Data Dependent Priors for Stable Learning

John Shawe-Taylor
University College London

Work with Emilio Parrado-Hernández, Amiran Ambroladze, Francois Laviolette, Guy Lever and Shiliang Sun

PAC-Bayesian Workshop, NIPS 2017
Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks
Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks

The stability analysis of Bousquet and Elisseeff provides an inspiration for this approach
Background

- Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks
- The stability analysis of Bousquet and Elisseeff provides an inspiration for this approach
- Link between stability and data distribution priors that could point the way to further analysis of stable learning
Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks

The stability analysis of Bousquet and Elisseeff provides an inspiration for this approach

Link between stability and data distribution priors that could point the way to further analysis of stable learning

Show that SVM weight vectors produced by random training sets are concentrated
Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks

The stability analysis of Bousquet and Elisseeff provides an inspiration for this approach

Link between stability and data distribution priors that could point the way to further analysis of stable learning

Show that SVM weight vectors produced by random training sets are concentrated

Gives tighter bounds based on data distribution defined prior
Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks

The stability analysis of Bousquet and Elisseeff provides an inspiration for this approach

Link between stability and data distribution priors that could point the way to further analysis of stable learning

Show that SVM weight vectors produced by random training sets are concentrated

Gives tighter bounds based on data distribution defined prior

Begin by reviewing PAC-Bayes and introducing data dependence
The PAC-Bayes theorem involves a class of classifiers $C$ together with a prior distribution $P$ and posterior $Q$ over $C$.
The PAC-Bayes theorem involves a class of classifiers $C$ together with a prior distribution $P$ and posterior $Q$ over $C$.

The distribution $P$ must be chosen before learning, but the bound holds for all choices of $Q$, hence $Q$ does not need to be the classical Bayesian posterior.
The PAC-Bayes theorem involves a class of classifiers $C$ together with a prior distribution $P$ and posterior $Q$ over $C$. The distribution $P$ must be chosen before learning, but the bound holds for all choices of $Q$, hence $Q$ does not need to be the classical Bayesian posterior. The bound holds for all (prior) choices of $P$ – hence it’s validity is not affected by a poor choice of $P$ though the quality of the resulting bound may be – contrast with standard Bayes analysis which only holds if the prior assumptions are correct.
Being a frequentist (PAC) style result we assume an unknown distribution $D$ on the input space $X$. 
Definitions for main result

Error measures

- Being a frequentist (PAC) style result we assume an unknown distribution $\mathcal{D}$ on the input space $X$.
- $\mathcal{D}$ is used to generate the labelled training samples i.i.d., i.e. $S \sim \mathcal{D}^m$
Definitions for main result

Error measures

- Being a frequentist (PAC) style result we assume an unknown distribution \( \mathcal{D} \) on the input space \( X \).
- \( \mathcal{D} \) is used to generate the labelled training samples i.i.d., i.e. \( S \sim \mathcal{D}^m \).
- It is also used to measure generalisation error \( c_{\mathcal{D}} \) of a classifier \( c \):
  \[
  c_{\mathcal{D}} = \Pr_{(x,y) \sim \mathcal{D}}(c(x) \neq y)
  \]
Being a frequentist (PAC) style result we assume an unknown distribution $\mathcal{D}$ on the input space $X$.

$\mathcal{D}$ is used to generate the labelled training samples i.i.d., i.e. $S \sim \mathcal{D}^m$.

It is also used to measure generalisation error $c_D$ of a classifier $c$:

$$c_D = \Pr_{(x,y) \sim \mathcal{D}}(c(x) \neq y)$$

The empirical generalisation error is denoted $\hat{c}_S$:

$$\hat{c}_S = \frac{1}{m} \sum_{(x,y) \in S} I[c(x) \neq y]$$

where $I[\cdot]$ indicator function.
The result is concerned with bounding the performance of a probabilistic classifier that given a test input $x$ chooses a classifier $c \sim Q$ (the posterior) and returns $c(x)$.
The result is concerned with bounding the performance of a probabilistic classifier that given a test input $x$ chooses a classifier $c \sim Q$ (the posterior) and returns $c(x)$.

We are interested in the relation between two quantities:

$$Q_D = \mathbb{E}_{c \sim Q}[c_D]$$

the true error rate of the probabilistic classifier and

$$\hat{Q}_S = \mathbb{E}_{c \sim Q}[\hat{c}_S]$$

its empirical error rate.
Note that this does not bound the posterior average but we have

$$\Pr_{(x,y) \sim D}(\text{sgn}(\mathbb{E}_{c \sim Q}[c(x)]) \neq y) \leq 2Q_D.$$ 

since for any point $x$ misclassified by $\text{sgn}(\mathbb{E}_{c \sim Q}[c(x)])$ the probability of a random $c \sim Q$ misclassifying is at least 0.5.
PAC-Bayes Theorem

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

$$\text{KL}(\hat{Q}_S \| Q_D) \leq \frac{\text{KL}(Q \| P) + \ln((m + 1) / \delta)}{m}$$

where $\text{KL}$ is the KL divergence between distributions

$$\text{KL}(Q \| P) = \mathbb{E}_{c \sim Q} \left[ \ln \frac{Q(c)}{P(c)} \right]$$

with $\hat{Q}_S$ and $Q_D$ considered as distributions on $\{0, +1\}$. 
We will choose the prior and posterior distributions to be Gaussians with unit variance.
Linear classifiers

- We will choose the prior and posterior distributions to be Gaussians with unit variance.
- The prior $P$ will be centered at the origin with unit variance.
We will choose the prior and posterior distributions to be Gaussians with unit variance.

The prior $P$ will be centered at the origin with unit variance.

The specification of the centre for the posterior $Q(\mathbf{w}, \mu)$ will be by a unit vector $\mathbf{w}$ and a scale factor $\mu$. 
PAC-Bayes Bound for SVM (1/2)

Prior $P$ is Gaussian $\mathcal{N}(0, 1)$
PAC-Bayes Bound for SVM (1/2)

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

Prior $P$ is Gaussian $\mathcal{N}(0, 1)$
Posterior is in the direction $w$
PAC-Bayes Bound for SVM (1/2)

- **Prior** $P$ is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the direction $w$
- at distance $\mu$ from the origin
PAC-Bayes Bound for SVM (1/2)

- **Prior** \( P \) is Gaussian \( \mathcal{N}(0, 1) \)
- Posterior is in the **direction** \( w \)
- at **distance** \( \mu \) from the origin
- **Posterior** \( Q \) is Gaussian
Form of the SVM bound

Note that bound holds for all posterior distributions so that we can choose $\mu$ to optimise the bound.
Form of the SVM bound

- Note that bound holds for all posterior distributions so that we can choose $\mu$ to optimise the bound.
- If we define the inverse of the KL by

$$KL^{-1}(q, A) = \max\{\rho : KL(q||\rho) \leq A\}$$

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

$$\Pr(\langle w, \phi(x) \rangle \neq y) \leq 2 \min_{\mu} KL^{-1}\left(\mathbb{E}_m[\tilde{F}(\mu \gamma(x, y))], \frac{\mu^2/2 + \ln \frac{m+1}{\delta}}{m}\right)$$
Learning the prior (1/3)

- Bound depends on the **distance between prior and posterior**
Learning the prior (1/3)

- Bound depends on the distance between prior and posterior
- Better prior (closer to posterior) would lead to tighter bound
Learning the prior (1/3)

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior $P$ with part of the data
Learning the prior (1/3)

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tightly bound**
- **Learn** the prior $P$ with part of the data
- Introduce the learnt prior **in the bound**
Learning the prior (1/3)

- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior $P$ with part of the data
- Introduce the learnt prior **in the bound**
- Compute stochastic error with **remaining data**
Solve SVM with \textit{subset of patterns}
New prior for the SVM (3/3)

- Solve SVM with **subset of patterns**
- Prior in the direction $w_r$

![Diagram showing a subset of patterns and a prior direction $w_r$.]
New prior for the SVM (3/3)

- Solve SVM with **subset of patterns**
- Prior in the **direction** \( w_r \)
- **Posterior** like PAC-Bayes Bound
New prior for the SVM (3/3)

- Solve SVM with **subset of patterns**
- Prior in the direction $w_r$
- **Posterior** like PAC-Bayes Bound
- **New bound** proportional to $\text{KL}(P||Q)$
SVM performance may be tightly bounded by

\[
\text{KL}(\hat{Q}_S(w, \mu) \parallel Q_D(w, \mu)) \leq \frac{0.5 \| \mu w - \eta w_r \|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]
New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

\[
\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}
\]

- \(Q_D(w, \mu)\) true performance of the classifier
SVM performance may be **tightly** bounded by

\[
\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\| \mu w - \eta w_r \|^2 + \ln \left( \frac{(m-r+1)J}{\delta} \right)}{m-r}
\]
SVM performance may be **tightly** bounded by

\[
\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \left(\frac{(m-r+1)J}{\delta}\right)}{m - r}
\]

- \(\hat{Q}_S(w, \mu)\) stochastic measure of the training error on remaining data

\[
\hat{Q}(w, \mu)_S = \mathbb{E}_{m-r}[\tilde{F}(\mu \gamma(x, y))]
\]
New Bound for the SVM (2/3)

SVM performance may be tightly bounded by

$$\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\|\mu w - \eta w_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m - r}$$
New Bound for the SVM (2/3)

SVM performance may be \textit{tightly} bounded by

$$\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\|\mu w - \eta w_r\|^2}{m - r} + \ln \left(\frac{(m-r+1)J}{\delta}\right)$$

- \(0.5\|\mu w - \eta w_r\|^2\) distance between prior and posterior
New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

\[
\text{KL}(\hat{Q}_S(\mathbf{w}, \mu) \| Q_D(\mathbf{w}, \mu)) \leq \frac{0.5\|\mu \mathbf{w} - \eta \mathbf{w}_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m-r}
\]
SVM performance may be **tightly** bounded by

\[
\text{KL}(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{0.5\| \mu w - \eta w_r \|^2 + \ln \frac{(m-r+1)J}{\delta}}{m-r}
\]

- Penalty term only dependent on the remaining data \(m - r\)
Determine the **prior** with a subset of the training examples to obtain $w_r$
1. Determine the **prior** with a subset of the training examples to obtain $w_r$.

2. Solve optimisation to minimise bound: **p-SVM** giving $w$. 

---

**p-SVM**

John Shawe-Taylor University College London

Data Dependent Priors for Stable Learning
1. Determine the **prior** with a subset of the training examples to obtain $w_r$

2. Solve optimisation to minimise bound: **p-SVM** giving $w$

3. **Margin** for the stochastic classifier $\hat{Q}_s$

$$\gamma(x_j, y_j) = \frac{y_j w^T \phi(x_j)}{\|\phi(x_j)\| \|w\|} \quad j = 1, \ldots, m - r$$
1. Determine the **prior** with a subset of the training examples to obtain $w_r$.

2. Solve optimisation to minimise bound: **p-SVM** giving $w$.

3. **Margin** for the stochastic classifier $\hat{Q}_s$

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

4. **Linear search** to obtain the optimal value of $\mu$. This introduces an insignificant extra penalty term.
Bound for $\eta$-prior-SVM

- Prior is elongated along the line of $w_r$ but spherical with variance 1 in other directions

$$\text{KL}(\hat{Q}_S \setminus R(w, \mu) \| Q_D(w, \mu)) \leq 0.5 (\ln(\tau^2) + \tau^{-2} - 1 + P\|w_r(\mu w - w_r)^2/\tau^2 + P\perp w_r(\mu w)^2) + \ln(m - r + 1)$$
Bound for $\eta$-prior-SVM

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

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

$$
\text{KL}(\hat{Q}_{S \setminus R}(w, \mu) \parallel Q_D(w, \mu)) \leq \frac{0.5(\ln(\tau^2) + \tau^{-2} - 1 + P_{w_r}(\mu w - w_r)^2/\tau^2 + P_{w_r}^\perp(\mu w)^2) + \ln(\frac{m-r+1}{\delta})}{m - r}
$$
Model Selection with the new bound: setup

Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
Model Selection with the new bound: setup

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

UCI datasets

Select $C$ and $\sigma$ that lead to minimum Classification Error (CE)

For X-F XV select the pair that minimize the validation error
Model Selection with the new bound: setup

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

<table>
<thead>
<tr>
<th>Problem</th>
<th>SVM</th>
<th>( \eta \text{Prior SVM} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2FCV</td>
<td>10FCV</td>
</tr>
<tr>
<td>digits</td>
<td>Bound</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.007</td>
</tr>
<tr>
<td>waveform</td>
<td>Bound</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.090</td>
</tr>
<tr>
<td>pima</td>
<td>Bound</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.244</td>
</tr>
<tr>
<td>ringnorm</td>
<td>Bound</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.016</td>
</tr>
<tr>
<td>spam</td>
<td>Bound</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>0.066</td>
</tr>
</tbody>
</table>
The idea of using a data distribution defined prior was pioneered by Catoni who looked at these distributions:

\[
P(h) := \frac{1}{Z} e^{-\gamma \text{risk}(h)}
\]

\[
Q(h) := \frac{1}{Z} e^{-\gamma \hat{\text{risk}}(h)}
\]

These distributions are hard to work with since we cannot apply the bound to a single weight vector, but the bounds can be very tight:
The idea of using a data distribution defined prior was pioneered by Catoni who looked at these distributions:

- $P$ and $Q$ are Gibbs-Boltzmann distributions

$$p(h) := \frac{1}{Z'} e^{-\gamma \text{risk}(h)}$$

$$q(h) := \frac{1}{Z} e^{-\gamma \hat{\text{risk}}_S(h)}$$
The idea of using a data distribution defined prior was pioneered by Catoni who looked at these distributions:
- \( P \) and \( Q \) are Gibbs-Boltzmann distributions

\[
p(h) := \frac{1}{Z} e^{-\gamma \text{risk}(h)} \quad q(h) := \frac{1}{Z} e^{-\gamma \text{risk}_S(h)}
\]

These distributions are hard to work with since we cannot apply the bound to a single weight vector, but the bounds can be very tight:

\[
KL_+(\hat{Q}_S(\gamma) \parallel Q_D(\gamma)) \leq \frac{1}{m} \left( \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{8\sqrt{m}}{\delta}} + \frac{\gamma^2}{4m} + \ln \frac{4\sqrt{m}}{\delta} \right)
\]

as it appears we can choose \( \gamma \) small even for complex classes.
Let’s try something simple to motivate the idea.

Consider the Gaussian prior centred on the weight vector: $w \sim \mathcal{N}(\mu, \Sigma)$. Note that we do not know this vector, but it is nonetheless fixed independently of the training sample. We can compute a sample-based estimate of this vector as $\hat{w} \sim \mathcal{N}(\mu_{\text{est}}, \Sigma_{\text{est}})$.
Let’s try something simple to motivate the idea

Consider the Gaussian prior centred on the weight vector:

$$w_p = \mathbb{E}[y\phi(x)]$$
Let’s try something simple to motivate the idea
Consider the Gaussian prior centred on the weight vector:

\[ w_p = \mathbb{E}[y\phi(x)] \]

Note that we do not know this vector, but it is nonetheless fixed independently of the training sample.
Let’s try something simple to motivate the idea

Consider the Gaussian prior centred on the weight vector:

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

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

We can compute a sample based estimate of this vector as

$$\hat{w}_p = \mathbb{E}_S[y\phi(x)]$$
With probability $1 - \frac{\delta}{2}$ we have

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

- With probability $1 - \delta/2$ we have
  $$\|\hat{w}_p - w_p\| \leq \frac{R}{\sqrt{m}} \left( 2 + \sqrt{2 \ln \frac{2}{\delta}} \right).$$

- Proof relies on independence of examples and the fact the vector is a simple sum.
With probability $1 - \delta/2$ we have
\[
\|\hat{w}_p - w_p\| \leq \frac{R}{\sqrt{m}} \left( 2 + \sqrt{2 \ln \frac{2}{\delta}} \right).
\]

Proof relies on independence of examples and the fact the vector is a simple sum

We can therefore w.h.p. upper bound KL divergence between prior $P$, an isotropic Gaussian at $w_p$, and posterior $Q$, an isotropic Gaussian at $w$ by
\[
\frac{1}{2} \left( \|w - \hat{w}_p\| + \frac{R}{\sqrt{m}} \left( 2 + \sqrt{2 \ln \frac{2}{\delta}} \right) \right)^2.
\]
Resulting bound

- Giving the following bound on generalisation:

\[
KL_+(\hat{Q}_S(w, \mu) \| Q_D(w, \mu)) \leq \frac{1}{2} \left( \| \mu w - \eta \hat{w}_p \| + \eta \frac{R}{\sqrt{m}} \left( 2 + \sqrt{2 \ln \frac{2}{\delta}} \right) \right)^2 + \ln \frac{2(m+1)}{\delta} \\
m
\]

with probability $1 - \delta$.

- Values of the bounds for an SVM.

<table>
<thead>
<tr>
<th>Prob.</th>
<th>PAC-Bayes</th>
<th>PrPAC</th>
<th>(\tau)-PrPAC</th>
<th>(\mathbb{E}) PrPAC</th>
<th>(\tau)-(\mathbb{E}) PrPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>han</td>
<td>0.175 ± 0.001</td>
<td>0.107 ± 0.004</td>
<td>0.108 ± 0.005</td>
<td>0.157 ± 0.001</td>
<td>0.176 ± 0.001</td>
</tr>
<tr>
<td>wav</td>
<td>0.203 ± 0.001</td>
<td>0.185 ± 0.005</td>
<td>0.184 ± 0.005</td>
<td>0.202 ± 0.001</td>
<td>0.205 ± 0.001</td>
</tr>
<tr>
<td>pim</td>
<td>0.424 ± 0.003</td>
<td>0.420 ± 0.015</td>
<td>0.423 ± 0.014</td>
<td>0.428 ± 0.003</td>
<td>0.433 ± 0.003</td>
</tr>
<tr>
<td>rin</td>
<td>0.203 ± 0.000</td>
<td>0.110 ± 0.004</td>
<td>0.110 ± 0.004</td>
<td>0.201 ± 0.001</td>
<td>0.204 ± 0.000</td>
</tr>
<tr>
<td>spa</td>
<td>0.254 ± 0.001</td>
<td>0.198 ± 0.006</td>
<td>0.198 ± 0.006</td>
<td>0.249 ± 0.001</td>
<td>0.255 ± 0.001</td>
</tr>
</tbody>
</table>
Consider the Gaussian prior (with isotropic variance $1$) centred on the weight vector:

$$w_p = \mathbb{E}_{S \sim D^m}[A_S]$$
Expected SVM as prior

- Consider the Gaussian prior (with isotropic variance $1$) centred on the weight vector:

$$w_p = E_{S \sim D^m}[A_S]$$

- Following Bousquet et al we use the SVM with hinge loss:

$$A_S = \arg \min_w \frac{1}{m} \sum_{i=1}^{m} \ell(g_w, (x_i, y_i)) + \frac{\lambda}{2} \|w\|^2 \quad (1)$$
Consider the Gaussian prior (with isotropic variance 1) centred on the weight vector:

\[ \mathbf{w}_p = \mathbb{E}_{S \sim \mathcal{D}^m}[A_S] \]

Following Bousquet et al we use the SVM with hinge loss:

\[ A_S = \arg \min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \ell(g_{\mathbf{w}}, (x_i, y_i)) + \frac{\lambda}{2} \| \mathbf{w} \|^2 \]  

(1)

Loss function is 1-Lipschitz and \( \lambda > 0 \) gives concentration of SVM weight vectors: with prob at least \( 1 - \delta \)

\[ g(S) = \| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \| \leq \frac{1}{\lambda \sqrt{m}} \left( 3 + \sqrt{\frac{1}{2} \ln \frac{1}{\delta}} \right) \]
Proof outline

First use McDiarmid inequality on

\[ g(S) = \| A_S - \mathbb{E}_\tilde{S}[A_{\tilde{S}}] \| \]

to show this is concentrated around its expectation - follows from Bousquet et al’s results
Proof outline

- First use McDiarmid inequality on

$$g(S) = \| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \|$$

- to show this is concentrated around its expectation - follows from Bousquet et al’s results

- Next step is to bound $\mathbb{E} \left[ \| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \| \right]$
Proof outline

- First use McDiarmid inequality on

\[ g(S) = \|A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}]\| \]

to show this is concentrated around its expectation - follows from Bousquet et al’s results

- Next step is to bound \( \mathbb{E} [\|A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}]\|] \)

- Would like to use the same idea as for the sum of random vectors
Proof outline

First use McDiarmid inequality on

\[ g(S) = \| A_S - \mathbb{E}_S[A_S] \| \]

to show this is concentrated around its expectation - follows from Bousquet et al’s results

Next step is to bound \( \mathbb{E} [\| A_S - \mathbb{E}_\tilde{S}[A_{\tilde{S}}] \|] \)

Would like to use the same idea as for the sum of random vectors

- observe SVM weight has dual representation as sum, but dual variables vary
Proof outline

- First use McDiarmid inequality on
  \[ g(S) = \| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \| \]
  to show this is concentrated around its expectation - follows from Bousquet et al’s results
- Next step is to bound \[ \mathbb{E} \left[ \| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \| \right] \]
- Would like to use the same idea as for the sum of random vectors
  - observe SVM weight has dual representation as sum, but dual variables vary
  - can bound sum of expected values of dual variables
Proof outline

- First use McDiarmid inequality on

\[ g(S) = \|A_S - \mathbb{E}_S[A_S]\| \]

- to show this is concentrated around its expectation - follows from Bousquet et al’s results

- Next step is to bound \( \mathbb{E} [\|A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}]\|] \)

- Would like to use the same idea as for the sum of random vectors
  - observe SVM weight has dual representation as sum, but dual variables vary
  - can bound sum of expected values of dual variables
  - can also show this sum is close to true SVM vector
Resulting bound

We obtain a bound for which the KL term is $O(1/m^2)$: with probability $1 - \delta$:

$$\text{KL}_+ (\hat{Q}_S(A_S, 1) \| Q_D(A_S, 1)) \leq \frac{1}{2\lambda^2 m^2} \left( 3 + \sqrt{\frac{1}{2} \ln \frac{2}{\delta}} \right)^2 + \frac{1}{m} \ln \left( \frac{2(m + 1)}{\delta} \right)$$
Resulting bound

- We obtain a bound for which the KL term is $O(1/m^2)$: with probability $1 - \delta$:

$$\text{KL}_+(\hat{Q}_S(A_S, 1) \| Q_D(A_S, 1)) \leq \frac{1}{2\lambda^2 m^2} \left( 3 + \sqrt{\frac{1}{2} \ln \frac{2}{\delta}} \right)^2 + \frac{1}{m} \ln \left( \frac{2(m + 1)}{\delta} \right)$$

- Compared with Bousquet et al bound:

$$R \leq R_{\text{emp}} + \frac{1}{\lambda m} + \left( 1 + \frac{2}{\lambda} \right) \sqrt{\frac{\ln(1/\delta)}{2m}}$$
Implications

- Cost of generalisation is expected difference between average weight vector from random training sets and specific training set

This suggests we may be able to learn in very flexible spaces such as those used in Deep Learning provided we can show weights are concentrated around an expected value. Given the many equivalent solutions in deep architectures, this will not be true from the beginning of learning but stability suggests it will hold after initial 'burn in'.
Implications

- Cost of generalisation is expected difference between average weight vector from random training sets and specific training set.
- This suggests we may be able to learn in very flexible spaces such as those used in Deep Learning provided we can show weights are concentrated around an expected value.

Given the many equivalent solutions in deep architectures, this will not be true from the beginning of learning but stability suggests it will hold after initial 'burn in'.
Implications

- Cost of generalisation is expected difference between average weight vector from random training sets and specific training set.
- This suggests we may be able to learn in very flexible spaces such as those used in Deep Learning provided we can show weights are concentrated around an expected value.
- Given the many equivalent solutions in deep architectures this will not be true from the beginning of learning but stability suggests will hold after initial ‘burn in’.
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
Concluding remarks

- Investigation of learning of the prior of the distribution of classifiers
- Data distribution defined priors considered:

Data distribution defined priors considered:

- Ideal Gibbs-Boltzmann distribution
- Simple expectation of $y \phi(x)$
- Expectation of complete SVM
- For complete SVM use stability analysis to show that weight vectors are concentrated around their expectation
- Suggests we might be able to extend the analysis to the weight updates given by SGD in Deep Learning
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
- Data distribution defined priors considered:
  - ideal Gibbs-Boltzmann distribution
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
- Data distribution defined priors considered:
  - ideal Gibbs-Boltzmann distribution
  - simple expectation of $y\phi(x)$,
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
- Data distribution defined priors considered:
  - ideal Gibbs-Boltzmann distribution
  - simple expectation of $y\phi(x)$,
  - expectation of complete SVM
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
- Data distribution defined priors considered:
  - ideal Gibbs-Boltzmann distribution
  - simple expectation of $y\phi(x)$,
  - expectation of complete SVM
- For complete SVM use stability analysis to show that weight vectors are concentrated around their expectation
Concluding remarks

- Investigation of **learning of the prior** of the distribution of classifiers
- Data distribution defined priors considered:
  - ideal Gibbs-Boltzmann distribution
  - simple expectation of $y\phi(x)$,
  - expectation of complete SVM
- For complete SVM use stability analysis to show that weight vectors are concentrated around their expectation
- Suggests we might be able to extend the analysis to the weight updates given by SGD in Deep Learning