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

Background

- Renewed interest in stability in connection with Stochastic Gradient Descent for training Deep Networks

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

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

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

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
- Show that SVM weight vectors produced by random training sets are concentrated
- Gives tighter bounds based on data distribution defined prior

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
- 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

Definitions for main result
Prior and posterior distributions

- The PAC-Bayes theorem involves a class of classifiers $\mathcal{C}$ together with a prior distribution P and posterior Q over $\mathcal{C}$

Definitions for main result
Prior and posterior distributions

- The PAC-Bayes theorem involves a class of classifiers $\mathcal{C}$ together with a prior distribution P and posterior Q over $\mathcal{C}$
- 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

Definitions for main result
Prior and posterior distributions

- The PAC-Bayes theorem involves a class of classifiers $\mathcal{C}$ together with a prior distribution P and posterior Q over $\mathcal{C}$
- 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

Definitions for main result
Error measures

- Being a frequentist (PAC) style result we assume an unknown distribution $\mathcal{D}$ 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:

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

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:

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

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

$$\hat{c}_S = \frac{1}{m} \sum_{(x,y)\in S} I[c(x) \neq y] \quad \text{where } I[\cdot] \text{ indicator function.}$$

Definitions for main result
Assessing the posterior

- 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 $c(x)$

Definitions for main result
Assessing the posterior

- 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:

$$Q_{\mathcal{D}} = \mathbb{E}_{c \sim Q}[c_{\mathcal{D}}]$$

the true error rate of the probabilistic classifier and

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

its empirical error rate

Definitions for main result
Generalisation error

- Note that this does not bound the posterior average but we have
$$\Pr_{(x,y)\sim\mathcal{D}}(\operatorname{sgn}(\mathbb{E}_{c\sim Q}[c(x)]) \neq y) \leq 2Q_{\mathcal{D}}.$$
since for any point x misclassified by $\operatorname{sgn}(\mathbb{E}_{c\sim Q}[c(x)])$ 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 δ, then with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$, all posteriors Q satisfy

$$\mathrm{KL}(\hat{Q}_S \| Q_\mathcal{D}) \leq \frac{\mathrm{KL}(Q\|P) + \ln((m+1)/\delta)}{m}$$

where KL is the KL divergence between distributions

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

with $\hat{Q}_S$ and $Q_\mathcal{D}$ considered as distributions on $\{0, +1\}$.

Linear classifiers

- We will choose the prior and posterior distributions to be Gaussians with unit variance.

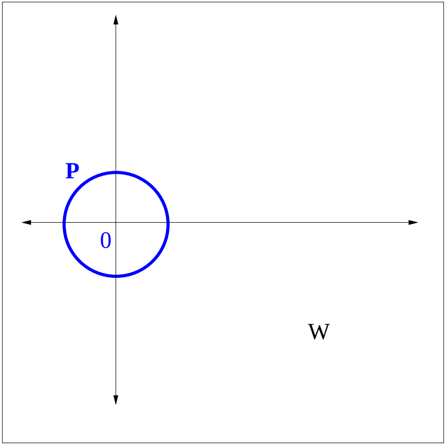
Linear classifiers

- 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

Linear classifiers

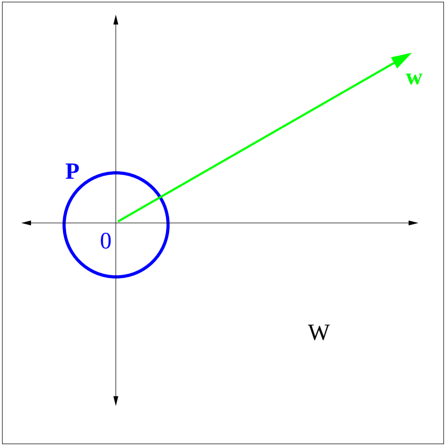
- We will choose the prior and posterior distributions to be Gaussians with unit variance.
- The prior P will be centered at the origin with unit variance
- The specification of the centre for the posterior $Q(\boldsymbol{w}, \mu)$ will be by a unit vector $\mathbf{w}$ and a scale factor μ.

PAC-Bayes Bound for SVM (1/2)



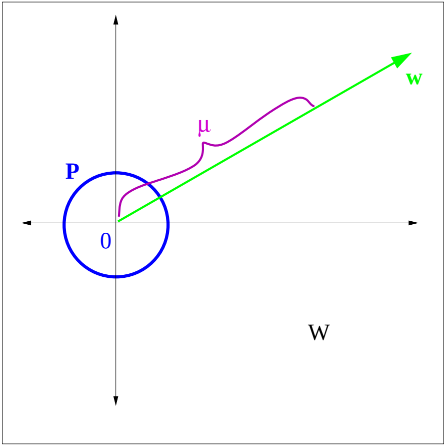
- **Prior** P is Gaussian $\mathcal{N}(0,1)$
-
-
-

PAC-Bayes Bound for SVM (1/2)



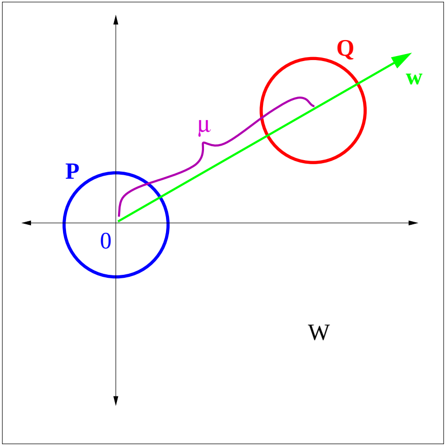
- **Prior** P is Gaussian $\mathcal{N}(0,1)$
- Posterior is in the **direction w**
-
-

PAC-Bayes Bound for SVM (1/2)



- **Prior** P is Gaussian $\mathcal{N}(0, 1)$
- Posterior is in the **direction w**
- at **distance** μ 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** μ 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 μ to optimise the bound

Form of the SVM bound

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

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

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

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

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

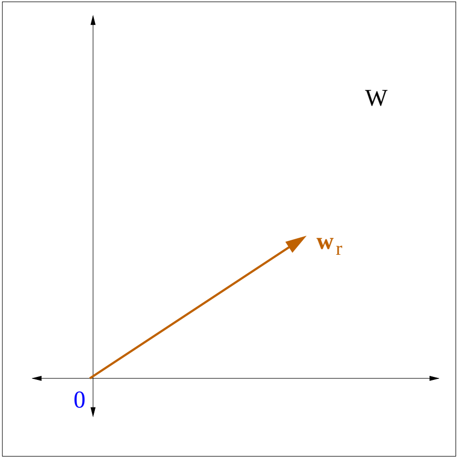
Learning the prior (1/3)

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

Learning the prior (1/3)

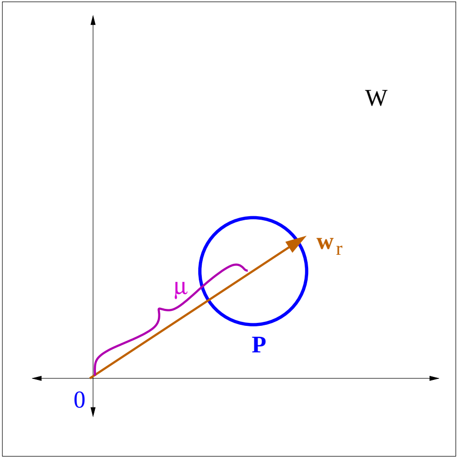
- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior P with part of the data
- Introduce the learnt prior **in the bound**
- Compute stochastic error with **remaining data**

New prior for the SVM (3/3)



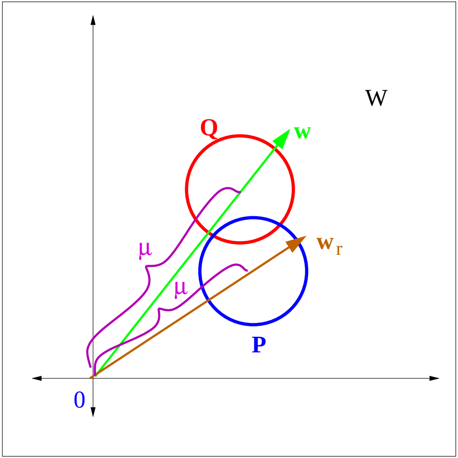
- Solve SVM with **subset of patterns**
-
-
-

New prior for the SVM (3/3)



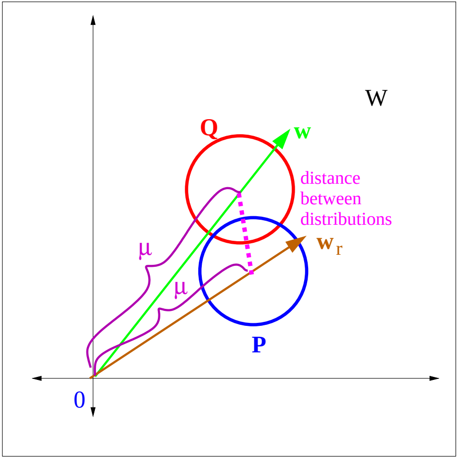
- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
-
-

New prior for the SVM (3/3)



- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
- **Posterior** like PAC-Bayes Bound
-

New prior for the SVM (3/3)



- Solve SVM with **subset of patterns**
- Prior in the **direction** $\mathbf{w}_r$
- **Posterior** like PAC-Bayes Bound
- **New bound** proportional to $\mathrm{KL}(P\|Q)$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|\boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \le \frac{0.5\|\mu\boldsymbol{w}-\eta\boldsymbol{w}_r\|^2+\ln\frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu) \| \boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \leq \frac{0.5\|\mu\boldsymbol{w} - \eta\boldsymbol{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

- $Q_{\mathcal{D}}(\boldsymbol{w},\mu)$ true performance of the classifier

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\boxed{\hat{Q}_S(\boldsymbol{w}, \mu)} \| Q_{\mathcal{D}}(\boldsymbol{w}, \mu)) \leq \frac{0.5\|\mu \boldsymbol{w} - \eta \boldsymbol{w}_r\|^2 + \ln \frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\boxed{\hat{Q}_S(\boldsymbol{w},\mu)}\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{0.5\|\mu\boldsymbol{w}-\eta\boldsymbol{w}_r\|^2+\ln\frac{(m-r+1)J}{\delta}}{m-r}$$

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

$$\hat{Q}(\boldsymbol{w},\mu)_S = \mathbb{E}_{m-r}[\tilde{F}(\mu\gamma(\boldsymbol{x},y))]$$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{0.5\|\mu\boldsymbol{w} - \eta\boldsymbol{w}_r\|^2} + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{0.5\|\mu\boldsymbol{w}-\eta\boldsymbol{w}_r\|^2} + \ln\frac{(m-r+1)J}{\delta}}{m-r}$$

- $0.5\|\mu\boldsymbol{w}-\eta\boldsymbol{w}_r\|^2$ distance between prior and posterior

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{0.5\|\mu\boldsymbol{w}-\eta\boldsymbol{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{\boxed{m-r}}$$

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{0.5\|\mu\boldsymbol{w} - \eta\boldsymbol{w}_r\|^2 + \ln\frac{(m-r+1)J}{\delta}}{\boxed{m-r}}$$

- Penalty term only dependent on the remaining data $m - r$

p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\boldsymbol{w}_r$

p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\boldsymbol{w}_r$
2. Solve optimisation to minimise bound: **p-SVM** giving $\boldsymbol{w}$

p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\boldsymbol{w}_r$
2. Solve optimisation to minimise bound: **p-SVM** giving $\boldsymbol{w}$
3. **Margin** for the stochastic classifier $\hat{Q}_s$

$$\gamma(\boldsymbol{x}_j, y_j) = \frac{y_j \boldsymbol{w}^T \phi(\boldsymbol{x}_j)}{\|\phi(\boldsymbol{x}_j)\| \|\boldsymbol{w}\|} \qquad j = 1, \ldots, m - r$$

p-SVM

1. Determine the **prior** with a subset of the training examples to obtain $\boldsymbol{w}_r$
2. Solve optimisation to minimise bound: **p-SVM** giving $\boldsymbol{w}$
3. **Margin** for the stochastic classifier $\hat{Q}_s$

$$\gamma(\boldsymbol{x}_j, y_j) = \frac{y_j \boldsymbol{w}^T \phi(\boldsymbol{x}_j)}{\|\phi(\boldsymbol{x}_j)\| \|\boldsymbol{w}\|} \qquad j = 1, \dots, m - r$$

4. **Linear search** to obtain the optimal value of μ. This introduces an insignificant extra penalty term

Bound for η-prior-SVM

- Prior is elongated along the line of $\mathbf{w}_r$ but spherical with variance 1 in other directions

Bound for η-prior-SVM

- Prior is elongated along the line of $\mathbf{w}_r$ but spherical with variance 1 in other directions
- Optimisation costs only distance from line defined by $\mathbf{w}_r$
- Posterior again on the line of solution $\mathbf{w}$ at a distance μ chosen to optimise the bound.

Bound for η-prior-SVM

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

$$\mathrm{KL}(\hat{Q}_{S\setminus R}(\mathbf{w},\mu)\|Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq$$
$$\frac{0.5(\ln(\tau^2) + \tau^{-2} - 1 + P^{\|}_{\mathbf{w}_r}(\mu\mathbf{w} - \mathbf{w}_r)^2/\tau^2 + P^{\perp}_{\mathbf{w}_r}(\mu\mathbf{w})^2) + \ln(\frac{m-r+1}{\delta})}{m-r}$$

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 Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ 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 σ 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

		Classifier					
		SVM				ηPrior SVM	
Problem		2FCV	10FCV	PAC	PrPAC	PrPAC	τ-PrPAC
digits	Bound	–	–	0.175	0.107	0.050	0.047
	CE	0.007	0.007	0.007	0.014	0.010	0.009
waveform	Bound	–	–	0.203	0.185	0.178	0.176
	CE	0.090	0.086	0.084	0.088	0.087	0.086
pima	Bound	–	–	0.424	0.420	0.428	0.416
	CE	0.244	0.245	0.229	0.229	0.233	0.233
ringnorm	Bound	–	–	0.203	0.110	0.053	0.050
	CE	0.016	0.016	0.018	0.018	0.016	0.016
spam	Bound	–	–	0.254	0.198	0.186	0.178
	CE	0.066	0.063	0.067	0.077	0.070	0.072

Defining the prior through the data distribution

- The idea of using a data distribution defined prior was pioneered by Catoni who looked at these distributions:

Defining the prior through the data distribution

- 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

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h)} \qquad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_S(h)}$$

Defining the prior through the data distribution

- 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

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h)} \qquad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_S(h)}$$

- 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:

$$KL_+(\hat{Q}_S(\gamma)||Q_{\mathcal{D}}(\gamma)) \leq \frac{1}{m}\left(\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{8\sqrt{m}}{\delta}} + \frac{\gamma^2}{4m} + \ln\frac{4\sqrt{m}}{\delta}\right)$$

as it appears we can choose γ small even for complex classes.

Data distribution dependent prior

- Let's try something simple to motivate the idea

Data distribution dependent prior

- Let's try something simple to motivate the idea
- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_p = \mathbb{E}[y\phi(\mathbf{x})]$$

Data distribution dependent prior

- Let's try something simple to motivate the idea
- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_p = \mathbb{E}[y\phi(\mathbf{x})]$$

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

Data distribution dependent prior

- Let's try something simple to motivate the idea
- Consider the Gaussian prior centred on the weight vector:

$$\mathbf{w}_p = \mathbb{E}[y\phi(\mathbf{x})]$$

- 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

$$\hat{\mathbf{w}}_p = \mathbb{E}_S[y\phi(\mathbf{x})]$$

Estimating the KL divergence

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

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

Estimating the KL divergence

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

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

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

Estimating the KL divergence

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

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

- 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 $\mathbf{w}_p$, and posterior Q, an isotropic Gaussian at $\mathbf{w}$ by

$$\frac{1}{2} \left(\|\mathbf{w} - \hat{\mathbf{w}}_p\| + \frac{R}{\sqrt{m}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}} \right) \right)^2$$

Resulting bound

- Giving the following bound on generalisation:

$$KL_+(\hat{Q}_S(\mathbf{w},\mu)\|Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq$$

$$\frac{\frac{1}{2}\left(\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p\|+\eta\frac{R}{\sqrt{m}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right)\right)^2+\ln\frac{2(m+1)}{\delta}}{m}$$

  with probability $1-\delta$.
- Values of the bounds for an SVM.

Prob.	PAC-Bayes	PrPAC	τ-PrPAC	$\mathbb{E}$ PrPAC	τ-$\mathbb{E}$ PrPAC
han	0.175 ± 0.001	$\mathbf{0.107 \pm 0.004}$	$\mathbf{0.108 \pm 0.005}$	0.157 ± 0.001	0.176 ± 0.001
wav	0.203 ± 0.001	$\mathbf{0.185 \pm 0.005}$	$\mathbf{0.184 \pm 0.005}$	0.202 ± 0.001	0.205 ± 0.001
pim	0.424 ± 0.003	0.420 ± 0.015	0.423 ± 0.014	0.428 ± 0.003	0.433 ± 0.003
rin	0.203 ± 0.000	$\mathbf{0.110 \pm 0.004}$	$\mathbf{0.110 \pm 0.004}$	0.201 ± 0.001	0.204 ± 0.000
spa	0.254 ± 0.001	$\mathbf{0.198 \pm 0.006}$	$\mathbf{0.198 \pm 0.006}$	0.249 ± 0.001	0.255 ± 0.001

Expected SVM as prior

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

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

Expected SVM as prior

- 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:

$$A_S = \arg\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \ell(g_{\mathbf{w}}, (\mathbf{x}_i, y_i)) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \qquad (1)$$

Expected SVM as prior

- 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:

$$A_S = \arg\min_{\mathbf{w}} \frac{1}{m}\sum_{i=1}^{m} \ell(g_{\mathbf{w}}, (\mathbf{x}_i, y_i)) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \qquad (1)$$

- 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

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

  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

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

  to show this is concentrated around its expectation - follows from Bousquet et al's results
- Next step is to bound $\mathbb{E}\left[\left\| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \right\| \right]$
- Would like to use the same idea as for the sum of random vectors

Proof outline

- First use McDiarmid inequality on

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

  to show this is concentrated around its expectation - follows from Bousquet et al's results
- Next step is to bound $\mathbb{E}\left[\left\| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \right\|\right]$
- 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

Proof outline

- First use McDiarmid inequality on

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

  to show this is concentrated around its expectation - follows from Bousquet et al's results
- Next step is to bound $\mathbb{E}\left[\left\|A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}]\right\|\right]$
- 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

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

  to show this is concentrated around its expectation - follows from Bousquet et al's results
- Next step is to bound $\mathbb{E}\left[\left\| A_S - \mathbb{E}_{\tilde{S}}[A_{\tilde{S}}] \right\|\right]$
- 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
  - 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 $1-\delta$:

$$
\begin{aligned}
\mathrm{KL}_+(\hat{Q}_S(A_S,1)\|Q_{\mathcal{D}}(A_S,1)) \quad &\leq \quad \frac{1}{2\lambda^2 m^2}\left(3+\sqrt{\frac{1}{2}\ln\frac{2}{\delta}}\right)^2 + \\
&+\frac{1}{m}\ln\left(\frac{2(m+1)}{\delta}\right)
\end{aligned}
$$

Resulting bound

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

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

- Compared with Bousquet et al bound:

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

Implications

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

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

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:

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 $y\phi(\mathbf{x})$,

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 $y\phi(\mathbf{x})$,
  - 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 $y\phi(\mathbf{x})$,
  - 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 $y\phi(\mathbf{x})$,
  - 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