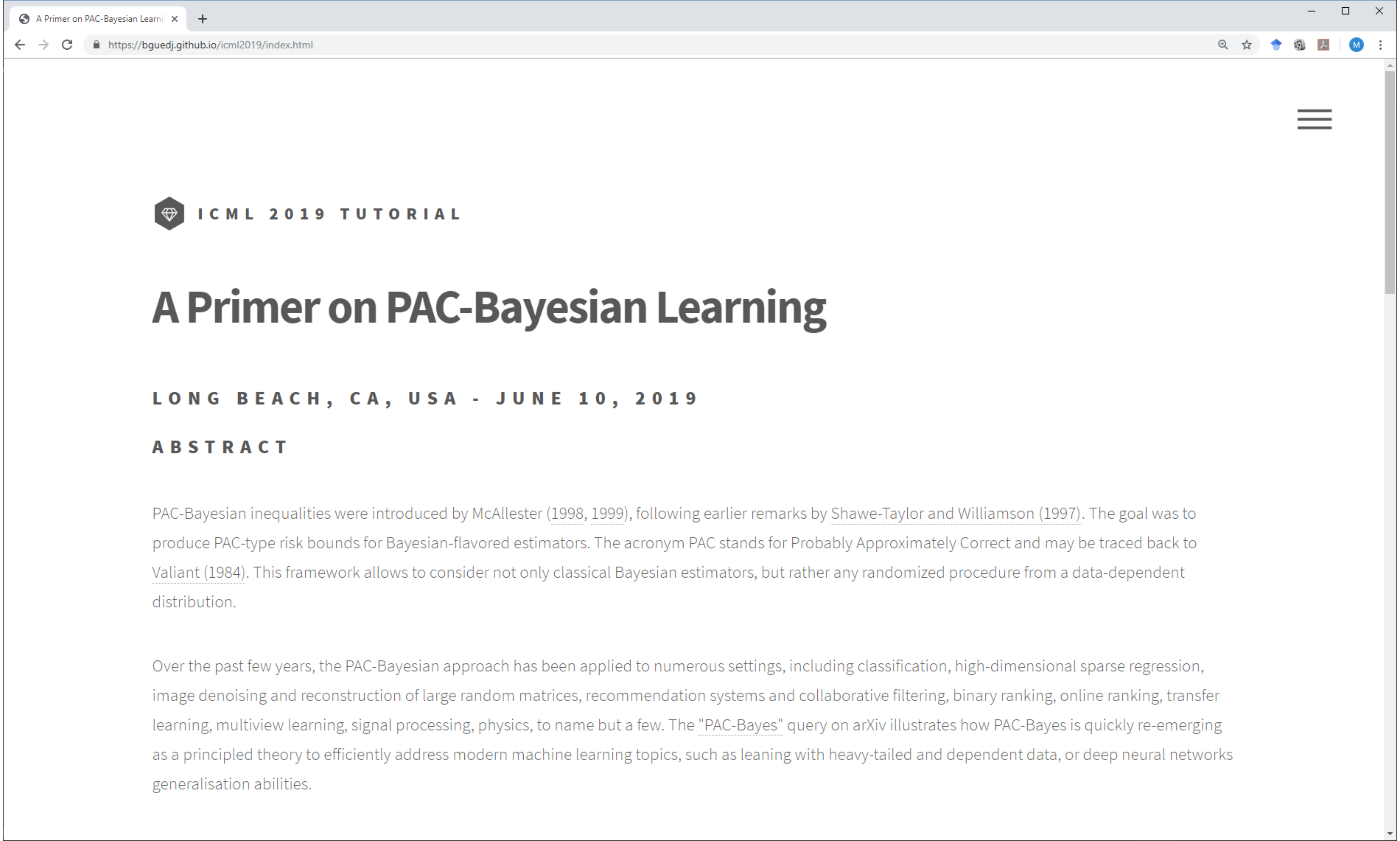
Guest Lecture
by
Marius Popescu

Teacher Transfer → Style Transfer

PAC-Bayes


A Primer on PAC-Bayesian Learni
https://bguedj.github.io/icml2019/index.html

ICML 2019 TUTORIAL

A Primer on PAC-Bayesian Learning

LONG BEACH, CA, USA - JUNE 10, 2019

ABSTRACT

PAC-Bayesian inequalities were introduced by McAllester (1998, 1999), following earlier remarks by Shawe-Taylor and Williamson (1997). The goal was to produce PAC-type risk bounds for Bayesian-flavored estimators. The acronym PAC stands for Probably Approximately Correct and may be traced back to Valiant (1984). This framework allows to consider not only classical Bayesian estimators, but rather any randomized procedure from a data-dependent distribution.

Over the past few years, the PAC-Bayesian approach has been applied to numerous settings, including classification, high-dimensional sparse regression, image denoising and reconstruction of large random matrices, recommendation systems and collaborative filtering, binary ranking, online ranking, transfer learning, multiview learning, signal processing, physics, to name but a few. The "PAC-Bayes" query on arXiv illustrates how PAC-Bayes is quickly re-emerging as a principled theory to efficiently address modern machine learning topics, such as leaning with heavy-tailed and dependent data, or deep neural networks generalisation abilities.

PAC-Bayes: The Framework

- Let $\mathcal{H}$ be a hypothesis class and P a prior distribution over $\mathcal{H}$. That is, we assign a probability (or density if $\mathcal{H}$ is continuous) $P(h) \geq 0$ for each $h \in \mathcal{H}$.
- Following the Bayesian reasoning approach, the output of the learning algorithm is not necessarily a single hypothesis. Instead, the learning process defines a posterior probability Q over $\mathcal{H}$.
- In the context of a supervised learning problem, where $\mathcal{H}$ contains functions from X to Y, one can think of Q as defining a randomized prediction rule as follows: whenever we get a new instance $\mathbf{x}$, we randomly pick a hypothesis $h \in \mathcal{H}$ according to Q and predict $h(\mathbf{x})$.

Example
Bayesian Regression

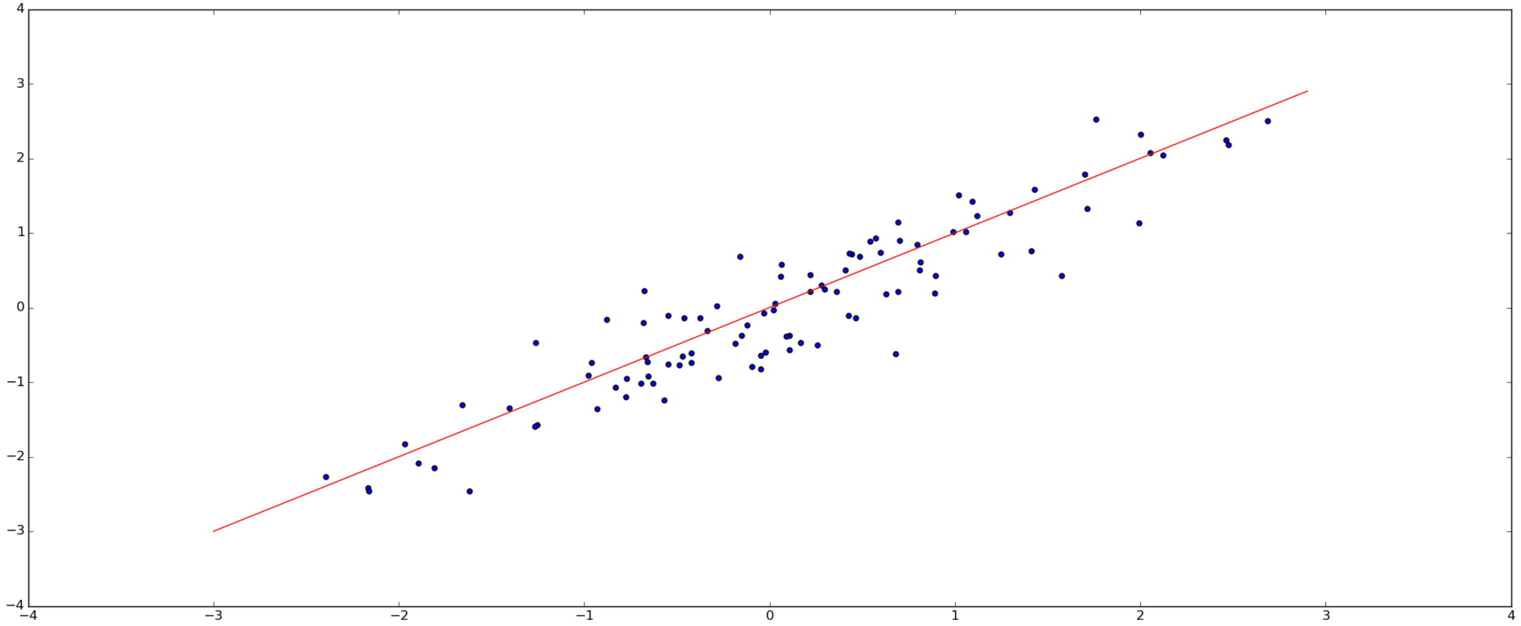
Let's assume we have a one-dimensional dataset:

The goal is to predict as a function of

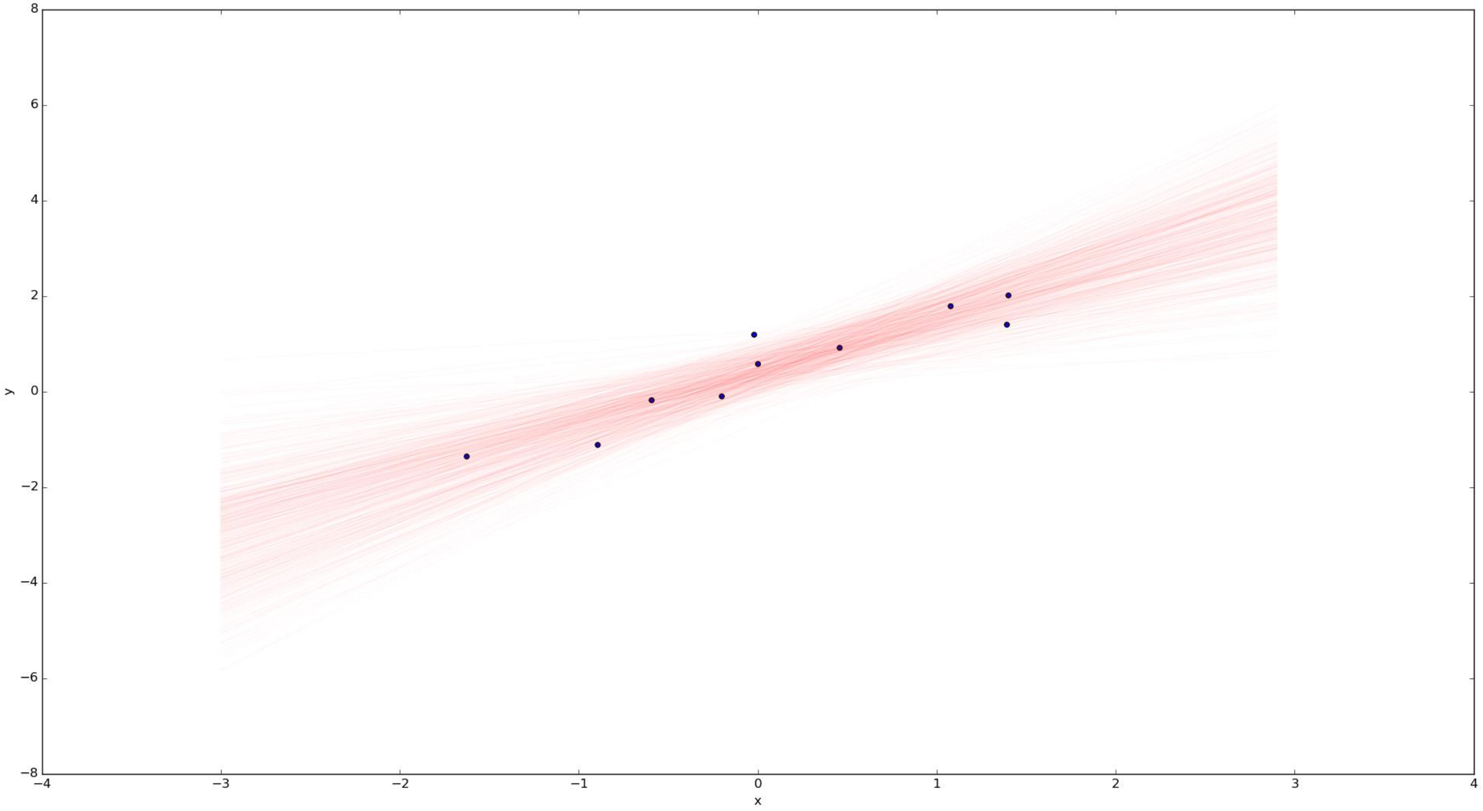
Our model describing is:

where and are unknown parameters, and is the statistical noise (a random variable)

Our goal will be to compute a *posterior* on that represents our degree of belief that any particular is the "correct" one



A Sample of Regression Lines (n = 10)



PAC-Bayes: The Framework

We define the loss of Q on an example z to be:

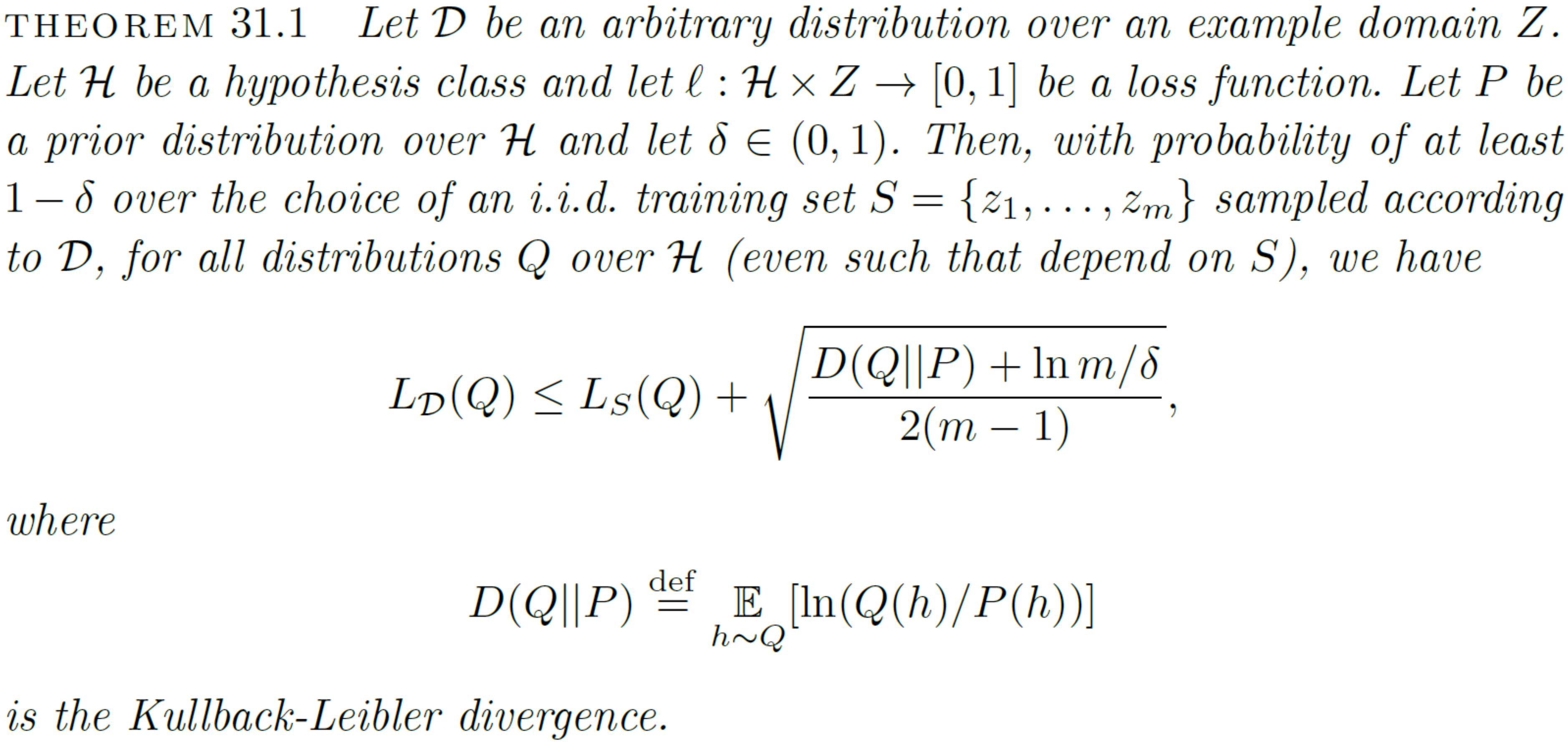
$$\ell(Q,z) = \underset{h\sim Q}{\mathrm{E}}[\ell(h,z)]$$

By the linearity of expectation, the generalization loss and training loss of Q can be written a s:

$$L_D(Q) = \underset{h\sim Q}{\mathrm{E}}[L_D(h)]$$

$$L_S(Q) = \underset{h\sim Q}{\mathrm{E}}[L_S(h)]$$

A PAC-Bayes Bound



THEOREM 31.1 *Let $\mathcal{D}$ be an arbitrary distribution over an example domain Z. Let $\mathcal{H}$ be a hypothesis class and let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. Then, with probability of at least $1-\delta$ over the choice of an i.i.d. training set $S = \{z_1, \ldots, z_m\}$ sampled according to $\mathcal{D}$, for all distributions Q over $\mathcal{H}$ (even such that depend on S), we have*

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{D(Q||P) + \ln m/\delta}{2(m-1)}},$$

where

$$D(Q||P) \overset{\text{def}}{=} \underset{h \sim Q}{\mathbb{E}}[\ln(Q(h)/P(h))]$$

is the Kullback-Leibler divergence.

Proof: in the book (chapter 31)

A PAC-Bayes Bound

- The theorem tells us that the difference between the generalization loss and the empirical loss of a posterior Q is bounded by an expression that depends on the Kullback-Leibler divergence between Q and the prior distribution P.
- The theorem suggests that if we would like to minimize the generalization loss of Q, we should jointly minimize both the empirical loss of Q and the Kullback-Leibler distance between Q and the prior distribution.

Remark

Suppose that $\mathcal{H}$ is a finite hypothesis class, set the prior to be uniform over $\mathcal{H}$, and set the posterior to be $Q(h_S) = 1$ for some h_S and $Q(h) = 0$ for all other $h \in \mathcal{H}$. Show that

$$L_{\mathcal{D}}(h_S) \leq L_S(h) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(m/\delta)}{2(m-1)}}.$$

$\Updownarrow$?

Remember

Uniform convergence for finite classes:

$$m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

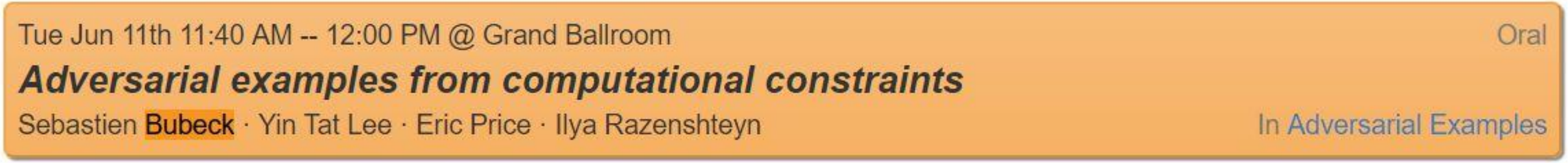
Some References

- D. A. McAllester. Some PAC-Bayesian theorems. Machine Learning, 37(3):355–363, 1999
- Guedj, Benjamin. A primer on PAC-Bayesian learning. arXiv preprint arXiv:1901.05353, 2019

Adversarially Robust Generalization

A PAC Hot Topic

ICML 2019



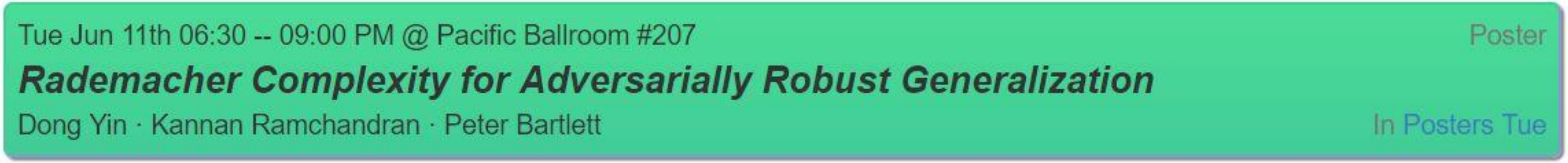
Tue Jun 11th 11:40 AM -- 12:00 PM @ Grand Ballroom

Oral

Adversarial examples from computational constraints

Sebastien Bubeck · Yin Tat Lee · Eric Price · Ilya Razenshteyn

In Adversarial Examples



Tue Jun 11th 06:30 -- 09:00 PM @ Pacific Ballroom #207

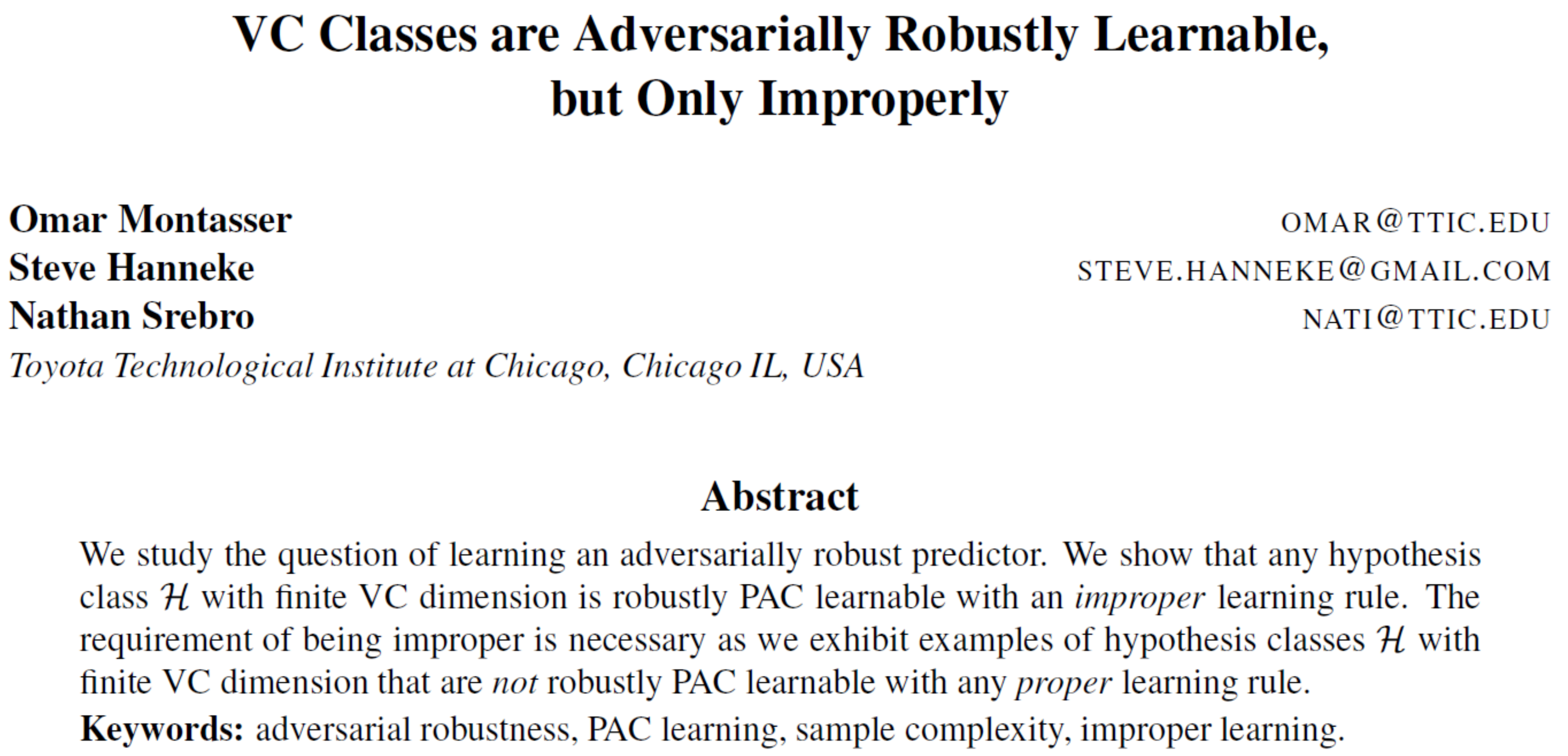
Poster

Rademacher Complexity for Adversarially Robust Generalization

Dong Yin · Kannan Ramchandran · Peter Bartlett

In Posters Tue

COLT 2019



VC Classes are Adversarially Robustly Learnable, but Only Improperly

Omar Montasser OMAR@TTIC.EDU
Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago, Chicago IL, USA

Abstract

We study the question of learning an adversarially robust predictor. We show that any hypothesis class $\mathcal{H}$ with finite VC dimension is robustly PAC learnable with an *improper* learning rule. The requirement of being improper is necessary as we exhibit examples of hypothesis classes $\mathcal{H}$ with finite VC dimension that are *not* robustly PAC learnable with any *proper* learning rule.

Keywords: adversarial robustness, PAC learning, sample complexity, improper learning.