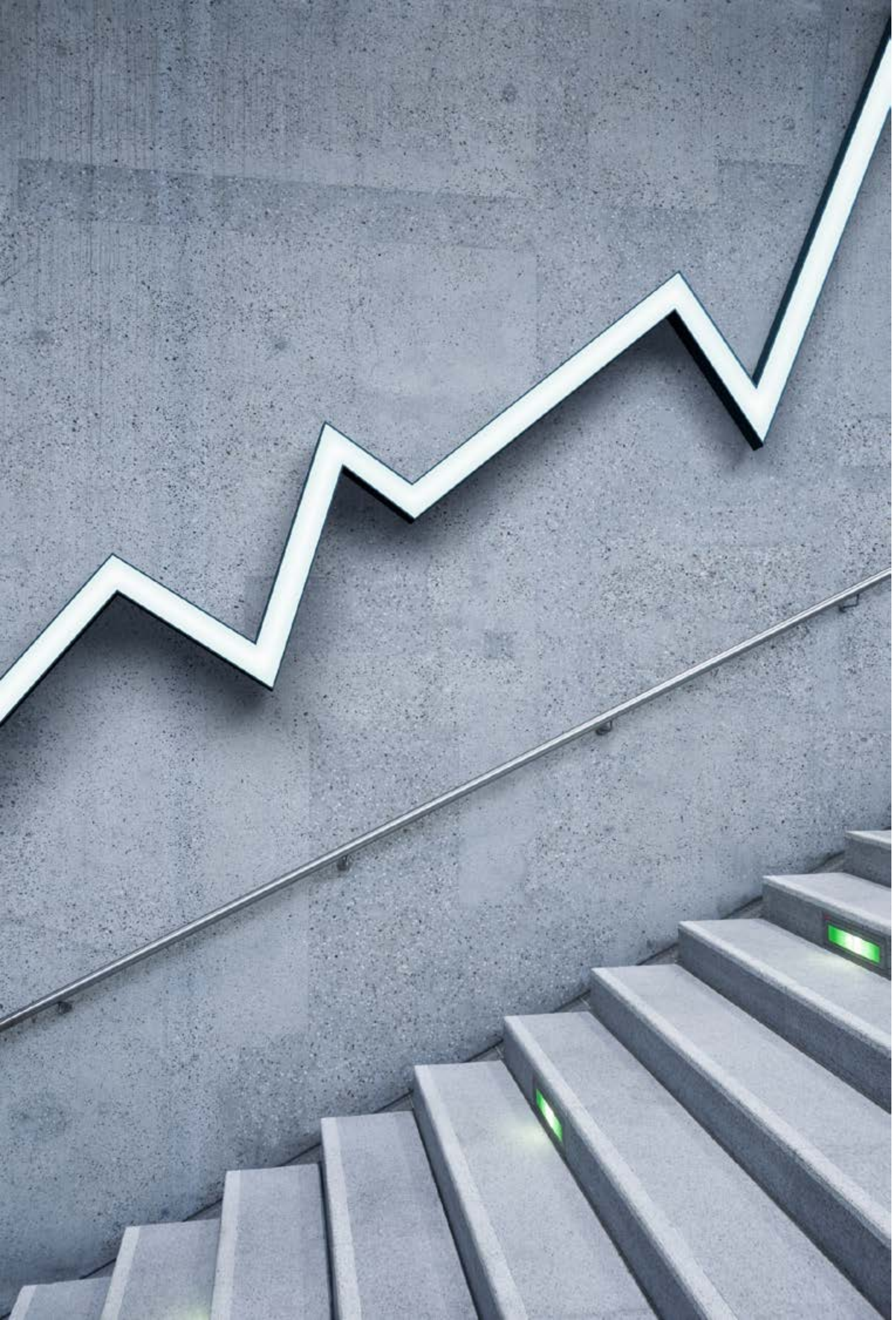
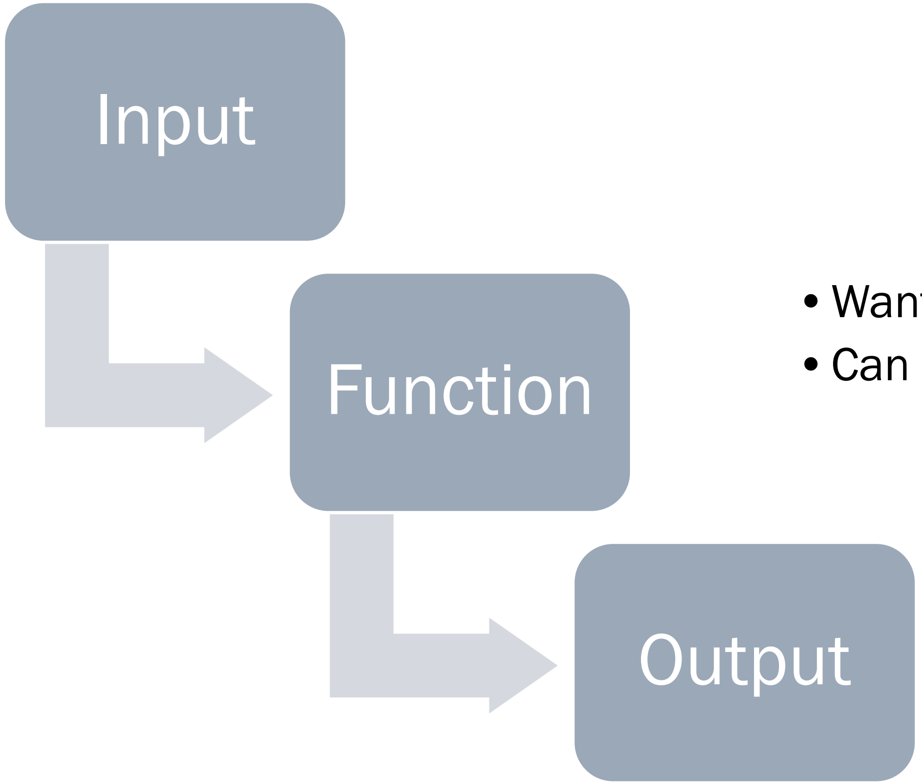


Are Loss Functions All the Same?

NOV. 11, 2003. L. ROSASCO, E. DE VITO, A. CAPONNETTO, M. PIANA

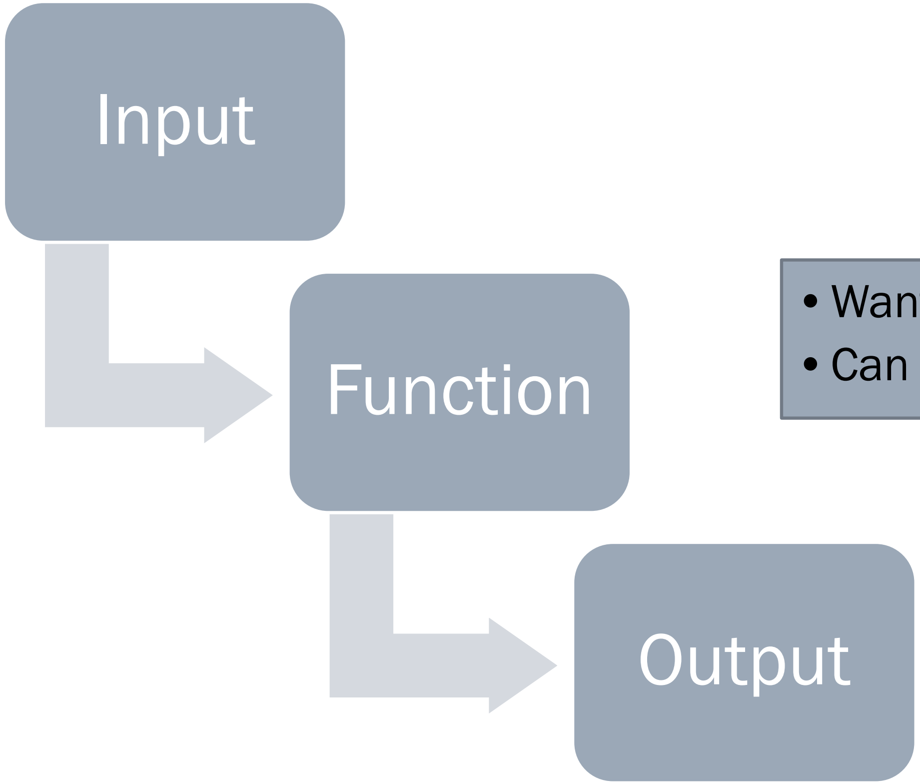
{PRESENTED BY AHMED ABOUSERIE AND MEGAN STICKLER}

Machine Learning Process



- Want: Function that minimizes $R_{L,P}$
- Can Find: Function that minimizes $R_{L,D}$

Machine Learning Process



- Want: Function that minimizes $R_{L,P}$
- Can Find: Function that minimizes $R_{L,D}$

Loss functions and Risk

Theoretically, our expected risk is

$$R[f] = \int_{X \times Y} L(f(x), y) p(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

Where $L(f(x), y)$ is the loss function and $p(x, y)$ is the probability distribution on $X \times Y$ and our solution is the function $f_0 : X \to \mathbb{R}$ that minimizes this.

In practice, $p(x, y)$ is not minutely known and so instead we use our data, which consists of N samples drawn from $p(x, y)$ and find the empirical risk

$$R_{emp}[f] = \frac{1}{l} \sum_{i=1}^{l} L(f(x_i), y_i)$$

And its minimizing argument f_D.

Loss functions and Risk

Theoretically, our expected risk is

$$R[f] = \int_{X \times Y} L(f(x), y) p(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

Where $L(f(x), y)$ is the loss function and $p(x, y)$ is the probability distribution on $X \times Y$ and our solution is the function $f_0 \colon X \to \mathbb{R}$ that minimizes this.

In practice, $p(x, y)$ is not minutely known and so instead we use our data, which consists of N samples drawn from $p(x, y)$ and find the empirical risk

$$R_{emp}[f] = \frac{1}{l} \sum_{i=1}^{l} L(f(x_i), y_i)$$

And its minimizing argument f_D.

Error Source!! → $R \neq R_{emp} \Rightarrow f_0 \neq f_D$

Loss functions and Risk

Additionally, attempting to approximate f_0 from a finite data set is an ill-posed problem, so we regularize the problem by imposing smoothness constraints on the set from which f_0 is drawn. That is, we use an RKHS H. We further restrict this space by using a threshold $C > 0$:

$$H_C = \{f \in H: \|f\|_H \leq C\}$$

Hence, the minimizer we're actually finding is f_c, which minimizes over H_C, not f_0, which minimizes over the space of measureable functions F for which $R[f]$ is well-defined.

Loss functions and Risk

Additionally, attempting to approximate f_0 from a finite data set is an ill-posed problem, so we regularize the problem by imposing smoothness constraints on the set from which f_0 is drawn. That is, we use an RKHS H. We further restrict this space by using a threshold $C > 0$:

$$H_C = \{f \in H: \|f\|_H \leq C\}$$

Hence, the minimizer we're actually finding is f_c, which minimizes over H_C, not f_0, which minimizes over the space of measureable functions F for which $R[f]$ is well-defined.

Error Source!!

$$F \neq H_C \Rightarrow f_0 \neq f_C$$

Error!

$$R[f_D] - R[f_0] = (R[f_D] - R[f_C]) + (R[f_C] - R[f_0])$$

Error!

$$R[f_D] - R[f_0] = (R[f_D] - R[f_C]) + (R[f_C] - R[f_0])$$

The second part, $(R[f_C] - R[f_0])$, is called the «approximation error» and can be minimized by choosing a sufficiently rich hypothesis space H_C.

Error!

$$R[f_D] - R[f_0] = (R[f_D] - R[f_C]) + (R[f_C] - R[f_0])$$

The second part, $(R[f_C] - R[f_0])$, is called the «approximation error» and can be minimized by choosing a sufficiently rich hypothesis space H_C.

The first part, $(R[f_D] - R[f_C])$, is called the «sample» or «estimation error».

Error!

$$R[f_D] - R[f_0] = (R[f_D] - R[f_C]) + (R[f_C] - R[f_0])$$

The second part, $(R[f_C] - R[f_0])$, is called the «approximation error» and can be minimized by choosing a sufficiently rich hypothesis space H_C.

The first part, $(R[f_D] - R[f_C])$, is called the «sample» or «estimation error».

Question: What effect does the choice of loss function have on the sample error?

Loss Functions

Part of what this paper does is explicitly require loss functions to be convex. As a result, it is able to use the convexity and its results, Lipschitz continuity and boundedness at 0, in its analysis.

L_M: Lipschitz constant for $M > 0$

C_0: bound at 0

$$L(0, y) \leq C_0$$

Loss Functions

It then compares loss functions for regression problems and loss functions for classification problems:

<u>Regression Losses</u>

Square Loss: $L(x, y) = (x - y)^2$

Abs. Value Loss: $L(x, y) = |x - y|$

ε-insensitive Loss: $L(x, y) = \max\{|x - y| - \varepsilon, 0\}$

<u>Classification Losses</u>

Square Loss: $L(x, y) = (x - y)^2 = (1 - xy)^2$

Hinge Loss: $L(x, y) = \max\{1 - xy, 0\}$

Logistic Loss: $L(x, y) = (\ln 2)^{-1} \ln(1 + e^{-xy})$

L_M : Lipschitz constant

C_0: bound at 0.

For regression problems on interval $[a, b] \subset \mathbb{R}$,

$\delta = \max\{|a|, |b|\}$

problem	loss	L_M	C_0
regr	quad	$2M + \delta$	δ^2
regr	abs val	1	δ
regr	ϵ-insensitive	1	δ
class	quad	$2M + 2$	1
class	hinge	1	1
class	logistic	$(\ln 2)^{-1} e^M / (1 + e^M)$	1

Bound on Sample Error

One of the first things this paper does is extend a result from (Cucker and Smale 2002b) to provide a bound on the estimation error .

Lemma: Let $M = \|f\|_\infty C$ and $B = L_M M + C_0$. For all $\varepsilon > 0$,

$$P\left\{D \in (X \times Y)^l : \sup_{f \in Hc} \left| R[f] - R_{emp}[f] \right| \leq \varepsilon \right\} \geq 1 - 2N\left(\frac{\varepsilon}{4L_M}\right) \mathrm{e}^{\left(-\frac{l\varepsilon^2}{8B^2}\right)}$$

Bound on Sample Error

One of the first things this paper does is extend a result from (Cucker and Smale 2002b) to provide a bound on the estimation error .

Theorem: Given $0 < \eta < 1, l \in \mathbb{N}, C > 0$, then with probability at least $1 - \eta$

$$R[f_D] \leq R_{emp}[f_D] + \varepsilon(\eta, l, C)$$

And $|R[f_D] - R[f_C]| \leq 2\varepsilon(\eta, l, C)$

With $\lim_{l \to \infty} \varepsilon(\eta, l, C) = 0.$

Bounds on Sample Error

Using this and the table results, they get the following convergence rates:

<u>Regression</u>

Square: $2N\left(\frac{\varepsilon}{4(2C+\delta)}\right)\exp\left(-\frac{l\varepsilon^2}{8(C(x+\delta)+\delta^2)^2}\right)$

Abs. Value and ε-insensitive: $2N\left(\frac{\varepsilon}{4}\right)\exp\left(-\frac{l\varepsilon^2}{8(C+\delta)^2}\right)$

problem	loss	L_M	C_0
regr	quad	$2M+\delta$	δ^2
regr	abs val	1	δ
regr	ϵ-insensitive	1	δ
class	quad	$2M+2$	1
class	hinge	1	1
class	logistic	$(\ln 2)^{-1}e^M/(1+e^M)$	1

Bounds on Sample Error

Using this and the table results, they get the following convergence rates:

problem	loss	L_M	C_0
regr	quad	$2M+\delta$	δ^2
regr	abs val	1	δ
regr	ϵ-insensitive	1	δ
class	quad	$2M+2$	1
class	hinge	1	1
class	logistic	$(\ln 2)^{-1}e^M/(1+e^M)$	1

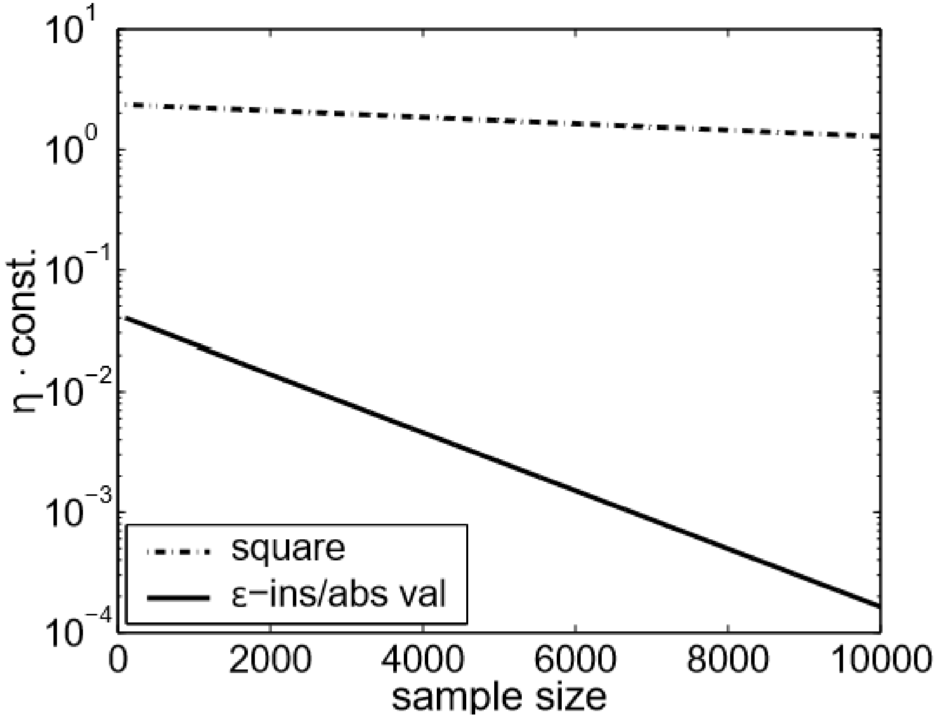
<u>Classification:</u>

Square: $2N\left(\frac{\varepsilon}{4(2C+\delta)}\right)\exp\left(-\frac{l\varepsilon^2}{8(C(x+\delta)+\delta^2)^2}\right)$

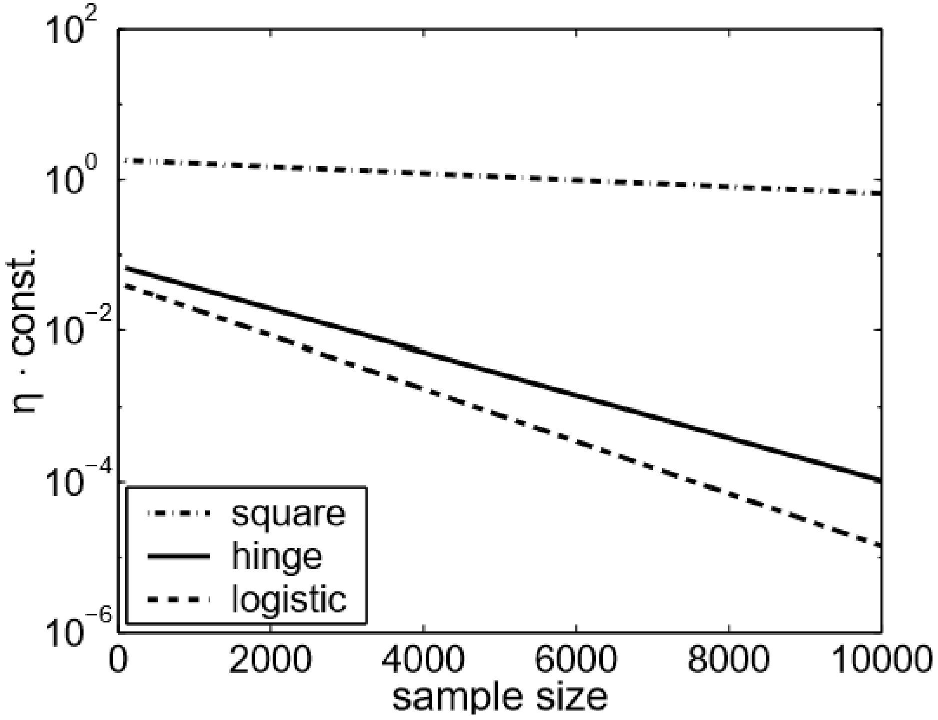
Hinge: $2N\left(\frac{\varepsilon}{4}\right)\exp\left(\frac{-l\varepsilon^2}{8(C+1)^2}\right)$

Logistic: $2N\left(\frac{\varepsilon}{4}\right)\left(\frac{\ln 2(1+\mathrm{e}^C)}{\mathrm{e}^C}\right)\exp\left(-\frac{l\varepsilon^2}{8\left(C\left((\ln 2)^{-1}\mathrm{e}^C/\left(\mathrm{e}^C+1\right)\right)+1\right)^2}\right)$

Bounds on Sample Error



Regression Losses



Classification Losses

Further Bounds on Sample Error for Classification Problems

They show also that the Bayes Optimal solution f_b is equivalent to the sign of f_0 ($\text{sgn}(f_0)$):

Assume that the loss function $L(x, y) = L(xy)$ is convex and that it is decreasing in a neighborhood of 0. If $f_0(x) \neq 0$, then

$$\text{sgn}(f_0) = f_b = \begin{cases} 1 & \text{if } p(1|\mathbf{x}) > p(-1|\mathbf{x}) \\ -1 & \text{if } p(1|\mathbf{x}) < p(-1|\mathbf{x}). \end{cases}$$

Bounds on Estimation Error for Classification Problems

In terms of minimizing total error, for classification problems we would like to bound

$$R[\mathrm{sgn}(f_D)] - R[f_b]$$

A result from (Lin et al., 2003) shows that specifically for hinge loss, $R[f_0] = R[f_b]$.

They combine this with the previously derived bounds to show that in the case of hinge loss,

for $0 < \eta < 1, C > 0$, with probability at least $1 - \eta$

$$0 \leq R[\mathrm{sgn}(f_D)] - R[f_b] \leq R[f_D] - R[f_0] \leq 2\varepsilon(\eta, l, C)$$