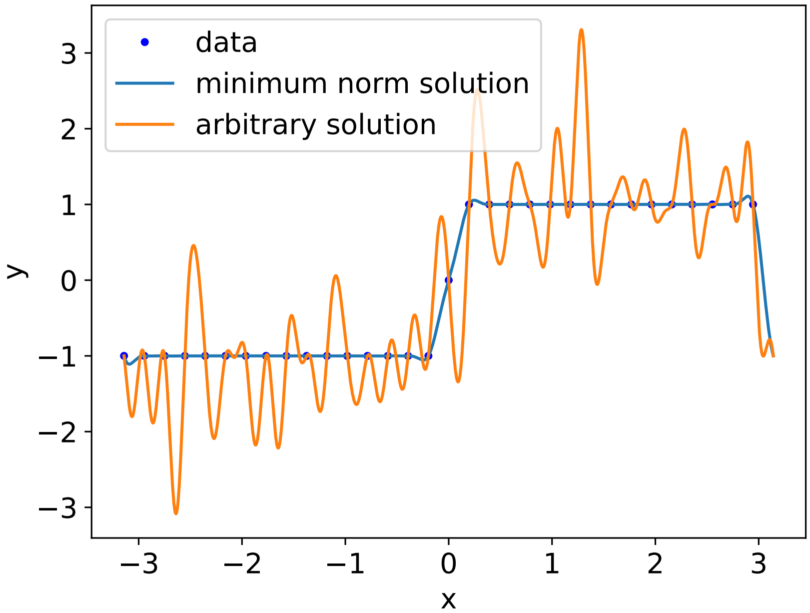
Inductive bias and regularization

COMS 4771 Fall 2023

Minimum norm solutions

Normal equations $(A^\top A)w = A^\top b$ can have infinitely-many solutions

$$\varphi(x) = \left(1, \cos(x), \sin(x), \frac{\cos(2x)}{2}, \frac{\sin(2x)}{2}, \ldots, \frac{\cos(32x)}{32}, \frac{\sin(32x)}{32}\right)$$



Norm of w is a measure of "steepness"

$$\underbrace{|w^\top \varphi(x) - w^\top \varphi(x')|}_{\text{change in output}} \leq \|w\| \times \underbrace{\|\varphi(x) - \varphi(x')\|}_{\text{change in input}}$$

(Cauchy-Schwarz inequality)

- Note: Data does not provide a reason to prefer short w over long w
- Preference for short w is example of inductive bias (tie-breaking rule)

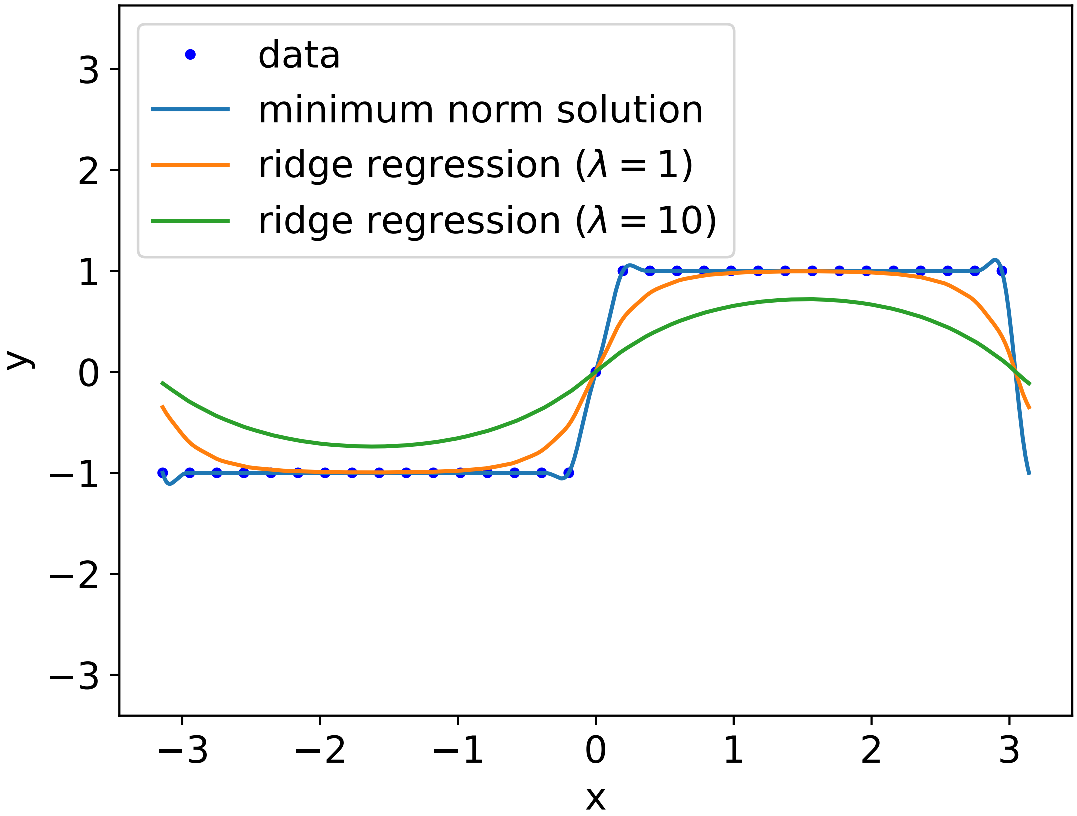
Ridge regression

Ridge regression: "balance" two concerns by minimizing

$$\|Aw - b\|^2 + \lambda \|w\|^2$$

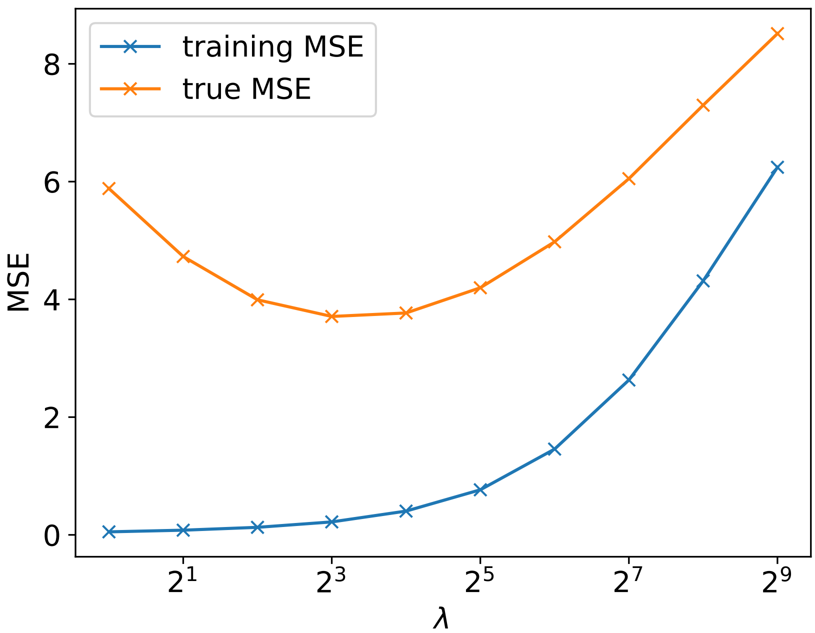
where $\lambda \geq 0$ is hyperparameter

- Concern 1: "data fitting term" $\|Aw - b\|^2$ (involves training data)
- Concern 2: regularizer $\lambda \|w\|^2$ (doesn't involve training data)
- $\lambda = 0$ corresponds to objective in OLS
- $\lambda \to 0^+$ gives minimum norm solution



Example: $n = d = 100$, $((X^{(i)}, Y^{(i)}))_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} (X, Y)$, where $X \sim \mathrm{N}(0, I)$, and conditional distribution of Y given $X = x$ is $\mathrm{N}(\sum_{j=1}^{10} x_j, 1)$

- Normal equations have unique solution, but OLS performs poorly



Different interpretation of ridge regression objective

$$
\begin{aligned}
&\|Aw - b\|^2 + \lambda\|w\|^2 \\
&= \|Aw - b\|^2 + \|(\sqrt{\lambda} I)w - 0\|^2
\end{aligned}
$$

- Second term is MSE on d additional "fake examples"

$$
\begin{aligned}
(x^{(n+1)}, y^{(n+1)}) &= \underline{\hspace{4cm}} \\
(x^{(n+2)}, y^{(n+2)}) &= \underline{\hspace{4cm}} \\
&\vdots \\
(x^{(n+d)}, y^{(n+d)}) &= \underline{\hspace{4cm}}
\end{aligned}
$$

"Augmented" dataset in matrix notation:

$$
\widetilde{A} = \begin{bmatrix} \longleftarrow & (x^{(1)})^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n)})^\mathsf{T} & \longrightarrow \\ \longleftarrow & (x^{(n+1)})^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n+d)})^\mathsf{T} & \longrightarrow \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}
$$

so

$$
\|Aw - b\|^2 + \lambda\|w\|^2 = \|\widetilde{A}w - \tilde{b}\|^2
$$

What are "normal equations" for ridge regression objective (in terms of $\widetilde{A}$, $\tilde{b}$)?

Other forms of regularization

Regularization using **domain-specific data augmentation**

Create "fake examples" from existing data by applying transformations that do not change appropriateness of corresponding label, e.g.,

- Image data: rotations, rescaling
- Audio data: change playback rate
- Text data: replace words with synonyms

Functional penalties (e.g., norm on w)

- Ridge: (squared) ℓ^2 norm

$$\|w\|^2$$

- Lasso: ℓ^1 norm

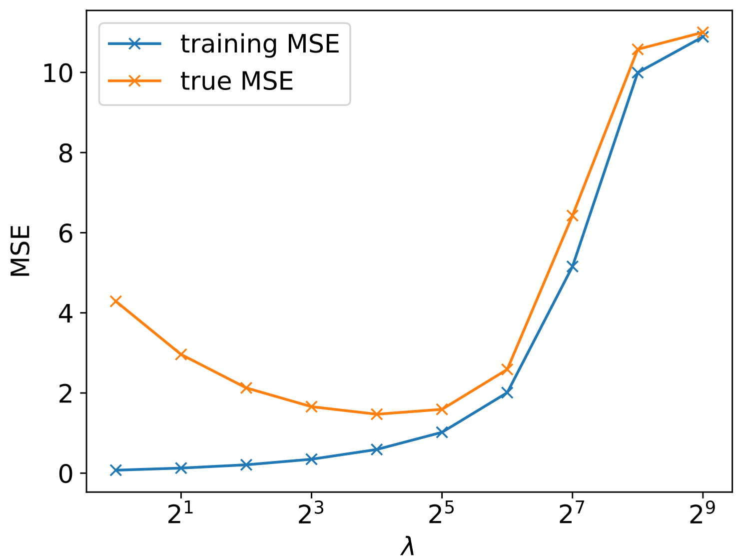
$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

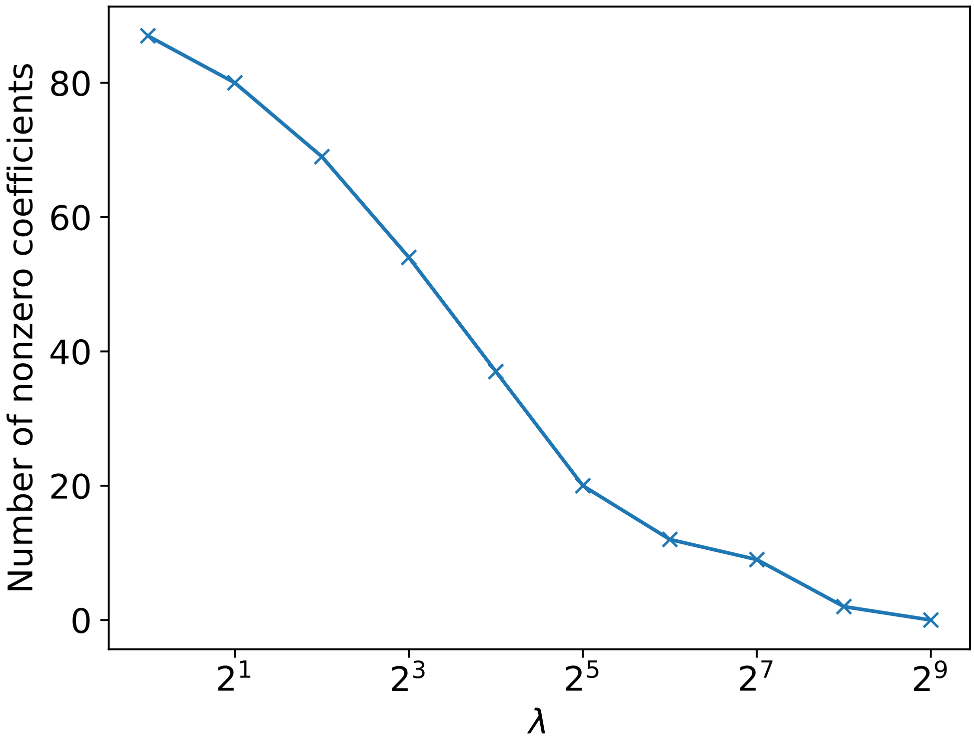
- Sparse regularization: ℓ^0 "norm" (not really a norm)

$$\|w\|_0 = \# \text{ coefficients in } w \text{ that are non-zero}$$

Example: $n = d = 100$, $((X^{(i)}, Y^{(i)}))_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X, Y)$, where $X \sim \mathrm{N}(0, I)$, and conditional distribution of Y given $X = x$ is $\mathrm{N}(\sum_{j=1}^{10} x_j, 1)$

- Minimize $\|Aw - b\|^2 + \lambda \|w\|_1$ (Lasso)





Weighted (squared) ℓ^2 norm:

$$\sum_{i=1}^{d} c_i \, w_i^2$$

for some "costs" $c_1, \ldots, c_d \geq 0$

- Motivation: make it more "costly" (in regularizer) to use certain features
- Where do costs come from?

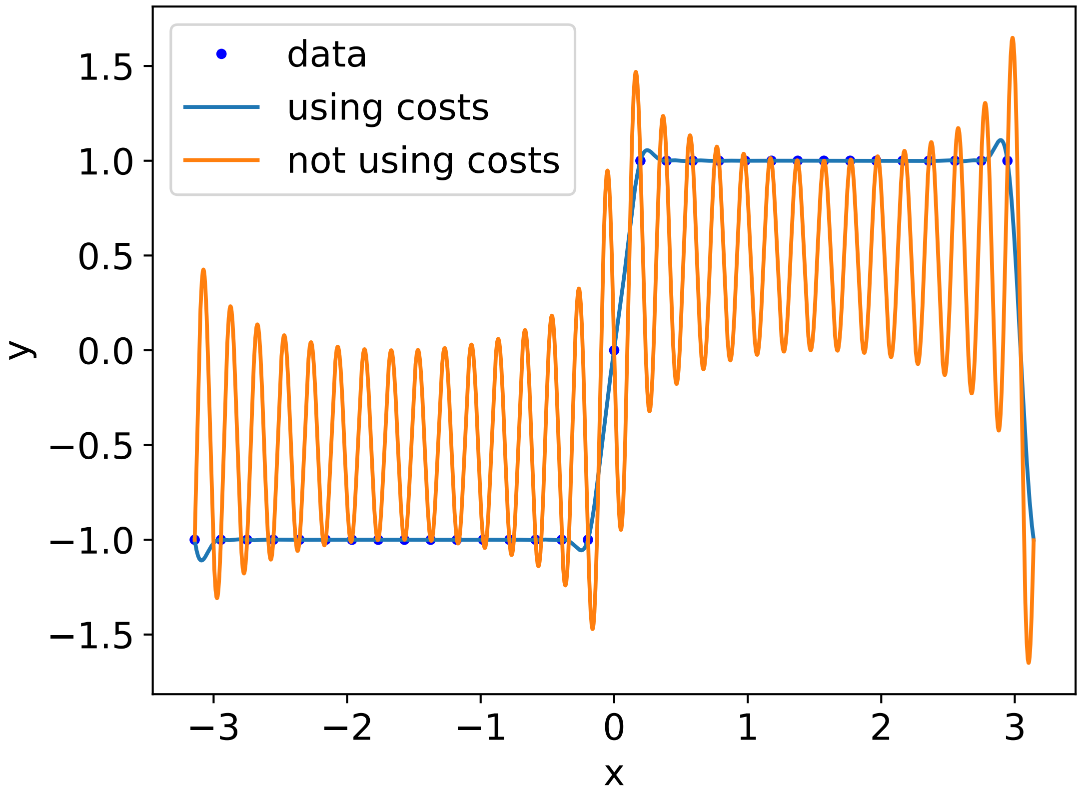
Example:

$$\varphi(x) = (1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots, \cos(32x), \sin(32x))$$

with regularizer on $w = (w_0, w_{\cos,1}, w_{\sin,1}, \ldots, w_{\cos,32}, w_{\sin,32})$

$$w_0^2 + \sum_{j=1}^{d} j^2 \times \left(w_{\cos,j}^2 + w_{\sin,j}^2\right)$$

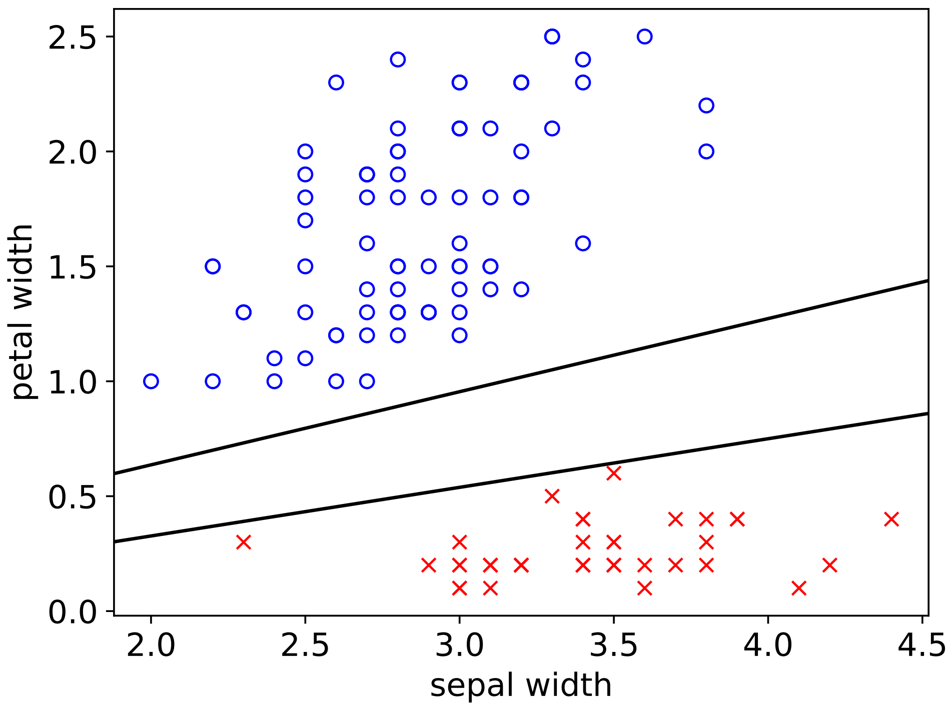
(More expensive to use "high frequency" features)



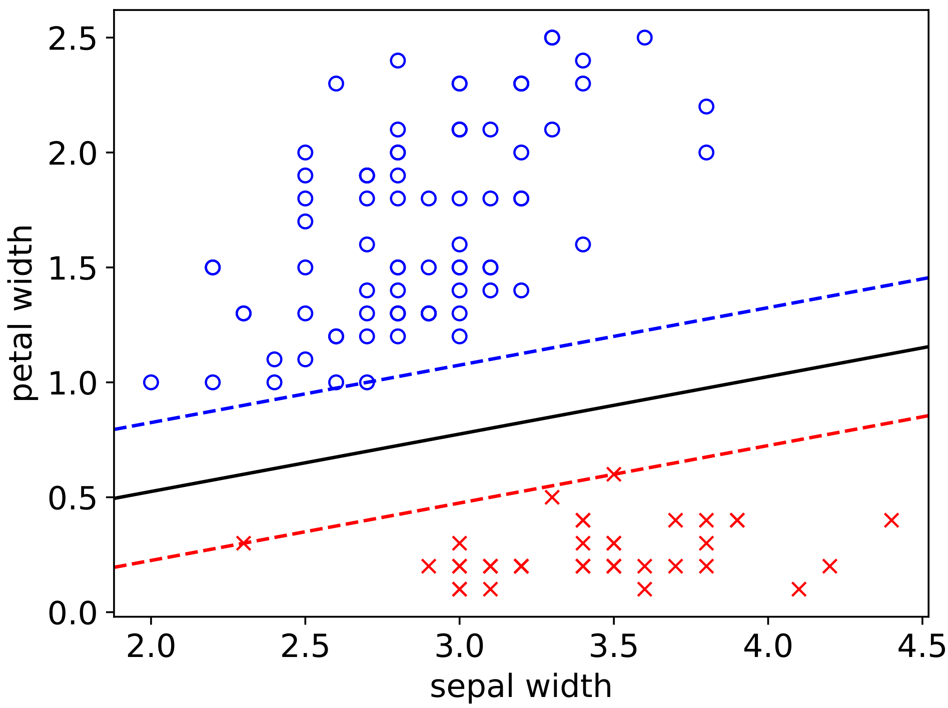
Question: Can effect of costs be achieved using (original) ridge regularization by changing φ?

Margins and support vector machines

Many linear classifiers with same training error rate



Possible inductive bias: largest "margin", i.e., most "wiggle room"



For notational convenience, use $\mathcal{Y} = \{-1, 1\}$ instead of $\mathcal{Y} = \{0, 1\}$

- $f_{w,b}(x) = \text{sign}(w^{\mathsf{T}} x + b)$
- $f_{w,b}(x) = y$ can be written as

$$y(w^{\mathsf{T}} x + b) > 0$$

- If it is possible to satisfy

$$y(w^{\mathsf{T}} x + b) > 0 \quad \text{for all } (x, y) \in \mathcal{S},$$

  then can rescale w and b so that

$$\min_{(x,y)\in\mathcal{S}} y(w^{\mathsf{T}} x + b) = 1$$

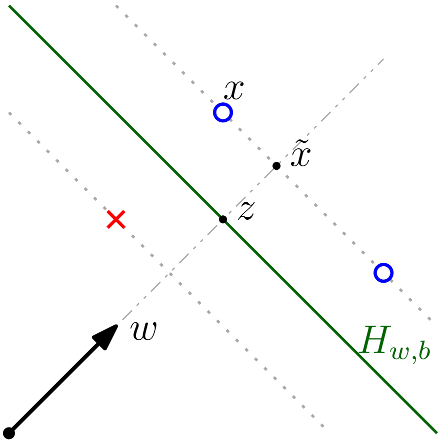
Say linear classifier $f_{w,b}$ achieves margin γ on example (x, y) if:

- $f_{w,b}(x) = y$
- Distance from x to decision boundary of $f_{w,b}$ is γ

Say $f_{w,b}$ achieves margin γ on dataset $\mathcal{S}$ if it achieves margin at least γ on every example $(x, y) \in \mathcal{S}$

- I.e., γ is "worst" margin achieved on a training example

How to find linear classifier $f_{w,b}$ with largest margin on dataset $\mathcal{S}$?



Let $z \in \operatorname{span}\{w\} \cap H_{w,b}$

For $(x, y) \in \mathcal{S}$ satisfying $y(w^\top x + b) = 1$, let $\tilde{x}$ be orthoprojection of x to $\operatorname{span}\{w\}$, so

$$w^\top x + b = w^\top \tilde{x} + b = y$$

Therefore

$$|w^\top(\tilde{x} - z)| = \underline{\qquad}$$

So distance from x to $H_{w,b}$ is $\underline{\qquad\qquad}$

How to find linear classifier $f_{w,b}$ with largest margin on dataset $\mathcal{S}$?

Solution: find $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ that satisfy

$$\min_{(x,y)\in\mathcal{S}} y(w^\top x + b) = 1$$

and that maximizes $\dfrac{1}{\|w\|}$

Support Vector Machine (SVM) optimization problem

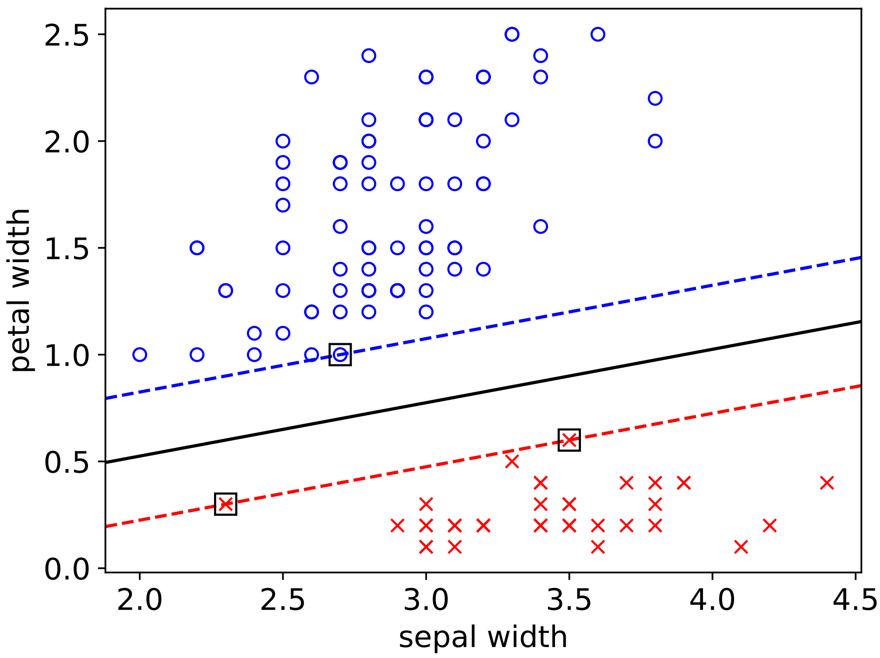
$$\min_{(w,b)\in\mathbb{R}^d\times\mathbb{R}} \quad \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y(w^\top x + b) \geq 1 \quad \text{for all } (x,y) \in \mathcal{S}$$

(Recall, labels are from $\{-1, 1\}$ instead of $\{0, 1\}$ here)

Examples $(x,y) \in \mathcal{S}$ for which $y(w^\top x + b) = 1$ are called support vectors

Iris dataset, treating versicolor and virginica as a single class, using features

$x_1 =$ sepal width, $x_2 =$ petal width



Soft-margin SVM: for datasets that are not linearly separable

$$\min_{(w,b)\in\mathbb{R}^d\times\mathbb{R}} \quad \frac{1}{2}\|w\|^2 + C\sum_{(x,y)\in\mathcal{S}}[1-y(w^\intercal x+b)]_+$$

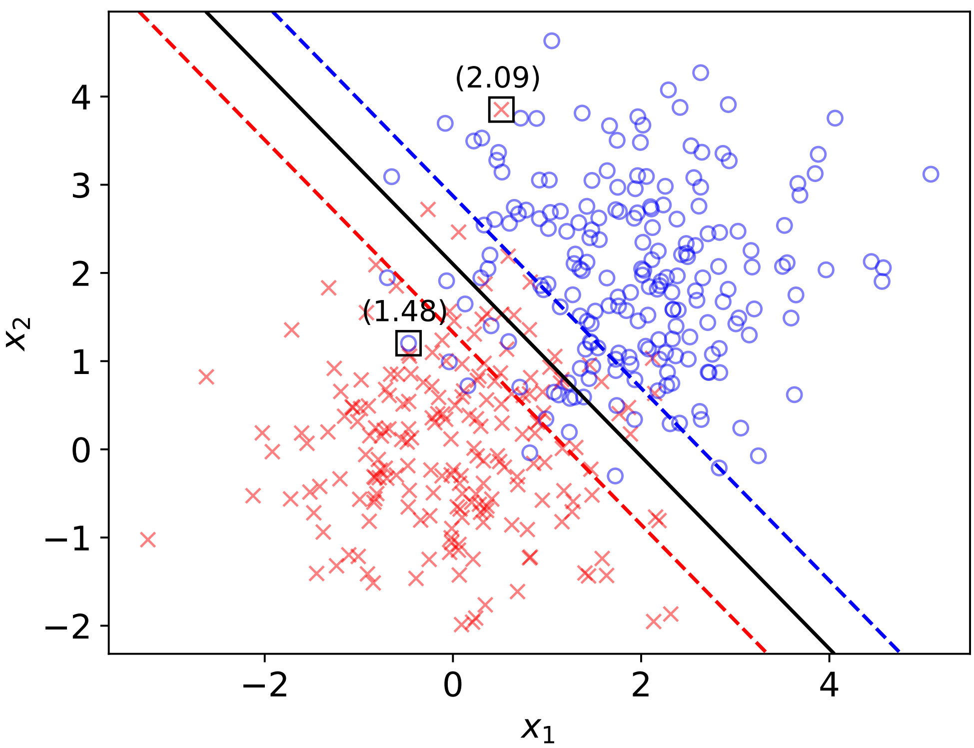
where $[z]_+ = \max\{0, z\}$ (and $C > 0$ is hyperparameter)

Term in summation corresponding to $(x, y) \in \mathcal{S}$:

- Zero if $y(w^\intercal x + b) \geq 1$
- Otherwise, proportional to distance that x must be moved in order to satisfy $y(w^\intercal x + b) = 1$

Synthetic example with normal feature vectors

- Two classes; class 0: $\mathrm{N}((0,0), I)$, class 1: $\mathrm{N}((2,2), I)$
- 200 training data from each class
- Solved soft-margin SVM problem with $C = 10$ to obtain (w, b)


(2.09)
(1.48)
4
3
2
1
0
−1
−2
−2
0
2
4
x_1
x_2