Generalization of the PAC-Bayesian Theory

and Applications to Semi-Supervised Learning

Pascal Germain

INRIA Paris (SIERRA Team)

Modal Seminar

INRIA Lille

January 24, 2017

Dans la vie, l'essentiel est de porter sur tout des jugements a priori.

— Boris Vian

Plan

Plan

Definitions

Learning example

An example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a **description-label** pair.

Data generating distribution

Each example is an **observation from distribution** D on $\mathcal{X} \times \mathcal{Y}$.

Learning sample

$$S \;=\; \{\, (x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n) \,\} \;\sim\; D^n$$

Predictors (or hypothesis)

$$h : \mathcal{X} \to \mathcal{Y}, \quad h \in \mathcal{H}$$

Learning algorithm

$$A(S) \longrightarrow h$$

Loss function

$$\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Empirical loss

$$\widehat{\mathcal{L}}_S^{\,\ell}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, x_i, y_i)$$

Generalization loss

$$\mathcal{L}_D^{\,\ell}(h) = \mathop{\mathbf{E}}_{(x,y)\sim D} \ell(h, x, y)$$

PAC-Bayesian Theory

Initiated by McAllester (1999), the PAC-Bayesian theory gives **PAC** generalization guarantees to "**Bayesian** like" algorithms.

PAC guarantees (Probably Approximately Correct)

With probability at least "$1-\delta$", the loss of predictor h is less than "ε"

$$\Pr_{S\sim D^n}\left(\mathcal{L}_D^\ell(h) \leq \varepsilon(\widehat{\mathcal{L}}_S^\ell(h), n, \delta, \ldots)\right) \geq 1-\delta$$

Bayesian flavor

Given:

- A **prior** distribution P on $\mathcal{H}$.
- A **posterior** distribution Q on $\mathcal{H}$.

$$\Pr_{S\sim D^n}\left(\mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^\ell(h) \leq \varepsilon(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^\ell(h), n, \delta, P, \ldots)\right) \geq 1-\delta$$

A *Classical* PAC-Bayesian Theorem

PAC-Bayesian theorem (adapted from McAllester 1999, 2003)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set of predictors $\mathcal{H}$, for any loss $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, we have,

$$\Pr_{S \sim D^n}\left(\forall\, Q \text{ on } \mathcal{H} : \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell}(h) \leq \mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}_S^{\ell}(h) + \sqrt{\frac{1}{2n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]}\right) \geq 1-\delta\,,$$

where $\mathrm{KL}(Q\|P) = \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}$ is the **Kullback-Leibler divergence**.

Training bound

- Gives generalization guarantees **not based on testing sample**.

Valid *for all* posterior Q on $\mathcal{H}$

- Inspiration for conceiving **new learning algorithms**.

Plan

Plan

Majority Vote Classifiers

Consider a binary classification problem, where $\mathcal{Y} = \{-1, +1\}$ and the set $\mathcal{H}$ contains **binary voters** $h : \mathcal{X} \to \{-1, +1\}$

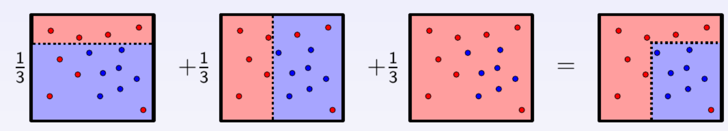
Weighted majority vote

To predict the label of $x \in \mathcal{X}$, the classifier asks for the *prevailing opinion*

$$B_Q(x) \;=\; \operatorname{sgn}\left(\mathop{\mathbf{E}}_{h\sim Q} h(x) \right)$$

Many learning algorithms output majority vote classifiers

AdaBoost, Random Forests, Bagging, ...



A Surrogate Loss

Majority vote risk

$$R_D(B_Q) \;=\; \Pr_{(x,y)\sim D}\Big(B_Q(x) \neq y\Big) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D} \mathrm{I}\Big[\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x) \leq 0\Big]$$

where $\mathrm{I}[a] = 1$ if predicate a is *true*; $\mathrm{I}[a] = 0$ otherwise.

Gibbs Risk / Linear Loss

The stochastic Gibbs classifier $G_Q(x)$ draws $h' \in \mathcal{H}$ according to Q and output $h'(x)$.

$$R_D(G_Q) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D}\mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\Big[h(x) \neq y\Big]$$

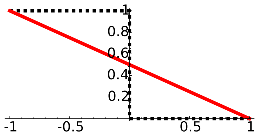
$$\;=\; \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell_{01}}(h)\,,$$

where $\ell_{01}(h,x,y) = \mathrm{I}\big[h(x) \neq y\big]$.

Factor two

It is well-known that

$$R_D(B_Q) \leq 2 \times R_D(G_Q)$$



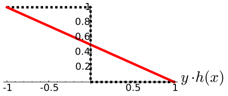
$\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x)$

From the *Factor 2* to the $\mathcal{C}$-bound

From Markov's inequality ($\Pr(X \geq a) \leq \frac{\mathbf{E}\,X}{a}$), we obtain:

Factor 2 bound

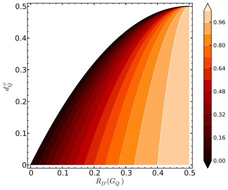
$$\begin{aligned} R_D(B_Q) &= \Pr_{(x,y)\sim D}\big(1 - y\cdot h(x) \geq 1\big) \\ &\leq \mathop{\mathbf{E}}_{(x,y)\sim D}\big(1 - y\cdot h(x)\big) \;=\; 2\,R_D(G_Q)\,. \end{aligned}$$



From Chebyshev's inequality ($\Pr(X - \mathbf{E}\,X \geq a) \leq \frac{\mathbf{Var}\,X}{a^2 + \mathbf{Var}\,X}$), we obtain:

The $\mathcal{C}$-bound (Lacasse et al., 2006)

$$R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;=\; 1 - \frac{\big(1 - 2\cdot R_D(G_Q)\big)^2}{1 - 2\cdot d_Q^D}$$



where d_Q^D is the **expected disagreement**:

$$d_Q^D \;=\; \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\; \mathop{\mathbf{E}}_{h_i\sim Q}\; \mathop{\mathbf{E}}_{h_j\sim Q} \mathrm{I}\Big[h_i(x) \neq h_j(x)\Big] \;=\; \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\left[\mathop{\mathbf{E}}_{h\sim Q} h(x)\right]^2\right).$$

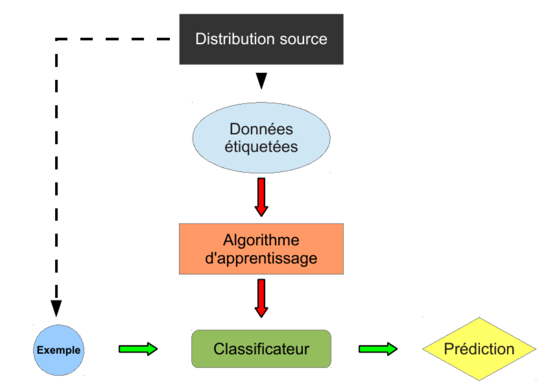
Plan

I.I.D. Assumption

Assumption

Examples are generated *i.i.d.* by a distribution D on $\mathcal{X} \times \mathcal{Y}$.

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \sim D^n$$



A General PAC-Bayesian Theorem

Δ-function: «distance» between $\widehat{R}_S(G_Q)$ et $R_D(G_Q)$

Convex function $\Delta : [0,1] \times [0,1] \to \mathbb{R}$.

General theorem (Bégin et al. 2014, 2016; Germain 2015)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of voters, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H}: \quad \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(n)}{\delta}\right],$$

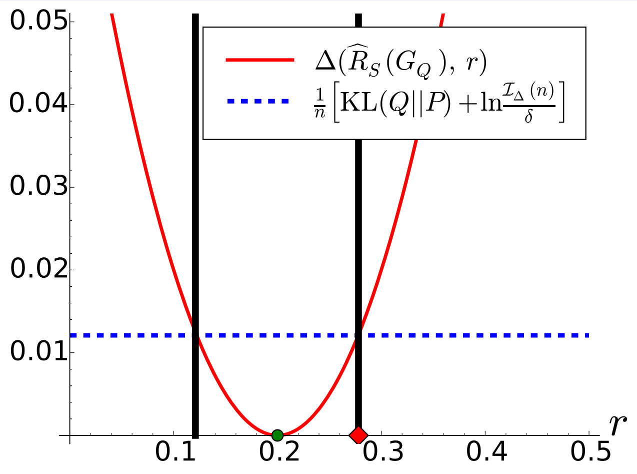
where

$$\mathcal{I}_\Delta(n) \;=\; \sup_{r\in[0,1]} \left[\, \sum_{k=0}^{n} \underbrace{\binom{n}{k} r^k (1-r)^{n-k}}_{\mathbf{Bin}\left(k;n,r\right)} e^{n\Delta\left(\frac{k}{n},\, r\right)} \right].$$

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \ \ge\ 1-\delta\,.$$

Interpretation.



General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Proof ideas.

Change of Measure Inequality

For any P and Q on $\mathcal{H}$, and for any measurable function $\phi:\mathcal{H}\to\mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\phi(h) \;\leq\; \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Markov's inequality

$$\Pr\left(X \geq a\right) \leq \tfrac{\mathbf{E}\,X}{a} \quad\Longleftrightarrow\quad \Pr\left(X \leq \tfrac{\mathbf{E}\,X}{\delta}\right) \geq 1-\delta\,.$$

Probability of observing k misclassifications among n examples

Given a voter h, consider a **binomial variable** of n trials with **success** $\mathcal{L}_D^{\ell_{01}}(h)$:

$$\begin{aligned}\Pr_{S\sim D^n}\left(\widehat{\mathcal{L}}_S^{\ell_{01}}(h) = \tfrac{k}{n}\right) &= \binom{n}{k}\Big(\mathcal{L}_D^{\ell_{01}}(h)\Big)^k\Big(1-\mathcal{L}_D^{\ell_{01}}(h)\Big)^{n-k}\\ &= \mathbf{Bin}\Big(k; n, \mathcal{L}_D^{\ell_{01}}(h)\Big)\end{aligned}$$

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Proof.

$$n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\Big)$$

Jensen's Inequality $\leq \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{\mathcal{L}}_S^{\ell}(h), \mathcal{L}_D^{\ell}(h)\Big)$

Change of measure $\leq \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{\mathcal{L}}_S^{\ell}(h), \mathcal{L}_D^{\ell}(h)\right)}$

Markov's Inequality $\leq_{1-\delta} \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^n}\mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h), \mathcal{L}_D^{\ell}(h))}$

Expectation swap $= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim D^n} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h), \mathcal{L}_D^{\ell}(h))}$

Binomial law $= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{n}\mathbf{Bin}\big(k; n, \mathcal{L}_D^{\ell}(h)\big)\, e^{n\cdot\Delta(\frac{k}{n}, \mathcal{L}_D^{\ell}(h))}$

Supremum over risk $\leq \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{n}\mathbf{Bin}\big(k; n, r\big)\, e^{n\Delta(\frac{k}{n}, r)}\right]$

$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathcal{I}_\Delta(n)\,.$ $\square$

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}: \ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Corollary

[...] with probability at least $1-\delta$ over the choice of $S\sim D^n$, for all Q on $\mathcal{H}$:

(a) $\mathrm{kl}\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]$, *(Langford and Seeger 2001)*

(b) $R_D(G_Q) \leq \widehat{R}_S(G_Q) + \sqrt{\frac{1}{2n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]}$, *(McAllester 1999, 2003)*

(c) $R_D(G_Q) \leq \frac{1}{1-e^{-c}}\left(c\cdot\widehat{R}_S(G_Q) + \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right)$, *(Catoni 2007)*

(d) $R_D(G_Q) \leq \widehat{R}_S(G_Q) + \frac{1}{\lambda}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} + f(\lambda, n)\right]$. *(Alquier et al. 2015)*

$$\begin{aligned}
\mathrm{kl}(q,p) &= q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \geq 2(q-p)^2\,,\\
\Delta_c(q,p) &= -\ln[1-(1-e^{-c})\cdot p] - c\cdot q\,,\\
\Delta_\lambda(q,p) &= \frac{\lambda}{n}(p-q)\,.
\end{aligned}$$

Plan

Bounding the Expected Disagreement

Expected disagreement

$$d_Q^D \;=\; \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\ \mathop{\mathbf{E}}_{h_i\sim Q}\ \mathop{\mathbf{E}}_{h_j\sim Q} \mathrm{I}\Big[h_i(x) \neq h_j(x)\Big] \;=\; \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\ \mathop{\mathbf{E}}_{h_{ij}\sim Q^2} \ell_d(h_{ij}, x, \cdot)\,,$$

where $Q^2(h_{ij}) \;=\; Q(h_i)\,Q(h_j) \quad\Longrightarrow\quad \mathrm{KL}(Q^2\|P^2) \;=\; 2\,\mathrm{KL}(Q\|P)\,.$

General theorem

[...] with probability at least $1{-}\delta$ *over the choice of* $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H}: \quad \Delta\Big(\widehat{d}_Q^S,\, d_Q^D\Big) \;\le\; \frac{1}{n}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right].$$

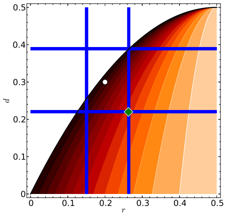
Corollary

(a) $\mathrm{kl}\Big(\widehat{d}_Q^S,\, d_Q^D\Big) \le \frac{1}{n}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right].$

The $\mathcal{C}$-bound

The $\mathcal{C}$-bound (Lacasse et al., 2006)

$$R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;=\; 1 - \frac{\Big(1 - 2\cdot R_D(G_Q)\Big)^2}{1 - 2\cdot d_Q^D}\,.$$



PAC-Bayes $\mathcal{C}$-bound 1

[...] with probability at least $1{-}\delta$ *over the choice of* $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H}: \quad R_D(B_Q) \;\leq\; 1 - \frac{\big(1{-}2{\cdot}\overline{r}\big)^2}{1{-}2{\cdot}\underline{d}}\,,$$

with

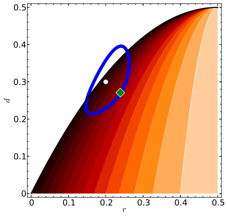
$$\overline{r} = \max_{r\in[0,\frac{1}{2}]}\left\{\Delta\Big(\widehat{R}_S(G_Q),\, r\Big) \;\leq\; \frac{1}{n}\Big[2\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta/2}\Big]\right\},$$

$$\underline{d} = \min_{d\in[0,\frac{1}{2}]}\left\{\Delta\Big(\widehat{d}_Q^S,\, d\Big),\;\leq\; \frac{1}{n}\Big[2\,\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta/2}\Big]\right\}.$$

The $\mathcal{C}$-bound

The $\mathcal{C}$-bound (Lacasse et al., 2006)

$$R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;=\; 1 - \frac{\Big(1 - 2 \cdot R_D(G_Q)\Big)^2}{1 - 2 \cdot d_Q^D}\,.$$



PAC-Bayes $\mathcal{C}$-bound 2

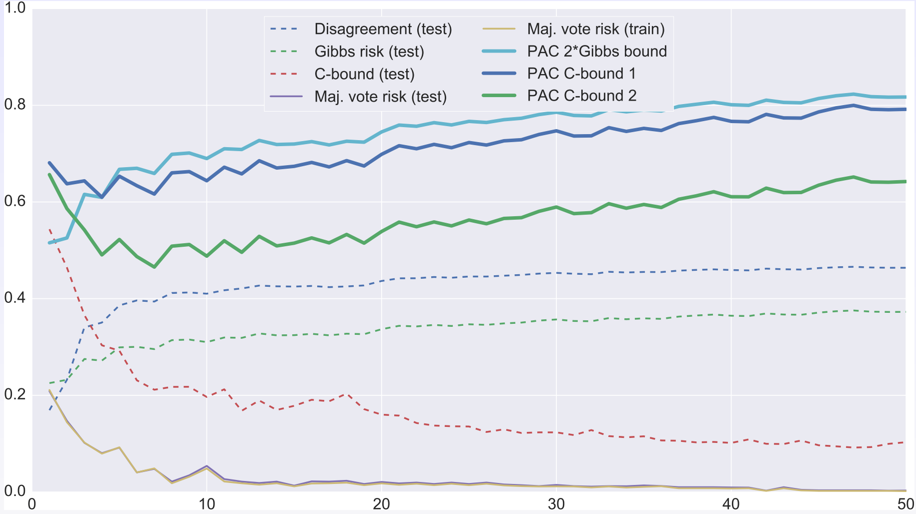
[...] with probability at least $1-\delta$ *over the choice of* $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H}: \quad R_D(B_Q) \;\leq\; \sup_{(r,d)\in\mathcal{A}} \left\{ 1 - \frac{\big(1-2\cdot r\big)^2}{1-2\cdot d} \right\},$$

with

$$\mathcal{A} = \left\{ (r,d) \in [0, \tfrac{1}{2}] \;\middle|\; \Delta_2\Big((\widehat{R}_S(G_Q),\, r), (\widehat{d}_Q^S,\, d) \Big) \;\leq\; \frac{1}{n}\Big[2\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_{\Delta_2}(n)}{\delta} \Big] \right\}.$$

Bounds Values (adaboost iterates)



Plan

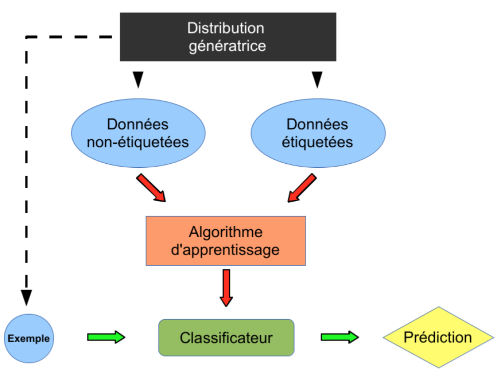
Plan

Semi-Supervised Learning

Assumption

Examples are generated *i.i.d.* by a distribution D on $\mathcal{X} \times \mathcal{Y}$.

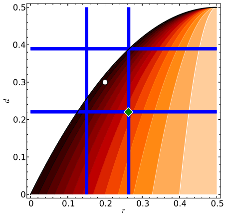
$$\begin{array}{rclllll} S &=& \{\,(x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n)\,\} & \sim D^n \\ U &=& \{\,(x_{n+1}, \cdot), & (x_{n+2}, \cdot), & \ldots, & (x_{n+n'}, \cdot)\,\} & \sim D^{n'} \end{array}$$



Semi-Supervised Learning

The $\mathcal{C}$-bound (Lacasse et al., 2006)

$$R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;=\; 1 - \frac{\Big(1 - 2\cdot R_D(G_Q)\Big)^2}{1 - 2\cdot d_Q^D}\,.$$



PAC $\mathcal{C}$-bound semi-supervised

[...] with probability at least $1{-}\delta$ *over the choice of* $S \sim D^n$ *and* $U \sim D^{n'}$,

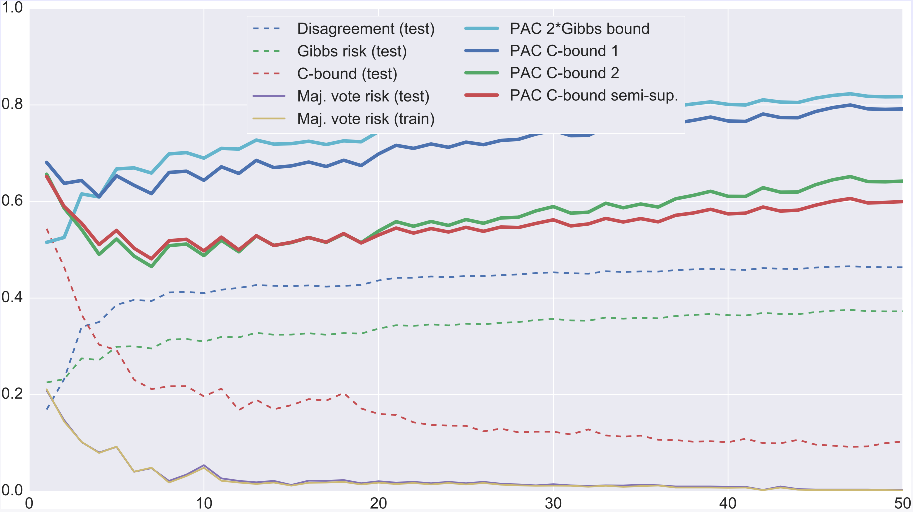
$$\forall\, Q \text{ on } \mathcal{H}: \quad R_D(B_Q) \;\leq\; 1 - \frac{\big(1{-}2{\cdot}\overline{r}\big)^2}{1{-}2{\cdot}\underline{d}}\,,$$

with

$$\overline{r} = \max_{r\in[0,\frac{1}{2}]}\left\{\Delta\Big(\widehat{R}_S(G_Q),\, r\Big) \;\leq\; \frac{1}{n}\Big[2\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta/2}\Big]\right\},$$

$$\underline{d} = \min_{d\in[0,\frac{1}{2}]}\left\{\Delta\Big(\widehat{d}_Q^{S\cup U},\, d\Big) \;\leq\; \frac{1}{n{+}n'}\Big[2\,\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n{+}n')}{\delta/2}\Big]\right\}.$$

Bounds Values (adaboost iterates)



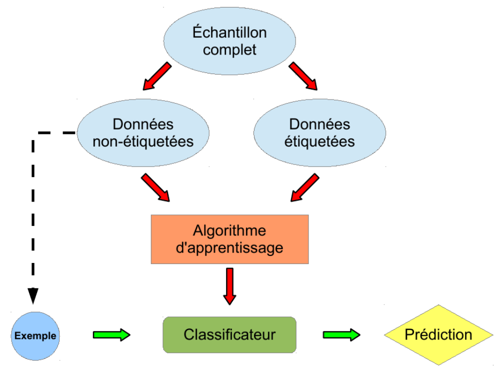
Plan

Transductive Learning

Assumption

Examples are drawn *without replacement* from a finite set Z of size N.

$$\begin{array}{rcllll} S & = & \{\, (x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n) \,\} \subset Z \\ U & = & \{\, (x_{n+1}, \cdot), & (x_{n+2}, \cdot), & \ldots, & (x_N, \cdot) \,\} = Z \setminus S \end{array}$$



Transductive Learning

Assumption

Examples are drawn *without replacement* from a finite set Z of size N.

$$\begin{array}{rcllllll} S &=& \{\, (x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n) \,\} & \subset Z \\ U &=& \{\, (x_{n+1}, \cdot), & (x_{n+2}, \cdot), & \ldots, & (x_N, \cdot) \,\} & = Z \setminus S \end{array}$$

Inductive learning: n draws with replacement according to $D \Rightarrow$ Binomial law.

Transductive learning: n draws without replacement in $Z \Rightarrow$ Hypergeometric law.

Theorem (Bégin et al. 2014)

For any set Z of N examples, [...] with probability at least $1-\delta$ over the choice of n examples among Z,

$$\forall\, Q \text{ on } \mathcal{H}: \quad \Delta\Big(\widehat{R}_S(G_Q), \widehat{R}_Z(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\Delta(n,N)}{\delta}\right],$$

where

$$\mathcal{T}_\Delta(n,N) = \max_{K=0\ldots N}\left[\sum_{k=\max[0,K+n-N]}^{\min[n,K]} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} e^{n\Delta\left(\frac{k}{n},\frac{K}{N}\right)}\right].$$

Theorem

$$\Pr_{S\sim[Z]^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), \widehat{R}_Z(G_Q)\Big) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\Delta(n,N)}{\delta}\right]\right) \ge 1-\delta\,.$$

Proof.

$$n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^\ell(h), \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_Z^\ell(h)\Big)$$

Jensen's inequality $\le \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{\mathcal{L}}_S^\ell(h), \widehat{\mathcal{L}}_Z^\ell(h)\Big)$

Change of measure $\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{\mathcal{L}}_S^\ell(h), \widehat{\mathcal{L}}_Z^\ell(h)\right)}$

Markov's inequality $\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim[Z]^n}\mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^\ell(h), \widehat{\mathcal{L}}_Z^\ell(h))}$

Expectations swap $= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim[Z]^n} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^\ell(h), \widehat{\mathcal{L}}_Z^\ell(h))}$

Hypergeometric law $= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_k \frac{\binom{N\cdot\widehat{\mathcal{L}}_Z^\ell(h)}{k}\binom{N-N\cdot\widehat{\mathcal{L}}_Z^\ell(h)}{n-k}}{\binom{N}{n}} e^{n\cdot\Delta(\frac{k}{n}, \widehat{\mathcal{L}}_Z^\ell(h))}$

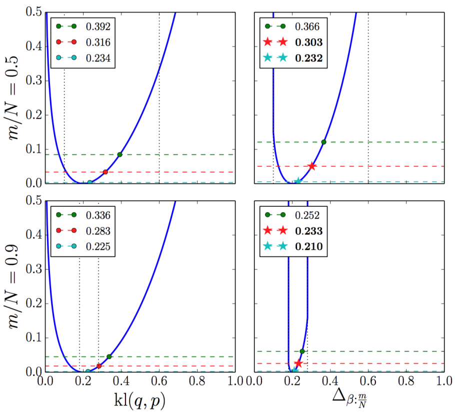
Supremum over risk $\le \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\max_{K=0\ldots N}\left[\sum_k \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} e^{n\Delta(\frac{k}{n}, \frac{K}{N})}\right]$

$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathcal{T}_\Delta(n,N)\,.$ $\square$

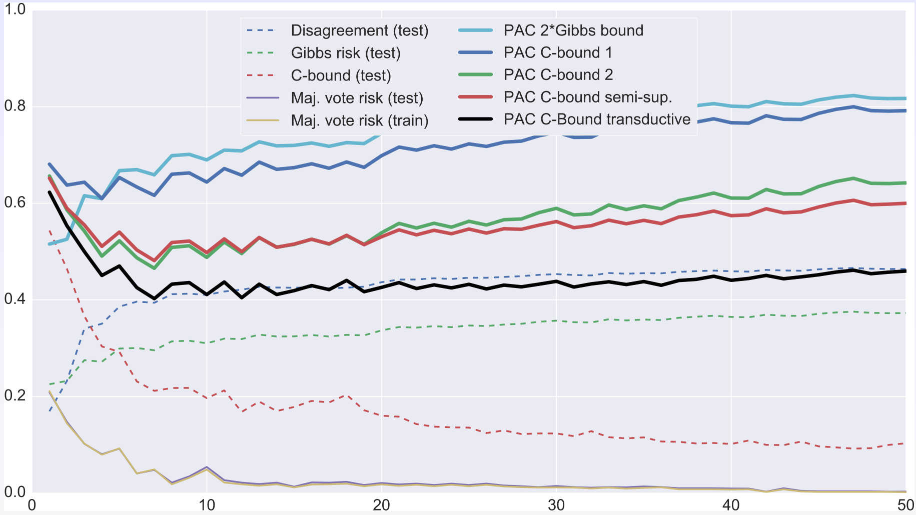
A New Transductive Bound

A new Δ-function (Bégin et al. 2014)

$$\Delta_\beta(q,p) \;=\; \mathrm{kl}(q,p) + \frac{1-\beta}{\beta}\,\mathrm{kl}\!\left(\frac{p-\beta q}{1-\beta}, p\right).$$



Bounds Values (adaboost iterates)



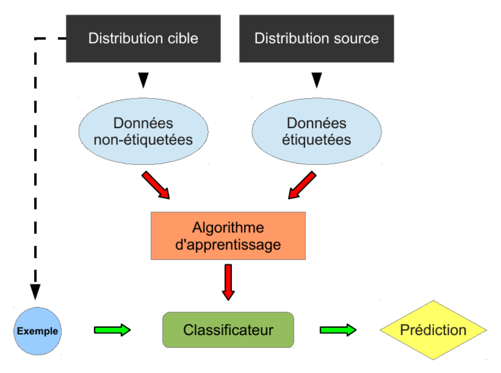
Plan

Domain Adaptation

Assumption

Source and target examples are generated by different distributions.

$$\begin{array}{rcllllll} S & = & \{\,(x_1,y_1), & (x_2,y_2), & \ldots, & (x_n,y_n)\,\} & \sim & (D_S)^n \\ T & = & \{\,(x_1,\cdot), & (x_2,\cdot), & \ldots, & (x_n,\cdot)\,\} & \sim & (D_T)^{n'} \end{array}$$



Our Domain Adaptation Setting

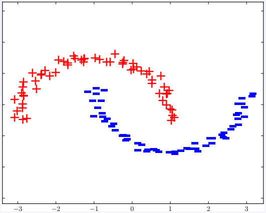
Assumption

Source and target examples are generated by different distributions.

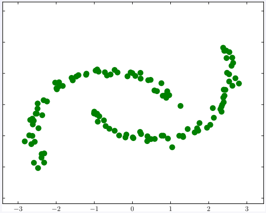
$$\begin{array}{lcllllcl} \color{red}{S} & = & \{\, (x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n)\,\} & \sim & (D_\mathrm{S})^n \\ \color{green}{T} & = & \{\, (x_1, \cdot), & (x_2, \cdot), & \ldots, & (x_n, \cdot)\,\} & \sim & (D_\mathrm{T})^{n'} \end{array}$$

A **domain adaptation** learning algorithm is provided with

a **labeled source sample**



an **unlabeled target sample**



The goal is to build a classifier with a low **target risk** $R_{D_\mathrm{T}}(h)$

Generalization Bound

Observation

$$R_D(G_Q) \;=\; \tfrac{1}{2} d_Q^D + e_Q^D \;=\; \tfrac{1}{2} d_Q^D + \mathop{\mathbf{E}}_{(x,y)\sim D} \mathop{\mathbf{E}}_{h_i \sim Q} \mathop{\mathbf{E}}_{h_j \sim Q} \mathrm{I}\Big[h_i(x) \neq y\Big] \mathrm{I}\Big[h_j(x) \neq y\Big].$$

PAC-Bayesian DA Bound (Germain, Habrard, et al. 2016)

For any prior P over $\mathcal{H}$, any $\delta \in (0,1]$, any real numbers $b > 1$ and $c > 1$, with a probability at least $1-\delta$ over the choices of $S \sim (D_S)^n$ and $T \sim (D_T)^{n'}$, we have

$\forall Q$ on $\mathcal{H}$,

$$R_{D_T}(G_Q) \leq \; c \times \tfrac{1}{2} \overbrace{\widehat{d}_Q^T}^{\substack{\text{empirical} \\ \text{target disagreement}}} + b \times \underbrace{\beta_\infty(D_T \| D_S)}_{\text{domain divergence}} \overbrace{\widehat{e}_Q^S}^{\substack{\text{empirical} \\ \text{source joint error}}} + \overbrace{O\Big(\mathrm{KL}(Q\|P) + \ln \tfrac{1}{\delta}\Big)}^{\text{complexity term}}.$$

Linear trade-off between $\widehat{d}_Q^T$ and $\widehat{e}_Q^S$:

$\Rightarrow$ We consider $\beta_\infty(D_T \| D_S) = \sup_{\mathbf{x}} \frac{D_T(\mathbf{x})}{D_S(\mathbf{x})}$ as a parameter to tune.

Learning algorithm for Linear Classifiers

As many PAC-Bayesian works (since Langford and Shawe-Taylor 2002) :

- We consider the set $\mathcal{H}$ of all linear classifiers $h_{\mathbf{v}}$ in $\mathcal{X} = \mathbb{R}^d$:

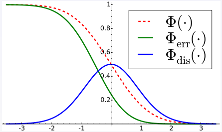
$$h_{\mathbf{v}}(\mathbf{x}) \;=\; \mathrm{sgn}(\mathbf{v}\cdot\mathbf{x})\,.$$

- Let $Q_{\mathbf{w}}$ on $\mathcal{H}$ be a Gaussian distribution centered on $\mathbf{w}$ (with $\Sigma = \mathbf{I}_d$):

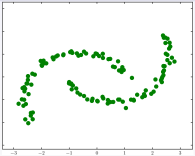
$$h_{\mathbf{w}}(\mathbf{x}) = \mathrm{sgn}\left[\mathop{\mathbf{E}}_{\mathbf{v}\sim Q_{\mathbf{w}}} h_{\mathbf{v}}(\mathbf{x})\right].$$

Given $T = \{\mathbf{x}_i\}_{i=1}^{n'}$ and $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n}$, find $\mathbf{w} \in \mathbb{R}^d$ that minimizes:

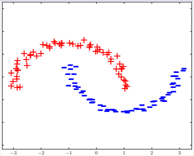
$$\underbrace{C \times \widehat{d}_Q^{\,T}}_{\frac{C}{n'}\sum_i \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)} \;+\; \underbrace{B \times \widehat{e}_Q^{\,S}}_{\frac{B}{n}\sum_j \Phi_{\mathrm{err}}\left(y_j\frac{\mathbf{w}\cdot\mathbf{x}_j}{\|\mathbf{x}_j\|}\right)} + \underbrace{\mathrm{KL}(Q_{\mathbf{w}}\|P_{\mathbf{0}})}_{\frac{1}{2}\|\mathbf{w}\|^2}.$$



Margin $y\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$ losses



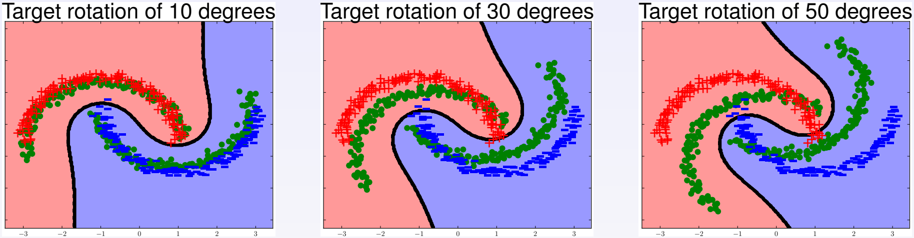
Low density region on target



Classification accuracy on source

Toy Experiment

- RBF kernel
- $B = 1$
- $C = 1$



Empirical results on Amazon Dataset

- Linear kernel
- Hyper-parameter selection by reverse cross-validation

	svm	dasvm	coda	PBDA	**DALC**
books→DVDs	*0.179*	0.193	0.181	0.183	**0.178**
books→electro	0.290	*0.226*	0.232	0.263	**0.212**
books→kitchen	0.251	**0.179**	0.215	0.229	*0.194*
DVDs→books	0.203	0.202	0.217	*0.197*	**0.186**
DVDs→electro	0.269	**0.186**	*0.214*	0.241	0.245
DVDs→kitchen	0.232	0.183	*0.181*	0.186	**0.175**
electro→books	0.287	0.305	0.275	**0.232**	*0.240*
electro→DVDs	0.267	**0.214**	0.239	*0.221*	0.256
electro→kitchen	*0.129*	0.149	0.134	0.141	**0.123**
kitchen→books	0.267	0.259	*0.247*	*0.247*	**0.236**
kitchen→DVDs	0.253	**0.198**	0.238	0.233	*0.225*
kitchen→electro	0.149	0.157	0.153	**0.129**	*0.131*
Average	0.231	*0.204*	0.210	0.208	**0.200**

Plan

Short Summary

An original PAC-Bayesian approach

- General theorem from which we recover existing results;
- Modular proof, easy to adapt to various frameworks (*i.e.*, transductive learning).

The virtue of disagreement

- Second-order statistic from unlabeled sample
(useful in semi-supervised setting and variants);
- Allows to reduce the value of PAC-Bayesian bounds ($\mathcal{C}$-bound);
- Also used in a domain adaptation learning algorithm.

Some (self-)pointers

Our 74 pages journal paper (JMLR)

- Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm (Germain, Lacasse, et al. 2015)

My PhD thesis (in french)

- Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine (Germain 2015)
  `http://www.di.ens.fr/~germain/publis/these.pdf`

Recent papers

- ICML: A New PAC-Bayesian Perspective on Domain Adaptation (Germain, Habrard, et al. 2016)
- AISTATS: PAC-Bayesian Bounds Based on the Rényi Divergence (Bégin et al. 2016)
- NIPS: PAC-Bayesian Theory Meets Bayesian Inference (Germain, Bach, et al. 2016)

Even some code on my GitHub....

- `https://github.com/pgermain`

References I

Alquier, Pierre, James Ridgway, and Nicolas Chopin (2015). "On the properties of variational approximations of Gibbs posteriors". In: *ArXiv e-prints*. url: http://arxiv.org/abs/1506.04091.

Bégin, Luc, Pascal Germain, François Laviolette, and Jean-Francis Roy (2014). "PAC-Bayesian Theory for Transductive Learning". In: *AISTATS*.

— (2016). "PAC-Bayesian Bounds based on the Rényi Divergence". In: *AISTATS*.

Catoni, Olivier (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Vol. 56. Inst. of Mathematical Statistic.

Germain, Pascal (2015). "Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine." PhD thesis. Université Laval. url: http://www.theses.ulaval.ca/2015/31774/.

Germain, Pascal, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien (2016). "PAC-Bayesian Theory Meets Bayesian Inference". In: *ArXiv e-prints*. url: http://arxiv.org/abs/1605.08636.

Germain, Pascal, Amaury Habrard, François Laviolette, and Emilie Morvant (2016). "A New PAC-Bayesian Perspective on Domain Adaptation". In: *ICML*. url: http://arxiv.org/abs/1506.04573.

Germain, Pascal, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy (2015). "Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm". In: *JMLR* 16.

References II

Langford, John and Matthias Seeger (2001). *Bounds for averaging classifiers*. Tech. rep. Carnegie Mellon, Departement of Computer Science.

Langford, John and John Shawe-Taylor (2002). "PAC-Bayes & Margins". In: *NIPS*.

McAllester, David (1999). "Some PAC-Bayesian Theorems". In: *Machine Learning* 37.3.

— (2003). "PAC-Bayesian Stochastic Model selection". In: *Machine Learning* 51.1.