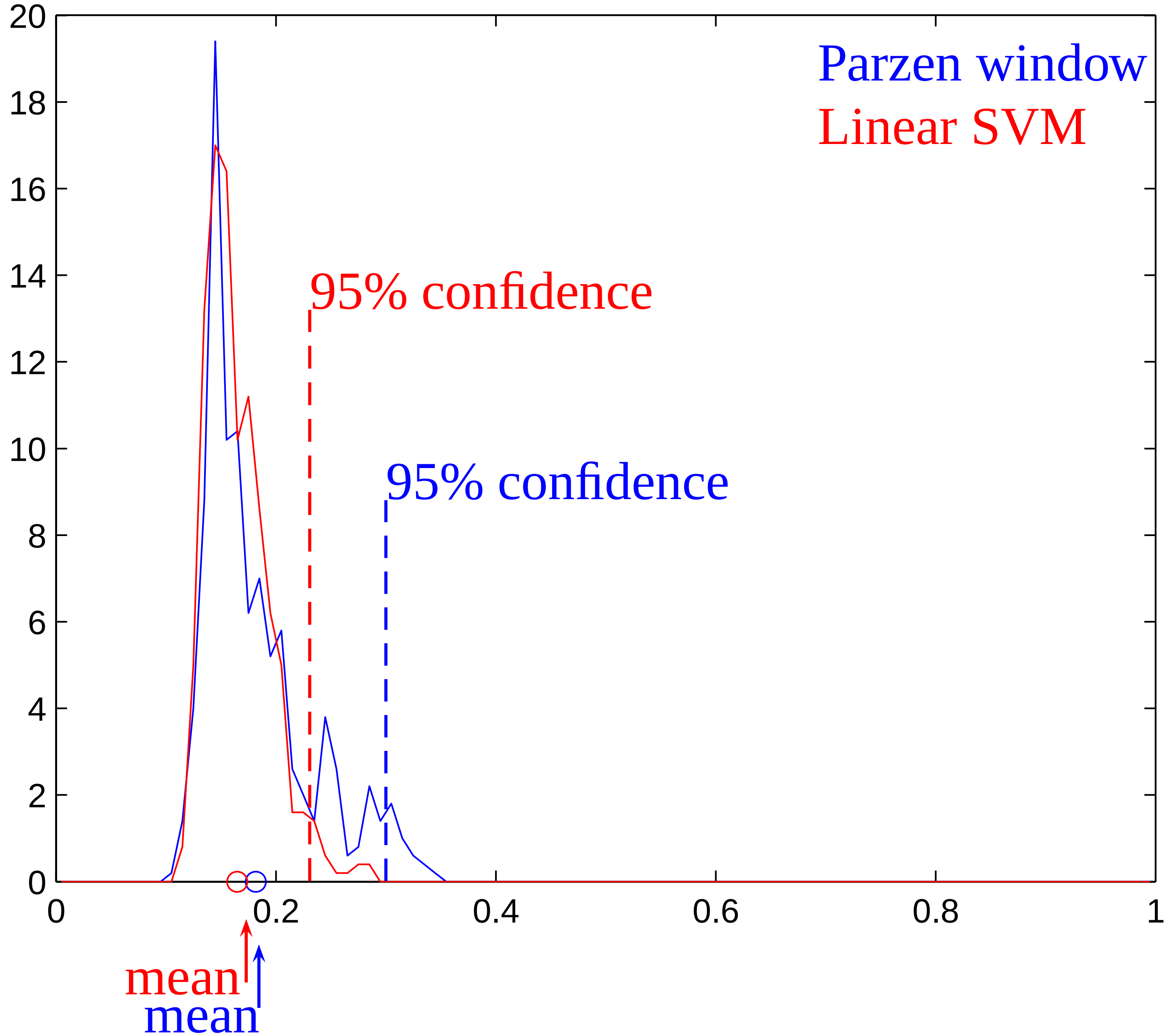
Statistical Learning Theory: A Hitchhiker's Guide

John Shawe-Taylor UCL
Omar Rivasplata UCL / DeepMind

December 2018

Why SLT

Error distribution picture

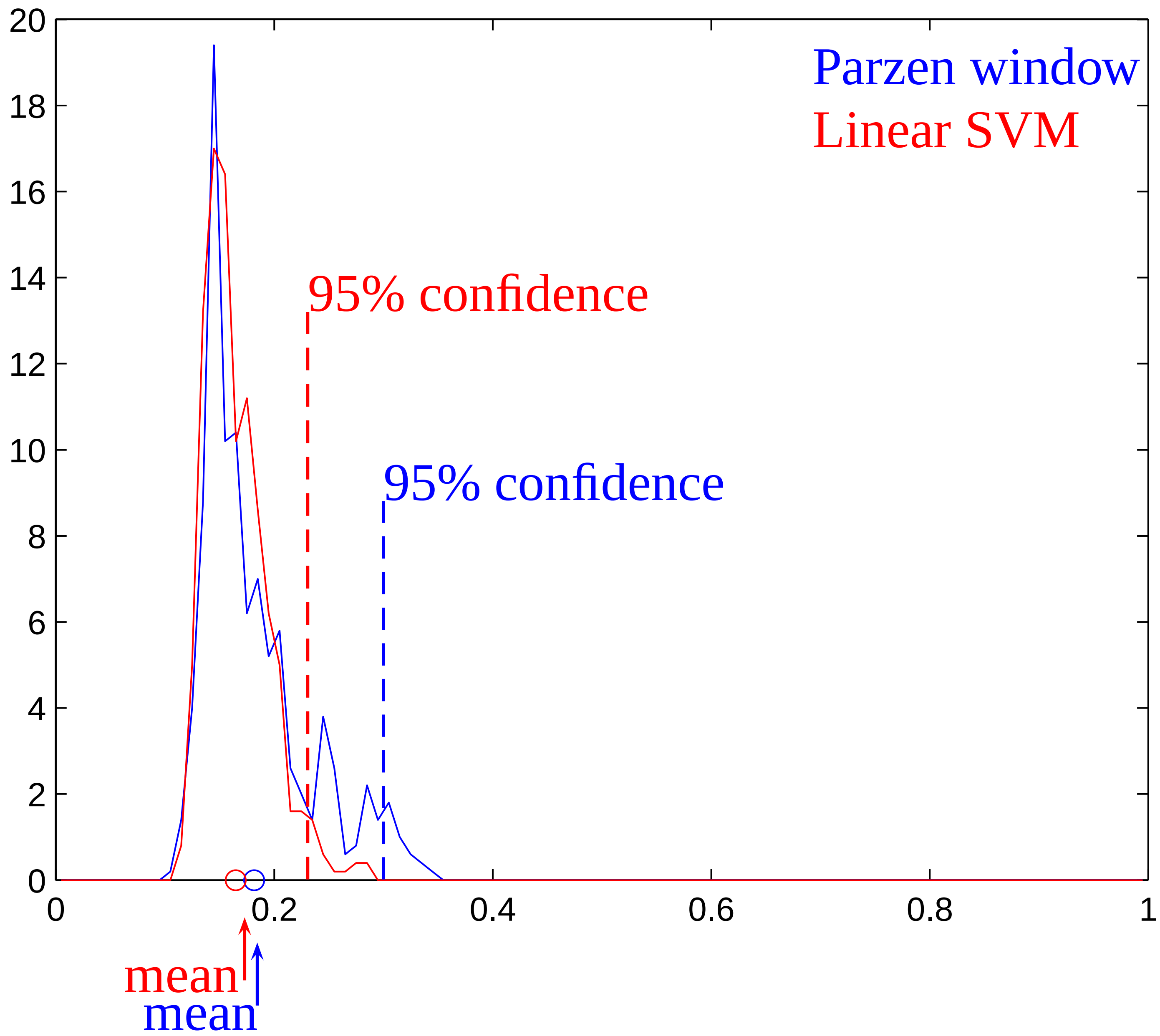


SLT is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?
  ▹ can be misleading: learner only has one sample
- Statistical Learning Theory: tail of the distribution
  ▹ finding bounds which hold with high probability over random samples of size m
- Compare to a statistical test – at 99% confidence level
  ▹ chances of the conclusion not being true are less than 1%
- PAC: probably approximately correct
  Use a 'confidence parameter' δ: $\mathbb{P}^m[\text{large error}] \leq \delta$
  δ is probability of being misled by the training set
- Hence high confidence: $\mathbb{P}^m[\text{approximately correct}] \geq 1 - \delta$

Error distribution picture



Overview

The Plan

- Definitions and Notation: (John)
  - ▹ risk measures, generalization
- First generation SLT: (Omar)
  - ▹ worst-case uniform bounds
  - ▹ Vapnik-Chervonenkis characterization
- Second generation SLT: (John)
  - ▹ hypothesis-dependent complexity
  - ▹ SRM, Margin, PAC-Bayes framework
- Next generation SLT? (Omar)
  - ▹ Stability. Deep NN's. Future directions

What to expect

We will...

- Focus on aims / methods / key ideas
  - Outline some proofs
    - Hitchhiker's guide!

We will not...

- Detailed proofs / full literature (apologies!)
  - Complete history / other learning paradigms
    - Encyclopaedic coverage of SLT

Definitions and Notation

$\boxed{\text{Learning algorithm} \quad \mathsf{A} : \mathcal{Z}^m \to \mathcal{H}}$

- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

  $\mathcal{X}$ = set of inputs

  $\mathcal{Y}$ = set of labels

- $\mathcal{H}$ = hypothesis class

  = set of predictors

  (e.g. classifiers)

Training set (aka sample): $S_m = ((X_1, Y_1), \ldots, (X_m, Y_m))$
a finite sequence of input-label examples.

SLT assumptions:

- A data-generating distribution $\mathbb{P}$ over $\mathcal{Z}$.
- Learner doesn't know $\mathbb{P}$, only sees the training set.
- The training set examples are *i.i.d.* from $\mathbb{P}$: $S_m \sim \mathbb{P}^m$

▹ these can be relaxed (but beyond the scope of this tutorial)

What to achieve from the sample?

Use the available sample to:

(1) learn a predictor
(2) certify the predictor's performance

Learning a predictor:

- algorithm driven by some learning principle
- informed by prior knowledge resulting in inductive bias

Certifying performance:

- what happens beyond the training set
- generalization bounds

Actually $\boxed{\text{these two goals interact}}$ with each other!

Risk (aka error) measures

A loss function $\ell(h(X), Y)$ is used to measure the discrepancy between a predicted label $h(X)$ and the true label Y.

Empirical risk: $R_{\text{in}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(X_i), Y_i)$
(in-sample)

Theoretical risk: $R_{\text{out}}(h) = \mathbb{E}[\ell(h(X), Y)]$
(out-of-sample)

Examples:

- $\ell(h(X), Y) = \mathbf{1}[h(X) \neq Y]$: 0-1 loss (classification)
- $\ell(h(X), Y) = (Y - h(X))^2$: square loss (regression)
- $\ell(h(X), Y) = (1 - Yh(X))_+$: hinge loss
- $\ell(h(X), Y) = -\log(h(X))$: log loss (density estimation)

If classifier h does well on the in-sample (X, Y) pairs...

...will it still do well on out-of-sample pairs?

Generalization gap: $\Delta(h) = R_{\text{out}}(h) - R_{\text{in}}(h)$

Upper bounds: w.h.p. $\boxed{\Delta(h) \leq \epsilon(m, \delta)}$

- $\boxed{R_{\text{out}}(h) \leq R_{\text{in}}(h) + \epsilon(m, \delta)}$

Lower bounds: w.h.p. $\boxed{\Delta(h) \geq \tilde{\epsilon}(m, \delta)}$

Flavours:

- distribution-free
- algorithm-free
- distribution-dependent
- algorithm-dependent

First generation SLT

For one fixed (non data-dependent) h:

$$\mathbb{E}[R_{\text{in}}(h)] = \mathbb{E}\Big[\tfrac{1}{m}\textstyle\sum_{i=1}^{m} \ell(h(X_i), Y_i)\Big] = R_{\text{out}}(h)$$

- $\mathbb{P}^m[\Delta(h) > \epsilon] = \mathbb{P}^m\big[\mathbb{E}[R_{\text{in}}(h)] - R_{\text{in}}(h) > \epsilon\big]$ deviation ineq.
- $\ell(h(X_i), Y_i)$ are independent r.v.'s
- If $0 \leq \ell(h(X), Y) \leq 1$, using Hoeffding's inequality:

$$\mathbb{P}^m[\Delta(h) > \epsilon] \leq \exp\left\{-2m\epsilon^2\right\} = \delta$$

- Given $\delta \in (0, 1)$, equate RHS to δ, solve equation for ϵ, get

$$\mathbb{P}^m\left[\Delta(h) > \sqrt{(1/2m)\log(1/\delta)}\right] \leq \delta$$

- with probability $\geq 1 - \delta$, $\boxed{R_{\text{out}}(h) \leq R_{\text{in}}(h) + \sqrt{\frac{1}{2m}\log\left(\frac{1}{\delta}\right)}}$

Finite function class

Algorithm $\mathsf{A} : \mathcal{Z}^m \to \mathcal{H}$ Function class $\mathcal{H}$ with $|\mathcal{H}| < \infty$

Aim for a uniform bound: $\mathbb{P}^m[\forall f \in \mathcal{H},\ \Delta(f) \leq \epsilon] \geq 1 - \delta$

Basic tool: $\boxed{\mathbb{P}^m(E_1 \text{ or } E_2 \text{ or } \cdots) \leq \mathbb{P}^m(E_1) + \mathbb{P}^m(E_2) + \cdots}$

known as the union bound (aka countable sub-additivity)

$$\mathbb{P}^m\Big[\exists f \in \mathcal{H},\ \Delta(f) > \epsilon\Big] \leq \sum_{f \in \mathcal{H}} \mathbb{P}^m\Big[\Delta(f) > \epsilon\Big]$$

$$\leq |\mathcal{H}| \exp\left\{-2m\epsilon^2\right\} = \delta$$

w.p. $\geq 1 - \delta$, $\boxed{\forall h \in \mathcal{H},\ \ R_{\text{out}}(h) \leq R_{\text{in}}(h) + \sqrt{\frac{1}{2m} \log\left(\frac{|\mathcal{H}|}{\delta}\right)}}$

Uncountably infinite function class?

Algorithm $\mathsf{A} : \mathcal{Z}^m \to \mathcal{H}$ Function class $\mathcal{H}$ with $|\mathcal{H}| \geq |\mathbb{N}|$

Double sample trick: a second 'ghost sample'

- true error $\leftrightarrow$ empirical error on the 'ghost sample'
- hence reduce to a finite number of behaviours
- make union bound, but bad events grouped together

Symmetrization:

- bound the probability of good performance on one sample but bad performance on the other sample
- swapping examples between actual and ghost sample

Growth function of class $\mathcal{H}$:

- $G_{\mathcal{H}}(m)$ = largest number of dichotomies (± 1 labels) generated by the class $\mathcal{H}$ on any m points.

VC dimension of class $\mathcal{H}$:

- $VC(\mathcal{H})$ = largest m such that $G_{\mathcal{H}}(m) = 2^m$

VC upper bound

Vapnik & Chervonenkis: For any m, for any $\delta \in (0,1)$,

w.p. $\geq 1-\delta$, $\boxed{\forall h \in \mathcal{H}, \quad \Delta(h) \leq \sqrt{\frac{8}{m}\log\left(\frac{4G_{\mathcal{H}}(2m)}{\delta}\right)}}$

growth function

- Bounding the growth function → Sauer's Lemma
- If $d = VC(\mathcal{H})$ finite, then $G_{\mathcal{H}}(m) \leq \sum_{k=0}^{d} \binom{m}{k}$ for all m implies $G_{\mathcal{H}}(m) \leq (em/d)^d$ (polynomial in m)

For $\mathcal{H}$ with $d = VC(\mathcal{H})$ finite, for any m, for any $\delta \in (0,1)$,

w.p. $\geq 1-\delta$, $\boxed{\forall h \in \mathcal{H}, \quad \Delta(h) \leq \sqrt{\frac{8d}{m}\log\left(\frac{2em}{d}\right) + \frac{8}{m}\log\left(\frac{4}{\delta}\right)}}$

VC upper bound:

- Note that the bound is:
  the same for all functions in the class (uniform over $\mathcal{H}$)
  and the same for all distributions (uniform over $\mathbb{P}$)

VC lower bound:

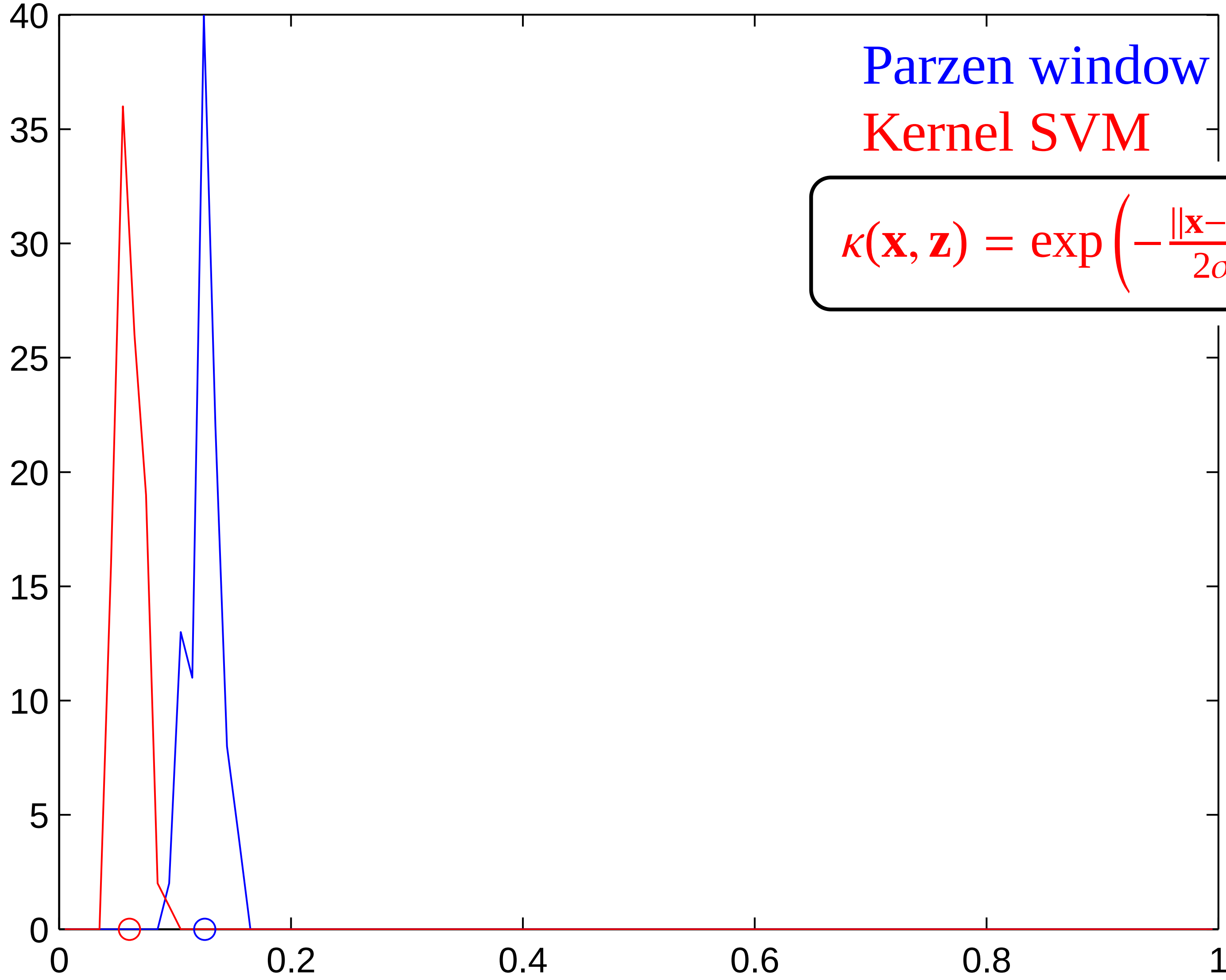
- VC dimension *characterises* learnability in PAC setting:
  there exist distributions such that with large probability over m random examples, the gap between the risk and the best possible risk achievable over the class is at least

$$\sqrt{\frac{d}{m}}.$$

Limitations of the VC framework

- The theory is certainly valid and tight – lower and upper bounds match!
- VC bounds motivate Empirical Risk Minimization (ERM), as apply to a hypothesis space, not hypothesis-dependent
- Practical algorithms often do not search a fixed hypothesis space but regularise to trade complexity with empirical error, e.g. k-NN or SVMs or DNNs
- Mismatch between theory and practice
- Let's illustrate this with SVMs...

SVM with Gaussian kernel



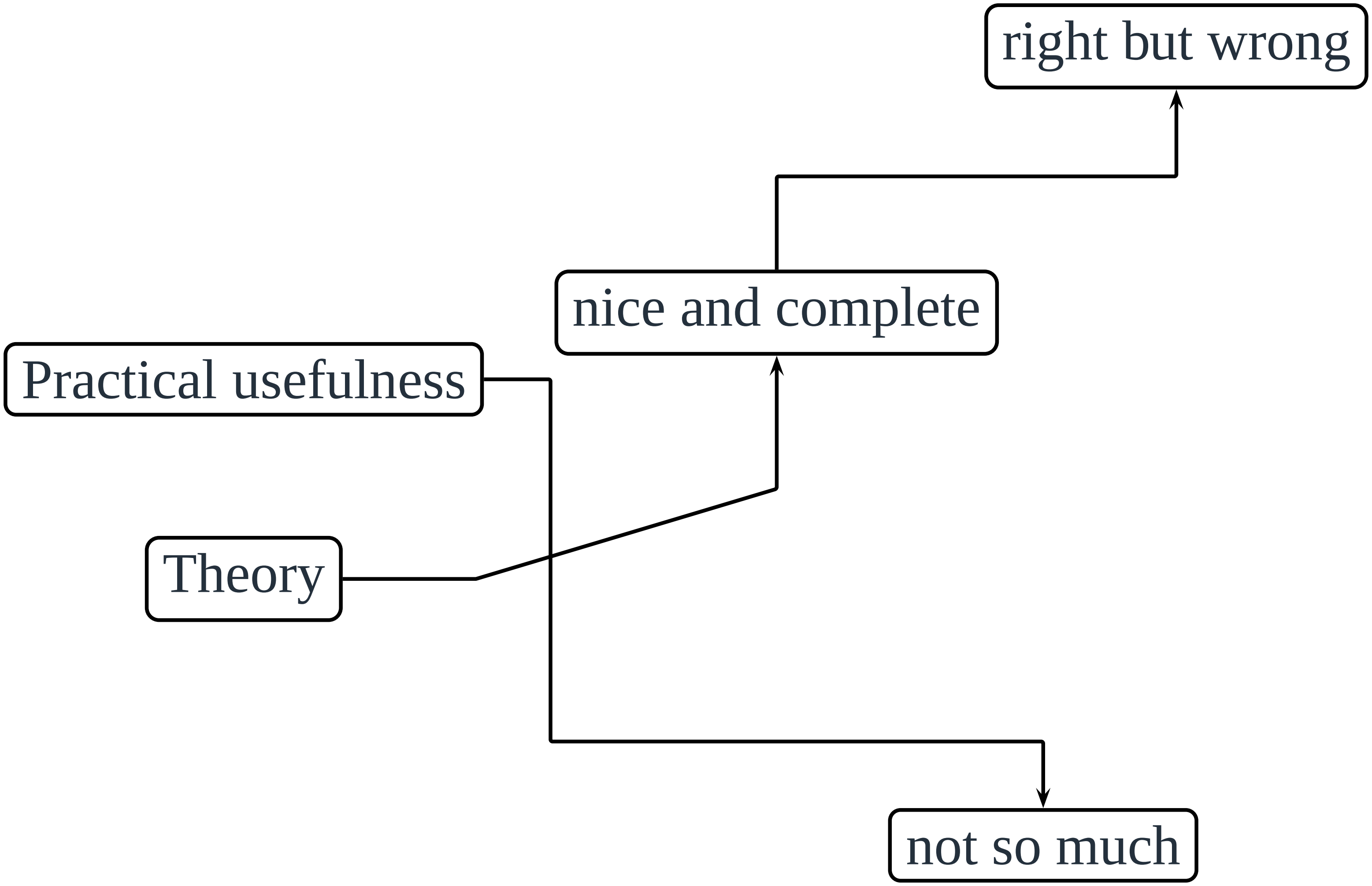
$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right)$$

SVM with Gaussian kernel: A case study

- VC dimension $\longrightarrow$ infinite
- but observed performance is often excellent
- VC bounds aren't able to explain this
- lower bounds appear to contradict the observations
- How to resolve this apparent contradiction?

Coming up...

- large margin ▹ distribution may not be worst-case


right but wrong
nice and complete
Practical usefulness
Theory
not so much

Second generation SLT

Recap and what's coming

We saw...

- SLT bounds the tail of the error distribution
  - giving high confidence bounds on generalization
    - VC gave uniform bounds over a set of classifiers
      - and worst-case over data-generating distributions
        - VC characterizes learnability (for a fixed class)

Coming up...

- exploiting non worst-case distributions
  - bounds that depend on the chosen function
    - new proof techniques
      - approaches for deep learning and future directions

First step towards non-uniform learnability.

$\mathcal{H} = \bigcup_{k\in\mathbb{N}} \mathcal{H}_k$ (countable union), each $d_k = VC(\mathcal{H}_k)$ finite.
Use a weighting scheme: w_k weight of class $\mathcal{H}_k$, $\sum_k w_k \leq 1$.
For each k, $\mathbb{P}^m[\exists f \in \mathcal{H}_k,\ \Delta(f) > \epsilon_k] \leq w_k\delta$, then union bound:

Hence, w.p. $\geq 1-\delta$, $\boxed{\forall k \in \mathbb{N},\ \forall h \in \mathcal{H}_k,\ \ \Delta(h) \leq \epsilon_k}$

Comments:

- First attempt to introduce hypothesis-dependence (i.e. complexity depends on the chosen function)
- The bound leads to a bound-minimizing algorithm:

$$k(h) := \min\{k : h \in \mathcal{H}_k\}, \quad \text{return} \quad \arg\min_{h\in\mathcal{H}} \left\{R_{\text{in}}(h) + \epsilon_{k(h)}\right\}$$

Detecting benign distributions

- SRM detects ‘right’ complexity for the particular problem, but must define the hierarchy a priori
- need to have more nuanced ways to detect how benign a particular distribution is
- SVM uses the margin: appears to detect ‘benign’ distribution in the sense that data unlikely to be near decision boundary $\rightarrow$ easier to classify
- Audibert & Tsybakov: minimax asymptotic rates for the error for class of distributions with reduced margin density
- Marchand and S-T showed how sparsity can also be an indicator of a benign learning problem
- All examples of luckiness framework that shows how SRM can be made data-dependent

Case study: Margin

- Maximising the margin frequently makes it possible to obtain good generalization despite high VC dimension
- The lower bound implies that SVMs must be taking advantage of a benign distribution, since we know that in the worst case generalization will be bad.
- Hence, we require a theory that can give bounds that are sensitive to serendipitous distributions, with the margin an indication of such ‘luckiness’.
- One intuition: if we use real-valued function classes, the margin will give an indication of the accuracy with which we need to approximate the functions

Three proof techniques

We will give an introduction to three proof techniques

- First is motivated by approximation accuracy idea:
  - ▹ Covering Numbers
- Second again uses real value functions but reduces to how well the class can align with random labels:
  - ▹ Rademacher Complexity
- Finally, we introduce an approach inspired by Bayesian inference that maintains distributions over the functions:
  - ▹ PAC-Bayes Analysis

Covering numbers

- As with VC bound use the double-sample trick to reduce the problem to a finite set of points (actual & ghost sample)
- find a set of functions that cover the performances of the function class on that set of points, up to the accuracy of the margin
- In the cover there is a function close to the learned function and because of the margin it will have similar performance on train and test, so can apply symmetrisation
- Apply the union bound over the cover
- Effective complexity is the log of the covering numbers
- This can be bounded by a generalization of the VC dimension, known as the fat-shattering dimension

Starts from considering the uniform (over the class) bound on the gap:

$$\mathbb{P}^m[\forall h \in \mathcal{H},\ \ \Delta(h) \leq \epsilon] = \mathbb{P}^m[\sup_{h\in\mathcal{H}} \Delta(h) \leq \epsilon]$$

Original sample: $S = (Z_1, \ldots, Z_m)$, $\quad \Delta(h) = R_{\mathrm{out}}(h) - R_{\mathrm{in}}(h, S)$

Ghost sample: $S' = (Z'_1, \ldots, Z'_m)$, $\quad R_{\mathrm{out}}(h) = \mathbb{E}^m[R_{\mathrm{in}}(h, S')]$

$$\mathbb{E}^m\Big[\sup_{h\in\mathcal{H}} \Delta(h)\Big] \leq \mathbb{E}^{2m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}[\ell(h, Z'_i) - \ell(h, Z_i)]\right]$$

$$= \mathbb{E}^{2m}\,\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i[\ell(h, Z'_i) - \ell(h, Z_i)]\right]$$

symmetrization

σ_i's i.i.d. symmetric $\{\pm 1\}$-valued

Rademacher r.v.'s

$$\leq 2\mathbb{E}^m\,\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\,\ell(h, Z_i)\right]$$

▹ Rademacher complexity of a class

Generalization bound from RC

Empirical
Rademacher complexity: $\boxed{\mathfrak{R}(\mathcal{H}, S_m) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell(h(X_i), Y_i) \right]}$

Rademacher complexity: $\boxed{\mathfrak{R}(\mathcal{H}) = \mathbb{E}^m [\mathfrak{R}(\mathcal{H}, S_m)]}$

- Symmetrization ▹ $\mathbb{E}^m \left[\sup_{h \in \mathcal{H}} \Delta(h) \right] \leq 2\mathfrak{R}(\mathcal{H})$
- McDiarmid's ineq. (w.p. $\geq 1 - \delta$) ▹ $\sup_{h \in \mathcal{H}} \Delta(h) \leq \mathbb{E}^m \left[\sup_{h \in \mathcal{H}} \Delta(h) \right] + \sqrt{\frac{1}{2m} \log\left(\frac{1}{\delta}\right)}$
- McDiarmid's ineq. (w.p. $\geq 1 - \delta$) ▹ $\mathfrak{R}(\mathcal{H}) \leq \mathfrak{R}(\mathcal{H}, S_m) + \sqrt{\frac{1}{2m} \log\left(\frac{1}{\delta}\right)}$

For any m, for any $\delta \in (0, 1)$,

w.p. $\geq 1 - \delta$, $\boxed{\forall h \in \mathcal{H}, \quad \Delta(h) \leq 2\mathfrak{R}(\mathcal{H}, S_m) + 3\sqrt{\frac{1}{2m} \log(\frac{2}{\delta})}}$

Rademacher Complexity of SVM

- Let $\mathcal{F}(\kappa, B)$ be the class of real-valued functions in a feature space defined by kernel κ with 2-norm of the weight vector $\mathbf{w}$ bounded by B

$$\Re(\mathcal{F}(\kappa, B), S_m) = \frac{B}{m}\sqrt{\sum_{i=1}^{m} \kappa(\mathbf{x}_i, \mathbf{x}_i)}$$

- Hence, control complexity by regularizing with the 2-norm, while keeping outputs at ± 1: gives SVM optimisation with hinge loss to take real valued to classification
- Rademacher complexity controlled as hinge loss is a Lipschitz function
- putting pieces together gives bound that motivates the SVM algorithm with slack variables ξ_i and margin $\gamma = 1/||\mathbf{w}||$

Error bound for SVM

- Upper bound on the generalization error:

$$\frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{m\gamma}\sqrt{\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)} + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

- For the Gaussian kernel this reduces to

$$\frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{\sqrt{m}\gamma} + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

Comments on RC approach

This gives a plug-and-play that we can use to derive bounds based on Rademacher Complexity for other kernel-based (2-norm regularised) algorithms, e.g.

- kernel PCA
- kernel CCA
- one-class SVM
- multiple kernel learning
- regression

Approach can also be used for 1-norm regularised methods as Rademacher complexity is not changed by taking the convex hull of a set of functions, e.g. LASSO and boosting

The PAC-Bayes framework

- Before data, fix a distribution $Q_0 \in M_1(\mathcal{H})$ ▹ ‘prior’
- Based on data, learn a distribution $Q \in M_1(\mathcal{H})$ ▹ ‘posterior’
- Predictions:
  - draw $h \sim Q$ and predict with the chosen h.
  - each prediction with a fresh random draw.

The risk measures $R_{\text{in}}(h)$ and $R_{\text{out}}(h)$ are extended by averaging:

$$\boxed{R_{\text{in}}(Q) \equiv \int_{\mathcal{H}} R_{\text{in}}(h)\, dQ(h)} \qquad \boxed{R_{\text{out}}(Q) \equiv \int_{\mathcal{H}} R_{\text{out}}(h)\, dQ(h)}$$

Typical PAC-Bayes bound:

Fix Q_0. For any sample size m, for any $\delta \in (0,1)$, w.p. $\geq 1-\delta$,

$$\boxed{\forall Q \quad KL(R_{\text{in}}(Q)\|R_{\text{out}}(Q)) \leq \frac{KL(Q\|Q_0) + \log(\frac{m+1}{\delta})}{m}}$$

$$\boxed{W_m = \mathsf{A}_{\mathrm{SVM}}(S_m), \quad \hat{W}_m = W_m / \|W_m\|}$$

For any m, for any $\delta \in (0,1)$,

w.p. $\geq 1 - \delta$, $\boxed{KL(R_{\mathrm{in}}(Q_\mu) \| R_{\mathrm{out}}(Q_\mu)) \leq \frac{\frac{1}{2}\mu^2 + \log(\frac{m+1}{\delta})}{m}}$

Gaussian randomization:

- $Q_0 = \mathcal{N}(0, I)$
- $Q_\mu = \mathcal{N}(\mu \hat{W}_m, I)$

- $KL(Q_\mu \| Q_0) = \frac{1}{2}\mu^2$

$R_{\mathrm{in}}(Q_\mu) = \mathbb{E}^m[\tilde{F}(\mu\gamma(\mathbf{x}, y))]$ where $\tilde{F}(t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \mathrm{d}x$

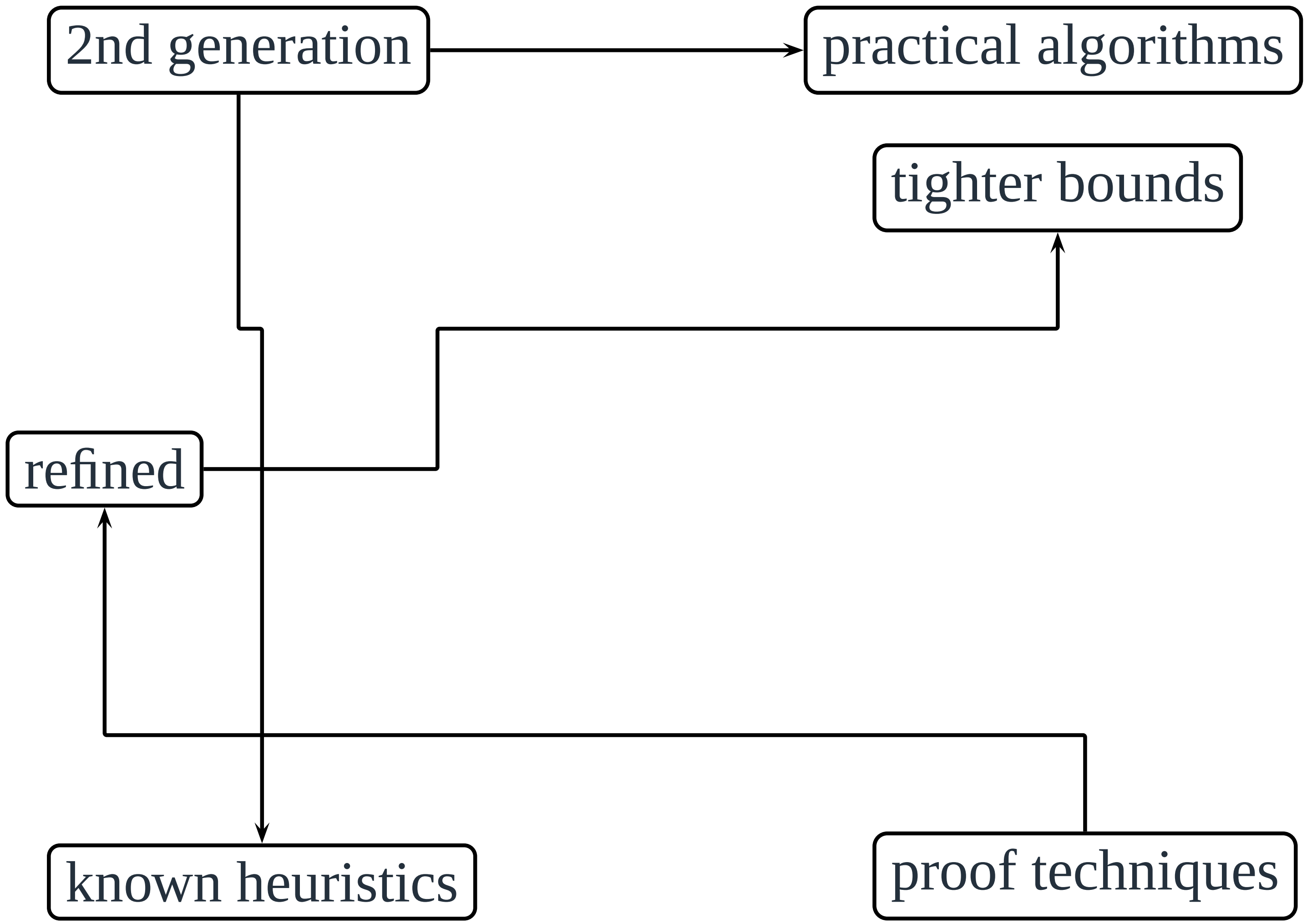
SVM generalization error $\leq 2 \min_\mu R_{\mathrm{out}}(Q_\mu)$

Results

		Classifier					
		SVM				ηPrior SVM	
Problem		2FCV	10FCV	PAC	PrPAC	PrPAC	τ-PrPAC
digits	Bound	–	–	0.175	0.107	0.050	**0.047**
	CE	**0.007**	**0.007**	**0.007**	0.014	0.010	0.009
waveform	Bound	–	–	0.203	0.185	0.178	**0.176**
	CE	0.090	0.086	**0.084**	0.088	0.087	0.086
pima	Bound	–	–	0.424	0.420	0.428	**0.416**
	CE	0.244	0.245	**0.229**	**0.229**	0.233	0.233
ringnorm	Bound	–	–	0.203	0.110	0.053	**0.050**
	CE	**0.016**	**0.016**	0.018	0.018	**0.016**	**0.016**
spam	Bound	–	–	0.254	0.198	0.186	**0.178**
	CE	0.066	**0.063**	0.067	0.077	0.070	0.072

PAC-Bayes bounds vs. Bayesian learning

- **Prior**
  - PAC-Bayes bounds: bounds hold even if prior incorrect
  - Bayesian: inference must assume prior is correct
- **Posterior**
  - PAC-Bayes bounds: bound holds for all posteriors
  - Bayesian: posterior computed by Bayesian inference
- **Data distribution**
  - PAC-Bayes bounds: can be used to define prior, hence no need to be known explicitly: see below
  - Bayesian: input effectively excluded from the analysis: randomness in the noise model generating the output


2nd generation
practical algorithms
tighter bounds
refined
known heuristics
proof techniques

Next generation SLT

Performance of deep NNs

- Deep learning has thrown down a challenge to SLT: very good performance with extremely complex hypothesis classes
- Recall that we can think of the margin as capturing an accuracy with which we need to estimate the weights
- If we have a deep network solution with a wide basin of good performance we can take a similar approach using PAC-Bayes with a broad posterior around the solution
- Dziugaite and Roy have derived useful bounds in this way
- There have also been suggestions that stability of SGD is important in obtaining good generalization
- We present stability approach combining with PAC-Bayes and argue this results in a new learning principle linked to recent analysis of information stored in weights

Uniform hypothesis sensitivity β at sample size m:

$$\| \mathsf{A}(z_{1:m}) - \mathsf{A}(z'_{1:m}) \| \leq \beta \sum_{i=1}^{m} \mathbf{1}[z_i \neq z'_i]$$

$(z_1, \ldots, z_m)$ $(z'_1, \ldots, z'_m)$

- $\mathsf{A}(z_{1:m}) \in \mathcal{H}$ normed space
- $w_m = \mathsf{A}(z_{1:m})$ ‘weight vector’
- Lipschitz
- smoothness

Uniform loss sensitivity β at sample size m:

$$|\ell(\mathsf{A}(z_{1:m}), z) - \ell(\mathsf{A}(z'_{1:m}), z)| \leq \beta \sum_{i=1}^{m} \mathbf{1}[z_i \neq z'_i]$$

- worst-case
- data-insensitive
- distribution-insensitive
- Open: data-dependent?

Generalization from Stability

If A has sensitivity β at sample size m, then for any $\delta \in (0, 1)$,

w.p. $\geq 1 - \delta$, $\boxed{R_{\text{out}}(h) \leq R_{\text{in}}(h) + \epsilon(\beta, m, \delta)}$

(e.g. Bousquet & Elisseeff)

- the intuition is that if individual examples do not affect the loss of an algorithm then it will be concentrated
- can be applied to kernel methods where β is related to the regularisation constant, but bounds are quite weak
- question: algorithm output is highly concentrated $\Longrightarrow$ stronger results?

Distribution-dependent priors

- The idea of using a data distribution defined prior was pioneered by Catoni who looked at these distributions:
- Q_0 and Q are Gibbs-Boltzmann distributions

$$Q_0(h) := \frac{1}{Z'} e^{-\gamma \operatorname{risk}(h)} \qquad Q(h) := \frac{1}{Z} e^{-\gamma \hat{\operatorname{risk}}_S(h)}$$

- These distributions are hard to work with since we cannot apply the bound to a single weight vector, but the bounds can be very tight:

$$KL_+(\hat{Q}_S(\gamma) \| Q_{\mathcal{D}}(\gamma)) \leq \frac{1}{m}\left(\frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{8\sqrt{m}}{\delta}} + \frac{\gamma^2}{4m} + \ln \frac{4\sqrt{m}}{\delta} \right)$$

as it appears we can choose γ small even for complex classes.

If A has uniform hypothesis stability β at sample size n, then for any $\delta \in (0,1)$, w.p. $\geq 1 - 2\delta$,

$$\boxed{KL(R_{\text{in}}(Q) \| R_{\text{out}}(Q)) \leq \frac{\frac{n\beta^2}{2\sigma^2}\left(1 + \sqrt{\frac{1}{2}\log(\frac{1}{\delta})}\right)^2 + \log(\frac{n+1}{\delta})}{n}}$$

Gaussian randomization

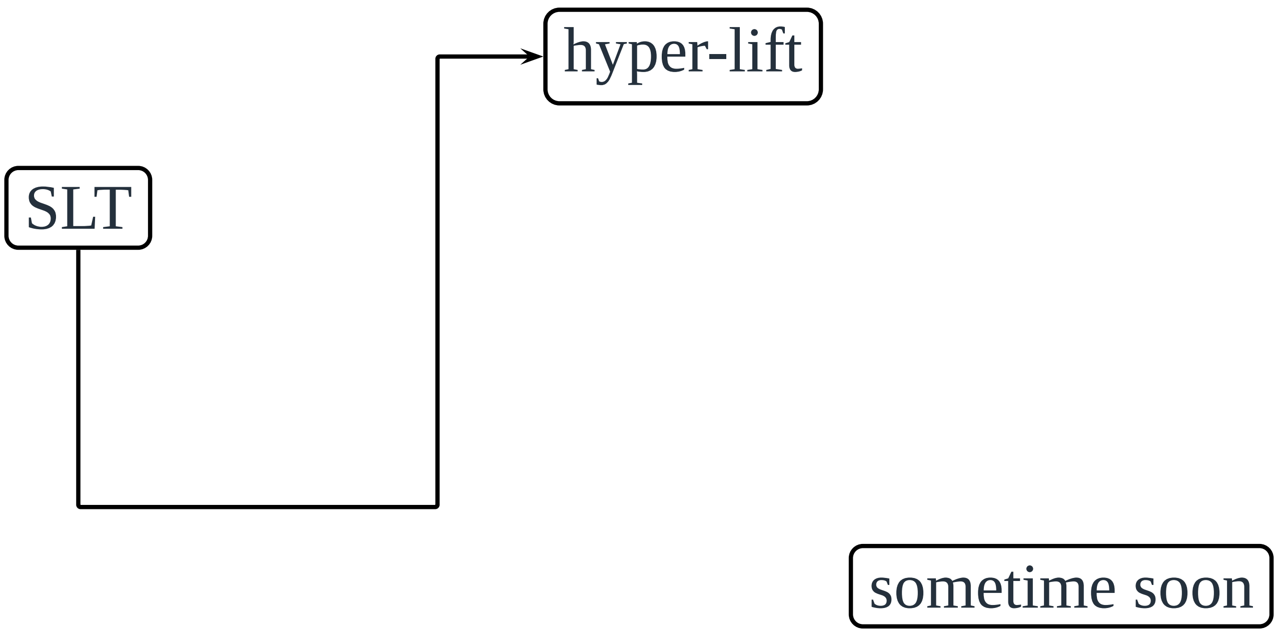
- $Q_0 = \mathcal{N}(\mathbb{E}[W_n], \sigma^2 I)$
- $Q = \mathcal{N}(W_n, \sigma^2 I)$
- $KL(Q \| Q_0) = \frac{1}{2\sigma^2} \|W_n - \mathbb{E}[W_n]\|^2$

Main proof components:

- w.p. $\geq 1 - \delta$, $\quad KL(R_{\text{in}}(Q) \| R_{\text{out}}(Q)) \leq \frac{KL(Q\|Q_0) + \log\left(\frac{n+1}{\delta}\right)}{n}$
- w.p. $\geq 1 - \delta$, $\quad \|W_n - \mathbb{E}[W_n]\| \leq \sqrt{n}\,\beta\left(1 + \sqrt{\frac{1}{2}\log(\frac{1}{\delta})}\right)$

Information about Training Set

- Achille and Soatto studied the amount of information stored in the weights of deep networks
- Overfitting is related to information being stored in the weights that encodes the particular training set, as opposed to the data generating distribution
- This corresponds to reducing the concentration of the distribution of weight vectors output by the algorithm
- They argue that the Information Bottleneck criterion can control this information: hence could potentially lead to a tighter PAC-Bayes bound
- potential for algorithms that optimize the bound


hyper-lift
SLT
sometime soon

Thank you!

Acknowledgements

John gratefully acknowledges support from:

- UK Defence Science and Technology Laboratory (Dstl) Engineering and Physical Research Council (EPSRC). Collaboration between: US DOD, UK MOD, UK EPSRC under the Multidisciplinary University Research Initiative.

Omar gratefully acknowledges support from:

- DeepMind

References

- Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018
- N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM*, 44(4):615–631, 1997
- M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999
- M. Anthony and N. Biggs. *Computational Learning Theory*, volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1992
- Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. https://arxiv.org/abs/math/0507180v3, 2011
- P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998
- P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002
- Shai Ben-David and Shai Shalev-Shwartz. *Understanding Machine Learning: from Theory to Algorithms*. Cambridge University Press, Cambridge, UK, 2014
- Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2008
- O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002
- Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, 2007
- Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems*, 2013
- Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008, 2017
- Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayes risk bounds for general loss functions. In *Proceedings of the 2006 conference on Neural Information Processing Systems (NIPS-06), accepted*, 2006
- Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayes risk bounds for general loss functions. In *Proceedings of the 2006 conference on Neural Information Processing Systems (NIPS-06), accepted*, 2006
- W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Stat. Assoc.*, 58:13–30, 1963
- M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994
- Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2011
- V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, pages 443 – 459, 2000

References

- J. Langford and J. Shawe-Taylor. PAC bayes and margins. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press
- Mario Marchand and John Shawe-Taylor. The set covering machine. *JOURNAL OF MACHINE LEARNING REASEARCH*, 3:2002, 2002
- Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004
- David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), 2003
- David McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2003
- C. McDiarmid. On the method of bounded differences. In 141 London Mathematical Society Lecture Notes Series, editor, *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2018
- Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13(1):3507–3531, December 2012
- T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991
- R. Schapire, Y. Freund, P. Bartlett, and W. Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998. (To appear. An earlier version appeared in: D.H. Fisher, Jr. (ed.), Proceedings ICML97, Morgan Kaufmann.)
- Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July 2001
- Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalization Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003
- John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1998
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004
- John Shawe-Taylor, Christopher K. I. Williams, Nello Cristianini, and Jaz S. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51:2510–2522, 2005
- Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000
- V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998
- V. Vapnik and A. Chervonenkis. Uniform convergence of frequencies of occurence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:915 – 918, 1968
- V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971
- Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002