نظریه یادگیری

(*Learning Theory, LT*)

نامساوی‌های تمرکز

(*Concentration Inequalities*)

(بخش ۱)

دانشگاه فردوسی مشهد
دانشکده مهندسی
رضا منصفی

نامساوی‌های تمرکز (*Concentration inequalities*)

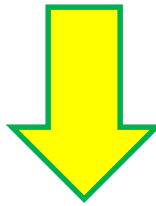
تعریف: **نابرابری‌های تمرکز**، برای **متغیر تصادفی** که **پیرامون میانگین، متمرکز شده** و یا **انحراف** از **میانگین** و یا **دیگر مقدار** دارد، **کران‌های احتمالاتی** می‌دهد.

Concentration inequalities give probability bounds for a random variable to be concentrated around its mean, or for it to deviate from its mean or some other value.

نامساوی‌های تمرکز (*Concentration inequalities*)

Few of these inequalities are:-

- ***Markov's Inequality***
- ***Chebyshev's Inequality***
- ***Hoeffding's Inequality***
- ***McDiarmid's Inequality***
- ……



یادآوری (*Recapitulate*)

Note: *For a function* $h \in H$ *the* ***Empirical Risk function*** *is*

$$\hat{R}(h) = error_s = \frac{1}{m}\sum_{i=1}^{m} l(h(X_i), Y_i)$$

and the ***Empirical Risk Minimizing (ERM) function*** *is*

$$h_m = \arg\min_{h \in H} \hat{R}(h)$$

Note: *Often, interested in the* ***true performance*** *of* h_m

$$R(h_m) = error_{true}(h_m) = E[l(h_m(X), Y)]$$

Example: *This might correspond to the* ***performance*** *of the* ***hinge-loss cost function*** *using an* ***SVM****.* ***(click here)***

Note: *For finite classes* H *the statement can be shown implies that*

$$\forall h \in H \quad P(R(h) \geq \hat{R}(h) + \varepsilon) \leq e^{-c\varepsilon^2 m}$$

$$\delta \leq |H| e^{-2m\varepsilon^2} \qquad m \geq \frac{1}{2\varepsilon^2}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

$$\Pr[(\exists h \in H)(error_{\hat{D}}(h) > error_D(h) + \varepsilon)] \leq |H| e^{-2m\varepsilon^2}$$

$$P(\exists h \in H : R(h) \geq \hat{R}(h) + \varepsilon) \leq |H| e^{-c\varepsilon^2 m}$$

or equivalently, with probability $\geq 1 - \delta, \forall h \in H$,

$$R(h) \leq \hat{R}(h) + c\sqrt{\frac{\ln|H|}{m} + \frac{\ln(1/\delta)}{m}}$$

مروری بر نابرابری‌ها (*Recapitulate of Inequalities*)

کاربرد: برای نشان دادن این‌که

امید ریاضی (ریسک مورد انتظار) ***R(h)*** **(*expected risk*)**

مقداری نزدیک

به

میانگین نمونه ***Ř(h)*** **(*average sample*)**

دارد، از

نابرابری‌های تمرکز **(*concentration inequalities*)**

استفاده می‌کنیم.

توجه: استفاده از **۲** نابرابری ساده:-

۱) **نابرابری مارکوف** **(*Markov's Inequality*)**

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a} \qquad \textit{for } X \geq 0$$

$$P(|X| \geq a) \leq \frac{\mathrm{E}(|X|)}{a}$$

۲) **نابرابری چبیشف** **(*Chebyshev's Inequality*)**

$$P(|X - \mathrm{E}(X)| \geq a) = P((X - \mathrm{E}(X))^2 \geq a^2) \leq \frac{\mathrm{E}[(X - \mathrm{E}(X))^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

Proof:- Markov's Inequalities نابرابری مارکوف

Suppose ***X*** *is* ***continuous*** ($X \geq 0$) *with* ***density f*** .

$$E[X] = \int_0^\infty xf(x)\,dx$$

$$= \int_0^a xf(x)\,dx + \int_a^\infty xf(x)\,dx$$

$$\geq \int_a^\infty xf(x)\,dx$$

$$\geq \int_a^\infty af(x)\,dx$$

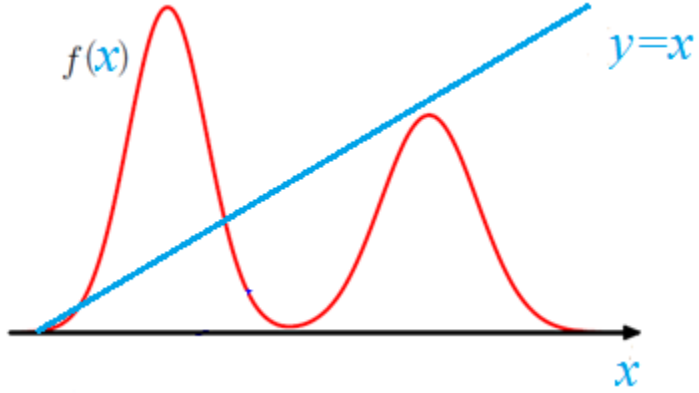
$$= a\int_a^\infty f(x)\,dx$$

$$= aP\{X \geq a\}$$

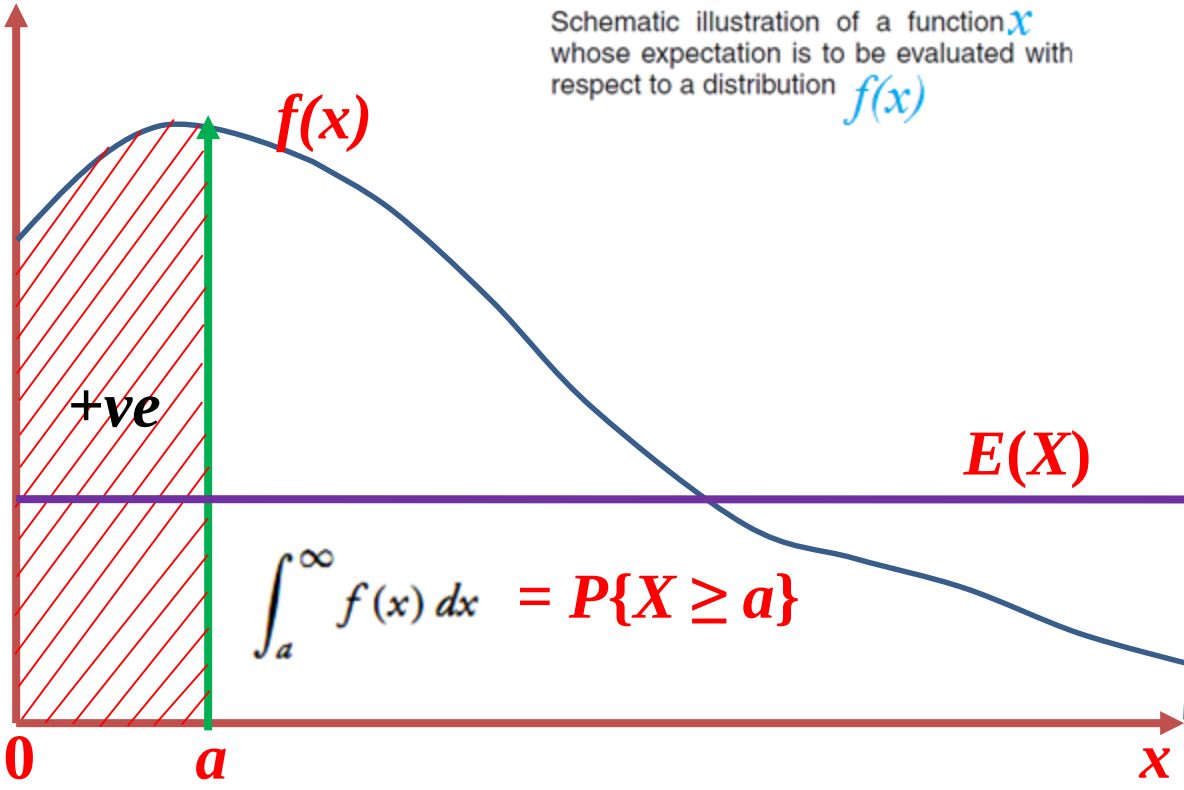
$$E[X] \geq aP\{X \geq a\}$$

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Q.E.D



Schematic illustration of a function x whose expectation is to be evaluated with respect to a distribution $f(x)$



۱) نابرابری مارکف (*Markov's Inequalities*)

قضیه نابرابری مارکُف:

در **نظریه‌ی احتمالات** برای **متغیرهای تصادفی** X با **مقادیر نامنفی** $X \geq 0$، و $E[X]$ وجود داشته باشد،

و

a مثبت باشد

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

کران بالا

برای

تخمین احتمال

می‌دهد.

مثال:

امکان ندارد بیش از **۲۰** درصد از **افراد جامعه**

درآمدی بیش از **۵ برابر متوسط درآمد جامعه**

را داشته باشند.

درآمد افراد جامعه X و $X \geq 0$

متوسط درآمد افراد جامعه $E[X]$

درصد افراد جامعه a

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

<درصد افراد جامعه> <امکان ندارد> (< متوسط درآمد جامعه *5> <بیش‌تر و یا مساوی> <درآمد افراد جامعه>) <احتمال>

$$P\ (\ X\ \geq\ 5E[X]\ )\ \leq\ 0.2$$

۱) نابرابری مارکف (*Markov's Inequalities*) (ادامه)

توجه: **نابرابری مارکف**

عموماً برای

تخمین احتمالات

به‌کار می‌رود و این

کران بالا از

دقت بالایی برخوردار

نیست!!

تعریف: اگر X **متغیر تصادفی نامنفی** ($X \geq 0$) و $a > 0$ باشد:

نتیجه این نابرابری این است که اگر

X متغیر تصادفی دلخواه،

$E[X]$ وجود داشته باشد و

$a > 0$ باشد:

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

$$P(|X| \geq a) \leq \frac{\mathrm{E}(|X|)}{a}$$

و از این می‌توان نتیجه گرفت: $$P(|X-\mathrm{E}(X)|\geq a)=P((X-\mathrm{E}(X))^2\geq a^2)\leq \frac{\mathrm{E}[(X-\mathrm{E}(X))^2]}{a^2}=\frac{\mathrm{Var}(X)}{a^2}$$

که از این طریق می‌توان به نابرابری دیگری به نام

۲) نابرابری چبیشف (***Chebyshev's inequality***)

دست یافت.

مثال:

- ***Markov's inequality*** *is used to prove* ***Chebyshev's inequality***.
- ***Markov's inequality*** *can be used to show that, for a* ***nonnegative random variable***, *the* ***mean (μ)*** *and a* ***median (m)*** *are such that* $m \leq 2\mu$.

جمع‌بندی:-

$$P(X\geq a)\leq \frac{\mathrm{E}(X)}{a}$$

$$P(|X|\geq a)\leq \frac{\mathrm{E}(|X|)}{a}$$

$$P(|X-\mathrm{E}(X)|\geq a)=P((X-\mathrm{E}(X))^2\geq a^2)\leq \frac{\mathrm{E}[(X-\mathrm{E}(X))^2]}{a^2}=\frac{\mathrm{Var}(X)}{a^2}$$

۲) نابرابری چبیشف (*Chebyshev's Inequality*)

If X *is a* ***random variable*** *with* ***mean*** $\mu = E[X]$ *and* ***variance*** $\sigma^2 = Var(X)$,

then for any ***value*** $k > 0$

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

$P(|Z| \geq k) \leq 1/k^2$

$Z = (X - \mu)/\sigma$

$P(|Z| > 1) \leq 1$
$P(|Z| > 2) \leq (1/4 = 0.25)$
$P(|Z| > 3) \leq (1/9 = 0.1111)$

Proof:-

Since $(X - \mu)^2$ *is a* ***nonnegative random*** *variable,*

we can apply ***Markov's inequality***

(with $a = k^2$*) to obtain*

$$P\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2}$$

$P(|Z| > 1) \leq (1 - 0.68 = 0.32)$
$P(|Z| > 2) \leq (1 - 0.95 = 0.05)$
$P(|Z| > 3) \leq (1 - 0.9997 = 0.0003)$

But since $(X - \mu)^2$ *if and only if* $|X - \mu| \geq k$, *we have*

$$P\{|X - \mu| \geq k\} \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

$$P(|X - \mathrm{E}(X)| \geq a) = P((X - \mathrm{E}(X))^2 \geq a^2) \leq \frac{\mathrm{E}[(X - \mathrm{E}(X))^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

Q.E.D

$$P(|X - \mathrm{E}(X)| \geq a) = P((X - \mathrm{E}(X))^2 \geq a^2) \leq \frac{\mathrm{E}[(X - \mathrm{E}(X))^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

The Weak Law of Large Numbers

Note: *Let* $X_1, X_2, \ldots,$ *be a*

***sequence** of **independent** and **identically distributed (iid) random variables**,*

each having

mean $E[X_i] = \mu$.

Then, for any $\varepsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} \to 0 \quad \text{as } n \to \infty$$

Proof:-

We shall prove the result only under the additional assumption that the

random variables

have a

finite variance σ^2.

Now, as

$$E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{and} \quad \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

*it follows from **Chebyshev's inequality** that*

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

Q.E.D

By replacing: k by $k\sigma$, we can write **Chebyshev's inequality** as

$$P\{|X - \mu| > k\sigma\} \leq 1/k^2$$

It states: **The probability a random variable differs** from its **mean** by **more** than **k standard deviations** is **bounded** by $1/k^2$.

$$P\{|X - \mu| \geq k\} \leq \sigma^2/k^2$$

$P(|Z| \geq k) \leq 1/k^2$

$Z = (X - \mu)/\sigma$

$P(|Z| > 1) \leq 1$

$P(|Z| > 2) \leq 1/4$

$P(|Z| > 3) \leq 1/9$

Weak law of large numbers: By using **Chebyshev's inequality** to prove the **weak law** of **large numbers**, which states that the **probability** that the **average** of the **first n terms** in a **sequence** of **independent** and **identically distributed (iid)** **random variables differs** by its **mean** by **more** than ε goes to **0** as **n** goes to **infinity**.

$P(|Z| > 1) \leq 1$

$P(|Z| > 2) \leq (1/4 = 0.25)$

$P(|Z| > 3) \leq (1/9 = 0.1111)$

$P(|Z| > 1) \leq (1 - 0.68 = 0.32)$

$P(|Z| > 2) \leq (1 - 0.95 = 0.05)$

$P(|Z| > 3) \leq (1 - 0.9997 = 0.0003)$

A Fact: *The importance of*

Markov's

and

Chebyshev's inequalities

is that they enable us to derive

bounds *on* ***probabilities***

when only the

mean,

or both the

mean *and the* ***variance***,

of the probability distribution are known.

Of course: *If the*

actual distribution

were known, then the

desired probabilities

could be

exactly computed

and we would not need to

resort to bounds.

$P(|Z| > 1) \leq 1$

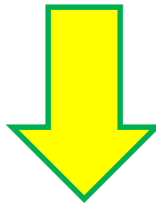
$P(|Z| > 2) \leq (1/4 = 0.25)$

$P(|Z| > 3) \leq (1/9 = 0.1111)$

$P(|Z| > 1) \leq (1 - 0.68 = 0.32)$

$P(|Z| > 2) \leq (1 - 0.95 = 0.05)$

$P(|Z| > 3) \leq (1 - 0.9997 = 0.0003)$



Example:-

Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean **50**.

(a) *What can be said about the probability that this week's production will exceed* **75**?

(b) *If the* ***variance*** *of a week's production is known to be equal to* **25**, *then*

what can be said about the ***probability*** *that this week's production*

will be between **40** *and* **60**?

Solution:-

Let ***X*** *be the number of items that will be produced in a week:*

(a) *By Markov's inequality*

$$P\{X > 75\} \leq \frac{E[X]}{75} = \frac{50}{75} = \frac{2}{3}$$

(b) *By Chebyshev's inequality*

$$P\{|X - 50| \geq 10\} \leq \frac{\sigma^2}{10^2} = \frac{1}{4}$$

$$P\{|X - 50| < 10\} \geq 1 - \frac{1}{4} = \frac{3}{4}$$

and so the probability that this week's production will be between **40** *and* **60** *is at least* **0.75**.

Chebyshev's inequality

Let **m** *and* **s** *be the* **sample mean** *and* **sample standard deviation** *of a data set. Assuming that*

s > 0, ***Chebyshev's inequality*** *states that for any value of*

$k \geq 1$, *greater than* **$100(1 - 1/k^2)$** *percent*

of the data lie within the interval from **$m - ks$** *to* **$m + ks$.**

Thus, by

letting **k = 1.5**,

we obtain from ***Chebyshev's*** *inequality that*

greater *than* **100(5/9) = 55.56** *percent of the*

data from any data set lies within a

distance **1.5s** *of the sample mean* **m**;

letting **k = 2**

shows that greater than **75** *percent of the*

data lies within **2s** *of the sample mean; and*

letting **k = 3**

shows that greater than **800/9 ≈ 88.9** *percent of the*

data lies within **3** *sample standard deviations of* **m**.

$P(|Z| > 1) \leq 1$

$P(|Z| > 2) \leq (1/4 = 0.25)$

$P(|Z| > 3) \leq (1/9 = 0.1111)$

$P(|Z| > 1) \leq (1 - 0.68 = 0.32)$

$P(|Z| > 2) \leq (1 - 0.95 = 0.05)$

$P(|Z| > 3) \leq (1 - 0.9997 = 0.0003)$

مروری بر نابرابری‌ها (*Recapitulate of Inequalities*) ادامه

۲ نابرابری فوق خیلی کلی است،

علاقمندیم

کران‌ها،

هم‌گرایی قوی‌تر (**نمایی**) (***exponential***) ***convergence***) (***stronger***

بدهند! ($1 - x \leq exp\ (- x)$)

آن‌گاه

۱) **نابرابری هافدینگ** (***Hoeffding's Inequality***)

برای **مجموع** از **متغیرهای مستقل کران‌دار**

(***sums of independent bounded variables***)

معرفی شده و نشان داده خواهد شد **هم‌گرایی نمایی دست‌یافتنی** است،

و به‌دنبال

۲) **نابرابری مک‌دیارمید** (***McDiarmid's***)

با نام

تعمیم نابرابری هافدینگ

(***generalization of Hoeffding's inequality***)

تحت عنوان

اختلاف‌های کران‌دار (***Bounded Differences***) **و**

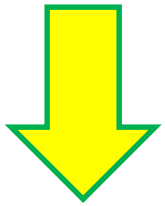
هافدینگ/آزوما (***Hoeffding/Azuma***)

معرفی خواهد شد.

Some general tools for error analysis and bounds:-

Layout of the rest of Lecture notes

- ***Hoeffding's inequality*** (***additive***)
  - ***Hoeffding's Lemma***
    - ***Convexity***
    - ***Taylor's series expansion***
  - ***Hoeffding's Theorem***
    - ***Chernoff's Bounding techniques***
    - ***Markov's inequality***
    - ***Hoeffding Lemma***
  - ***Hoeffding's inequality***
    - ***Hoeffding's Theorem***
    - ***One-sided and two sided inequalities based on Chernoff's bounding techniques***



- ***Chernoff's (bounds) inequality*** (***multiplicative***)
  - ***Chernoff's bounding techniques***
- ***McDiarmid's inequality*** (***more general***)
  - ***Hoeffding's Lemma***
  - ***Hoeffding's inequality***

Hoeffding's Inequality

A fact: *Most of the **methods** that have been introduced are somehow **heuristically**, in the sense that we have not **rigorously proven** that they **actually work**!*

Roughly speaking: *In **supervised learning** we have taken the following **strategy**:*

- *Pick some **class of functions h(x)** (**decision trees**, **linear functions**, etc.)*
- *Pick some **loss function**, measuring how we would like **h(x)** to **perform** on **test data**.*
- *Fit **h(x)** so that it has a **good average loss** on **training data**. (Perhaps using **cross-validation** to **regularize**).*

Question: *What is **missing** here is a **proof** that the **performance** on **training data** is **indicative** of **performance** on **test data**.*

Define powerful:

*We **intuitively** know that the more "**powerful**"*
*the **class** of **functions** is,*
*the **more training data***
*we will **tend** to need,*
but we have not made the
***definition** of "**powerful**",*
nor this
***relationship** "**precise**".*

Demonstration:

*Will study **2** of the most basic ways of*
***characterizing** the*
*"**power**" of a **set** of **functions**.*
Will look at some
rigorous bounds
***confirming** and **quantifying** our above **intuition**,*
*that is, for a **specific** set of **functions**, how much*
training data
*is needed to **prove** that the*
***function** will work **well** on*
***test data**?*

Hoeffding's inequality: *The basic tool that will be used to* ***understand generalization,*** *is* ***Hoeffding's inequality****.*

This is a general result in ***probability theory****.*

It is ***extremely widely*** *used in* ***machine learning theory****.*

There are several equivalent forms of it,

and it is worth understanding these in detail.

Theorem: ***Hoeffding's Inequality-1***

Let $X_1, ..., X_n$ *be*

random i.i.d variables*,*

such that $0 \leq X_i \leq 1$.

Then,

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2)$$

The intuition for this result is very simple.

When we ***average a bunch of random variables*** X_i *,*

we should ***usually*** *get something* ***close*** *to the* ***expected value****.*

Hoeffding quantifies **1)** *"****usually****" and* **2)** *"****close****" for us.*

Note: *The general form of* ***Hoeffding's inequality*** *is for* ***random variables*** *in some* ***range*** $a \leq X_i \leq b$

Main Concern: *As we will mostly be worrying about the* **0/1** ***classification error***,
the above form is fine for our purposes.

Note: *Can also* ***rescale*** *our* ***variables*** *to lie between* **0** *and* **1**,
and then apply the above proof) .

Examples: *Compare* ***Hoeffding's inequality*** *to the*
true probability *of deviating from the*
mean *by more than* ε *for* ***binomial distributed variables***,
with $E[X] = P$.

For $P = \frac{1}{2}$ *the* ***bound*** *is not bad. However,*

For $P = {}^{1}/_{10}$ *the* ***bound*** *is not good at all.*

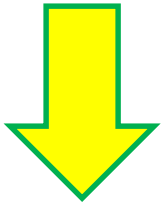
What is happening is that ***Hoeffding's inequality*** *does not make use of any properties of the* ***distribution***, *such as its*

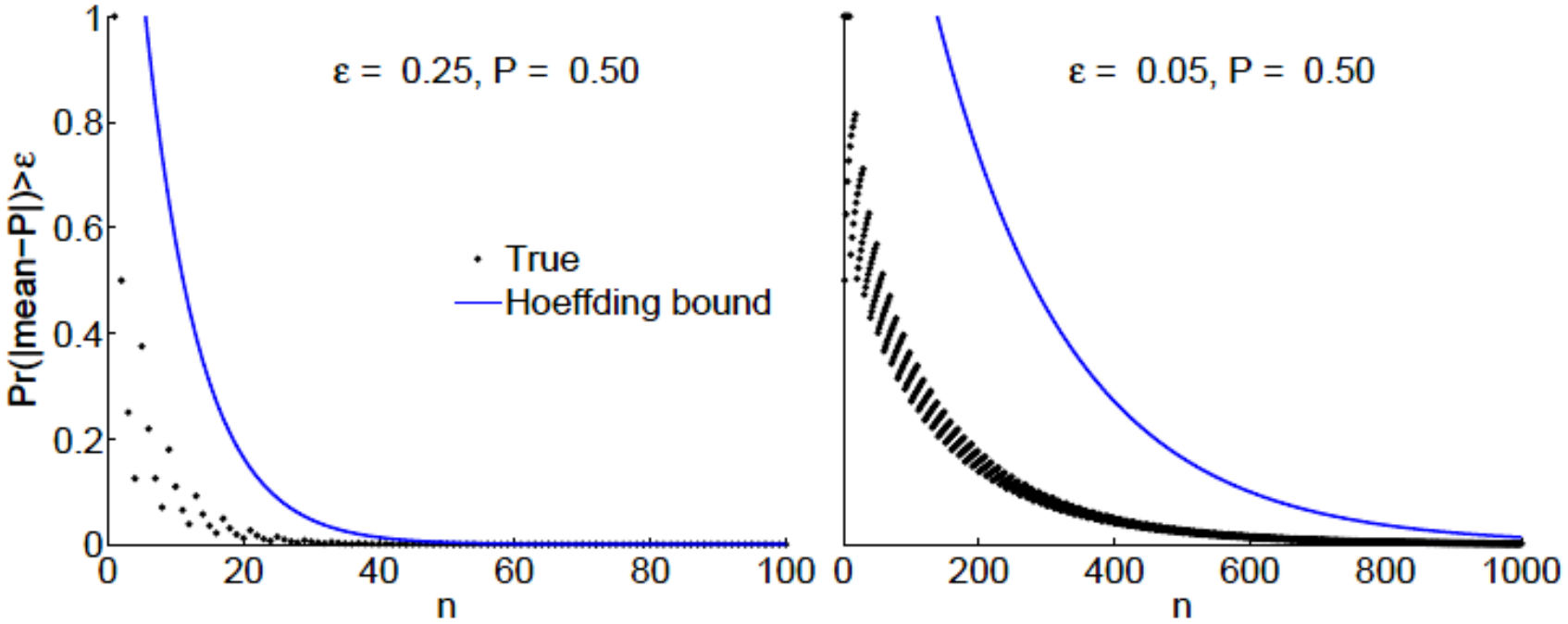
mean *or* ***variance***.

In a way, this is great, since we can calculate it just from **n** *and* ε.

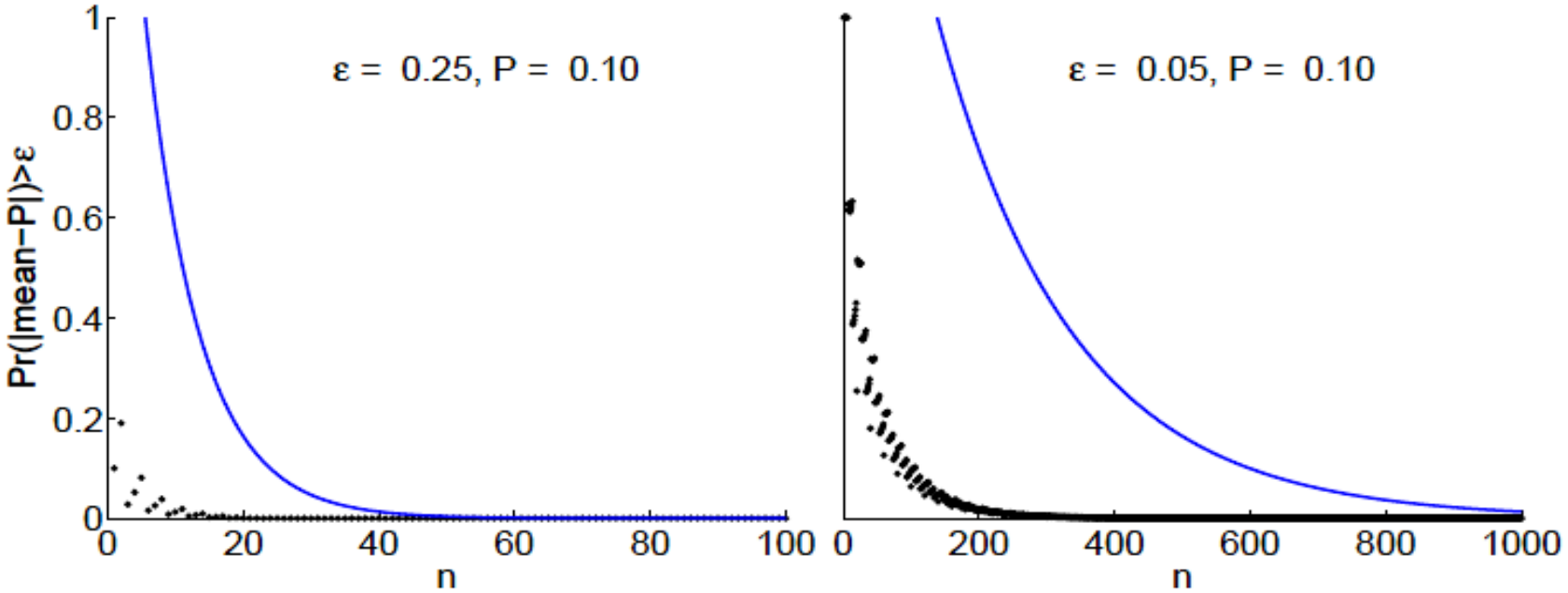
The price we pay for this generality is that some
distributions *will converge to their*
means *much faster than*
Hoeffding *is* ***capable*** *of knowing.*

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2)$$





Binomial *distributed variables, with* ***E[X] = P***



$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2)$$

Question: *How will these figures look with* ***P = 0.9?!!***

Will ***Hoeffding*** *be* ***loose*** *or* ***tight*** ***?!!***

Hoeffding-2: *Another form of* ***Hoeffding's inequality*** *of the following:-*

Theorem: ***Hoeffding-2***

Suppose we choose $n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$

Then, with ***probability*** *at* ***least*** $1 - \delta$, (*click here*)

the ***difference*** *between the* ***empirical mean*** $\frac{1}{n} \sum_{i=1}^{n} \zeta_i$

and the ***true mean*** $E[X]$ *is at* ***most****.*

Hoeffding-2: *This second form is very useful.*

*To understand it, we can think of setting two "****slop****" parameters:*

- *The "****accuracy*** ε*" says how far we are willing to allow the* ***empirical*** ***mean*** *to be from the* ***true*** ***mean****.*
- *The "****confidence*** δ*" says what probability we are willing to allow of "****failure****".*

(That is, a deviation larger than ε*)*

If we choose these two parameters, equation $n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$

*tells us how much it will "****cost****" in terms of* ***samples****.*

Note: *Informally,* ***accuracy*** *is* ***expensive****,*
while ***confidence*** *is* ***cheap****.*

Example: *Explicitly, if we find* ***n*** *for some* ε *and* δ
then we may decide that we want
10 *times more confidence.*

Example1: *We can calculate that for* $\delta' = \delta / 10$*, we will need*

?

$$n' = \frac{1}{2}\left(\frac{1}{\epsilon}\right)^2 \log \frac{2 \cdot 10}{\delta} = n + \frac{1}{2}\left(\frac{1}{\epsilon}\right)^2 \log(10) = n + C(\epsilon)$$

samples *to achieve this. Thus,*
we can just add ***C(ε)*** ***(a constant number)*** *for* ***extra samples****.*

?

Example2: *If we would like* **100** *times* ***more confidence****,*
we can just add **2C(ε)** *extra samples.*

Another way of looking at this is that $n \propto \log \frac{2}{\delta}$ *or* $\delta \propto \frac{2}{\exp(n)}$

To emphasize: *This is great, this is the best we could imagine.*
We will see below that the
*"****cheapness****" of* ***confidence*** *turns out*
to be key to our ***goal*** *of*
creating learning bounds*.*

$$n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$$

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2)$$

Example: *On the other hand, **accuracy** is quite **expensive**.*

Example1: *Suppose that we decide we want **10** times more **accuracy**.*

*We can **calculate** that for $\epsilon' = \epsilon/10$, we will need **100n** **samples**.*

*An **increase** of a **factor** of **100**.*

Example2: *If we want $\epsilon' = \epsilon/100$, i.e., **100** times **more accuracy**,*

*we will need a **factor** of **10,000n** (**times** more)**samples**.*

Another way of looking at this is that $\epsilon \propto \frac{1}{\sqrt{n}}$

*Yet another way of stating **Hoeffding's inequality** is*

Theorem: **Hoeffding-3**

*If we draw **n** **samples**, then with **probability** at least* $1 - \delta$,

*the **difference** between the **empirical mean*** $\frac{1}{n}\sum_{i=1}^{n} X_i$

*and the true **mean E[X]** is at most **ε**, where* $\epsilon \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$

*With only this **simple tool**,*

*we can actually **derive quite** a*

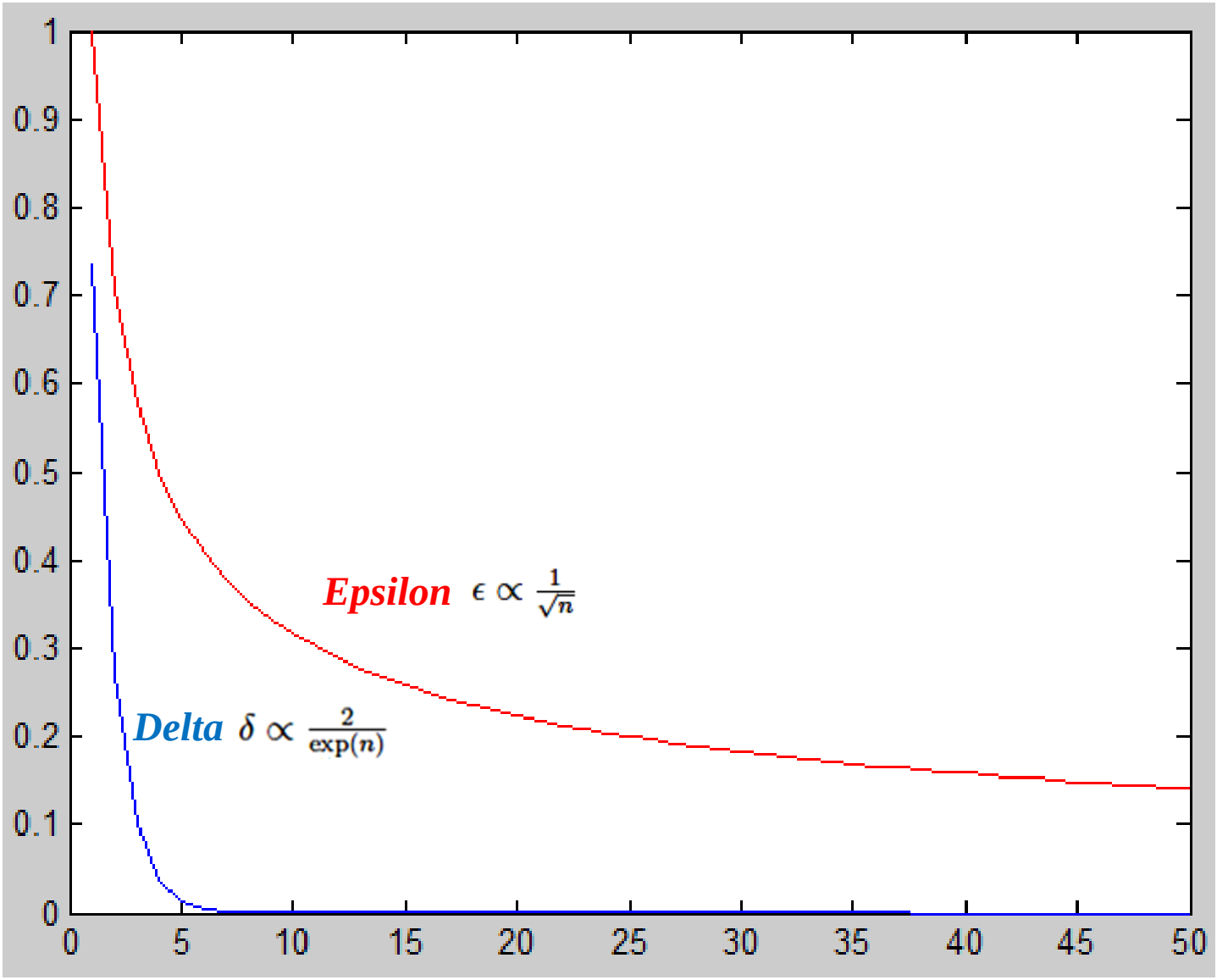
***lot about learning theory**.*

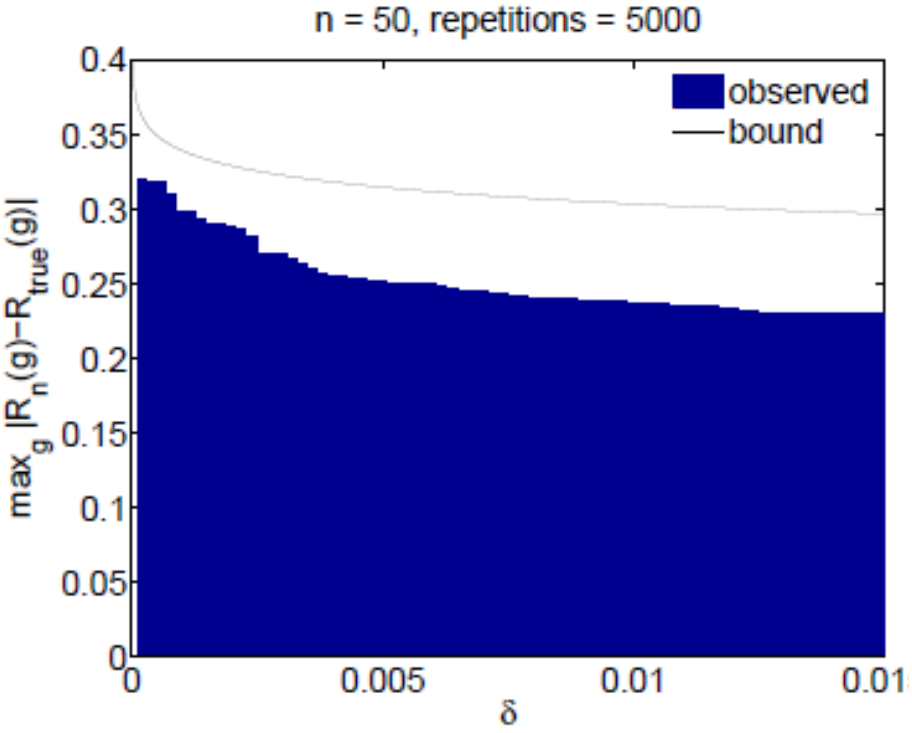
$$\delta \propto \frac{2}{\exp(n)}$$

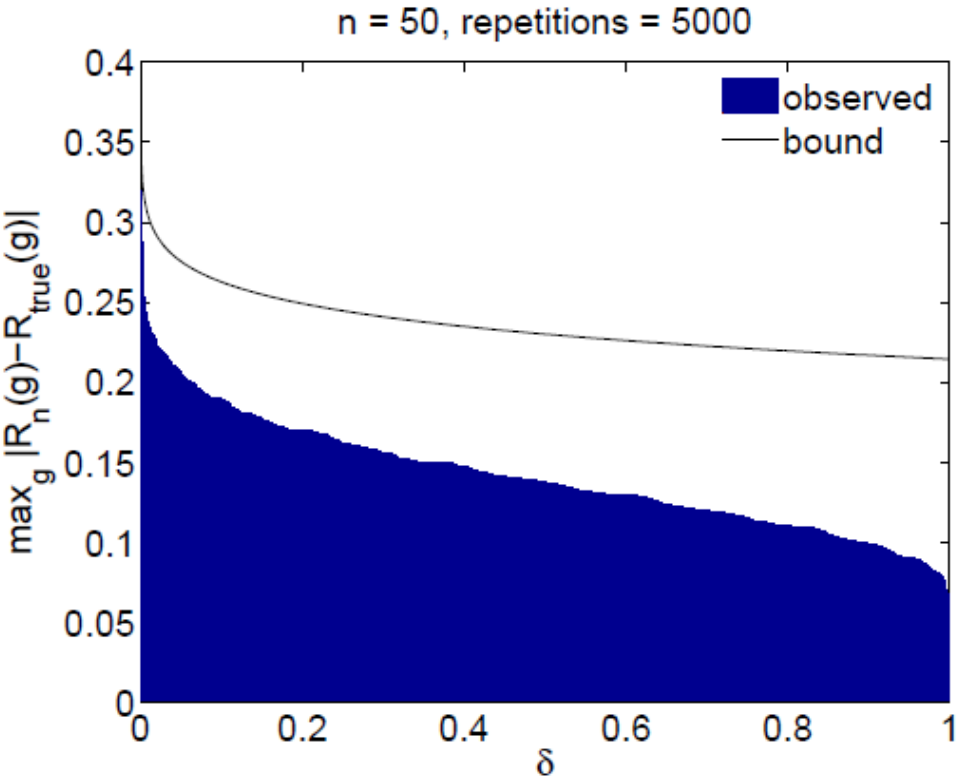
$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}$$

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2)$$

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp(-2n\epsilon^2) \qquad n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}$$




n = 50, repetitions = 5000
observed
bound
$\max_g |R_n(g) - R_{true}(g)|$
δ
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.005
0.01


n = 50, repetitions = 5000
observed
bound
$\max_g |R_n(g) - R_{true}(g)|$
δ
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.2
0.4
0.6
0.8
1

Discussion

A fact: *The **fundamental weakness** of the above **bounds** is their **looseness**.*

In practice: *The **bound** on the **difference between** the **training risk** and **true risk** is often **hundreds** of times **higher** than the **true difference**.*

Good Enough: *There are many worst-case assumptions leading to the **bound** that are often not so bad in practice.*

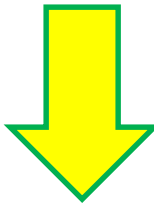
Open Research: *__Tightening__ the **bounds remains** an **open area** of **research**.*

Quite Good: *On the other **hand**, the **bound** can sometimes work well in **practice despite** its **looseness**.*

The reason: *Is that we are **fundamentally interested** in **performing model selection**, not **bounding test errors**.*

SRM: *The model selected by **Structural Risk Minimization** (**SRM**) is sometimes **quite good**, despite the **looseness** of the **bound**.*

خاتمه بخش ۱ جزوه



Hinge loss

In machine learning, the ***hinge loss*** *is a* ***loss function*** *used for* ***training classifiers****.*

*The hinge loss is used for "****maximum-margin****" ****classification****,*

most notably for ***support vector machines*** *(****SVMs****).*

For an ***intended output*** $y = \pm 1$ *and a* ***classifier score*** $f(x)$*, the*

hinge loss *of the* ***prediction*** $f(x)$ *is defined as* $L_{oss}(f(x)) = \max(0 , 1 - y * f(x))$

Note that $f(x)$ *should be the*

$$\ell(y) = \max(0, 1 - t \cdot y)$$

*"****raw****" ****output*** *of the* ***classifier's decision function****,*

not the ***predicted class label****.*

e.g., in linear ***SVMs****.*

It can be seen that when y *and* $f(x)$

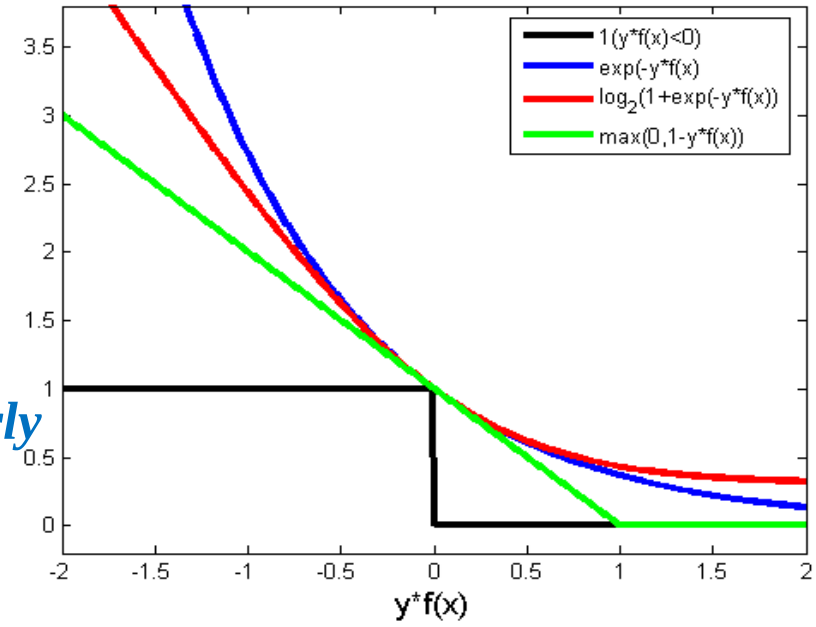
have the ***same sign*** *(****meaning*** $f(x)$

predicts *the* ***right class****) and , the*

hinge loss *, but when they have*

opposite sign*,* ***increases linearly***

with $f(x)$ *(****one-sided error****).*

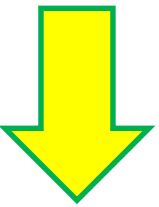


return

$$n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$$

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right] \leq \delta = 2\exp\left(-2n\epsilon^2\right)$$

$$\begin{aligned} &\leq 2\exp(-2\frac{1}{2\epsilon^2}\log\frac{2}{\delta}\epsilon^2) \\ &= 2\exp(-\log\frac{2}{\delta}). \\ &= \delta \end{aligned}$$



return