The Learning Problem and Regularization

Tomaso Poggio

9.520 Class 02

September 2013

Computational Learning

Statistical Learning Theory

Learning is viewed as a generalization/inference problem from usually **small** sets of **high dimensional**, noisy data.

Learning Tasks and Models

- Supervised
- Semisupervised
- Unsupervised
- Online
- Transductive
- Active
- Variable Selection
- Reinforcement
- .....

One can consider the data to be created in a deterministic, stochastic or even adversarial way.

Where to Start?

Statistical and Supervised Learning

- Statistical Models are essentially to deal with noise sampling and other sources of uncertainty.
- Supervised Learning is by far the most understood class of problems

Regularization

- Regularization provides a a fundamental framework to solve learning problems and design learning algorithms.
- **We present a set of ideas and tools which are at the core of several developments in supervised learning and beyond it.**
- This is a theory and associated algorithms which work in practice, eg in products, such as in vision systems for cars. Later in the semester we will learn about ongoing research combining neuroscience and learning. The latter research is at the frontier on approaches that may work in practice or may not (similar to Bayes techniques: still unclear how well they work beyond toy or special problems).

Where to Start?

Statistical and Supervised Learning

- Statistical Models are essentially to deal with noise sampling and other sources of uncertainty.
- Supervised Learning is by far the most understood class of problems

Regularization

- Regularization provides a a fundamental framework to solve learning problems and design learning algorithms.
- **We present a set of ideas and tools which are at the core of several developments in supervised learning and beyond it.**
- This is a theory and associated algorithms which work in practice, eg in products, such as in vision systems for cars. Later in the semester we will learn about ongoing research combining neuroscience and learning. The latter research is at the frontier on approaches that may work in practice or may not (similar to Bayes techniques: still unclear how well they work beyond toy or special problems).

Remarks on Foundations of Learning Theory

Intelligent behavior (at least learning) consists of optimizing under constraints. Constraints are key for solving computational problems; constraints are key for prediction. Constraints may correspond to rather general symmetry properties of the problem (eg time invariance, space invariance, invariance to physical units (pai theorem), universality of numbers and metrics implying normalization, etc.)

- Key questions at the core of learning theory:
  - generalization and predictivity *not* explanation
  - probabilities are unknown, only data are given
  - which constraints are needed to ensure generalization (therefore which hypotheses spaces)?
  - regularization techniques result usually in *computationally "nice"* and *well-posed* optimization problems

Plan

- Part I: Basic Concepts and Notation
- Part II: Foundational Results
- Part III: Algorithms

Problem at a Glance

Given a training set of *input-output* pairs

$$S_n = (x_1, y_1), \ldots, (x_n, y_n)$$

find f_S such that

$$f_S(x) \sim y.$$

e.g. the x's are vectors and the y's discrete labels in classification and real values in regression.

Learning is Inference

For the above problem to make sense we need to assume input and output to be related!

Statistical and Supervised Learning

- Each input-output pairs is a sample from a fixed but unknown distribution $p(x, y)$.
- Under general condition we can write

$$p(x, y) = p(y|x)p(x).$$

- the training set S_n is a set of **identically and independently distributed** samples.

Learning is Inference

For the above problem to make sense we need to assume input and output to be related!

Statistical and Supervised Learning

- Each input-output pairs is a sample from a fixed but unknown distribution $p(x, y)$.
- Under general condition we can write

$$p(x, y) = p(y|x)p(x).$$

- the training set S_n is a set of **identically and independently distributed** samples.

Again: Data Generated By A Probability Distribution

We assume that there are an "input" space X and an "output" space Y. We are given a **training set** S consisting n samples drawn i.i.d. from the probability distribution $\mu(z)$ on $Z = X \times Y$:

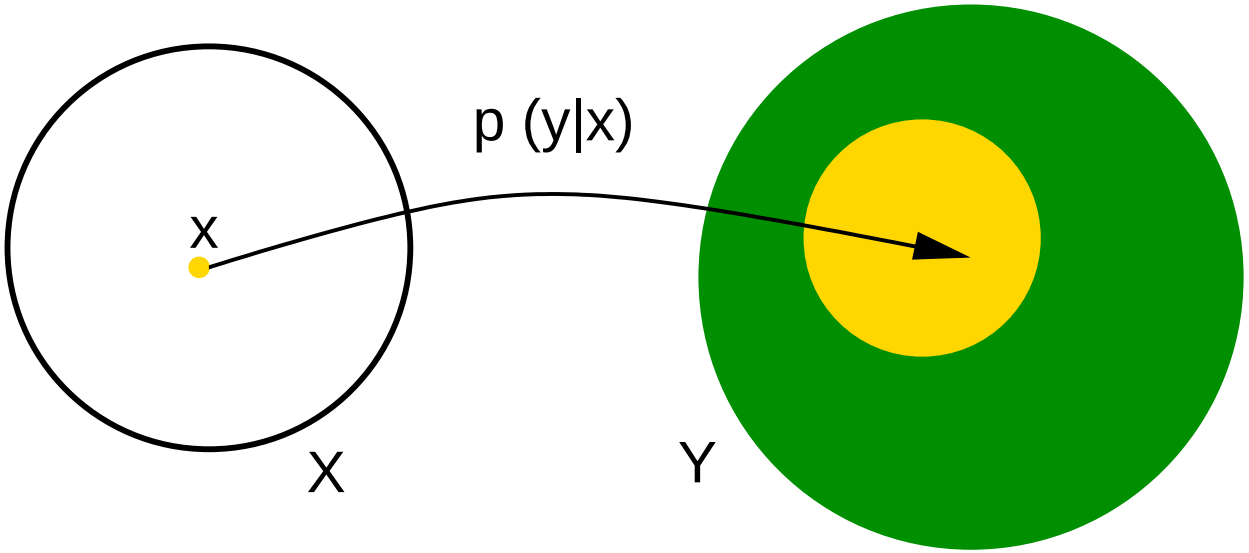
$$(x_1, y_1), \ldots, (x_n, y_n)$$

that is $z_1, \ldots, z_n$
We will use the **conditional probability of y given x**, written $p(y|x)$:

$$\mu(z) = p(x, y) = p(y|x) \cdot p(x)$$

It is crucial to note that we view $p(x, y)$ as **fixed** but **unknown**.

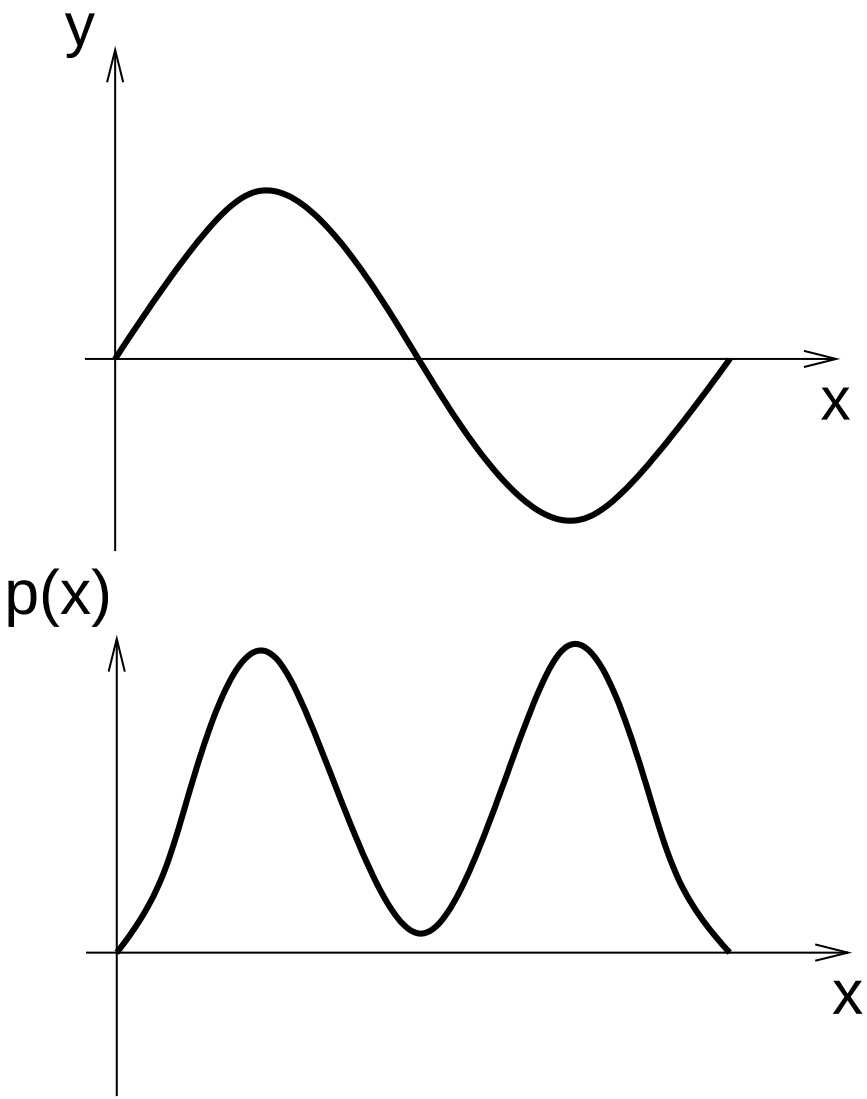
Noise ...



the same x can generate different y (according to $p(y|x)$):

- the underlying process is deterministic, but there is **noise** in the measurement of y;
- the underlying process is **not deterministic**;
- the underlying process is deterministic, but only **incomplete** information is available.

...and Sampling

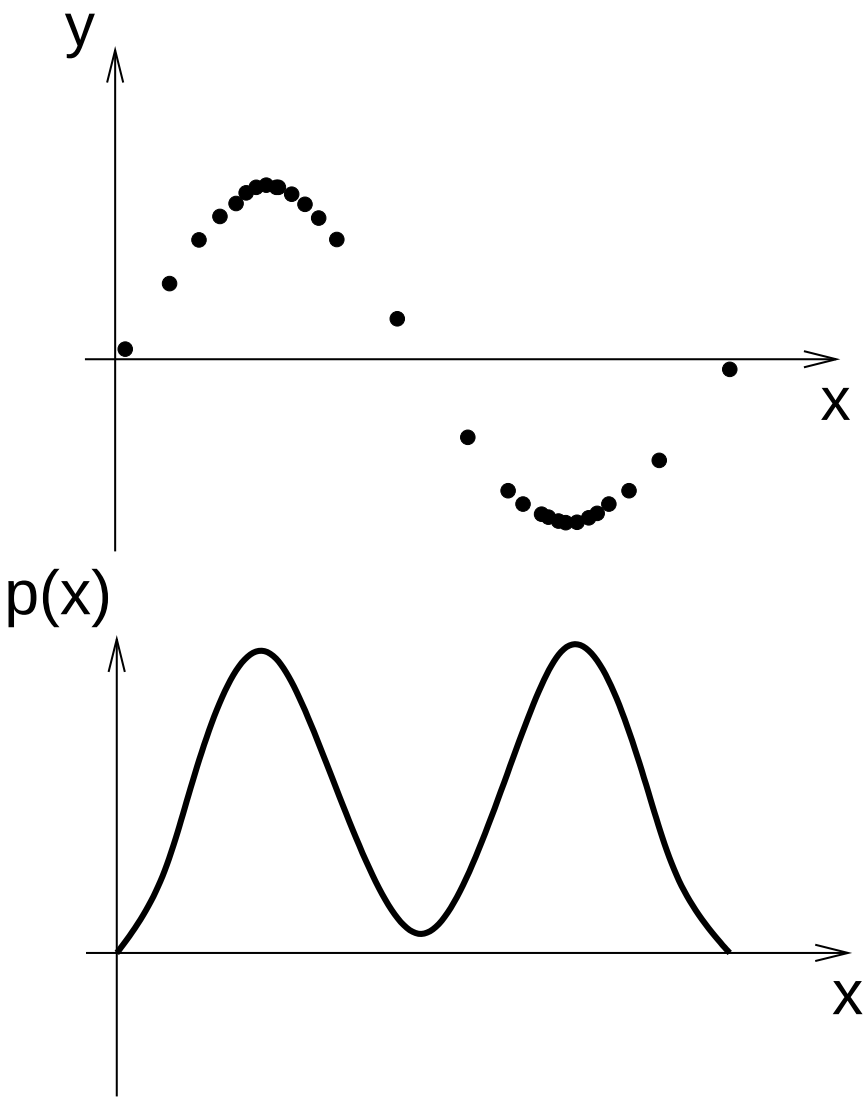


even in a noise free case we have to deal with sampling

the marginal $p(x)$ distribution might model

- errors in the location of the input points;
- discretization error for a given grid;
- presence or absence of certain input instances

...and Sampling

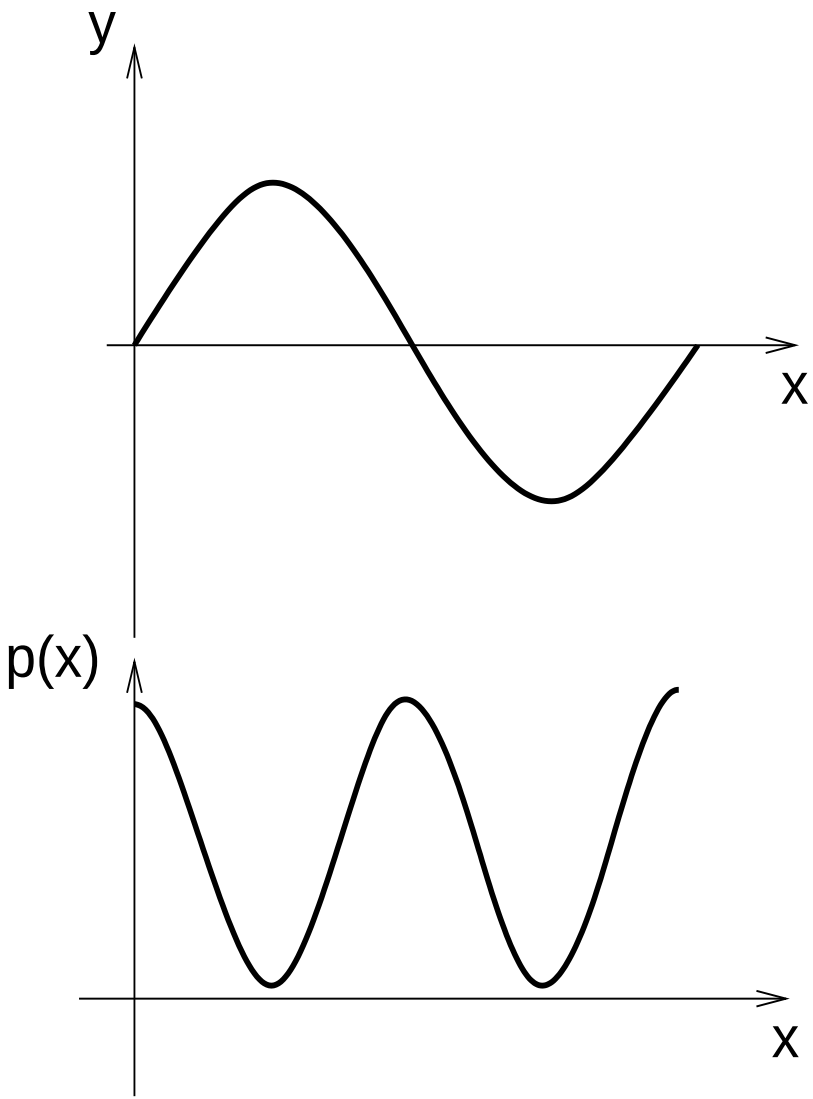


even in a noise free case we have to deal with sampling

the marginal $p(x)$ distribution might model

- errors in the location of the input points;
- discretization error for a given grid;
- presence or absence of certain input instances

...and Sampling

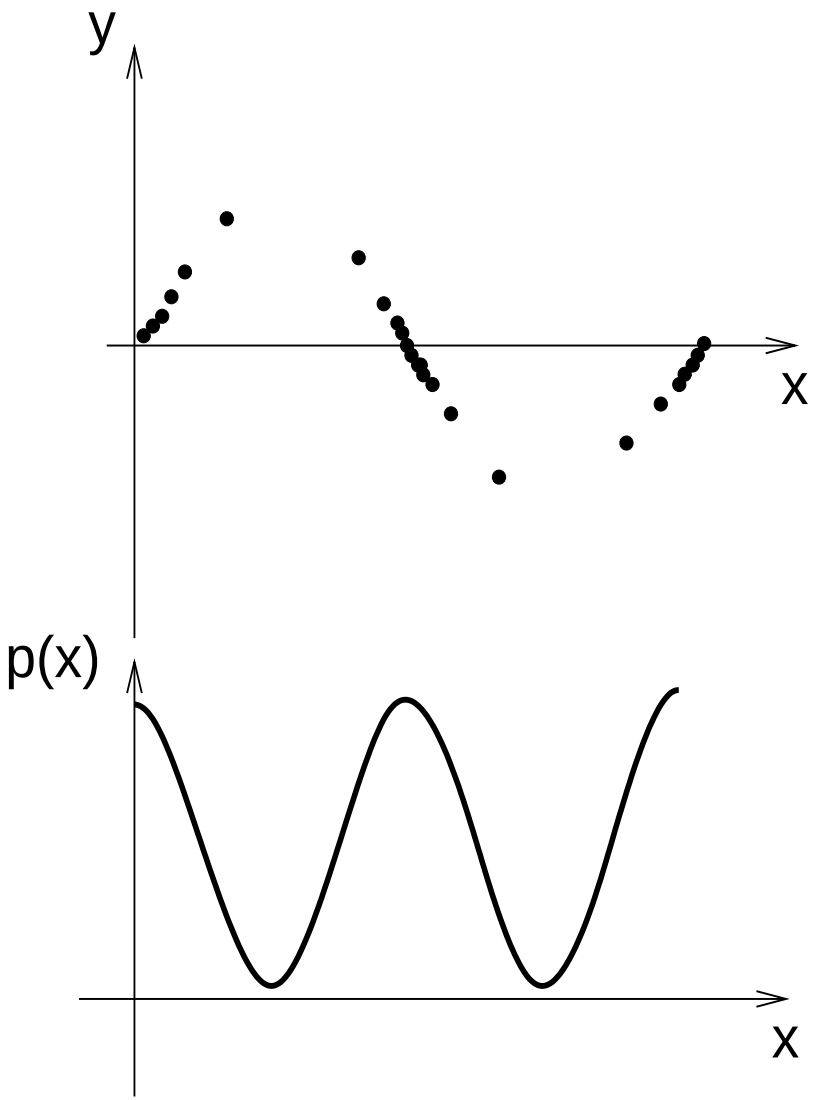


even in a noise free case we have to deal with sampling

the marginal $p(x)$ distribution might model

- errors in the location of the input points;
- discretization error for a given grid;
- presence or absence of certain input instances

...and Sampling



even in a noise free case we have to deal with sampling

the marginal $p(x)$ distribution might model

- errors in the location of the input points;
- discretization error for a given grid;
- presence or absence of certain input instances

Problem at a Glance

Given a training set of *input-output* pairs

$$S_n = (x_1, y_1), \ldots, (x_n, y_n)$$

find f_S such that

$$f_S(x) \sim y.$$

e.g. the x's are vectors and the y's discrete labels in classification and real values in regression.

Learning, Generalization and Overfitting

Predictivity or Generalization

Given the data, the goal is to learn how to make decisions/predictions about future data / data not belonging to the training set. **Generalization** is the key requirement emphasized in Learning Theory. This emphasis makes it different from traditional statistics (especially explanatory statistics) or Bayesian freakonomics.

The problem is often: **Avoid overfitting!!**

Loss functions

As we look for a deterministic estimator in stochastic environment we expect to incur into errors.

Loss function

A loss function $V : \mathbf{R} \times Y$ determines the price $V(f(x), y)$ we pay, predicting $f(x)$ when in fact the true output is y.

Loss functions

As we look for a deterministic estimator in stochastic environment we expect to incur into errors.

Loss function

A loss function $V : \mathbf{R} \times Y$ determines the price $V(f(x), y)$ we pay, predicting $f(x)$ when in fact the true output is y.

Loss functions for regression

- The most common is the **square loss** or L_2 **loss**

$$V(f(x), y) = (f(x) - y)^2$$

- **Absolute value** or L_1 **loss**:

$$V(f(x), y) = |f(x) - y|$$

- Vapnik's ϵ-**insensitive loss**:

$$V(f(x), y) = (|f(x) - y| - \epsilon)_+$$

Loss functions for (binary) classification

- The most intuitive one: $0-1$**-loss**:

$$V(f(x), y) = \theta(-yf(x))$$

(θ is the step function)

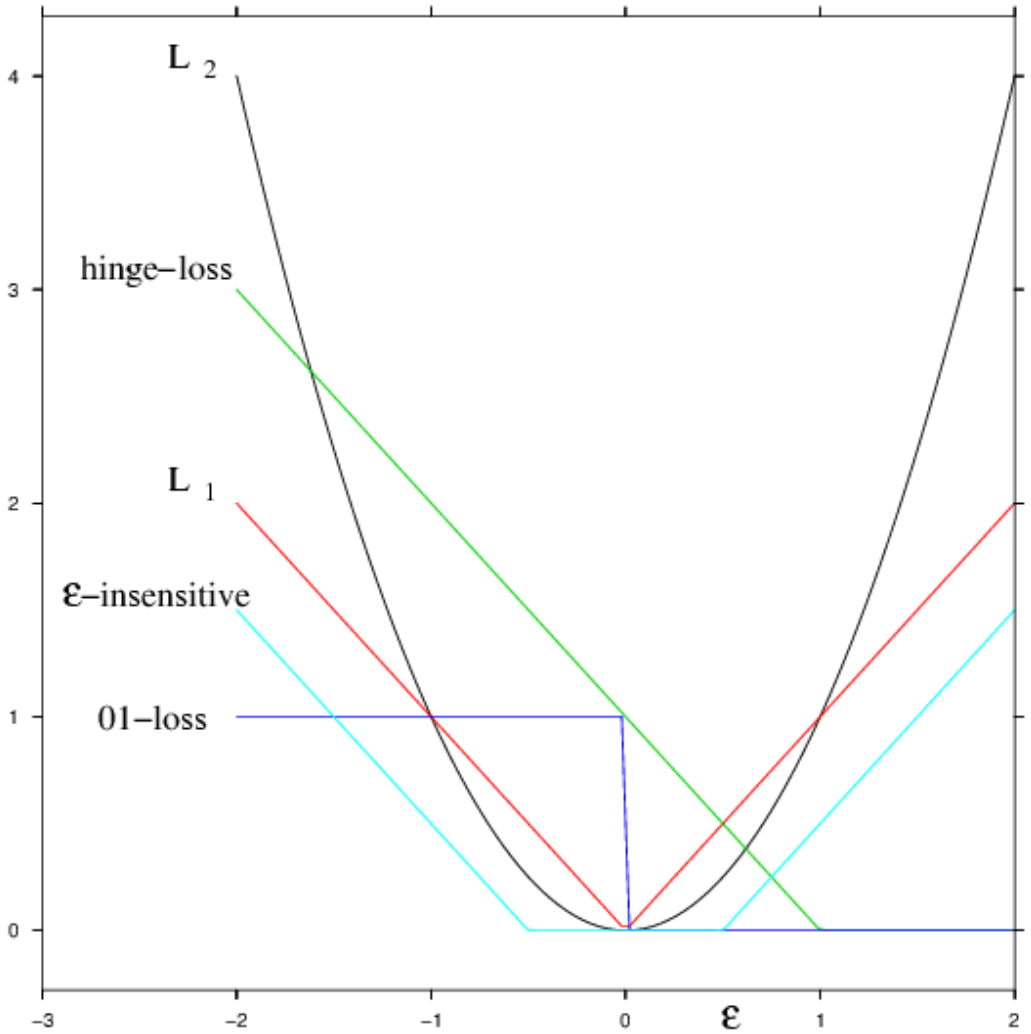
- The more tractable **hinge loss**:

$$V(f(x), y) = (1 - yf(x))_+$$

- And again the **square loss** or L_2 **loss**

$$V(f(x), y) = (1 - yf(x))^2$$

Loss functions



Expected Risk

A good function – we will also speak about *hypothesis* – should incur in only *a few* errors. We need a way to quantify this idea.

Expected Risk

The quantity

$$I[f] = \int_{X \times Y} V(f(x), y) p(x, y) dxdy.$$

is called the expected error and measures the loss averaged over the unknown distribution.

A good function should have small expected risk.

Expected Risk

A good function – we will also speak about *hypothesis* – should incur in only *a few* errors. We need a way to quantify this idea.

Expected Risk

The quantity

$$I[f] = \int_{X \times Y} V(f(x), y) p(x, y) dxdy.$$

is called the expected error and measures the loss averaged over the unknown distribution.

A good function should have small expected risk.

Target Function

The expected risk is usually defined on some large space $\mathcal{F}$ possible dependent on $p(x, y)$.

The best possible error is

$$\inf_{f \in \mathcal{F}} I[f]$$

The infimum is often achieved at a minimizer f_* that we call *target function*.

Learning Algorithms and Generalization

A learning algorithm can be seen as a map

$$S_n \to f_n$$

from the training set to the a set of candidate functions.

Basic definitions

- $p(x, y)$ probability distribution,
- S_n training set,
- $V(f(x), y)$ loss function,
- $I_n[f] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)$, empirical risk,
- $I[f] = \int_{X \times Y} V(f(x), y) p(x, y) dxdy$, expected risk,

Reminder

Convergence in probability

Let $\{X_n\}$ be a sequence of bounded random variables. Then

$$\lim_{n\to\infty} X_n = X \quad \text{in probability}$$

if

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \epsilon\} = 0$$

Convergence in Expectation

Let $\{X_n\}$ be a sequence of bounded random variables. Then

$$\lim_{n\to\infty} X_n = X \quad \text{in expectation}$$

if

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0$$

. Convergence in the mean implies convergence in probability.

Reminder

Convergence in probability

Let $\{X_n\}$ be a sequence of bounded random variables. Then

$$\lim_{n\to\infty} X_n = X \quad \text{in probability}$$

if

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \epsilon\} = 0$$

Convergence in Expectation

Let $\{X_n\}$ be a sequence of bounded random variables. Then

$$\lim_{n\to\infty} X_n = X \quad \text{in expectation}$$

if

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0$$

. Convergence in the mean implies convergence in probability.

Consistency and Universal Consistency

A requirement considered of basic importance in classical statistics is for the algorithm to get better as we get more data (in the context of machine learning consistency is less immediately critical than *generalization*)...

Consistency

We say that an algorithm is consistent if

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Universal Consistency

We say that an algorithm is universally consistent if for all probability p,

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Consistency and Universal Consistency

A requirement considered of basic importance in classical statistics is for the algorithm to get better as we get more data (in the context of machine learning consistency is less immediately critical than *generalization*)...

Consistency

We say that an algorithm is consistent if

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Universal Consistency

We say that an algorithm is universally consistent if for all probability p,

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Consistency and Universal Consistency

A requirement considered of basic importance in classical statistics is for the algorithm to get better as we get more data (in the context of machine learning consistency is less immediately critical than *generalization*)...

Consistency

We say that an algorithm is consistent if

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Universal Consistency

We say that an algorithm is universally consistent if for all probability p,

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\{I[f_n] - I[f_*] \geq \epsilon\} = 0$$

Sample Complexity and Learning Rates

The above requirements are asymptotic.

Error Rates

A more practical question is, how fast does the error decay?
This can be expressed as

$$\mathbb{P}\{I[f_n] - I[f_*]\} \leq \epsilon(n, \delta)\} \geq 1 - \delta.$$

Sample Complexity

Or equivalently, 'how many point do we need to achieve an error ϵ with a prescribed probability δ?'
This can expressed as

$$\mathbb{P}\{I[f_n] - I[f_*] \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Sample Complexity and Learning Rates

The above requirements are asymptotic.

Error Rates

A more practical question is, how fast does the error decay?
This can be expressed as

$$\mathbb{P}\{I[f_n] - I[f_*]\} \leq \epsilon(n, \delta)\} \geq 1 - \delta.$$

Sample Complexity

Or equivalently, 'how many point do we need to achieve an error ϵ with a prescribed probability δ?'
This can expressed as

$$\mathbb{P}\{I[f_n] - I[f_*] \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Sample Complexity and Learning Rates

The above requirements are asymptotic.

Error Rates

A more practical question is, how fast does the error decay?
This can be expressed as

$$\mathbb{P}\{I[f_n] - I[f_*]\} \leq \epsilon(n, \delta)\} \geq 1 - \delta.$$

Sample Complexity

Or equivalently, 'how many point do we need to achieve an error ϵ with a prescribed probability δ?'
This can expressed as

$$\mathbb{P}\{I[f_n] - I[f_*] \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Sample Complexity and Learning Rates

The above requirements are asymptotic.

Error Rates

A more practical question is, how fast does the error decay? This can be expressed as

$$\mathbb{P}\{I[f_n] - I[f_*]\} \leq \epsilon(n, \delta)\} \geq 1 - \delta.$$

Sample Complexity

Or equivalently, 'how many point do we need to achieve an error ϵ with a prescribed probability δ?'

This can expressed as

$$\mathbb{P}\{I[f_n] - I[f_*] \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Sample Complexity and Learning Rates

The above requirements are asymptotic.

Error Rates

A more practical question is, how fast does the error decay?
This can be expressed as

$$\mathbb{P}\{I[f_n] - I[f_*]\} \leq \epsilon(n, \delta)\} \geq 1 - \delta.$$

Sample Complexity

Or equivalently, 'how many point do we need to achieve an error ϵ with a prescribed probability δ?'
This can expressed as

$$\mathbb{P}\{I[f_n] - I[f_*] \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Empirical risk and Generalization

How do we design learning algorithms that work? One of the most natural ideas is ERM...

Empirical Risk

The empirical risk is a natural proxy (how good?) for the expected risk

$$I_n[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i).$$

Generalization Error

The effectiveness of such an approximation error is captured by the generalization error,

$$\mathbb{P}\{|I[f_n] - I_n[f_n]| \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Empirical risk and Generalization

How do we design learning algorithms that work? One of the most natural ideas is ERM...

Empirical Risk

The empirical risk is a natural proxy (how good?) for the expected risk

$$I_n[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i).$$

Generalization Error

The effectiveness of such an approximation error is captured by the generalization error,

$$\mathbb{P}\{|I[f_n] - I_n[f_n]| \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Empirical risk and Generalization

How do we design learning algorithms that work? One of the most natural ideas is ERM...

Empirical Risk

The empirical risk is a natural proxy (how good?) for the expected risk

$$I_n[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i).$$

Generalization Error

The effectiveness of such an approximation error is captured by the generalization error,

$$\mathbb{P}\{|I[f_n] - I_n[f_n]| \leq \epsilon\} \geq 1 - \delta,$$

for $n = n(\epsilon, \delta)$.

Some (Theoretical and Practical) Questions

- How do we go from data to an actual algorithm or class of algorithms?
- Is minimizing error on the data a good idea?
- Are there fundamental limitations in what we can and cannot learn?

Some (Theoretical and Practical) Questions

- How do we go from data to an actual algorithm or class of algorithms?
- Is minimizing error on the data a good idea?
- Are there fundamental limitations in what we can and cannot learn?

Some (Theoretical and Practical) Questions

- How do we go from data to an actual algorithm or class of algorithms?
- Is minimizing error on the data a good idea?
- Are there fundamental limitations in what we can and cannot learn?

Plan

- Part I: Basic Concepts and Notation
- **Part II: Foundational Results**
- Part III: Algorithms

No Free Lunch Theorem Devroye et al.

Universal Consistency

Since classical statistics worries so much about consistency let us start here even if it is not the practically important concept. Can we learn consistently any problem? Or equivalently do universally consistent algorithms exist?

YES! Neareast neighbors, Histogram rules, SVM with (so called) universal kernels...

No Free Lunch Theorem

Given a number of points (and a confidence), can we always achieve a prescribed error?

NO!

The last statement can be interpreted as follows: inference from finite samples can effectively performed if and only if the problem satisfies some a priori condition.

No Free Lunch Theorem Devroye et al.

Universal Consistency

Since classical statistics worries so much about consistency let us start here even if it is not the practically important concept.
Can we learn consistently any problem? Or equivalently do universally consistent algorithms exist?
YES! Neareast neighbors, Histogram rules, SVM with (so called) universal kernels...

No Free Lunch Theorem

Given a number of points (and a confidence), can we always achieve a prescribed error?
NO!

The last statement can be interpreted as follows: inference from finite samples can effectively performed if and only if the problem satisfies some a priori condition.

No Free Lunch Theorem Devroye et al.

Universal Consistency

Since classical statistics worries so much about consistency let us start here even if it is not the practically important concept.
Can we learn consistently any problem? Or equivalently do universally consistent algorithms exist?
YES! Neareast neighbors, Histogram rules, SVM with (so called) universal kernels...

No Free Lunch Theorem

Given a number of points (and a confidence), can we always achieve a prescribed error?
NO!

The last statement can be interpreted as follows: inference from finite samples can effectively performed if and only if the problem satisfies some a priori condition.

No Free Lunch Theorem Devroye et al.

Universal Consistency

Since classical statistics worries so much about consistency let us start here even if it is not the practically important concept.
Can we learn consistently any problem? Or equivalently do universally consistent algorithms exist?
YES! Neareast neighbors, Histogram rules, SVM with (so called) universal kernels...

No Free Lunch Theorem

Given a number of points (and a confidence), can we always achieve a prescribed error?
NO!

The last statement can be interpreted as follows: inference from finite samples can effectively performed if and only if the problem satisfies some a priori condition.

Hypotheses Space

Learning does not happen in void. In statistical learning a first prior assumption amounts to choosing a suitable space of hypotheses $\mathcal{H}$.

The **hypothesis space** $\mathcal{H}$ is the space of functions that we allow our algorithm to "look at". For many algorithms (such as optimization algorithms) it is the space the algorithm is allowed to search. As we will see in future classes, it is often important to choose the hypothesis space as a function of the amount of data n available.

Hypotheses Space

Learning does not happen in void. In statistical learning a first prior assumption amounts to choosing a suitable space of hypotheses $\mathcal{H}$.

The **hypothesis space** $\mathcal{H}$ is the space of functions that we allow our algorithm to "look at". For many algorithms (such as optimization algorithms) it is the space the algorithm is allowed to search. As we will see in future classes, it is often important to choose the hypothesis space as a function of the amount of data n available.

Hypotheses Space

Examples: linear functions, polynomial, RBFs, Sobolev Spaces...

Learning algorithm

A learning algorithm A is then a map from the data space to $\mathcal{H}$,

$$A(S_n) = f_n \in \mathcal{H}.$$

Hypotheses Space

Examples: linear functions, polynomial, RBFs, Sobolev Spaces...

Learning algorithm

A learning algorithm A is then a map from the data space to $\mathcal{H}$,

$$A(S_n) = f_n \in \mathcal{H}.$$

Empirical Risk Minimization

How do we choose $\mathcal{H}$? How do we design A?

ERM

A prototype algorithm in statistical learning theory is Empirical Risk Minimization:

$$\min_{f \in \mathcal{H}} I_n[f].$$

Reminder: Expected error, empirical error

Given a function f, a loss function V, and a probability distribution μ over Z, the **expected or true error** of f is:

$$I[f] = \mathbb{E}_z V[f, z] = \int_Z V(f, z) d\mu(z)$$

which is the **expected loss** on a new example drawn at random from μ.

We would like to make $I[f]$ small, but in general we do not know μ.

Given a function f, a loss function V, and a training set S consisting of n data points, the **empirical error** of f is:

$$I_S[f] = \frac{1}{n} \sum V(f, z_i)$$

Reminder: Generalization

A natural requirement for f_S is distribution independent **generalization**

$$\lim_{n\to\infty} |I_S[f_S] - I[f_S]| = 0 \ \textit{in probability}$$

This is equivalent to saying that for each n there exists a ε_n and a $\delta(\varepsilon)$ such that

$$\mathbb{P}\left\{|I_{S_n}[f_{S_n}] - I[f_{S_n}]| \geq \varepsilon_n\right\} \leq \delta(\varepsilon_n), \tag{1}$$

with ε_n and δ going to zero for $n \to \infty$.
In other words, the training error for the solution must converge to the expected error and thus be a "proxy" for it. Otherwise the solution would not be "predictive".
A desirable additional requirement is **consistency**

$$\varepsilon > 0 \ \lim_{n\to\infty} \mathbb{P}\left\{ I[f_S] - \inf_{f\in\mathcal{H}} I[f] \geq \varepsilon \right\} = 0.$$

A learning algorithm should be well-posed, eg stable

In addition to the key property of generalization, a "good" learning algorithm should also be *stable*: f_S should depend continuously on the training set S. In particular, changing one of the training points should affect less and less the solution as n goes to infinity. Stability is a good requirement for the learning problem and, in fact, for any mathematical problem. We open here a small parenthesis on stability and well-posedness.

General definition of Well-Posed and Ill-Posed problems

A problem is **well-posed** if its solution:

- exists
- is unique
- depends continuously on the data (e.g. it is *stable*)

A problem is **ill-posed** if it is not well-posed. In the context of this class, well-posedness is mainly used to mean *stability* of the solution.

More on well-posed and ill-posed problems

Hadamard introduced the definition of ill-posedness. Ill-posed problems are typically inverse problems.
As an example, assume g is a function in Y and u is a function in X, with Y and X Hilbert spaces. Then given the linear, continuous operator L, consider the equation

$$g = Lu.$$

The direct problem is is to compute g given u; the inverse problem is to compute u given the data g. In the learning case L is somewhat similar to a "sampling" operation and the inverse problem becomes the problem of finding a function that takes the values

$$f(x_i) = y_i, i = 1, \ldots n$$

The inverse problem of finding u is well-posed when

- the solution exists,
- is unique and
- is *stable*, that is depends continuously on the initial data g.

Given a training set S and a function space $\mathcal{H}$, empirical risk minimization as we have seen is the class of algorithms that look at S and select f_S as

$$f_S = \arg\min_{f \in \mathcal{H}} I_S[f]$$

.

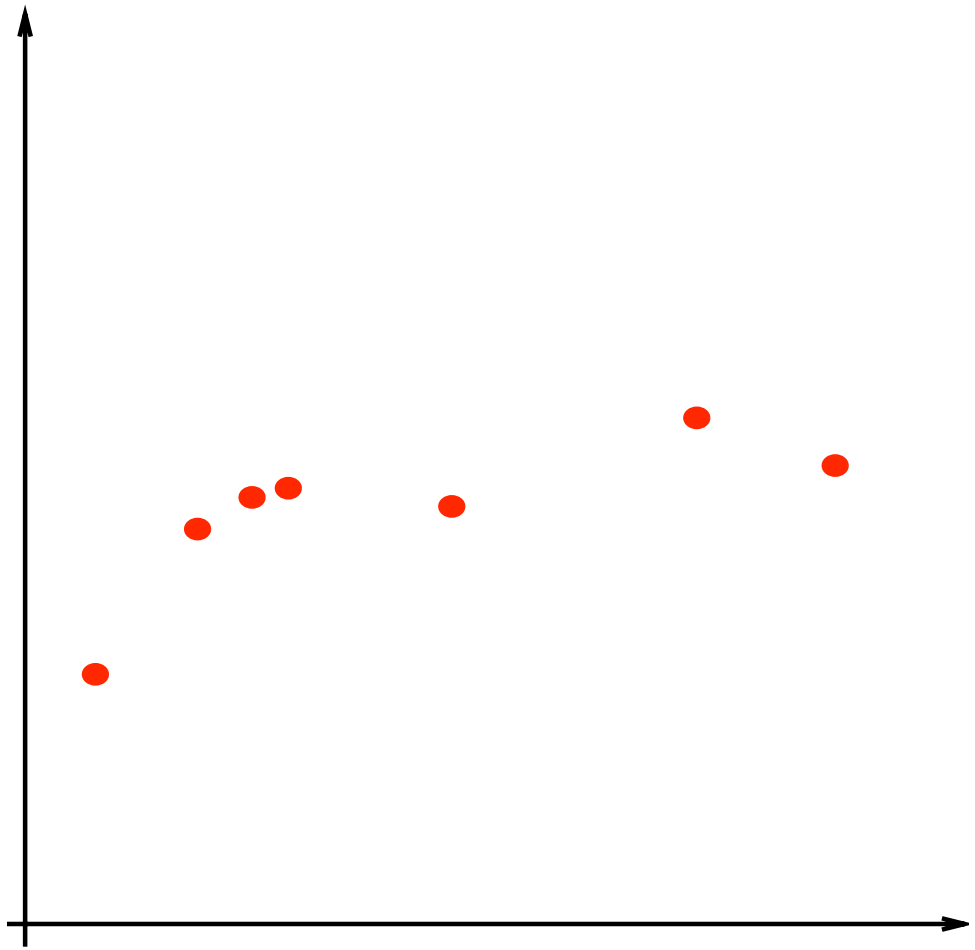
For example linear regression is ERM when $V(z) = (f(x) - y)^2$ and H is space of linear functions $f = ax$.

Generalization and Well-posedness of Empirical Risk Minimization

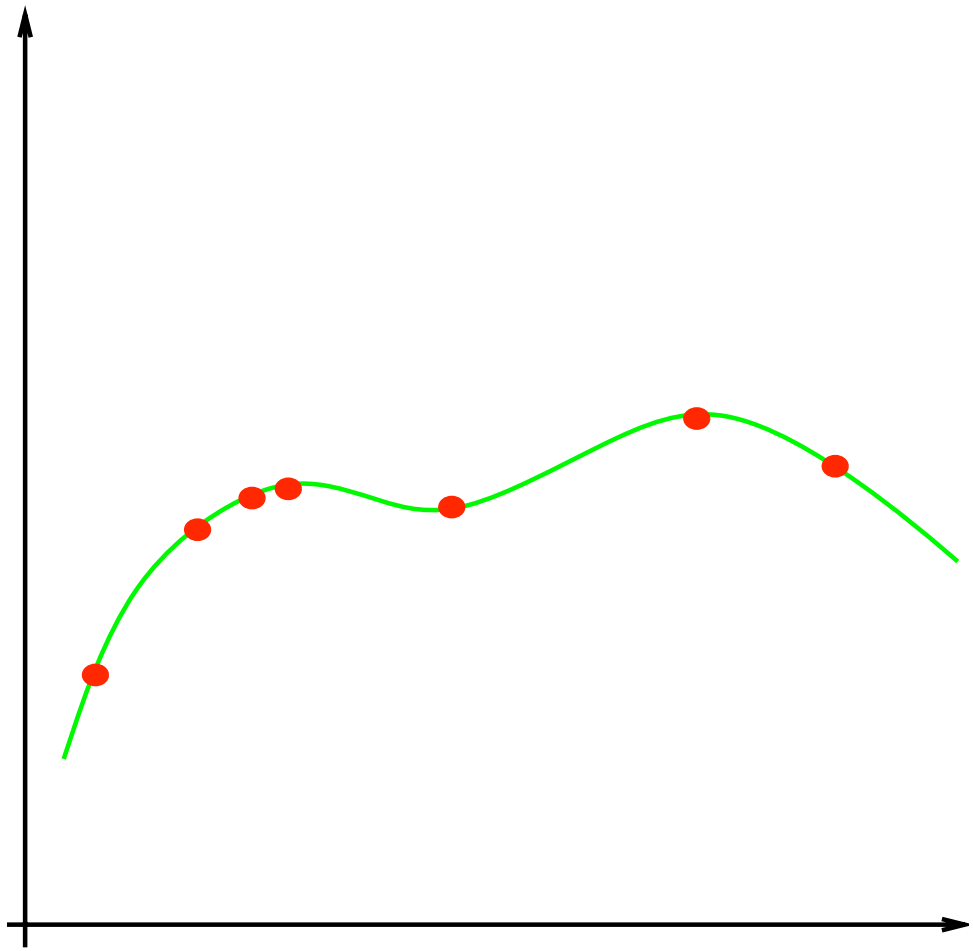
For ERM to represent a “good” class of learning algorithms, the solution should

- *generalize*
- exist, be unique and – especially – be *stable* (well-posedness), according to *some* definition of stability.

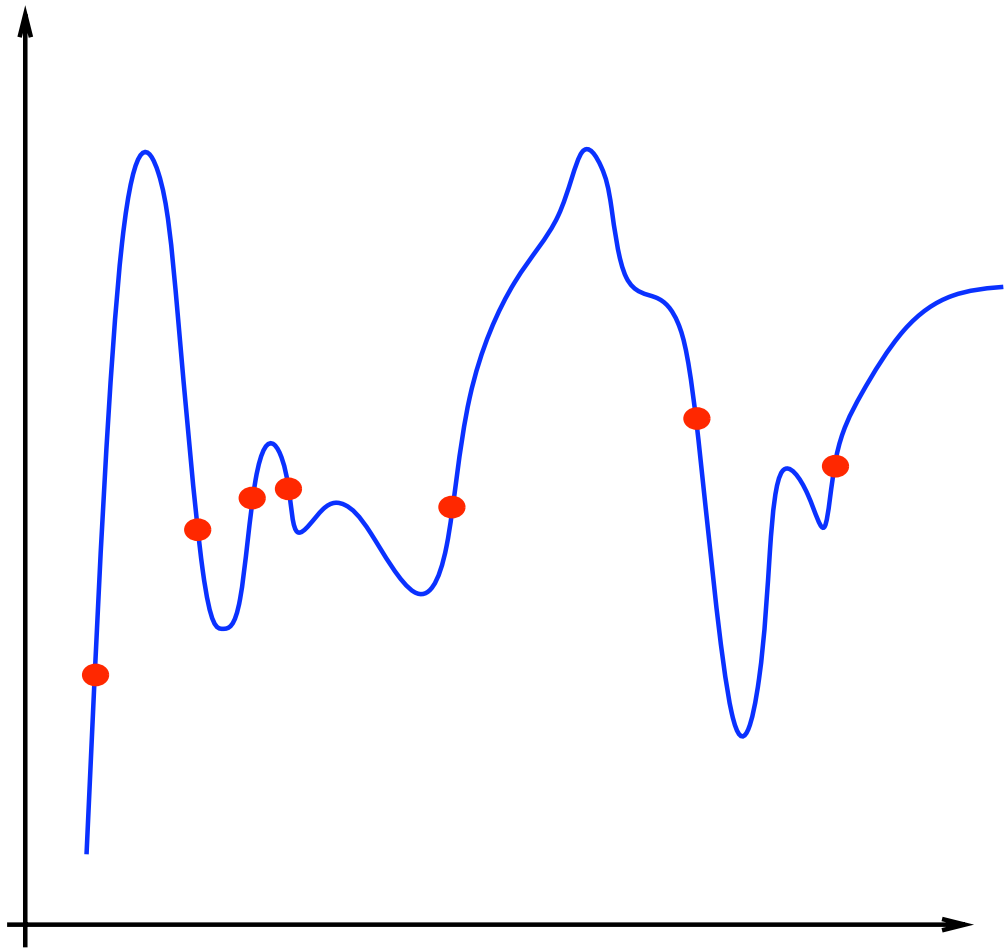
ERM and generalization: given a certain number of samples...



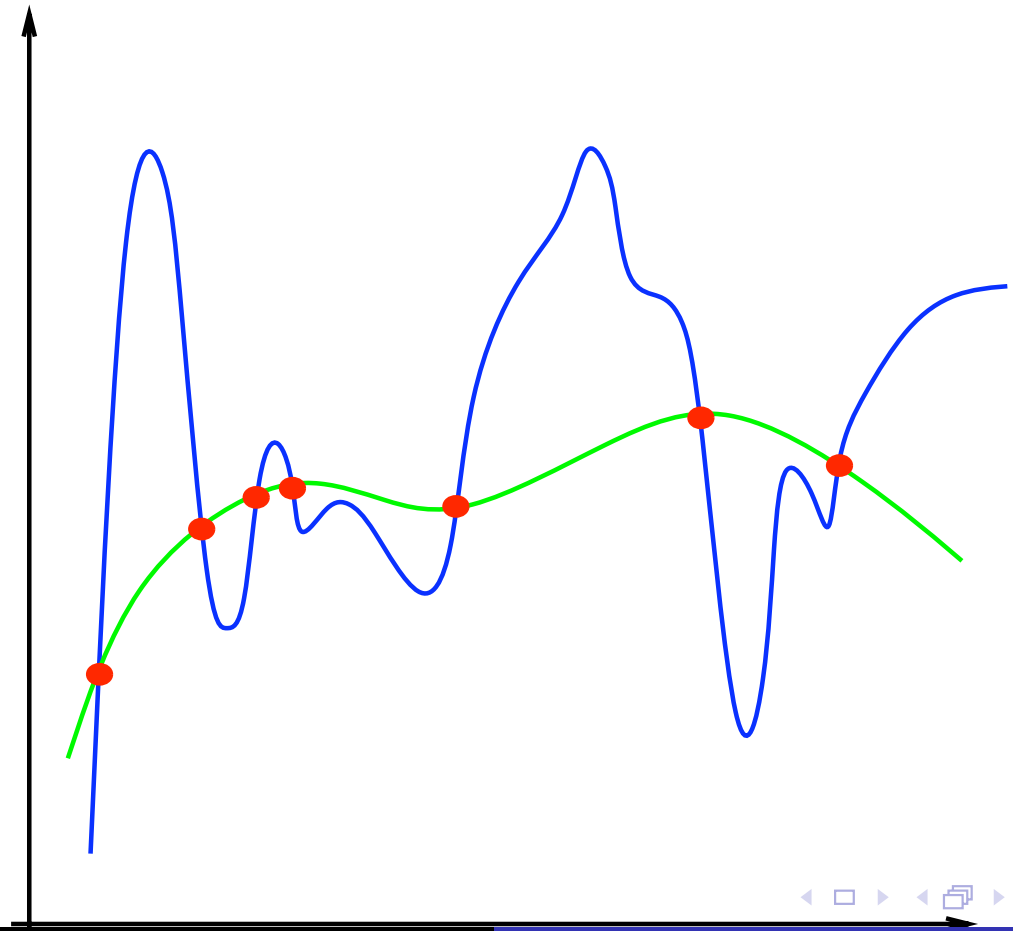
...suppose this is the “true” solution...



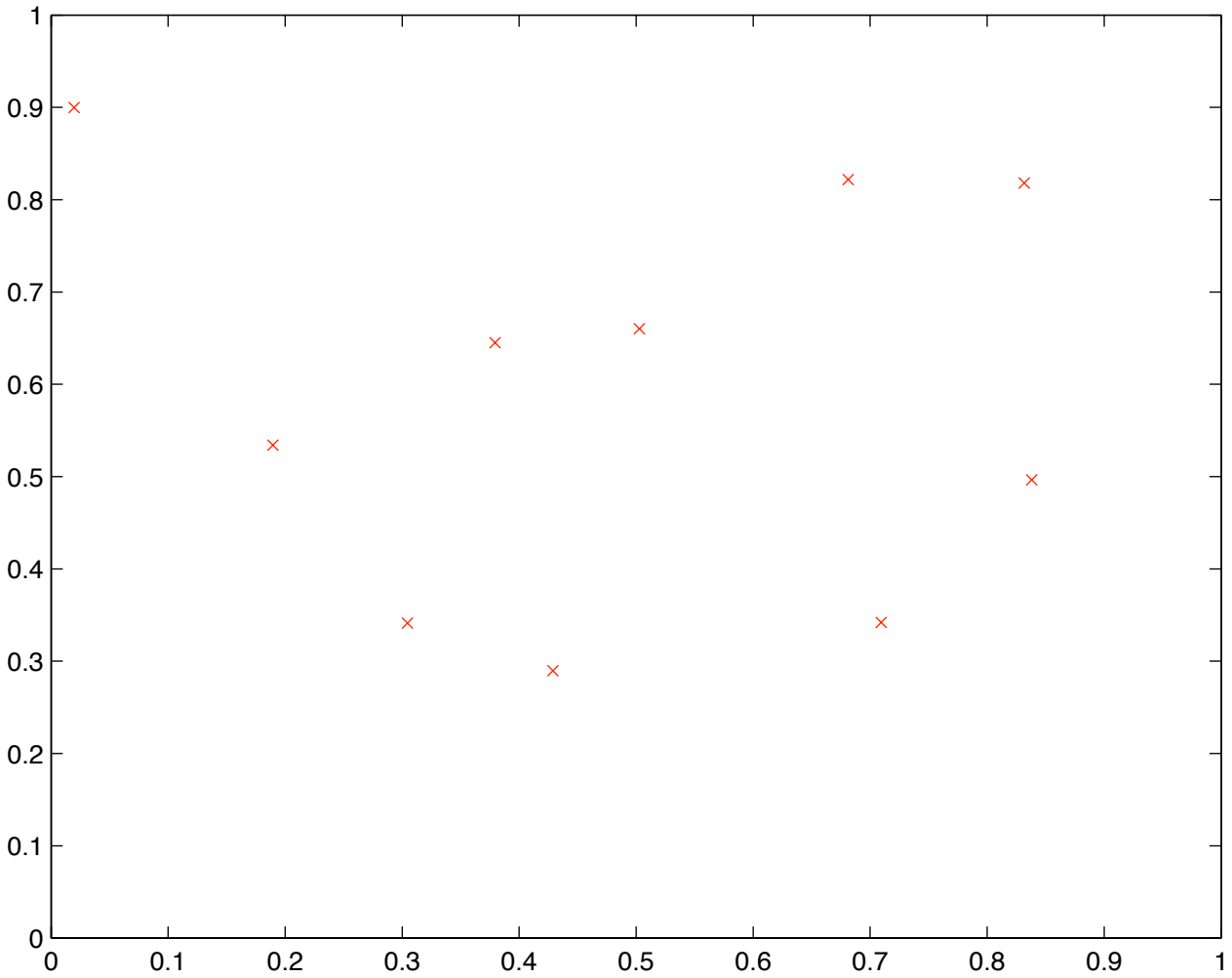
... but suppose ERM gives this solution.



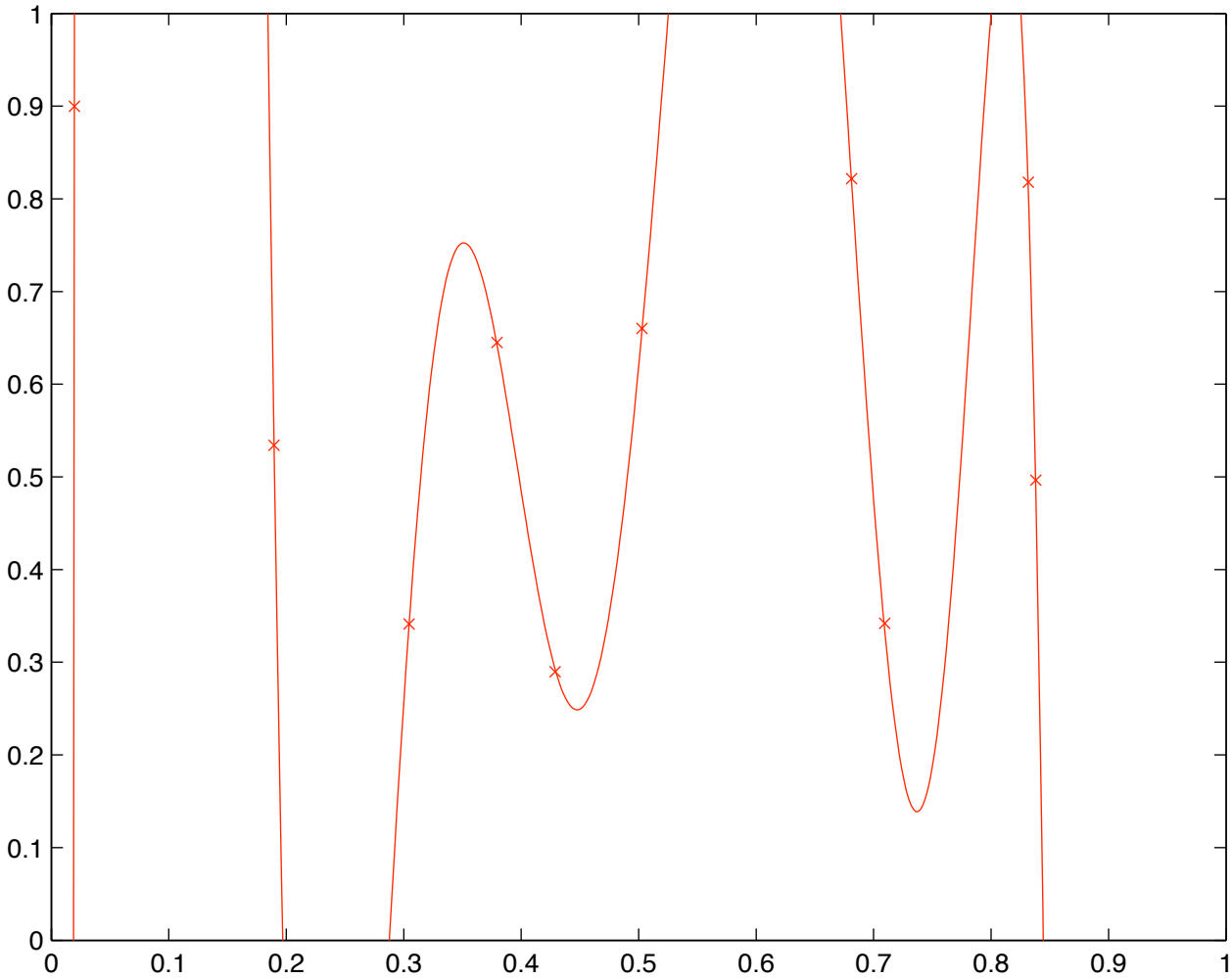
Under which conditions the ERM solution converges with increasing number of examples to the true solution? In other words...what are the conditions for generalization of ERM?



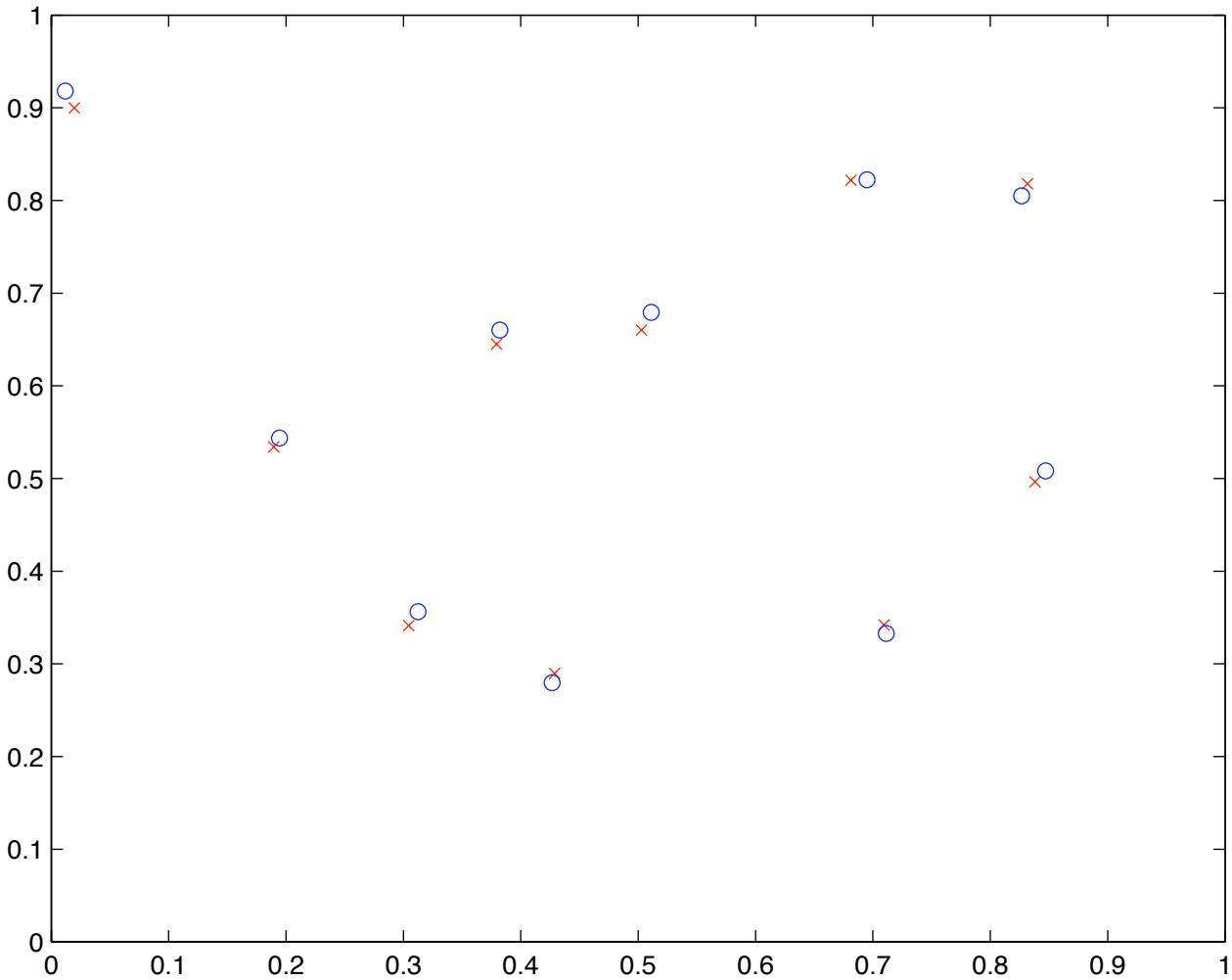
ERM and stability: given 10 samples...



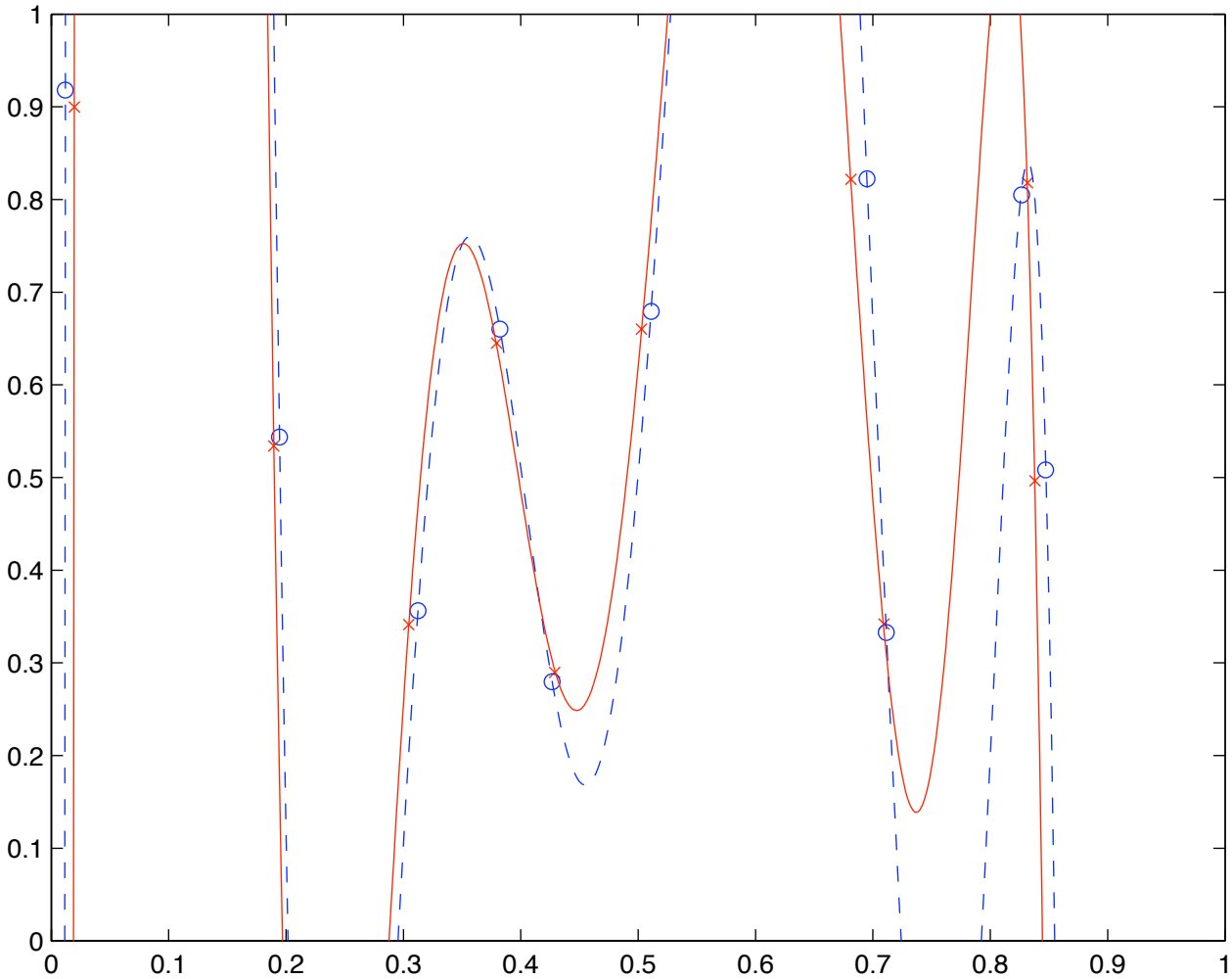
...we can find the smoothest interpolating polynomial (which degree?).



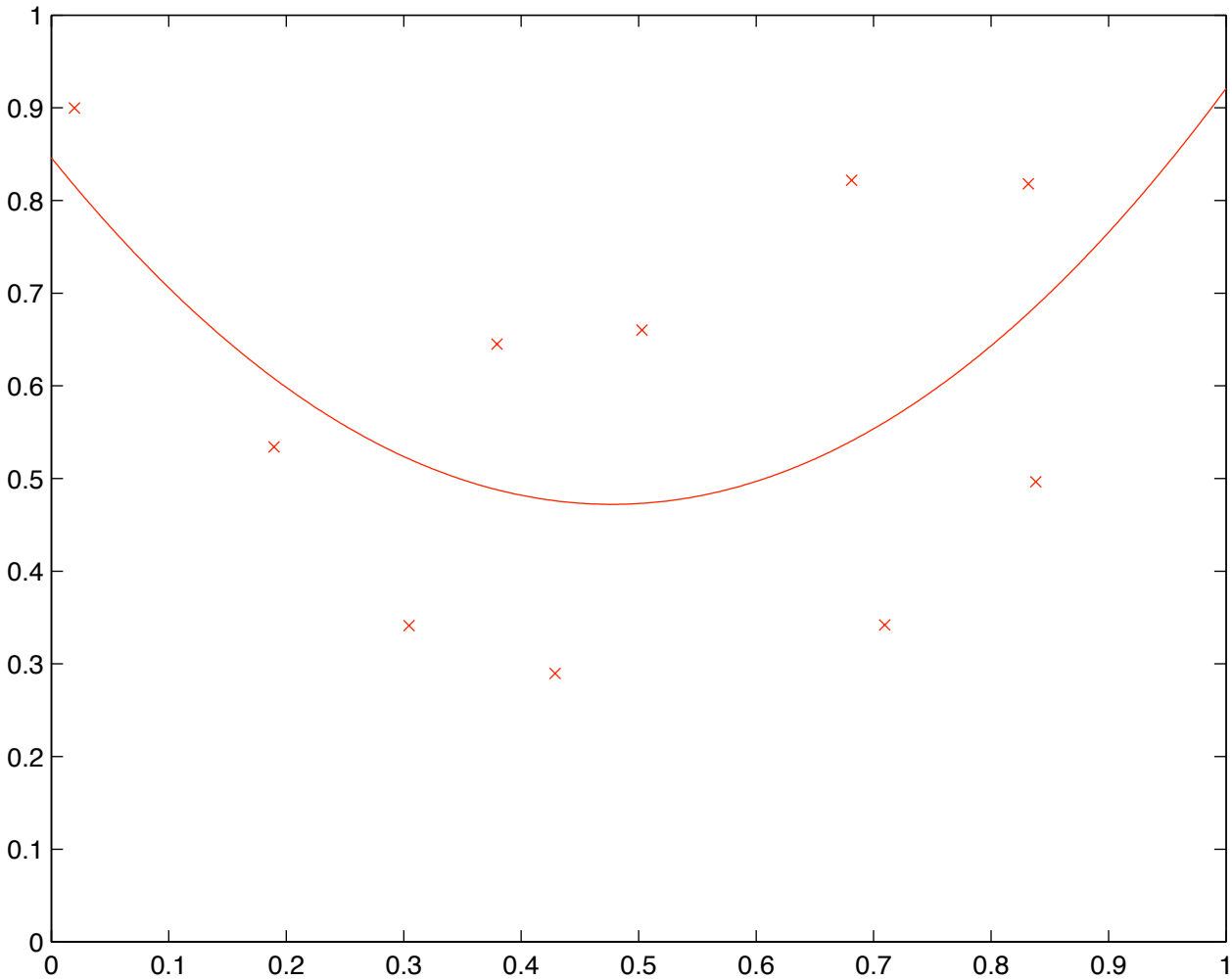
But if we perturb the points slightly...



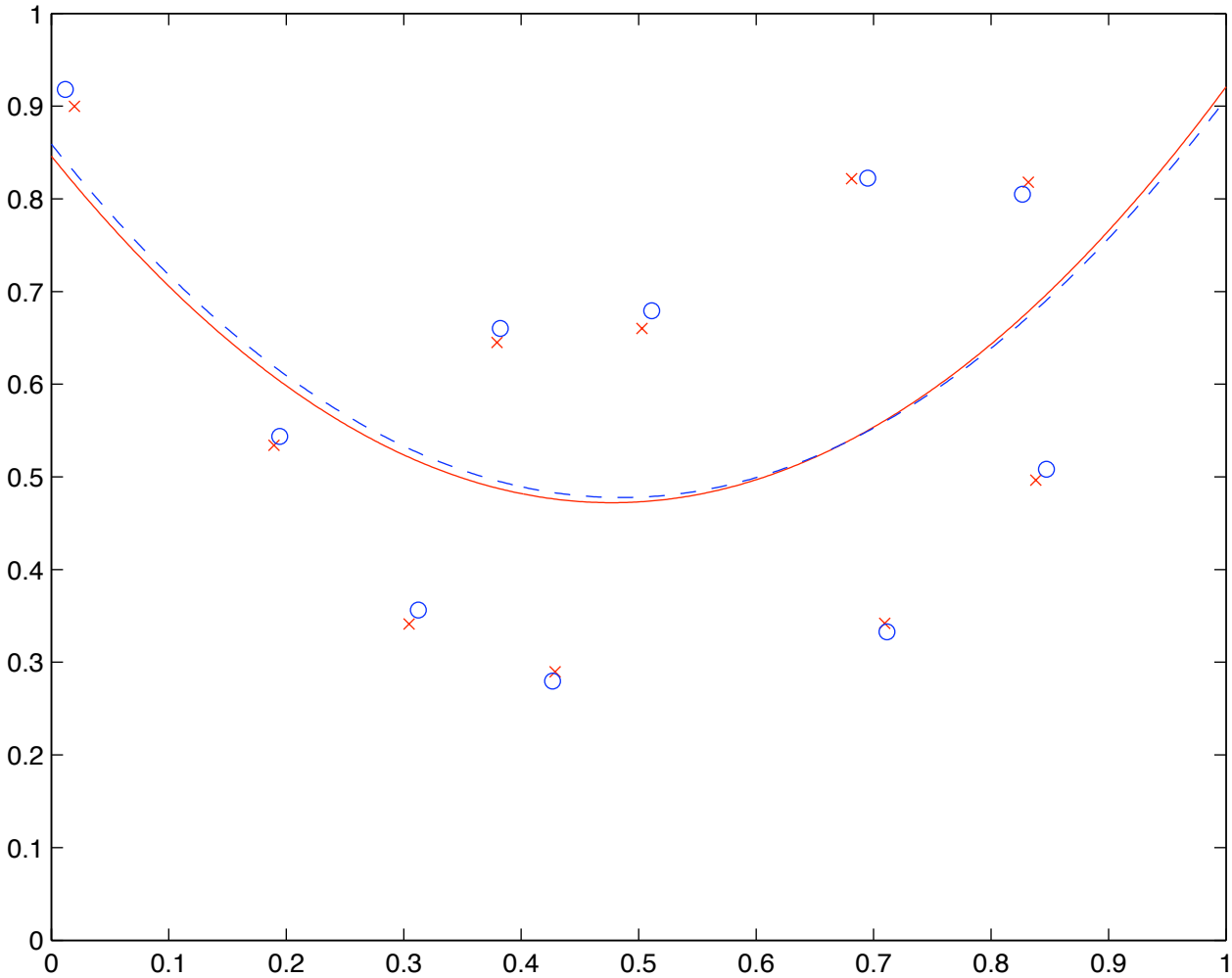
...the solution changes a lot!



If we restrict ourselves to degree two polynomials...



...the solution varies only a small amount under a small perturbation.



ERM: conditions for well-posedness (stability) and predictivity (generalization)

Since Tikhonov, it is well-known that a generally ill-posed problem such as ERM, can be guaranteed to be well-posed and therefore *stable* by an appropriate choice of $\mathcal{H}$. For example, compactness of $\mathcal{H}$ guarantees stability.
It seems intriguing that the *classical conditions for consistency of ERM* – thus quite a different property – consist of appropriately restricting $\mathcal{H}$. It seems that the same restrictions that make the approximation of the data stable, may provide solutions that generalize...

ERM: conditions for well-posedness (stability) and predictivity (generalization)

We would like to have a hypothesis space that yields generalization. Loosely speaking this would be a H for which the solution of ERM, say f_S is such that $|I_S[f_S] - I[f_S]|$ converges to zero in probability for n increasing.
Note that the above requirement is NOT the law of large numbers; the requirement for a fixed f that $|I_S[f] - I[f]|$ converges to zero in probability for n increasing IS the law of large numbers.

ERM: conditions for well-posedness (stability) and predictivity (generalization) in the case of regression and classification

Theorem [Vapnik and Červonenkis (71), Alon et al (97), Dudley, Giné, and Zinn (91)]

A (necessary) and sufficient condition for generalization (and consistency) of ERM is that $\mathcal{H}$ is uGC.

Definition

$\mathcal{H}$ is a (weak) uniform Glivenko-Cantelli (uGC) class if

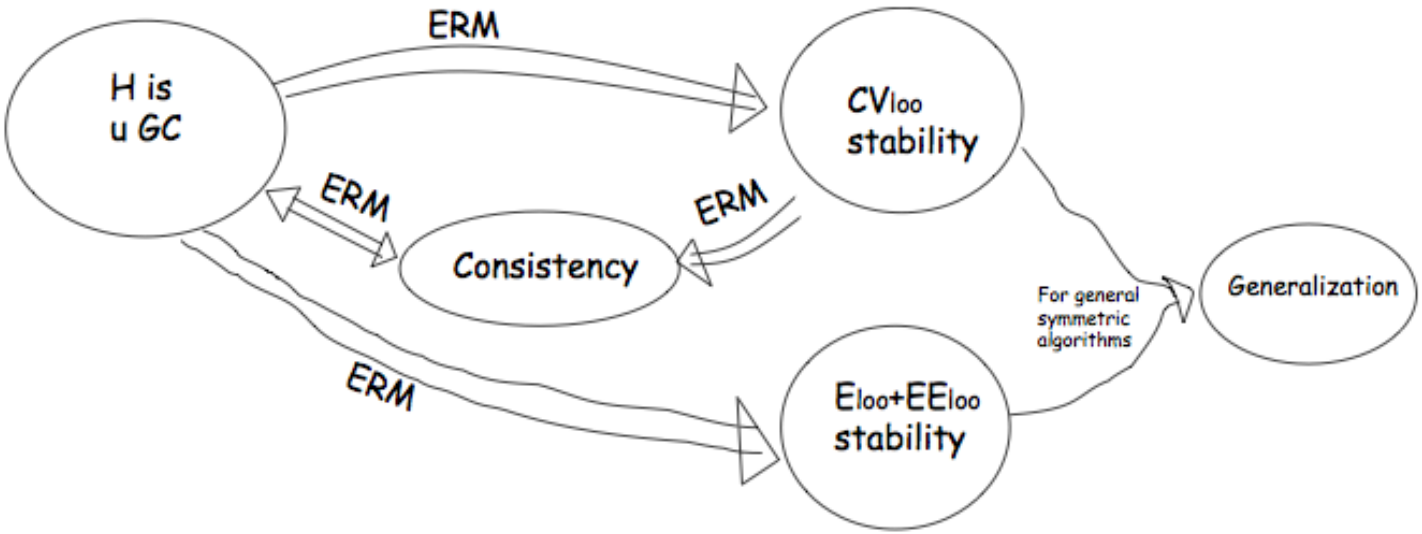
$$\forall \varepsilon > 0 \lim_{n \to \infty} \sup_{\mu} \mathbb{P}_S \left\{ \sup_{f \in \mathcal{H}} |I[f] - I_S[f]| > \varepsilon \right\} = 0.$$

ERM: conditions for well-posedness (stability) and predictivity (generalization) in the case of regression and classification

- The theorem (Vapnik et al.) says that a proper choice of the hypothesis space $\mathcal{H}$ ensures generalization of ERM (and consistency since for ERM generalization is necessary and sufficient for consistency and viceversa). Other results characterize uGC classes in terms of measures of complexity or capacity of H (such as VC dimension).
- A separate theorem (Niyogi, Poggio et al.) guarantees also stability (defined in a specific way) of ERM (for supervised learning). Thus with the appropriate definition of stability, *stability and generalization are equivalent for ERM.*

Thus the two desirable conditions for a supervised learning algorithm – generalization and stability – are equivalent (and they correspond to the same constraints on $\mathcal{H}$).

Key Theorem(s) Illustrated



Key Theorem(s)

Uniform Glivenko-Cantelli Classes

We say that $\mathcal{H}$ is a uniform Glivenko-Cantelli (uGC) class, if for all p,

$$\forall \epsilon > 0 \lim_{n \to \infty} \mathbb{P} \left\{ \sup_{f \in \mathcal{H}} |I[f] - I_n[f]| > \epsilon \right\} = 0.$$

A necessary and sufficient condition for consistency of ERM is that $\mathcal{H}$ is uGC.

See: [Vapnik and Červonenkis (71), Alon et al (97), Dudley, Giné, and Zinn (91)].

In turns the UGC property is equivalent to requiring $\mathcal{H}$ to have finite capacity: V_γ dimension in general and VC dimension in classification.

Key Theorem(s)

Uniform Glivenko-Cantelli Classes

We say that $\mathcal{H}$ is a uniform Glivenko-Cantelli (uGC) class, if for all p,

$$\forall \epsilon > 0 \lim_{n \to \infty} \mathbb{P} \left\{ \sup_{f \in \mathcal{H}} |I[f] - I_n[f]| > \epsilon \right\} = 0.$$

A necessary and sufficient condition for consistency of ERM is that $\mathcal{H}$ is uGC.
See: [Vapnik and Červonenkis (71), Alon et al (97), Dudley, Giné, and Zinn (91)].

In turns the UGC property is equivalent to requiring $\mathcal{H}$ to have finite capacity: V_γ dimension in general and VC dimension in classification.

Stability

$z = (x, y)$
$S = z_1, ..., z_n$
$S^i = z_1, ..., z_{i-1}, z_{i+1}, ...z_n$

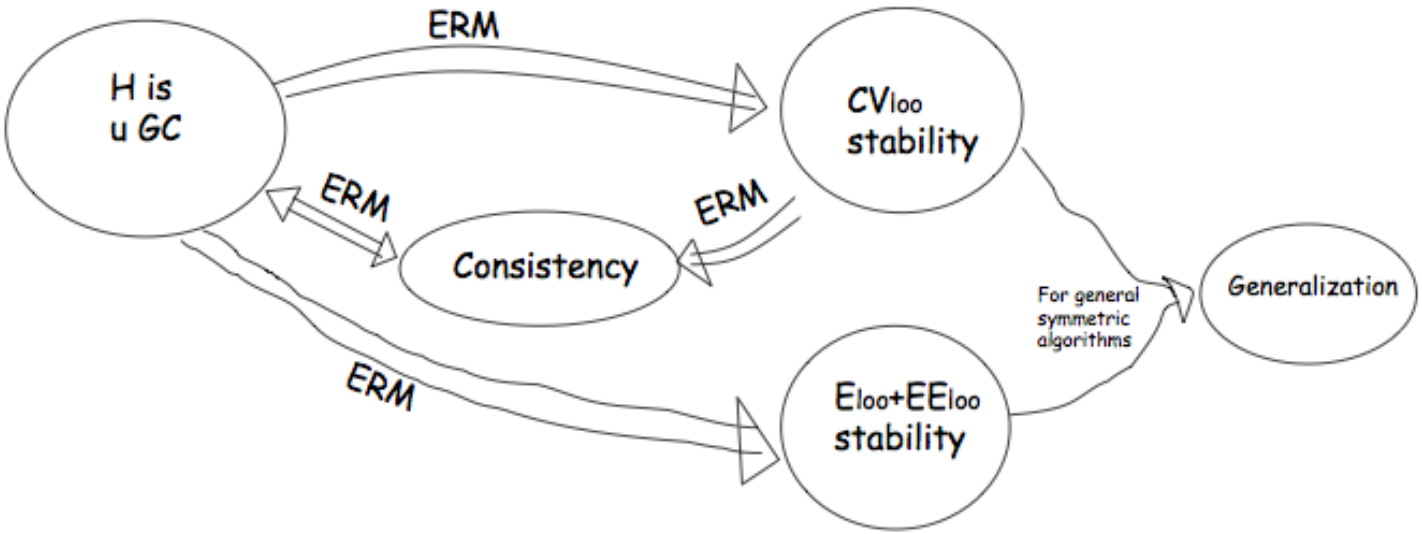
CV Stability

A learning algorithm A is CV_{loo} *stability* **if** for each n there exists a $\beta_{CV}^{(n)}$ and a $\delta_{CV}^{(n)}$ such that for all p

$$\mathbb{P}\left\{|V(f_{S^i}, z_i) - V(f_S, z_i)| \leq \beta_{CV}^{(n)}\right\} \geq 1 - \delta_{CV}^{(n)},$$

with $\beta_{CV}^{(n)}$ and $\delta_{CV}^{(n)}$ going to zero for $n \to \infty$.

Key Theorem(s) Illustrated



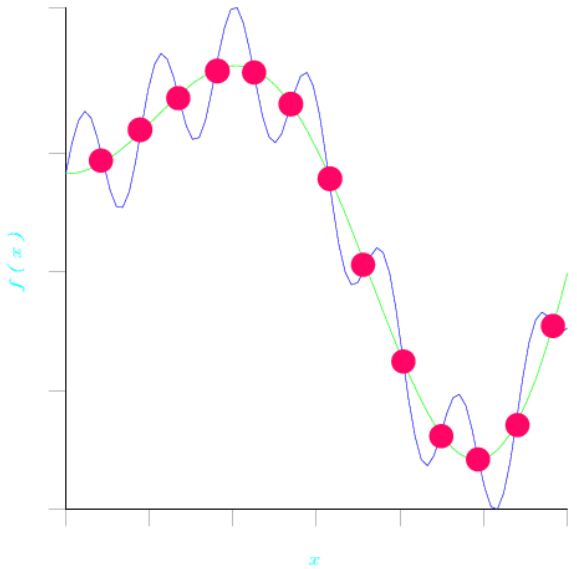
ERM and ill-posedness

Ill posed problems often arise if one tries to infer general laws from few data

- the hypothesis space is too large
- there are not enough data

In general ERM leads to ill-posed solutions because

- the solution may be too complex
- it may be not unique
- it may change radically when leaving one sample out



Regularization

Regularization is the classical way to restore well posedness (and ensure generalization). Regularization (originally introduced by Tikhonov independently of the learning problem) ensures *well-posedness* and (because of the above argument) *generalization* of ERM by constraining the hypothesis space $\mathcal{H}$. The direct way – minimize the empirical error subject to f in a ball in an appropriate $\mathcal{H}$ – is called *Ivanov regularization*. The indirect way is *Tikhonov regularization* (which is not strictly ERM).

Ivanov and Tikhonov Regularization

ERM finds the function in $(\mathcal{H})$ which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

which in general – for arbitrary hypothesis space $\mathcal{H}$ – is *ill-posed*.

- Ivanov regularizes by finding the function that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

  while satisfying $\mathcal{R}(f) \leq A$.

- Tikhonov regularization minimizes over the hypothesis space $\mathcal{H}$, for a fixed positive parameter γ, the regularized functional

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \gamma\mathcal{R}(f). \tag{2}$$

$\mathcal{R}(f)$ is the regulirizer, a penalization on f. In this course we will mainly discuss the case $\mathcal{R}(f) = \|f\|_K^2$ where $\|f\|_K^2$ is the norm in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, defined by the kernel K.

Tikhonov Regularization

As we will see in future classes

- Tikhonov regularization ensures well-posedness eg existence, uniqueness and especially *stability* (in a very strong form) of the solution
- Tikhonov regularization ensures generalization
- Tikhonov regularization is closely related to – but different from – Ivanov regularization, eg ERM on a hypothesis space $\mathcal{H}$ which is a ball in a RKHS.

Remarks on Foundations of Learning Theory

Intelligent behavior (at least learning) consists of optimizing under constraints. Constraints are key for solving computational problems; constraints are key for prediction. Constraints may correspond to rather general symmetry properties of the problem (eg time invariance, space invariance, invariance to physical units (pai theorem), universality of numbers and metrics implying normalization, etc.)

- Key questions at the core of learning theory:
  - generalization and predictivity *not* explanation
  - probabilities are unknown, only data are given
  - which constraints are needed to ensure generalization (therefore which hypotheses spaces)?
  - regularization techniques result usually in *computationally "nice"* and *well-posed* optimization problems

Statistical Learning Theory and Bayes

The Bayesian approach tends to ignore

- the issue of generalization (following the tradition in statistics of explanatory statistics);
- that probabilities are not known and that only data are known: assuming a specific distribution is a very strong – unconstrained by any Bayesian theory – seat-of-the-pants guess;
- the question of which priors are needed to ensure generalization;
- that the resulting optimization problems are often *computationally intractable* and possibly ill-posed optimization problems (for instance not unique).

The last point may be quite devastating for Bayesonomics: Montecarlo techniques etc. may just hide hopeless exponential computational complexity for the Bayesian approach to real-life problems, like exhastive search did initially for AI. A possibly interesting conjecture suggested by our stability results and the last point above, is that *ill-posed optimization problems or their ill-conditioned approximative solutions may not be predictive*!

Plan

- Part I: Basic Concepts and Notation
- Part II: Foundational Results
- Part III: Algorithms

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$\sum_{i,j=1}^{p}$$

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$p$$

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

Hypotheses Space

We are going to look at hypotheses spaces which are reproducing kernel Hilbert spaces.

- RKHS are **Hilbert spaces** of **point-wise defined** functions.
- They can be defined via a **reproducing kernel**, which is a symmetric positive definite function.

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

- functions in the space are (the completion of) linear combinations

$$f(x) = \sum_{i=1}^{p} K(x, x_i) c_i.$$

- the norm in the space is a natural measure of complexity

$$p$$

Examples of pd kernels

Very common examples of symmetric pd kernels are

- **Linear kernel**

$$K(x, x') = x \cdot x'$$

- **Gaussian kernel**

$$K(x, x') = \mathrm{e}^{-\frac{\|x-x'\|^2}{\sigma^2}}, \qquad \sigma > 0$$

- **Polynomial kernel**

$$K(x, x') = (x \cdot x' + 1)^d, \qquad d \in \mathbb{N}$$

For specific applications, designing an effective kernel is a challenging problem.

Kernel and Features

Often times kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \ldots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

setting

$$K(x, x') = \sum_{i=1}^{p} \phi_j(x)\phi_j(x').$$

Ivanov regularization

We can regularize by explicitly restricting the hypotheses space $\mathcal{H}$ — for example to a ball of radius R.

Ivanov regularization

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)$$

subject to

$$\|f\|_{\mathcal{H}}^2 \leq R.$$

The above algorithm corresponds to a constrained optimization problem.

Tikhonov regularization

Regularization can also be done implicitly via penalization

Tikhonov regularizarion

$$\arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 .$$

λ is the regularization parameter trading-off between the two terms.

The above algorithm can be seen as the Lagrangian formulation of a constrained optimization problem.

The Representer Theorem

An important result

The minimizer over the RKHS $\mathcal{H}$, f_S, of the regularized empirical functional

$$I_S[f] + \lambda \|f\|_{\mathcal{H}}^2,$$

can be represented by the expression

$$f_n(x) = \sum_{i=1}^{n} c_i K(x_i, x),$$

for some $(c_1, \ldots, c_n) \in \mathbb{R}$.

Hence, minimizing over the (possibly infinite dimensional) Hilbert space, *boils down to minimizing over* $\mathbb{R}^n$.

SVM and RLS

The way the coefficients $\mathbf{c} = (c_1, \ldots, c_n)$ are computed depend on the loss function choice.

- RLS: Let Let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{K}_{i,j} = K(x_i, x_j)$ then $\mathbf{c} = (\mathbf{K} + \lambda n I)^{-1}\mathbf{y}$.
- SVM: Let $\alpha_i = y_i c_i$ and $\mathbf{Q}_{i,j} = y_i K(x_i, x_j) y_j$

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2}\alpha^T \mathbf{Q}\alpha$$

$$\text{subject to :} \quad \sum_{i=1}^n y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, n$$

SVM and RLS

The way the coefficients $\mathbf{c} = (c_1, \ldots, c_n)$ are computed depend on the loss function choice.

- RLS: Let Let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{K}_{i,j} = K(x_i, x_j)$ then $\mathbf{c} = (\mathbf{K} + \lambda n I)^{-1}\mathbf{y}$.
- SVM: Let $\alpha_i = y_i c_i$ and $\mathbf{Q}_{i,j} = y_i K(x_i, x_j) y_j$

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2}\alpha^T \mathbf{Q}\alpha$$

$$\text{subject to :} \quad \sum_{i=1}^n y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, n$$

Bayes Interpretation

- **Empirical risk minimization** is ML.

$$p(\mathbf{Y}|f,\mathbf{X}) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

- **Linear RLS** is MAP.

$$p(\mathbf{Y},f|\mathbf{X}) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - \langle x_i,\theta\rangle)^2} \cdot e^{-\frac{\lambda}{2}\theta^T\theta}$$

- **Kernel RLS** is also MAP.

$$p(\mathbf{Y},f|\mathbf{X}) \propto e^{-\frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i)^2} \cdot e^{-\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2}$$

Regularization approach

More generally we can consider:

$$I_n(f) + \lambda R(f)$$

where, $R(f)$ is a regularizing functional.

- Sparsity based methods
- Manifold learning
- Multiclass
- ...

Summary

- statistical learning as framework to deal with uncertainty in data.
- non free lunch and key theorem: no prior, no learning.
- regularization as a fundamental tool for enforcing a prior and ensure stability and generalization

Kernel and Data Representation

In the above reasoning the kernel and the hypotheses space define a representation/parameterization of the problem and hence play a special role.

Where do they come from?

- There are a few off the shelf choices (Gaussian, polynomial etc.)
- Often they are the product of problem specific engineering.

Are there principles– applicable in a wide range of situations– to design effective data representation?

Kernel and Data Representation

In the above reasoning the kernel and the hypotheses space define a representation/parameterization of the problem and hence play a special role.

Where do they come from?

- There are a few off the shelf choices (Gaussian, polynomial etc.)
- Often they are the product of problem specific engineering.

Are there principles– applicable in a wide range of situations– to design effective data representation?

Kernel and Data Representation

In the above reasoning the kernel and the hypotheses space define a representation/parameterization of the problem and hence play a special role.

Where do they come from?

- There are a few off the shelf choices (Gaussian, polynomial etc.)
- Often they are the product of problem specific engineering.

Are there principles– applicable in a wide range of situations– to design effective data representation?

Kernel and Data Representation

In the above reasoning the kernel and the hypotheses space define a representation/parameterization of the problem and hence play a special role.

Where do they come from?

- There are a few off the shelf choices (Gaussian, polynomial etc.)
- Often they are the product of problem specific engineering.

Are there principles– applicable in a wide range of situations– to design effective data representation?