PAC-Bayesian Analysis and its Applications

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

François Laviolette  Université Laval

John Shawe-Taylor  University College London

ECML-PKDD-2012 Tutorial
Outline of the Tutorial

Part I

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

Yevgeny

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

▶ 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

François
PAC (Probably Approximately Correct) Learning Framework  (*Valiant, 1984*)

**Approximately**

Provide guarantees on the approximation error of empirical estimates...

**Probably**

... that hold with high probability with respect to representativeness of the observed sample.
Supervised Learning: Some Basic Definitions

\( \mathcal{X} \) - sample space

\( \mathcal{Y} \) - label space

\( \ell(y, y') \) - 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 $\rho$ 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 $\rho$ 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 $\rho$

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

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

Let $\rho$ 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 $\rho$

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

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

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

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

Let $\rho$ and $\pi$ be two distributions over $\mathcal{H}$

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

\[
= \langle \rho, \ln \frac{\rho}{\pi} \rangle = \begin{cases} 
\sum_h \rho(h) \ln \frac{\rho(h)}{\pi(h)}, & \text{Discrete } \mathcal{H} \\
\int_\mathcal{H} \ln \left( \frac{\rho(h)}{\pi(h)} \right) \rho(h) dh, & \text{Continuous } \mathcal{H}
\end{cases}
\]
PAC-Bayes-Hoeffding Inequality \textit{(McAllester, 1998, 1999)}

**Theorem (Simplified version)**

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

$$L(\rho) \lesssim \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}.$$
PAC-Bayes-Hoeffding Inequality  \((\text{McAllester, 1998, 1999})\)

**Theorem (Simplified version)**

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

\[
L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}.
\]

For comparison: Hoeffding’s inequality for individual \(h\)

\[
L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.
\]
PAC-Bayes-Hoeffding Inequality  (*McAllester, 1998, 1999*)

**Theorem (Simplified version)**

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

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

For comparison: Hoeffding’s inequality for individual $h$

- If $\rho = \pi$, then $\text{KL}(\rho\|\pi) = 0$

$$L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$
PAC-Bayes-Hoeffding Inequality  \textit{(McAllester, 1998, 1999)}

Theorem (Simplified version)

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

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

For comparison: Hoeffding’s inequality for individual $h$

- If $\rho = \pi$, then $\text{KL}(\rho\|\pi) = 0$
- If $\mathcal{H}$ is finite and $\pi(h) = \frac{1}{|\mathcal{H}|}$, then $\text{KL}(\rho\|\pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle$
Theorem (Simplified version)

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

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

For comparison: Hoeffding’s inequality for individual $h$

- If $\rho = \pi$, then $\text{KL}(\rho\|\pi) = 0$
- If $\mathcal{H}$ is finite and $\pi(h) = \frac{1}{|\mathcal{H}|}$, then
  $$\text{KL}(\rho\|\pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle$$
PAC-Bayes-Hoeffding Inequality (McAllester, 1998, 1999)

Theorem (Simplified version)

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

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

For comparison: Hoeffding’s inequality for individual $h$

- If $\rho = \pi$, then $\text{KL}(\rho\|\pi) = 0$
- If $\mathcal{H}$ is finite and $\pi(h) = \frac{1}{|\mathcal{H}|}$, then
  \[ \text{KL}(\rho\|\pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle \]
  \[ = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle \]
  \[ = \ln |\mathcal{H}| - H(\rho) \leq \ln |\mathcal{H}| \]
Theorem (Simplified version)

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

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

For comparison: Hoeffding’s inequality for individual \( h \)

\[
L(h) \leq \hat{L}(h) + \sqrt{\ln \frac{1}{\delta}}.
\]

- If \( \rho = \pi \), then \( KL(\rho\|\pi) = 0 \)
- If \( \mathcal{H} \) is finite and \( \pi(h) = \frac{1}{|\mathcal{H}|} \), then

\[
KL(\rho\|\pi) = \langle \ln \frac{\rho}{\pi}, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle = \ln |\mathcal{H}| - H(\rho) \leq \ln |\mathcal{H}|
\]

(we recover the union bound)
Intuition Behind the Bound

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

\[ \text{KL}(\rho\|\pi) = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle - \text{H}(\rho) \]

- Description length
- Entropy
Intuition Behind the Bound

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

\[ \text{KL}(\rho\|\pi) = \langle \ln \frac{1}{\pi}, \rho \rangle + \langle \ln \rho, \rho \rangle = \langle \ln \frac{1}{\pi}, \rho \rangle - \text{H}(\rho) \]

\[ \text{Description length} \quad \text{Entropy} \]

Trade-off

Pick \( \rho \) 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

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

Relation

1. Explicit way to incorporate prior information (via \( \pi(h) \))
Relation and Difference with Bayesian Learning

\[ L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{KL(\rho\|\pi) + \ln\frac{1}{\delta}}{2m}}. \]

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

\[ L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{2m}}. \]

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

\[ L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}. \]

**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

\[ L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}. \]

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

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

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 \( \rho \) (including the Bayes posterior)
Relation and Difference with VC-theory and Rademacher complexities

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

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

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

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
3. The bound is defined for randomized classifiers \( \rho \) (not individual \( h \)); but workarounds exist in many cases
Relation to Statistical Physics

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

- Rewrite as a parameterized trade-off

\[ \mathcal{F}(\rho, \beta) = \beta m \hat{L}(\rho) + \text{KL}(\rho\|\pi) \]
Relation to Statistical Physics

\[ L(\rho) \preceq \hat{L}(\rho) + \sqrt{\frac{\KL(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}. \]

- Rewrite as a parameterized trade-off

\[ \mathcal{F}(\rho, \beta) = \beta m \hat{L}(\rho) + \KL(\rho\|\pi) \]

- The bound provides the optimal temperature to study the system depending on
  - The size of the sample \( m \)
  - Empirical properties of the system \( \langle \hat{L}, \rho \rangle \)
Theorem (Simplified version)

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

$$L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\text{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{2m}}.$$
Proof Idea: Basis

Theorem (Variational Definition of KL-divergence \((Donsker and Varadhan, 1975)\))

$$\text{KL}(\rho\|\pi) = \sup_{f} \left( \langle f, \rho \rangle - \ln \langle e^f, \pi \rangle \right)$$
Proof Idea: Basis

Theorem (Variational Definition of KL-divergence (Donsker and Varadhan, 1975))

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

Corollary (Change of Measure Inequality)

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

$$\langle f, \rho \rangle \leq \text{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$
Proof Idea: 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 $1 - \delta$:

$$Z \leq \frac{1}{\delta} \mathbb{E}[Z]$$
Proof Idea: 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 $Z_1, \ldots, Z_m$ be i.i.d., such that $Z_i \in [0, 1]$. Then for any $\lambda$:

$$\mathbb{E} \left[ e^{\lambda \frac{1}{m} \sum_{i=1}^{m} (\mathbb{E}[Z_i] - Z_i)} \right] \leq e^{\lambda^2 / (8m)}$$
Proof Idea

Step 1: Change of Measure Inequality
For any function $f: \mathcal{H} \rightarrow \mathbb{R}$ and any $\rho$ and $\pi$:

$$\langle f, \rho \rangle \leq KL(\rho \| \pi) + \ln \langle e^f, \pi \rangle$$
Proof Idea

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \rightarrow \mathbb{R}$ and any $\rho$ and $\pi$:

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

Step 2: Take $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$. Bound $\langle e^f, \pi \rangle$. 
Proof Idea

Step 1: Change of Measure Inequality
For any function \( f : \mathcal{H} \to \mathbb{R} \) and any \( \rho \) and \( \pi \):

\[
\langle f, \rho \rangle \leq \text{KL}(\rho\|\pi) + \ln \langle e^f, \pi \rangle
\]

Step 2: Take \( f(h) = \lambda \left( L(h) - \hat{L}(h) \right) \). Bound \( \langle e^f, \pi \rangle \).

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

Step 1: Change of Measure Inequality

For any function $f : \mathcal{H} \rightarrow \mathbb{R}$ and any $\rho$ and $\pi$:

$$\langle f, \rho \rangle \leq KL(\rho \parallel \pi) + \ln \langle e^f, \pi \rangle$$

Step 2: Take $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$. Bound $\langle e^f, \pi \rangle$.

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

$$= \frac{1}{\delta} \left\langle \mathbb{E} \left[ e^f \right], \pi \right\rangle \quad \text{(Linearity of } \mathbb{E}; \pi \text{ is deterministic)}$$
Proof Idea

Step 1: Change of Measure Inequality
For any function $f : \mathcal{H} \to \mathbb{R}$ and any $\rho$ and $\pi$:

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

Step 2: Take $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$. Bound $\langle e^f, \pi \rangle$.

$$\langle e^f, \pi \rangle \leq \frac{1}{\delta} \mathbb{E} \left[ \langle e^f, \pi \rangle \right]$$

(w.p. $\geq 1 - \delta$; Markov)

$$= \frac{1}{\delta} \left\langle \mathbb{E} \left[ e^f \right], \pi \right\rangle$$

(Linearity of $\mathbb{E}$; $\pi$ is deterministic)

$$\leq \frac{1}{\delta} \left\langle e^{\lambda^2/(8m)}, \pi \right\rangle$$

(Hoeffding)
Proof Idea

Step 1: Change of Measure Inequality
For any function $f : \mathcal{H} \rightarrow \mathbb{R}$ and any $\rho$ and $\pi$:

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

Step 2: Take $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$. Bound $\langle e^f, \pi \rangle$.

$$\langle e^f, \pi \rangle \leq \frac{1}{\delta} \mathbb{E} \left[ \langle e^f, \pi \rangle \right]$$

$$= \frac{1}{\delta} \langle \mathbb{E} [e^f], \pi \rangle$$

$$\leq \frac{1}{\delta} \langle e^{\lambda^2/(8m)}, \pi \rangle$$

$$= \frac{1}{\delta} e^{\lambda^2/(8m)}$$

(w.p. $\geq 1 - \delta$; Markov)

(Linearity of $\mathbb{E}$; $\pi$ is deterministic)

(Hoeffding)
Proof Idea

Step 1: Change of Measure Inequality
For any function $f : \mathcal{H} \to \mathbb{R}$ and any $\rho$ and $\pi$:

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

Step 2: Take $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$, by Markov&Hoeffding

$$\ln \langle e^f, \pi \rangle \leq \ln \frac{1}{\delta} + \frac{\lambda^2}{8m}$$
Proof Idea

**Step 1: Change of Measure Inequality**
For any function $f : \mathcal{H} \to \mathbb{R}$ and any $\rho$ and $\pi$:

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

**Step 2: Take** $f(h) = \lambda \left( L(h) - \hat{L}(h) \right)$, by Markov & Hoeffding

$$\ln \langle e^f, \pi \rangle \leq \ln \frac{1}{\delta} + \frac{\lambda^2}{8m}$$

**Step 3: Substitute and normalize by $\lambda$**

$$\langle L(h) - \hat{L}(h), \rho \rangle \leq \frac{\text{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda}{8m}$$
Proof Idea

Step 1: Change of Measure Inequality
For any function \( f : \mathcal{H} \to \mathbb{R} \) and any \( \rho \) and \( \pi \):

\[ \langle f, \rho \rangle \leq KL(\rho\|\pi) + \ln \langle e^f, \pi \rangle \]

Step 2: Take \( f(h) = \lambda \left( L(h) - \hat{L}(h) \right) \), by Markov&Hoeffding

\[ \ln \langle e^f, \pi \rangle \leq \ln \frac{1}{\delta} + \frac{\lambda^2}{8m} \]

Step 3: Substitute and normalize by \( \lambda \)

\[ \langle L(h) - \hat{L}(h), \rho \rangle \leq \frac{KL(\rho\|\pi) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda}{8m} \]

Step 4: Optimize over \( \lambda \)
PAC-Bayes-Hoeffding Inequality

Theorem (Simplified version)

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

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

Outine of the Tutorial

Part I

- 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

$$\rho(y | x_1, x_2) = \sum_{c_1, c_2} \rho(y | c_1, c_2) \rho(c_1 | x_1) \rho(c_2 | x_2)$$
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
- \( \pi \) - 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 \(|C_i|\) \(i \in \{1, 2\}\)
Prior Construction

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

- $|X_i|$ possibilities to choose $|C_i|$ ($i \in \{1, 2\}$)
- $\leq |X_i||C_i|^{-1}$ possibilities to choose the sizes of the clusters
- $\left(\frac{|X_i|}{n_1^i, \ldots, n_{|C_i|}^i}\right) \leq e^{|X_i|H(C_i)}$ 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 $|C_i|$ \( (i \in \{1, 2\}) \)
- $\leq |X_i||C_i|^{-1}$ possibilities to choose the sizes of the clusters
- \( \binom{|X_i|}{n_i^1, \ldots, n_i^{|C_i|}} \leq e^{|X_i|H(C_i)} \) 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 \(|C_i|\) \((i \in \{1, 2\})\)
- \(|X_i|^{|C_i|^{-1}}\) possibilities to choose the sizes of the clusters
- \(\left(\begin{array}{c} |X_i| \\ n_1^i, \ldots, n_{|C_i|}^i \end{array}\right) \leq e^{X_i |H(C_i)}\) possibilities to assign \(x_i\)-s to \(c_i\)-s
- \(|Y|^{|C_1||C_2|}\) possibilities to assign labels to partition cells

\[
\pi(h) \geq \exp \left( \sum_{i=1}^{2} (-|X_i|H_h(C_i) - |C_i| \ln |X_i|) - |C_1||C_2| \ln |Y| \right)
\]

\[
\binom{4}{2} = 6 \text{ balanced partitions} \quad 4 \text{ unbalanced partitions}
\]
Bounding $\text{KL}(\rho \| \pi)$

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

$$\rho = \{ \rho(c_1|x_1), \rho(c_2|x_2), \rho(y|x_1, x_2) \}$$
Bounding $\text{KL}(\rho \| \pi)$

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

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

After some calculations...

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

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

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

After some calculations...

$$\text{KL}(\rho \| \pi) \leq \sum_{i=1}^{2} (|X_i|I_{\rho}(X_i; C_i) + |C_i| \ln |X_i|) + |C_1||C_2| \ln |Y|$$

$$\rho(x_i, c_i) = \frac{1}{|X_i|} \rho(c_i|x_i)$$
With probability $\geq 1 - \delta$, for all $\rho$:

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

- **Lowest Complexity**
  \(I_\rho(X_i; C_i) = 0\)
  - Grid with no black squares

- **Lower Complexity**
  - Grid with 1 black square

- **Higher Complexity**
  - Grid with 2 black squares

- **Highest Complexity**
  \(I_\rho(X_i; C_i) = \ln |X_i|\)
  - Grid with all black squares
Two Types of Prior Knowledge

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

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

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:

\[ L(\rho) \leq \hat{L}(\rho) + \sqrt{\frac{\sum_{i=1}^{2} (|X_i|I_\rho(X_i; C_i) + |C_i| \ln |X_i|) + |C_1|^2|C_2|^2 \ln |Y| + \ln \frac{1}{\delta} + \nu(\rho)}}}{2m} \]

Replace with a trade-off and apply linear search over \( \beta \)

\[ \mathcal{F}(\rho, \beta) = \beta m \hat{L}(\rho) + \sum_{i=1}^{2} |X_i|I_\rho(X_i; C_i) \]
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.
- $\frac{1}{\beta} \sum_{i=1}^{2} |X_i| I_\rho(X_i; C_i)$ 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


Outline of the Tutorial

Part I

▶ 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

Yevgeny
Acknowledgements

Many inputs to the presentation, but special thanks to:

- Emilio Parado-Hernandez
- Guy Lever
- Shiliang Sun
The small $\text{kl}$ divergence

- Let $p$ and $q$ be biases of two Bernoulli random variables.

$$\text{kl}(q\|p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p} = \text{KL}([q, 1 - q] \| [p, 1 - p])$$
The small $\text{kl}$ divergence

- Let $p$ and $q$ be biases of two Bernoulli random variables.

\[
\text{kl}(q\|p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p} = \text{KL}([q, 1 - q]\|\pmb{1} - [p, 1 - p])
\]

- By Pinsker’s inequality:

\[
\text{kl}(q\|p) \geq 2(q - p)^2
\]
The small kl divergence

- Let \( p \) and \( q \) be biases of two Bernoulli random variables.

\[
kl(q\|p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p} = KL([q, 1 - q] \| [p, 1 - p])
\]

- By Pinsker’s inequality:

\[
kl(q\|p) \geq 2(q - p)^2
\]

Here is a comparison between \( kl(q\|p) \) and \( 2(q - p)^2 \) when \( p \) varies

- a) when \( q = .5 \)

- b) when \( q = .01 \)
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

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

\[ \langle L, \rho \rangle = \mathbb{E}_{(x, y) \sim \mathcal{D}, c \sim \rho} [\ell(y, c(x))] = \Pr_{(x, y) \sim \mathcal{D}, c \sim \rho} (c(x) \neq y) \]

\[ \langle \hat{L}, \rho \rangle = \Pr_{(x, y) \sim \mathcal{S}, c \sim \rho} (c(x) \neq y) \]
Seeger version of the bound

- We consider the 0-1 loss

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

\[
\langle L, \rho \rangle = \mathbb{E}_{(x,y) \sim D, c \sim \rho} [\ell(y, c(x))] = \Pr_{(x,y) \sim D, c \sim \rho} (c(x) \neq y)
\]

\[
\langle \hat{L}, \rho \rangle = \Pr_{(x,y) \sim S, c \sim \rho} (c(x) \neq y)
\]

- **Seeger’s PAC-Bayesian Theorem**  Fix an arbitrary \( D \), arbitrary prior \( \pi \), and confidence \( \delta \), then with probability at least \( 1 - \delta \) over samples \( S \sim D^m \), all posteriors \( \rho \) satisfy

\[
\text{kl}(\langle \hat{L}, \rho \rangle \| \langle L, \rho \rangle) \leq \frac{\text{KL}(\rho\|\pi) + \ln((m + 1)/\delta)}{m}
\]
Seeger version of the bound

- We consider the 0-1 loss

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

\[
\langle L, \rho \rangle = \mathbb{E}_{(x, y) \sim D, c \sim \rho} [\ell(y, c(x))] = Pr_{(x, y) \sim D, c \sim \rho}(c(x) \neq y)
\]

\[
\langle \hat{L}, \rho \rangle = Pr_{(x, y) \sim S, c \sim \rho}(c(x) \neq y)
\]

- **Seeger’s PAC-Bayesian Theorem**  
  Fix an arbitrary \( D \), arbitrary prior \( \pi \), and confidence \( \delta \), then with probability at least \( 1 - \delta \) over samples \( S \sim D^m \), all posteriors \( \rho \) satisfy

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

- Gives a tighter bound than PAC-Bayes-Hoeffding, particularly for small empirical error rates.
Linear classifiers

- We consider linear classifiers in a kernel $\kappa$ defined feature space:

$$\mathcal{F} = \{c_w : x \mapsto \text{sgn}(\langle w, \phi(x) \rangle)\}$$

where $\langle \phi(x), \phi(z) \rangle = \kappa(x, z)$.
Linear classifiers

- We consider linear classifiers in a kernel $\kappa$ defined feature space:

$$\mathcal{F} = \{ c_w : x \mapsto \text{sgn} (\langle w, \phi(x) \rangle) \}$$

where $\langle \phi(x), \phi(z) \rangle = \kappa(x, z)$.

- The mapping $\phi$ embeds the input space into a Hilbert space, which is usually specified by the kernel $\kappa$ satisfying the positive semi-definite property.
We consider linear classifiers in a kernel \( \kappa \) defined feature space:

\[
\mathcal{F} = \{ c_w : x \mapsto \text{sgn} ( \langle w, \phi(x) \rangle ) \}
\]

where \( \langle \phi(x), \phi(z) \rangle = \kappa(x, z) \).

The mapping \( \phi \) embeds the input space into a Hilbert space, which is usually specified by the kernel \( \kappa \) 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 $\kappa$ defined feature space:

$$\mathcal{F} = \{ c_w : x \mapsto \text{sgn}(\langle w, \phi(x) \rangle) \}$$

where $\langle \phi(x), \phi(z) \rangle = \kappa(x, z)$.

- The mapping $\phi$ embeds the input space into a Hilbert space, which is usually specified by the kernel $\kappa$ 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 $\pi$ 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 $\pi$ will be centered at the origin with unit variance.
- The specification of the centre for the posterior $\rho(w, \mu)$ will be by a unit vector $w$ and a scale factor $\mu$. 
PAC-Bayes Bound for SVM

- Prior $\pi$ 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 $\pi$ is Gaussian $\mathcal{N}(0,1)$
- Posterior is in the direction $w$
- at distance $\mu$ from the origin
PAC-Bayes Bound for SVM

- Prior $\pi$ is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the direction $\mathbf{w}$
- at distance $\mu$ from the origin
- Posterior $\rho$ is Gaussian
PAC-Bayes Bound for SVM

Linear classifiers performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

**Example**

\[
\langle L, \rho(w, \mu) \rangle \text{ true performance of the stochastic classifier}
\]

\[
\mathbb{E}_{c \sim \rho(w, \mu)}[c(x) \neq y]
\]
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

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

- SVM is deterministic classifier that exactly corresponds to
  \[ \text{sgn} \left( E_{c \sim \rho(w, \mu)}[c(x)] \right) \neq y \] 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

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi)}{m} + \ln \frac{m+1}{\delta}
\]

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

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]
Linear classifiers performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

\[
\langle \hat{L}, \rho(w, \mu) \rangle \text{ stochastic measure of the training error}
\]
Linear classifiers performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

- \langle \hat{L}, \rho(w, \mu) \rangle \text{ stochastic measure of the training error}
- \langle \hat{L}, \rho(w, \mu) \rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu \gamma(x_j, y_j))
Linear classifiers performance may be bounded by

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

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

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}
\]

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

Linear classifiers performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

- \langle \hat{L}, \rho(w, \mu) \rangle \text{ stochastic measure of the training error}
- \langle \hat{L}, \rho(w, \mu) \rangle = \frac{1}{m} \sum_{j=1}^{m} \tilde{F}(\mu \gamma(x_j, y_j))
- \text{where } \tilde{F}(\mu \gamma(x, y)) \text{ is probability of error of stochastic classifier on example } (x, y)
- \text{where } \gamma(x, y) = (y w^T \phi(x)) / (\|\phi(x)\|\|w\|)
- \text{and } \tilde{F}(t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu)\|\pi) + \ln \frac{m+1}{\delta}}{m}
\]
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi)}{m} + \ln \frac{m+1}{\delta}
\]

- Prior \( \pi \equiv \text{Gaussian centered on the origin} \)
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi)}{m} + \ln \frac{m+1}{\delta}
\]

- Prior \( \pi \equiv \text{Gaussian centered on the origin} \)
- Posterior \( \rho \equiv \text{Gaussian along } w \text{ at a distance } \mu \text{ from the origin} \)
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

- Prior \( \pi \equiv \text{Gaussian centered on the origin} \)
- Posterior \( \rho \equiv \text{Gaussian along } w \text{ at a distance } \mu \text{ from the origin} \)
- \( \text{KL}(\rho \| \pi) = \mu^2 / 2 \)
PAC-Bayes Bound for SVM

**Linear classifiers** performance may be bounded by

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

$\delta$ is the confidence

The bound holds with probability $1 - \delta$ over the random i.i.d. selection of the training data.
Linear classifiers performance may be bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle || \langle L, \rho(w, \mu) \rangle) \leq \frac{\text{KL}(\rho(w, \mu) || \pi) + \ln \frac{m+1}{\delta}}{m}
\]

\(\delta\) is the confidence


**PAC-Bayes Bound for SVM**

*Linear classifiers* performance may be bounded by

\[
kl(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{KL(\rho(w, \mu) \| \pi) + \ln \frac{m+1}{\delta}}{m}
\]

- $\delta$ 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 $\mu$ to optimise the bound
Form of the SVM bound

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

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

then have with probability at least \( 1 - \delta \)

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

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

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

<table>
<thead>
<tr>
<th>Problem</th>
<th>SVM</th>
<th>2FCV</th>
<th>10FCV</th>
<th>PAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>digits</td>
<td>Bound</td>
<td>–</td>
<td>–</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>waveform</td>
<td>Bound</td>
<td>–</td>
<td>–</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.090</td>
<td>0.086</td>
<td>0.084</td>
</tr>
<tr>
<td>pima</td>
<td>Bound</td>
<td>–</td>
<td>–</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.244</td>
<td>0.245</td>
<td>0.229</td>
</tr>
<tr>
<td>ringnorm</td>
<td>Bound</td>
<td>–</td>
<td>–</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.016</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>spam</td>
<td>Bound</td>
<td>–</td>
<td>–</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.066</td>
<td>0.063</td>
<td>0.067</td>
</tr>
</tbody>
</table>
Outline of the Tutorial

Part I

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

- 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

\[ \text{kl}((\hat{L}, \rho)\| (L, \rho)) \leq \frac{\text{KL}(\rho\|\pi) + \ln((m + 1)/\delta)}{m} \]

Relation

1. Explicit way to incorporate prior information (via \( \pi \))
Relation and Difference with Bayesian Learning

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

Relation

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

Difference

1. Explicit high-probability guarantee on the expected performance
Relation and Difference with Bayesian Learning

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

Relation

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

Difference

1. Explicit high-probability guarantee on the expected performance
2. No belief in prior correctness (frequentist bound)
Relation and Difference with Bayesian Learning

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

Relation

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

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

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

Relation

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

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

\[
\text{kl}(⟨\hat{L}, ρ⟩∥⟨L, ρ⟩) \leq \frac{\text{KL}(ρ∥π) + \ln((m + 1)/δ)}{m}
\]

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 \(π(h)\) vs. evidence \(\hat{L}(h)\)
5. Holds for any distribution \(ρ\) (including the Bayes posterior)
Outline of the Tutorial

Part I

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

- 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

- Learn the prior $\pi$ with part of the data

- Introduce the learnt prior in the bound

- Compute stochastic error with remaining data
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 $\pi$ 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 $\pi$ 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 $\pi$ 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* $w_r$. 

[Diagram showing a vector $w_r$ and a circle]
New prior for the SVM

- Solve SVM with **subset of patterns**
- Prior in the **direction** $w_r$
- **Posterior** like PAC-Bayes Bound
New prior for the SVM

- Solve SVM with subset of patterns
- Prior in the direction $w_r$
- Posterior like PAC-Bayes Bound
- New bound proportional to $\text{KL}(\rho||\pi)$
New Bound for the SVM

SVM performance may be tightly bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]
New Bound for the SVM

SVM performance may be tightly bounded by

$$\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \left(\frac{(m-r+1)J}{\delta}\right)}{m-r}$$

- $\langle L, \rho(w, \mu) \rangle$ true performance of the classifier
SVM performance may be **tightly** bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \left( \frac{(m-r+1)J}{\delta} \right)}{m-r}
\]
New Bound for the SVM

SVM performance may be tightly bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \left(\frac{(m-r+1)J}{\delta}\right)}{m - r}
\]

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

\[
\hat{\rho}(w, \mu)_S = \frac{1}{m - r} \sum_{j=r+1}^{m} \tilde{F}(\mu \gamma(x_j, y_j))
\]
New Bound for the SVM

SVM performance may be **tightly** bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]
New Bound for the SVM

SVM performance may be **tightly** bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]

\(\triangleright\) 0.5\(\| \mu w - \eta w_r \|^2\) distance between prior and posterior
New Bound for the SVM

SVM performance may be **tightly** bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \frac{(m-r+1)J}{\delta}}{m-r}
\]
New Bound for the SVM

SVM performance may be **tightly** bounded by

\[
\text{kl}(\hat{L}, \rho(w, \mu)) \| L, \rho(w, \mu) \rangle \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]

- Penalty term only dependent on the remaining data \( m - r \)
New Bound for the SVM

SVM performance may be **tightly** bounded by

\[
\text{kl}(\langle \hat{L}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \frac{(m-r+1)}{\delta} J}{m - r}
\]
SVM performance may be tightly bounded by

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

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

<table>
<thead>
<tr>
<th>Problem</th>
<th>Bound</th>
<th>2FCV</th>
<th>10FCV</th>
<th>PAC</th>
<th>PrPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>digits</td>
<td></td>
<td>–</td>
<td>–</td>
<td>0.175</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.014</td>
</tr>
<tr>
<td>waveform</td>
<td></td>
<td>–</td>
<td>–</td>
<td>0.203</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.090</td>
<td>0.086</td>
<td>0.084</td>
<td>0.088</td>
</tr>
<tr>
<td>pima</td>
<td></td>
<td>–</td>
<td>–</td>
<td>0.424</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.244</td>
<td>0.245</td>
<td>0.229</td>
<td>0.229</td>
</tr>
<tr>
<td>ringnorm</td>
<td></td>
<td>–</td>
<td>–</td>
<td>0.203</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.016</td>
<td>0.016</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td>spam</td>
<td></td>
<td>–</td>
<td>–</td>
<td>0.254</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.066</td>
<td>0.063</td>
<td>0.067</td>
<td>0.077</td>
</tr>
</tbody>
</table>
Prior-SVM

- New bound proportional to $\|\mu w - \eta w_r\|^2$
Prior-SVM

- New bound proportional to $\|\mu w - \eta w_r\|^2$
- Classifier that **optimises the bound**
Prior-SVM

- New bound proportional to \( \|\mu w - \eta w_r\|^2 \)
- Classifier that **optimises the bound**
- Optimisation problem to determine the \( p \)-SVM

\[
\begin{align*}
\min_{w, \xi_i} & \quad \left[ \frac{1}{2} \|w - w_r\|^2 + C \sum_{i=r+1}^{m} \xi_i \right] \\
\text{s.t.} \quad y_i w^T \phi(x_i) & \geq 1 - \xi_i & i = r + 1, \ldots, m \\
\xi_i & \geq 0 & i = r + 1, \ldots, m
\end{align*}
\]
Prior-SVM

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

$$
\min_{w, \xi_i} \left[ \frac{1}{2} \|w - w_r\|^2 + C \sum_{i=r+1}^{m} \xi_i \right]
\quad \text{s.t. } y_i w^T \phi(x_i) \geq 1 - \xi_i \quad i = r+1, \ldots, m
\quad \xi_i \geq 0 \quad i = r+1, \ldots, m
$$

- 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 $w_r$. 
Bound for p-SVM

1. Determine the prior with a subset of the training examples to obtain $w_r$
2. Solve p-SVM and obtain $w$
Bound for p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $w_r$

2. Solve **p-SVM** and obtain $w$

3. **Margin** for the stochastic classifier $c \sim \rho$

$$\gamma(x_j, y_j) = \frac{y_j w^T \phi(x_j)}{\|\phi(x_j)\| \|w\|} \quad j = r + 1, \ldots, m$$
Bound for p-SVM

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

$$\gamma(x_j, y_j) = \frac{y_j w^T \phi(x_j)}{\|\phi(x_j)\| \|w\|} \quad j = r + 1, \ldots, m$$

4. **Linear search** to obtain the optimal value of $\mu$. This introduces an insignificant extra penalty term.
Consider using a prior distribution $\pi$ that is elongated in the direction of $w_r$. 

$$\min_{v, \eta, \xi} \left[ \frac{1}{2} \|v\|_2^2 + C \sum_{i=r+1}^{m} \xi_i \right]$$

subject to

$$y_i (v + \eta w_r) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$ for $i = r+1, \ldots, m$
Consider using a prior distribution $\pi$ that is elongated in the direction of $w_r$.

This will mean that there is low penalty for large projections onto this direction.
Consider using a prior distribution $\pi$ that is elongated in the direction of $w_r$. This will mean that there is low penalty for large projections onto this direction. Translates into an optimisation:

$$\min_{v,\eta,\xi_i} \left[ \frac{1}{2} \|v\|^2 + C \sum_{i=r+1}^{m} \xi_i \right]$$
Consider using a prior distribution $\pi$ that is elongated in the direction of $w_r$.

This will mean that there is low penalty for large projections onto this direction.

Translates into an optimisation:

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

subject to

$$y_i(v + \eta w_r)^T \phi(x_i) \geq 1 - \xi_i, \quad i = r + 1, \ldots, m$$

$$\xi_i \geq 0, \quad i = r + 1, \ldots, m$$
Bound for $\eta$-prior-SVM

- Prior is elongated along the line of $w_r$ but spherical with variance 1 in other directions
Bound for \( \eta \)-prior-SVM

- Prior is elongated along the line of \( w_r \) but spherical with variance 1 in other directions
- Posterior again on the line of \( w \) at a distance \( \mu \) chosen to optimise the bound.
Bound for $\eta$-prior-SVM

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

$$ kl(\langle \hat{L}_{S_{m-r}}, \rho(w, \mu) \rangle \| \langle L, \rho(w, \mu) \rangle) \leq $$

$$ 0.5(\ln(\tau^2) + \tau^{-2} - 1 + P_{w_r}^{\parallel}(\mu w - w_r)^2/\tau^2 + P_{w_r}^{\perp}(\mu w)^2) + \ln\left(\frac{m-r+1}{\delta}\right) $$

$$ m - r $$
<table>
<thead>
<tr>
<th>Problem</th>
<th>Bound</th>
<th>2FCV</th>
<th>10FCV</th>
<th>PAC</th>
<th>PrPAC</th>
<th>PrPAC</th>
<th>τ-PrPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>digits</td>
<td>–</td>
<td>–</td>
<td>0.175</td>
<td>0.107</td>
<td>0.050</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.014</td>
<td>0.010</td>
<td>0.009</td>
</tr>
<tr>
<td>waveform</td>
<td>–</td>
<td>–</td>
<td>0.203</td>
<td>0.185</td>
<td>0.178</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.090</td>
<td>0.086</td>
<td>0.084</td>
<td>0.088</td>
<td>0.087</td>
<td>0.086</td>
</tr>
<tr>
<td>pima</td>
<td>–</td>
<td>–</td>
<td>0.424</td>
<td>0.420</td>
<td>0.428</td>
<td>0.416</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.244</td>
<td>0.245</td>
<td>0.229</td>
<td>0.229</td>
<td>0.233</td>
<td>0.233</td>
</tr>
<tr>
<td>ringnorm</td>
<td>–</td>
<td>–</td>
<td>0.203</td>
<td>0.110</td>
<td>0.053</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.016</td>
<td>0.016</td>
<td>0.018</td>
<td>0.018</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>spam</td>
<td>–</td>
<td>–</td>
<td>0.254</td>
<td>0.198</td>
<td>0.186</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.066</td>
<td>0.063</td>
<td>0.067</td>
<td>0.077</td>
<td>0.070</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Data distribution dependent prior

Consider the Gaussian prior centred on the weight vector:

\[ w_\pi = \mathbb{E}[y\phi(x)] \]
Consider the Gaussian prior centred on the weight vector:

$$w_\pi = \mathbb{E}[y\phi(x)]$$

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:

$$ w_\pi = \mathbb{E}[y\phi(x)] $$

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

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

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

- KL divergence is simple have the squared distance.

- With probability $1 - \delta/2$ we have

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

- We can therefore w.h.p. upper bound KL divergence between prior $\pi$, an isotropic Gaussian at $w_\pi$, and posterior $\rho$, an isotropic Gaussian at $w$ by

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

Giving the following bound on generalisation:

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

with probability \(1 - \delta\).

Values of the bounds for an SVM.

<table>
<thead>
<tr>
<th>Prob.</th>
<th>PAC-Bayes</th>
<th>PrPAC</th>
<th>(\tau)-PrPAC</th>
<th>(\mathbb{E}) PrPAC</th>
<th>(\tau-\mathbb{E}) PrPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>han</td>
<td>0.175</td>
<td>0.107</td>
<td>0.108</td>
<td>0.157</td>
<td>0.176</td>
</tr>
<tr>
<td>wav</td>
<td>0.203</td>
<td>0.185</td>
<td>0.184</td>
<td>0.202</td>
<td>0.205</td>
</tr>
<tr>
<td>pim</td>
<td>0.424</td>
<td>0.420</td>
<td>0.423</td>
<td>0.428</td>
<td>0.433</td>
</tr>
<tr>
<td>rin</td>
<td>0.203</td>
<td>0.110</td>
<td>0.110</td>
<td>0.201</td>
<td>0.204</td>
</tr>
<tr>
<td>spa</td>
<td>0.254</td>
<td>0.198</td>
<td>0.198</td>
<td>0.249</td>
<td>0.255</td>
</tr>
</tbody>
</table>
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 \((x, y) \in \mathcal{X} \times \{-1, +1\}\), is drawn iid acc. to \(D\).
Definitions often related to PAC-Bayes bound in supervised learning

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

\[
R(h) \overset{\text{def}}{=} \mathbb{E}_{(x, y) \sim D} I(h(x) \neq y) \quad ; \quad R_S(h) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i).
\]

where \(I(y' \neq y)\) is the so called 0 – 1 loss.
Definitions often related to PAC-Bayes bound in supervised learning

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

\[
R(h) \overset{\text{def}}{=} \mathbb{E}_{(x, y) \sim D} I(h(x) \neq y) ; \quad R_S(h) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i).
\]

where \(I(y' \neq y)\) is the so called \(0-1\) loss.
- The learner’s goal is to choose a **posterior distribution** \(\rho\) on a space \(\mathcal{H}\) of hypothesis such that the risk of the \(\rho\)-weighted majority vote \(B_\rho\) is as small as possible.
Definitions often related to PAC-Bayes bound in supervised learning

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

\[
R(h) \overset{\text{def}}{=} \mathbb{E}_{(x,y) \sim D} I(h(x) \neq y) ; \quad R_S(h) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i).
\]

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

\[
B_\rho(x) \overset{\text{def}}{=} \text{sgn} \left[ \mathbb{E}_{h \sim \rho} h(x) \right]
\]
Definitions often related to PAC-Bayes bound in supervised learning

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

\[
R(h) \overset{\text{def}}{=} \mathbb{E}_{(x,y)\sim D} I(h(x) \neq y) ; \quad R_S(h) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i) .
\]

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

\[
B_{\rho}(x) \overset{\text{def}}{=} \text{sgn} \left[ \mathbb{E}_{h \sim \rho} h(x) \right]
\]

- $B_{\rho}$ is also called the Bayes classifier.
PAC-Bayes approach does not directly bounds the risk of $B_\rho$. 

The Gibbs classifier $G_\rho$ draws $h$ from $H$ according to $\rho$, and predicts $h(x)$. The risk and the training error of $G_\rho$ are thus defined as:

$$R(G_\rho) = \mathbb{E}_{h \sim \rho} R(h)$$

$$R_S(G_\rho) = \mathbb{E}_{h \sim \rho} R_S(h)$$
The Gibbs classifier

- PAC-Bayes approach does not directly bounds the risk of $B_\rho$
- It bounds the risk of the **Gibbs classifier** $G_\rho$: 
The Gibbs classifier

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

It bounds the risk of the **Gibbs classifier** $G_\rho$:

- to predict the label of $x$, $G_\rho$ draws $h$ from $\mathcal{H}$ according to $\rho$, and predicts $h(x)$

The risk and the training error of $G_\rho$ are thus defined as:

$$R(G_\rho) = \mathbb{E}_{h \sim \rho} R(h) ; \quad R_S(G_\rho) = \mathbb{E}_{h \sim \rho} R_S(h).$$
If $B_\rho$ misclassifies $x$, then at least half of the hypothesis (under measure $\rho$) err on $x$. 
If $B_\rho$ misclassifies $x$, then at least half of the hypothesis (under measure $\rho$) err on $x$.

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

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)*

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

or in the context of this tutorial:

\[
\langle f, \rho \rangle \leq \text{KL}(\rho\|\pi) + \ln \langle e^f, \pi \rangle
\]
History

- Pre-pre-history: Variational Definition of $\text{KL}$-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 $\pi$ of support $\mathcal{H}$, any $\delta \in (0, 1]$, we have

$$
\Pr_{S \sim D^m} \left( \forall \rho \text{ on } \mathcal{H}: \frac{1}{2} (R_S(G_\rho) - R(G_\rho))^2 \leq \frac{1}{m} \left[ \text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \geq 1 - \delta
$$
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 $\pi$ of support $\mathcal{H}$, any $\delta \in (0, 1]$, we have

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

or

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

---

**Seeger Bound**

For any $D$, any $H$, any $\pi$ of support $H$, any $\delta \in (0, 1]$, we have

$$\Pr_{S \sim D} \left( \forall \rho : \text{KL}(\rho \| \pi) + \ln \frac{1}{\sqrt{m} \delta} \leq 1 - \delta \right),$$

where $\text{KL}(q \| p) \triangleq q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}$. 
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 $\pi$ of support $\mathcal{H}$, any $\delta \in (0, 1]$, we have

$$\Pr_{S \sim D^n} \left( \forall \rho \text{ on } \mathcal{H}: \right. \\
\left. \text{kl}(R_S(G_\rho)\|R(G_\rho)) \leq \frac{1}{m} \left[ \text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \right) \geq 1 - \delta,$$

where

$$\text{kl}(q\|p) \overset{\text{def}}{=} q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}.$$
Graphical illustration of the Seeger bound

\[
\text{kl}(0.1||R(Q))
\]
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)

- 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); 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)

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); 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 $k_l$ 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 $k_l$ 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

$$B_w(x) = \text{sgn}\left[ \mathbb{E}_{i \sim w} k(x_i, x) \right]$$

dealing with the hypothesis being here $h_i(\cdot) = k(x_i, \cdot)$. Problem:

• Recall once more that the prior is not allowed to depend on the training set.
• How a prior on a set of hypothesis can be data-independent?
• The trick: put a prior on the possible ways that hypothesis can be constructed when given the data.
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

\[ B_w(x) = \text{sgn} \left[ \mathbb{E}_{i \sim w} k(x_i, x) \right] \]

the hypothesis being here \( h_i(\cdot) = k(x_i, \cdot) \).

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

\[ B_w(x) = \text{sgn}\left[ \mathbb{E}_{i \sim w} k(x_i, x) \right] \]

the hypothesis being here \( h_i(\cdot) = k(x_i, \cdot) \).

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

\[ B_w(x) = \text{sgn} \left[ \mathbf{E}_{i \sim w} k(x_i, x) \right] \]

the hypothesis being here \( h_i(\cdot) = k(x_i, \cdot) \).

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 $\rho$ that minimizes the bound.

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

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

Catoni’s bound

\[
\Pr_{S \sim D^m} \left( \forall \rho \text{ on } \mathcal{H}: R(G_{\rho}) \leq \frac{1}{1-e^{-C}} \left\{ 1 - \exp\left[ -\left( C \cdot R_S(G_{\rho}) + \frac{1}{m} \left[ \text{KL}(\rho\|\pi) + \ln \frac{1}{\delta} \right] \right) \right\} \right) \geq 1 - \delta.
\]
When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior $\rho$ that minimizes the bound.

Catoni’s bound

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

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 $\rho$ that minimizes the bound.

Catoni’s bound

\[
\Pr_{S \sim D^n} \left( \forall \rho \text{ on } \mathcal{H}: R(G_\rho) \leq \frac{1}{1-e^{-c}} \left\{ 1 - \exp \left[ -\left( C \cdot R_S(G_\rho) + \frac{1}{m} \left[ KL(\rho\|\pi) + \ln \frac{1}{\delta} \right] \right) \right] \right\} \right) \geq 1 - \delta.
\]

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

- **KL-Regularized Adaboost**  Germain et al. (2009b)
Not only SVM has been rediscovered as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost**  \textit{Germain et al. (2009b)}
- **Kernel Ridge Regression**  \textit{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)
Not only SVM has been rediscovered 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:

\[ L(\rho) \leq \hat{L}(\rho) + \sqrt{\text{KL}(\rho\|\pi) + \ln \frac{\xi(m)}{\delta}} \cdot \frac{1}{2m} \]
What is a localized PAC-Bayesian bound?

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

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

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.)

Hence, the bound expresses a tradeoff to be followed for finding suitable choices of the posterior distribution \( \rho \).
What is a localized PAC-Bayesian bound?

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

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

- Hence, the bound expresses a tradeoff to be followed for finding suitable choices of the posterior distribution \( \rho \).
- 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:

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

- Hence, the bound expresses a tradeoff to be followed for finding suitable choices of the posterior distribution \( \rho \).
- 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:

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

- Hence, the bound expresses a tradeoff to be followed for finding suitable choices of the posterior distribution \( \rho \).
- 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:

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

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

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

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

- 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.


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, $\pi$ 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, $\pi$ will have to remain unknown!
  - But may be we can manage to nevertheless estimate $\text{KL}(\rho||\pi)$. 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, \( \pi \) will have to remain unknown!
  ▶ But maybe we can manage to nevertheless estimate \( \text{KL}(\rho||\pi) \).
    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, $\pi$ will have to remain unknown!
  ▶ But may be we can manage to nevertheless estimate $\text{KL}(\rho||\pi)$. 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: $w_p = E_{(x,y) \sim D}(y \varphi(x))$. 
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, $\pi$ will have to remain unknown!
  - But may be we can manage to nevertheless estimate $\text{KL}(\rho\|\pi)$. 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: $w_p = \mathbb{E}_{(x,y) \sim D} (y \varphi(x))$.

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, $\pi$ will have to remain unknown!
  - But may be we can manage to nevertheless estimate $\text{KL}(\rho||\pi)$. 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: $w_p = \mathbb{E}_{(x,y) \sim D} (y \phi(x))$.


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:

$$\text{KL}(\rho \| \pi) \leq \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}.$$
Localized PAC-Bayesian bounds:

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

The KL-term is bounded as follows:

\[
\text{KL}(\rho \| \pi) \leq \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}.
\]

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:

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

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:

\[
\text{KL}(\rho \| \pi) \leq \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}.
\]

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 \( \pi \) of support \( \mathcal{H} \), any \( \delta \in (0, 1] \), we have

\[
\Pr\left(\forall \rho \text{ on } \mathcal{H}: \text{kl}(R_S(G_{\rho}), R(G_{\rho})) \leq \frac{1}{m} \left[ \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{2\xi(m)}{\delta/2}} + \frac{\gamma^2}{4m} + \ln \frac{\xi(m)}{\delta/2} \right] \right) \geq 1 - \delta.
\]
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 $\rho$ is **aligned** on $\pi$ 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 $\rho$, there is a posterior $\rho'$, aligned on $\pi$ such that $B_\rho(x) = B_{\rho'}(x)$. So, same classification capacity if one restrict itself to aligned posterior. But then, the KL-term vanishes from the PAC-Bayesian bound !!! MAGIC !!!
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 $\rho$ is aligned on $\pi$ 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 $\rho$ is aligned on $\pi$ 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 $\rho$, there is a posterior $\rho'$, aligned on $\pi$ such that

$$B_\rho(x) = B_{\rho'}(x).$$
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 $\rho$ is **aligned** on $\pi$ 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 $\rho$, there is a posterior $\rho'$, aligned on $\pi$ such that

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

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 $\rho$ is **aligned** on $\pi$ 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 $\rho$, there is a posterior $\rho'$, aligned on $\pi$ such that

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

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 $\rho$ is **aligned** on $\pi$ 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 $\rho$, there is a posterior $\rho'$, aligned on $\pi$ such that

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

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)

KL(\(\rho\|\pi\)) arises when transforming the expectation over \(\pi\) to the expectation over \(\rho\):

\[
\ln \left[ \mathbb{E}_{h \sim \pi} e^{m \cdot 2 (R_S(h) - R(h))^2} \right] \\
\geq \ln \left[ \mathbb{E}_{h \sim \rho} \frac{\pi(h)}{\rho(h)} e^{m \cdot 2 (R_S(h) - R(h))^2} \right] \\
\geq \mathbb{E}_{h \sim \rho} \ln \left[ \frac{\pi(h)}{\rho(h)} e^{m \cdot 2 (R_S(h) - R(h))^2} \right] \\
= m \mathbb{E}_{h \sim \rho} 2 (R_S(h) - R(h))^2 - KL(\rho\|\pi)
\]
Absence of KL for Aligned Posteriors

**General theorem (McAllester)**

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

\[
\ln \left[ \mathbf{E}_{h \sim \pi} e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
\geq \ln \left[ \mathbf{E}_{h \sim \rho} \frac{\pi(h)}{\rho(h)} e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
\geq \mathbf{E}_{h \sim \rho} \ln \left[ \frac{\pi(h)}{\rho(h)} e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
= m \mathbf{E}_{h \sim \rho} 2(R_S(h) - R(h))^2 - \text{KL}(\rho \| \pi)
\]

**Aligned posterior theorem**

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

\[
\ln \left[ \mathbf{E}_{h \sim \pi} e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
= \ln \left[ \mathbf{E}_{h \sim \rho} e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
\geq \mathbf{E}_{h \sim \rho} \ln \left[ e^{m \cdot 2(R_S(h) - R(h))^2} \right] \\
= m \mathbf{E}_{h \sim \rho} 2(R_S(h) - R(h))^2
\]
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 \( h \in \mathcal{H}_1 : -h \in \mathcal{H}_2 \).

\[
\mathbb{E}_{h \sim \pi} e^{m \cdot 2(R_S(h) - R(h))^2} = \int_{h \in \mathcal{H}_1} d\pi(h) e^{m \cdot 2(R_S(h) - R(h))^2} + \int_{h \in \mathcal{H}_2} d\pi(h) e^{m \cdot 2(R_S(h) - R(h))^2}
\]

\[
= \int_{h \in \mathcal{H}_1} d\pi(h) e^{m \cdot 2(R_S(h) - R(h))^2} + \int_{h \in \mathcal{H}_1} d\pi(-h) e^{m \cdot 2((1 - R_S(h)) - (1 - R(h)))^2}
\]

\[
= \int_{h \in \mathcal{H}_1} d\pi(h) e^{m \cdot 2(R_S(h) - R(h))^2} + \int_{h \in \mathcal{H}_1} d\pi(-h) e^{m \cdot 2(R_S(h) - R(h))^2}
\]

\[
= \int_{h \in \mathcal{H}_1} (d\pi(h) + d\pi(-h)) e^{m \cdot 2(R_S(h) - R(h))^2}
\]

\[
= \int_{h \in \mathcal{H}_1} (d\rho(h) + d\rho(-h)) e^{m \cdot 2(R_S(h) - R(h))^2}
\]

\[
\vdots
\]

\[
= \mathbb{E}_{h \sim \rho} e^{m \cdot 2(R_S(h) - R(h))^2}.
\]
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.


Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for


Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of


John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over
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
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.

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

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i$</td>
<td>0.1</td>
<td>0.3</td>
<td>...</td>
<td>0.2</td>
</tr>
<tr>
<td>$m_i$</td>
<td>12</td>
<td>24</td>
<td>...</td>
<td>19</td>
</tr>
<tr>
<td>$\hat{p}_i = m_i/m$</td>
<td>12/100</td>
<td>24/100</td>
<td>...</td>
<td>19/100</td>
</tr>
</tbody>
</table>
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 $h : \mathcal{X} \rightarrow \mathcal{Z}$
- $p_h$ - 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 $\rho$ simultaneously:

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

- $\mathcal{X}$ - sample space
- $\mathcal{H}$ - hypothesis space
- $\mathcal{Z}$ - finite, $|\mathcal{Z}| = K$
- Each $h \in \mathcal{H}$ is a mapping $h : \mathcal{X} \rightarrow \mathcal{Z}$
- $p_h$ - 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 $\rho$ simultaneously:

$$KL(\langle \hat{p}_h, \rho \rangle || \langle p_h, \rho \rangle) \leq \frac{KL(\rho || \pi) + (K - 1) \ln(m + 1) + \ln \frac{1}{\delta}}{m}$$
Application Example: Density Estimation with Co-clustering

Input
Sample \((X_1^1, X_2^1), \ldots, (X_m^1, X_m^2)\)

Goal
Build an estimator \(\rho(x^1, x^2)\) that minimizes
\[-E_{p(X^1, X^2)} [\ln \rho(X^1, X^2)]\]
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 $(X_1^1, X_2^1), \ldots, (X_m^1, X_m^2)$

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

\[ \rho(x^1, x^2) = \sum_{c^1, c^2} \tilde{\rho}(c^1, c^2) \prod_{i=1}^{2} \frac{\tilde{p}(x^i)}{\tilde{p}_{\rho}(c^i)} \rho(c^i|x^i) \]
Application Example: Density Estimation with Co-clustering

Bound
\[ W.p. \geq 1 - \delta: \]
\[ -\mathbb{E}_{\rho(x^1,x^2)} \left[ \ln \rho(X^1, X^2) \right] \]
\[ \leq \left( \sum_{i=1}^{2} \hat{H}(X^i) \right) - \hat{I}_\rho(C^1; C^2) + \ln(||C^1||||C^2||) \sqrt{\frac{\sum_i |X^i|I_\rho(X^i; C_i) + \ldots}{2m}} + \ldots \]

Approximation by product of marginals

Added value of clustering

Complexity of clustering

\[ \hat{I}_\rho(C^1; C^2) = 0 \]
\[ I_\rho(X^i; C^i) = 0 \]

\[ \hat{I}_\rho(C^1; C^2) = \hat{I}(X^1; X^2) \]
\[ I_\rho(X^i; C^i) = \ln |X^i| \]
Further Reading

Discrete Density Estimation
- Graph clustering
- Topic models

Continuous Density Estimation
- 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

\( Z_1, \ldots, Z_n \) is a \textit{martingale difference sequence} if

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

Martingale

Let

\[ M_j = \sum_{i=1}^{j} Z_i \]

then \( M_1, \ldots, M_n \) is a martingale.

Examples

- Random walk
- Gambler’s capital
PAC-Bayesian Inequalities for Martingales

Example: Capital of multiple gamblers in a zero-sum game
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 $\lambda \in [0, \frac{1}{C}]$:

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

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

"certain technical condition"

we have

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

Game round

- 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
Applications

- Online advertisement
- Medical (and other) experiment design
- Adaptive routing
- ...

Exploration-exploitation trade-off

- Let $\hat{a}_t^*$ be empirically best action at time $t$
- Should we play $\hat{a}_t^*$ at round $t + 1$ or try another $a$?
<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>...</th>
<th>$a_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p(r</td>
<td>a_i, s_j)$</td>
<td></td>
</tr>
<tr>
<td>$s_N$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Setting

- $S$ - a set of states
- Each state corresponds to a multiarmed bandit
- States are drawn according to a fixed distribution $p(s)$
Importance Weighted Sampling
In Multiarmed Bandits

Define pseudo-rewards

\[ R_t^a = \begin{cases} \frac{1}{\rho_t(a)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise} \end{cases} \]
Importance Weighted Sampling
In Multiarmed Bandits

Define pseudo-rewards

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

$R_t^a$ is an unbiased estimate of $r(a)$

$$\mathbb{E}[R_t^a | \text{game history}] = \rho_t(a) \left( \frac{1}{\rho_t(a)} \mathbb{E}[R_t | \text{game history, } A_t = a] \right) + 0 = r(a)$$
Importance Weighted Sampling
In Multiarmed Bandits

Define pseudo-rewards

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

\( R_t^a \) is an unbiased estimate of \( r(a) \)

\[ \mathbb{E}[R_t^a | \text{game history}] = \rho_t(a) \left( \frac{1}{\rho_t(a)} \mathbb{E}[R_t | \text{game history}, A_t = a] \right) + 0 \]
\[ = r(a) \]

Martingales

\( (R_1^a - r(a)), (R_2^a - r(a)), \ldots \) is a martingale difference sequence
Variance of Importance Weighted Sampling

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

\[ \mathbb{E}[R^a_t | \text{game history}] = r(a) \]

Variance

\[ \mathbb{E} \left[ (R^a_t - r(a))^2 | \text{game history} \right] \leq \frac{1}{\rho_t(a)} \]
Multiarmed Bandits with Side Information

Hypothesis Space

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

Example:

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_2$</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>$s_4$</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>
Multiarmed Bandits with Side Information

**Game Round**

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$\ldots$</th>
<th>$a_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_N$</td>
<td>$p(r</td>
<td>a_i, s_j)$</td>
<td></td>
</tr>
</tbody>
</table>
Multiarmed Bandits with Side Information

Game Round

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>...</th>
<th>$a_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>$p(r</td>
<td>a_i, s_j)$</td>
<td></td>
</tr>
<tr>
<td>$s_N$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $R_t \sim p(r|A_t, S_t)$. 
Multiarmed Bandits with Side Information

Importance-Weighted Rewards

\[ R_{t}^{a,S_t} = \begin{cases} \frac{1}{\rho_t(a|S_t)} R_t, & \text{if } A_t = a \\ 0, & \text{otherwise.} \end{cases} \]
Multiarmed Bandits with Side Information

Importance-Weighted Rewards

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

\[ \hat{R}_t(h) = \sum_{i=1}^{t} R_i^{h(S_i),S_i} \]
Multiarmed Bandits with Side Information

Importance-Weighted 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

\[ \Delta(h) = R(h^*) - R(h) \]
\[ \hat{\Delta}_t(h) = \hat{R}_t(h^*) - \hat{R}_t(h). \]
Multiarmed Bandits with Side Information

Importance-Weighted 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

\[
|\langle M_n, \rho \rangle| \lesssim \sqrt{\langle V_n, \rho \rangle \left(\text{KL}(\rho \| \pi) + \ln \frac{1}{\delta}\right)}
\]
PAC-Bayesian Regret Bound

Reminder: PAC-Bayes-Bernstein Inequality for Martingales

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

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

Pick a combinatorial prior $\pi$ over $\mathcal{H}$, then:

$\text{KL}(\rho || \pi) \leq NI_\rho(S; A) + K \ln N + K \ln K$
Reminder: PAC-Bayes-Bernstein Inequality for Martingales

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

**Treating KL(\rho \| \pi)**

Pick a combinatorial prior \( \pi \) over \( \mathcal{H} \), then:

\[ \text{KL}(\rho \| \pi) \leq NI_\rho(S; A) + K \ln N + K \ln K \]

**Treating \( \langle V_n, \rho \rangle \)**

Smooth the playing strategies for all \( t < n \) by \( \varepsilon \).
PAC-Bayesian Regret Bound

\[ \langle \Delta, \rho_n \rangle = \frac{1}{n} \langle \left( n\Delta - \hat{\Delta}_n \right), \rho_n \rangle + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle \]

\[ \leq \sqrt{\langle V_n, \rho_n \rangle \left( N I_{\rho_n}(S; A) + K \ln N + \ldots \right) \ldots} + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle \]

- **Martingales**
- **Policy complexity**
- **Empirical Performance**
PAC-Bayesian Regret Bound

\[
\langle \Delta, \rho_n \rangle = \frac{1}{n} \langle \left( n\Delta - \hat{\Delta}_n \right), \rho_n \rangle + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle
\]

- Martingales

\[
\leq \sqrt{\langle V_n, \rho_n \rangle (NI_{\rho_n}(S; A) + K \ln N + ...)} + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle
\]

- Policy complexity

- Empirical Performance

Remarks

\[
0 \leq NI_{\rho_n}(S; A) \leq N \ln K
\]
PAC-Bayesian Regret Bound

\[ \langle \Delta, \rho_n \rangle = \frac{1}{n} \left\langle \left( n\Delta - \hat{\Delta}_n \right), \rho_n \right\rangle + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle \]

Martingales

\[ \leq \sqrt{\langle V_n, \rho_n \rangle \left( N I_{\rho_n} (S; A) + K \ln N + \ldots \right) \ldots} + \frac{1}{n} \langle \hat{\Delta}_n, \rho_n \rangle \]

Policy complexity

Empirical Performance

Remarks

\[ 0 \leq N I_{\rho_n} (S; A) \leq N \ln K \]

\[ \ln |\mathcal{H}| = \ln \left( K^N \right) = N \ln K \]
## Experiments

### Setting

#### Experiment 1

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>...</th>
<th>$a_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{100}$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

Setting

### Experiment 1

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>...</th>
<th>$a_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{100}$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

### Experiment 2

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>...</th>
<th>$a_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{33}$</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{66}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{100}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

Setting

Experiment 1

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>...</th>
<th>$a_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{100}$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

Experiment 2

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>...</th>
<th>$a_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{33}$</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{66}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$s_{100}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

Experiment 3

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

Experiment 4

$H(A^h) = \ln(20) \approx 4$
Experiments - Regret Graph

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

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

$H(A^{h^*}) = 2$

$H(A^{h^*}) = 3$

Baseline
Experiments - Bound

Bound on $\Delta(\rho_t^{\exp})$

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

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

$H(A^{h^*}) = 2$

$H(A^{h^*}) = 3$
Experiments - Mutual Information

\[ I(\rho_t) = (S;A) \]

\[
H(A^{h^*}) = \begin{align*}
0 \\
1 \\
2 \\
3
\end{align*}
\]

\[ x 10^6 \\
0 \\
0.5 \\
1 \\
1.5 \\
2 \\
2.5 \]

\[ t \]

\[ I(\rho_t) = (S;A) \]

\[
H(A^{h^*}) = \begin{align*}
0 \\
1 \\
2 \\
3
\end{align*}
\]

\[ x 10^6 \\
0 \\
0.5 \\
1 \\
1.5 \\
2 \\
2.5 \]

\[ t \]
Further Reading


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

\[ \langle f, \rho \rangle \leq KL(\rho\|\pi) + \ln\langle e^f, \pi \rangle \]

- Design a hypothesis space \( \mathcal{H} \)
- Design a reference measure \( \pi \) 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 \( \rho \)
- Bound \( KL(\rho\|\pi) \)
- Combine everything together
\[ \langle f, \rho \rangle \leq \text{KL}(\rho \| \pi) + \ln \langle e^f, \pi \rangle \]

### Choice of \( f \)

- **PAC-Bayes-Hoeffding**
  \[ f(h) = \lambda (L(h) - \hat{L}(h)) \]

- **PAC-Bayes-kl**
  \[ f(h) = n \text{kl}(\hat{L}(h) \| L(h)) \]

- **PAC-Bayes-Bernstein**
  \[ f(h) = \lambda (\hat{L}(h) - L(h)) - (e - 2)\lambda^2 V_n(h) \]

- **PAC-Bayes-KL**
  \[ f(h) = n \text{KL}(\hat{p}(h) \| p(h)) \]

- **Martingales**
  
  ...  

### Choice of \( \pi \)

- **Combinatorial**
  \[ \text{KL}(\rho \| \pi) \leq I_\rho(X; C) \]

- **Gaussian**
  \[ \text{KL}(\rho \| \pi) \leq \|w\|_2 \]

- **Laplacian**
  \[ \text{KL}(\rho \| \pi) \leq \|w\|_1 \]

- **Distribution-Dependent**
  \[ \text{KL}(\rho \| \pi) \leq \gamma \sqrt{\ln(\ldots)/m} + \frac{\gamma^2}{4m} \]

...
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