Large Margin and Kernel Methods

Arindam Banerjee
Consider a 2-class classification task

Training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$

$(x, y)$ sampled i.i.d. from fixed unknown distribution $D$
Learning Machines

Consider a 2-class classification task

- Training set \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \)
- \((x, y)\) sampled i.i.d. from fixed unknown distribution \( D \)

A hypothesis space \( \mathcal{H} \) is a class of learning machines

- Each machine \( f(x, \alpha) : \mathcal{X} \mapsto \{-1, +1\} \)
Learning Machines

Consider a 2-class classification task
- Training set \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \)
- \((x, y)\) sampled i.i.d. from fixed unknown distribution \( D \)

A hypothesis space \( \mathcal{H} \) is a class of learning machines
- Each machine \( f(x, \alpha) : \mathcal{X} \mapsto \{-1, +1\} \)

Given \( S \) (and \( \mathcal{H} \)), the problem is to find a “good” \( \alpha \)
Consider a 2-class classification task

- Training set \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \)
- \((x, y)\) sampled i.i.d. from fixed unknown distribution \( D \)

A hypothesis space \( \mathcal{H} \) is a class of learning machines

- Each machine \( f(x, \alpha) : \mathcal{X} \mapsto \{-1, +1\} \)

Given \( S \) (and \( \mathcal{H} \)), the problem is to find a “good” \( \alpha \)

How to quantify “good”?
Risk

Risk is the expected error of the learning machine $f(x, \alpha)$

$$R(\alpha) = E_D[|y - f(x, \alpha)|/2]$$
Risk

- Risk is the expected error of the learning machine \( f(x, \alpha) \)
  \[
  R(\alpha) = E_D[|y - f(x, \alpha)|/2]
  \]

- Empirical risk is the risk on training set
  \[
  R_{\text{emp}}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} |y_i - f(x_i, \alpha)|/2
  \]
Risk

- Risk is the expected error of the learning machine $f(x, \alpha)$

$$R(\alpha) = E_D[|y - f(x, \alpha)|/2]$$

- Empirical risk is the risk on training set

$$R_{\text{emp}}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} |y_i - f(x_i, \alpha)|/2$$

- If VC dimension of $\mathcal{H}$ is $d_\mathcal{H}$, with high probability

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_\mathcal{H}/m)$$
Risk

- Risk is the expected error of the learning machine $f(x, \alpha)$

$$R(\alpha) = E_D[|y - f(x, \alpha)|/2]$$

- Empirical risk is the risk on training set

$$R_{\text{emp}}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} |y_i - f(x_i, \alpha)|/2$$

- If VC dimension of $\mathcal{H}$ is $d_\mathcal{H}$, with high probability

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_\mathcal{H}/m)$$

- The bound does not depend on $D$
Structural Risk Minimization (SRM)

\[ R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_\mathcal{H}/m) \]

- Complicated hypothesis spaces have high \(d_\mathcal{H}\) but can achieved low empirical risk
Structural Risk Minimization (SRM)

\[ R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_H/m) \]

- Complicated hypothesis spaces have high \( d_H \) but can achieved low empirical risk
- Simple hypothesis spaces have high low \( d_H \) but may have high empirical error risk
Structural Risk Minimization (SRM)

\[ R(\alpha) \leq R_{\text{emp}}(\alpha) + g(\frac{d_{\mathcal{H}}}{m}) \]

- Complicated hypothesis spaces have high \( d_{\mathcal{H}} \) but can achieve low empirical risk

- Simple hypothesis spaces have high low \( d_{\mathcal{H}} \) but may have high empirical error risk

- A fundamental tradeoff for designing learning machines
  - \( d_{\mathcal{H}} \) depends on the entire hypothesis space \( \mathcal{H} \)
  - \( R_{\text{emp}} \) depends on a particular machine/function
Structural Risk Minimization (SRM)

\[ R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_{\mathcal{H}}/m) \]

- Complicated hypothesis spaces have high \( d_{\mathcal{H}} \) but can achieve low empirical risk
- Simple hypothesis spaces have high low \( d_{\mathcal{H}} \) but may have high empirical error risk
- A fundamental tradeoff for designing learning machines
  - \( d_{\mathcal{H}} \) depends on the entire hypothesis space \( \mathcal{H} \)
  - \( R_{\text{emp}} \) depends on a particular machine/function

- Structural Risk Minimization
  - Form a nested set of hypothesis spaces \( \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3 \subseteq \ldots \)
  - Find the function/machine that minimizes the bound
Structural Risk Minimization (SRM)

\[ R(\alpha) \leq R_{\text{emp}}(\alpha) + g(d_{\mathcal{H}}/m) \]

- Complicated hypothesis spaces have high \( d_{\mathcal{H}} \) but can achieve low empirical risk
- Simple hypothesis spaces have high low \( d_{\mathcal{H}} \) but may have high empirical error risk
- A fundamental tradeoff for designing learning machines
  - \( d_{\mathcal{H}} \) depends on the entire hypothesis space \( \mathcal{H} \)
  - \( R_{\text{emp}} \) depends on a particular machine/function
- Structural Risk Minimization
  - Form a nested set of hypothesis spaces \( \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3 \subseteq \ldots \)
  - Find the function/machine that minimizes the bound
- Think of increasingly complicated kernels in the SVM setting
Linear Support Vector Machines

- Linear SVM with separable data
- The prediction: \( f(x) = \mathbf{w}^T x + b \)
- The separability assumption: \( \exists \mathbf{w}, \forall i, y_i f(x_i) \geq 1 \)
- Maximum margin problem can be posed as

\[ \min \frac{1}{2} \| \mathbf{w} \|^2 \quad \text{such that} \quad y_i f(x_i) \geq 1, \forall i \]
Example (To Refresh Memory)

minimize \( w^T w \)

subject to \( A w \leq c \)

- Lagrange dual

\[
L^*(\alpha) = \inf_w (w^T w + \alpha^T (A w - c)) = -\frac{1}{4} \alpha^T A A^T \alpha - c^T \alpha
\]

- Dual problem

maximize \(-\frac{1}{4} \alpha^T A A^T \alpha - c^T \alpha\)

subject to \( \alpha \geq 0 \)

- From Slater’s condition, it is sufficient to solve the dual
Linear SVM: Separable Case

The Lagrangian

\[ L([w \ b], \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i (w^T x_i + b) + \sum_i \alpha_i \]

Setting gradient w.r.t. \([w \ b]\) to 0, we get

\[ w = \sum_i \alpha_i y_i x_i \quad \sum_i \alpha_i y_i = 0 \]

Substituting these back, we get the Lagrange dual (\(\alpha \geq 0\))

\[ L^*(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \]

Recall complementary slackness \(\alpha_i g_i(x) = 0\) for \(g_i(x) \leq 0\)

\[ \alpha_i > 0 \Rightarrow y_i (w^T x_i + b) = 1 \quad x_i \text{ is a support vector} \]

Otherwise \(y_i (w^T x_i + b) > 1 \quad x_i \text{ is not a support vector} \]
Linear SVM: Non-separable Case

We made the separability assumption: \( \exists w, \forall i \ y_i f(x_i) \geq 1 \)
We made the separability assumption: $\exists \mathbf{w}, \forall i \ y_i f(\mathbf{x}_i) \geq 1$

If not true, the problem formulation is infeasible
Linear SVM: Non-separable Case

- We made the separability assumption: \( \exists \mathbf{w}, \forall i \ y_i f(x_i) \geq 1 \)

- If not true, the problem formulation is infeasible

- So we are going to cut some slack

\[
  y_i\left(\mathbf{w}^T \mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i
\]
Linear SVM: Non-separable Case

We made the separability assumption: \( \exists w, \forall i \ y_i f(x_i) \geq 1 \)

If not true, the problem formulation is infeasible

So we are going to cut some slack

\[
y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i
\]

Note that \( \sum_i \xi_i \) is an upper bound on the training error
Linear SVM: Non-separable Case

- We made the separability assumption: $\exists w, \forall i \ y_i f(x_i) \geq 1$

- If not true, the problem formulation is infeasible

- So we are going to cut some slack

$$y_i (w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i$$

- Note that $\sum_i \xi_i$ is an upper bound on the training error

- In general, the problem can be formulated as

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \quad \text{such that} \quad y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0$$
Linear SVM: Non-separable Case

- We made the separability assumption: \( \exists w, \forall i \ y_i f(x_i) \geq 1 \)
- If not true, the problem formulation is infeasible
- So we are going to cut some slack
  \[
  y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i
  \]
- Note that \( \sum_i \xi_i \) is an upper bound on the training error
- In general, the problem can be formulated as
  \[
  \min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \quad \text{such that} \quad y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0
  \]
- Form the Lagrangian, setup the KKT conditions, and solve it
Upper Bounds on Training Error

\[ y_f(x) \]

\[ \text{Loss} \]

- 0-1 loss
- exp loss
- logistic loss
- hinge loss
SVM maximizes minimum margin
SVM is a $L_2$ regularized fit using hinge loss
Logistic and Hindge losses are very similar
Non-linear SVMs

All important equations have $\langle x_i, x_j \rangle$.
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using $\Phi : \mathbb{R}^d \mapsto H$
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using $\Phi : \mathbb{R}^d \mapsto H$
- The derived feature vectors are $\Phi(x_i), \forall i$
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using $\Phi : \mathbb{R}^d \mapsto H$
- The derived feature vectors are $\Phi(x_i), \forall i$
- The dot products are $\langle \Phi(x_i), \Phi(x_j) \rangle$
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using
  $\Phi : \mathbb{R}^d \mapsto H$
- The derived feature vectors are $\Phi(x_i), \forall i$
- The dot products are $\langle \Phi(x_i), \Phi(x_j) \rangle$
- A kernel function allows the dot-product computation without explicitly mapping the points

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using
  $\Phi : \mathbb{R}^d \mapsto H$
- The derived feature vectors are $\Phi(x_i), \forall i$
- The dot products are $\langle \Phi(x_i), \Phi(x_j) \rangle$
- A kernel function allows the dot-product computation without explicitly mapping the points
  
  $$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$

- Learn a linear max margin separator in $H$
Non-linear SVMs

- All important equations have $\langle x_i, x_j \rangle$
- Map $x$ to some other (higher dimensional) space using $\Phi : \mathbb{R}^d \mapsto H$
- The derived feature vectors are $\Phi(x_i), \forall i$
- The dot products are $\langle \Phi(x_i), \Phi(x_j) \rangle$
- A kernel function allows the dot-product computation without explicitly mapping the points

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$

- Learn a linear max margin separator in $H$
- The final prediction function

$$f(x) = \sum_{i=1}^{N_s} \alpha_i y_i \langle \Phi(x_i), \Phi(x) \rangle + b = \sum_{i=1}^{N_s} \alpha_i y_i K(x_i, x) + b$$
The Kernel Trick

- Reduces non-linear SVM learning to linear SVM learning
The Kernel Trick

- Reduces non-linear SVM learning to linear SVM learning
- What functions $K(x_i, x_j)$ are valid kernels? Mercer kernels, RKHS and all that
The Kernel Trick

- Reduces non-linear SVM learning to linear SVM learning

- What functions $K(x_i, x_j)$ are valid kernels? Mercer kernels, RKHS and all that

- Examples:
  - Polynomial Kernel: $K(x_i, x_j) = (x_i^T x_j + 1)^p$
  - RBF Kernel: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2)$
The Kernel Trick

- Reduces non-linear SVM learning to linear SVM learning

- What functions $K(x_i, x_j)$ are valid kernels? Mercer kernels, RKHS and all that

- Examples:
  - Polynomial Kernel: $K(x_i, x_j) = (x_i^T x_j + 1)^p$
  - RBF Kernel: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2)$

- How to choose a kernel for a given application?
The Kernel Trick

- Reduces non-linear SVM learning to linear SVM learning

- What functions $K(x_i, x_j)$ are valid kernels? Mercer kernels, RKHS and all that

- Examples:
  - Polynomial Kernel: $K(x_i, x_j) = (x_i^T x_j + 1)^p$
  - RBF Kernel: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2)$

- How to choose a kernel for a given application?

- Isn’t the kernel trick increasing the VC dimension?
The prediction $f(x) = w^T x + b$
The prediction \( f(x) = w^T x + b \)

The (primal) non-separable case

\[
\min_w \left\{ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \right\} \quad \text{such that} \quad y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0
\]
SVM loss, Revisited

- The prediction $f(x) = w^T x + b$

- The (primal) non-separable case

$$\min_{w} \left\{ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \right\} \text{ such that } y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0$$

- Alternative viewpoint as a regularized hinge loss

$$\min_{w} \left\{ \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i f(x_i)\} + \lambda \|w\|^2 \right\}$$
SVM loss, Revisited

- The prediction \( f(x) = w^T x + b \)

- The (primal) non-separable case

\[
\min_w \left\{ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \right\} \quad \text{such that} \quad y_i f(x_i) \geq 1 - \xi_i, \xi_i \geq 0
\]

- Alternative viewpoint as a regularized hinge loss

\[
\min_w \left\{ \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i f(x_i)\} + \lambda \|w\|^2 \right\}
\]

- Regularized loss minimization with two terms
  - First term: Margin loss on the training set
  - Second term: Regularization
Regularization

A large class of linear models minimize

$$\sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|_p$$
Regularization

A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|_p \]

- \( L \) is a loss function
Regularization

- A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \|w\|_p \]

- \( L \) is a loss function
  - Regression: Squared loss, Bregman divergences (GLMs)
Regularization

A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|^p \]

\( L \) is a loss function
- Regression: Squared loss, Bregman divergences (GLMs)
- Classification: 0-1 loss, Exp loss, Log loss, Hindge loss
Regularization

A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|_p \]

\( L \) is a loss function
- Regression: Squared loss, Bregman divergences (GLMs)
- Classification: 0-1 loss, Exp loss, Log loss, Hindge loss
- Encompasses both training and margin losses
A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|_p^p \]

- \( L \) is a loss function
  - Regression: Squared loss, Bregman divergences (GLMs)
  - Classification: 0-1 loss, Exp loss, Log loss, Hinge loss
  - Encompasses both training and margin losses

- \( \| w \|_p^p \) is the regularization term
  - Regression: \( p = 1 \) for Lasso, \( p = 2 \) for Ridge regression
  - Classification: \( p = 1 \) for Boosting, \( p = 2 \) for SVMs
  - The weight vector is constrained in the \( L_p \) ball
A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, w^T x_i) + \lambda \| w \|_p \]

- **L** is a loss function
  - Regression: Squared loss, Bregman divergences (GLMs)
  - Classification: 0-1 loss, Exp loss, Log loss, Hinge loss
  - Encompasses both training and margin losses

- $\| w \|_p$ is the regularization term
  - Regression: $p = 1$ for Lasso, $p = 2$ for Ridge regression
  - Classification: $p = 1$ for Boosting, $p = 2$ for SVMs
  - The weight vector is constrained in the $L_p$ ball

- **Bayesian interpretation:** loss comes from conditional, regularization comes from prior
Regularization

A large class of linear models minimize

\[ \sum_{i=1}^{n} L(y_i, \mathbf{w}^T \mathbf{x}_i) + \lambda \| \mathbf{w} \|_p \]

\( L \) is a loss function
- Regression: Squared loss, Bregman divergences (GLMs)
- Classification: 0-1 loss, Exp loss, Log loss, Hindge loss
- Encompasses both training and margin losses

\( \| \mathbf{w} \|_p \) is the regularization term
- Regression: \( p = 1 \) for Lasso, \( p = 2 \) for Ridge regression
- Classification: \( p = 1 \) for Boosting, \( p = 2 \) for SVMs
- The weight vector is constrained in the \( L_p \) ball

Bayesian interpretation: loss comes from conditional, regularization comes from prior

Special case of more general theory of regularizers
For a distribution $W$ on $(x, y)$, $\|x\| = 1$, $y \in \{-1, +1\}$, the error rate of $w$ at margin $\gamma$

$$\ell_\gamma(w, W) = \Pr_{(x, y) \sim W}[y \langle w, x \rangle \leq \gamma]$$
Theory: Basic Definitions

For a distribution $W$ on $(x, y)$, $\|x\| = 1$, $y \in \{-1, +1\}$, the error rate of $w$ at margin $\gamma$

$$
\ell_{\gamma}(w, W) = \Pr_{(x, y) \sim W}[y\langle w, x \rangle \leq \gamma]
$$

For a stochastic classifier that maintains $Q$ over all possible $w$, the error rate at margin $\gamma$

$$
\ell_{\gamma}(Q, W) = E_{w \sim Q, (x, y) \sim W}[y\langle w, x \rangle \leq \gamma] = E_{w \sim Q}[\ell_{\gamma}(w, W)]
$$
Theory: Basic Definitions

For a distribution $W$ on $(x, y)$, $\|x\| = 1$, $y \in \{-1, +1\}$, the error rate of $w$ at margin $\gamma$

$$\ell_\gamma(w, W) = \Pr_{(x, y) \sim W}[y\langle w, x \rangle \leq \gamma]$$

For a stochastic classifier that maintains $Q$ over all possible $w$, the error rate at margin $\gamma$

$$\ell_\gamma(Q, W) = E_{w \sim Q, (x, y) \sim W}[y\langle w, x \rangle \leq \gamma] = E_{w \sim Q}[\ell_\gamma(w, W)]$$

We are interested in $W = S$, the sample distribution, and $W = D$, the true distribution
Theory: Basic Definitions

- For a distribution $W$ on $(x, y)$, $\|x\| = 1$, $y \in \{-1, +1\}$, the error rate of $w$ at margin $\gamma$

$$\ell_\gamma(w, W) = \Pr_{(x,y) \sim W}[y\langle w, x \rangle \leq \gamma]$$

- For a stochastic classifier that maintains $Q$ over all possible $w$, the error rate at margin $\gamma$

$$\ell_\gamma(Q, W) = E_{w \sim Q, (x,y) \sim W}[y\langle w, x \rangle \leq \gamma] = E_{w \sim Q} [\ell_\gamma(w, W)]$$

- We are interested in $W = S$, the sample distribution, and $W = D$, the true distribution

- We want to get a margin bound of the form

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + h(m, \gamma)$$
Theory: Basic Definitions

For a distribution $W$ on $(x, y)$, $\|x\| = 1$, $y \in \{-1, +1\}$, the error rate of $w$ at margin $\gamma$

$$\ell_\gamma(w, W) = \Pr_{(x, y) \sim W}[y<w, x> \leq \gamma]$$

For a stochastic classifier that maintains $Q$ over all possible $w$, the error rate at margin $\gamma$

$$\ell_\gamma(Q, W) = E_{w \sim Q, (x, y) \sim W}[y<w, x> \leq \gamma] = E_{w \sim Q}[\ell_\gamma(w, W)]$$

We are interested in $W = S$, the sample distribution, and $W = D$, the true distribution

We want to get a margin bound of the form

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + h(m, \gamma)$$

Recall (and differentiate with) the VC dimension based bound

$$\ell_0(w, D) \leq \ell_0(w, S) + g(d/m)$$
Consider the stochastic classification setting
Step 0: The PAC-Bayes Theorem

- Consider the stochastic classification setting
- Let $P$ be a prior distribution over $w$
Consider the stochastic classification setting

Let $P$ be a prior distribution over $w$

Let $Q$ be a posterior distribution over $w$ after observing the training set $S$, with $|S| = m$
Step 0: The PAC-Bayes Theorem

- Consider the stochastic classification setting

- Let $P$ be a prior distribution over $w$

- Let $Q$ be a posterior distribution over $w$ after observing the training set $S$, with $|S| = m$

- With probability at least $(1 - \delta)$ over the choice of $S \sim D^m$,

$$KL(\ell_\gamma(Q, S) \| \ell_\gamma(Q, D)) \leq \frac{KL(Q \| P) + \ln \frac{m+1}{\delta}}{m}$$
Step 0: The PAC-Bayes Theorem

- Consider the stochastic classification setting

- Let $P$ be a prior distribution over $w$

- Let $Q$ be a posterior distribution over $w$ after observing the training set $S$, with $|S| = m$

- With probability at least $(1 - \delta)$ over the choice of $S \sim D^m$,

$$KL(\ell_\gamma(Q, S) \| \ell_\gamma(Q, D)) \leq \frac{KL(Q \| P) + \ln \frac{m+1}{\delta}}{m}$$

- Gives an upper bound on the “difference” between training set error rate and true error rate
Step 0: The PAC-Bayes Theorem

Consider the stochastic classification setting

Let $P$ be a prior distribution over $w$

Let $Q$ be a posterior distribution over $w$ after observing the training set $S$, with $|S| = m$

With probability at least $(1 - \delta)$ over the choice of $S \sim D^m$,

$$KL(\ell_\gamma(Q, S) \| \ell_\gamma(Q, D)) \leq \frac{KL(Q \| P) + \ln \frac{m+1}{\delta}}{m}$$

Gives an upper bound on the “difference” between training set error rate and true error rate

How do we

- Introduce margins into the bound
- Get bounds for deterministic linear classifiers
Step 1: Gaussian Distribution

Consider the normal distribution $N(w; \mu, 1)$
Step 1: Gaussian Distribution

- Consider the normal distribution \( N(w; \mu, 1) \)

- For \( w \sim N(w; \mu, 1) \), what is \( \Pr(w \mu \leq 0) \)?
Consider the normal distribution $N(w; \mu, 1)$

For $w \sim N(w; \mu, 1)$, what is $Pr(w \mu \leq 0)$?

More generally, consider $Q$ to be the multivariate $N(x; \mu, \mathbb{1})$.
Step 1: Gaussian Distribution

- Consider the normal distribution $N(w; \mu, 1)$
- For $w \sim N(w; \mu, 1)$, what is $\Pr(w \mu \leq 0)$?
- More generally, consider $Q$ to be the multivariate $N(x; \mu, I)$
- A direct calculation shows that

$$\Pr_{x \sim Q}[x^T \mu \leq 0] \leq \frac{1}{2\|\mu\|} \exp(-\|\mu\|^2/2)$$
Step 1: Gaussian Distribution

- Consider the normal distribution $N(w; \mu, 1)$

- For $w \sim N(w; \mu, 1)$, what is $\Pr(w \mu \leq 0)$?

- More generally, consider $Q$ to be the multivariate $N(x; \mu, I)$

- A direct calculation shows that

  $$\Pr_{x \sim Q}[x^T \mu \leq 0] \leq \frac{1}{2\|\mu\|} \exp(-\|\mu\|^2/2)$$

- The prior $P$ is chosen to be $N(x; 0, I)$
Step 2: Margins

Let $W$ be any distribution over $(x, y)$
Step 2: Margins

- Let $W$ be any distribution over $(x, y)$
- Let $Q$ be $N(w; w^\dagger, I)$ over $w$
Step 2: Margins

- Let \( W \) be any distribution over \((x, y)\)
- Let \( Q \) be \( N(w; w^\dagger, I) \) over \( w \)
- Then, \( \forall \gamma > 0, \beta \in \mathbb{R} \)

\[
\ell_\beta(Q, W) \leq \ell_{\beta+\gamma}(w^\dagger, W) + \frac{1}{2\gamma} \exp(-\gamma^2/2)
\]
Step 2: Margins

Let \( W \) be any distribution over \((x, y)\)

Let \( Q \) be \( N(w; w^\dagger, I) \) over \( w \)

Then, \( \forall \gamma > 0, \beta \in \mathbb{R} \)

\[
\ell_\beta(Q, W) \leq \ell_{\beta + \gamma}(w^\dagger, W) + \frac{1}{2\gamma} \exp(-\gamma^2/2)
\]

Further,

\[
\ell_\beta(w^\dagger, W) \leq \ell_{\beta + \gamma}(Q, W) + \frac{1}{2\gamma} \exp(-\gamma^2/2)
\]
Step 2: Margins

- Let $W$ be any distribution over $(x, y)$
- Let $Q$ be $N(w; w^\dagger, \I)$ over $w$
- Then, $\forall \gamma > 0, \beta \in \mathbb{R}$
  
  $$\ell_\beta(Q, W) \leq \ell_{\beta+\gamma}(w^\dagger, W) + \frac{1}{2\gamma} \exp(-\gamma^2/2)$$

- Further,
  
  $$\ell_\beta(w^\dagger, W) \leq \ell_{\beta+\gamma}(Q, W) + \frac{1}{2\gamma} \exp(-\gamma^2/2)$$

- The Gaussian tail (Step 1) is used to derive these
Step 3: Putting It Together

- The true error rate of the classifier $w^\dagger$

$$\ell_0(w^\dagger, D) \leq \ell_{\gamma/2}(Q, D) + \frac{1}{\gamma} \exp(-\gamma^2/8)$$
Step 3: Putting It Together

- The true error rate of the classifier $\hat{w}$

$$\ell_0(\hat{w}, D) \leq \ell_{\gamma/2}(Q, D) + \frac{1}{\gamma} \exp(-\gamma^2/8)$$

- The train set error rate of the stochastic classifier $Q$

$$\ell_{\gamma/2}(Q, S) \leq \ell_{\gamma}(\hat{w}, S) + \frac{1}{\gamma} \exp(-\gamma^2/8)$$
Step 3: Putting It Together

- The true error rate of the classifier $\mathbf{w}^\dagger$

\[ \ell_0(\mathbf{w}^\dagger, D) \leq \ell_{\gamma/2}(Q, D) + \frac{1}{\gamma} \exp(-\gamma^2/8) \]

- The train set error rate of the stochastic classifier $Q$

\[ \ell_{\gamma/2}(Q, S) \leq \ell_\gamma(\mathbf{w}^\dagger, S) + \frac{1}{\gamma} \exp(-\gamma^2/8) \]

- From the PAC Bayes theorem

\[ KL(\ell_\gamma(\mathbf{w}^\dagger, S) \| \ell_{\gamma/2}(Q, D)) \leq \frac{KL(Q \| P) + \ln \frac{m+1}{\delta}}{m} \]
Step 3: Putting It Together

The true error rate of the classifier $w^\dagger$

$$\ell_0(w^\dagger, D) \leq \ell_{\gamma/2}(Q, D) + \frac{1}{\gamma} \exp(-\gamma^2/8)$$

The train set error rate of the stochastic classifier $Q$

$$\ell_{\gamma/2}(Q, S) \leq \ell_{\gamma}(w^\dagger, S) + \frac{1}{\gamma} \exp(-\gamma^2/8)$$

From the PAC Bayes theorem

$$KL(\ell_{\gamma}(w^\dagger, S)\|\ell_{\gamma/2}(Q, D)) \leq \frac{KL(Q\|P) + \ln \frac{m+1}{\delta}}{m}$$

Since $P = N(w; 0, I)$ and $Q = N(w; w^\dagger, I)$, $KL(Q\|P) = \frac{||w^\dagger||^2}{2}$
The normalized margin $\tilde{\gamma} = \frac{\langle w^\dagger, x \rangle}{\|w^\dagger\| \|x\|}$
The normalized margin $\tilde{\gamma} = \frac{\langle w^\dagger, x \rangle}{\|w^\dagger\| \|x\|}$

Note that $\tilde{\gamma} \in [0, 1]$ and $k = m\tilde{\gamma}^2 \in [0, m]$
The normalized margin $\tilde{\gamma} = \frac{\langle w^+, x \rangle}{\|w^+\| \|x\|}$

Note that $\tilde{\gamma} \in [0, 1]$ and $k = m\tilde{\gamma}^2 \in [0, m]$

Consider (polynomial) large number of specific values of $\tilde{\gamma} \in [0, 1]$
The normalized margin $\tilde{\gamma} = \frac{\langle w^\dagger, x \rangle}{\|w^\dagger\| \|x\|}$

Note that $\tilde{\gamma} \in [0, 1]$ and $k = m\tilde{\gamma}^2 \in [0, m]$

Consider (polynomial) large number of specific values of $\tilde{\gamma} \in [0, 1]$

For any $\tilde{\gamma}$ is this large set

$$\ell_0(w^\dagger, D) \leq \ell_{\gamma}(w^\dagger, S) + \frac{1}{k\sqrt{2\ln k}} + \frac{4\ln k}{k} + O\left(\sqrt{\frac{\ln \frac{m}{\delta}}{m}}\right)$$

$$+ \sqrt{\left(\ell_{\gamma}(w^\dagger, S) + \frac{1}{k\sqrt{2\ln(k)}}\right) \frac{4\ln(k)}{k}}$$
Margin Bound (After “Simplification”)

- The normalized margin \( \tilde{\gamma} = \frac{\langle w^\dagger, x \rangle}{\|w^\dagger\| \|x\|} \)

- Note that \( \tilde{\gamma} \in [0, 1] \) and \( k = m\tilde{\gamma}^2 \in [0, m] \)

- Consider (polynomial) large number of specific values of \( \tilde{\gamma} \in [0, 1] \)

- For any \( \tilde{\gamma} \) this large set

\[
\ell_0(w^\dagger, D) \leq \ell_\gamma(w^\dagger, S) + \frac{1}{k\sqrt{2\ln k}} + \frac{4\ln k}{k} + O\left(\sqrt{\frac{\ln m}{m}}\right)
\]

\[
+ \sqrt{\left(\ell_\gamma(w^\dagger, S) + \frac{1}{k\sqrt{2\ln(k)}}\right) \frac{4\ln(k)}{k}}
\]

- Similar results can be derived for all \( \tilde{\gamma} \in [0, 1] \)
For the chosen $\gamma$, $\ell_\gamma(w^\dagger, S') = 0$
The Realizable Case

- For the chosen $\gamma$, $\ell_{\gamma}(w^\dagger, S') = 0$

- Consider the case $m \to \infty$ holding $k = m\tilde{\gamma}^2$ constant
The Realizable Case

- For the chosen $\gamma$, $\ell_{\gamma}(w^\dagger, S) = 0$

- Consider the case $m \to \infty$ holding $k = m\bar{\gamma}^2$ constant

- The true error rate

$$\ell_0(w^\dagger, D) \leq \frac{1}{k\sqrt{2\ln k}} + \frac{4\ln k}{k} + \frac{\sqrt{8\ln k}}{k} = O\left(\frac{\ln k}{k}\right)$$
The Realizable Case

- For the chosen \( \gamma \), \( \ell_\gamma(w^\dagger, S) = 0 \)

- Consider the case \( m \to \infty \) holding \( k = m\bar{\gamma}^2 \) constant

- The true error rate

\[
\ell_0(w^\dagger, D) \leq \frac{1}{k\sqrt{2\ln k}} + \frac{4\ln k}{k} + \frac{\sqrt{8\ln k}}{k} = O\left(\frac{\ln k}{k}\right)
\]

- Recall the (realizable) convex classifier bound

\[
\ell_0(w^\dagger, D) \leq O\left(\sqrt{\frac{d_H}{k}}\right)
\]

- p.20
For $k = m\gamma^2$, we obtained bounds

$$
\ell_0(w, D) \leq \ell_\gamma(w, S) + h(k)
$$
For \( k = m\gamma^2 \), we obtained bounds

\[ \ell_0(w, D) \leq \ell_\gamma(w, S) + h(k) \]

The function \( h(k) \) measures complexity of hypothesis class.
For $k = m\bar{\gamma}^2$, we obtained bounds

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + h(k)$$

- The function $h(k)$ measures complexity of hypothesis class
- Note that $h(k)$ is a decreasing function of $k$ (and $\gamma$)
For $k = m \gamma^2$, we obtained bounds

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + h(k)$$

The function $h(k)$ measures complexity of hypothesis class.

Note that $h(k)$ is a decreasing function of $k$ (and $\gamma$).

The basic trade-off (for a fixed $m$)
- Large $\gamma \Rightarrow$ low complexity classifier, more margin error
- Small $\gamma \Rightarrow$ high complexity classifier, low margin error
For $k = m\bar{\gamma}^2$, we obtained bounds

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + h(k)$$

- The function $h(k)$ measures complexity of hypothesis class
- Note that $h(k)$ is a decreasing function of $k$ (and $\gamma$)
- The basic trade-off (for a fixed $m$)
  - Large $\gamma \Rightarrow$ low complexity classifier, more margin error
  - Small $\gamma \Rightarrow$ high complexity classifier, low margin error
- In practice, trade-off is between the two terms in the objective
The Big Picture

- For $k = m\gamma^2$, we obtained bounds
  \[ \ell_0(w, D) \leq \ell_\gamma(w, S) + h(k) \]

- The function $h(k)$ measures complexity of hypothesis class

- Note that $h(k)$ is a decreasing function of $k$ (and $\gamma$)

- The basic trade-off (for a fixed $m$)
  - Large $\gamma \Rightarrow$ low complexity classifier, more margin error
  - Small $\gamma \Rightarrow$ high complexity classifier, low margin error

- In practice, trade-off is between the two terms in the objective

- So that $\lambda$ (or $C'$) is important