq (\mathcal{F}, \lambda_2) \subseteq \cdots \subseteq (\mathcal{F}, \lambda_k) \subseteq \cdots$$

Think of $\mathcal{H}_k = \{f \in \mathcal{F} : \ \Omega(f) \leq A_k\}$

Then, one implements the SRM principle as

$$\min_{f \in \mathcal{F}} R_{emp}(f) + \lambda_k \Omega(f)$$

To choose the tuning parameter, λ_k^*, use validation, cross-validation


Risk of Estimator
Complexity of $\mathcal{H}_k$
λ_k
Bound on $\mathbf{E}[\hat{f}_n^{\mathcal{H}_k}]$
Confidence Interval
Empirical Risk