PAC-Bayes theory in supervised Learning

Université Laval, Québec, Canada

François Laviolette

March 22, 2010

Summary

Today, I intend to

- present some basic mathematics that underlies the PAC-Bayes theory
- look for PAC-Bayes bound minimization algorithms and compare them with existing ones.

Summary

Today, I intend to

- present some basic mathematics that underlies the PAC-Bayes theory
- look for PAC-Bayes bound minimization algorithms and compare them with existing ones.

Definitions

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

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

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

Definitions

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

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

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

Definitions

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

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

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

Definitions

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

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

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

Definitions

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

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

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_Q
- It bounds the risk of the **Gibbs classifier** G_Q:
  - to predict the label of x, G_Q draws h from $\mathcal{H}$ and predicts $h(x)$
- The risk and the training error of G_Q are thus defined as:

$$R(G_Q) = \mathop{\mathbf{E}}_{h\sim Q} R(h) \quad ; \quad R_S(G_Q) = \mathop{\mathbf{E}}_{h\sim Q} R_S(h) .$$

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_Q
- It bounds the risk of the **Gibbs classifier** G_Q:
  - to predict the label of $\mathbf{x}$, G_Q draws h from $\mathcal{H}$ and predicts $h(\mathbf{x})$
- The risk and the training error of G_Q are thus defined as:

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

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_Q
- It bounds the risk of the **Gibbs classifier** G_Q:
  - to predict the label of $\mathbf{x}$, G_Q draws h from $\mathcal{H}$ and predicts $h(\mathbf{x})$
- The risk and the training error of G_Q are thus defined as:

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

The Gibbs clasifier

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

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

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

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies **x**, then at least half of the classifiers (under measure Q) err on **x**.
  - Hence: $R(B_Q) \leq 2R(G_Q)$
  - **Thus, an upper bound on** $R(G_Q)$ **gives rise to an upper bound on** $R(B_Q)$
- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.
- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}.$$

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies $\mathbf{x}$, then at least half of the classifiers (under measure Q) err on $\mathbf{x}$.
  - Hence: $R(B_Q) \leq 2R(G_Q)$
  - **Thus, an upper bound on $R(G_Q)$ gives rise to an upper bound on $R(B_Q)$**
- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.
- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}\,.$$

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies $\mathbf{x}$, then at least half of the classifiers (under measure Q) err on $\mathbf{x}$.
  - Hence: $R(B_Q) \leq 2R(G_Q)$
  - **Thus, an upper bound on $R(G_Q)$ gives rise to an upper bound on $R(B_Q)$**
- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.
- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\mathrm{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}.$$

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies $\mathbf{x}$, then at least half of the classifiers (under measure Q) err on $\mathbf{x}$.
  - Hence: $R(B_Q) \leq 2R(G_Q)$
  - **Thus, an upper bound on $R(G_Q)$ gives rise to an upper bound on $R(B_Q)$**
- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.
- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}.$$

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies $\mathbf{x}$, then at least half of the classifiers (under measure Q) err on $\mathbf{x}$.
    - Hence: $R(B_Q) \leq 2R(G_Q)$
    - **Thus, an upper bound on $R(G_Q)$ gives rise to an upper bound on $R(B_Q)$**
- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.
- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)} .$$

A PAC-Bayes bound to rule them all !

J.R.R. Tolkien, roughly
or John Langford, less roughly.

Theorem 1 Germain et al. 2009

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of classifiers, for any prior distribution P of support $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, we have

$$\Pr_{S \sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ \mathcal{D}(R_S(G_Q), R(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q \| P) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim D} \mathop{\mathbf{E}}_{h \sim P} e^{m\mathcal{D}(R_S(h), R(h))} \right) \right] \right) \geq 1 - \delta\,.$$

A PAC-Bayes bound to rule them all !

J.R.R. Tolkien, roughly
or John Langford, less roughly.

Theorem 1^+ Lever et al (2010)

For any functions $A(h)$, $B(h)$ over $\mathcal{H}$, either of which may be a statistic of a sample S of size n, any distributions P over $\mathcal{H}$, any $\delta \in (0,1]$, any $t > 0$, and convex function $\mathcal{D} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, we have

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ \mathcal{D}\left(\mathop{\mathbf{E}}_{h\in Q} A(h), \mathop{\mathbf{E}}_{h\in Q} B(h)\right) \leq \frac{1}{t}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D}\mathop{\mathbf{E}}_{h\sim P} e^{t\cdot\mathcal{D}(A(h),B(h))}\right)\right]\right) \geq 1-\delta\,.$$

Proof of Theorem 1

- Since $\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$ is a non-negative r.v., Markov's inequality gives

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} \le \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right) \ge 1-\delta\,.$$

- Hence, by taking the logarithm on each side of the inequality and by transforming the expectation over P into an expectation over Q:

$$\Pr_{S\sim D^m}\left(\forall Q: \ln\left[\mathop{\mathbf{E}}_{h\sim Q} \frac{P(h)}{Q(h)} e^{m\mathcal{D}(R_S(h),R(h))}\right] \le \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

- Then, exploiting the fact that the logarithm is a concave function, by an application of Jensen's inequality, we obtain

$$\Pr_{S\sim D^m}\left(\forall Q: \mathop{\mathbf{E}}_{h\sim Q} \ln\left[\frac{P(h)}{Q(h)} e^{m\mathcal{D}(R_S(h),R(h))}\right] \le \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

Proof of Theorem 1

- Since $\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$ is a non-negative r.v., Markov's inequality gives

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} \le \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right) \ge 1-\delta\,.$$

- Hence, by taking the logarithm on each side of the inequality and by transforming the expectation over P into an expectation over Q:

$$\Pr_{S\sim D^m}\left(\forall Q:\ \ln\left[\mathop{\mathbf{E}}_{h\sim Q}\frac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right] \le \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

- Then, exploiting the fact that the logarithm is a concave function, by an application of Jensen's inequality, we obtain

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathop{\mathbf{E}}_{h\sim Q}\ln\left[\frac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right] \le \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

Proof of Theorem 1

- Since $\mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$ is a non-negative r.v., Markov's inequality gives

$$\Pr_{S\sim D^m}\left(\mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} \leq \frac{1}{\delta}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right) \geq 1-\delta\,.$$

- Hence, by taking the logarithm on each side of the inequality and by transforming the expectation over P into an expectation over Q :

$$\Pr_{S\sim D^m}\left(\forall Q:\ \ln\left[\mathbf{E}_{h\sim Q}\frac{P(h)}{Q(h)} e^{m\mathcal{D}(R_S(h),R(h))}\right] \leq \ln\left[\frac{1}{\delta}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \geq 1-\delta\,.$$

- Then, exploiting the fact that the logarithm is a concave function, by an application of Jensen's inequality, we obtain

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathbf{E}_{h\sim Q}\ln\left[\frac{P(h)}{Q(h)} e^{m\mathcal{D}(R_S(h),R(h))}\right] \leq \ln\left[\frac{1}{\delta}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \geq 1-\delta\,.$$

Proof of Theorem 1 (cont)

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right]\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

- From basic logarithm properties, and from the fact that $\mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}\right]\stackrel{\text{def}}{=}-\mathrm{KL}(Q\|P)$, we now have

$$\Pr_{S\sim D^m}\left(\forall Q:\ -\mathrm{KL}(Q\|P)+\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h))\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

- Then, since $\mathcal{D}$ has been supposed convexe, again by the Jensen inequality, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h)) = m\,\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim Q}\ R_S(h),\mathop{\mathbf{E}}_{h\sim Q}\ R(h)\right),$$

which immediately implies the result. □

Proof of Theorem 1 (cont)

$$\Pr_{S\sim D^m}\left(\forall Q: \mathop{\mathbf{E}}_{h\sim Q} \ln\left[\frac{P(h)}{Q(h)} e^{m\mathcal{D}(R_S(h),R(h))}\right] \le \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

- From basic logarithm properties, and from the fact that $\mathop{\mathbf{E}}_{h\sim Q} \ln\left[\frac{P(h)}{Q(h)}\right] \stackrel{\text{def}}{=} -\mathrm{KL}(Q\|P)$, we now have

$$\Pr_{S\sim D^m}\left(\forall Q: -\mathrm{KL}(Q\|P) + \mathop{\mathbf{E}}_{h\sim Q} m\mathcal{D}(R_S(h),R(h)) \le \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right) \ge 1-\delta\,.$$

- Then, since $\mathcal{D}$ has been supposed convexe, again by the Jensen inequality, we have

$$\mathop{\mathbf{E}}_{h\sim Q} m\mathcal{D}(R_S(h),R(h)) = m\,\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R(h)\right),$$

which immediately implies the result. □

Proof of Theorem 1 (cont)

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right]\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

- From basic logarithm properties, and from the fact that $\mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}\right]\overset{\text{def}}{=}-\mathrm{KL}(Q\|P)$, we now have

$$\Pr_{S\sim D^m}\left(\forall Q:\ -\mathrm{KL}(Q\|P)+\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h))\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

- Then, since $\mathcal{D}$ has been supposed convexe, again by the Jensen inequality, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h))\ =\ m\,\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim Q}\ R_S(h),\mathop{\mathbf{E}}_{h\sim Q}\ R(h)\right),$$

which immediately implies the result. □

Applicability of Theorem 1

How can we estimate $\ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim D^m} \mathop{\mathbf{E}}_{h \sim P} e^{m \mathcal{D}(R_S(h), R(h))} \right]$?

The Seeger's bound (2002)

Seeger Bound

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

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ \mathrm{kl}(R_S(G_Q), R(G_Q)) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta}\right]\right) \ge 1-\delta\,,$$

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

and where $\quad \xi(m) \overset{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

- Note: $\xi(m) \in \Theta(\sqrt{m})$ and $\xi(m) \le m+1$

The Seeger's bound (2002)

Seeger Bound

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

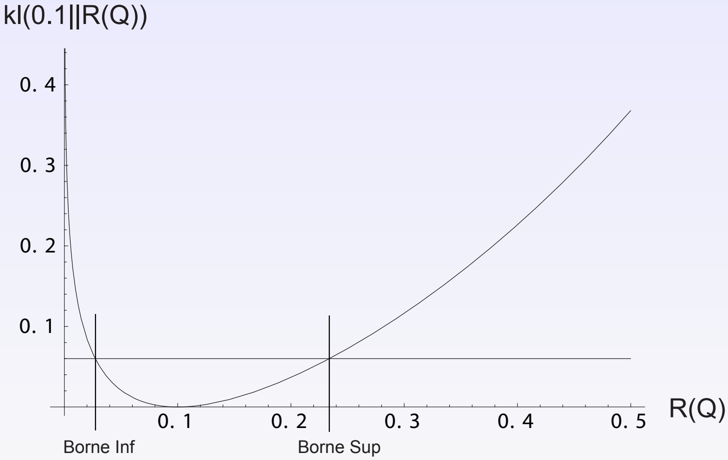
$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ \mathrm{kl}(R_S(G_Q), R(G_Q)) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta}\right]\right) \ge 1-\delta\,,$$

where $\mathrm{kl}(q,p) \stackrel{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$,

and where $\xi(m) \stackrel{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

- Note: $\xi(m) \in \Theta(\sqrt{m})$ and $\xi(m) \le m+1$

Graphical illustration of the Seeger bound



Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \left(\frac{R_S(h)}{R(h)}\right)^{mR_S(h)} \left(\frac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \Pr_{S\sim D^m}\left(R_S(h)=\frac{k}{m}\right) \left(\frac{\frac{k}{m}}{R(h)}\right)^{k} \left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \sum_{k=0}^{m} \binom{m}{k} (k/m)^k (1-k/m)^{m-k}\,, \qquad (1) \\
&\leq m+1\,. \qquad \square
\end{aligned}
$$

- Note that, in Line (1) of the proof, $\Pr_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \left(\frac{R_S(h)}{R(h)}\right)^{mR_S(h)} \left(\frac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \textstyle\sum_{k=0}^{m} \Pr\limits_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right) \left(\frac{\frac{k}{m}}{R(h)}\right)^k \left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \textstyle\sum_{k=0}^{m} \binom{m}{k} (k/m)^k (1-k/m)^{m-k}\,, \qquad (1) \\
&\leq m+1\,.
\end{aligned}$$

$\square$

- Note that, in Line (1) of the proof, $\Pr\limits_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \left(\frac{R_S(h)}{R(h)}\right)^{mR_S(h)} \left(\frac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \textstyle\sum_{k=0}^{m} \displaystyle\Pr_{S\sim D^m}\left(R_S(h){=}\tfrac{k}{m}\right) \left(\frac{\frac{k}{m}}{R(h)}\right)^{k} \left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \textstyle\sum_{k=0}^{m} \binom{m}{k} (k/m)^k (1-k/m)^{m-k}\,, \qquad (1) \\
&\leq m+1\,. \qquad \square
\end{aligned}
$$

- Note that, in Line (1) of the proof, $\displaystyle\Pr_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \left(\tfrac{R_S(h)}{R(h)}\right)^{mR_S(h)} \left(\tfrac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \textstyle\sum_{k=0}^{m} \Pr\limits_{S\sim D^m}\left(R_S(h)=\frac{k}{m}\right)\left(\frac{\frac{k}{m}}{R(h)}\right)^{k}\left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \textstyle\sum_{k=0}^{m} \binom{m}{k}(k/m)^k(1-k/m)^{m-k}\,, \qquad (1) \\
&\leq m+1\,.
\end{aligned}
$$

$\square$

- Note that, in Line (1) of the proof, $\Pr\limits_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P}\ \mathop{\mathbf{E}}_{S\sim D^m}\ \left(\frac{R_S(h)}{R(h)}\right)^{mR_S(h)}\left(\frac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P}\ \sum_{k=0}^{m} \Pr_{S\sim D^m}\left(R_S(h)=\tfrac{k}{m}\right)\left(\frac{\frac{k}{m}}{R(h)}\right)^{k}\left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \sum_{k=0}^{m} \binom{m}{k}(k/m)^k(1-k/m)^{m-k}\,, \qquad (1) \\
&\leq m+1\,. \qquad \square
\end{aligned}
$$

- Note that, in Line (1) of the proof, $\Pr_{S\sim D^m}\left(R_S(h)=\frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S\sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

The McAllester's bound (1998)

Put $\mathcal{D}(q,p) = \frac{1}{2}(q-p)^2$, Theorem 1 then gives

McAllester Bound

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

$$\Pr_{S\sim D^m}\Bigg(\forall\, Q \text{ on } \mathcal{H}\colon\ \frac{1}{2}(R_S(G_Q), R(G_Q))^2 \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta}\right]\Bigg) \geq 1-\delta\,,$$

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

and where $\xi(m) \stackrel{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

- Note: $\xi(m) \in \Theta(\sqrt{m})$ and $\xi(m) \leq m+1$

The McAllester's bound (1998)

Put $\mathcal{D}(q,p) = \frac{1}{2}(q-p)^2$, Theorem 1 then gives

McAllester Bound

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

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

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

and where $\xi(m) \overset{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

- Note: $\xi(m) \in \Theta(\sqrt{m})$ and $\xi(m) \le m+1$

The Catoni's bound (2004)

In Theorem 1, let $\mathcal{D}(q,p) = \mathcal{F}(p) - C \cdot q$., then

Catoni's bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, and any positive real number C, we have

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

- Because,

$$\mathop{\mathbf{E}}_{S \sim D^m} \mathop{\mathbf{E}}_{h \sim P} e^{m\, \mathcal{D}(R_S(h), R(h))} = \mathop{\mathbf{E}}_{h \sim P} e^{m \mathcal{F}(R(h))} \left(R(h) e^{-C} + (1 - R(h)) \right)^m .$$

The Catoni's bound (2004)

In Theorem 1, let $\mathcal{D}(q,p) = \mathcal{F}(p) - C \cdot q.$, then

Catoni's bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, and any positive real number C, we have

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

- Because,

$$\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\,\mathcal{D}(R_S(h),R(h))} = \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{F}(R(h))}\left(R(h)e^{-C} + (1-R(h))\right)^m.$$

The Catoni's bound (2004)

In Theorem 1, let $\mathcal{D}(q,p) = \mathcal{F}(p) - C \cdot q.$, then

Catoni's bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, and any positive real number C, we have

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

- Because,

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\,\mathcal{D}(R_S(h),R(h))} \ = \ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{F}(R(h))}\Big(R(h)e^{-C}+(1-R(h))\Big)^m.$$

Observations about Catoni's bound

- G_Q is minimizing the Catoni's bound iff it minimizes the following cost function (linear in $R_S(G_Q)$):

$$C\, m\, R_S(G_Q) + \mathrm{KL}(Q \| P)$$

- We have a **hyperparameter** C to tune (in contrast with the Seeger' bound).
- Seeger' bound gives a bound which is always tighter except for a narrow range of C values.
  - In fact, if we would replace $\xi(m)$ by one, LS-bound would always be a tighter.

Observations about Catoni's bound

- G_Q is minimizing the Catoni's bound iff it minimizes the following cost function (linear in $R_S(G_Q)$):

$$C\, m\, R_S(G_Q) + \mathrm{KL}(Q\|P)$$

- We have a **hyperparameter** C to tune (in contrast with the Seeger' bound).
- Seeger' bound gives a bound which is always tighter except for a narrow range of C values.
  - In fact, if we would replace $\xi(m)$ by one, LS-bound would always be a tighter.

Observations about Catoni's bound

- G_Q is minimizing the Catoni's bound iff it minimizes the following cost function (linear in $R_S(G_Q)$):

$$C\, m\, R_S(G_Q) + \mathrm{KL}(Q \| P)$$

- We have a **hyperparameter** C to tune (in contrast with the Seeger' bound).
- Seeger' bound gives a bound which is always tighter except for a narrow range of C values.
    - In fact, if we would replace $\xi(m)$ by one, LS-bound would always be a tighter.

Observations about Catoni's bound (cont)

- Given any prior P, the posterior Q^* minimizing the bound of Catoni's bound is given by the Boltzman distribution:

$$Q^*(h) \;=\; \frac{1}{Z} P(h) e^{-C \cdot m R_S(h)} .$$

- We could sample Q^* by Markov Chain Monté Carlo.
  - But the mixing time being unknown, we have few control over the precision of the approximation.

- To avoid MCMC, let us analyse the case where Q is chosen from a **parameterized set of distributions** over the (continuous) space of **linear classifiers**.

Observations about Catoni's bound (cont)

- Given any prior P, the posterior Q^* minimizing the bound of Catoni's bound is given by the Boltzman distribution:

$$Q^*(h) \;=\; \frac{1}{Z} P(h) e^{-C \cdot m R_S(h)}\,.$$

- We could sample Q^* by Markov Chain Monté Carlo.
  - But the mixing time being unknown, we have few control over the precision of the approximation.
- To avoid MCMC, let us analyse the case where Q is chosen from a **parameterized set of distributions** over the (continuous) space of **linear classifiers**.

Observations about Catoni's bound (cont)

- Given any prior P, the posterior Q^* minimizing the bound of Catoni's bound is given by the Boltzman distribution:

$$Q^*(h) \; = \; \frac{1}{Z} P(h) e^{-C \cdot m R_S(h)} \, .$$

- We could sample Q^* by Markov Chain Monté Carlo.
    - But the mixing time being unknown, we have few control over the precision of the approximation.
- To avoid MCMC, let us analyse the case where Q is chosen from a **parameterized set of distributions** over the (continuous) space of **linear classifiers**.

Bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Bounding $\mathbf{E}_{S\sim\tilde{D}} \mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Bounding $\mathbf{E}_{S\sim\tilde{D}} \mathbf{E}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1. (See later on in this workshop)

Supervised learning in the non iid case

- Given a training set of m examples

$$S \stackrel{\text{def}}{=} \{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m)\}$$

where each generated according to a (unknown) distribution $\tilde{D}$ over the set $(\mathcal{X} \times \mathcal{Y})^m$ of all possible labeled examples.

- in the traditionnal iid case, the goal of the **learner** is, to try to find a **classifier** h with the smallest possible **risk** $R(h)$

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{S \sim D} \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} I(h(\mathbf{x}) \neq y) \quad \left(\neq \Pr_{(\mathbf{x},y) \sim D} \{h(\mathbf{x}) \neq y\}\right).$$

- **And the question is again: What should the learner optimize on S to obtain a classifier h having the smallest possible risk $R(h)$?**

Supervised learning in the non iid case

- Given a training set of m examples

$$S \stackrel{\text{def}}{=} \{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m)\}$$

  where each generated according to a (unknown) distribution $\tilde{D}$ over the set $(\mathcal{X} \times \mathcal{Y})^m$ of all possible labeled examples.

- in the traditionnal iid case, the goal of the **learner** is, to try to find a **classifier** h with the smallest possible **risk** $R(h)$

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{S \sim D} \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} I(h(\mathbf{x}) \neq y) \quad \Big(\neq \Pr_{(\mathbf{x},y) \sim D} \big\{ h(\mathbf{x}) \neq y \big\} \Big).$$

- And the question is again: What should the learner optimize on S to obtain a classifier h having the smallest possible risk $R(h)$?

Supervised learning in the non iid case

- Given a training set of m examples

$$S \stackrel{\text{def}}{=} \{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m)\}$$

where each generated according to a (unknown) distribution $\tilde{D}$ over the set $(\mathcal{X} \times \mathcal{Y})^m$ of all possible labeled examples.

- in the traditionnal iid case, the goal of the **learner** is, to try to find a **classifier** h with the smallest possible **risk** $R(h)$

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{S \sim D} \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} I(h(\mathbf{x}) \neq y) \quad \left(\neq \Pr_{(\mathbf{x},y) \sim D} \left\{ h(\mathbf{x}) \neq y \right\}\right).$$

- **And the question is again: What should the learner optimize on S to obtain a classifier h having the smallest possible risk $R(h)$?**

The problem of bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$

Theorem 1

For any distribution D_0, for any set $\mathcal{H}$ of classifiers, for any prior distribution P of support $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, we have

$$\Pr_{S\sim D}\Bigg(\forall\, Q \text{ on } \mathcal{H}\colon\ \mathcal{D}(R_S(G_Q), R(G_Q)) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right)\right]\Bigg) \ge 1-\delta\,.$$

- We will here restrict ourself to the particular non iid case where there exists a function g, and an integer $n \le m$ such that the $\tilde{D}$-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s.

The problem of bounding $\mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$

Theorem 1

For any distribution D_0, for any set $\mathcal{H}$ of classifiers, for any prior distribution P of support $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, we have

$$\Pr_{S\sim D}\Bigg(\forall\, Q \text{ on } \mathcal{H}\colon\; \mathcal{D}(R_S(G_Q), R(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim\tilde{D}} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}\right)\right]\Bigg) \geq 1-\delta\,.$$

- We will here restrict ourself to the particular non iid case where there exists a function g, and an integer $n \leq m$ such that the $\tilde{D}$-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s.

The fractional chromatic number of the dependency graph

- Another approach is to directly take advantage of the assumption that there exists a function g, and an integer $n \leq m$ such that the D-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s,
- Indeed, we can then subdivise S in various iid subsets S_j, togheter with weights ω_j such that each example $(\mathbf{x}_i, y_i)$, the total of the weights associate with the S_j's that contain $(\mathbf{x}_i, y_i)$ is 1.
- This is the idea of Ralaivola et al. (2008)
- Based on this idea, Theorem 1 can be restated as follows.

The fractional chromatic number of the dependency graph

- Another approach is to directly take advantage of the assumption that there exists a function g, and an integer $n \leq m$ such that the D-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s,
- Indeed, we can then subdivise S in various iid subsets S_j, togheter with weights ω_j such that each example $(\mathbf{x}_i, y_i)$, the total of the weights associate with the S_j's that contain $(\mathbf{x}_i, y_i)$ is 1.
- This is the idea of Ralaivola et al. (2008)
- Based on this idea, Theorem 1 can be restated as follows.

The fractional chromatic number of the dependency graph

- Another approach is to directly take advantage of the assumption that there exists a function g, and an integer $n \leq m$ such that the D-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s,
- Indeed, we can then subdivise S in various iid subsets S_j, togheter with weights ω_j such that each example $(\mathbf{x}_i, y_i)$, the total of the weights associate with the S_j's that contain $(\mathbf{x}_i, y_i)$ is 1.
- This is the idea of Ralaivola et al. (2008)
- Based on this idea, Theorem 1 can be restated as follows.

The fractional chromatic number of the dependency graph

- Another approach is to directly take advantage of the assumption that there exists a function g, and an integer $n \leq m$ such that the D-drawing of a training set is of the form $S = g(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ for some pairewise independent random variables $\mathbf{Z}_i \in \mathcal{Z}$'s,
- Indeed, we can then subdivise S in various iid subsets S_j, togheter with weights ω_j such that each example $(\mathbf{x}_i, y_i)$, the total of the weights associate with the S_j's that contain $(\mathbf{x}_i, y_i)$ is 1.
- This is the idea of Ralaivola et al. (2008)
- Based on this idea, Theorem 1 can be restated as follows.

Theorem 1 (revisited)

- Suppose that from any training set S drawn according to D, there is a $(S_j, \omega_j)_{j=1,..n}$ that are only defined based on the indices of elements of S is such that
  - S_j is iid and a subset of S for all $j = 1, .., n$
  - $\sum_{i=1}^{n} \omega_j \, I((\mathbf{x}_i, y_i) \in S_j) = 1 \quad$ for all $i = 1, .., m$.

Theorem 1 (revisited for the non iid case)

For any distribution D, for any set $\mathcal{H}$ of classifiers, for any prior distribution $P1,..,P_n$ of support $\mathcal{H}$, for any $\delta \in (0, 1]$, and for any convex function $\mathcal{D} : [0, 1] \times [0, 1] \to \mathbb{R}$, we have

$$\Pr_{S \sim D}\Bigg(\forall\, Q_1, ..Q_n \text{ on } \mathcal{H}\colon\; \mathcal{D}\left(\sum_{j=1}^{n} \frac{\omega_j}{\sum \omega_j} R_S(G_{Q_j}), \sum_{j=1}^{n} \frac{\omega_j}{\sum \omega_j} R(G_{Q_j}) \right) \leq$$

$$\frac{\sum_{j=1}^{n} \omega_j}{m} \left[\frac{\omega_j}{\sum_{j=1}^{n} \omega_j} \mathrm{KL}(Q_j \| P_j) + \ln\left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim D} \mathop{\mathbf{E}}_{h \sim P} \sum_{j=1}^{n} e^{m|S_j| \mathcal{D}(R_{S_j}(h_j), R(h_j))} \right) \right] \Bigg) \geq 1 - \delta\,.$$

The problem of bounding $R(G_Q)$ instead of $R(B_Q)$

The main problem PAC-Bayes theory is the fact that it allows us to bound the Gibbs risk but, most of the time, it is the Bayes risk we are in. To this problem I will discuss here two possible answers:

- Answer#1: if a non too small "part" of the classifier of $\mathcal{H}$ are strong, then one can obtained a quiet tight bound (exemple: if $\mathcal{H}$ is the set of all linear classifiers in a high-dimensional feature vectors space, like in SVM)
- Answer#2: otherwise, extend the PAC-Bayes bound to something else than the Gibbs's Risk

The problem of bounding $R(G_Q)$ instead of $R(B_Q)$

The main problem PAC-Bayes theory is the fact that it allows us to bound the Gibbs risk but, most of the time, it is the Bayes risk we are in. To this problem I will discuss here two possible answers:

- Answer#1: if a non too small "part" of the classifier of $\mathcal{H}$ are strong, then one can obtained a quiet tight bound (exemple: if $\mathcal{H}$ is the set of all linear classifiers in a high-dimensional feature vectors space, like in SVM)
- Answer#2: otherwise, extend the PAC-Bayes bound to something else than the Gibbs's Risk

Specialization to Linear classifiers

- Each $\mathbf{x}$ is mapped to a high-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$:

$$\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x})) .$$

- $\boldsymbol{\phi}$ is often implicitly given by a Mercer kernel

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}') .$$

- The output $h_{\mathbf{v}}(\mathbf{x})$ of linear classifier $h_{\mathbf{v}}$ with weight vector $\mathbf{v}$ is given by

$$h_{\mathbf{v}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})) .$$

- Let us moreover suppose that each posterior $Q_{\mathbf{w}}$ is an isotropic Gaussian centered on $\mathbf{w}$:

$$Q_{\mathbf{w}}(\mathbf{v}) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2\right)$$

Specialization to Linear classifiers

- Each $\mathbf{x}$ is mapped to a high-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$:

$$\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))\,.$$

- $\boldsymbol{\phi}$ is often implicitly given by a Mercer kernel

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}')\,.$$

- The output $h_{\mathbf{v}}(\mathbf{x})$ of linear classifier $h_{\mathbf{v}}$ with weight vector $\mathbf{v}$ is given by

$$h_{\mathbf{v}}(\mathbf{x}) \;=\; \operatorname{sgn}\left(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})\right)\,.$$

- Let us moreover suppose that each posterior $Q_{\mathbf{w}}$ is an isotropic Gaussian centered on $\mathbf{w}$:

$$Q_{\mathbf{w}}(\mathbf{v}) \;=\; \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\!\left(-\tfrac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2\right)$$

Specialization to Linear classifiers

- Each $\mathbf{x}$ is mapped to a high-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$:

$$\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))\,.$$

- $\boldsymbol{\phi}$ is often implicitly given by a Mercer kernel

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}')\,.$$

- The output $h_{\mathbf{v}}(\mathbf{x})$ of linear classifier $h_{\mathbf{v}}$ with weight vector $\mathbf{v}$ is given by

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

- Let us moreover suppose that each posterior $Q_{\mathbf{w}}$ is an isotropic Gaussian centered on $\mathbf{w}$:

$$Q_{\mathbf{w}}(\mathbf{v}) \;=\; \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\!\left(-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2\right)$$

Specialization to Linear classifiers

- Each $\mathbf{x}$ is mapped to a high-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$:

$$\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))\,.$$

- $\boldsymbol{\phi}$ is often implicitly given by a Mercer kernel

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}')\,.$$

- The output $h_{\mathbf{v}}(\mathbf{x})$ of linear classifier $h_{\mathbf{v}}$ with weight vector $\mathbf{v}$ is given by

$$h_{\mathbf{v}}(\mathbf{x}) \;=\; \operatorname{sgn}\left(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})\right)\,.$$

- Let us moreover suppose that each posterior $Q_{\mathbf{w}}$ is an isotropic Gaussian centered on $\mathbf{w}$:

$$Q_{\mathbf{w}}(\mathbf{v}) \;=\; \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\!\left(-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2\right)$$

Bayes-equivalent classifiers

- With this choice for $Q_{\mathbf{w}}$, the majority vote $B_{Q_{\mathbf{w}}}$ is the same classifier as $h_{\mathbf{w}}$ since:

$$B_{Q_{\mathbf{w}}}(\mathbf{x}) = \operatorname{sgn}\left(\mathop{\mathbf{E}}_{\mathbf{v}\sim Q_{\mathbf{w}}} \operatorname{sgn}\left(\mathbf{v}\cdot\phi(\mathbf{x})\right)\right) = \operatorname{sgn}\left(\mathbf{w}\cdot\phi(\mathbf{x})\right) = h_{\mathbf{w}}(\mathbf{x})\,.$$

- Thus $R(h_{\mathbf{w}}) = R(B_{Q_{\mathbf{w}}}) \leq 2R(G_{Q_{\mathbf{w}}})$: an upper bound on $R(G_{Q_{\mathbf{w}}})$ also provides an upper bound on $R(h_{\mathbf{w}})$.

- The prior $P_{\mathbf{w}_p}$ is also an isotropic Gaussian centered on $\mathbf{w}_p$. Consequently:

$$\mathrm{KL}(Q_{\mathbf{w}}\|P_{\mathbf{w}_p}) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2\,.$$

Bayes-equivalent classifiers

- With this choice for $Q_{\mathbf{w}}$, the majority vote $B_{Q_{\mathbf{w}}}$ is the same classifier as $h_{\mathbf{w}}$ since:

$$B_{Q_{\mathbf{w}}}(\mathbf{x}) \;=\; \operatorname{sgn}\left(\mathop{\mathbf{E}}_{\mathbf{v}\sim Q_{\mathbf{w}}} \operatorname{sgn}\left(\mathbf{v}\cdot\phi(\mathbf{x})\right)\right) \;=\; \operatorname{sgn}\left(\mathbf{w}\cdot\phi(\mathbf{x})\right) \;=\; h_{\mathbf{w}}(\mathbf{x})\,.$$

- Thus $R(h_{\mathbf{w}}) = R(B_{Q_{\mathbf{w}}}) \leq 2R(G_{Q_{\mathbf{w}}})$: an upper bound on $R(G_{Q_{\mathbf{w}}})$ also provides an upper bound on $R(h_{\mathbf{w}})$.

- The prior $P_{\mathbf{w}_p}$ is also an isotropic Gaussian centered on $\mathbf{w}_p$. Consequently:

$$\mathrm{KL}(Q_{\mathbf{w}}\|P_{\mathbf{w}_p}) \;=\; \frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2\,.$$

Bayes-equivalent classifiers

- With this choice for $Q_{\mathbf{w}}$, the majority vote $B_{Q_{\mathbf{w}}}$ is the same classifier as $h_{\mathbf{w}}$ since:

$$B_{Q_{\mathbf{w}}}(\mathbf{x}) = \mathrm{sgn}\left(\mathop{\mathbf{E}}_{\mathbf{v}\sim Q_{\mathbf{w}}} \mathrm{sgn}\left(\mathbf{v}\cdot\phi(\mathbf{x})\right)\right) = \mathrm{sgn}\left(\mathbf{w}\cdot\phi(\mathbf{x})\right) = h_{\mathbf{w}}(\mathbf{x})\,.$$

- Thus $R(h_{\mathbf{w}}) = R(B_{Q_{\mathbf{w}}}) \leq 2R(G_{Q_{\mathbf{w}}})$: an upper bound on $R(G_{Q_{\mathbf{w}}})$ also provides an upper bound on $R(h_{\mathbf{w}})$.

- The prior $P_{\mathbf{w}_p}$ is also an isotropic Gaussian centered on $\mathbf{w}_p$. Consequently:

$$\mathrm{KL}(Q_{\mathbf{w}}\|P_{\mathbf{w}_p}) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2\,.$$

Gibbs' risk

We need to compute Gibb's risk $R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}})$ on $(\mathbf{x}, y)$ since:

$$R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) \stackrel{\text{def}}{=} \int_{\mathbb{R}^N} Q_\mathbf{w}(\mathbf{v})\, I(y\mathbf{v}\cdot\boldsymbol{\phi}(\mathbf{x})<0)\; d\mathbf{v}$$

we have:

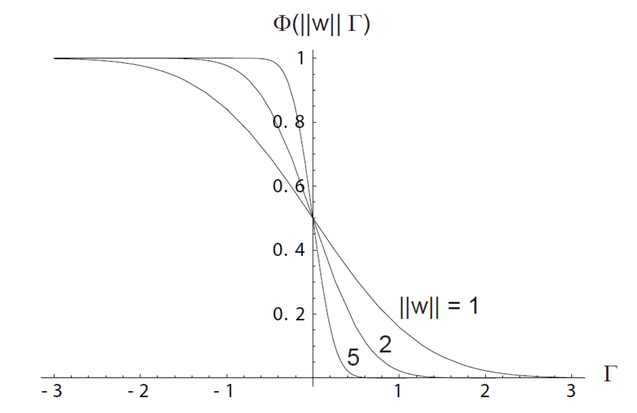
$$R(G_{Q_\mathbf{w}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) \quad \text{and} \quad R_S(G_{Q_\mathbf{w}}) = \tfrac{1}{m}\sum_{i=1}^m R_{(\mathbf{x}_i,y_i)}(G_{Q_\mathbf{w}})\,.$$

Moreover, as in Langford (2005), the Gaussian integral gives:

$$R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) = \Phi\Big(\|\mathbf{w}\|\,\Gamma_\mathbf{w}(\mathbf{x},y)\Big)$$

where: $\Gamma_\mathbf{w}(\mathbf{x},y) \stackrel{\text{def}}{=} \frac{y\mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x})}{\|\mathbf{w}\|\,\|\boldsymbol{\phi}(\mathbf{x})\|}$ and $\Phi(a) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}}\int_a^\infty \exp\left(-\frac{1}{2}x^2\right) dx\,.$

Probit loss



Objective function from Catoni's bound

Recall that, to minimize the Catoni's bound, for fixed C and $\mathbf{w}_p$, we need to find $\mathbf{w}$ that minimizes:

$$C\, m\, R_S(G_{Q_\mathbf{w}}) + \mathrm{KL}(Q_\mathbf{w} \| P_{\mathbf{w}_p})$$

Which, according to preceding slides, corresponds of minimizing

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Objective function from Catoni's bound

Recall that, to minimize the Catoni's bound, for fixed C and $\mathbf{w}_p$, we need to find $\mathbf{w}$ that minimizes:

$$C\, m\, R_S(G_{Q_{\mathbf{w}}}) + \mathrm{KL}(Q_{\mathbf{w}} \| P_{\mathbf{w}_p})$$

Which, according to preceding slides, corresponds of minimizing

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C\sum_{i=1}^{m}\Phi\left(\frac{y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right)+\frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C\sum_{i=1}^{m}\max\Big(0,1-y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)\Big)+\frac{1}{2}\|\mathbf{w}\|^2,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C \sum_{i=1}^{m} \max\Big(0, 1 - y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)\Big) + \frac{1}{2}\|\mathbf{w}\|^2 ,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C \sum_{i=1}^{m} \max\Big(0, 1 - y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)\Big) + \frac{1}{2}\|\mathbf{w}\|^2 ,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C\sum_{i=1}^{m}\Phi\left(\frac{y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right)+\frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C\sum_{i=1}^{m}\max\Big(0,1-y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)\Big)+\frac{1}{2}\|\mathbf{w}\|^2\,,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C\sum_{i=1}^{m}\Phi\left(\frac{y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right)+\frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C\sum_{i=1}^{m}\max\Big(0, 1 - y_i\mathbf{w}\cdot\phi(\mathbf{x}_i)\Big)+\frac{1}{2}\|\mathbf{w}\|^2,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Numerical result [ICML09]

Dataset				(s) SVM		(1) PBGD1			(2) PBGD2			(3) PBGD3	
Name	$\lvert S\rvert$	$\lvert T\rvert$	n	$R_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$
Usvotes	235	200	16	0.055	0.370	0.080	0.117	0.244	**0.050**	0.050	0.153	0.075	0.085
Credit-A	353	300	15	0.183	0.591	**0.150**	0.196	0.341	**0.150**	0.152	0.248	0.160	0.267
Glass	107	107	9	0.178	0.571	**0.168**	0.349	0.539	0.215	0.232	0.430	**0.168**	0.316
Haberman	144	150	3	0.280	0.423	0.280	0.285	0.417	0.327	0.323	0.444	**0.253**	0.250
Heart	150	147	13	0.197	0.513	0.190	0.236	0.441	**0.184**	0.190	0.400	0.197	0.246
Sonar	104	104	60	0.163	0.599	0.250	0.379	0.560	0.173	0.231	0.477	**0.144**	0.243
BreastCancer	343	340	9	**0.038**	0.146	0.044	0.056	0.132	0.041	0.046	0.101	0.047	0.051
Tic-tac-toe	479	479	9	0.081	0.555	0.365	0.369	0.426	0.173	0.193	0.287	**0.077**	0.107
Ionosphere	176	175	34	0.097	0.531	0.114	0.242	0.395	0.103	0.151	0.376	**0.091**	0.165
Wdbc	285	284	30	0.074	0.400	0.074	0.204	0.366	**0.067**	0.119	0.298	0.074	0.210
MNIST:0vs8	500	1916	784	**0.003**	0.257	0.009	0.053	0.202	0.007	0.015	0.058	0.004	0.011
MNIST:1vs7	500	1922	784	0.011	0.216	0.014	0.045	0.161	**0.009**	0.015	0.052	0.010	0.012
MNIST:1vs8	500	1936	784	0.011	0.306	0.014	0.066	0.204	0.011	0.019	0.060	**0.010**	0.024
MNIST:2vs3	500	1905	784	**0.020**	0.348	0.038	0.112	0.265	0.028	0.043	0.096	0.023	0.036
Letter:AvsB	500	1055	16	**0.001**	0.491	0.005	0.043	0.170	0.003	0.009	0.064	**0.001**	0.408
Letter:DvsO	500	1058	16	0.014	0.395	0.017	0.095	0.267	0.024	0.030	0.086	**0.013**	0.031
Letter:OvsQ	500	1036	16	0.015	0.332	0.029	0.130	0.299	0.019	0.032	0.078	**0.014**	0.045
Adult	1809	10000	14	**0.159**	0.535	0.173	0.198	0.274	0.180	0.181	0.224	0.164	0.174
Mushroom	4062	4062	22	**0.000**	0.213	0.007	0.032	0.119	0.001	0.003	0.011	**0.000**	0.001

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
  - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
  - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
  - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
  - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
  - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
  - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
  - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
  - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
    - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
    - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Answer # 1

- Suppose $\mathcal{H} = \{h_1, .., h_n, h_{n+1}, .., h_{2n}\}$ with $h_{i+n} = -h_i$,
- and consider instead, the set of *all the majority votes* over $\mathcal{H}$

$$\mathcal{H}^{MV} \stackrel{\text{def}}{=} \{\operatorname{sgn}(\mathbf{v} \cdot \phi(\mathbf{x})) : \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}\}$$

where $\phi(\mathbf{x}) \stackrel{\text{def}}{=} (h_1(\mathbf{x}), \ldots, h_{2n}(\mathbf{x}))$.

- Then we are back to the linear classifier specialization.

Answer # 1

- Suppose $\mathcal{H} = \{h_1, .., h_n, h_{n+1}, .., h_{2n}\}$ with $h_{i+n} = -h_i$,
- and consider instead, the set of *all the majority votes* over $\mathcal{H}$

$$\mathcal{H}^{MV} \stackrel{\text{def}}{=} \{\mathrm{sgn}\left(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})\right) \,:\, \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}\}$$

where $\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (h_1(\mathbf{x}), \ldots, h_{2n}(\mathbf{x}))$.

- Then we are back to the linear classifier specialization.

Answer # 1

- Suppose $\mathcal{H} = \{h_1, .., h_n, h_{n+1}, .., h_{2n}\}$ with $h_{i+n} = -h_i$,
- and consider instead, the set of *all the majority votes* over $\mathcal{H}$

$$\mathcal{H}^{MV} \stackrel{\text{def}}{=} \{\mathrm{sgn}\left(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})\right) \,:\, \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}\}$$

  where $\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (h_1(\mathbf{x}), \ldots, h_{2n}(\mathbf{x}))$.
- Then we are back to the linear classifier specialization.

Numerical result [ICML09], with decision stumps as weak learners

Dataset				(a) AdaBoost		(1) PBGD1			(2) PBGD2			(3) PBGD3		
Name	$\|S\|$	$\|T\|$	n	$R_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd
Usvotes	235	200	16	**0.055**	0.346	0.085	0.103	0.207	0.060	0.058	0.165	0.060	0.057	0.261
Credit-A	353	300	15	0.170	0.504	0.177	0.243	0.375	0.187	0.191	0.272	**0.143**	0.159	0.420
Glass	107	107	9	0.178	0.636	0.196	0.346	0.562	0.168	0.176	0.395	**0.150**	0.226	0.581
Haberman	144	150	3	**0.260**	0.590	0.273	0.283	0.422	0.267	0.287	0.465	0.273	0.386	0.424
Heart	150	147	13	0.259	0.569	**0.170**	0.250	0.461	0.190	0.205	0.379	0.184	0.214	0.473
Sonar	104	104	60	0.231	0.644	0.269	0.376	0.579	0.173	0.168	0.547	**0.125**	0.209	0.622
BreastCancer	343	340	9	0.053	0.295	**0.041**	0.058	0.129	0.047	0.054	0.104	0.044	0.048	0.190
Tic-tac-toe	479	479	9	0.357	0.483	0.294	0.384	0.462	**0.207**	0.208	0.302	**0.207**	0.217	0.474
Ionosphere	176	175	34	0.120	0.602	0.120	0.223	0.425	0.109	0.129	0.347	**0.103**	0.125	0.557
Wdbc	285	284	30	0.049	0.447	0.042	0.099	0.272	0.049	0.048	0.147	**0.035**	0.051	0.319
MNIST:0vs8	500	1916	784	0.008	0.528	0.015	0.052	0.191	0.011	0.016	0.062	**0.006**	0.011	0.262
MNIST:1vs7	500	1922	784	**0.013**	0.541	0.020	0.055	0.184	0.015	0.016	0.050	0.016	0.017	0.233
MNIST:1vs8	500	1936	784	0.025	0.552	0.037	0.097	0.247	0.027	0.030	0.087	**0.018**	0.037	0.305
MNIST:2vs3	500	1905	784	0.047	0.558	0.046	0.118	0.264	0.040	0.044	0.105	**0.034**	0.048	0.356
Letter:AvsB	500	1055	16	0.010	0.254	0.009	0.050	0.180	**0.007**	0.011	0.065	**0.007**	0.044	0.180
Letter:DvsO	500	1058	16	0.036	0.378	0.043	0.124	0.314	0.033	0.039	0.090	**0.024**	0.038	0.360
Letter:OvsQ	500	1036	16	**0.038**	0.431	0.061	0.170	0.357	0.053	0.053	0.106	0.042	0.049	0.454
Adult	1809	10000	14	**0.149**	0.394	0.168	0.196	0.270	0.169	0.169	0.209	0.159	0.160	0.364
Mushroom	4062	4062	22	**0.000**	0.200	0.046	0.065	0.130	0.016	0.017	0.030	0.002	0.004	0.150

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q}\, y h(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(x,y)) \geq I(M_Q(x,y)<0) \quad \forall Q, x, y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q}\, y h(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)<0) \quad \forall Q, \mathbf{x}, y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q} yh(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)<0) \quad \forall Q,\mathbf{x},y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q} yh(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

  and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)<0) \quad \forall Q,\mathbf{x},y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q} yh(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

  and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)<0) \quad \forall Q,\mathbf{x},y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

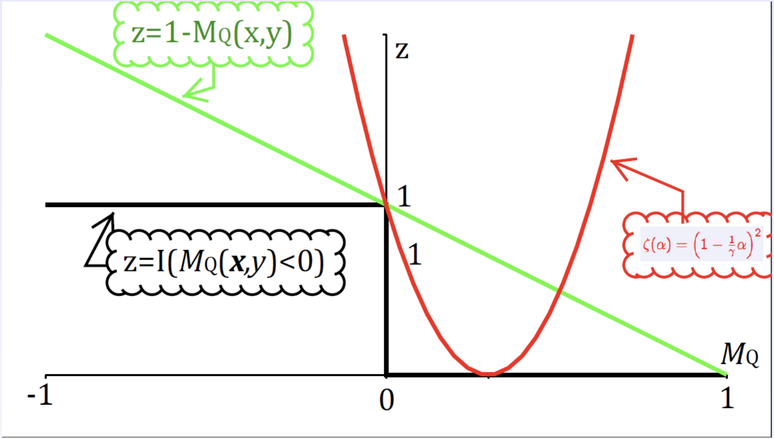
- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q}\, y h(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

  and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)<0) \quad \forall Q,\mathbf{x},y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$



Note: $1 - M_Q(\mathbf{x}, y) = 2R(G_Q)$

Thus the green and the black curves illustrate: $R(B_Q) \leq 2R(G_Q)$

Catoni's bound for a general loss

If we define

$$\begin{aligned}
\zeta_Q &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \zeta_Q(M_Q(\mathbf{x},y)) \\
\widehat{\zeta_Q} &\stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \zeta_Q(M_Q(\mathbf{x}_i,y_i)) \\
c_a &\stackrel{\text{def}}{=} \zeta(1) \\
\bar{k} &= \zeta'(1)
\end{aligned}$$

Catoni's bound become :

Theorem 3.2. *For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, any positive real number C', any loss function $\zeta_Q(\mathbf{x},y)$ defined above, we have*

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ \zeta_Q \le\ g(c_a, C') + \frac{C'}{1-e^{-C'}}\left[\widehat{\zeta_Q} + \frac{2c_a}{mC'}\left[\bar{k}\cdot \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right]\right) \ge 1-\delta,$$

where $g(c_a, C') \stackrel{\text{def}}{=} 1 - c_a + \frac{C'}{1-e^{-C'}}\cdot(c_a - 1)$.

Answer # 2 (cont)

The trick !

- $\zeta_Q(\mathbf{x}, y)$ can be expressed in terms of the risk on example $(\mathbf{x}, y)$ of a Gibbs classifier described by a *transformed* posterior $\overline{Q}$ on $\mathbb{N} \times \mathcal{H}^\infty$

$$\zeta_Q(M_Q(\mathbf{x}, y)) \;=\; c_a \left[M_{\overline{Q}}(\mathbf{x}, y) \right] \,,$$

where $c_a \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k$ and where

$$R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}}) \stackrel{\text{def}}{=} \frac{1}{c_a} \sum_{k=1}^{\infty} |a_k| \mathop{\mathbf{E}}_{h_1 \sim Q} \ldots \mathop{\mathbf{E}}_{h_k \sim Q} I\Big((-y)^k h_1(\mathbf{x}) \ldots h_k(\mathbf{x}) \;=\; -\text{sgn}(a_k) \Big) \,.$$

- Since $R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}})$ is the expectation of boolean random variable, the Catoni's bound holds if we replace (P, Q) by $(\overline{P}, \overline{Q})$

Answer # 2 (cont)

The trick !

- $\zeta_Q(\mathbf{x}, y)$ can be expressed in terms of the risk on example $(\mathbf{x}, y)$ of a Gibbs classifier described by a *transformed* posterior $\overline{Q}$ on $\mathbb{N} \times \mathcal{H}^\infty$

$$\zeta_Q(M_Q(\mathbf{x}, y)) \;=\; c_a \left[M_{\overline{Q}}(\mathbf{x}, y) \right] ,$$

where $c_a \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k$ and where

$$R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}}) \stackrel{\text{def}}{=} \frac{1}{c_a} \sum_{k=1}^{\infty} |a_k| \mathop{\mathbf{E}}_{h_1 \sim Q} \ldots \mathop{\mathbf{E}}_{h_k \sim Q} I\left((-y)^k h_1(\mathbf{x}) \ldots h_k(\mathbf{x}) \;=\; -\mathrm{sgn}(a_k) \right) .$$

- Since $R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}})$ is the expectation of boolean random variable, the Catoni's bound holds if we replace (P, Q) by $(\overline{P}, \overline{Q})$

Answer # 2 (cont)

The trick !

- $\zeta_Q(\mathbf{x}, y)$ can be expressed in terms of the risk on example $(\mathbf{x}, y)$ of a Gibbs classifier described by a *transformed* posterior $\overline{Q}$ on $\mathbb{N}\times\mathcal{H}^\infty$

$$\zeta_Q(M_Q(\mathbf{x}, y)) \;=\; c_a\left[M_{\overline{Q}}(\mathbf{x}, y)\right] \;,$$

where $c_a \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k$ and where

$$R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}}) \stackrel{\text{def}}{=} \frac{1}{c_a}\sum_{k=1}^{\infty} |a_k| \mathop{\mathbf{E}}_{h_1\sim Q} \ldots \mathop{\mathbf{E}}_{h_k\sim Q} I\left((-y)^k h_1(\mathbf{x})\ldots h_k(\mathbf{x}) \;=\; -\mathrm{sgn}(a_k)\right) .$$

- Since $R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}})$ is the expectation of boolean random variable, the Catoni's bound holds if we replace (P, Q) by $(\overline{P}, \overline{Q})$

Minimizing Catoni's bound for a general loss

Minimizing this version of the Catoni's bound is equivalent to finding Q that minimizes

$$f(Q) \stackrel{\text{def}}{=} C \sum_{i=1}^{m} \zeta_Q(\mathbf{x}_i, y_i) + \mathrm{KL}(Q \| P)\,,$$

here: $C \stackrel{\text{def}}{=} C'/(2c_a\overline{k})$.

Minimizing Catoni's bound for a general loss

- To compare the proposed learning algorithms with AdaBoost, we will consider, for $\zeta_Q(\mathbf{x}, y)$, the *exponential loss* given by

$$\exp\left(-\frac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x})\right) = \exp\left(\frac{1}{\gamma}\left[M_Q(\mathbf{x},y)\right]\right).$$

- Because of its simplicity, let us also consider, for $\zeta_Q(\mathbf{x}, y)$, the *quadratic loss* given by

$$\left(\frac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x}) - 1\right)^2 = \left(\frac{1}{\gamma}\, M_Q(\mathbf{x},y) - 1\right)^2.$$

Minimizing Catoni's bound for a general loss

- To compare the proposed learning algorithms with AdaBoost, we will consider, for $\zeta_Q(\mathbf{x}, y)$, the *exponential loss* given by

$$\exp\left(-\frac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x})\right) = \exp\left(\frac{1}{\gamma}\,[M_Q(\mathbf{x},y)]\right).$$

- Because of its simplicity, let us also consider, for $\zeta_Q(\mathbf{x}, y)$, the *quadratic loss* given by

$$\left(\frac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x}) - 1\right)^2 = \left(\frac{1}{\gamma}\, M_Q(\mathbf{x},y) - 1\right)^2.$$

Empirical results (Nips[09])

Dataset				(1) **AdB**	(2) **RR**		(3) **KL-EL**			(4) **KL-QL**		
Name	$\lvert S\rvert$	$\lvert T\rvert$	a	R_T	R_T	C	R_T	C	γ	R_T	C	γ
BreastCancer	343	340	9	0.053	0.050	10	**0.047**	0.1	0.1	**0.047**	0.02	0.4
Liver	170	175	6	0.320	0.309	5	0.360	0.5	0.02	**0.286**	0.02	0.3
Credit-A	353	300	15	0.170	**0.157**	2	0.227	0.1	0.2	0.183	0.02	0.05
Glass	107	107	9	**0.178**	0.206	5	0.187	500	0.01	0.196	0.02	0.01
Haberman	144	150	3	0.260	0.273	100	**0.253**	500	0.2	0.260	0.02	0.5
Heart	150	147	13	0.252	0.197	1	0.211	0.2	0.1	**0.177**	0.05	0.2
Ionosphere	176	175	34	0.120	0.131	0.05	0.120	20	0.0001	**0.097**	0.2	0.1
Letter:AB	500	1055	16	0.010	**0.004**	0.5	0.006	0.1	0.02	0.006	1000	0.1
Letter:DO	500	1058	16	0.036	0.026	0.05	**0.019**	500	0.01	0.020	0.02	0.05
Letter:OQ	500	1036	16	**0.038**	0.045	0.5	0.043	10	0.0001	0.047	0.1	0.05
MNIST:0vs8	500	1916	784	0.008	0.015	0.05	**0.006**	500	0.001	0.015	0.2	0.02
MNIST:1vs7	500	1922	784	0.013	**0.012**	1	0.014	500	0.02	0.014	1000	0.1
MNIST:1vs8	500	1936	784	0.025	0.024	0.2	**0.016**	0.2	0.001	0.031	1	0.02
MNIST:2vs3	500	1905	784	0.047	0.033	0.2	0.035	500	0.0001	**0.029**	0.02	0.05
Mushroom	4062	4062	22	**0.000**	0.001	0.5	**0.000**	10	0.001	**0.000**	1000	0.02
Ringnorm	3700	3700	20	0.043	0.037	0.05	**0.025**	500	0.01	0.039	0.05	0.05
Sonar	104	104	60	0.231	0.192	0.05	0.135	500	0.05	**0.115**	1000	0.1
Usvotes	235	200	16	**0.055**	0.060	2	0.060	0.5	0.1	**0.055**	1000	0.05
Waveform	4000	4000	21	0.085	**0.079**	0.02	0.080	0.2	0.05	0.080	0.02	0.05
Wdbc	285	284	30	0.049	0.049	0.2	**0.039**	500	0.02	0.046	1000	0.1

From $\mathrm{KL}(Q\|P)$ to ℓ_2 regularization

We can recover ℓ_2 regularization if we upper-bound $\mathrm{KL}(Q\|P)$ by a quadratic function. Indeed, if we use

$$q \ln q + \left(\tfrac{1}{n} - q\right) \ln\left(\tfrac{1}{n} - q\right) \le \tfrac{1}{n} \ln \tfrac{1}{2n} + 4n\left(q - \tfrac{1}{2n}\right)^2 \ \forall q \in [0, 1/n]\,,$$

Moreover, if we suppose we have

- $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ with $h_{i+n} = -h_i$
- a uniform prior $\left(P(h_i) = 1/(2n)\right)$
- a posterior distribution Q aligned on the prior P. $\left(Q(h_i) + Q(h_{i+n}) = 1/n \right)$
- and defined: $w_j \stackrel{\text{def}}{=} Q(h_j) - Q(h_{j+n})$

Then,

$$\begin{aligned} \mathrm{KL}(Q\|P) &= \ln(2n) + \textstyle\sum_{i=1}^{n} \left[Q_i \ln Q_i + \left(\tfrac{1}{n} - Q_i\right) \ln\left(\tfrac{1}{n} - Q_i\right) \right] \\ &\le 4n \textstyle\sum_{i=1}^{n} \left(Q_i - \tfrac{1}{2n}\right)^2 \\ &= n \textstyle\sum_{i=1}^{n} w_i^2 \,. \end{aligned}$$

From $\mathrm{KL}(Q\|P)$ to ℓ_2 regularization

We can recover ℓ_2 regularization if we upper-bound $\mathrm{KL}(Q\|P)$ by a quadratic function. Indeed, if we use

$$q \ln q + \left(\tfrac{1}{n} - q\right) \ln\left(\tfrac{1}{n} - q\right) \leq \tfrac{1}{n} \ln \tfrac{1}{2n} + 4n\left(q - \tfrac{1}{2n}\right)^2 \ \forall q \in [0, 1/n]\,,$$

Moreover, if we suppose we have

- $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ with $h_{i+n} = -h_i$
- a uniform prior $\left(P(h_i) = 1/(2n)\right)$
- a posterior distribution Q aligned on the prior P. $\left(Q(h_i) + Q(h_{i+n}) = 1/n \right)$
- and defined: $w_j \stackrel{\text{def}}{=} Q(h_j) - Q(h_{j+n})$

Then,

$$\begin{aligned} \mathrm{KL}(Q\|P) &= \ln(2n) + \sum_{i=1}^{n} \left[Q_i \ln Q_i + \left(\tfrac{1}{n} - Q_i\right) \ln\left(\tfrac{1}{n} - Q_i\right)\right] \\ &\leq 4n \sum_{i=1}^{n} \left(Q_i - \tfrac{1}{2n}\right)^2 \\ &= n \sum_{i=1}^{n} w_i^2\,. \end{aligned}$$

From $\mathrm{KL}(Q\|P)$ to ℓ_2 regularization

We can recover ℓ_2 regularization if we upper-bound $\mathrm{KL}(Q\|P)$ by a quadratic function. Indeed, if we use

$$q \ln q + \left(\tfrac{1}{n} - q\right) \ln\left(\tfrac{1}{n} - q\right) \leq \tfrac{1}{n} \ln \tfrac{1}{2n} + 4n\left(q - \tfrac{1}{2n}\right)^2 \ \forall q \in [0, 1/n],$$

Moreover, if we suppose we have

- $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ with $h_{i+n} = -h_i$
- a uniform prior $\left(P(h_i) = 1/(2n)\right)$
- a posterior distribution Q aligned on the prior P. $\left(Q(h_i) + Q(h_{i+n}) = 1/n \right)$
- and defined: $w_j \stackrel{\text{def}}{=} Q(h_j) - Q(h_{j+n})$

Then,

$$\begin{aligned} \mathrm{KL}(Q\|P) &= \ln(2n) + \textstyle\sum_{i=1}^n \left[Q_i \ln Q_i + \left(\tfrac{1}{n} - Q_i\right) \ln\left(\tfrac{1}{n} - Q_i\right) \right] \\ &\leq 4n \textstyle\sum_{i=1}^n \left(Q_i - \tfrac{1}{2n}\right)^2 \\ &= n \textstyle\sum_{i=1}^n w_i^2 . \end{aligned}$$

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes
$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w}\cdot\mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$
subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1, \ldots, n\}$.

- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.

- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes

$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w}\cdot \mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$

  subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1,\ldots,n\}$.

- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.

- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes
$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w}\cdot\mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$
subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1,\ldots,n\}$.

- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.

- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes
$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w}\cdot\mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$
subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1,\ldots,n\}$.
- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.
- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes

$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w} \cdot \mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$

  subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1, \ldots, n\}$.

- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.

- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

Answer#2 and kernel methods

- Note that in contrast with the approach Answer#1, the approach (Answer#2) can not, as it is presently stated, construct kernel based algorithm.
- For that we need to extend the PAC-Bayes theorem to the sample compression setting (see presentation of Pascal Germain).

Answer#2 and kernel methods

- Note that in contrast with the approach Answer#1, the approach (Answer#2) can not, as it is presently stated, construct kernel based algorithm.
- For that we need to extend the PAC-Bayes theorem to the sample compression setting (see presentation of Pascal Germain).

Conclusion

- Theorem 1, being relatively simple, represent a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
  - this is nice
  - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as "extended".

Conclusion

- Theorem 1, being relatively simple, represent a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
  - this is nice
  - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as "extended".

Conclusion

- Theorem 1, being relatively simple, represent a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
  - this is nice
  - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as "extended".

Conclusion

- Theorem 1, being relatively simple, represent a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
  - this is nice
  - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as “extended”.

Conclusion

- Theorem 1, being relatively simple, represent a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
    - this is nice
    - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as “extended”.

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

QUESTIONS ?