A Strongly Quasiconvex PAC-Bayesian Bound

Yevgeny Seldin

NIPS-2017 Workshop on (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights

Based on joint work with Niklas Thiemann, Christian Igel, and Olivier Wintenberger, ALT 2017
Quick Summary

Two major ways to convexify classification with 0-1 loss

- Convexify the loss
- Work in the space of distributions over $\mathcal{H}$ (PAC-Bayes)
Quick Summary

Two major ways to convexify classification with 0-1 loss

- Convexify the loss
- Work in the space of distributions over $\mathcal{H}$ (PAC-Bayes)

We propose

- A relaxation of the PAC-Bayes-kl bound (Seeger, 2002) and an alternating minimization procedure
- Sufficient conditions for strong quasiconvexity of the bound
  - which guarantee convergence to the global minimum
- Construction of a hypothesis space tailored for the bound
- In our experiments rigorous minimization of the bound was competitive with cross-validation in tuning the trade-off between complexity and empirical performance
Outline

A Very Quick Recap of PAC-Bayesian Analysis

A Strongly Quasiconvex PAC-Bayesian Bound

Construction of a Hypothesis Set

Experiments
Randomized Classifiers

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

Randomized Classifiers

At each round of the game:

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

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

Randomized Classifiers
At each round of the game:

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

Expected loss of $\rho$

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

Empirical loss of $\rho$ on a sample $S$

$$\mathbb{E}_{h \sim \rho}[\hat{L}(h, S)] = \mathbb{E}_{\rho} \left[ \hat{L}(h, S) \right]$$
Approximation-Estimation Perspective
(Bias-Variance)

\[
\hat{L}(h, S) \approx \hat{L}(h', S) \quad \text{and} \quad \pi(h) \approx \pi(h') \quad \Rightarrow \quad \rho(h) \approx \rho(h')
\]

- Reduced variance at the same bias level

Selection from a small \( \mathcal{H} \)
Selection from a large \( \mathcal{H} \)
Approximation-Estimation Perspective
(Bias-Variance)

Randomized Classification

- Avoid selection when not necessary
  - If $\hat{L}(h, S) \approx \hat{L}(h', S)$ and $\pi(h) \approx \pi(h')$, take $\rho(h) \approx \rho(h')$
  - Reduced variance at the same bias level
Kullback-Leibler (KL) divergence $= \text{Relative Entropy}$

**KL divergence**

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

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

**Binary kl divergence**

For two Bernoulli random variables with biases $p$ and $q$

$$\text{kl}(p \| q) = KL(\left[ p, 1 - p \right] \| \left[ q, 1 - q \right])$$
PAC-Bayes-kl Inequality

Theorem (Seeger, 2002)

For any prior $\pi$ over $\mathcal{H}$ and any $\delta \in (0, 1)$, with probability greater than $1 - \delta$ over a random draw of a sample $S$, for all distributions $\rho$ over $\mathcal{H}$ simultaneously:

$$\text{kl} \left( \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] \middle\| \mathbb{E}_\rho \left[ L(h) \right] \right) \leq \frac{\text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{n}}{\delta}}{n}.$$
PAC-Bayes-kl Inequality

Theorem (Seeger, 2002)

For any prior \( \pi \) over \( \mathcal{H} \) and any \( \delta \in (0, 1) \), with probability greater than \( 1 - \delta \) over a random draw of a sample \( S \), for all distributions \( \rho \) over \( \mathcal{H} \) simultaneously:

\[
\text{kl} \left( \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] \bigg\| \mathbb{E}_\rho [L(h)] \right) \leq \frac{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{n}.
\]

Challenge

- The bound is not convex in \( \rho \)
- Common heuristic: replace with a parametrized tradeoff \( \beta n \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] + \text{KL}(\rho \| \pi) \) and tune \( \beta \) by cross-validation
Outline

A Very Quick Recap of PAC-Bayesian Analysis

A Strongly Quasiconvex PAC-Bayesian Bound

Construction of a Hypothesis Set

Experiments
Theorem (PAC-Bayes-\(\lambda\) Inequality)

For any prior \(\pi\) and any \(\delta \in (0, 1)\), with probability greater than \(1 - \delta\), for all \(\rho\) and \(\lambda \in (0, 2)\) simultaneously:

\[
\mathbb{E}_\rho [L(h)] \leq \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{KL(\rho\|\pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right)n}.
\]
Relaxation of PAC-Bayes-kl
Based on refined Pinsker’s inequality

**Theorem (PAC-Bayes-λ Inequality)**

For any prior $\pi$ and any $\delta \in (0, 1)$, with probability greater than $1 - \delta$, for all $\rho$ and $\lambda \in (0, 2)$ simultaneously:

$$
\mathbb{E}_\rho [L(h)] \leq \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] \leq \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n}.
$$

For the optimal $\lambda$ this leads to

$$
\mathbb{E}_\rho [L(h)] \leq \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] + \sqrt{\frac{2\mathbb{E}_\rho \left[ \hat{L}(h, S) \right] \left(\text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{n}}{\delta}\right)}{n} + \frac{2 \left(\text{KL}(\rho\|\pi) + \ln \frac{2\sqrt{n}}{\delta}\right)}{n}}
$$

“Fast convergence rate”
Alternating Minimization of PAC-Bayes-λ

\[ \mathbb{E}_{\rho} [L(h)] \leq \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho || \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]

\[ \mathcal{F}(\rho, \lambda) \]
Alternating Minimization of PAC-Bayes-\(\lambda\)

\[
\mathbb{E}_\rho [L(h)] \leq \frac{\mathbb{E}_\rho [\hat{L}(h, S)]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta} + \frac{\lambda}{2} n}{\lambda (1 - \frac{\lambda}{2}) n}
\]

\(\mathcal{F}(\rho, \lambda)\)

- For a fixed \(\lambda\) the bound is convex in \(\rho\) and minimized by

\[
\rho\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h,S)}}{\mathbb{E}_\pi [e^{-\lambda n\hat{L}(h',S)}]}
\]
Alternating Minimization of PAC-Bayes-$\lambda$

$$\mathbb{E}_\rho [L(h)] \leq \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] \leq \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left( 1 - \frac{\lambda}{2} \right) n} \quad \mathcal{F}(\rho, \lambda)$$

- For a fixed $\lambda$ the bound is convex in $\rho$ and minimized by

$$\rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n\hat{L}(h', S)} \right]}$$

- For a fixed $\rho$ the bound is convex in $\lambda$ and minimized by

$$\lambda = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}} + 1 + 1}$$
Alternating Minimization of PAC-Bayes-$\lambda$

$$
\mathbb{E}_\rho [L(h)] \leq \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left( 1 - \frac{\lambda}{2} \right) n}
$$

\[\mathcal{F}(\rho, \lambda)\]

- For a fixed $\lambda$ the bound is convex in $\rho$ and minimized by

$$
\rho_{\lambda}(h) = \frac{\pi(h) e^{\lambda n \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{\lambda n \hat{L}(h', S)} \right]}
$$

- For a fixed $\rho$ the bound is convex in $\lambda$ and minimized by

$$
\lambda = \frac{2}{\sqrt{\frac{2n \mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}} + 1 + 1}}
$$

- $\mathcal{F}(\rho, \lambda)$ is not necessarily jointly convex in $\rho$ and $\lambda$
Simplification 1

$$\mathcal{F}(\rho, \lambda) = \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] + \frac{\text{KL}(\rho \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda (1 - \frac{\lambda}{2}) n}$$

$$\rho \lambda(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n \hat{L}(h', S)} \right]}$$
Simplification 1

\[ F(\rho, \lambda) = \mathbb{E}_\rho \left[ \hat{L}(h, S) \right] + \frac{\text{KL}(\rho || \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]

\[ \rho_\lambda(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n \hat{L}(h', S)} \right]} \]

\[ F(\lambda) = F(\rho_\lambda, \lambda) = \frac{\mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho_\lambda || \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]
Simplification 1

\[ \mathcal{F}(\rho, \lambda) = \frac{\mathbb{E}_\rho \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]

\[ \rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h', S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n\hat{L}(h', S)} \right]} \]

\[ \mathcal{F}(\lambda) = \mathcal{F}(\rho_\lambda, \lambda) = \frac{\mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho_\lambda \parallel \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]

One-dimensional function
Simplification 2

\[ \mathcal{F}(\lambda) = \mathcal{F}(\rho_{\lambda}, \lambda) = \frac{\mathbb{E}_{\rho_{\lambda}} \left[ \hat{L}(h, S') \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho_{\lambda} \parallel \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right) n} \]

\[ \rho_{\lambda}(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\mathbb{E}_{\pi} \left[ e^{-\lambda n \hat{L}(h', S)} \right]} \]
Simplification 2

\[ \mathcal{F}(\lambda) = \mathcal{F}(\rho_\lambda, \lambda) = \frac{\mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho_\lambda \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left(1 - \frac{\lambda}{2}\right)n} \]

\[ \rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h,S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n\hat{L}(h',S)} \right]} \]

\[ \text{KL}(\rho_\lambda \| \pi) = \mathbb{E}_{\rho_\lambda} \left[ \ln \frac{\rho_\lambda(h)}{\pi(h)} \right] = \mathbb{E}_{\rho_\lambda} \left[ \ln \frac{e^{-n\lambda \hat{L}(h,S)}}{\mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h',S)} \right]} \right] \]

\[ = -n\lambda \mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] - \ln \mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h,S)} \right] \]
Simplification 2

\[
\mathcal{F}(\lambda) = \mathcal{F}(\rho_\lambda, \lambda) = \frac{\mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho_\lambda \| \pi) + \ln \frac{2\sqrt{n}}{\delta}}{\lambda \left( 1 - \frac{\lambda}{2} \right) n}
\]

\[
\rho_\lambda(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n \hat{L}(h', S)} \right]}
\]

\[
\text{KL}(\rho_\lambda \| \pi) = \mathbb{E}_{\rho_\lambda} \left[ \ln \frac{\rho_\lambda(h)}{\pi(h)} \right] = \mathbb{E}_{\rho_\lambda} \left[ \ln \frac{e^{-n\lambda \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h', S)} \right]} \right]
\]

\[
= -n\lambda \mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] - \ln \mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h, S)} \right]
\]

\[
\mathcal{F}(\lambda) = \frac{-\ln \mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h, S)} \right] + \ln \frac{2\sqrt{n}}{\delta}}{n\lambda(1 - \lambda/2)}
\]
Strong Quasiconvexity - Sufficient Condition

Theorem (Strong Quasiconvexity)

If at least one of the two conditions

\[ 2 \text{KL}(\rho_\lambda \| \pi) + \ln \frac{4n}{\delta^2} > \lambda^2 n^2 \text{Var}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] \]

or

\[ \mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] > (1 - \lambda)n \text{Var}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] \]

is satisfied for all \( \lambda \in \left[ \sqrt{\frac{\ln \frac{2\sqrt{n}}{\delta}}{n}}, 1 \right] \), then \( F(\lambda) \) is strongly quasiconvex for \( \lambda \in (0, 1] \) and alternating minimization converges to the global minimum of \( F \).
\[ \mathcal{F}(\lambda) = \frac{-\ln \mathbb{E}_\pi \left[ e^{-n\lambda \hat{L}(h,S)} \right]}{n\lambda(1 - \lambda/2)} + \ln \frac{2\sqrt{n}}{\delta} \]
“Weak Separation” Sufficient Condition for Strong Quasiconvexity

\[
\hat{L}(h^*, S) + a \quad \hat{L}(h^*, S) + b
\]
“Weak Separation” Sufficient Condition for Strong Quasiconvexity

\[ \hat{L}(h^*, S) + a \quad \hat{L}(h^*, S) + b \]

\[ \hat{L}(h^*, S) \quad 1 \]

**Theorem (Weak Separation)**

Let \( \mathcal{H} \) be finite with \( |\mathcal{H}| = m \) and \( \pi(h) \) uniform. Let \( a = \frac{\sqrt{\ln \frac{4n}{\delta^2}}}{n\sqrt{3}} \) and \( b \approx \frac{\ln(3mn)}{\sqrt{n \ln \frac{2\sqrt{n}}{\delta}}} \). If the number of hypotheses for which

\[ \hat{L}(h, S) \in \left( \hat{L}(h^*, S) + a, \hat{L}(h^*, S) + b \right) \]

is at most \( \frac{e^2}{12} \ln \frac{4n}{\delta^2} \) then \( \mathcal{F}(\lambda) \) is strongly quasiconvex and alternating minimization converges to the global minimum.
Proof Highlights

By the Strong Quasiconvexity Theorem, if $\text{Var}_{\rho \lambda} \left[ \hat{L}(h, S) \right]$ is “small” then $\mathcal{F}(\lambda)$ is strongly quasiconvex.

Let $\Delta_h = \hat{L}(h, S) - \hat{L}(h^*, S)$

$$
\text{Var}_{\rho \lambda} \left[ \hat{L}(h, S) \right] \leq \mathbb{E}_{\rho \lambda} \left[ \Delta_h^2 \right] = \sum_h \rho \lambda(h) \Delta_h^2
$$

$$
= \sum_h \Delta_h^2 e^{-n\lambda \Delta_h} \Bigg/ \sum_h e^{-n\lambda \Delta_h}
$$
Breaking the Quasiconvexity

- It is possible to break the quasiconvexity...
- ... but one has to work hard for it
It is possible to break the quasiconvexity...

... but one has to work hard for it

For example, taking $n = 200$, $\delta = 0.25$, $m = 2.7 \cdot 10^6$, $\Delta_h = 0.1$ and uniform $\pi$ breaks it
It is possible to break the quasiconvexity...

... but one has to work hard for it

For example, taking $n = 200$, $\delta = 0.25$, $m = 2.7 \cdot 10^6$, $\Delta_h = 0.1$ and uniform $\pi$ breaks it

In all our experiments $\mathcal{F}(\lambda)$ was convex even when the “weak separation” sufficient condition was violated

So it might be possible to relax the sufficient condition further
Outline

A Very Quick Recap of PAC-Bayesian Analysis

A Strongly Quasiconvex PAC-Bayesian Bound

Construction of a Hypothesis Set

Experiments
Computation of the normalization of $\rho_\lambda$ can be prohibitively expensive

$$\rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n \hat{L}(h,S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n \hat{L}(h',S)} \right]}$$
Challenge

Computation of the normalization of $\rho_\lambda$ can be prohibitively expensive

$$\rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h,S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n\hat{L}(h',S)} \right]}$$

Parametrization of $\rho$ may break the convexity
**Challenge**

Computation of the normalization of $\rho_\lambda$ can be prohibitively expensive

$$
\rho_\lambda(h) = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\mathbb{E}_\pi \left[ e^{-\lambda n \hat{L}(h', S)} \right]} = \frac{\pi(h) e^{-\lambda n \hat{L}(h, S)}}{\sum_{h'} \pi(h') e^{-\lambda n \hat{L}(h', S)}}
$$

Parametrization of $\rho$ may break the convexity

**Solution**

- Work with finite $\mathcal{H}$
- We need a “powerful” finite $\mathcal{H}$
Construction of a finite sample-dependent $\mathcal{H}$

- Select $m = |\mathcal{H}|$ subsamples of $r$ points each
- Train a model $h$ on $r$ points and validate on $n - r$ points
- Validation loss: $\hat{L}^{\text{val}}(h)$

Adapted Bound

$E_{\rho}[L(h)] \leq E_{\rho}[\hat{L}^{\text{val}}(h, S)] + \frac{1}{2} KL(\rho \parallel \pi) + \ln n - r + 1 \delta(n - r) \lambda(1 - \lambda^2)$

Special Case: $k$-fold cross-validation

Most computational advantage is achieved by "inverse CV"
Construction of a finite sample-dependent $\mathcal{H}$

- Select $m = |\mathcal{H}|$ subsamples of $r$ points each
- Train a model $h$ on $r$ points and validate on $n - r$ points
- Validation loss: $\hat{L}^{\text{val}}(h)$

Adapted Bound

$$
\mathbb{E}_\rho [L(h)] \leq \frac{\mathbb{E}_\rho [\hat{L}^{\text{val}}(h, S)]}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(\rho \| \pi) + \ln \frac{n-r+1}{\delta}}{(n-r)\lambda (1 - \frac{\lambda}{2})}
$$
Construction of a finite sample-dependent $\mathcal{H}$

- Select $m = |\mathcal{H}|$ subsamples of $r$ points each
- Train a model $h$ on $r$ points and validate on $n - r$ points
- Validation loss: $\hat{L}_{\text{val}}(h)$

Adapted Bound

$$\mathbb{E}_\rho [L(h)] \leq \mathbb{E}_\rho \left[ \hat{L}_{\text{val}}(h, S) \right] + \frac{\text{KL}(\rho || \pi) + \ln \frac{n-r+1}{\delta}}{(n-r)\lambda \left(1 - \frac{\lambda}{2}\right)}$$

Special Case: $k$-fold cross-validation
Most computational advantage is achieved by “inverse CV”
Outline

A Very Quick Recap of PAC-Bayesian Analysis

A Strongly Quasiconvex PAC-Bayesian Bound

Construction of a Hypothesis Set

Experiments
Experiments

We compare

- Kernel-SVM trained by cross-validation
- $\rho$-weighting of multiple “weak” SVMs trained on $d + 1$ samples

More precisely, we apply $\rho$-weighted aggregation

$$MV_\rho(x) = \text{sign}\left(\sum h_\rho(h(x))\right)$$

but in our case there was no significant difference between $L(MV_\rho)$ and $E_\rho[L(h)]$. 
Experiments

We compare

- Kernel-SVM trained by cross-validation
- $\rho$-weighting of multiple “weak” SVMs trained on $d + 1$ samples
  - More precisely, we apply $\rho$-weighted aggregation

$$MV_\rho(x) = \text{sign} \left( \sum_h \rho(h) h(x) \right)$$

but in our case there was no significant difference between $L(MV_\rho)$ and $\mathbb{E}_\rho [L(h)]$
Rough Runtime Comparison

$k$-fold cross-validation of kernel SVMs

\[ k \left( \left( n^{2+} \right)_{\text{training}} + V_{\text{validation}} \right) \approx kn^{2+} \]
Rough Runtime Comparison

\( k \)-fold cross-validation of kernel SVMs

\[
k \left( \frac{n^{2+}}{\text{training}} + \frac{V}{\text{validation}} \right) \approx kn^{2+}
\]

PAC-Bayesian aggregation of kernel SVMs

For \( r = d + 1 \) and \( m = n \):

\[
m \left( \frac{r^{2+}}{\text{training}} + \frac{rn}{\text{validation}} + \frac{A}{\text{aggregation}} \right) \approx mrn \approx dn^2
\]
Rough Runtime Comparison

\( k \)-fold cross-validation of kernel SVMs

\[
k \left( n^{2+} + V \right) \approx k n^{2+}
\]

PAC-Bayesian aggregation of kernel SVMs

For \( r = d + 1 \) and \( m = n \):

\[
m \left( r^{2+} + rn + A \right) \approx mrn \approx dn^2
\]

Computational Speed-up!
Experiments

(a) Ionosphere $n = 200, r = d + 1 = 35$.

(b) Waveform $n = 2000, r = d + 1 = 41$.

(c) Breast cancer $n = 340, r = d + 1 = 11$.

(d) AvsB $n = 1000, r = d + 1 = 17$. 
Summary

We proposed

- A relaxation of the PAC-Bayes-kl bound (Seeger, 2002)
- An alternating minimization procedure
- Sufficient conditions for strong quasiconvexity
  - which guarantee convergence to the global minimum
- Construction of $\mathcal{H}$
- In our experiments rigorous minimization of the bound was competitive with cross-validation in tuning the trade-off between complexity and empirical performance
We proposed

- A relaxation of the PAC-Bayes-kl bound (Seeger, 2002)
- An alternating minimization procedure
- Sufficient conditions for strong quasiconvexity
  - which guarantee convergence to the global minimum
- Construction of $\mathcal{H}$
- In our experiments rigorous minimization of the bound was competitive with cross-validation in tuning the trade-off between complexity and empirical performance

Rigorous minimization of a theoretical bound competitive with cross-validation!
What’s next?

Improved Sufficient Conditions

- In practice the bound was strongly convex even when the “weak separation” sufficient condition was violated.
- Relax the sufficient condition
  - We have dropped some terms when going from the Strong Quasiconvexity Theorem to the Weak Separation Condition.
Theorem (Strong Quasiconvexity)

If at least one of the two conditions

\[ 2 \text{KL}(\rho_\lambda \| \pi) + \ln \frac{4n}{\delta^2} > \lambda^2 n^2 \text{Var}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] \]

or

\[ \mathbb{E}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] > (1 - \lambda) n \text{Var}_{\rho_\lambda} \left[ \hat{L}(h, S) \right] \]

is satisfied for all \( \lambda \in \left[ \sqrt{\frac{\ln \frac{2\sqrt{n}}{\delta}}{n}}, 1 \right] \), then \( F(\lambda) \) is strongly quasiconvex for \( \lambda \in (0, 1] \) and alternating minimization converges to the global minimum of \( F \).
What’s next?

Improved Sufficient Conditions

- In practice the bound was strongly convex even when the “weak separation” sufficient condition was violated.
- Relax the sufficient condition
  - We have dropped some terms when going from the Strong Quasiconvexity Theorem to the Weak Separation Condition
What’s next?

Improved Sufficient Conditions

- In practice the bound was strongly convex even when the “weak separation” sufficient condition was violated.
- Relax the sufficient condition
  - We have dropped some terms when going from the Strong Quasiconvexity Theorem to the Weak Separation Condition

Improved Analysis of the Weighted Majority Vote

- Combine the results with improved analysis of weighted majority vote (the “C-bound”)
  - Lacasse, Laviolette, Marchand, Germain, and Usunier, NIPS, 2007
  - Laviolette, Marchand, Roy, ICML, 2011
  - Germain, Lacasse, Laviolette, Marchand, Roy, JMLR, 2015