PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement Learning

Yevgeny Seldin Max Planck Institute for Intelligent Systems and University College London

François Laviolette Université Laval

John Shawe-Taylor University College London

ICML-2012 Tutorial

Outline of the Tutorial

Part I | Yevgeny

- Introduction
- PAC-Bayes-Hoeffding Inequality
- Application in a finite domain (co-clustering)

John

- Application in a continuous domain (SVM)
- Relation between Bayesian learning and PAC-Bayesian analysis
- Learning the prior in PAC-Bayesian bounds

Outline of the Tutorial

Part II François

- A Bit of PAC-Bayesian History
- Localized PAC-Bayesian bounds

Yevgeny

- PAC-Bayesian bounds for unsupervised learning and density estimation
- PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning
- Summary

PAC (Probably Approximately Correct) Learning

Provide guarantees on the expected error (approximately) of prediction rules that hold with high probability (probably) with respect to representativeness of the observed sample.

Some Basic Definitions

$\ell(y, y')$ - loss function

$\mathcal{H}$ - hypothesis space

$h(x)$ - prediction of hypothesis $h \in \mathcal{H}$ on sample x

$L(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$ - expected loss of h

$\hat{L}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(x_i))$ - empirical loss of h

Randomized Classifiers

Let ρ be a distribution over $\mathcal{H}$

Randomized Classifiers

At each round of the game:

1. Pick $h \in \mathcal{H}$ according to $\rho(h)$
2. Observe x
3. Return $h(x)$

Randomized Classifiers

Let ρ be a distribution over $\mathcal{H}$

Randomized Classifiers

At each round of the game:

1. Pick $h \in \mathcal{H}$ according to $\rho(h)$
2. Observe x
3. Return $h(x)$

Loss of ρ

$$
\begin{aligned}
L(\rho) &= \mathbb{E}_{(x,y)\sim\mathcal{D}, h\sim\rho}[\ell(y, h(x))] \\
&= \mathbb{E}_{h\sim\rho}[L(h)] = \langle L, \rho \rangle = \begin{cases} \sum_{h\in\mathcal{H}} L(h)\rho(h), & \text{Discrete } \mathcal{H} \\ \int_{\mathcal{H}} L(h)\rho(h)dh, & \text{Continuous } \mathcal{H} \end{cases}
\end{aligned}
$$

Randomized Classifiers

Let ρ be a distribution over $\mathcal{H}$

Randomized Classifiers

At each round of the game:

1. Pick $h \in \mathcal{H}$ according to $\rho(h)$
2. Observe x
3. Return $h(x)$

Loss of ρ

$$L(\rho) = \mathbb{E}_{(x,y)\sim\mathcal{D},h\sim\rho}[\ell(y,h(x))]$$

$$= \mathbb{E}_{h\sim\rho}[L(h)] = \langle L,\rho\rangle = \begin{cases} \sum_{h\in\mathcal{H}} L(h)\rho(h), & \text{Discrete } \mathcal{H} \\ \int_{\mathcal{H}} L(h)\rho(h)dh, & \text{Continuous } \mathcal{H} \end{cases}$$

$$\hat{L}(\rho) = \mathbb{E}_{h\sim\rho}[\hat{L}(h)] = \langle \hat{L},\rho\rangle$$

KL-divergence

Let ρ and π be two distributions over $\mathcal{H}$

$$\begin{aligned}
\mathrm{KL}(\rho\|\pi) &= \mathbb{E}_\rho\left[\ln\frac{\rho}{\pi}\right] \\
&= \langle\rho, \ln\frac{\rho}{\pi}\rangle = \begin{cases} \sum_h \rho(h)\ln\frac{\rho(h)}{\pi(h)}, & \text{Discrete } \mathcal{H} \\ \int_{\mathcal{H}} \ln\left(\frac{\rho(h)}{\pi(h)}\right)\rho(h)dh, & \text{Continuous } \mathcal{H} \end{cases}
\end{aligned}$$

PAC-Bayes-Hoeffding Inequality

Theorem (Approximate version)

Assume that $\ell(y, y') \in [0,1]$. *Fix a reference distribution* π *over* $\mathcal{H}$. *Then for any* $\delta \in (0,1)$ *with probability greater than* $1-\delta$ *over the sample, for all distributions* ρ *simultaneously:*

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Intuition Behind the Bound

$$\langle L, \rho \rangle \lesssim \langle \hat{L}, \rho \rangle + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

$$\mathrm{KL}(\rho \| \pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle = \underbrace{\langle \ln \frac{1}{\pi}, \rho \rangle}_{\text{Description length}} - \underbrace{\mathrm{H}(\rho)}_{\text{Entropy}}$$

Intuition Behind the Bound

$$\langle L, \rho \rangle \lesssim \langle \hat{L}, \rho \rangle + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

$$\mathrm{KL}(\rho \| \pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle = \underbrace{\langle \ln \frac{1}{\pi}, \rho \rangle}_{\text{Description length}} - \underbrace{\mathrm{H}(\rho)}_{\text{Entropy}}$$

Trade-off

Pick ρ that minimizes the trade-off between:

1. The empirical error $\hat{L}(h)$
2. The complexity (description length, prior belief) $\ln \frac{1}{\pi(h)}$
3. And has maximum entropy

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Difference

1. Explicit high-probability guarantee on the expected performance

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function
4. Different weighting of prior belief $\pi(h)$ vs. evidence $\hat{L}(h)$

Relation and Difference with Bayesian Learning

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit way to incorporate prior information (via $\pi(h)$)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function
4. Different weighting of prior belief $\pi(h)$ vs. evidence $\hat{L}(h)$
5. Holds for *any* distribution ρ (including the Bayes posterior)

Relation and Difference with VC-theory and Rademacher complexities

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit high-probability guarantee on the expected performance
2. Explicit dependence on the loss function

Relation and Difference with VC-theory and Rademacher complexities

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}.$$

Relation

1. Explicit high-probability guarantee on the expected performance
2. Explicit dependence on the loss function

Difference

1. Complexity is defined individually for each h via $\pi(h)$ (rather than "complexity of a hypothesis class")
2. Explicit way to incorporate prior knowledge

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Origin: Donsker and Varadhan (1975) - Variational definition of relative entropy

$$\mathrm{KL}(\rho \| \pi) = \sup_f \Big(\langle f, \rho \rangle - \ln \langle e^f, \pi \rangle \Big)$$

The supremum is achieved by $f(h) = \ln \frac{\rho(h)}{\pi(h)}$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof.

$$\langle f, \rho \rangle = \langle \ln \left(\frac{\rho}{\pi} \frac{\pi}{\rho} e^f \right), \rho \rangle$$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof.

$$\begin{aligned}
\langle f, \rho \rangle &= \langle \ln \left(\frac{\rho}{\pi} \frac{\pi}{\rho} e^f \right), \rho \rangle \\
&= \langle \ln \frac{\rho}{\pi}, \rho \rangle + \langle \ln \left(\frac{\pi}{\rho} e^f \right), \rho \rangle
\end{aligned}$$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof.

$$\begin{aligned}
\langle f, \rho \rangle &= \langle \ln \left(\frac{\rho}{\pi} \frac{\pi}{\rho} e^f \right), \rho \rangle \\
&= \langle \ln \frac{\rho}{\pi}, \rho \rangle + \langle \ln \left(\frac{\pi}{\rho} e^f \right), \rho \rangle \\
&\leq \mathrm{KL}(\rho \| \pi) + \ln \langle \frac{\pi}{\rho} e^f, \rho \rangle \qquad \text{(Definition of } \mathrm{KL} + \text{Jensen)}
\end{aligned}$$

Proof Base: Change of Measure Inequality

Theorem (Change of Measure Inequality)

For any function $f : \mathcal{H} \to \mathbb{R}$ and any pair of distributions ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof.

$$\begin{aligned}
\langle f, \rho \rangle &= \langle \ln \left(\frac{\rho}{\pi} \frac{\pi}{\rho} e^f \right), \rho \rangle \\
&= \langle \ln \frac{\rho}{\pi}, \rho \rangle + \langle \ln \left(\frac{\pi}{\rho} e^f \right), \rho \rangle \\
&\leq \mathrm{KL}(\rho \| \pi) + \ln \langle \frac{\pi}{\rho} e^f, \rho \rangle && \text{(Definition of } \mathrm{KL} + \text{Jensen)} \\
&= \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle
\end{aligned}$$

$\square$

Some More Background

Theorem (Markov's inequality)

Let $Z \geq 0$ be a random variable and $\delta \in (0,1)$. Then with probability greater than $1-\delta$:

$$Z \leq \frac{1}{\delta}\mathbb{E}[Z]$$

Theorem (Hoeffding's inequality)

Let $Z_1, \ldots, Z_n$ be i.i.d. random variables, such that $Z_i \in [0,1]$. Then for any λ:

$$\mathbb{E}\left[e^{\lambda \frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}[Z_i]-Z_i)}\right] \leq e^{\lambda^2/2m}$$

PAC-Bayes-Hoeffding Inequality

Theorem (Approximate version)

Assume that $\ell(y, y') \in [0, 1]$*. Fix a reference distribution* π *over* $\mathcal{H}$*. Then for any* $\delta \in (0, 1)$ *with probability greater than* $1 - \delta$ *for all distributions* ρ *simultaneously:*

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\langle e^{\lambda f}, \pi \rangle \leq \frac{1}{\delta} \mathbb{E}\left[\langle e^{\lambda f}, \pi \rangle \right] \qquad \text{(w.p. } \geq 1 - \delta\text{; Markov)}$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \le \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\begin{aligned} \langle e^{\lambda f}, \pi \rangle &\le \frac{1}{\delta} \mathbb{E}\left[\langle e^{\lambda f}, \pi \rangle \right] && \text{(w.p. } \ge 1-\delta\text{; Markov)} \\ &= \frac{1}{\delta} \langle \mathbb{E}\left[e^{\lambda f} \right], \pi \rangle && \text{(Linearity of } \mathbb{E}\text{; } \pi \text{ is deterministic)} \end{aligned}$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\begin{aligned}
\langle e^{\lambda f}, \pi \rangle &\leq \frac{1}{\delta} \mathbb{E}\left[\langle e^{\lambda f}, \pi \rangle \right] && \text{(w.p. } \geq 1-\delta\text{; Markov)} \\
&= \frac{1}{\delta} \langle \mathbb{E}\left[e^{\lambda f} \right], \pi \rangle && \text{(Linearity of } \mathbb{E}\text{; } \pi \text{ is deterministic)} \\
&\leq \frac{1}{\delta} \langle e^{\lambda^2/2m}, \pi \rangle = \frac{1}{\delta} e^{\lambda^2/2m} && \text{(Hoeffding)}
\end{aligned}$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\langle e^{\lambda f}, \pi \rangle \leq \frac{1}{\delta} e^{\lambda^2/2m}$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\langle e^{\lambda f}, \pi \rangle \leq \frac{1}{\delta} e^{\lambda^2/2m}$$

Step 3: Combine and optimize over λ

$$L(\rho) - \hat{L}(\rho) = \langle L - \hat{L}, \rho \rangle \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda}{2m}$$

Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \to \mathbb{R}$ and any ρ and π:

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Markov + Hoeffding

Take $f(h) = \lambda(L(h) - \hat{L}(h))$

$$\langle e^{\lambda f}, \pi \rangle \leq \frac{1}{\delta} e^{\lambda^2/2m}$$

Step 3: Combine and optimize over λ

$$L(\rho) - \hat{L}(\rho) = \langle L - \hat{L}, \rho \rangle \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda}{2m}$$

Take $\lambda \approx \sqrt{2m(\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta})}$.

PAC-Bayes-Hoeffding Inequality

Theorem (Approximate version)

Assume that $\ell(y, y') \in [0, 1]$*. Fix a reference distribution* π *over* $\mathcal{H}$*. Then for any* $\delta \in (0, 1)$ *with probability greater than* $1 - \delta$ *for all distributions* ρ *simultaneously:*

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$

Further Reading

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 2012. Preprint available on arxiv.

Outline of the Tutorial

Part I Yevgeny

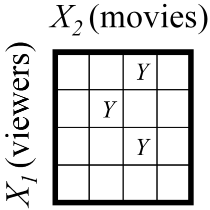
- Introduction
- PAC-Bayes-Hoeffding Inequality
- **Application in a finite domain (co-clustering)**

John

- Application in a continuous domain (SVM)
- Relation between Bayesian learning and PAC-Bayesian analysis
- Learning the prior in PAC-Bayesian bounds

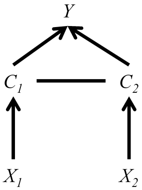
Discriminative Prediction Based on Co-clustering

Example: Collaborative Filtering



Model

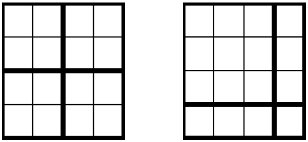
$$\rho(y|x_1, x_2) = \sum_{c_1, c_2} \rho(y|c_1, c_2)\rho(c_1|x_1)\rho(c_2|x_2)$$



PAC-Bayesian Analysis of Co-clustering

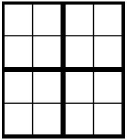
$$\rho(y|x_1, x_2) = \sum_{c_1, c_2} \rho(y|c_1, c_2)\rho(c_1|x_1)\rho(c_2|x_2)$$

- $\mathcal{H}$ - all hard partitions + labels for partition cells
- π - combinatorial (next slide)
- $\rho = \{\rho(c_1|x_1), \rho(c_2|x_2), \rho(y|x_1, x_2)\}$



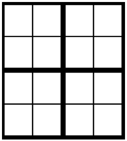
Prior Construction

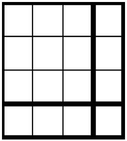
- $|X_i|$ possibilities to choose $|C_i| \quad (i \in \{1,2\})$



Prior Construction

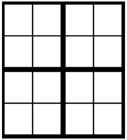
- $|X_i|$ possibilities to choose $|C_i| \quad (i \in \{1,2\})$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose the sizes of the clusters



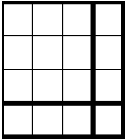


Prior Construction

- $|X_i|$ possibilities to choose $|C_i|$ $\quad (i \in \{1,2\})$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose the sizes of the clusters
- $\binom{|X_i|}{n_1^i,\ldots,n_{|C_i|}^i} \leq e^{|X_i| H(C_i)}$ possibilities to assign x_i-s to c_i-s



$\binom{4}{2} = 6$ balanced partitions



4 unbalanced partitions

Prior Construction

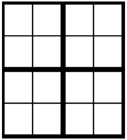
- $|X_i|$ possibilities to choose $|C_i| \quad (i \in \{1,2\})$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose the sizes of the clusters
- $\binom{|X_i|}{n_1^i,\dots,n_{|C_i|}^i} \leq e^{|X_i|H(C_i)}$ possibilities to assign x_i-s to c_i-s
- $|Y|^{|C_1||C_2|}$ possibilities to assign labels to partition cells

$\binom{4}{2} = 6$ balanced partitions   4 unbalanced partitions

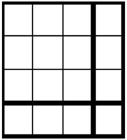
Prior Construction

- $|X_i|$ possibilities to choose $|C_i| \quad (i \in \{1, 2\})$
- $\leq |X_i|^{|C_i|-1}$ possibilities to choose the sizes of the clusters
- $\binom{|X_i|}{n_1^i, \dots, n_{|C_i|}^i} \leq e^{|X_i| H(C_i)}$ possibilities to assign x_i-s to c_i-s
- $|Y|^{|C_1||C_2|}$ possibilities to assign labels to partition cells

$$\pi(h) \geq \exp\left(\sum_{i=1}^{2} \left(-|X_i| \mathrm{H}_h(C_i) - |C_i| \ln |X_i|\right) - |C_1||C_2| \ln |Y|\right)$$



$\binom{4}{2} = 6$ balanced partitions

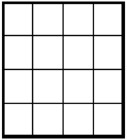


4 unbalanced partitions

Bounding $\mathrm{KL}(\rho\|\pi)$

$$\pi(h) \geq \exp\left(\sum_{i=1}^{2}\left(-|X_i|\mathrm{H}_h(C_i) - |C_i|\ln|X_i|\right) - |C_1||C_2|\ln|Y|\right)$$

$$\rho = \{\rho(c_1|x_1), \rho(c_2|x_2), \rho(y|x_1, x_2)\}$$



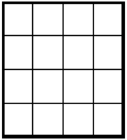
Bounding $\mathrm{KL}(\rho\|\pi)$

$$\pi(h) \geq \exp\left(\sum_{i=1}^{2}\left(-|X_i|\mathrm{H}_h(C_i) - |C_i|\ln|X_i|\right) - |C_1||C_2|\ln|Y|\right)$$

$$\rho = \{\rho(c_1|x_1), \rho(c_2|x_2), \rho(y|x_1, x_2)\}$$

After some calculations...

$$\mathrm{KL}(\rho\|\pi) \leq \sum_{i=1}^{2}\left(|X_i|\mathrm{I}_\rho(X_i;C_i) + |C_i|\ln|X_i|\right) + |C_1||C_2|\ln|Y|$$



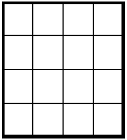
Bounding $\mathrm{KL}(\rho\|\pi)$

$$\pi(h) \geq \exp\left(\sum_{i=1}^{2}\left(-|X_i|\mathrm{H}_h(C_i) - |C_i|\ln|X_i|\right) - |C_1||C_2|\ln|Y|\right)$$

$$\rho = \{\rho(c_1|x_1), \rho(c_2|x_2), \rho(y|x_1, x_2)\}$$

After some calculations...

$$\mathrm{KL}(\rho\|\pi) \leq \sum_{i=1}^{2}\left(|X_i|\mathrm{I}_\rho(X_i; C_i) + |C_i|\ln|X_i|\right) + |C_1||C_2|\ln|Y|$$

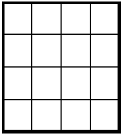


$$\rho(x_i, c_i) = \frac{1}{|X_i|}\rho(c_i|x_i)$$

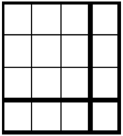
PAC-Bayesian Bound for Co-clustering

With probability $\geq 1 - \delta$, for all ρ:

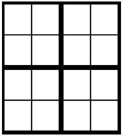
$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\sum_{i=1}^{2} \left(|X_i| \mathrm{I}_\rho(X_i; C_i) + |C_i| \ln |X_i|\right) + |C_1||C_2| \ln |Y| + \ln \frac{1}{\delta} + \nu(\rho)}{2m}}$$



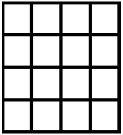
Lowest
Complexity
$\mathrm{I}_\rho(X_i; C_i) = 0$



Lower
Complexity



Higher
Complexity

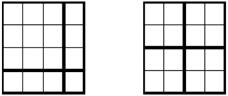


Highest
Complexity
$\mathrm{I}_\rho(X_i; C_i) = \ln |X_i|$

Two Types of Prior Knowledge

With probability $\geq 1-\delta$, for all ρ:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\sum_{i=1}^{2}\left(|X_i|\mathrm{I}_\rho(X_i;C_i) + |C_i|\ln|X_i|\right) + |C_1||C_2|\ln|Y| + \ln\frac{1}{\delta} + \nu(\rho)}{2m}}$$



Structural Prior Knowledge

Exploits symmetries in the hypothesis space

Prior Knowledge about the Distribution

Breaks the structural symmetries

Application: Collaborative Filtering

MovieLens Dataset

- 100,000 ratings on a five-star scale
- 80,000 ratings for training and 20,000 ratings for testing (5-fold)
- 943 viewers; 1680 movies
- State-of-the-art Mean Absolute Error 0.72

Application: Collaborative Filtering

MovieLens Dataset

- 100,000 ratings on a five-star scale
- 80,000 ratings for training and 20,000 ratings for testing (5-fold)
- 943 viewers; 1680 movies
- State-of-the-art Mean Absolute Error 0.72

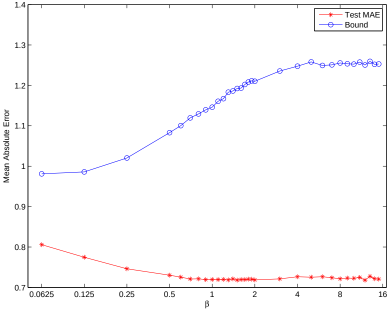
Bound:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\sum_{i=1}^{2} \left(|X_i| \mathrm{I}_\rho(X_i; C_i) + |C_i| \ln |X_i|\right) + |C_1||C_2| \ln |Y| + \ln \frac{1}{\delta} + \nu(\rho)}{2m}}$$

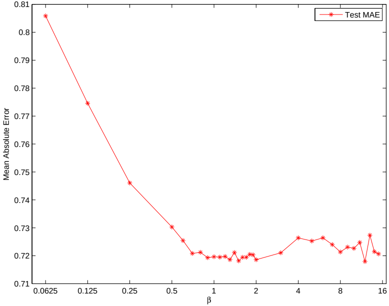
Replace with a trade-off and apply linear search over β

$$\mathcal{F}(\rho, \beta) = \beta m \hat{L}(\rho) + \sum_{i=1}^{2} |X_i| \mathrm{I}_\rho(X_i; C_i)$$

13x6 Clusters

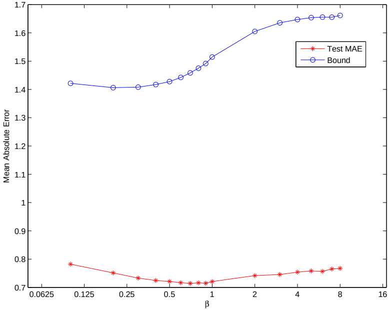


(a) Bound

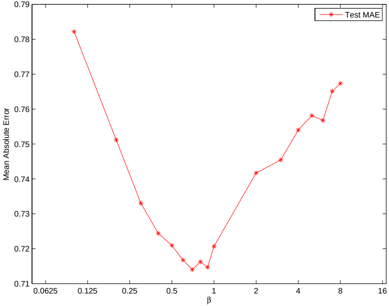


(b) Test Loss (zoom into (a))

50x50 Clusters

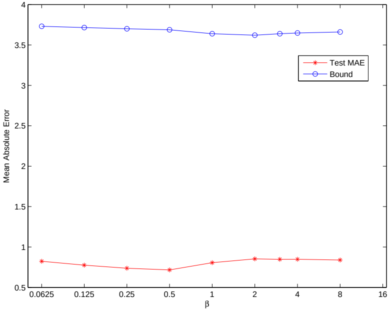


(a) Bound

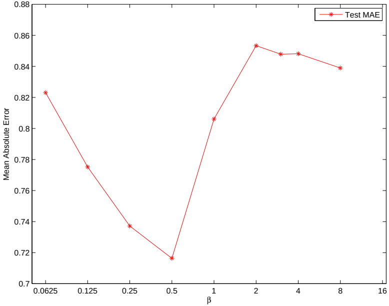


(b) Test Loss (zoom into (a))

283x283 Clusters



(a) Bound



(b) Test Loss (zoom into (a))

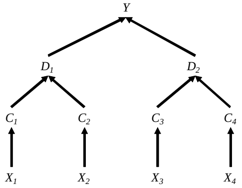
Summary of the Experiments

- The optimal performance is achieved even with 283x283 clusters
- $\frac{1}{\beta}\sum_{i=1}^{2}|X_i|\mathrm{I}_\rho(X_i;C_i)$ has a complete control over the model complexity
- The bound is meaningful, even though not tight

Further Reading

The results can be extended to:

- Matrix tri-factorization $A = LMR$
- Tree-shaped graphical models



Further Reading

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 2010.

Outline of the Tutorial

Part I Yevgeny

- Introduction
- PAC-Bayes-Hoeffding Inequality
- Application in a finite domain (co-clustering)

John

- **Application in a continuous domain (SVM)**
- Relation between Bayesian learning and PAC-Bayesian analysis
- Learning the prior in PAC-Bayesian bounds

Acknowledgements

Many inputs to the presentation, but special thanks to:

- Emilio Parado-Hernandez
- Guy Lever
- Shiliang Sun

Seeger version of the bound

- We consider the 0-1 loss

$$\ell(y, y') = \begin{cases} 0; & \text{if } y = y' \\ 1; & \text{otherwise.} \end{cases}$$

Seeger version of the bound

- We consider the 0-1 loss

$$\ell(y, y') = \begin{cases} 0; & \text{if } y = y' \\ 1; & \text{otherwise.} \end{cases}$$

- Recall

$$\begin{aligned} \langle L, \rho \rangle &= \mathbb{E}_{(x,y)\sim\mathcal{D}, c\sim\rho}[\ell(y, c(x))] = Pr_{(x,y)\sim\mathcal{D}, c\sim\rho}(c(x) \neq y) \\ \langle \hat{L}, \rho \rangle &= Pr_{(x,y)\sim S, c\sim\rho}(c(x) \neq y) \end{aligned}$$

Seeger version of the bound

- We consider the 0-1 loss

$$\ell(y, y') = \begin{cases} 0; & \text{if } y = y' \\ 1; & \text{otherwise.} \end{cases}$$

- Recall

$$\begin{aligned} \langle L, \rho \rangle &= \mathbb{E}_{(x,y)\sim\mathcal{D}, c\sim\rho}[\ell(y, c(x))] = Pr_{(x,y)\sim\mathcal{D}, c\sim\rho}(c(x) \neq y) \\ \langle \hat{L}, \rho \rangle &= Pr_{(x,y)\sim S, c\sim\rho}(c(x) \neq y) \end{aligned}$$

- For classification there is a tighter version that bounds the difference between true and empirical error rates measured by kl, the KL divergence between $\langle \hat{L}, \rho \rangle$ and $\langle L, \rho \rangle$ considered as distributions on $\{0, +1\}$.

Seeger version of the bound

- We consider the 0-1 loss

$$\ell(y, y') = \begin{cases} 0; & \text{if } y = y' \\ 1; & \text{otherwise.} \end{cases}$$

- Recall

$$\begin{aligned} \langle L, \rho \rangle &= \mathbb{E}_{(x,y)\sim\mathcal{D}, c\sim\rho}[\ell(y, c(x))] = Pr_{(x,y)\sim\mathcal{D}, c\sim\rho}(c(x) \neq y) \\ \langle \hat{L}, \rho \rangle &= Pr_{(x,y)\sim S, c\sim\rho}(c(x) \neq y) \end{aligned}$$

- For classification there is a tighter version that bounds the difference between true and empirical error rates measured by kl, the KL divergence between $\langle \hat{L}, \rho \rangle$ and $\langle L, \rho \rangle$ considered as distributions on $\{0, +1\}$.
- Hence,

$$\mathrm{kl}(\hat{p}, p) = \hat{p} \ln \frac{\hat{p}}{p} + (1 - \hat{p}) \ln \frac{1 - \hat{p}}{1 - p}$$

Seeger version of the bound

- We consider the 0-1 loss

$$\ell(y, y') = \begin{cases} 0; & \text{if } y = y' \\ 1; & \text{otherwise.} \end{cases}$$

- Recall

$$\begin{aligned} \langle L, \rho \rangle &= \mathbb{E}_{(x,y)\sim\mathcal{D}, c\sim\rho}[\ell(y, c(x))] = Pr_{(x,y)\sim\mathcal{D}, c\sim\rho}(c(x) \neq y) \\ \langle \hat{L}, \rho \rangle &= Pr_{(x,y)\sim S, c\sim\rho}(c(x) \neq y) \end{aligned}$$

- For classification there is a tighter version that bounds the difference between true and empirical error rates measured by kl, the KL divergence between $\langle \hat{L}, \rho \rangle$ and $\langle L, \rho \rangle$ considered as distributions on $\{0, +1\}$.
- Hence,

$$\mathrm{kl}(\hat{p}, p) = \hat{p} \ln \frac{\hat{p}}{p} + (1 - \hat{p}) \ln \frac{1 - \hat{p}}{1 - p}$$

- Gives a tighter bound than would be obtained using Pinsker's inequality, particularly for small empirical error rates.

PAC-Bayes Theorem (Seeger version)

- Fix an arbitrary $\mathcal{D}$, arbitrary prior π, and confidence δ, then with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, all posteriors ρ satisfy

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

where KL is the KL divergence between distributions

$$\mathrm{KL}(\rho \| \pi) = \mathbb{E}_{c \sim \rho}\left[\ln \frac{\rho(c)}{\pi(c)}\right]$$

Linear classifiers

- We consider linear classifiers in a kernel κ defined feature space:
$$\mathcal{F} = \{c_{\mathbf{w}} : \mathbf{x} \mapsto \operatorname{sgn}\left(\langle \mathbf{w}, \phi(\mathbf{x})\rangle\right)\}$$
where $\langle \phi(\mathbf{x}), \phi(\mathbf{z})\rangle = \kappa(\mathbf{x}, \mathbf{z})$.

Linear classifiers

- We consider linear classifiers in a kernel κ defined feature space:
$$\mathcal{F} = \{c_{\mathbf{w}} : \mathbf{x} \mapsto \mathrm{sgn}\left(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle\right)\}$$
where $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \kappa(\mathbf{x}, \mathbf{z})$.
- The mapping ϕ embeds the input space into a Hilbert space, which is usually specified by the kernel κ satisfying the positive semi-definite property.

Linear classifiers

- We consider linear classifiers in a kernel κ defined feature space:
$$\mathcal{F} = \{c_{\mathbf{w}} : \mathbf{x} \mapsto \operatorname{sgn}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle)\}$$
where $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \kappa(\mathbf{x}, \mathbf{z})$.
- The mapping ϕ embeds the input space into a Hilbert space, which is usually specified by the kernel κ satisfying the positive semi-definite property.
- We will be considering deterministic classifiers such as SVMs, but the bounds will be using stochastic classifiers defined through distributions over $\mathcal{F}$

Linear classifiers

- We consider linear classifiers in a kernel κ defined feature space:
$$\mathcal{F} = \{c_{\mathbf{w}} : \mathbf{x} \mapsto \operatorname{sgn}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle)\}$$
where $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \kappa(\mathbf{x}, \mathbf{z})$.
- The mapping ϕ embeds the input space into a Hilbert space, which is usually specified by the kernel κ satisfying the positive semi-definite property.
- We will be considering deterministic classifiers such as SVMs, but the bounds will be using stochastic classifiers defined through distributions over $\mathcal{F}$
- Note that any threshold must be represented and learnt through inclusion of a constant feature.

Linear classifiers

- We will choose the prior and posterior distributions over $\mathcal{F}$ to be Gaussians with unit variance.

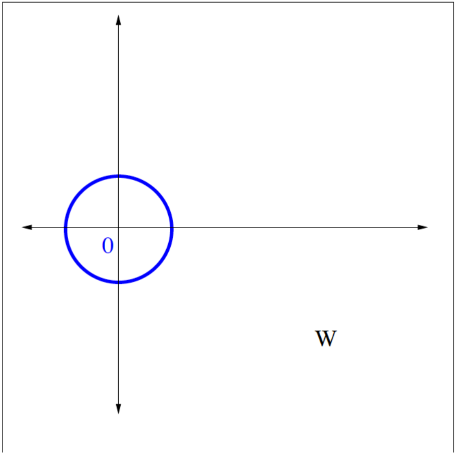
Linear classifiers

- We will choose the prior and posterior distributions over $\mathcal{F}$ to be Gaussians with unit variance.
- The prior π will be centered at the origin with unit variance

Linear classifiers

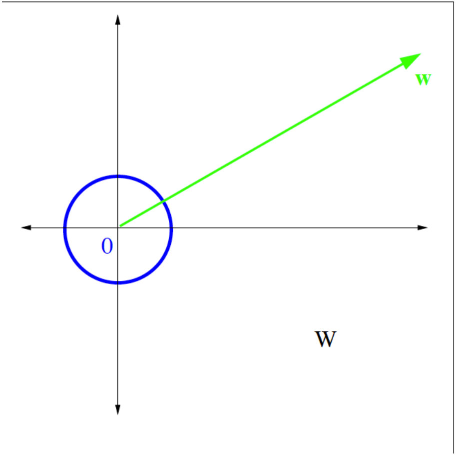
- We will choose the prior and posterior distributions over $\mathcal{F}$ to be Gaussians with unit variance.
- The prior π will be centered at the origin with unit variance
- The specification of the centre for the posterior $\rho(\mathbf{w}, \mu)$ will be by a unit vector $\mathbf{w}$ and a scale factor μ.

PAC-Bayes Bound for SVM



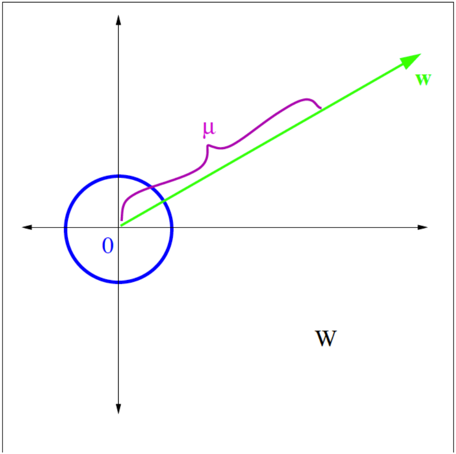
- **Prior** π is Gaussian $\mathcal{N}(0, 1)$
-
-
-

PAC-Bayes Bound for SVM



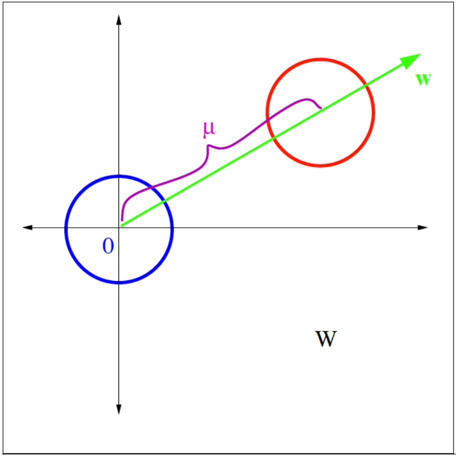
- **Prior** π is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the **direction** $\mathbf{w}$
-
-

PAC-Bayes Bound for SVM



- **Prior** π is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the **direction** $\mathbf{w}$
- at **distance** μ from the origin
-

PAC-Bayes Bound for SVM



- **Prior** π is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the **direction** $\mathbf{w}$
- at **distance** μ from the origin
- **Posterior** ρ is Gaussian

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \boxed{\langle L, \rho(\mathbf{w}, \mu)\rangle}) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}$$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle \| \boxed{\langle L, \rho(\mathbf{w}, \mu) \rangle}) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle L, \rho(\mathbf{w}, \mu) \rangle$ true performance of the stochastic classifier $\mathbb{E}_{c \sim \rho(\mathbf{w}, \mu)}[c(x) \neq y]$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \boxed{\langle L, \rho(\mathbf{w}, \mu)\rangle}) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle L, \rho(\mathbf{w}, \mu)\rangle$ true performance of the stochastic classifier $\mathbb{E}_{c \sim \rho(\mathbf{w}, \mu)}[c(x) \neq y]$
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c \sim \rho(\mathbf{w}, \mu)}[c(x)]\right) \neq y$ as centre of the Gaussian gives the same classification as halfspace with more weight.

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \boxed{\langle L, \rho(\mathbf{w}, \mu)\rangle}) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle L, \rho(\mathbf{w}, \mu)\rangle$ true performance of the stochastic classifier $\mathbb{E}_{c\sim\rho(\mathbf{w},\mu)}[c(x) \neq y]$
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c\sim\rho(\mathbf{w},\mu)}[c(x)]\right) \neq y$ as centre of the Gaussian gives the same classification as halfspace with more weight.
- Hence its error bounded by $2\langle L, \rho(\mathbf{w}, \mu)\rangle$, since if x misclassified at least half of $c \sim \rho$ err.

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}$$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}\left(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle} \,\|\, \langle L, \rho(\mathbf{w}, \mu) \rangle\right) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle$ stochastic measure of the training error

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}\left(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \,\|\, \langle L, \rho(\mathbf{w}, \mu)\rangle\right) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle$ stochastic measure of the training error
- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu\gamma(\mathbf{x}_j, y_j))$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle$ stochastic measure of the training error
- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu\gamma(\mathbf{x}_j, y_j))$
- where $\tilde{F}(\mu\gamma(\mathbf{x}, y))$ is probability of error of stochastic classifier on example $(\mathbf{x}, y)$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle} \| \langle L, \rho(\mathbf{w}, \mu) \rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle$ stochastic measure of the training error
- $\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu \gamma(\mathbf{x}_j, y_j))$
- where $\tilde{F}(\mu \gamma(\mathbf{x}, y))$ is probability of error of stochastic classifier on example $(\mathbf{x}, y)$
- where $\gamma(\mathbf{x}, y) = (y \mathbf{w}^T \phi(\mathbf{x})) / (\|\phi(\mathbf{x})\| \|\mathbf{w}\|)$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle$ stochastic measure of the training error
- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu\gamma(\mathbf{x}_j, y_j))$
- where $\tilde{F}(\mu\gamma(\mathbf{x}, y))$ is probability of error of stochastic classifier on example $(\mathbf{x}, y)$
- where $\gamma(\mathbf{x}, y) = (y\mathbf{w}^T\phi(\mathbf{x}))/(\|\phi(\mathbf{x})\|\|\mathbf{w}\|)$
- and $\tilde{F}(t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \mathsf{d}x$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\boxed{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi)} + \ln \frac{m+1}{\delta}}{m}$$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \langle L, \rho(\mathbf{w},\mu)\rangle) \leq \frac{\boxed{\mathsf{KL}(\rho(\mathbf{w},\mu)\|\pi)} + \ln\frac{m+1}{\delta}}{m}$$

- Prior $\pi \equiv$ Gaussian centered on the origin

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \langle L, \rho(\mathbf{w},\mu)\rangle) \leq \frac{\boxed{\mathsf{KL}(\rho(\mathbf{w},\mu)\|\pi)} + \ln \frac{m+1}{\delta}}{m}$$

- Prior $\pi \equiv$ Gaussian centered on the origin
- Posterior $\rho \equiv$ Gaussian along $\mathbf{w}$ at a distance μ from the origin

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \langle L, \rho(\mathbf{w},\mu)\rangle) \leq \frac{\boxed{\mathsf{KL}(\rho(\mathbf{w},\mu)\|\pi)} + \ln \frac{m+1}{\delta}}{m}$$

- Prior $\pi \equiv$ Gaussian centered on the origin
- Posterior $\rho \equiv$ Gaussian along $\mathbf{w}$ at a distance μ from the origin
- $\mathsf{KL}(\rho\|\pi) = \mu^2/2$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\boxed{\delta}}}{m}$$

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\boxed{\delta}}}{m}$$

- δ is the confidence

PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\mathsf{KL}(\rho(\mathbf{w}, \mu)\|\pi) + \ln \frac{m+1}{\boxed{\delta}}}{m}$$

- δ is the confidence
- The bound holds with probability $1 - \delta$ over the random i.i.d. selection of the training data.

Form of the SVM bound

- Note that bound holds for all posterior distributions so that we can choose μ to optimise the bound

Form of the SVM bound

- Note that bound holds for all posterior distributions so that we can choose μ to optimise the bound
- If we define the inverse of the kl by

$$\mathrm{kl}^{-1}(q, A) = \max\{p : \mathrm{kl}(q\|p) \leq A\}$$

then have with probability at least $1 - \delta$

$$Pr\left(\mathrm{sgn}\left(\langle \mathbf{w}, \phi(\mathbf{x})\rangle\right) \neq y\right) \leq 2 \min_{\mu} \mathrm{kl}^{-1}\left(\frac{1}{m}\sum_{j=1}^{m} \tilde{F}(\mu\gamma(\mathbf{x}_j, y_j)), \frac{\mu^2/2 + \ln \frac{m+1}{\delta}}{m}\right)$$

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)
  - For X-F XV select the pair that minimize the validation error

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)
  - For X-F XV select the pair that minimize the validation error
  - For PAC-Bayes Bound and Prior PAC-Bayes Bound select the pair that minimize the bound

Description of the Datasets

Problem	# samples	input dim.	Pos/Neg
Handwritten-digits	5620	64	2791 / 2829
Waveform	5000	21	1647 / 3353
Pima	768	8	268 / 500
Ringnorm	7400	20	3664 / 3736
Spam	4601	57	1813 / 2788

Table: Description of datasets in terms of number of patterns, number of input variables and number of positive/negative examples.

Results

		Classifier		
		SVM		
Problem		2FCV	10FCV	PAC
digits	Bound	–	–	0.175
	CE	0.007	0.007	0.007
waveform	Bound	–	–	0.203
	CE	0.090	0.086	0.084
pima	Bound	–	–	0.424
	CE	0.244	0.245	0.229
ringnorm	Bound	–	–	0.203
	CE	0.016	0.016	0.018
spam	Bound	–	–	0.254
	CE	0.066	0.063	0.067

Outline of the Tutorial

Part I Yevgeny

- Introduction
- PAC-Bayes-Hoeffding Inequality
- Application in a finite domain (co-clustering)

John

- Application in a continuous domain (SVM)
- **Relation between Bayesian learning and PAC-Bayesian analysis**
- Learning the prior in PAC-Bayesian bounds

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho\rangle \| \langle L, \rho\rangle) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Difference

1. Explicit high-probability guarantee on the expected performance

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function
4. Different weighting of prior belief $\pi(h)$ vs. evidence $\hat{L}(h)$

Relation and Difference with Bayesian Learning

$$\mathrm{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + \ln((m+1)/\delta)}{m}$$

Relation

1. Explicit way to incorporate prior information (via π)

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
3. Explicit dependence on the loss function
4. Different weighting of prior belief $\pi(h)$ vs. evidence $\hat{L}(h)$
5. Holds for *any* distribution ρ (including the Bayes posterior)

Outline of the Tutorial

Part I | Yevgeny

- Introduction
- PAC-Bayes-Hoeffding Inequality
- Application in a finite domain (co-clustering)

John

- Application in a continuous domain (SVM)
- Relation between Bayesian learning and PAC-Bayesian analysis
- **Learning the prior in PAC-Bayesian bounds**

Learning the prior

- Bound depends on the **distance between prior and posterior**

Learning the prior

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**

Learning the prior

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior π with part of the data

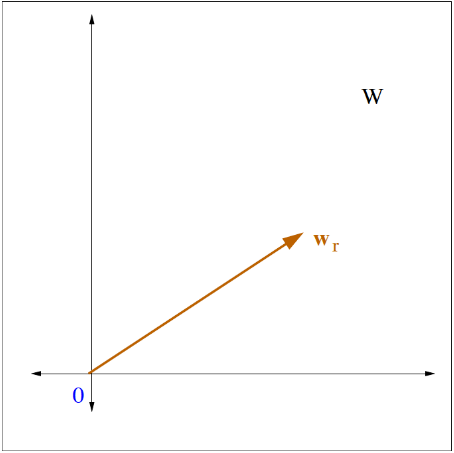
Learning the prior

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior π with part of the data
- Introduce the learnt prior **in the bound**

Learning the prior

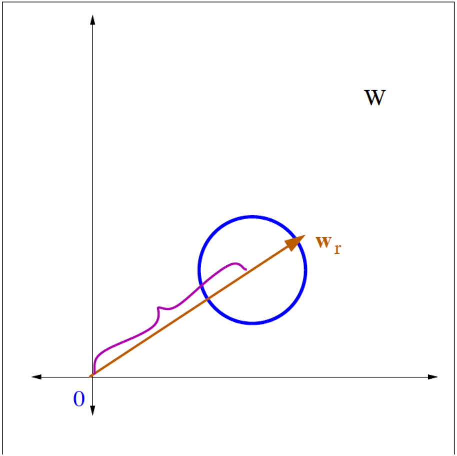
- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior π with part of the data
- Introduce the learnt prior **in the bound**
- Compute stochastic error with **remaining data**

New prior for the SVM



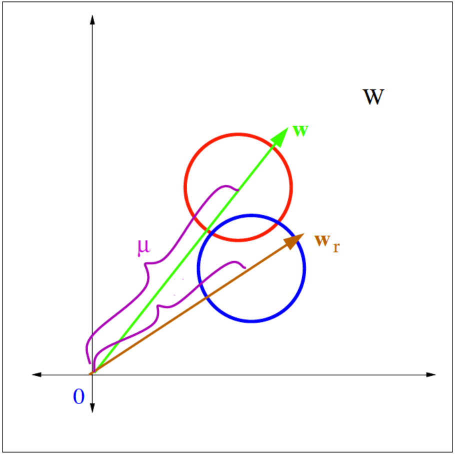
- Solve SVM with **subset of patterns**
-
-
-

New prior for the SVM



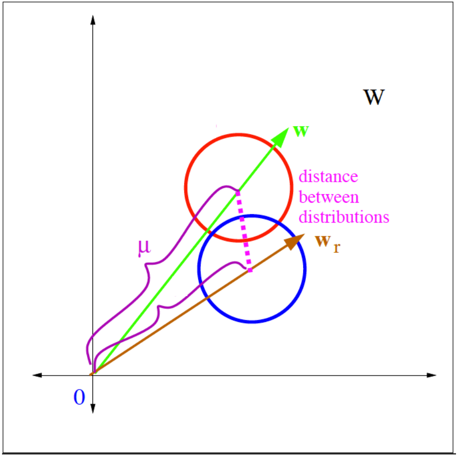
- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
-
-

New prior for the SVM



- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
- **Posterior** like PAC-Bayes Bound
-

New prior for the SVM



- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
- **Posterior** like PAC-Bayes Bound
- **New bound** proportional to $\mathrm{KL}(\rho\|\pi$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \boxed{\langle L, \rho(\mathbf{w},\mu)\rangle}) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \boxed{\langle L, \rho(\mathbf{w}, \mu)\rangle}) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m-r}$$

- $\langle L, \rho(\mathbf{w}, \mu)\rangle$ true performance of the classifier

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\boxed{\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle} \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

- $\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle$ stochastic measure of the training error on remaining data

$$\hat{\rho}(\mathbf{w}, \mu)_S = \frac{1}{m-r}\sum_{j=r+1}^{m} \tilde{F}(\mu\gamma(\mathbf{x}_j, y_j))$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \langle L, \rho(\mathbf{w},\mu)\rangle) \leq \frac{\boxed{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2} + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w},\mu)\rangle \| \langle L, \rho(\mathbf{w},\mu)\rangle) \leq \frac{\boxed{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2} + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

- $0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2$ distance between prior and posterior

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu) \rangle \| \langle L, \rho(\mathbf{w}, \mu) \rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{\boxed{m-r}}$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{\boxed{m-r}}$$

- Penalty term only dependent on the remaining data $m - r$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln \frac{(m-r+1)\boxed{J}}{\delta}}{m - r}$$

New Bound for the SVM

SVM performance may be **tightly** bounded by

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2 + \ln \frac{(m-r+1)\boxed{J}}{\delta}}{m-r}$$

- Must apply the bound for each of J different priors.

Results

		Classifier			
		SVM			
Problem		2FCV	10FCV	PAC	PrPAC
digits	Bound	–	–	0.175	0.107
	CE	0.007	0.007	0.007	0.014
waveform	Bound	–	–	0.203	0.185
	CE	0.090	0.086	0.084	0.088
pima	Bound	–	–	0.424	0.420
	CE	0.244	0.245	0.229	0.229
ringnorm	Bound	–	–	0.203	0.110
	CE	0.016	0.016	0.018	0.018
spam	Bound	–	–	0.254	0.198
	CE	0.066	0.063	0.067	0.077

Prior-SVM

- New bound proportional to $\|\mu \mathbf{w} - \eta \mathbf{w}_r\|^2$

Prior-SVM

- New bound proportional to $\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2$
- Classifier that **optimises the bound**

Prior-SVM

- New bound proportional to $\|\mu \mathbf{w} - \eta \mathbf{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\mathbf{w},\xi_i} \left[\frac{1}{2}\|\mathbf{w} - \mathbf{w}_r\|^2 + C \sum_{i=r+1}^{m} \xi_i \right]$$

$$\begin{aligned} \text{s.t. } & y_i \mathbf{w}^T \phi(\mathbf{x}_i) \geq 1 - \xi_i & i = r+1, \ldots, m \\ & \xi_i \geq 0 & i = r+1, \ldots, m \end{aligned}$$

Prior-SVM

- New bound proportional to $\|\mu\mathbf{w} - \eta\mathbf{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\mathbf{w},\xi_i} \left[\frac{1}{2}\|\mathbf{w} - \mathbf{w}_r\|^2 + C\sum_{i=r+1}^{m} \xi_i\right]$$

$$\text{s.t.}\quad y_i\mathbf{w}^T\phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = r+1,\ldots,m$$

$$\xi_i \geq 0 \qquad i = r+1,\ldots,m$$

- The p-SVM is only solved with the **remaining points**

Bound for p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\mathbf{w}_r$

Bound for p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\mathbf{w}_r$
2. Solve **p-SVM** and obtain $\mathbf{w}$

Bound for p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\mathbf{w}_r$
2. Solve **p-SVM** and obtain $\mathbf{w}$
3. **Margin** for the stochastic classifier $c \sim \rho$

$$\gamma(\mathbf{x}_j, y_j) = \frac{y_j \mathbf{w}^T \phi(\mathbf{x}_j)}{\|\phi(\mathbf{x}_j)\| \|\mathbf{w}\|} \qquad j = r+1, \ldots, m$$

Bound for p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\mathbf{w}_r$
2. Solve **p-SVM** and obtain $\mathbf{w}$
3. **Margin** for the stochastic classifier $c \sim \rho$

$$\gamma(\mathbf{x}_j, y_j) = \frac{y_j \mathbf{w}^T \phi(\mathbf{x}_j)}{\|\phi(\mathbf{x}_j)\| \|\mathbf{w}\|} \qquad j = r+1, \ldots, m$$

4. **Linear search** to obtain the optimal value of μ. This introduces an insignificant extra penalty term

η-Prior-SVM

- Consider using a prior distribution π that is elongated in the direction of $\mathbf{w}_r$

η-Prior-SVM

- Consider using a prior distribution π that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction

η-Prior-SVM

- Consider using a prior distribution π that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2} \|\mathbf{v}\|^2 + C \sum_{i=r+1}^{m} \xi_i \right]$$

η-Prior-SVM

- Consider using a prior distribution π that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C \sum_{i=r+1}^{m} \xi_i \right]$$

- subject to

$$\begin{aligned} y_i(\mathbf{v} + \eta \mathbf{w}_r)^T \phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad & i = r+1, \ldots, m \\ \xi_i \geq 0 \qquad & i = r+1, \ldots, m \end{aligned}$$

Bound for η-prior-SVM

- Prior is elongated along the line of $\mathbf{w}_r$ but spherical with variance 1 in other directions

Bound for η-prior-SVM

- Prior is elongated along the line of $\mathbf{w}_r$ but spherical with variance 1 in other directions
- Posterior again on the line of $\mathbf{w}$ at a distance μ chosen to optimise the bound.

Bound for η-prior-SVM

- Prior is elongated along the line of $\mathbf{w}_r$ but spherical with variance 1 in other directions
- Posterior again on the line of $\mathbf{w}$ at a distance μ chosen to optimise the bound.
- Resulting bound depends on a benign parameter τ determining the variance in the direction $\mathbf{w}_r$

$$\mathrm{kl}(\langle \hat{L}_{S_{m-r}}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{0.5(\ln(\tau^2) + \tau^{-2} - 1 + P^{\|}_{\mathbf{w}_r}(\mu\mathbf{w} - \mathbf{w}_r)^2/\tau^2 + P^{\perp}_{\mathbf{w}_r}(\mu\mathbf{w})^2) + \ln(\frac{m-r+1}{\delta})}{m-r}$$

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

Data distribution dependent prior

- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_\pi = \mathbb{E}[y\phi(\mathbf{x})]$$

Data distribution dependent prior

- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_\pi = \mathbb{E}[y\phi(\mathbf{x})]$$

- Note that we do not know this vector, but it is nonetheless fixed independently of the training sample.

Data distribution dependent prior

- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_\pi = \mathbb{E}[y\phi(\mathbf{x})]$$

- Note that we do not know this vector, but it is nonetheless fixed independently of the training sample.
- We can compute a sample based estimate of this vector as

$$\hat{\mathbf{w}}_\pi = \mathbb{E}_S[y\phi(\mathbf{x})] = \frac{1}{m}\sum_{i=1}^{m} y_i\phi(\mathbf{x}_i)$$

Estimating the KL divergence

- KL divergence is simple have the squared distance.

Estimating the KL divergence

- KL divergence is simple have the squared distance.
- With probability $1 - \delta/2$ we have

$$\|\hat{\mathbf{w}}_\pi - \mathbf{w}_\pi\| \leq \frac{R}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right).$$

Estimating the KL divergence

- KL divergence is simple have the squared distance.
- With probability $1 - \delta/2$ we have

$$\|\hat{\mathbf{w}}_\pi - \mathbf{w}_\pi\| \leq \frac{R}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right).$$

- We can therefore w.h.p. upper bound KL divergence between prior π, an isotropic Gaussian at $\mathbf{w}_\pi$, and posterior ρ, an isotropic Gaussian at $\mathbf{w}$ by

$$\frac{1}{2} \left(\|\mathbf{w} - \hat{\mathbf{w}}_\pi\| + \frac{R}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right) \right)^2$$

Resulting bound

- Giving the following bound on generalisation:

$$\mathrm{kl}(\langle \hat{L}, \rho(\mathbf{w}, \mu)\rangle \| \langle L, \rho(\mathbf{w}, \mu)\rangle) \leq \frac{\frac{1}{2}\left(\|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_\pi\| + \eta\frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right)\right)^2 + \ln\frac{2(m+1)}{\delta}}{m}$$

  with probability $1 - \delta$.

- Values of the bounds for an SVM.

Prob.	PAC-Bayes	PrPAC	τ-PrPAC	$\mathbb{E}$ PrPAC	τ-$\mathbb{E}$ PrPAC
han	0.175	**0.107**	**0.108**	0.157	0.176
wav	0.203	**0.185**	**0.184**	0.202	0.205
pim	0.424	0.420	0.423	0.428	0.433
rin	0.203	**0.110**	**0.110**	0.201	0.204
spa	0.254	**0.198**	**0.198**	0.249	0.255

PAC-Bayesian Analysis in Supervised, Unsupervised, and Reinforcement Learning
Part II

Yevgeny Seldin Max Planck Institute for Intelligent Systems and University College London

François Laviolette Université Laval

John Shawe-Taylor University College London

ICML-2012 Tutorial

Outline of the Tutorial

Part II François

- **A bit of PAC-Bayesian history**
- Localized PAC-Bayesian bounds

Yevgeny

- PAC-Bayesian bounds for unsupervised learning and density estimation
- PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning
- Summary

Definitions often related to PAC-Bayes bound in supervised learning

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn iid acc. to D.

Definitions often related to PAC-Bayes bound in supervised learning

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn iid acc. to D.
- The (true) risk $R(h)$ and training error $R_S(h)$ are defined as:

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad ; \quad R_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq y_i) \, .$$

where $I(y' \neq y)$ is the so called $0 - 1$ loss.

Definitions often related to PAC-Bayes bound in supervised learning

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn iid acc. to D.
- The (true) risk $R(h)$ and training error $R_S(h)$ are defined as:

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad ; \quad R_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq y_i)\,.$$

  where $I(y' \neq y)$ is the so called $0-1$ loss.
- The learner's goal is to choose a **posterior distribution** ρ on a space $\mathcal{H}$ of hypothesis such that the risk of the ρ-weighted **majority vote** B_ρ is as small as possible.

Definitions often related to PAC-Bayes bound in supervised learning

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn iid acc. to D.
- The (true) risk $R(h)$ and training error $R_S(h)$ are defined as:

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad ; \quad R_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq y_i)\,.$$

  where $I(y' \neq y)$ is the so called $0-1$ loss.
- The learner's goal is to choose a **posterior distribution** ρ on a space $\mathcal{H}$ of hypothesis such that the risk of the ρ-weighted **majority vote** B_ρ is as small as possible.

$$B_\rho(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right]$$

Definitions often related to PAC-Bayes bound in supervised learning

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn iid acc. to D.
- The (true) risk $R(h)$ and training error $R_S(h)$ are defined as:

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad ; \quad R_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq y_i)\,.$$

  where $I(y' \neq y)$ is the so called $0-1$ loss.
- The learner's goal is to choose a **posterior distribution** ρ on a space $\mathcal{H}$ of hypothesis such that the risk of the ρ-weighted **majority vote** B_ρ is as small as possible.

$$B_\rho(\mathbf{x}) \;\stackrel{\text{def}}{=}\; \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right]$$

- B_ρ is also called the *Bayes classifier*.

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_ρ

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_ρ
- It bounds the risk of the **Gibbs classifier** G_ρ:

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_ρ
- It bounds the risk of the **Gibbs classifier** G_ρ:
  - to predict the label of $\mathbf{x}$, G_ρ draws h from $\mathcal{H}$ according to ρ, and predicts $h(\mathbf{x})$

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_ρ
- It bounds the risk of the **Gibbs classifier** G_ρ:
  - to predict the label of $\mathbf{x}$, G_ρ draws h from $\mathcal{H}$ according to ρ, and predicts $h(\mathbf{x})$

- The risk and the training error of G_ρ are thus defined as:

$$R(G_\rho) \ = \ \mathop{\mathbf{E}}_{h\sim\rho} R(h) \quad ; \quad R_S(G_\rho) \ = \ \mathop{\mathbf{E}}_{h\sim\rho} R_S(h)\,.$$

G_ρ, B_ρ, and $\mathrm{KL}(\rho \| \pi)$

- If B_ρ misclassifies $\mathbf{x}$, then at least half of the hypothesis (under measure ρ) err on $\mathbf{x}$.

G_ρ, B_ρ, and $\mathrm{KL}(\rho\|\pi)$

- If B_ρ misclassifies $\mathbf{x}$, then at least half of the hypothesis (under measure ρ) err on $\mathbf{x}$.
  - Hence: $R(B_\rho) \leq 2R(G_\rho)$

G_ρ, B_ρ, and $\mathrm{KL}(\rho \| \pi)$

- If B_ρ misclassifies $\mathbf{x}$, then at least half of the hypothesis (under measure ρ) err on $\mathbf{x}$.
  - Hence: $R(B_\rho) \leq 2R(G_\rho)$
  - **Thus, an upper bound on $2R(G_\rho)$ gives rise to an upper bound on $R(B_\rho)$**

History

- **Pre-pre-history: Variational Definition of KL-divergence**
  Donsker and Varadhan (1975)

History

- **Pre-pre-history: Variational Definition of KL-divergence**
  Donsker and Varadhan (1975)

$$\mathbb{E}_\rho[\Phi] \;\leq\; \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_\pi[e^\Phi]$$

or in the context of this tutorial:

$$\langle f, \rho\rangle \leq \mathrm{KL}(\rho\|\pi) + \ln\langle e^f, \pi\rangle$$

History

- **Pre-pre-history: Variational Definition of KL-divergence**
  Donsker and Varadhan (1975)

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*

McAllester Bound

For any D, any $\mathcal{H}$, any π of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\forall\,\rho \text{ on } \mathcal{H}\colon\ \frac{1}{2}(R_S(G_\rho)-R(G_\rho))^2 \le \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]\right) \ge 1-\delta$$

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*

McAllester Bound

For any D, any $\mathcal{H}$, any π of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\forall\,\rho \text{ on } \mathcal{H}\colon\ \frac{1}{2}(R_S(G_\rho)-R(G_\rho))^2 \le \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]\right) \ge 1-\delta$$

or

$$\Pr_{S\sim D^m}\left(\forall\,\rho \text{ on } \mathcal{H}\colon\ R(G_\rho) \le R_S(G_\rho)+\sqrt{\frac{\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]}{2m}}\right) \ge 1-\delta\,,$$

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*

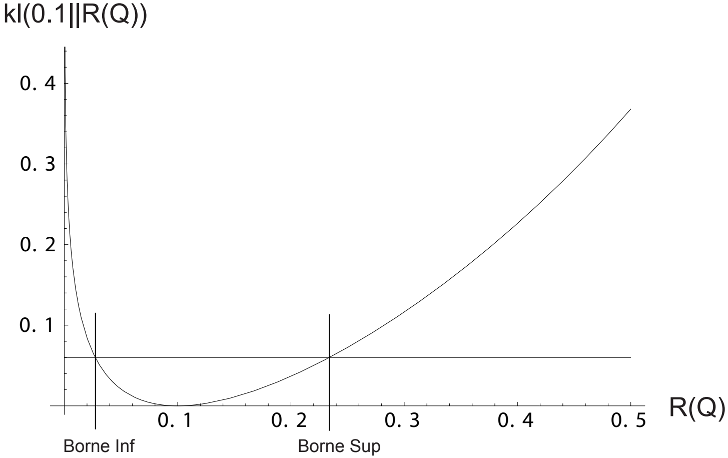
Seeger Bound

For any D, any $\mathcal{H}$, any π of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,\rho \text{ on } \mathcal{H}: \\ \mathrm{kl}(R_S(G_\rho)\|R(G_\rho)) \le \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]\end{array}\right) \ge 1-\delta\,,$$

where $\quad \mathrm{kl}(q\|p) \overset{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}.$

Graphical illustration of the Seeger bound



History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011)*

    This allows applications to ranking, U-statistic of higher order, bandit,...

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003);* **?**; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003);* **?**; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011)*
  - **sample compression setting** *Laviolette and Marchand (2005); Germain et al. (2011)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003);* **?**; ...
  - **Theory** *Catoni (2007); Audibert and Bousquet (2007a); Meir and Zhang (2003)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011)*
  - **sample compression setting** *Laviolette and Marchand (2005); Germain et al. (2011)*

PAC-Bayes and the sample compression setting

This is an important setting.

As example, in its dual version, the SVM can be viewed as a Bayes classifier of the form

$$B_{\mathbf{w}}(\mathbf{x}) \;=\; \mathrm{sgn}\left[\mathop{\mathbf{E}}_{i\sim\mathbf{w}} k(\mathbf{x}_i, \mathbf{x})\right]$$

the hypothesis being here $h_i(\cdot) = k(\mathbf{x}_i, \cdot)$.

PAC-Bayes and the sample compression setting

This is an important setting.

As example, in its dual version, the SVM can be viewed as a Bayes classifier of the form

$$B_{\mathbf{w}}(\mathbf{x}) \;=\; \operatorname{sgn}\left[\mathop{\mathbf{E}}_{i\sim\mathbf{w}} k(\mathbf{x}_i, \mathbf{x})\right]$$

the hypothesis being here $h_i(\cdot) = k(\mathbf{x}_i, \cdot)$.

Problem:

- Recall once more that the prior is not allowed to depend on the training set.

PAC-Bayes and the sample compression setting

This is an important setting.

As example, in its dual version, the SVM can be viewed as a Bayes classifier of the form

$$B_{\mathbf{w}}(\mathbf{x}) \;=\; \mathrm{sgn}\left[\mathop{\mathbf{E}}_{i\sim\mathbf{w}} k(\mathbf{x}_i,\mathbf{x})\right]$$

the hypothesis being here $h_i(\cdot) = k(\mathbf{x}_i,\cdot)$.

Problem:

- Recall once more that the prior is not allowed to depend on the training set.
- How a prior on a set of hypothesis construct from the data can be data-independent ?

PAC-Bayes and the sample compression setting

This is an important setting.

As example, in its dual version, the SVM can be viewed as a Bayes classifier of the form

$$B_{\mathbf{w}}(\mathbf{x}) \;=\; \mathrm{sgn}\left[\mathop{\mathbf{E}}_{i\sim\mathbf{w}} k(\mathbf{x}_i, \mathbf{x})\right]$$

the hypothesis being here $h_i(\cdot) = k(\mathbf{x}_i, \cdot)$.

Problem:

- Recall once more that the prior is not allowed to depend on the training set.
- How a prior on a set of hypothesis construct from the data can be data-independent ?

- The trick : put a prior on the possible ways that hypothesis can be constructed when given the data

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*
- **Martingales & reinforcement learning** *Seldin et al. (2011, 2012)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*
- **Martingales & reinforcement learning** *Seldin et al. (2011, 2012)*

- **Sincere apologizes to everybody we could not fit on the slide...**

Algorithms derived from PAC-Bayes Bound

When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior ρ that minimizes the bound.

Algorithms derived from PAC-Bayes Bound

When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior ρ that minimizes the bound.

Catoni's bound

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,\rho \text{ on } \mathcal{H}: \\ R(G_\rho) \leq \frac{1}{1-e^{-C}}\Big\{1-\exp\Big[-\big(C\cdot R_S(G_\rho) \\ \qquad\qquad\qquad + \frac{1}{m}\big[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\big]\big)\Big]\Big\}\end{array}\right) \quad \geq \quad 1 - \delta.$$

Algorithms derived from PAC-Bayes Bound

When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior ρ that minimizes the bound.

Catoni's bound

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,\rho \text{ on } \mathcal{H}: \\ R(G_\rho) \le \frac{1}{1-e^{-C}}\Big\{1-\exp\Big[-\big(C\cdot R_S(G_\rho) \\ \qquad\qquad + \frac{1}{m}\big[\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}\big]\big)\Big]\Big\}\end{array}\right) \quad\ge\quad 1 - \delta.$$

Interestingly, minimizing the Catoni's bound (when prior and posterior are restricted to Gaussian) give rise to the SVM !

Algorithms derived from PAC-Bayes Bound

When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior ρ that minimizes the bound.

Catoni's bound

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,\rho \text{ on } \mathcal{H}: \\ R(G_\rho) \le \frac{1}{1-e^{-C}}\Big\{1-\exp\Big[-\big(C\cdot R_S(G_\rho) \\ \qquad\qquad + \frac{1}{m}\big[\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}\big]\big)\Big]\Big\}\end{array}\right) \;\ge\; 1-\delta.$$

Interestingly, minimizing the Catoni's bound (when prior and posterior are restricted to Gaussian) give rise to the SVM !
In fact to an SVM where the Hinge loss is replaced by the sigmoid loss.

Algorithms derived from PAC-Bayes Bound (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

Algorithms derived from PAC-Bayes Bound (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost** *Germain et al. (2009b)*

Algorithms derived from PAC-Bayes Bound (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost** *Germain et al. (2009b)*
- **Kernel Ridge Regression** *Germain et al. (2011)*

Algorithms derived from PAC-Bayes Bound (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost** *Germain et al. (2009b)*
- **Kernel Ridge Regression** *Germain et al. (2011)*
- **the proposed structured output algorithm of Cortes et al. (2007)** *Unpublished work of Giguère et al. (2012)*

Algorithms derived from PAC-Bayes Bound (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost** *Germain et al. (2009b)*
- **Kernel Ridge Regression** *Germain et al. (2011)*
- **the proposed structured output algorithm of Cortes et al. (2007)** *Unpublished work of Giguère et al. (2012)*

New algorithms have been found: *Ambroladze et al. (2007); Shawe-Taylor and Hardoon (2009); Germain et al. (2011); Laviolette et al. (2011)*, . . .

Outline of the Tutorial

Part II François

- A bit of PAC-Bayesian history
- **Localized PAC-Bayesian bounds**

Yevgeny

- PAC-Bayesian bounds for unsupervised learning and density estimation
- PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning
- Summary

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{\xi(m)}{\delta}}{2m}}.$$

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution ρ.

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \le \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution ρ.
- A tradeoff between "empirical accuracy" and "complexity"; the complexity being quantify by how far a posterior distributions is from our prior knowledge.

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution ρ.
- A tradeoff between "empirical accuracy" and "complexity"; the complexity being quantify by how far a posterior distributions is from our prior knowledge.
- Thus, some "luckiness argument" is involved here.

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \le \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution ρ.
- A tradeoff between "empirical accuracy" and "complexity"; the complexity being quantify by how far a posterior distributions is from our prior knowledge.
- Thus, some "luckiness argument" is involved here.
  This can be good, but one might want to have some guarantees that, even in unlucky situations, the bound does not degrade over some level.

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution ρ.
- A tradeoff between "empirical accuracy" and "complexity"; the complexity being quantify by how far a posterior distributions is from our prior knowledge.
- Thus, some "luckiness argument" is involved here.
  This can be good, but one might want to have some guarantees that, even in unlucky situations, the bound does not degrade over some level.
  (In general the KL-divergence can be very large ... even infinite)

Localized PAC-Bayesian bounds : a way to reduce the KL-complexity term

- If something can be done to ensure that the bound remains under control it has to be based on the choice of the prior.

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{\xi(m)}{\delta}}{2m}}.$$

Localized PAC-Bayesian bounds : a way to reduce the KL-complexity term

- If something can be done to ensure that the bound remains under control it has to be based on the choice of the prior.

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \ln \frac{\xi(m)}{\delta}}{2m}}.$$

- However, recall that the prior is not allowed to depend in any way on the training set.

Localized PAC-Bayesian bounds :

(1) Let us simply learn the prior !

- As stated in the first part of this tutorial: one may leave a part of the training set in order to learn the prior, and only use the remaining part of it to calculate the PAC-Bayesian bound.

Localized PAC-Bayesian bounds :

(1) Let us simply learn the prior !

- As stated in the first part of this tutorial: one may leave a part of the training set in order to learn the prior, and only use the remaining part of it to calculate the PAC-Bayesian bound.
  - A. Ambroladze, E. Parrado-Hernndez, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems* 18, (2006) Pages 9-16.
  - P. Germain, A. Lacasse, F. Laviolette and M. Marchand. PAC-Bayesian learning of linear classifiers, in *Proceedings of the 26nd International Conference on Machine Learning* (ICML'09, Montréal, Canada.). ACM Press (2009), 382, Pages 453-460.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(\rho \| \pi)$. This is all we need here.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(\rho \| \pi)$. This is all we need here.

This has been proposed in

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(\rho \| \pi)$. This is all we need here.

This has been proposed in

- the previous part of this tutorial, dedicated to linear separator when the chosen prior was: $\mathbf{w}_p = \mathbb{E}_{(\mathbf{x},y)\sim D}(y\,\boldsymbol{\phi}(\mathbf{x}))$.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(\rho \| \pi)$. This is all we need here.

This has been proposed in

- the previous part of this tutorial, dedicated to linear separator when the chosen prior was: $\mathbf{w}_p = \mathbb{E}_{(\mathbf{x},y)\sim D}(y\,\boldsymbol{\phi}(\mathbf{x}))$.
- O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, π will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(\rho \| \pi)$. This is all we need here.

This has been proposed in

- the previous part of this tutorial, dedicated to linear separator when the chosen prior was: $\mathbf{w}_p \ = \ \mathbb{E}_{(\mathbf{x},y)\sim D}(y\,\boldsymbol{\phi}(\mathbf{x}))$.
- O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.
- G. Lever, F. Laviolette, J. Shawe-Taylor. Distribution-Dependent PAC-Bayes Priors. Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT 2010), 119-133.

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

- in particular, Lever et al propose a distribution dependent prior of the form:

$$\pi(h) \;=\; \frac{1}{Z}\exp(-\gamma R(h))\,,$$

for some a priori chosen hyper-parameter gamma.

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

- in particular, Lever et al propose a distribution dependent prior of the form:

$$\pi(h) \;=\; \frac{1}{Z}\exp(-\gamma R(h))\,,$$

  for some a priori chosen hyper-parameter gamma.
- Such distribution dependent priors are designed to put more weight on accurate hypothesis and exponentially decrease the weight as the accuracies are decreasing. (A “wise” choice).

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

- in particular, Lever et al propose a distribution dependent prior of the form:

$$\pi(h) \;=\; \frac{1}{Z}\exp(-\gamma R(h))\,,$$

  for some a priori chosen hyper-parameter gamma.
- Such distribution dependent priors are designed to put more weight on accurate hypothesis and exponentially decrease the weight as the accuracies are decreasing. (A "wise" choice).
- Then, we can bound the KL-term under the restriction that the posterior is of the form

$$\rho(h) \;=\; \frac{1}{Z'}\exp(-\gamma R_S(h))\,.$$

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

- in particular, Lever et al propose a distribution dependent prior of the form:

$$\pi(h) \; = \; \frac{1}{Z} \exp(-\gamma R(h)) \, ,$$

  for some a priori chosen hyper-parameter gamma.
- Such distribution dependent priors are designed to put more weight on accurate hypothesis and exponentially decrease the weight as the accuracies are decreasing. (A “wise” choice).
- Then, we can bound the KL-term under the restriction that the posterior is of the form

$$\rho(h) \; = \; \frac{1}{Z'} \exp(-\gamma R_S(h)) \, .$$

  Again a suitable form for a posterior (and which this time is a known quantity).

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

The KL-term is bounded as follows:

$$\mathrm{KL}(\rho\|\pi) \;\leq\; \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}} \,+\, \frac{\gamma^2}{4m}\,.$$

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

The KL-term is bounded as follows:

$$\mathrm{KL}(\rho \| \pi) \;\leq\; \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{2\xi(m)}{\delta}} \;+\; \frac{\gamma^2}{4m}\,.$$

The trick: we apply a second PAC-bayesian bound and applied it to the KL-term.

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

The KL-term is bounded as follows:

$$\mathrm{KL}(\rho\|\pi) \;\leq\; \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}\,.$$

The trick: we apply a second PAC-bayesian bound and applied it to the KL-term.

This gives rise to a very *tight* localized PAC-Bayesian bound:

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

The KL-term is bounded as follows:

$$\mathrm{KL}(\rho\|\pi) \;\leq\; \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}} \;+\; \frac{\gamma^2}{4m}\,.$$

The trick: we apply a second PAC-bayesian bound and applied it to the KL-term.

This gives rise to a very *tight* localized PAC-Bayesian bound:

Lever et al. (2010)

For any D, any $\mathcal{H}$, any π of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\Bigg(\forall\,\rho \text{ on } \mathcal{H}\colon\;\; \mathrm{kl}(R_S(G_\rho), R(G_\rho)) \leq \\ \frac{1}{m}\left[\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta/2}} \;+\; \frac{\gamma^2}{4m} + \ln\frac{\xi(m)}{\delta/2}\right]\Bigg) \geq 1-\delta\,.$$

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

In other words, for any posterior ρ, there is a posterior ρ', aligned on π such that

$$B_\rho(\mathbf{x}) \;=\; B_{\rho'}(\mathbf{x})\,.$$

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

In other words, for any posterior ρ, there is a posterior ρ', aligned on π such that

$$B_\rho(\mathbf{x}) \;=\; B_{\rho'}(\mathbf{x})\,.$$

So, same classification capacity if one restrict itself to aligned posterior.

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

In other words, for any posterior ρ, there is a posterior ρ', aligned on π such that

$$B_\rho(\mathbf{x}) \;=\; B_{\rho'}(\mathbf{x})\,.$$

So, same classification capacity if one restrict itself to aligned posterior.

But then, the KL-term vanishes from the PAC-Bayesian bound !!!

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that ρ is **aligned** on π iff for all $h \in \mathcal{H}$, we have

$$\rho(h) + \rho(-h) = \pi(h) + \pi(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

In other words, for any posterior ρ, there is a posterior ρ', aligned on π such that

$$B_\rho(\mathbf{x}) \;=\; B_{\rho'}(\mathbf{x})\,.$$

So, same classification capacity if one restrict itself to aligned posterior.

But then, the KL-term vanishes from the PAC-Bayesian bound !!!

MAGIC !!!

Absence of KL for Aligned Posteriors

General theorem (McAllester)

$\mathrm{KL}(\rho\|\pi)$ arises when transforming the expectation over π to the expectation over ρ:

$$
\begin{aligned}
& \ln\left[\mathop{\mathbf{E}}_{h\sim\pi} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
\geq\ & \ln\left[\mathop{\mathbf{E}}_{h\sim\rho} \frac{\pi(h)}{\rho(h)} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
\geq\ & \mathop{\mathbf{E}}_{h\sim\rho} \ln\left[\frac{\pi(h)}{\rho(h)} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
=\ & m \mathop{\mathbf{E}}_{h\sim\rho} 2\left(R_S(h)-R(h)\right)^2 - \mathrm{KL}(\rho\|\pi) \\
\geq\ & m\cdot 2\left(\mathop{\mathbf{E}}_{h\sim\rho} R_S(h) - \mathop{\mathbf{E}}_{h\sim\rho} R(h)\right)^2 - \mathrm{KL}(\rho\|\pi) \\
=\ & m\cdot 2\left(R_S(G_\rho)-R(G_\rho)\right)^2 - \mathrm{KL}(\rho\|\pi)\ .
\end{aligned}
$$

Absence of KL for Aligned Posteriors

General theorem (McAllester)

$\mathrm{KL}(\rho\|\pi)$ arises when transforming the expectation over π to the expectation over ρ:

$$
\begin{aligned}
& \ln\left[\mathop{\mathbf{E}}_{h\sim\pi} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
\geq\ & \ln\left[\mathop{\mathbf{E}}_{h\sim\rho} \frac{\pi(h)}{\rho(h)} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
\geq\ & \mathop{\mathbf{E}}_{h\sim\rho} \ln\left[\frac{\pi(h)}{\rho(h)} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
=\ & m \mathop{\mathbf{E}}_{h\sim\rho} 2\left(R_S(h)-R(h)\right)^2 - \mathrm{KL}(\rho\|\pi) \\
\geq\ & m\cdot 2\left(\mathop{\mathbf{E}}_{h\sim\rho} R_S(h) - \mathop{\mathbf{E}}_{h\sim\rho} R(h)\right)^2 - \mathrm{KL}(\rho\|\pi) \\
=\ & m\cdot 2\left(R_S(G_\rho)-R(G_\rho)\right)^2 - \mathrm{KL}(\rho\|\pi)\ .
\end{aligned}
$$

Aligned posterior theorem

Here, we do the same operation for "free" (proof on next slide):

$$
\begin{aligned}
& \ln\left[\mathop{\mathbf{E}}_{h\sim\pi} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
=\ & \ln\left[\mathop{\mathbf{E}}_{h\sim\rho} e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
\geq\ & \mathop{\mathbf{E}}_{h\sim\rho} \ln\left[e^{m\cdot 2(R_S(h)-R(h))^2}\right] \\
=\ & m \mathop{\mathbf{E}}_{h\sim\rho} 2\left(R_S(h)-R(h)\right)^2 \\
\geq\ & m\cdot 2\left(\mathop{\mathbf{E}}_{h\sim\rho} R_S(h) - \mathop{\mathbf{E}}_{h\sim\rho} R(h)\right)^2 \\
=\ & m\cdot 2\left(R_S(G_\rho)-R(G_\rho)\right)^2\ .
\end{aligned}
$$

Absence of KL for Aligned Posteriors

Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ with $\mathcal{H}_1 \cap \mathcal{H}_2 = \emptyset$ such that for each $h \in \mathcal{H}_1 : \ -h \in \mathcal{H}_2$.

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{h\sim\pi} e^{m\cdot 2(R_S(h)-R(h))^2} \\
&= \int_{h\in\mathcal{H}_1} d\pi(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_2} d\pi(h) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&= \int_{h\in\mathcal{H}_1} d\pi(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_1} d\pi(-h) e^{m\cdot 2((1-R_S(h))-(1-R(h)))^2} \\
&= \int_{h\in\mathcal{H}_1} \left(d\pi(h) + d\pi(-h)\right) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&= \int_{h\in\mathcal{H}_1} \left(d\rho(h) + d\rho(-h)\right) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&= \int_{h\in\mathcal{H}_1} d\rho(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_2} d\rho(h) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&= \mathop{\mathbf{E}}_{h\sim\rho} e^{m\cdot 2(R_S(h)-R(h))^2} .
\end{aligned}
$$

Aknowledgements

A big thank's to Mario Marchand that initiated me to PAC-Bayes theory and that have been my main PAC-Bayes collaborator since then.

Thank's also to all members of my lab: the GRAAL.

Thank's also to Liva Ralaivola, David McAllester, Guy Lever and John Langford for more than insightful discussions about the subject.

Outline of the Tutorial

Part II François

- A Bit of PAC-Bayesian History
- Localized PAC-Bayesian bounds

Yevgeny

- **PAC-Bayesian bounds for unsupervised learning and density estimation**
- PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning
- Summary

PAC-Bayesian Inequality for Discrete Density Estimation

Lemma

Let $Z_1, \ldots, Z_m$ *be* m *random variables drawn according to an unknown distribution* p *on* $\{1, \ldots, K\}$*. Let* $\hat{p}$ *be the empirical distribution on* $\{1, \ldots, K\}$ *corresponding to the sample.*

$$\mathbb{E}\left[e^{m\mathrm{KL}(\hat{p}\|p)}\right] \leq (m+1)^{K-1}.$$

PAC-Bayesian Inequality for Discrete Density Estimation

Lemma

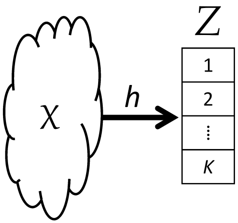
Let $Z_1, \ldots, Z_m$ be m random variables drawn according to an unknown distribution p on $\{1, \ldots, K\}$. Let $\hat{p}$ be the empirical distribution on $\{1, \ldots, K\}$ corresponding to the sample.

$$\mathbb{E}\left[e^{m\mathrm{KL}(\hat{p}\|p)}\right] \leq (m+1)^{K-1}.$$

	1	2	...	K
p_i	0.1	0.3	...	0.2
m_i	12	24	...	19
$\hat{p}_i = m_i/m$	12/100	24/100	...	19/100

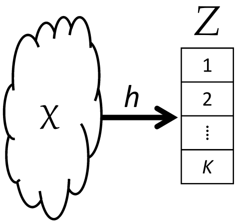
PAC-Bayes-KL Inequality

- $\mathcal{X}$ - sample space
- p - distribution over $\mathcal{X}$
- $\mathcal{H}$ - hypothesis space
- $\mathcal{Z}$ - finite, $|\mathcal{Z}| = K$
- Each $h \in \mathcal{H}$ is a mapping $h : \mathcal{X} \to \mathcal{Z}$
- p_h - induced distribution over $\mathcal{Z}$
- $\hat{p}_h$ - induced empirical distribution over $\mathcal{Z}$



PAC-Bayes-KL Inequality

- $\mathcal{X}$ - sample space
- p - distribution over $\mathcal{X}$
- $\mathcal{H}$ - hypothesis space
- $\mathcal{Z}$ - finite, $|\mathcal{Z}| = K$
- Each $h \in \mathcal{H}$ is a mapping $h : \mathcal{X} \to \mathcal{Z}$
- p_h - induced distribution over $\mathcal{Z}$
- $\hat{p}_h$ - induced empirical distribution over $\mathcal{Z}$



Theorem (PAC-Bayes-KL Inequality)

W.p. $\geq 1 - \delta$ *for all* ρ *simultaneously:*

$$\mathrm{KL}(\langle \hat{p}_h, \rho \rangle \| \langle p_h, \rho \rangle) \leq \frac{\mathrm{KL}(\rho \| \pi) + (K-1)\ln(m+1) + \ln \frac{1}{\delta}}{m}$$

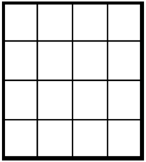
Application Example: Density Estimation with Co-clustering

Input

Sample $(X_1^1, X_1^2), \ldots, (X_m^1, X_m^2)$

Goal

Build an estimator $\rho(x^1, x^2)$ that minimizes $-\mathbb{E}_{p(X^1,X^2)}\left[\ln \rho(X^1, X^2)\right]$



Application Example: Density Estimation with Co-clustering

Input

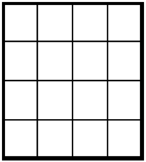
Sample $(X_1^1, X_1^2), \ldots, (X_m^1, X_m^2)$

Goal

Build an estimator $\rho(x^1, x^2)$ that minimizes $-\mathbb{E}_{p(X^1,X^2)}\left[\ln \rho(X^1, X^2)\right]$

Direct Estimation

Requires $\sim |X_1||X_2|$ samples



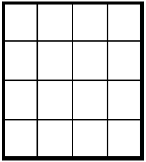
Application Example: Density Estimation with Co-clustering

Input

Sample $(X_1^1, X_1^2), \ldots, (X_m^1, X_m^2)$

Goal

Build an estimator $\rho(x^1, x^2)$ that minimizes $-\mathbb{E}_{p(X^1,X^2)}\left[\ln \rho(X^1, X^2)\right]$



Direct Estimation

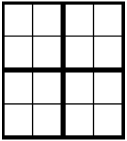
Requires $\sim |X_1||X_2|$ samples

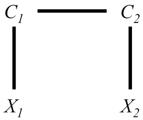
Can we do better?

Application Example: Density Estimation with Co-clustering

Idea

Try to find block structures





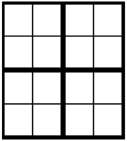
Model

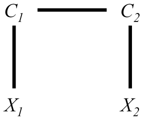
$\rho = \{\rho(c^1|x^1), \rho(c^2|x^2)\}$

Application Example: Density Estimation with Co-clustering

Idea

Try to find block structures





Model

$\rho = \{\rho(c^1|x^1), \rho(c^2|x^2)\}$

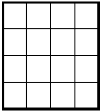
$\rho(x^1, x^2) = \sum_{c^1,c^2} \tilde{p}_\rho(c^1, c^2) \prod_{i=1}^2 \frac{\tilde{p}(x^i)}{\tilde{p}_\rho(c^i)} \rho(c^i|x^i)$

Application Example: Density Estimation with Co-clustering

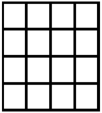
Bound

W.p. $\geq 1-\delta$:

$$-\mathbb{E}_{p(x^1,x^2)}\left[\ln \rho(X^1,X^2)\right] \leq \underbrace{\left(\sum_{i=1}^{2} \hat{\mathrm{H}}(X^i)\right)}_{\text{Approximation by product of marginals}} - \underbrace{\hat{\mathrm{I}}_\rho(C^1;C^2)}_{\text{Added value of clustering}} + \underbrace{\ln(|C^1||C^2|)\sqrt{\frac{\sum_i |X^i| \mathrm{I}_\rho(X_i;C_i) + \ldots}{2m}}}_{\text{Complexity of clustering}} + \ldots$$



$\hat{\mathrm{I}}_\rho(C^1;C^2) = 0$
$\mathrm{I}_\rho(X^i;C^i) = 0$



$\hat{\mathrm{I}}_\rho(C^1;C^2) = \hat{\mathrm{I}}(X^1;X^2)$
$\mathrm{I}_\rho(X^i;C^i) = \ln|X^i|$

Further Reading

Discrete Density Estimation

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 2010.

- Graph clustering
- Topic models

Continuous Density Estimation

Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *ALT*, 2010.

- Kernel density estimation

Outline of the Tutorial

Part II François

- A Bit of PAC-Bayesian History
- Localized PAC-Bayesian bounds

Yevgeny

- PAC-Bayesian bounds for unsupervised learning and density estimation
- **PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning**
- Summary

Martingales

Martingale difference sequence

$Z_1, \dots, Z_n$ is a *martingale difference sequence* if

$$\mathbb{E}[Z_i | Z_1, \dots, Z_{i-1}] = 0$$

Martingale

Let

$$M_j = \sum_{i=1}^{j} Z_i$$

then $M_1, \dots, M_n$ is a martingale.

Examples

- Random walk
- Gambler's capital

PAC-Bayesian Inequalities for Martingales

$$
\mathcal{H}\left\{
\begin{array}{ccccccc}
M_1(1) & \substack{\rightarrow \\ \searrow} & M_2(1) & \substack{\rightarrow \\ \searrow} & \cdots & \substack{\rightarrow \\ \searrow} & M_n(1) \\
\updownarrow & & \updownarrow & & & & \updownarrow \\
M_1(2) & \substack{\nearrow \\ \rightarrow \\ \searrow} & M_2(2) & \substack{\nearrow \\ \rightarrow \\ \searrow} & \cdots & \substack{\nearrow \\ \rightarrow \\ \searrow} & M_n(2) \\
\updownarrow & & \updownarrow & & & & \updownarrow \\
\vdots & & \vdots & & \ddots & & \vdots \\
\updownarrow & & \updownarrow & & & & \updownarrow \\
M_1(h) & \substack{\nearrow \\ \rightarrow \\ \searrow} & M_2(h) & \substack{\nearrow \\ \rightarrow \\ \searrow} & \cdots & \substack{\nearrow \\ \rightarrow \\ \searrow} & M_n(h) \\
\updownarrow & & \updownarrow & & & & \updownarrow \\
\vdots & & \vdots & & \ddots & & \vdots
\end{array}
\right.
$$

$$\xrightarrow[time]{\qquad\qquad\qquad}$$

$$\langle M_n, \rho \rangle \leq \ ???$$

Example: Capital of multiple gamblers in a zero-sum game

Background: Bernstein's Inequality for Martingales

Lemma (Bernstein's Inequality for Martingales)

Let $Z_1, \ldots, Z_n$ *be a martingale difference sequence, such that* $Z_i \leq C$ *for all* i.

Let $M_n = \sum_{i=1}^n Z_i$ *and* $V_n = \sum_{i=1}^n \mathbb{E}[Z_i^2 | Z_1, \ldots, Z_{i-1}]$.

Then for any fixed $\lambda \in [0, \frac{1}{C}]$:

$$\mathbb{E}\left[e^{\lambda M_n - (e-2)\lambda^2 V_n}\right] \leq 1.$$

PAC-Bayes-Bernstein Inequality for Martingales

Theorem (PAC-Bayes-Bernstein Inequality)

Assume that $|Z_i(h)| \leq C$ for all i and h with probability 1. Fix a reference distribution π over $\mathcal{H}$. Then, for any $\delta \in (0,1)$ with probability greater than $1-\delta$, simultaneously for all distributions ρ over $\mathcal{H}$ that satisfy

"certain technical condition"

we have

$$|\langle M_n, \rho \rangle| \lesssim \sqrt{\langle V_n, \rho \rangle \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta} \right)}$$

Application Example: Importance Weighted Sampling in Multiarmed Bandits

Multiarmed Bandits

- Given a set $\mathcal{A}$ of K actions
- Each action $a \in \mathcal{A}$ yields reward R distributed by $p(r|a)$ and bounded in $[0, 1]$
- $r(a) = \mathbb{E}_{R \sim p(r|a)}[R]$ - expected reward for playing a

Application Example: Importance Weighted Sampling in Multiarmed Bandits

Multiarmed Bandits

- Given a set $\mathcal{A}$ of K actions
- Each action $a \in \mathcal{A}$ yields reward R distributed by $p(r|a)$ and bounded in $[0, 1]$
- $r(a) = \mathbb{E}_{R \sim p(r|a)}[R]$ - expected reward for playing a

Limited Feedback

- At each round t the player plays action $A_t \in \mathcal{A}$
- The player obtains reward R_t for the action A_t
- Rewards for other actions are not observed

Importance Weighted Sampling

Definitions

$$R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases}$$

Importance Weighted Sampling

Definitions

$$R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{T}_t = \{A_1, \ldots, A_t, R_1, \ldots, R_t\} \quad - \quad \text{History of the game}$$

Importance Weighted Sampling

Definitions

$$R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{T}_t = \{A_1, \ldots, A_t, R_1, \ldots, R_t\} \quad - \quad \text{History of the game}$$

Observations

$$\mathbb{E}[R_t^a | \mathcal{T}_{t-1}] = \rho_t(a) \left(\frac{1}{\rho_t(a)} \mathbb{E}[R_t | A_t = a, \mathcal{T}_{t-1}] \right) + 0 = r(a)$$

Importance Weighted Sampling

Definitions

$$R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{T}_t = \{A_1, \ldots, A_t, R_1, \ldots, R_t\} \quad - \quad \text{History of the game}$$

Observations

$$\mathbb{E}[R_t^a | \mathcal{T}_{t-1}] = \rho_t(a) \left(\frac{1}{\rho_t(a)} \mathbb{E}[R_t | A_t = a, \mathcal{T}_{t-1}] \right) + 0 = r(a)$$

$(R_1^a - r(a)), (R_2^a - r(a)), \ldots$ – is a martingale difference sequence

Variance of Importance Weighted Sampling

$$R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}[R_t^a | \mathcal{T}_{t-1}] = r(a)$$

Variance

$$\mathbb{E}\left[(R_t^a - r(a))^2 \,\middle|\, \mathcal{T}_{t-1} \right] \le \frac{1}{\rho_t(a)}$$

Multiarmed Bandits with Side Information

	a_1	$\ldots$	a_K
s_1			
$\vdots$		$p(r \mid a_i, s_j)$	
s_N			

Multiarmed Bandits with Side Information

	a_1	$\dots$	a_K
s_1			
$\vdots$		$p(r\|a_i, s_j)$	
s_N			

Game Round

- Pick a policy $\rho_t(a|s)$
- Observe side information $S_t \sim p(s)$
- Play an action $A_t \sim \rho_t(a|S_t)$
- Obtain a reward $R_t \sim p(r|A_t, S_t)$.

Definitions

$\mathcal{H}$ - all possible deterministic strategies
Each $h \in \mathcal{H}$ assigns one action to each state $a = h(s)$
$|\mathcal{H}| = K^N$

Example:

	a_1	a_2	a_3
s_1	*		
s_2	*		
s_3		*	
s_4			*

Definitions

Rewards

$$R_t^{a,S_t} = \begin{cases} \frac{1}{\rho_t(a|S_t)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

Definitions

Rewards

$$R_t^{a,S_t} = \begin{cases} \frac{1}{\rho_t(a|S_t)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

$$\hat{R}_t(h) = \sum_{i=1}^{t} R_i^{h(S_i),S_i}$$

Definitions

Rewards

$$R_t^{a,S_t} = \begin{cases} \frac{1}{\rho_t(a|S_t)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

$$\hat{R}_t(h) = \sum_{i=1}^{t} R_i^{h(S_i),S_i}$$

Regret

$$\Delta(h) = R(h^*) - R(h)$$

$$\hat{\Delta}_t(h) = \hat{R}_t(h^*) - \hat{R}_t(h).$$

Definitions

Rewards

$$R_t^{a,S_t} = \begin{cases} \frac{1}{\rho_t(a|S_t)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases}$$

$$\hat{R}_t(h) = \sum_{i=1}^{t} R_i^{h(S_i),S_i}$$

Regret

$$\Delta(h) = R(h^*) - R(h)$$

$$\hat{\Delta}_t(h) = \hat{R}_t(h^*) - \hat{R}_t(h).$$

Martingales

$$\left(\hat{\Delta}_t(h) - t\Delta(h)\right)$$

PAC-Bayesian Regret Bound

Reminder: PAC-Bayes-Bernstein Inequality for Martingales

$$|\langle M_n, \rho \rangle| \lesssim \sqrt{\langle V_n, \rho \rangle \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta} \right)}$$

PAC-Bayesian Regret Bound

Reminder: PAC-Bayes-Bernstein Inequality for Martingales

$$|\langle M_n, \rho \rangle| \lesssim \sqrt{\langle V_n, \rho \rangle \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta} \right)}$$

Treating $\mathrm{KL}(\rho \| \pi)$

Pick a combinatorial prior π over $\mathcal{H}$, then:
$\mathrm{KL}(\rho \| \pi) \leq N \mathrm{I}_\rho(S; A) + K \ln N + K \ln K$

PAC-Bayesian Regret Bound

Reminder: PAC-Bayes-Bernstein Inequality for Martingales

$$|\langle M_n, \rho \rangle| \lesssim \sqrt{\langle V_n, \rho \rangle \left(\mathrm{KL}(\rho \| \pi) + \ln \frac{1}{\delta} \right)}$$

Treating $\mathrm{KL}(\rho \| \pi)$

Pick a combinatorial prior π over $\mathcal{H}$, then:
$\mathrm{KL}(\rho \| \pi) \leq N \mathrm{I}_\rho(S; A) + K \ln N + K \ln K$

Treating $\langle V_n, \rho \rangle$

Smooth the playing strategies for all $t < n$ by ε

PAC-Bayesian Regret Bound

$$\begin{aligned}\langle \Delta, \rho_n \rangle &= \frac{1}{n}\langle \underbrace{\left(n\Delta - \hat{\Delta}_n\right)}_{\text{Martingales}}, \rho_n \rangle + \frac{1}{n}\langle \hat{\Delta}_n, \rho_n \rangle \\ &\le \underbrace{\frac{\sqrt{\langle V_n, \rho_n \rangle (N \mathrm{I}_{\rho_n}(S;A) + K \ln N + \ldots) \ldots}}{n}}_{\text{Policy complexity}} + \underbrace{\frac{1}{n}\langle \hat{\Delta}_n, \rho_n \rangle}_{\substack{\text{Empirical} \\ \text{Performance}}}\end{aligned}$$

PAC-Bayesian Regret Bound

$$
\begin{aligned}
\langle \Delta, \rho_n \rangle &= \frac{1}{n} \langle \underbrace{\left(n\Delta - \hat{\Delta}_n \right)}_{\text{Martingales}}, \rho_n \rangle + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle \\
&\leq \underbrace{\frac{\sqrt{\langle V_n, \rho_n \rangle (N \mathrm{I}_{\rho_n}(S; A) + K \ln N + \ldots) \ldots}}{n}}_{\text{Policy complexity}} + \underbrace{\frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle}_{\substack{\text{Empirical} \\ \text{Performance}}}
\end{aligned}
$$

For Comparison

$$
\ln |\mathcal{H}| = \ln K^N = N \ln K
$$

$$
0 \leq N \mathrm{I}_{\rho_n}(S; A) \leq N \ln K
$$

Experiments

Setting

Experiment 1

	a_1	...	a_{20}
s_1	0.6	0.5	0.5
$\vdots$	0.6	0.5	0.5
s_{100}	0.6	0.5	0.5

$H(A^{h^*}) = \ln(1) = 0$

Experiments

Setting

Experiment 1

	a_1	...	a_{20}
s_1	0.6	0.5	0.5
$\vdots$	0.6	0.5	0.5
s_{100}	0.6	0.5	0.5

$H(A^{h^*}) = \ln(1) = 0$

Experiment 2

	a_1	a_2	a_3	...	a_{20}
s_1	0.6	0.5	0.5	0.5	0.5
$\vdots$	0.6	0.5	0.5	0.5	0.5
s_{33}	0.5	0.6	0.5	0.5	0.5
$\vdots$	0.5	0.6	0.5	0.5	0.5
s_{66}	0.5	0.5	0.6	0.5	0.5
$\vdots$	0.5	0.5	0.6	0.5	0.5
s_{100}	0.5	0.5	0.6	0.5	0.5

$H(A^{h^*}) = \ln(3) \approx 1$

Experiments

Setting

Experiment 1

	a_1	...	a_{20}
s_1	0.6	0.5	0.5
$\vdots$	0.6	0.5	0.5
s_{100}	0.6	0.5	0.5

$H(A^{h^*}) = \ln(1) = 0$

Experiment 2

	a_1	a_2	a_3	...	a_{20}
s_1	0.6	0.5	0.5	0.5	0.5
$\vdots$	0.6	0.5	0.5	0.5	0.5
s_{33}	0.5	0.6	0.5	0.5	0.5
$\vdots$	0.5	0.6	0.5	0.5	0.5
s_{66}	0.5	0.5	0.6	0.5	0.5
$\vdots$	0.5	0.5	0.6	0.5	0.5
s_{100}	0.5	0.5	0.6	0.5	0.5

$H(A^{h^*}) = \ln(3) \approx 1$

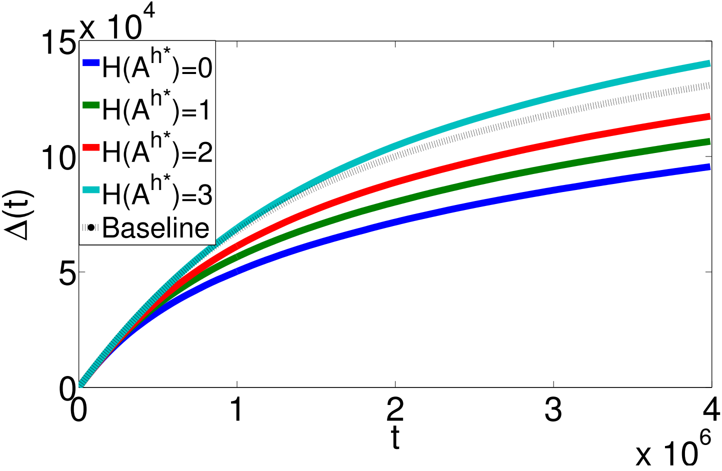
Experiment 3

$H(A^{h^*}) = \ln(7) \approx 3$

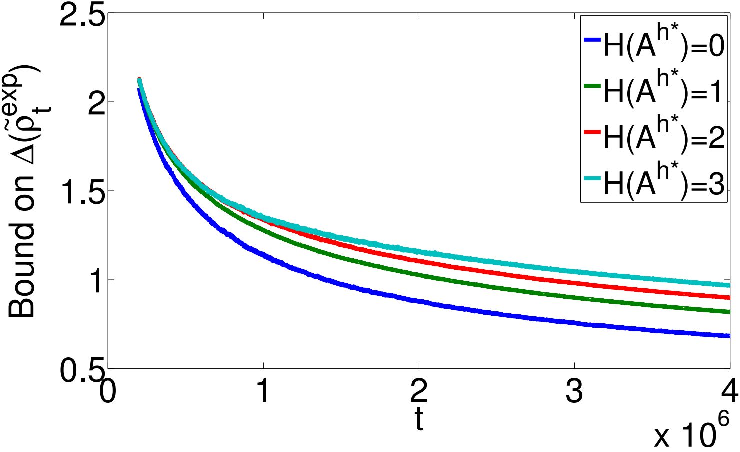
Experiment 4

$H(A^{h^*}) = \ln(20) \approx 4$

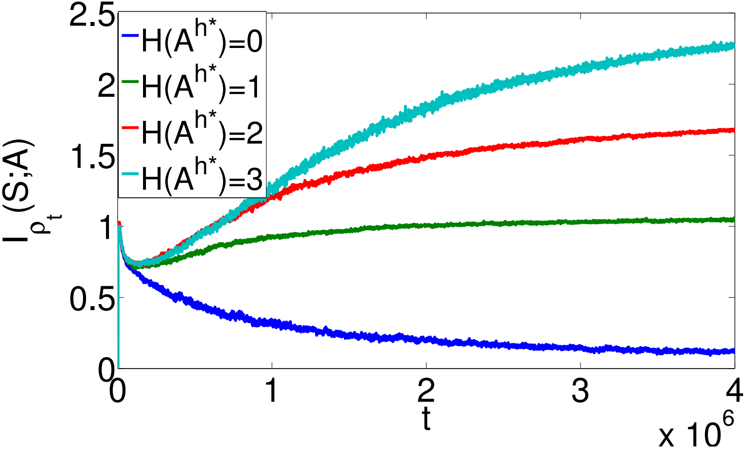
Experiments - Regret Graph



Experiments - Bound



Experiments - Mutual Information



Further Reading

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 2012. Preprint available on arxiv.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *NIPS*, 2011.

Outline of the Tutorial

Part II François

- A Bit of PAC-Bayesian History
- Localized PAC-Bayesian bounds

Yevgeny

- PAC-Bayesian bounds for unsupervised learning and density estimation
- PAC-Bayes-Bernstein inequality for martingales and its applications in reinforcement learning
- **Summary**

Summary: A General Workflow for Deriving a PAC-Bayesian Bound

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

- Design a hypothesis space $\mathcal{H}$
- Design a reference measure π over $\mathcal{H}$
- Pick $f(h)$
- Bound $\mathbb{E}[\langle e^f, \pi \rangle]$ *(usually, by bounding $\mathbb{E}[e^f]$)*
- Pick the form of ρ
- Bound $\mathrm{KL}(\rho \| \pi)$
- Combine everything together

Summary

$$\langle f, \rho \rangle \leq \mathrm{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$

Choice of f
PAC-Bayes-Hoeffding $f(h) = \lambda(L(h) - \hat{L}(h))$
PAC-Bayes-kl $f(h) = n\mathrm{kl}(\hat{L}(h) \| L(h))$
PAC-Bayes-Bernstein $f(h) = \lambda(\hat{L}(h) - L(h)) - (e-2)\lambda^2 V_n(h)$
PAC-Bayes-KL $f(h) = n\mathrm{KL}(\hat{p}(h) \| p(h))$
Martingales ...
...

Choice of π
Combinatorial $\mathrm{KL}(\rho \| \pi) \leq \mathrm{I}_\rho(X; C)$
Gaussian $\mathrm{KL}(\rho \| \pi) \leq \|w\|_2$
Laplacian $\mathrm{KL}(\rho \| \pi) \leq \|w\|_1$
Distribution-Dependent $\mathrm{KL}(\rho \| \pi) \leq \gamma\sqrt{\ln(..)/m} + \frac{\gamma^2}{4m}$
...

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes
- PAC ...
  - Strict generalization guarantees

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes
- PAC ...
  - Strict generalization guarantees
- ... and Bayesian
  - Easy way to incorporate prior knowledge
    both structural and distribution-dependent

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes
- PAC ...
  - Strict generalization guarantees
- ... and Bayesian
  - Easy way to incorporate prior knowledge
    both structural and distribution-dependent
- Bridges frequentist and Bayesian approaches

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes
- PAC ...
  - Strict generalization guarantees
- ... and Bayesian
  - Easy way to incorporate prior knowledge
    both structural and distribution-dependent
- Bridges frequentist and Bayesian approaches
- Tight bounds

Summary: PAC-Bayesian Analysis

A Natural and General Way to do Model Order Selection

- Generality
  - Supervised, Unsupervised, Reinforcement, ..., Learning
- Modularity
  - Any concentration inequality (Hoeffding/Bernstein/...)
    with any prior (Gaussian/Laplace/combinatorial/...)
  - For factorisable distributions (graphical models) KL factorizes
- PAC ...
  - Strict generalization guarantees
- ... and Bayesian
  - Easy way to incorporate prior knowledge
    both structural and distribution-dependent
- Bridges frequentist and Bayesian approaches
- Tight bounds
- Drives good algorithms

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J.-Y. Audibert. *Théorie Statistique de l?Apprentissage : une approche PAC-Bayésienne*. thèse de doctorat de l?Université Paris VI, 2004.

Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 2007a.

Jean-Yves Audibert and Olivier Bousquet. Combining pac-bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8:863–889, 2007b.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, 2007.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22, 2004.

Monroe D. Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009a.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Sara Shanian. From pac-bayes bounds to kl regularization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 603–610. 2009b. URL http://books.nips.cc/papers/files/nips22/NIPS2009_0456.pdf.

Pascal Germain, Alexandre Lacoste, François Laviolette, Mario Marchand, and Sara Shanian. A pac-bayes sample-compression approach to kernel methods. In *ICML*, pages 297–304, 2011.

Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for sample-compressed Gibbs classifiers. *Proc. of the 22nth International Conference on Machine Learning (ICML)*, pages 481–488, 2005.

François Laviolette, Mario Marchand, and Jean-Francis Roy. From pac-bayes bounds to quadratic programs for majority votes. In *ICML*, pages 649–656, 2011.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), 2003.

David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1998.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-IID data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 2010.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 2002.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11, 2010.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 2012. Accepted. Preprint available at http://arxiv.org/abs/1110.6886.

John Shawe-Taylor and David Hardoon. Pac-bayes analysis of maximum entropy classification. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1997.

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.

Monroe D. Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.