Data Dependent Priors in PAC-Bayes Bounds

John Shawe-Taylor
University College London

Joint work with Emilio Parrado-Hernández and Amiran Ambroladze

August, 2010

Evidence and generalisation

- Link between evidence and generalisation hypothesised by McKay
- First formal link was obtained by S-T & Williamson (1997): PAC Analysis of a Bayes Estimator
- Bound on generalisation in terms of the volume of the sphere that can be inscribed in the version space – included a dependence on the dimensionality of the space
- Used Luckiness framework – a data-dependent style of frequentist bound also used to bound generalisation of SVMs for which no dependence on the dimensionality is needed, just on the margin

Evidence and generalisation

- Link between evidence and generalisation hypothesised by McKay
- First formal link was obtained by S-T & Williamson (1997): PAC Analysis of a Bayes Estimator
- Bound on generalisation in terms of the volume of the sphere that can be inscribed in the version space – included a dependence on the dimensionality of the space
- Used Luckiness framework – a data-dependent style of frequentist bound also used to bound generalisation of SVMs for which no dependence on the dimensionality is needed, just on the margin

Evidence and generalisation

- Link between evidence and generalisation hypothesised by McKay
- First formal link was obtained by S-T & Williamson (1997): PAC Analysis of a Bayes Estimator
- Bound on generalisation in terms of the volume of the sphere that can be inscribed in the version space – included a dependence on the dimensionality of the space
- Used Luckiness framework – a data-dependent style of frequentist bound also used to bound generalisation of SVMs for which no dependence on the dimensionality is needed, just on the margin

Evidence and generalisation

- Link between evidence and generalisation hypothesised by McKay
- First formal link was obtained by S-T & Williamson (1997): PAC Analysis of a Bayes Estimator
- Bound on generalisation in terms of the volume of the sphere that can be inscribed in the version space – included a dependence on the dimensionality of the space
- Used Luckiness framework – a data-dependent style of frequentist bound also used to bound generalisation of SVMs for which no dependence on the dimensionality is needed, just on the margin

PAC-Bayes Theorem

- First version proved by McAllester in 1999
- Improved proof and bound due to Seeger in 2002 with application to Gaussian processes
- Application to SVMs by Langford and S-T also in 2002
- Excellent tutorial by Langford appeared in 2005 in JMLR

PAC-Bayes Theorem

- First version proved by McAllester in 1999
- Improved proof and bound due to Seeger in 2002 with application to Gaussian processes
- Application to SVMs by Langford and S-T also in 2002
- Excellent tutorial by Langford appeared in 2005 in JMLR

PAC-Bayes Theorem

- First version proved by McAllester in 1999
- Improved proof and bound due to Seeger in 2002 with application to Gaussian processes
- Application to SVMs by Langford and S-T also in 2002
- Excellent tutorial by Langford appeared in 2005 in JMLR

PAC-Bayes Theorem

- First version proved by McAllester in 1999
- Improved proof and bound due to Seeger in 2002 with application to Gaussian processes
- Application to SVMs by Langford and S-T also in 2002
- Excellent tutorial by Langford appeared in 2005 in JMLR

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

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

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

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
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
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
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 $c(x)$
- 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
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

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

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 $\{0, +1\}$.

Ingredients of proof (1/3)

1. $$\Pr_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{c\sim P}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}\leq\frac{m+1}{\delta}\right)\geq 1-\delta$$

- This follows from considering the expectation divided into probability of particular empirical error for any c:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}=\sum_k Pr_{S\sim\mathcal{D}^m}(\hat{c}_S=k)\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_{S'}=k)}=m+1.$$

Taking expectations wrt to c and reversing the expectations

$$\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{c\sim P}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}=m+1$$

and the result follows from Markov's inequality.

Ingredients of proof (1/3)

1.

$$\Pr_{S\sim\mathcal{D}^m}\left(\mathbb{E}_{c\sim P}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}\leq\frac{m+1}{\delta}\right)\geq 1-\delta$$

- This follows from considering the expectation divided into probability of particular empirical error for any c:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}=\sum_k Pr_{S\sim\mathcal{D}^m}(\hat{c}_S=k)\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_{S'}=k)}=m+1.$$

Taking expectations wrt to c and reversing the expectations

$$\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{c\sim P}\frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}=m+1$$

and the result follows from Markov's inequality.

Ingredients of proof (2/3)

1. $$\frac{\mathbb{E}_{c\sim Q} \ln \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}}{m} \geq \mathrm{KL}(\hat{Q}_S \| Q_{\mathcal{D}})$$

- This follows by considering the probabilities that the two empirical estimates are equal, applying the relative entropy Chernoff bound and then using the concavity of the KL divergence as a function of both arguments.

Ingredients of proof (2/3)

1. $$\frac{\mathbb{E}_{c\sim Q} \ln \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_S=\hat{c}_{S'})}}{m} \geq \mathrm{KL}(\hat{Q}_S \| Q_{\mathcal{D}})$$

- This follows by considering the probabilities that the two empirical estimates are equal, applying the relative entropy Chernoff bound and then using the concavity of the KL divergence as a function of both arguments.

Ingredients of proof (3/3)

1. Consider the distribution

$$P_G(c) = \frac{1}{\Pr_{S' \sim \mathcal{D}^m}(\hat{c}_{S'} = \hat{c}_S)\mathbb{E}_{d \sim P} \frac{1}{\Pr_{S' \sim \mathcal{D}^m}(\hat{d}_S = \hat{d}_{S'})}} P(c)$$

Ingredients of proof (2/3)

$$
\begin{aligned}
0 &\leq \mathrm{KL}(Q\|P_G) \\
&= \mathrm{KL}(Q\|P) - \mathbb{E}_{c\sim Q} \ln \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_{S'} = \hat{c}_S)} \\
&\qquad + \ln \mathbb{E}_{d\sim P} \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{d}_S = \hat{d}_{S'})}
\end{aligned}
$$

Ingredients of proof (3/3)

$$
\begin{aligned}
m\mathrm{KL}(\hat{Q}_S \| Q_{\mathcal{D}}) &\leq \mathbb{E}_{c\sim Q} \ln \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{c}_{S'} = \hat{c}_S)} \\
&\leq \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{d\sim P} \frac{1}{\Pr_{S'\sim\mathcal{D}^m}(\hat{d}_S = \hat{d}_{S'})} \\
&\leq \mathrm{KL}(Q\|P) + \frac{m+1}{\delta}
\end{aligned}
$$

with probability greater than $1-\delta$.

Finite Classes

- If we take a finite class of functions $h_1, \ldots, h_N$ with prior distribution $p_1, \ldots, p_N$ and assume that the posterior is concentrated on a single function, the generalisation is bounded by

$$\mathrm{KL}(\hat{\mathrm{err}}(h_i)\|\mathrm{err}(h_i)) \leq \frac{-\log(p_i) + \ln((m+1)/\delta)}{m}$$

- This is the standard result for finite classes with the slight refinement that it involves the KL divergence between empirical and true error and the extra $\log(m+1)$ term on the rhs.

Finite Classes

- If we take a finite class of functions $h_1, \ldots, h_N$ with prior distribution $p_1, \ldots, p_N$ and assume that the posterior is concentrated on a single function, the generalisation is bounded by

$$\mathrm{KL}(\hat{\mathrm{err}}(h_i) \| \mathrm{err}(h_i)) \leq \frac{-\log(p_i) + \ln((m+1)/\delta)}{m}$$

- This is the standard result for finite classes with the slight refinement that it involves the KL divergence between empirical and true error and the extra $\log(m+1)$ term on the rhs.

Other extensions/applications

- Matthias Seeger developed the theory for bounding the error of a Gaussian process classifier.
- Olivier Catoni has extended the result to exchangeable distributions enabling him to get a PAC-Bayes version of Vapnik-Chervonenkis bounds.
- Germain et al have extended to more general loss functions than just binary.
- David McAllester has extended the approach to structured output learning.

Other extensions/applications

- Matthias Seeger developed the theory for bounding the error of a Gaussian process classifier.
- Olivier Catoni has extended the result to exchangeable distributions enabling him to get a PAC-Bayes version of Vapnik-Chervonenkis bounds.
- Germain et al have extended to more general loss functions than just binary.
- David McAllester has extended the approach to structured output learning.

Other extensions/applications

- Matthias Seeger developed the theory for bounding the error of a Gaussian process classifier.
- Olivier Catoni has extended the result to exchangeable distributions enabling him to get a PAC-Bayes version of Vapnik-Chervonenkis bounds.
- Germain et al have extended to more general loss functions than just binary.
- David McAllester has extended the approach to structured output learning.

Other extensions/applications

- Matthias Seeger developed the theory for bounding the error of a Gaussian process classifier.
- Olivier Catoni has extended the result to exchangeable distributions enabling him to get a PAC-Bayes version of Vapnik-Chervonenkis bounds.
- Germain et al have extended to more general loss functions than just binary.
- David McAllester has extended the approach to structured output learning.

Linear classifiers and SVMs

- Focus in on linear function application (Langford & ST)
- How the application is made
- Extensions to learning the prior
- Some results on UCI datasets to give an idea of what can be achieved

Linear classifiers and SVMs

- Focus in on linear function application (Langford & ST)
- How the application is made
- Extensions to learning the prior
- Some results on UCI datasets to give an idea of what can be achieved

Linear classifiers and SVMs

- Focus in on linear function application (Langford & ST)
- How the application is made
- Extensions to learning the prior
- Some results on UCI datasets to give an idea of what can be achieved

Linear classifiers and SVMs

- Focus in on linear function application (Langford & ST)
- How the application is made
- Extensions to learning the prior
- Some results on UCI datasets to give an idea of what can be achieved

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.

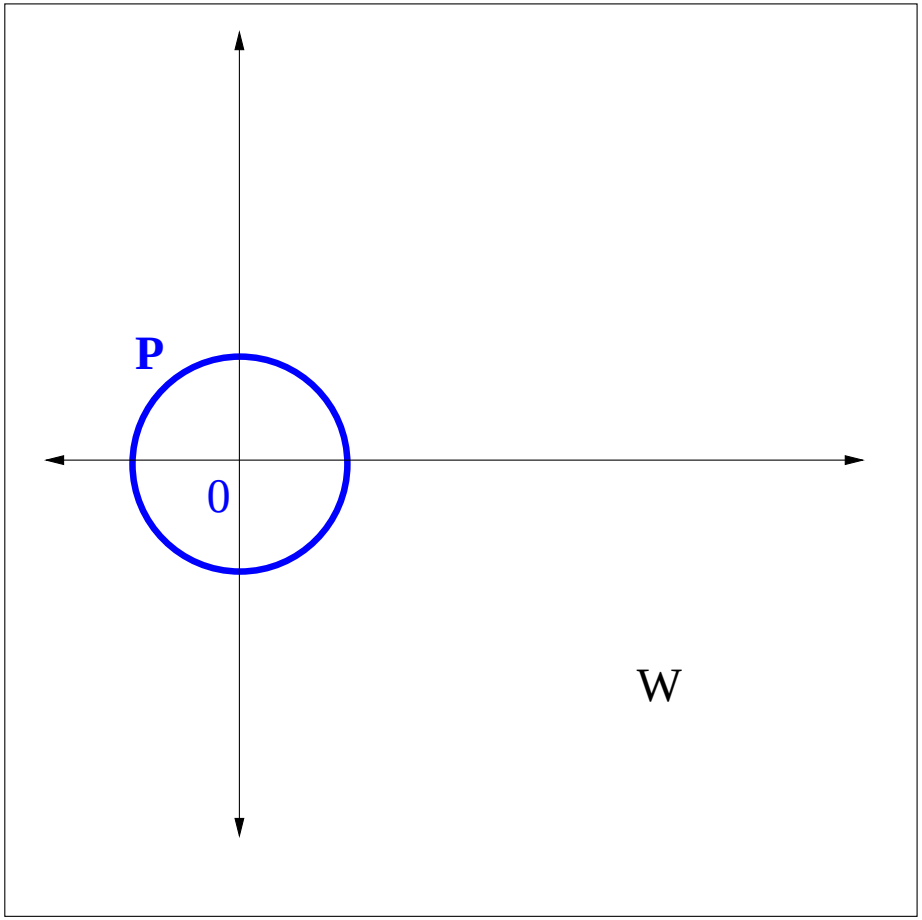
Linear classifiers

- 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.

Linear classifiers

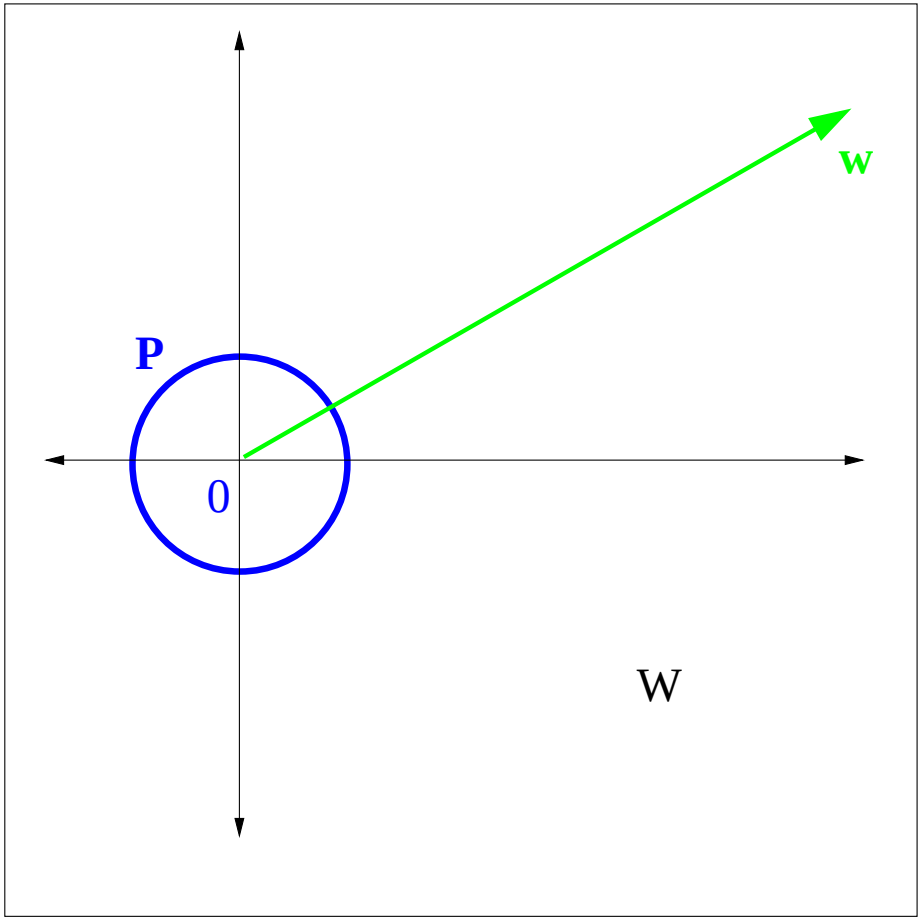
- We will choose the prior and posterior distributions to be Gaussians with unit variance.
- The prior P will be centered at the origin with unit variance
- The specification of the centre for the posterior $Q(\boldsymbol{w}, \mu)$ will be by a unit vector $\mathbf{w}$ and a scale factor μ.

PAC-Bayes Bound for SVM (1/2)



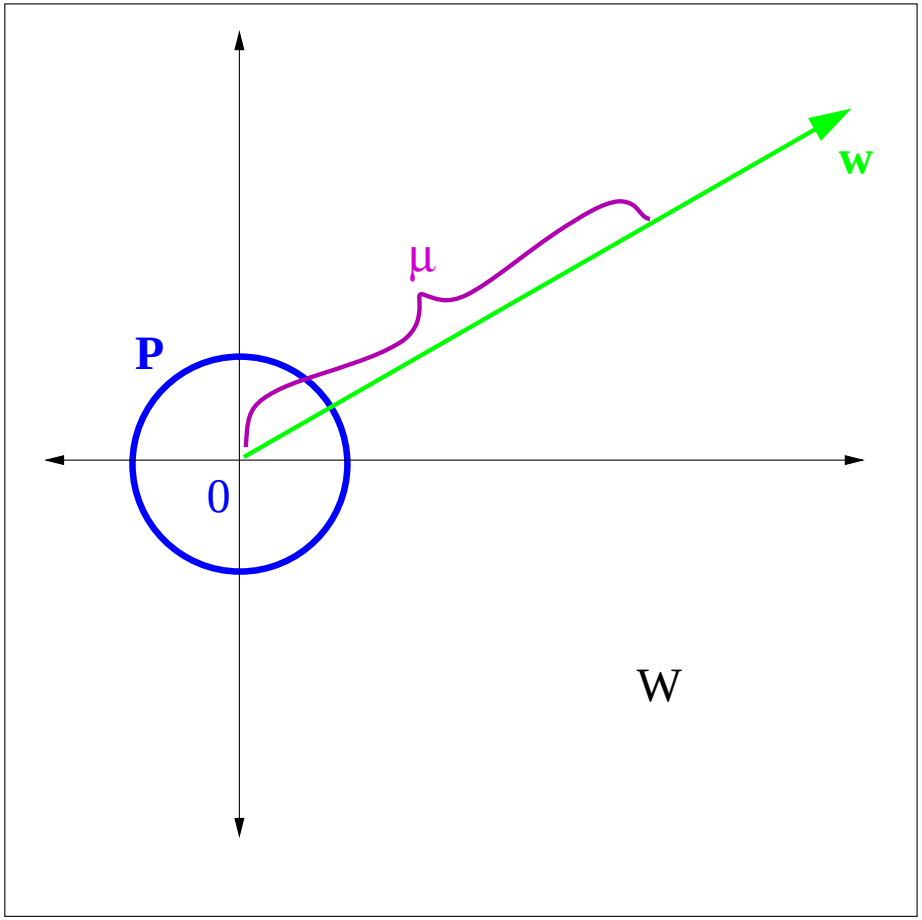
- **Prior** P is Gaussian $\mathcal{N}(0, 1)$
-
-
-

PAC-Bayes Bound for SVM (1/2)



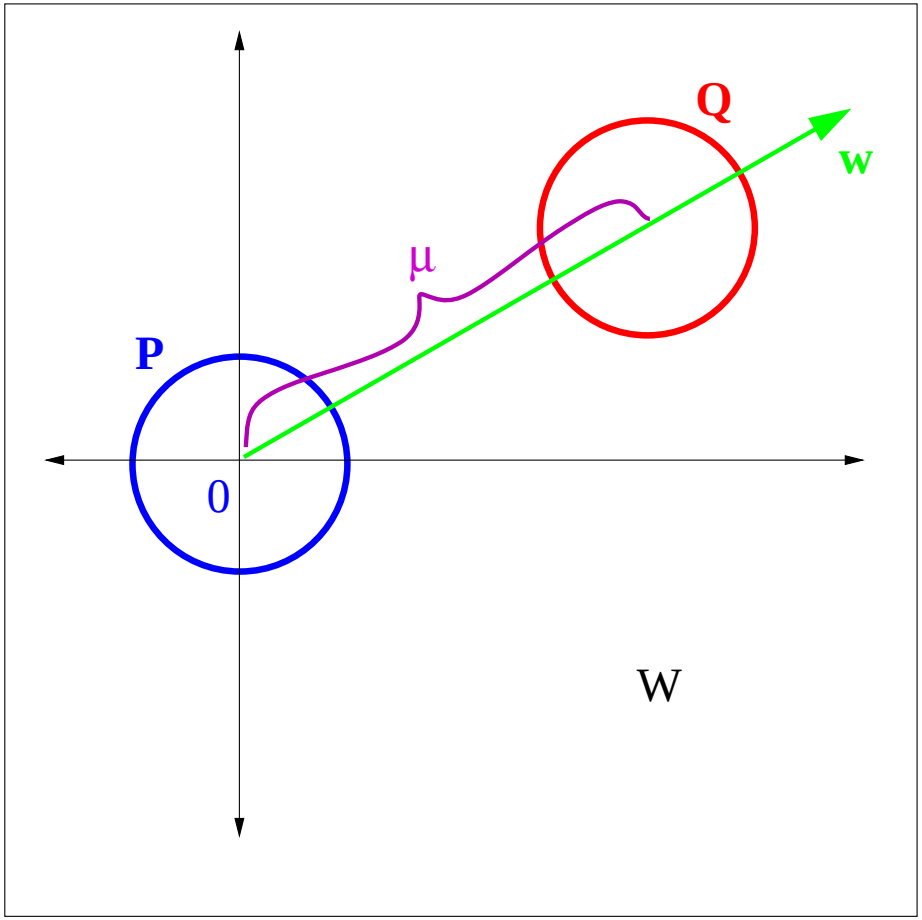
- **Prior** P is Gaussian $\mathcal{N}(0,1)$
- Posterior is in the **direction w**
-
-

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

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|\boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\delta}}{m}$$

- $Q_{\mathcal{D}}(\boldsymbol{w},\mu)$ true performance of the stochastic classifier
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c\sim Q(\mathbf{w},\mu)}[c(x)]\right)$ as centre of the Gaussian gives the same classification as halfspace with more weight.
- Hence its error bounded by $2Q_{\mathcal{D}}(\mathbf{w},\mu)$, since as observed above if x misclassified at least half of $c \sim Q$ err.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|\boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\delta}}{m}$$

- $Q_{\mathcal{D}}(\boldsymbol{w},\mu)$ true performance of the stochastic classifier
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c\sim Q(\mathbf{w},\mu)}[c(x)]\right)$ as centre of the Gaussian gives the same classification as halfspace with more weight.
- Hence its error bounded by $2Q_{\mathcal{D}}(\mathbf{w},\mu)$, since as observed above if x misclassified at least half of $c \sim Q$ err.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|\boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\delta}}{m}$$

- $Q_{\mathcal{D}}(\boldsymbol{w},\mu)$ true performance of the stochastic classifier
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c\sim Q(\mathbf{w},\mu)}[c(x)]\right)$ as centre of the Gaussian gives the same classification as halfspace with more weight.
- Hence its error bounded by $2Q_{\mathcal{D}}(\mathbf{w},\mu)$, since as observed above if x misclassified at least half of $c \sim Q$ err.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|\boxed{Q_{\mathcal{D}}(\boldsymbol{w},\mu)}) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\delta}}{m}$$

- $Q_{\mathcal{D}}(\boldsymbol{w},\mu)$ true performance of the stochastic classifier
- SVM is deterministic classifier that exactly corresponds to $\mathrm{sgn}\left(\mathbb{E}_{c\sim Q(\mathbf{w},\mu)}[c(x)]\right)$ as centre of the Gaussian gives the same classification as halfspace with more weight.
- Hence its error bounded by $2Q_{\mathcal{D}}(\mathbf{w},\mu)$, since as observed above if x misclassified at least half of $c \sim Q$ err.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

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

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

$$\hat{Q}_S(\boldsymbol{w},\mu) = \mathbb{E}_m[\tilde{F}(\mu\gamma(\boldsymbol{x},y))]$$
$$\gamma(\boldsymbol{x},y) = (y\boldsymbol{w}^T\phi(\boldsymbol{x}))/(\|\phi(\boldsymbol{x})\|\|\boldsymbol{w}\|)$$
$$\tilde{F}(t) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\mathrm{d}x$$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

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

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

$$\hat{Q}_S(\boldsymbol{w},\mu) = \mathbb{E}_m[\tilde{F}(\mu\gamma(\boldsymbol{x},y))]$$
$$\gamma(\boldsymbol{x},y) = (y\boldsymbol{w}^T\phi(\boldsymbol{x}))/(\|\phi(\boldsymbol{x})\|\|\boldsymbol{w}\|)$$
$$\tilde{F}(t) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\mathrm{d}x$$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu))} + \ln\frac{m+1}{\delta}}{m}$$

- Prior $P \equiv$ Gaussian centered on the origin
- Posterior $Q \equiv$ Gaussian along $\boldsymbol{w}$ at a distance μ from the origin
- $\mathrm{KL}(P\|Q) = \mu^2/2$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu))} + \ln\frac{m+1}{\delta}}{m}$$

- Prior $P \equiv$ Gaussian centered on the origin
- Posterior $Q \equiv$ Gaussian along $\boldsymbol{w}$ at a distance μ from the origin
- $\mathrm{KL}(P\|Q) = \mu^2/2$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu))} + \ln\frac{m+1}{\delta}}{m}$$

- Prior $P \equiv$ Gaussian centered on the origin
- Posterior $Q \equiv$ Gaussian along $\boldsymbol{w}$ at a distance μ from the origin
- $\mathrm{KL}(P\|Q) = \mu^2/2$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\boxed{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu))} + \ln\frac{m+1}{\delta}}{m}$$

- Prior $P \equiv$ Gaussian centered on the origin
- Posterior $Q \equiv$ Gaussian along $\boldsymbol{w}$ at a distance μ from the origin
- $\mathrm{KL}(P\|Q) = \mu^2/2$

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\boxed{\delta}}}{m}$$

- δ is the confidence
- The bound holds with probability $1-\delta$ over the random i.i.d. selection of the training data.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

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

- δ is the confidence
- The bound holds with probability $1-\delta$ over the random i.i.d. selection of the training data.

PAC-Bayes Bound for SVM (2/2)

Linear classifiers performance may be bounded by

$$\mathrm{KL}(\hat{Q}_S(\boldsymbol{w},\mu)\|Q_{\mathcal{D}}(\boldsymbol{w},\mu)) \leq \frac{\mathrm{KL}(P\|Q(\boldsymbol{w},\mu)) + \ln\frac{m+1}{\boxed{\delta}}}{m}$$

- δ is the confidence
- The bound holds with probability $1-\delta$ over the random i.i.d. selection of the training 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**
- Compute stochastic error with **remaining 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**
- Compute stochastic error with **remaining 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**
- Compute stochastic error with **remaining data**

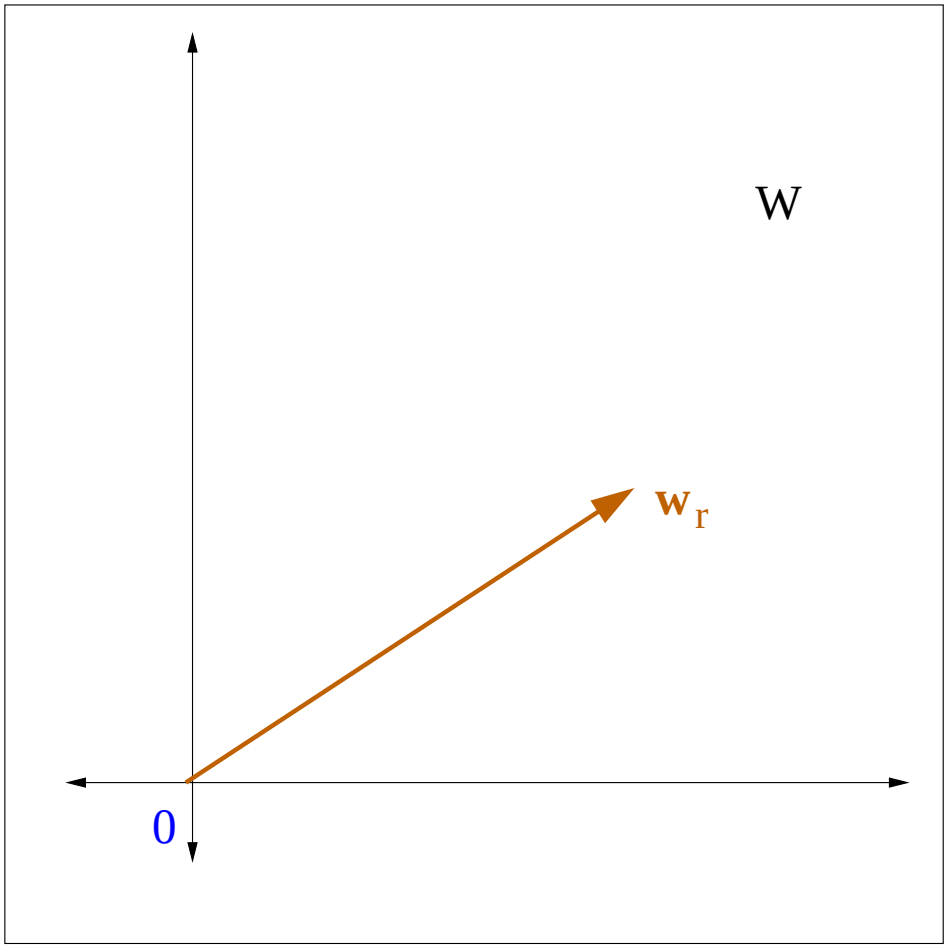
Learning the prior (1/3)

- 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**
- Compute stochastic error with **remaining data**

Learning the prior (1/3)

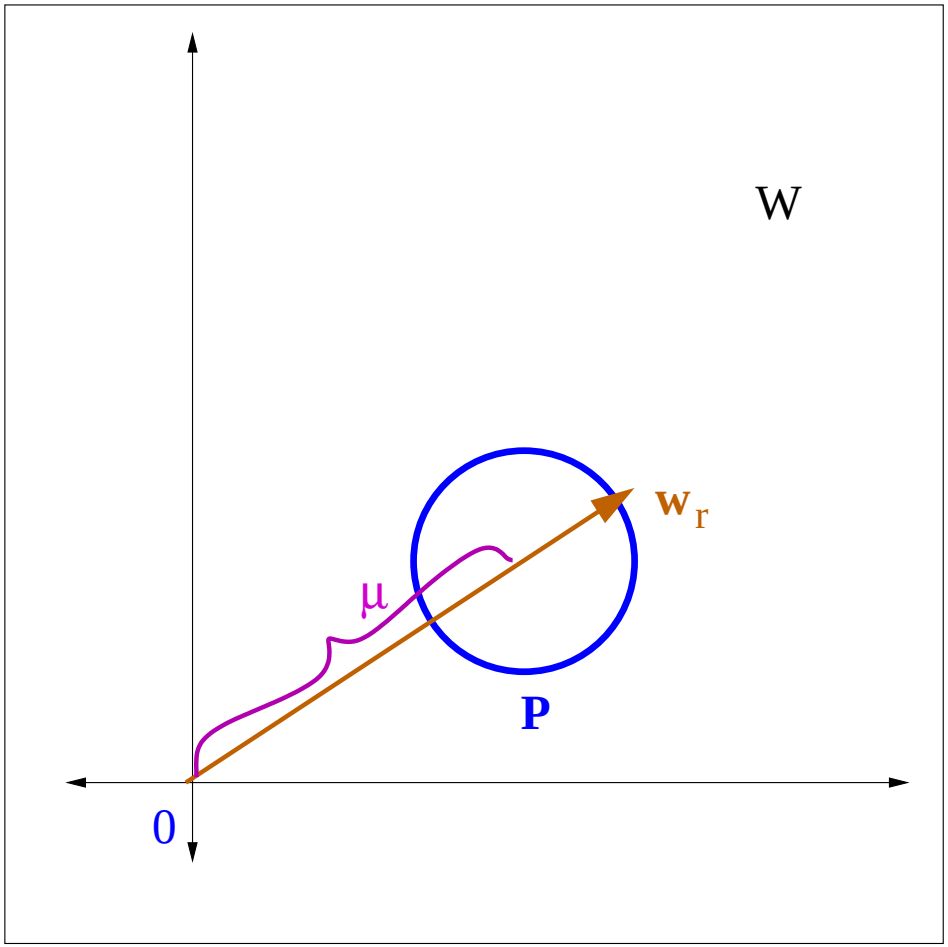
- Bound depends on the **distance between prior and posterior**
- Better prior (closer to posterior) would lead to **tighter bound**
- **Learn** the prior P with part of the data
- Introduce the learnt prior **in the bound**
- Compute stochastic error with **remaining data**

New prior for the SVM (3/3)



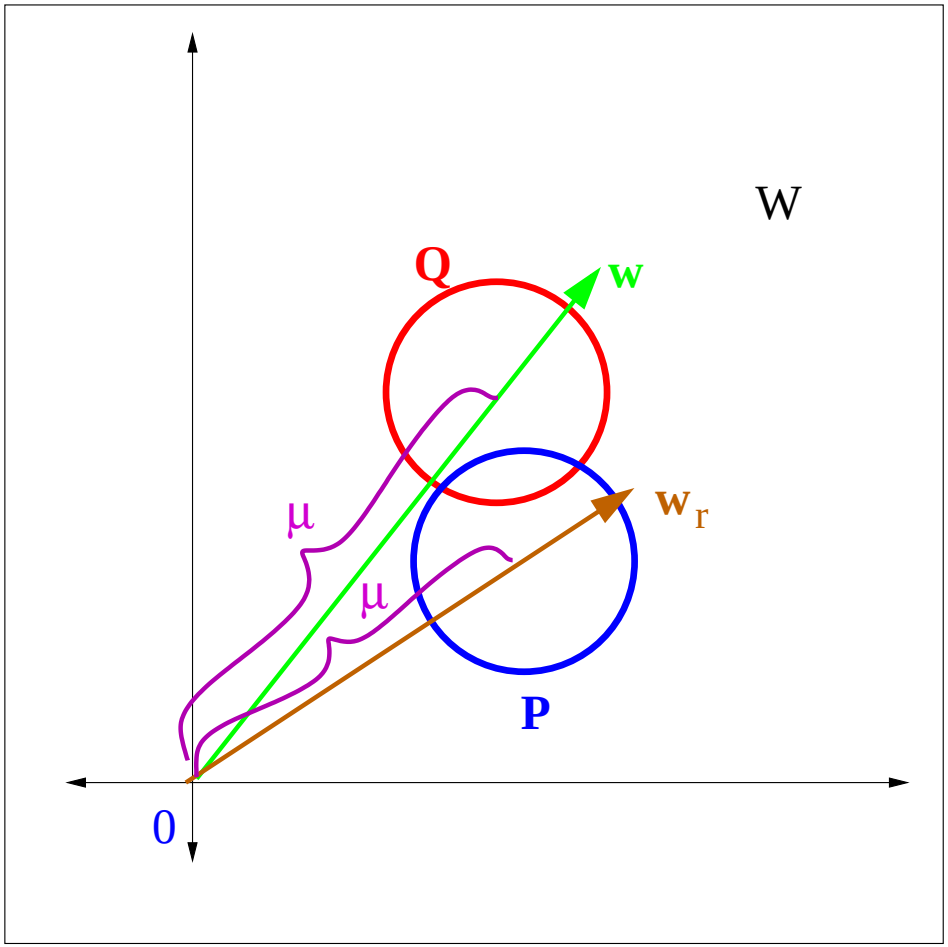
- Solve SVM with **subset of patterns**
-
-
-

New prior for the SVM (3/3)



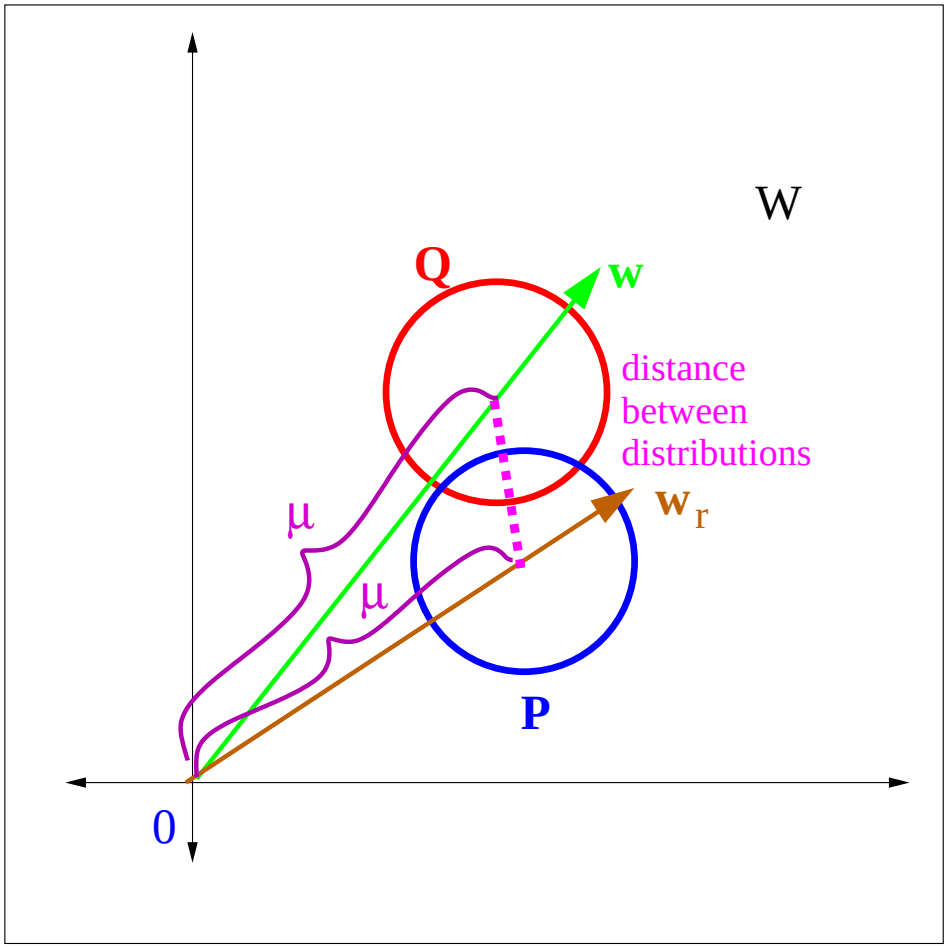
- Solve SVM with **subset of patterns**
- Prior in the **direction w_r**
-
-

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

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

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

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

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

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

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

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

- $\hat{Q}_S(\boldsymbol{w},\mu)$ 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

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

- $\hat{Q}_S(\boldsymbol{w},\mu)$ 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

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

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

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

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

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

New Bound for the SVM (2/3)

SVM performance may be **tightly** bounded by

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

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

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 $m-r$

Prior-SVM

- New bound proportional to $\|\mu \boldsymbol{w} - \eta \boldsymbol{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\boldsymbol{w},\xi_i} \left[\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_r\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

$$\text{s.t. } y_i \boldsymbol{w}^T \phi(\boldsymbol{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

- The p-SVM is only solved with the **remaining points**

Prior-SVM

- New bound proportional to $\|\mu \boldsymbol{w} - \eta \boldsymbol{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\boldsymbol{w},\xi_i} \left[\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_r\|^2 + C \sum_{i=1}^{m-r} \xi_i \right]$$

$$\text{s.t. } y_i \boldsymbol{w}^T \phi(\boldsymbol{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

- The p-SVM is only solved with the **remaining points**

Prior-SVM

- New bound proportional to $\|\mu \boldsymbol{w} - \eta \boldsymbol{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\boldsymbol{w},\xi_i} \left[\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_r\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

$$\text{s.t. } y_i \boldsymbol{w}^T \phi(\boldsymbol{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

- The p-SVM is only solved with the **remaining points**

Prior-SVM

- New bound proportional to $\|\mu \boldsymbol{w} - \eta \boldsymbol{w}_r\|^2$
- Classifier that **optimises the bound**
- Optimisation problem to determine the **p-SVM**

$$\min_{\boldsymbol{w},\xi_i} \left[\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_r\|^2 + C \sum_{i=1}^{m-r} \xi_i \right]$$

$$\text{s.t. } y_i \boldsymbol{w}^T \phi(\boldsymbol{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

- The p-SVM is only solved with the **remaining points**

Bound for p-SVM

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

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

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

Bound for p-SVM

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

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

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

Bound for p-SVM

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

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

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

Bound for p-SVM

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

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

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

η-Prior-SVM

- Consider using a prior distribution P that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

- subject to

$$y_i(\mathbf{v} + \eta\mathbf{w}_r)^T\phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

η-Prior-SVM

- Consider using a prior distribution P that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

- subject to

$$y_i(\mathbf{v} + \eta\mathbf{w}_r)^T \phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

η-Prior-SVM

- Consider using a prior distribution P that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

- subject to

$$y_i(\mathbf{v} + \eta\mathbf{w}_r)^T\phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

η-Prior-SVM

- Consider using a prior distribution P that is elongated in the direction of $\mathbf{w}_r$
- This will mean that there is low penalty for large projections onto this direction
- Translates into an optimisation:

$$\min_{\mathbf{v},\eta,\xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C\sum_{i=1}^{m-r} \xi_i \right]$$

- subject to

$$y_i(\mathbf{v} + \eta\mathbf{w}_r)^T\phi(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r$$
$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r$$

Bound for η-prior-SVM

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

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

Bound for η-prior-SVM

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

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

Bound for η-prior-SVM

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

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

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

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

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

Description of the Datasets

Problem	# samples	input dim.	Pos/Neg
Handwritten-digits	5620	64	2791 / 2829
Waveform	5000	21	1647 / 3353
Pima	768	8	268 / 500
Ringnorm	7400	20	3664 / 3736
Spam	4601	57	1813 / 2788

Table: Description of datasets in terms of number of patterns, number of input variables and number of positive/negative examples.

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

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**

Concluding remarks

- **Frequentist (PAC) and Bayesian** approaches to analysing learning lead to introduction of the PAC-Bayes bound
- Detailed look at the ingredients of the theory
- **Application to bound** the performance of an SVM
- Investigation of **learning of the prior** of the distribution of classifiers
- Experiments show the new bound can be **tighter** ...
- ...And reliable for **low cost model selection**
- p-SVM and η-p-SVM: classifiers that **optimise the new bound**