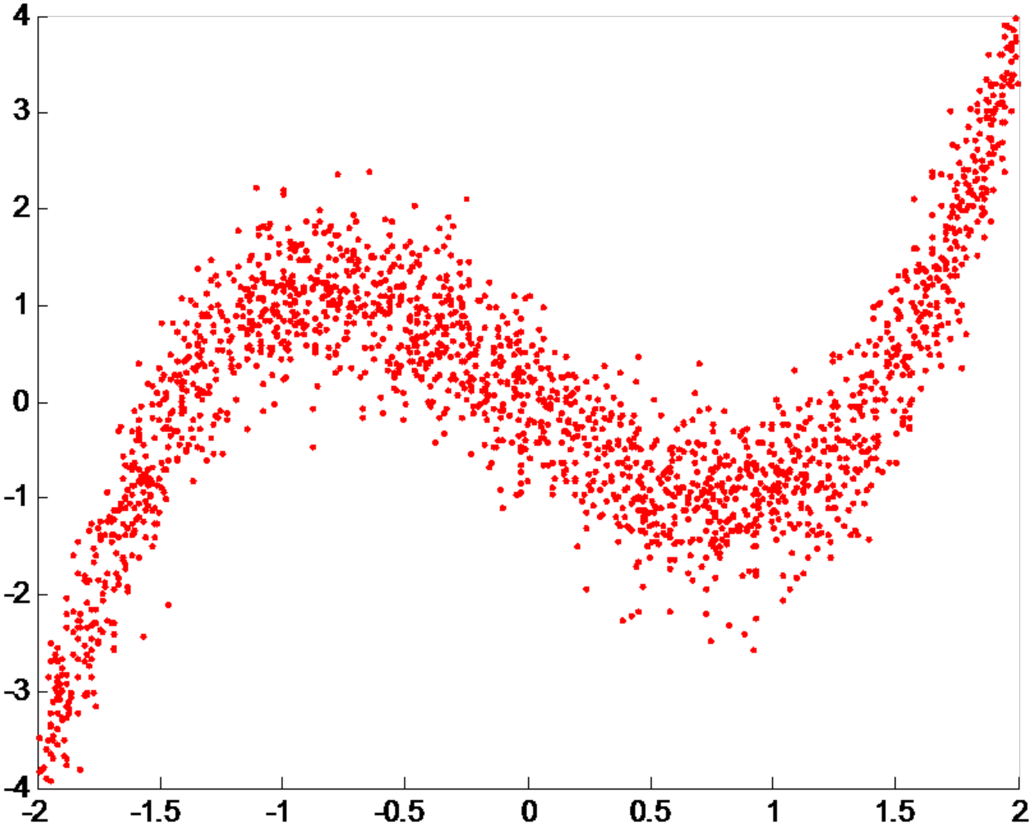
Statistical Learning Theory

- When a pattern or regularity is found in data, is it real or spurious?
- i.e., can it be used for prediction?
- Learning theories address this issue.
- Statistical Learning Theory was started by Vapnik and Chervonenkis in the 1960s (aka VC theory).
- Provides bounds on prediction error.
- An early motivation for SVMs.
- Main concepts: capacity and VC dimension.

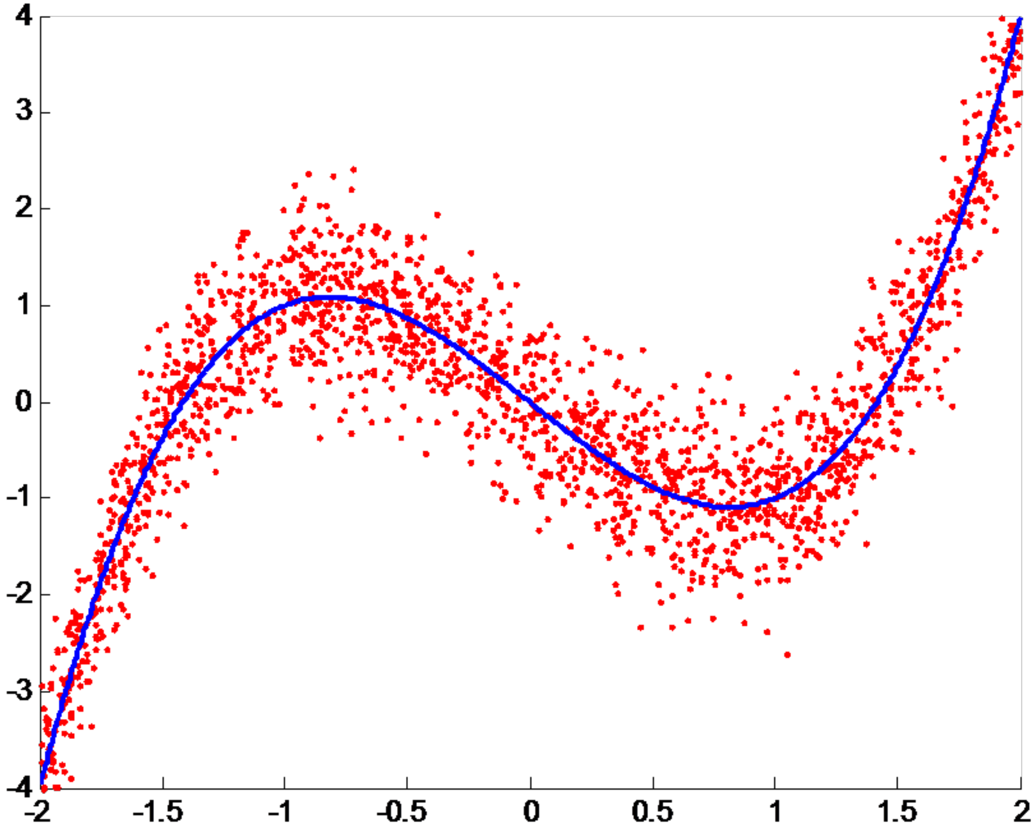
Prediction

- Problem: given x, predict y.
  - i.e., find a function f such that y = f(x).
- Usually, no function works perfectly.
- Instead, find a function that minimizes the expected error, which is called the *risk.*
- Error is measured by a cost function, c(x,y,z), where z = predicted value of y given x.
- Assume data comes from a distribution, P(x,y).

A data distribution for a regression problem



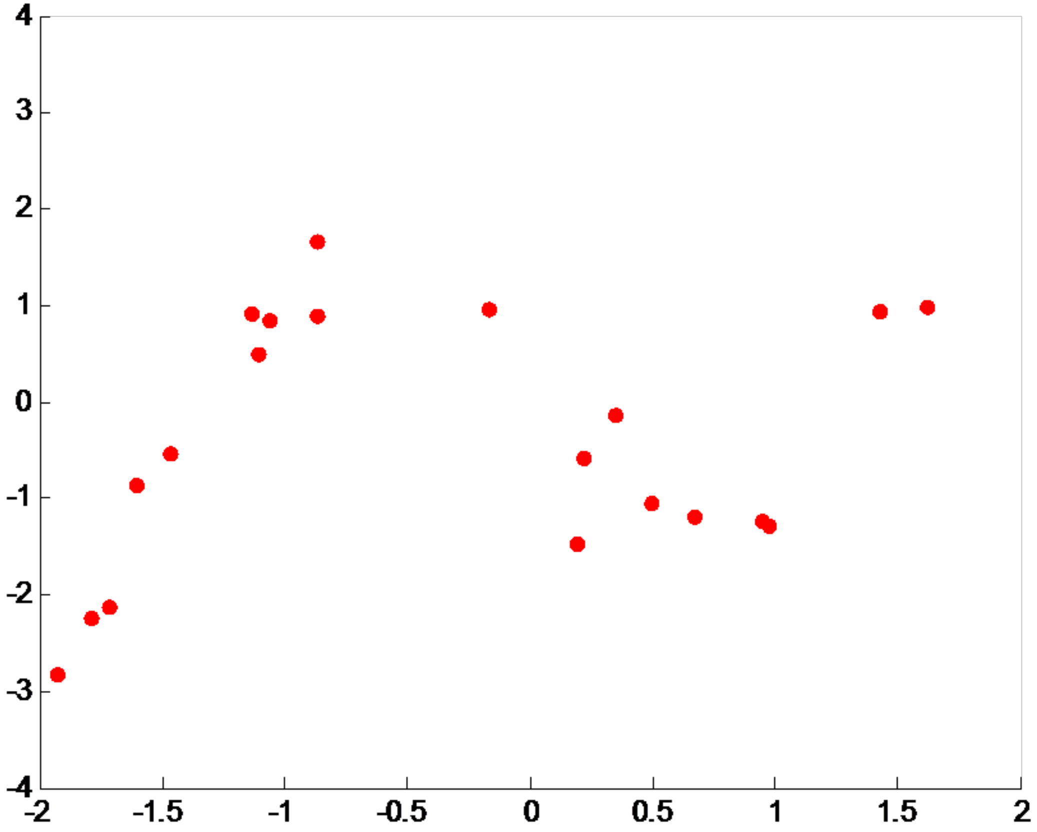
The function that minimizes risk (as measured by squared error)



Limitations

- We do not know the data distribution.
- We may not even know its form (e.g., Gaussian, Poisson, etc).
- We only have a sample of data from the distribution.

A random sample of 20 points



Actual risk v.s. Empirical risk

- R(f) denotes the actual risk of function f.
  - We cannot minimize actual risk, since we do not know the data distribution.
- R_{emp}(f) denotes the empirical risk of function f.
  - We can minimize empirical risk, since it depends only on the data sample.
- But we must be careful, since a function with low empirical risk can have high actual risk.
  - This is called *overfitting.*

No Free Lunch Theorem

If the class of functions is *completely unrestricted,* then

- Two functions can fit the training data perfectly but make completely opposite predictions.
- Since they behave identically on the training data, it is impossible to say which one makes better predictions.
- Learning and prediction are therefore impossible.

The Importance of the Set of Functions

What about allowing *all* functions from $\mathcal{X}$ to $\{\pm 1\}$?

Training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathcal{X} \times \{\pm 1\}$
Test patterns $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{\bar{m}} \in \mathcal{X}$,
such that $\{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{\bar{m}}\} \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} = \{\}$.

For any f there exists f^* s.t.:

1. $f^*(\mathbf{x}_i) = f(\mathbf{x}_i)$ for all i
2. $f^*(\bar{\mathbf{x}}_j) \neq f(\bar{\mathbf{x}}_j)$ for all j.

Based on the training set alone, there is *no* means of choosing which one is better. On the test set, however, they give *opposite* results. There is 'no free lunch' [32, 73].
$\longrightarrow$ a restriction must be placed on the *functions* that we allow

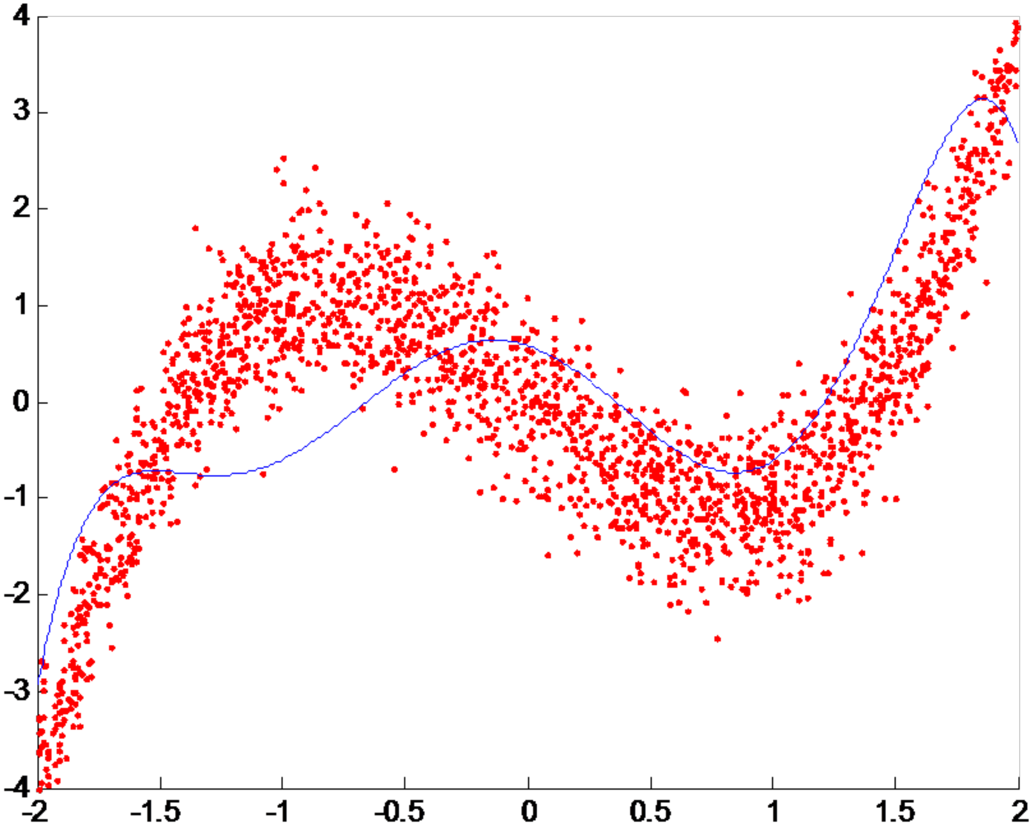
Restricting the class of functions

- To make accurate predictions, the class of functions must be restricted.
  - This is true of *any* approach to machine learning, statistical or non-statistical.
- In statistical learning theory, one limits the *capacity* of the class of functions (e.g., via the VC dimension).
- In Bayesian learning, one places *prior distributions* over the class of functions.

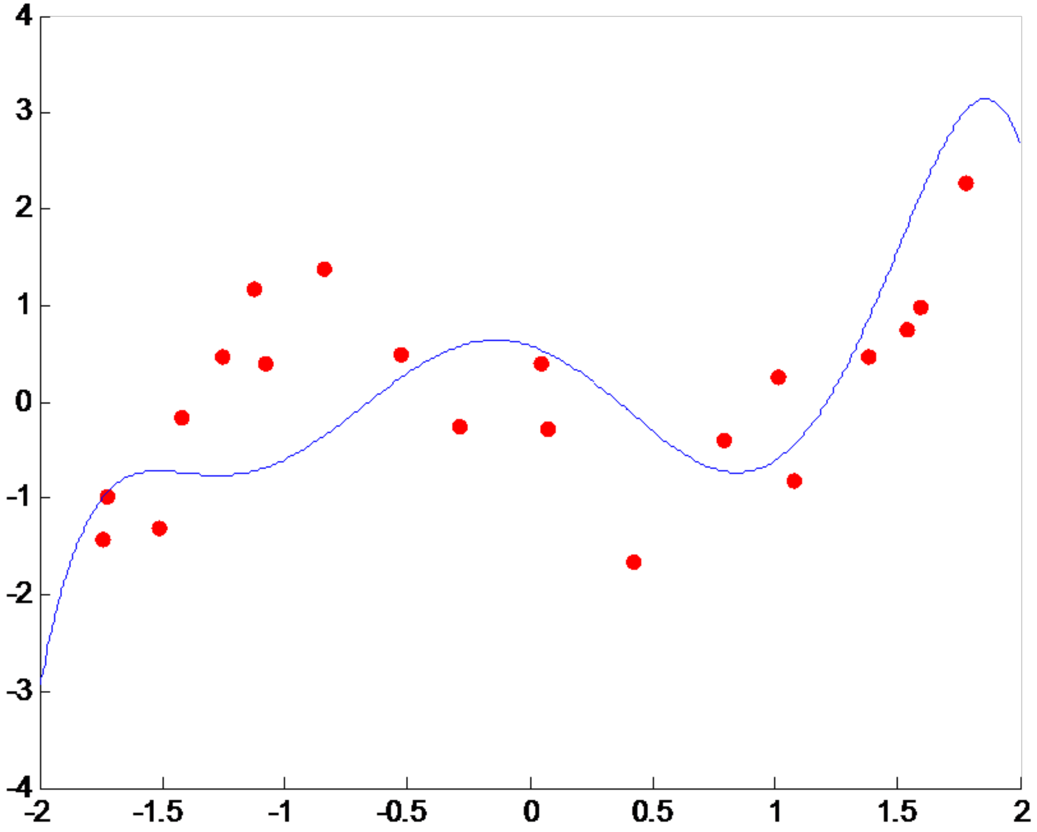
Empirical risk is a random variable

- Given a fixed function, f, the empirical risk depends only on the data set.
- Since the data set is random, the empirical risk is random.
- $R_{emp}(f)$ is therefore a random variable.
- As such, it has a mean and variance.

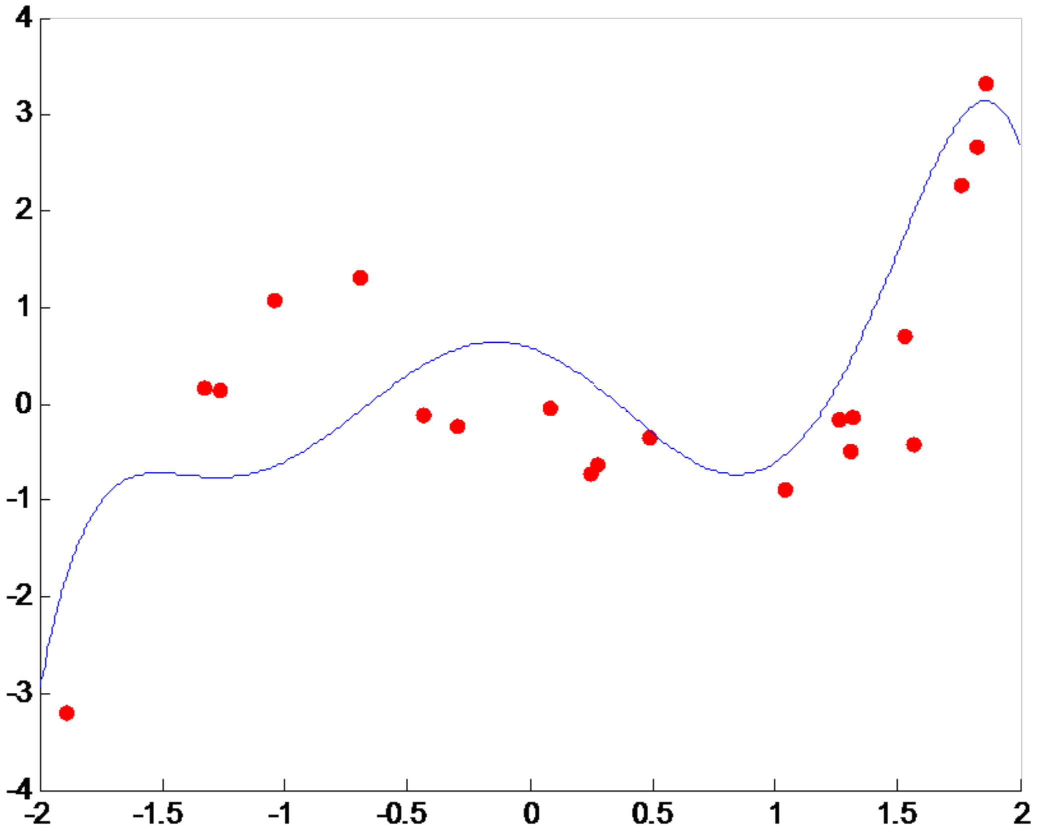
R(f): the actual risk of function f depends on the data distribution



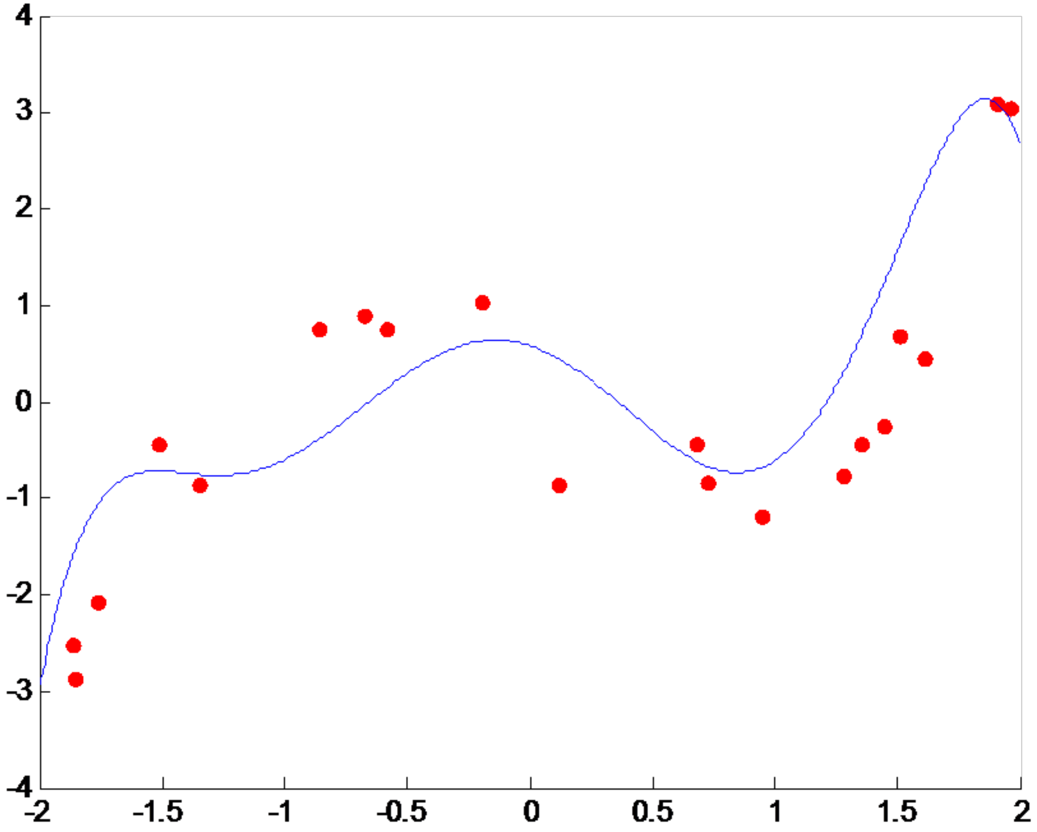
$R_{emp}(f)$: empirical risk of function f for data set 1



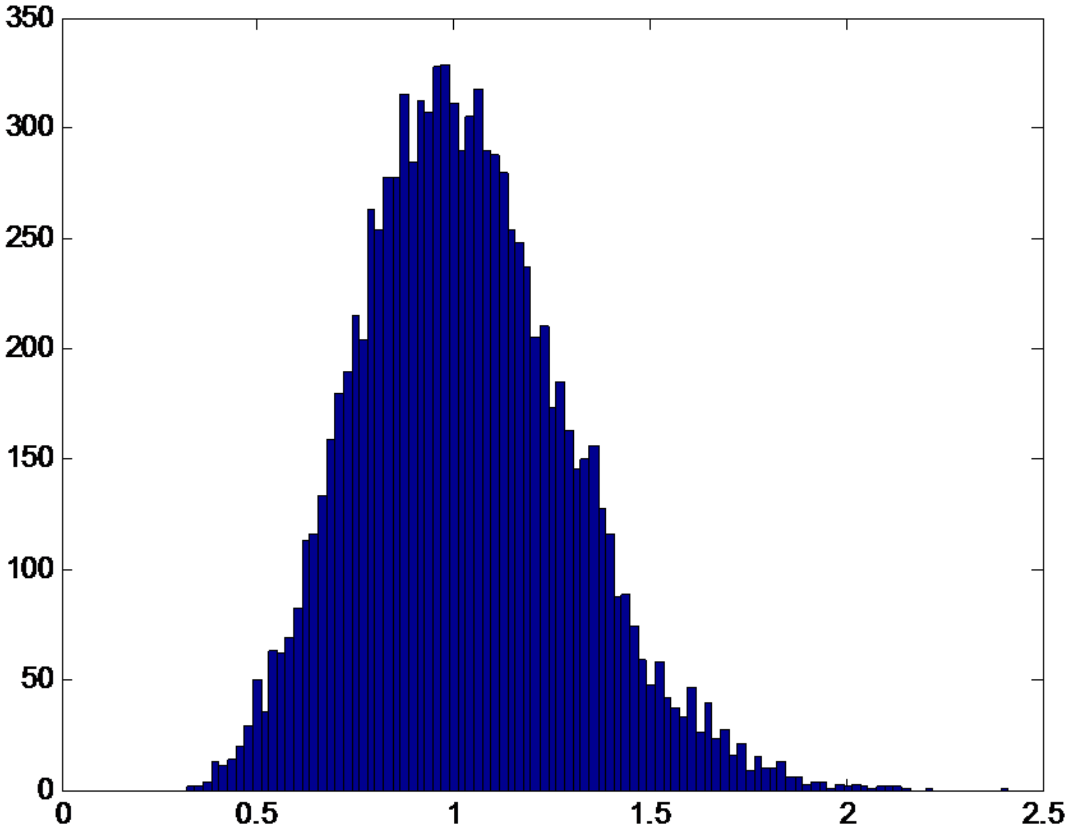
$R_{emp}(f)$: empirical risk of function f for data set 2



$R_{emp}(f)$: empirical risk of function f for data set 3



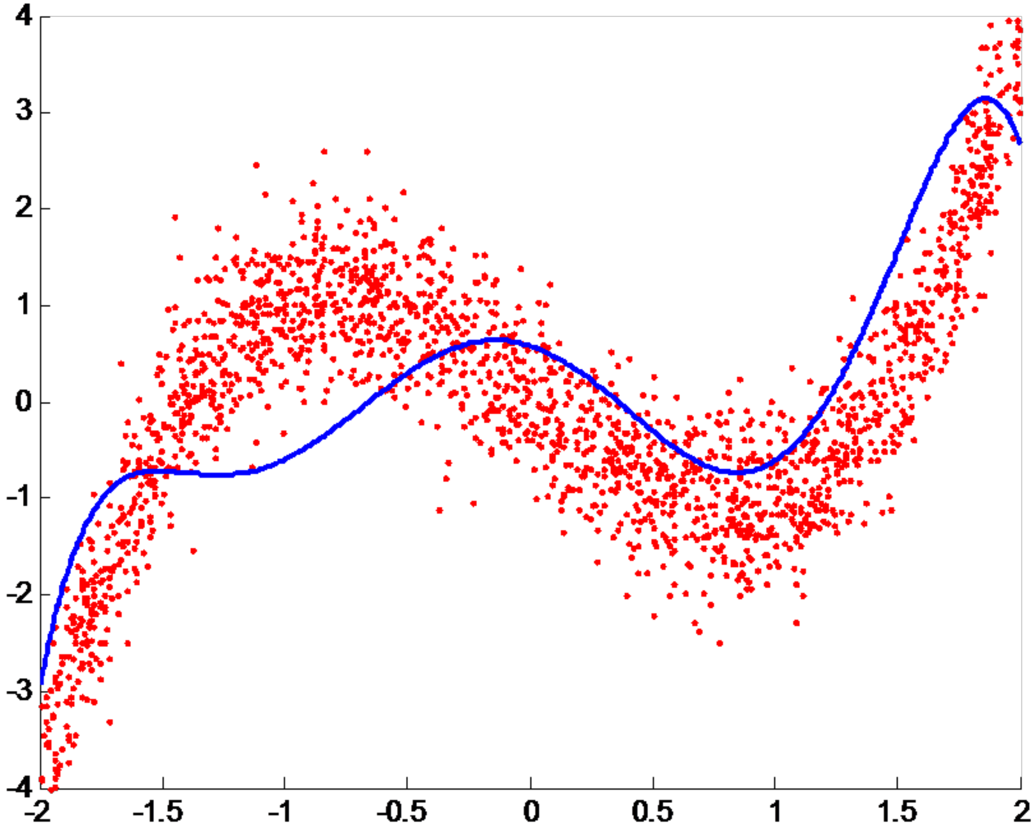
Histogram of $R_{emp}(f)$ for 10,000 data sets of 20 points each



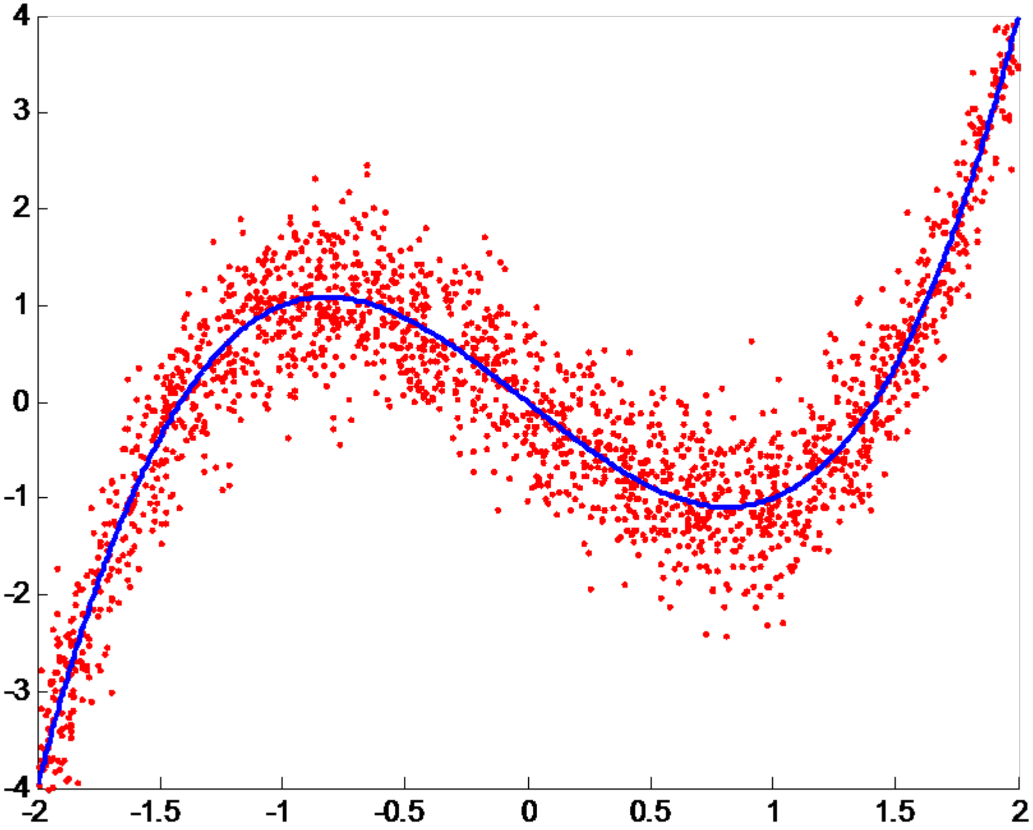
Different functions, different random variables

- Each function has its own actual risk and empirical risk.
- The empirical risks for different functions represent different random variables.
  - i.e., If f_1, f_2 and f_3 are different functions, then $R_{emp}(f_1)$, $R_{emp}(f_2)$ and $R_{emp}(f_3)$ are different random variables.

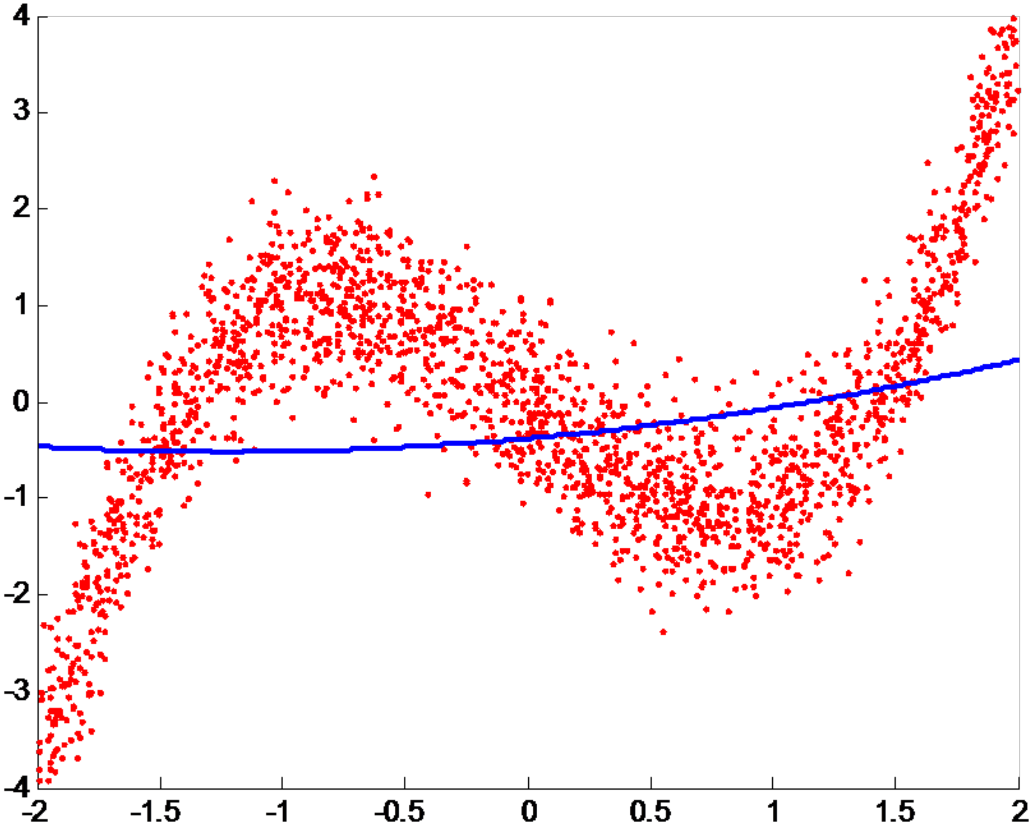
Function f_1



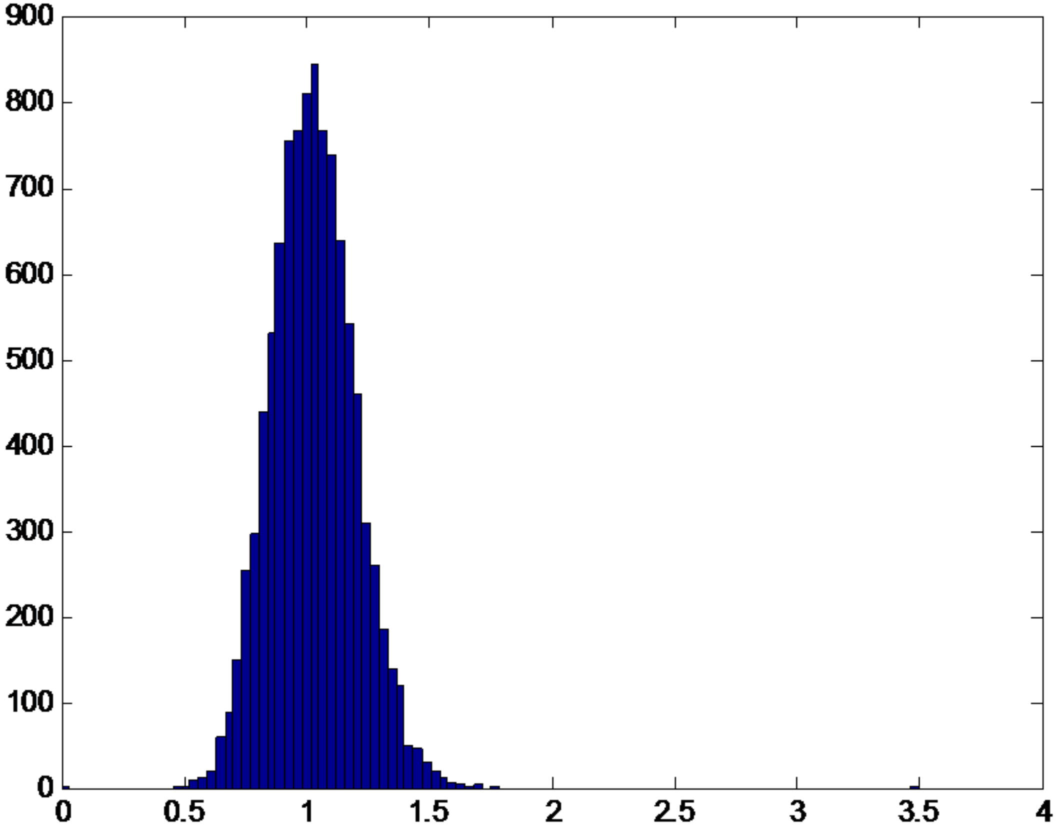
Function f_2



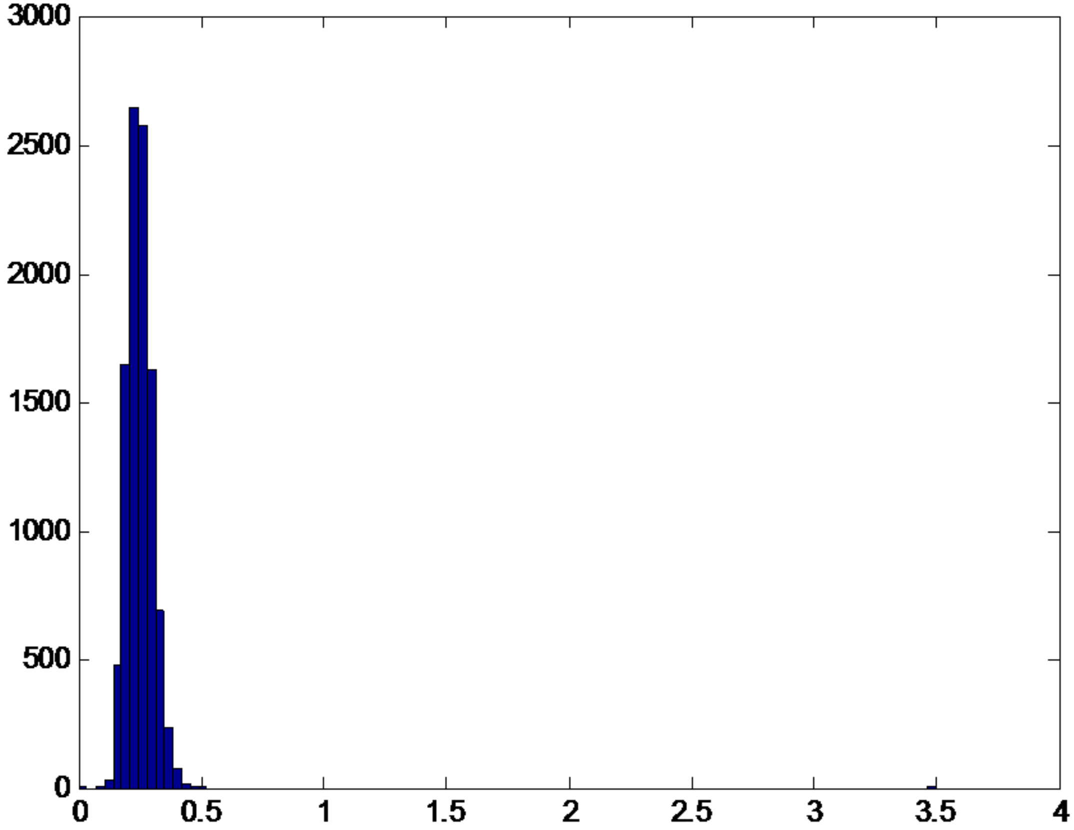
Function f_3



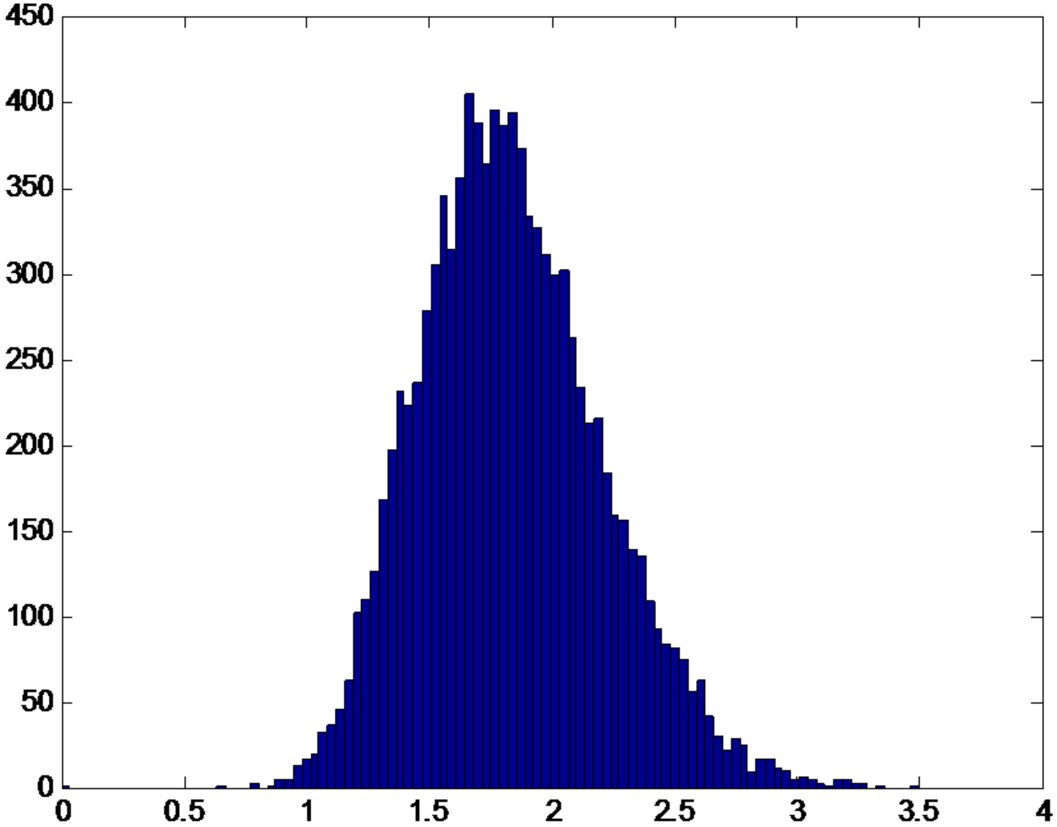
Histogram of empirical risk of f_1 for 10,000 data sets of 50 points each



Histogram of empirical risk of f_2 for 10,000 data sets of 50 points each



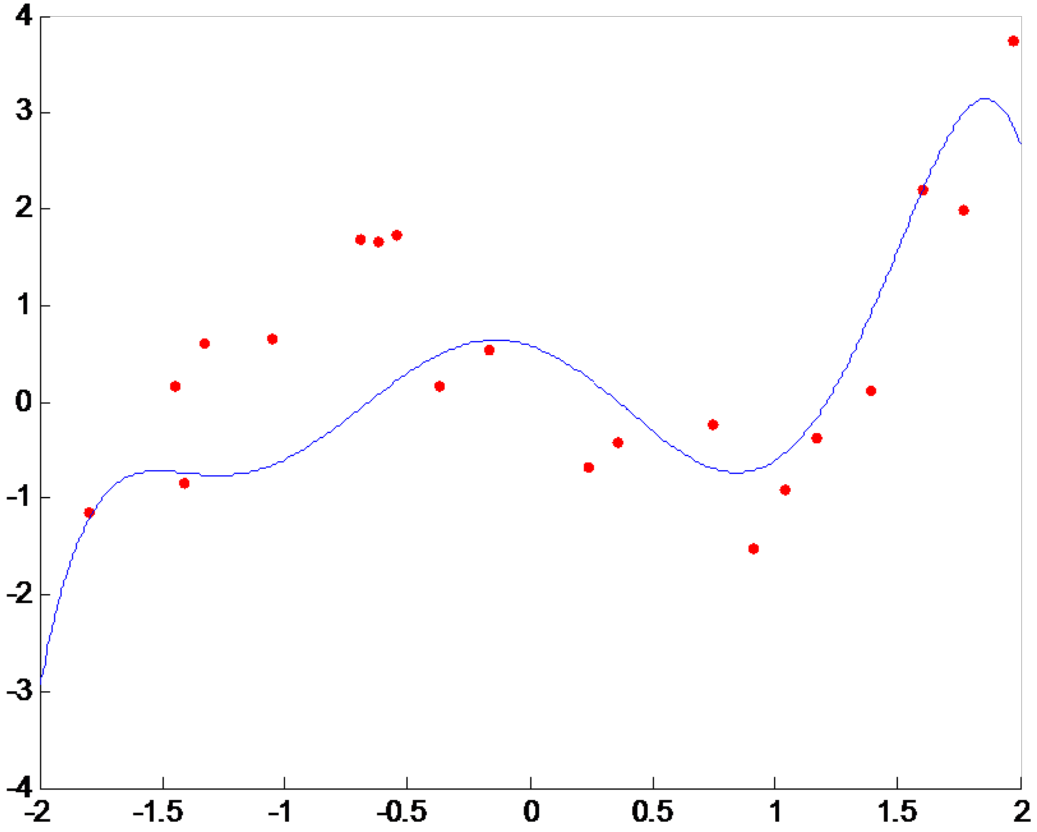
Histogram of empirical risk of f_3 for 10,000 data sets of 50 points each



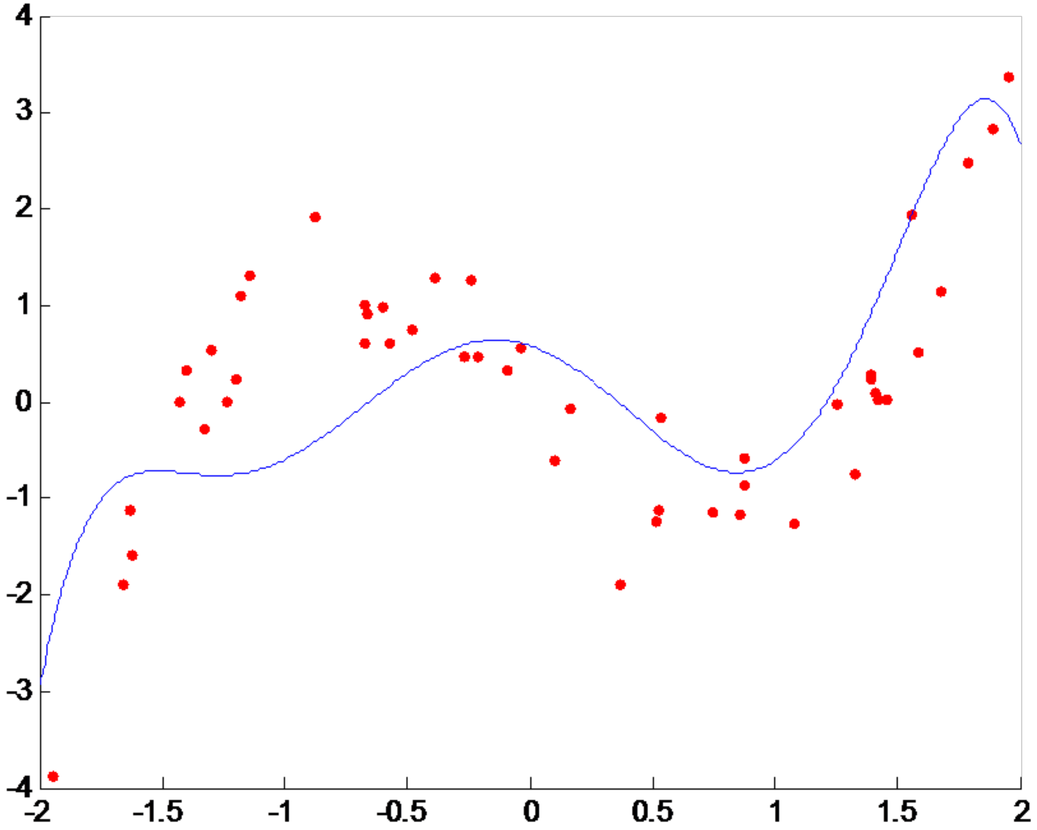
Properties of empirical risk

- Empirical risk is an *unbiased estimate* of actual risk: $E[R_{emp}(f)] = R(f)$.
- The variance of the empirical risk decreases as the size of the data sample increases.
- Thus, as the data sample gets larger, then with high probability, $R_{emp}(f)$ gets closer to $R(f)$.
- This is called *convergence in probability.*

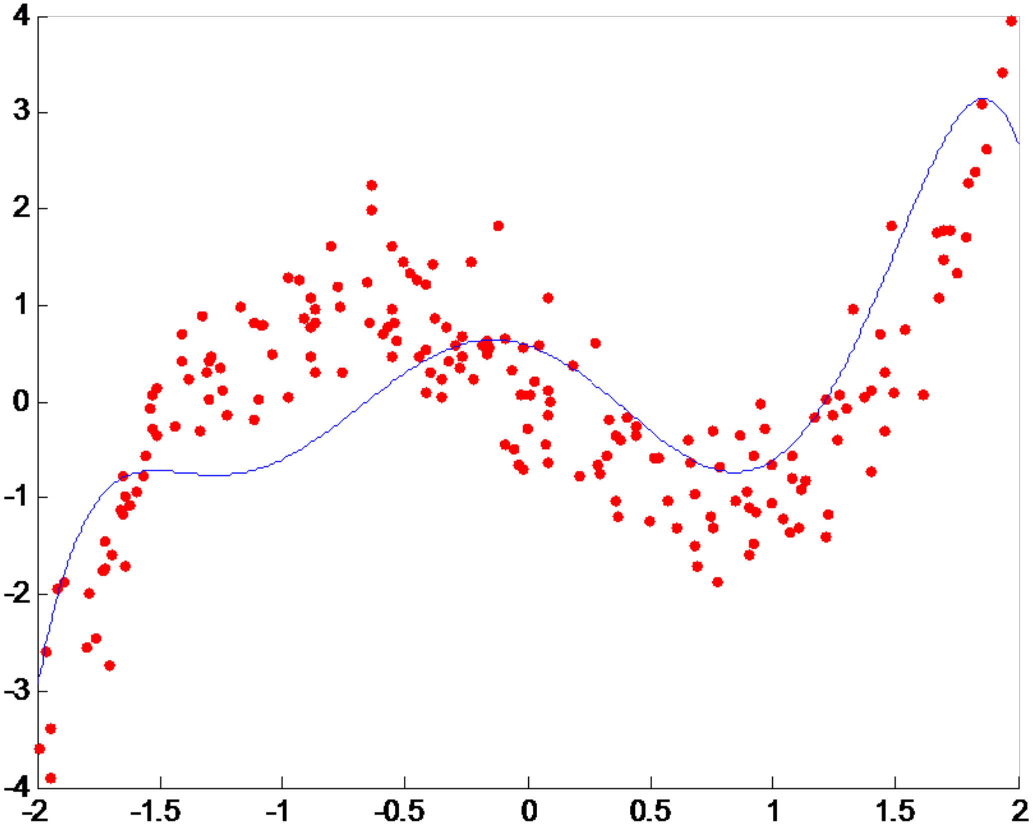
Function f and 20 data points



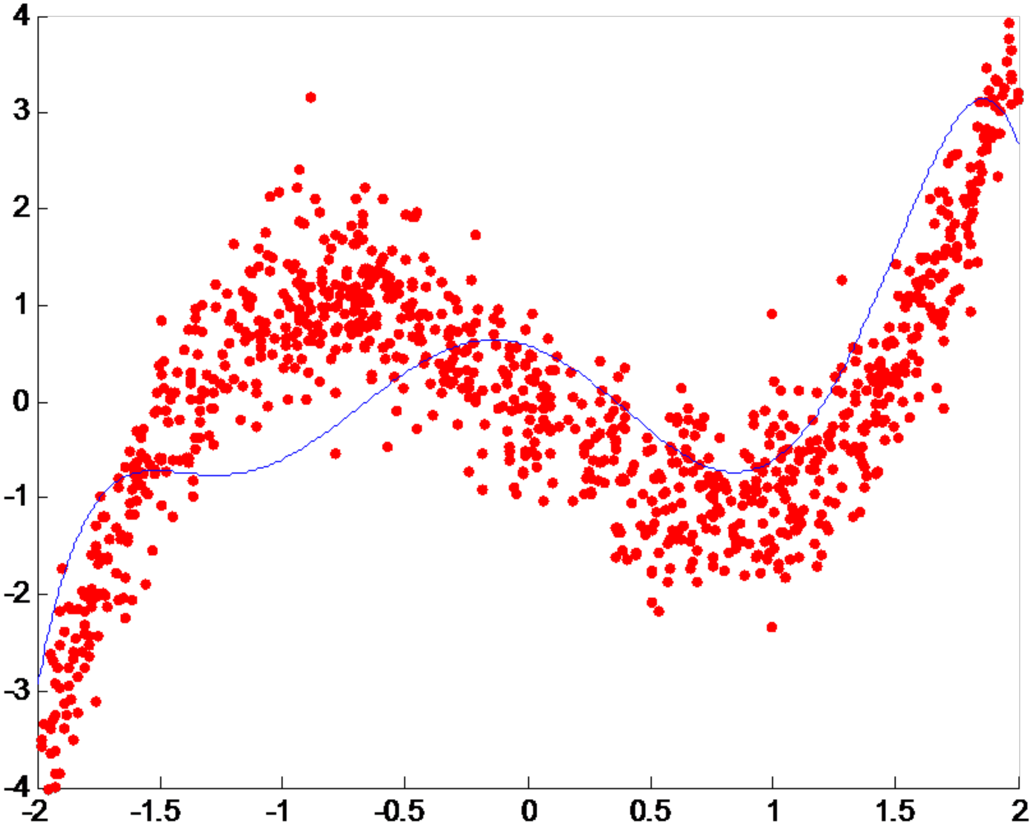
Function f and 50 data points



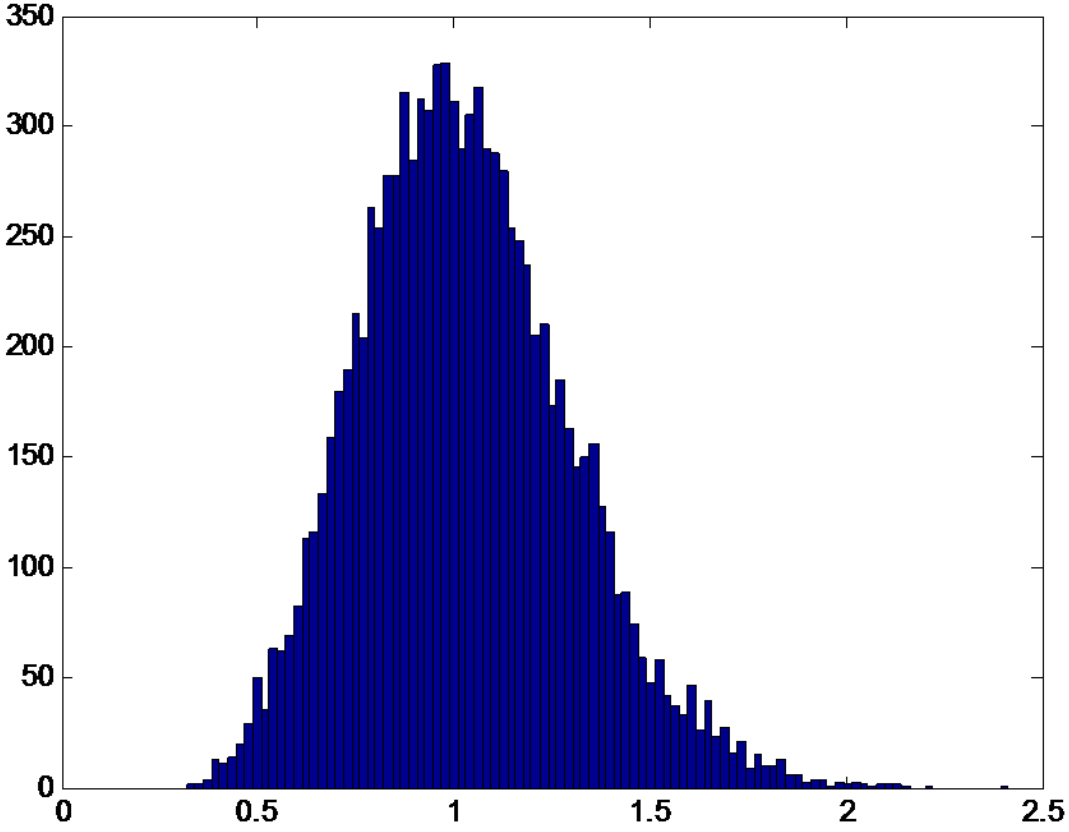
Function f and 200 data points



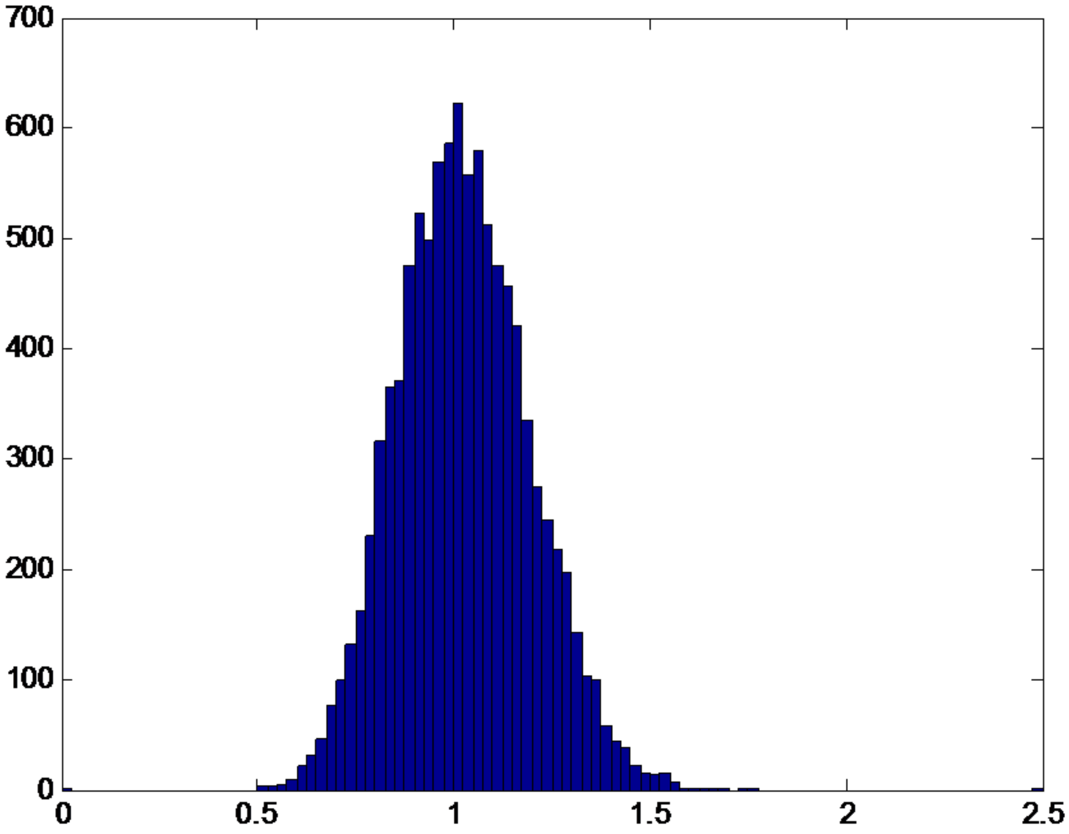
Function f and 1000 data points



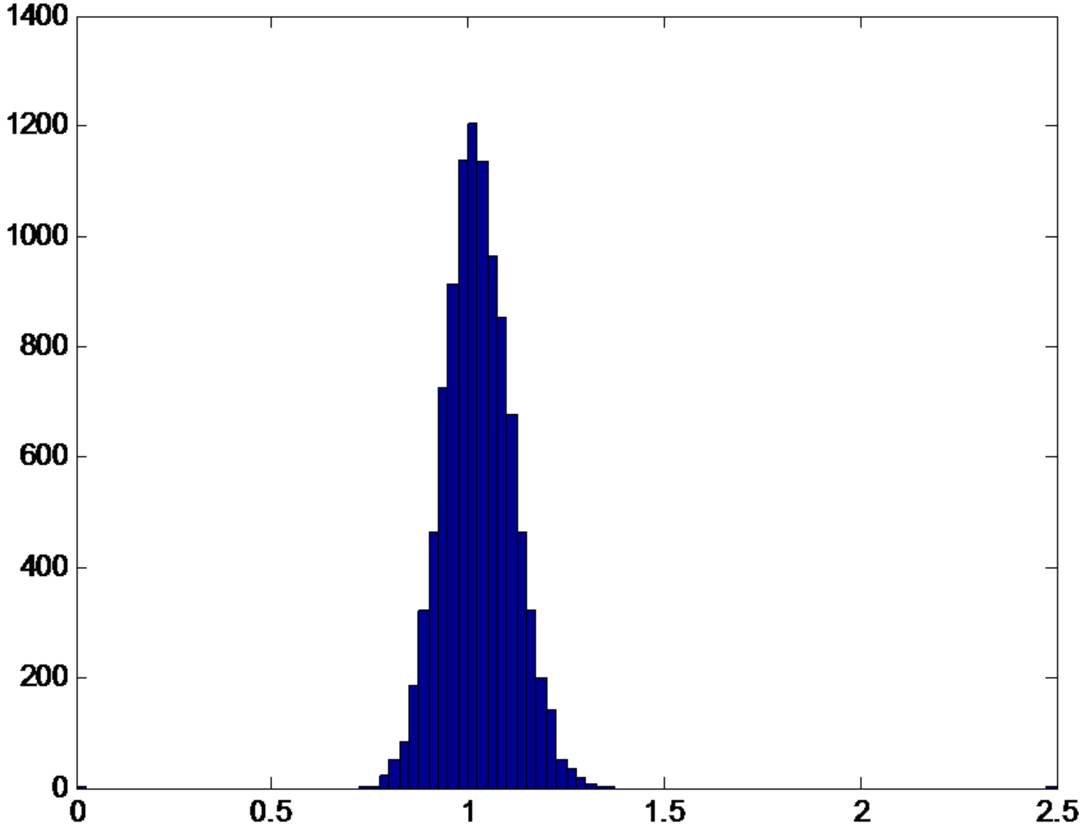
Histogram of empirical risk of f for 10,000 data sets of 20 points each



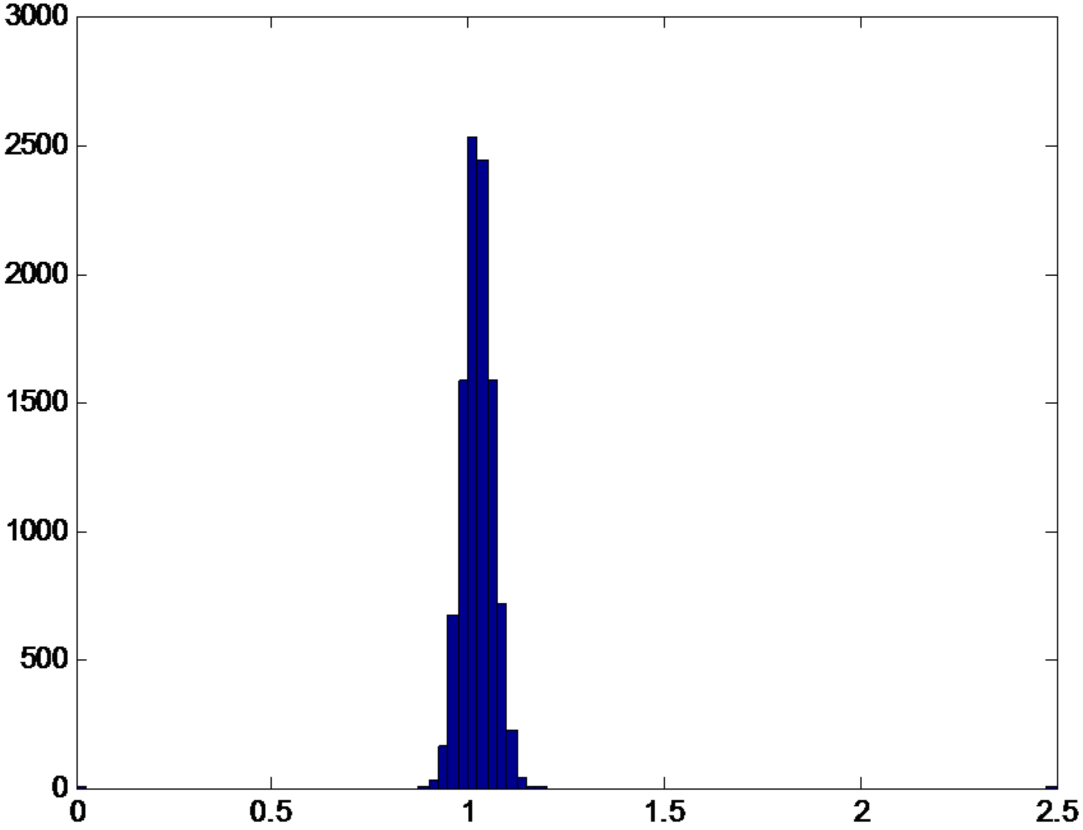
Histogram of empirical risk of f for 10,000 data sets of 50 points each



Histogram of empirical risk of f for 10,000 data sets of 200 points each



Histogram of empirical risk of f for 10,000 data sets of 1000 points each



Chernoff bound

- The law of large numbers tells us that $R_{emp}(f)$ converges in probability to R(f).
- The Chernoff bound tells us *how fast* it convergences.
- It also tells us something about actual risk by placing confidence intervals on it.
- All without knowing the data distribution!
- Note: f is any *fixed* function.

Learning by minimizing empirical risk

- Let f^m be a function that minimizes empirical risk for a given sample of m data points.
- Let f^{opt} be a function that minimizes actual risk.
- Does $R(f^m)$ converge in probability to $R(f^{opt})$?
- Our results so far do not answer this question since f^m is not fixed but depends on the data.

Detailed Analysis

- loss $\xi_i := \frac{1}{2}|f(x_i) - y_i|$ in $\{0, 1\}$
- the ξ_i are independent Bernoulli trials
- empirical mean $\frac{1}{m}\sum_{i=1}^{m} \xi_i$ (by def: equals $R_{\text{emp}}[f]$)
- expected value $\mathbf{E}[\xi]$ (equals $R[f]$)

Chernoff's Bound

$$\mathrm{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \mathbf{E}\left[\xi\right]\right| \geq \epsilon\right\} \leq 2\exp(-2m\epsilon^2)$$

- here, P refers to the probability of getting a sample $\xi_1, \ldots, \xi_m$ with the property $\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \mathbf{E}\left[\xi\right]\right| \geq \epsilon$ (is a product measure)

Useful corollary: Given a $2m$-sample of Bernoulli trials, we have

$$\mathrm{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i\right| \geq \epsilon\right\} \leq 4\exp\left(-\frac{m\epsilon^2}{2}\right).$$

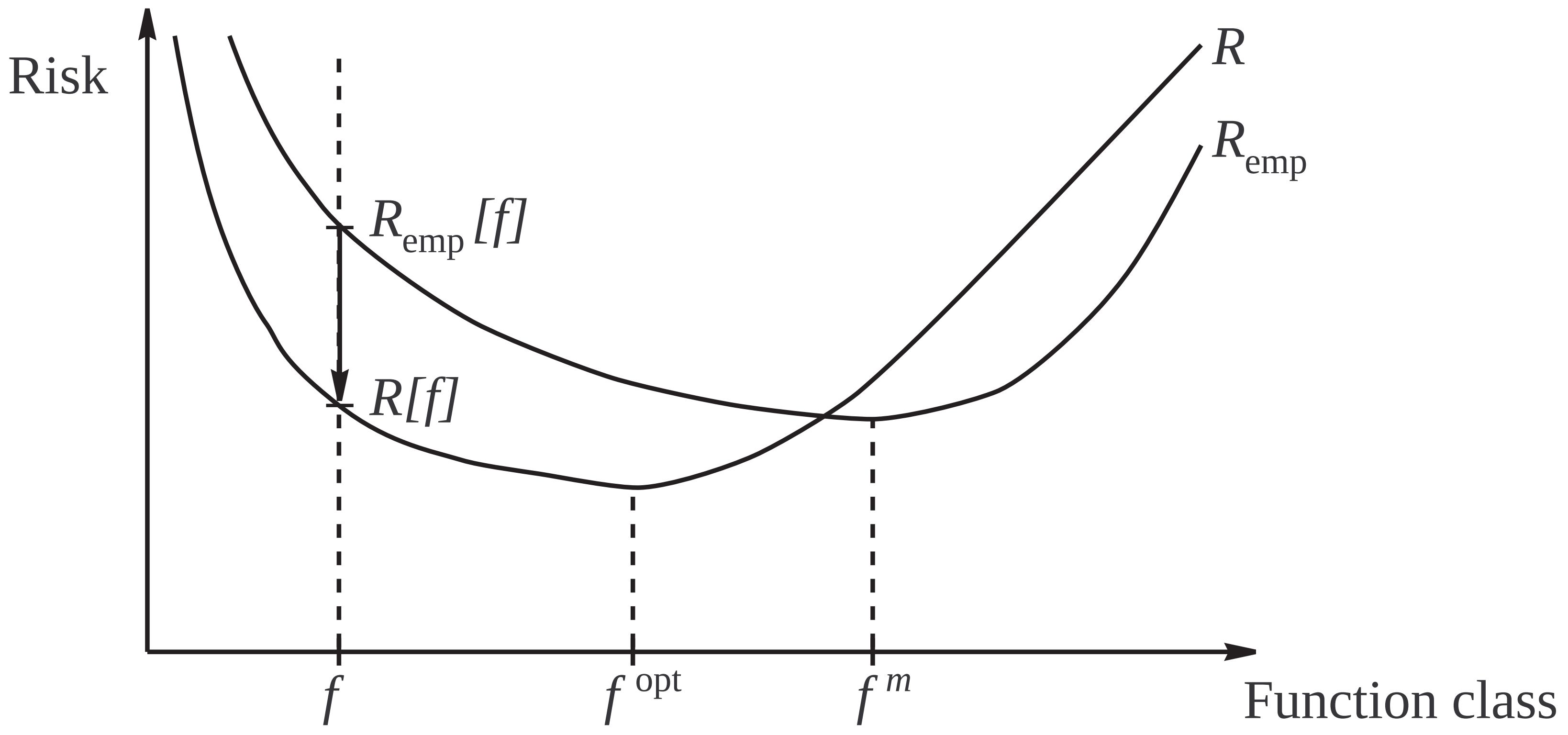
Chernoff's Bound, II

Translate this back into machine learning terminology: the probability of obtaining an m-sample where the training error and test error differ by more than $\epsilon > 0$ is bounded by

$$\mathrm{P}\left\{\left|R_{\mathrm{emp}}[f] - R[f]\right| \geq \epsilon\right\} \leq 2\exp(-2m\epsilon^2).$$

- refers to one fixed f
- not allowed to look at the data before choosing f, hence not suitable as a bound on the test error of a learning algorithm using empirical risk minimization

Consistency and Uniform Convergence



Two Observations

- denote the minimizer of R by f^{opt}. Then
$$R[f] - R[f^{\mathrm{opt}}] \geq 0$$
for all $f \in \mathcal{F}$.
- denote the minimizer of R_{emp} by f^m. Then
$$R_{\mathrm{emp}}[f] - R_{\mathrm{emp}}[f^m] \geq 0$$
for all $f \in \mathcal{F}$.
- In particular, we have
$$R[f^m] - R[f^{\mathrm{opt}}] \geq 0$$
and
$$R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R_{\mathrm{emp}}[f^m] \geq 0.$$

The sum of these two inequalities satisfies

$$\begin{aligned} 0 &\leq R[f^m] - R[f^{\mathrm{opt}}] + R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R_{\mathrm{emp}}[f^m] \\ &= R[f^m] - R_{\mathrm{emp}}[f^m] + R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R[f^{\mathrm{opt}}] \\ &\leq \sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) + (R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R[f^{\mathrm{opt}}]). \end{aligned}$$

- second half of RHS: f^{opt} is fixed (independent of training sample), hence by Chernoff: for all $\epsilon > 0$,

$$\lim_{m \to \infty} \mathrm{P}\{|R_{\mathrm{emp}}[f^{\mathrm{opt}}] - R[f^{\mathrm{opt}}]| > \epsilon\} = 0$$

(“convergence in probability”)

- If the first half of RHS also converges to zero (in probability), i.e.,

$$\lim_{m\to\infty} \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f])>\epsilon\}=0,$$

for all $\epsilon > 0$, then

$$\begin{aligned} R[f^m]-R[f^{\mathrm{opt}}] &\to 0 \\ R_{\mathrm{emp}}[f^{\mathrm{opt}}]-R_{\mathrm{emp}}[f^m] &\to 0 \end{aligned}$$

in probability — in this case, empirical risk minimization can be seen to be *consistent*.

Uniform Convergence (Vapnik & Chervonenkis)

Necessary and sufficient conditions for consistency of empirical risk minimization (ERM):
One-sided convergence, uniformly over all functions that can be implemented by the learning machine.

$$\lim_{m\to\infty} P\{\sup_{f\in\mathcal{F}} (R[f] - R_{\text{emp}}[f]) > \epsilon\} = 0$$

for all $\epsilon > 0$.

- note that this takes into account the whole set of functions that can be implemented by the learning machine
- this is hard to check for a learning machine

Are there properties of learning machines ($\equiv$ sets of functions) which ensure uniform convergence of risk?

How to Prove a VC Bound

Take a closer look at $\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\}$.
Plan:

- if the function class $\mathcal{F}$ contains only one function, then Chernoff's bound suffices:

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} \leq 2\exp(-2m\epsilon^2).$$

- if there are finitely many functions, we use the 'union bound'
- even if there are infinitely many, then *on any finite sample* there are effectively only finitely many (use *symmetrization* and *capacity concepts*)

The Case of Two Functions

Suppose $\mathcal{F} = \{f_1, f_2\}$. Rewrite

$$\mathrm{P}\{\sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2),$$

where

$$C_\epsilon^i := \{(x_1, y_1), \ldots, (x_m, y_m) \mid (R[f_i] - R_{\mathrm{emp}}[f_i]) > \epsilon\}$$

denotes the event that the risks of f_i differ by more than ϵ. The RHS equals

$$\begin{aligned} \mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2) &= \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2) - \mathrm{P}(C_\epsilon^1 \cap C_\epsilon^2) \\ &\leq \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2). \end{aligned}$$

Hence by Chernoff's bound

$$\begin{aligned} \mathrm{P}\{\sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} &\leq \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2) \\ &\leq 2 \cdot 2 \exp(-2m\epsilon^2). \end{aligned}$$

The Union Bound

Similarly, if $\mathcal{F} = \{f_1, \ldots, f_n\}$, we have

$$\mathrm{P}\{\sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup \cdots \cup C_\epsilon^n),$$

and

$$\mathrm{P}(C_\epsilon^1 \cup \cdots \cup C_\epsilon^n) \leq \sum_{i=1}^{n} \mathrm{P}(C_\epsilon^i).$$

Use Chernoff for each summand, to get an extra factor n in the bound.

Note: this becomes an equality if and only if all the events involved are *disjoint*.

Infinite Function Classes

- Note: empirical risk only refers to m points. On these points, the functions of $\mathcal{F}$ can take at most 2^m values
- for R_{emp}, the function class thus "looks" finite
- how about R?
- need to use a trick

Symmetrization

Lemma 1 (Vapnik & Chervonenkis (e.g., [63, 19]))
For $m\epsilon^2 \geq 2$ we have

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f]) > \epsilon\} \leq 2\mathrm{P}\{\sup_{f\in\mathcal{F}}(R_{\mathrm{emp}}[f]-R'_{\mathrm{emp}}[f]) > \epsilon/2\}$$

Here, the first P *refers to the distribution of iid samples of size m, while the second one refers to iid samples of size $2m$. In the latter case, R_{emp} measures the loss on the first half of the sample, and R'_{emp} on the second half.*

Shattering Coefficient

- Hence, we only need to consider the maximum size of $\mathcal{F}$ on $2m$ points. Call it $\mathcal{N}(\mathcal{F}, 2m)$.
- $\mathcal{N}(\mathcal{F}, 2m)$ = max. number of different outputs $(y_1, \ldots, y_{2m})$ that the function class can generate on $2m$ points — in other words, the max. number of different ways the function class can separate $2m$ points into two classes.
- $\mathcal{N}(\mathcal{F}, 2m) \leq 2^{2m}$
- if $\mathcal{N}(\mathcal{F}, 2m) = 2^{2m}$, then the function class is said to *shatter* $2m$ points.

Putting Everything Together

We now use (1) symmetrization, (2) the shattering coefficient, and (3) the union bound, to get

$$
\begin{aligned}
& \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f])>\epsilon\} \\
\leq\ & 2\mathrm{P}\{\sup_{f\in\mathcal{F}}(R_{\mathrm{emp}}[f]-R'_{\mathrm{emp}}[f])>\epsilon/2\} \\
=\ & 2\mathrm{P}\{(R_{\mathrm{emp}}[f_1]-R'_{\mathrm{emp}}[f_1])>\epsilon/2 \vee\ldots\vee (R_{\mathrm{emp}}[f_{\mathcal{N}(\mathcal{F},2m)}]-R'_{\mathrm{emp}}[f_{\mathcal{N}(\mathcal{F},2m)}])>\epsilon/2\} \\
\leq\ & \sum_{n=1}^{\mathcal{N}(\mathcal{F},2m)} 2\mathrm{P}\{(R_{\mathrm{emp}}[f_n]-R'_{\mathrm{emp}}[f_n])>\epsilon/2\}.
\end{aligned}
$$

ctd.

Use Chernoff's bound for each term:*

$$\mathrm{P}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i \geq \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{2}\right).$$

This yields

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f]) > \epsilon\} \leq 4\mathcal{N}(\mathcal{F},2m)\exp\left(-\frac{m\epsilon^2}{8}\right).$$

- provided that $\mathcal{N}(\mathcal{F},2m)$ does not grow exponentially in m, this is nontrivial
- such bounds are called *VC type inequalities*
- two types of randomness: (1) the P refers to the drawing of the training examples, and (2) $R[f]$ is an expectation over the drawing of test examples.

* A rigorous treatment would need to use a second randomization over permutations of the $2m$-sample, see [53].

Confidence Intervals

Rewrite the bound: specify the probability with which we want R to be close to R_{emp}, and solve for ϵ:
With a probability of at least $1 - \delta$,

$$R[f] \leq R_{\mathrm{emp}}[f] + \sqrt{\frac{8}{m}\left(\ln(\mathcal{N}(\mathcal{F}, 2m)) + \ln\frac{4}{\delta}\right)}.$$

This bound holds independent of f; in particular, it holds for the function f^m minimizing the empirical risk.

Discussion

- tighter bounds are available (better constants etc. — cf. Shahar Mendelson's tutorial)
- cannot minimize the bound over f
- other capacity concepts can be used

VC Entropy

On an example $(\mathbf{x}, y)$, f causes a loss

$$c(x, y, f(x)) := \frac{1}{2}|f(x) - y| \in \{0, 1\}.$$

For a larger sample $(x_1, y_1) \ldots, (x_m, y_m)$, the different functions $f \in \mathcal{F}$ lead to a *set* of loss vectors

$$\boldsymbol{\xi}_f = (c(x_1, y_1, f(x_1)), \ldots, c(x_m, y_m, f(x_m))),$$

whose cardinality we denote by

$$\mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right).$$

The *VC entropy* is defined as

$$H_{\mathcal{F}}(m) = \mathbf{E}\left[\ln \mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)\right],$$

where the expectation is taken over the random generation of the m-sample $(x_1, y_1) \ldots, (x_m, y_m)$ from P.

$H_{\mathcal{F}}(m)/m \to 0 \iff$ uniform convergence of risks (hence consistency)

Further PR Capacity Concepts

- exchange '**E**' and 'ln': *annealed entropy*.

  $H_{\mathcal{F}}^{\mathrm{ann}}(m)/m \to 0 \Longleftrightarrow$ exponentially fast uniform convergence

- take 'max' instead of 'E': *growth function*.
  Note that $G_{\mathcal{F}}(m) = \ln \mathcal{N}(\mathcal{F}, m)$.

  $G_{\mathcal{F}}(m)/m \to 0 \Longleftrightarrow$ exponential convergence for all underlying distributions P.

  $G_{\mathcal{F}}(m) = m \cdot \ln(2)$ for all $m \Longleftrightarrow$ for any m, all loss vectors can be generated, i.e., the m points can be chosen such that by using functions of the learning machine, they can be separated in all 2^m possible ways (*shattered*).

Structure of the Growth Function

Either $G_{\mathcal{F}}(m) = m \cdot \ln(2)$ for all $m \in \mathbb{N}$

Or there exists some *maximal* m for which the above is possible. Call this number the *VC-dimension*, and denote it by h. For $m > h$,

$$G_{\mathcal{F}}(m) \leq h\left(\ln \frac{m}{h} + 1\right).$$

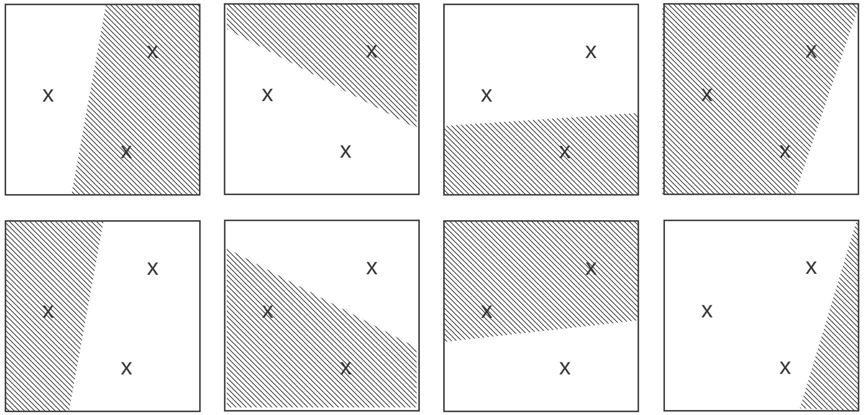
Nothing "in between" linear growth and logarithmic growth is possible.

VC-Dimension: Example

Half-spaces in $\mathbb{R}^2$:

$$f(x,y) = \mathrm{sgn}(a + bx + cy), \quad \text{with parameters } a, b, c \in \mathbb{R}$$

- Clearly, we can shatter three non-collinear points.
- But we can never shatter four points.
- Hence the VC dimension is $h = 3$ (in this case, equal to the number of parameters)



A Typical Bound for Pattern Recognition

For any $f \in \mathcal{F}$ and $m > h$, with a probability of at least $1 - \delta$,

$$R[f] \leq R_{\text{emp}}[f] + \phi\left(\frac{h}{m}, \frac{\log(\delta)}{m}\right)$$

holds, where the *confidence term* ϕ is defined as

$$\phi\left(\frac{h}{m}, \frac{\log(\delta)}{m}\right) = \sqrt{\frac{h\left(\log\frac{2m}{h} + 1\right) - \log(\delta/4)}{m}}.$$

- does this mean, that we can learn *any*thing?
- The study of the consistency of ERM has thus led to concepts and results which lets us formulate a better induction principle: we can use this bound to get a low risk!
- in practice: use as a guideline for designing algorithms

Examples of Induction Principles

- *Empirical risk minimization:* minimize

$$R_{\mathrm{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|f(\mathbf{x}_i) - y_i|$$

- *Minimum description length:* minimize some measure of the description length of the sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ by a function f.

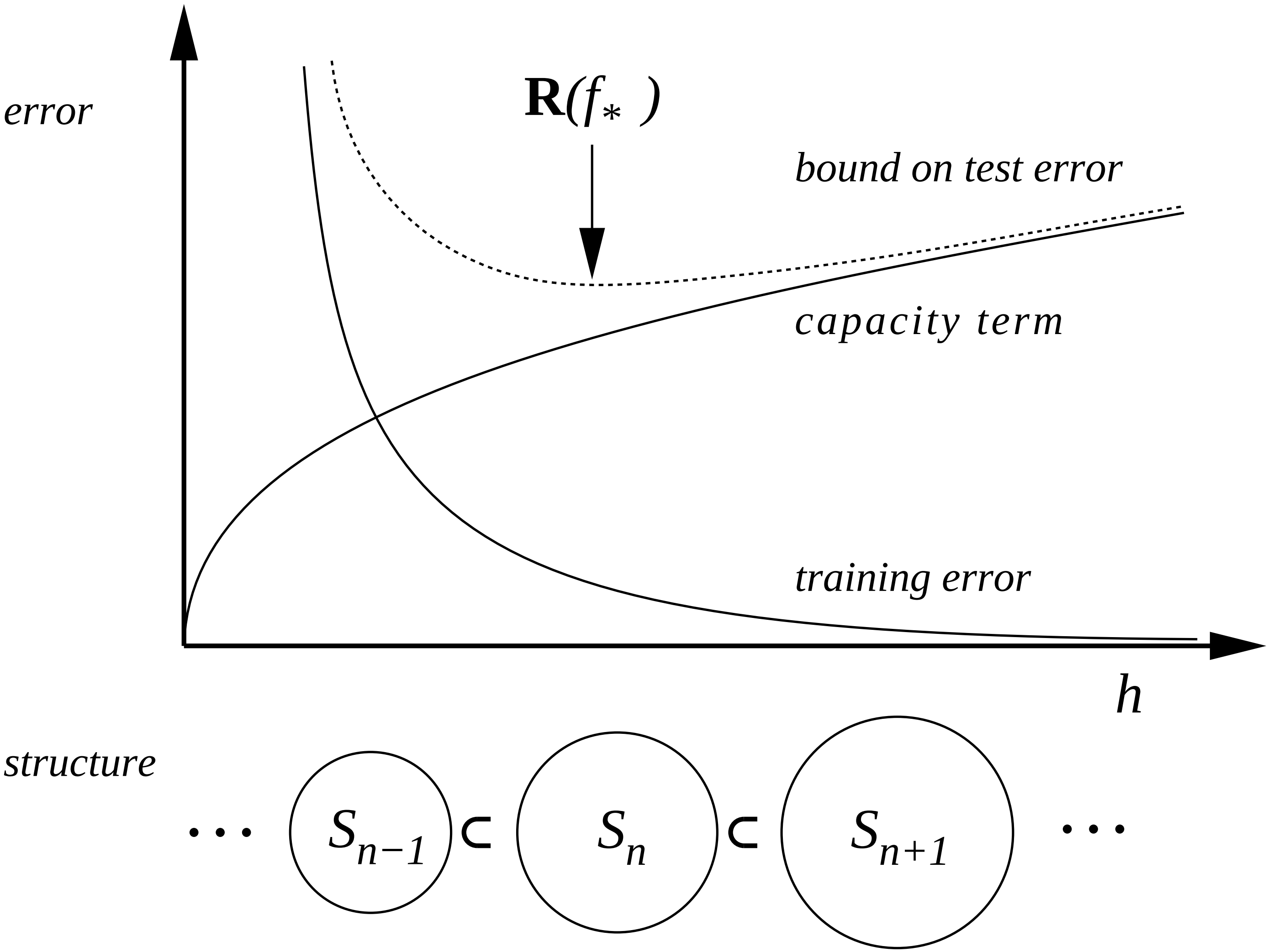
- *Structural risk minimization (SRM) [63]:* minimize the RHS of

$$R[f] \leq R_{\mathrm{emp}}[f] + \phi\left(\frac{h}{m}\right).$$

  To this end, introduce a structure on $\mathcal{F}$.

Learning machine $\equiv$ a set of functions and an induction principle

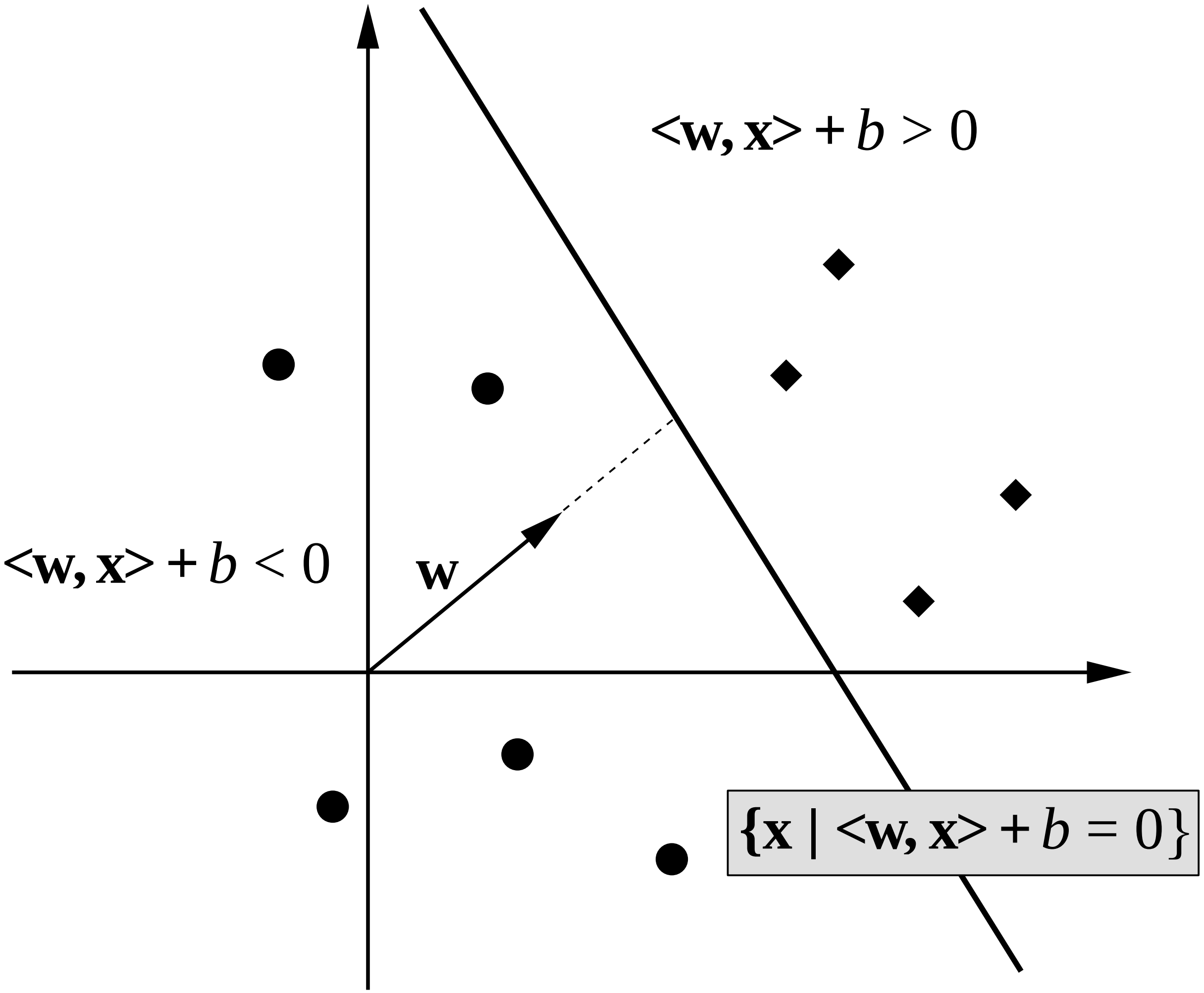
SRM: The Picture



Finding a Good Function Class

- recall: separating hyperplanes in $\mathbb{R}^2$ have a VC dimension of 3.
- more generally: separating hyperplanes in $\mathbb{R}^N$ have a VC dimension of $N + 1$.
- hence: separating hyperplanes in high-dimensional feature spaces have extremely large VC dimension, and may not generalize well
- however, *margin* hyperplanes can still have a small VC dimension

Separating Hyperplane


$\langle \mathbf{w}, \mathbf{x} \rangle + b > 0$
$\langle \mathbf{w}, \mathbf{x} \rangle + b < 0$
$\mathbf{w}$
$\{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$

Canonical Hyperplanes

Note: if $c \neq 0$, then

$$\{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\} = \{\mathbf{x} | \langle c\mathbf{w}, \mathbf{x} \rangle + cb = 0\}.$$

Hence $(c\mathbf{w}, cb)$ describes the same hyperplane as $(\mathbf{w}, b)$.

Definition: The hyperplane is in *canonical* form w.r.t. $X^* = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ if $\min_{\mathbf{x}_i \in X} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1$.

Note that for canonical hyperplanes, the distance of the closest point to the hyperplane ("margin") is $1/\|\mathbf{w}\|$:
$\min_{\mathbf{x}_i \in X} \left| \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \mathbf{x}_i \right\rangle + \frac{b}{\|\mathbf{w}\|} \right| = \frac{1}{\|\mathbf{w}\|}$.

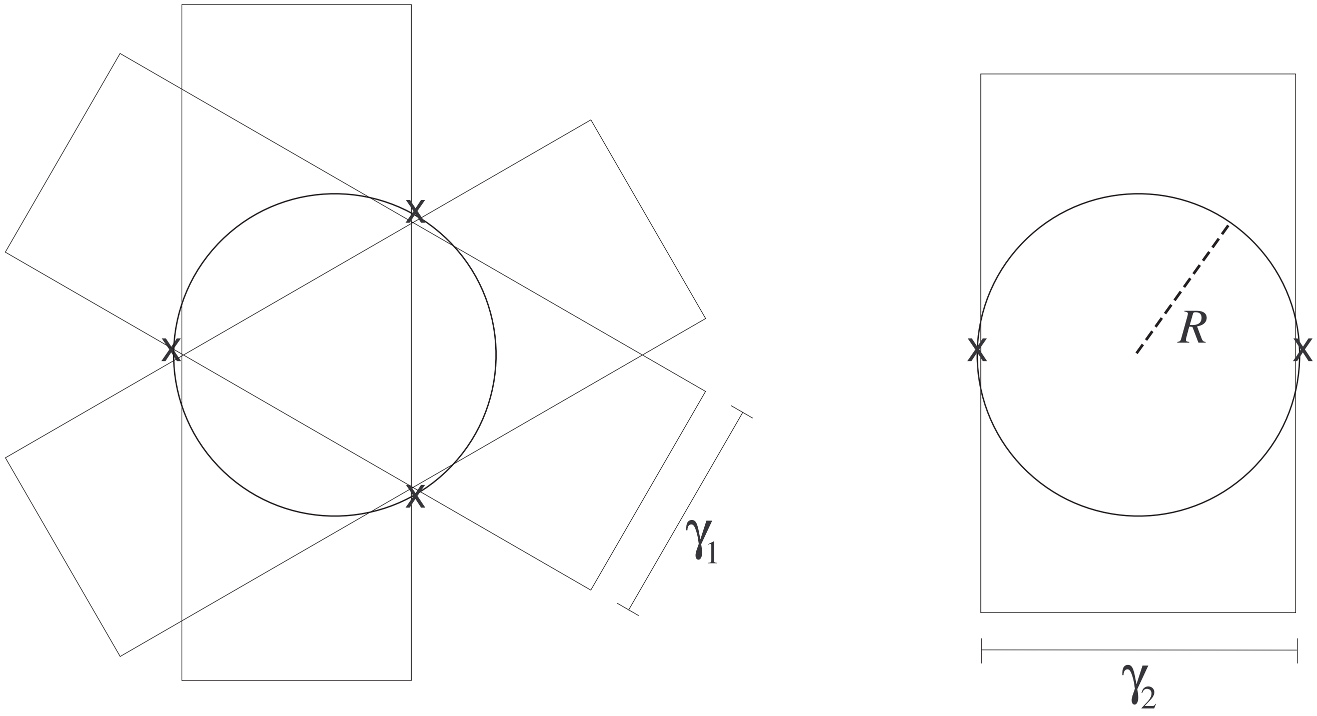
Theorem 2 (Vapnik [63]) *Consider hyperplanes* $\langle \mathbf{w}, \mathbf{x} \rangle = 0$ *where* $\mathbf{w}$ *is normalized such that they are in canonical form w.r.t. a set of points* $X^* = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$, *i.e.,*

$$\min_{i=1,\ldots,r} | \langle \mathbf{w}, \mathbf{x}_i \rangle | = 1.$$

The set of decision functions $f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn} \langle \mathbf{x}, \mathbf{w} \rangle$ *defined on* X^* *and satisfying the constraint* $\|\mathbf{w}\| \leq \Lambda$ *has a VC dimension satisfying*

$$h \leq R^2 \Lambda^2.$$

Here, R *is the radius of the smallest sphere around the origin containing* X^*.



recall $\gamma > 2/\Lambda$

Proof Strategy (Gurvits, 1997)

Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are shattered by canonical hyperplanes with $\|\mathbf{w}\| \leq \Lambda$, i.e., for all $y_1, \ldots, y_r \in \{\pm 1\}$, there exists a $\mathbf{w}$ such that

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \text{ for all } i = 1, \ldots, r. \tag{1}$$

Two steps:

- prove that the more points we want to shatter (1), the larger $\|\sum_{i=1}^r y_i \mathbf{x}_i\|$ must be
- upper bound the size of $\|\sum_{i=1}^r y_i \mathbf{x}_i\|$ in terms of R

Combining the two tells us how many points we can at most shatter.

Part I

Summing (1) over $i = 1, \ldots, r$ yields

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^{r} y_i \mathbf{x}_i \right) \right\rangle \geq r.$$

By the Cauchy-Schwarz inequality, on the other hand, we have

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^{r} y_i \mathbf{x}_i \right) \right\rangle \leq \|\mathbf{w}\| \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\| \leq \Lambda \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\| .$$

Combine both:

$$\frac{r}{\Lambda} \leq \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\| . \qquad (2)$$

Part II

Consider independent random labels $y_i \in \{\pm 1\}$, uniformly distributed (*Rademacher variables*).

$$\begin{aligned}
\mathbf{E}\left[\left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2\right] &= \sum_{i=1}^{r} \mathbf{E}\left[\left\langle y_i \mathbf{x}_i, \sum_{j=1}^{r} y_j \mathbf{x}_j \right\rangle\right] \\
&= \sum_{i=1}^{r} \mathbf{E}\left[\left\langle y_i \mathbf{x}_i, \left(\left(\sum_{j\neq i} y_j \mathbf{x}_j\right) + y_i \mathbf{x}_i\right)\right\rangle\right] \\
&= \sum_{i=1}^{r} \left(\left(\sum_{j\neq i} \mathbf{E}\left[\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle\right]\right) + \mathbf{E}\left[\langle y_i \mathbf{x}_i, y_i \mathbf{x}_i \rangle\right]\right) \\
&= \sum_{i=1}^{r} \mathbf{E}\left[\|y_i \mathbf{x}_i\|^2\right] = \sum_{i=1}^{r} \mathbf{E}\left[\|\mathbf{x}_i\|^2\right]
\end{aligned}$$

Part II, ctd.

Since $\|\mathbf{x}_i\| \leq R$, we get

$$\mathbf{E}\left[\left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2\right] \leq rR^2.$$

- This holds for the *expectation* over the random choices of the labels, hence there must be at least one set of labels for which it also holds true. Use this set.

Hence

$$\left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2 \leq rR^2.$$

Part I and II Combined

Part I: $\left(\frac{r}{\Lambda}\right)^2 \leq \left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2$

Part II: $\left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2 \leq r R^2$

Hence

$$\frac{r^2}{\Lambda^2} \leq r R^2,$$

i.e.,

$$r \leq R^2 \Lambda^2.$$