CS 480/680
Introduction to Machine Learning

Lecture 7
Support Vector Machines Part II
Soft Margin Classifier

Kathryn Simone

1 October 2024

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Interlude: Revising optimization with Lagrangian multipliers

Consider the constrained objective

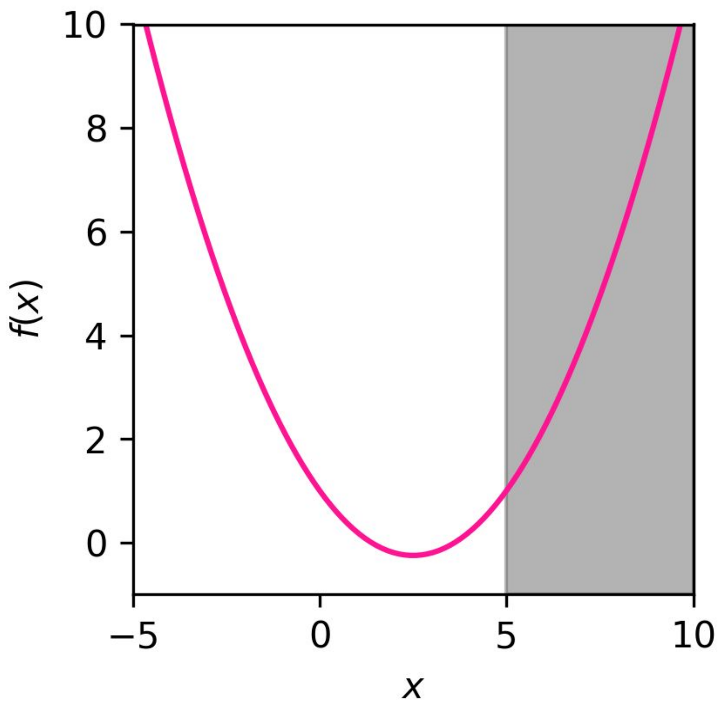
$$\min\, y(x) = 0.2x^2 - x + 1$$
$$\text{subject to: } x \geq 5$$

This can be transformed into the *dual* problem

$$\max\, \mathcal{L}(x, \lambda) = 0.2x^2 - x + 1 - \lambda(x - 5)$$
$$\text{subject to: } \lambda \geq 0$$

This can be transformed into the *dual* problem

$$\min_{x} \max_{\lambda \geq 0} \mathcal{L}(x, \lambda) = 0.2x^2 - x + 1 - \lambda(x - 5)$$



Interlude: Revising optimization with Lagrangian multipliers

Consider the constrained objective

$$\min\ y(x) = 0.2x^2 - x + 1$$
$$\text{subject to: } x \geq 5$$

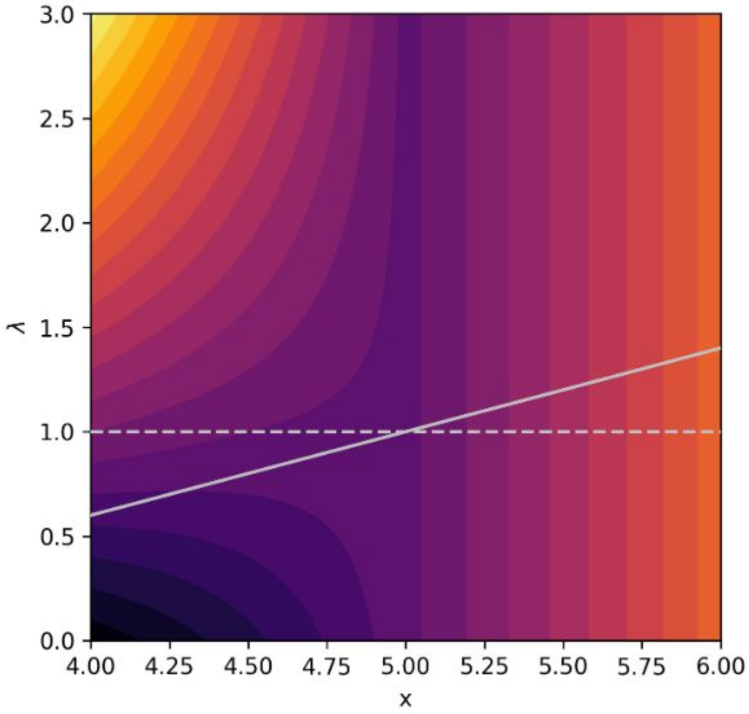
This can be transformed into the *dual* problem

$$\min_x \max_{\lambda \geq 0} \mathcal{L}(x, \lambda) = 0.2x^2 - x + 1 - \lambda(x - 5)$$

Where λ is a Lagrange multiplier. The dual problem is constructed by first minimizing the Lagrangian with respect to x, and then maximizing the result with respect to $\lambda \geq 0$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x} &= 2(0.2)x - 1 - \lambda \\
&= 0.4x - 1 - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= -(x - 5) = 0 \\
&\implies x = 5 \\
&\implies \lambda = 0.4(5) - 1 = 2 - 1 = 1
\end{aligned}$$

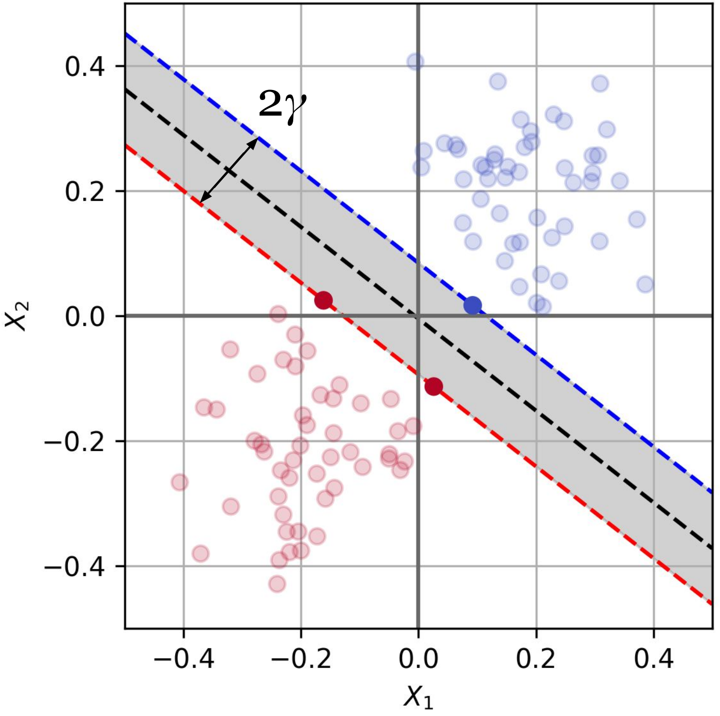
As $\lambda \geq 0$, $x = 5$ is a feasible solution to the dual problem.



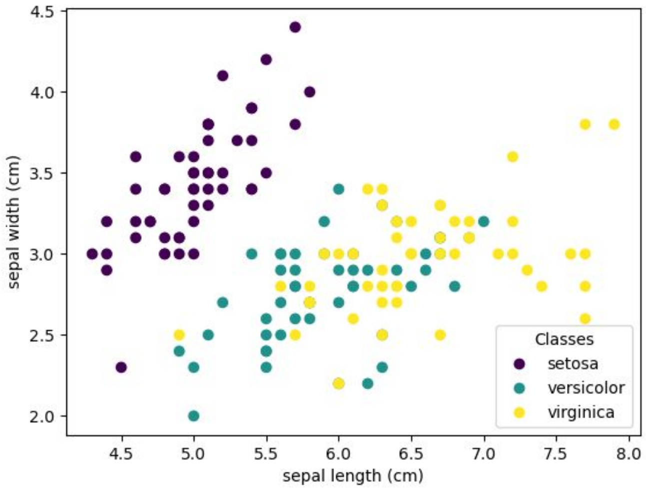
Optimal separating hyperplane with support vectors for linearly separable data

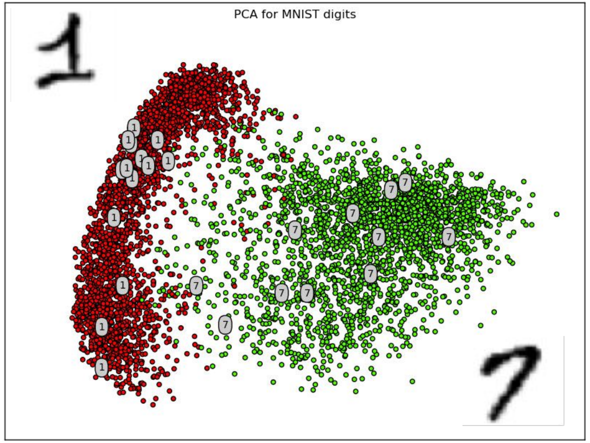
$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2$$

$$\text{subject to: } y_i(w^T x_i + b) \geq 1 \;\; \forall i$$



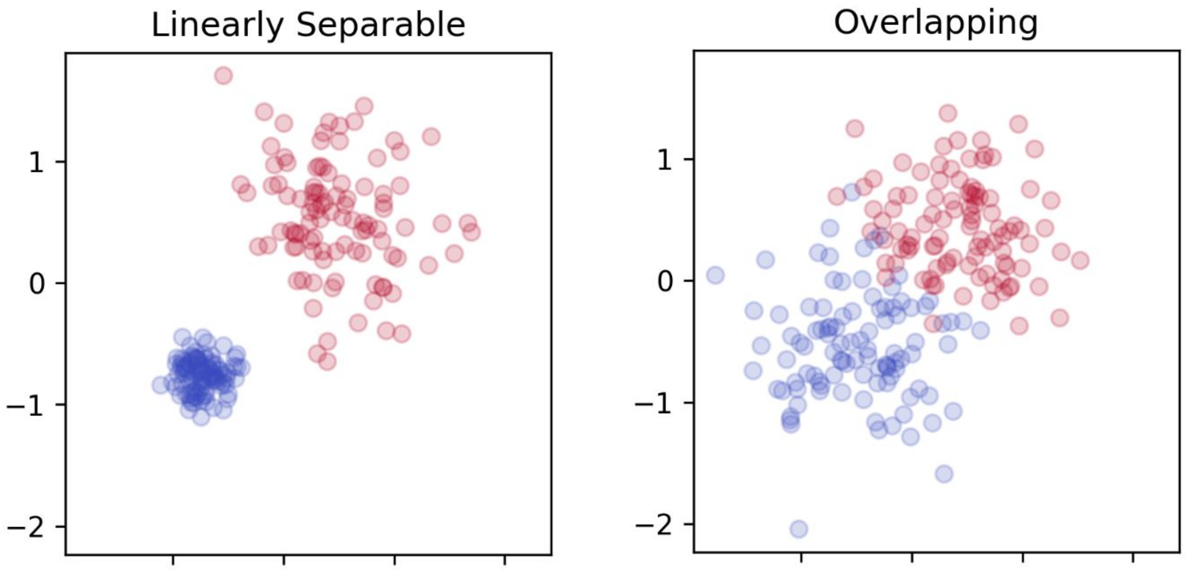
Real-world datasets are not usually linearly separable





Figures (clockwise from left): Sci-Kit Learn, Medium, ADBench.github.io)

How can the SVM be generalized to handle harder problems?



Key Questions

I. How can we relax the hard-margin constraints?

II. Can we gain any insights deriving the dual?

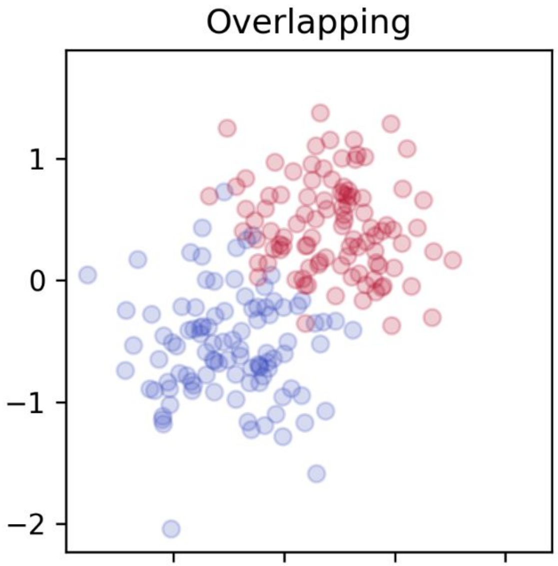
III. How do we optimize?

Key Questions

I. **How can we relax the hard-margin constraints?**

II. **Can we gain any insights deriving the dual?**

III. **How do we optimize?**


Overlapping
1
0
−1
−2

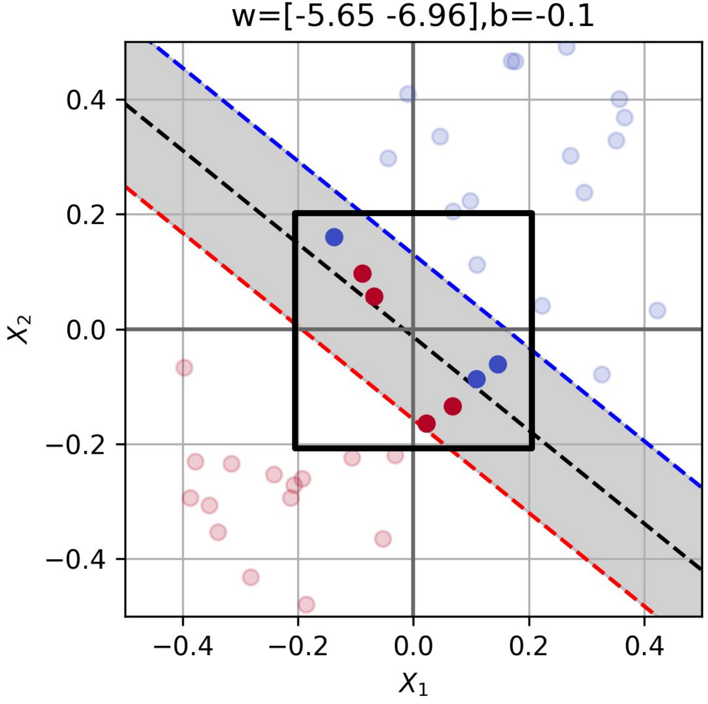
Soft-Margin SVM turns constraint into a cost

Hard Margin:

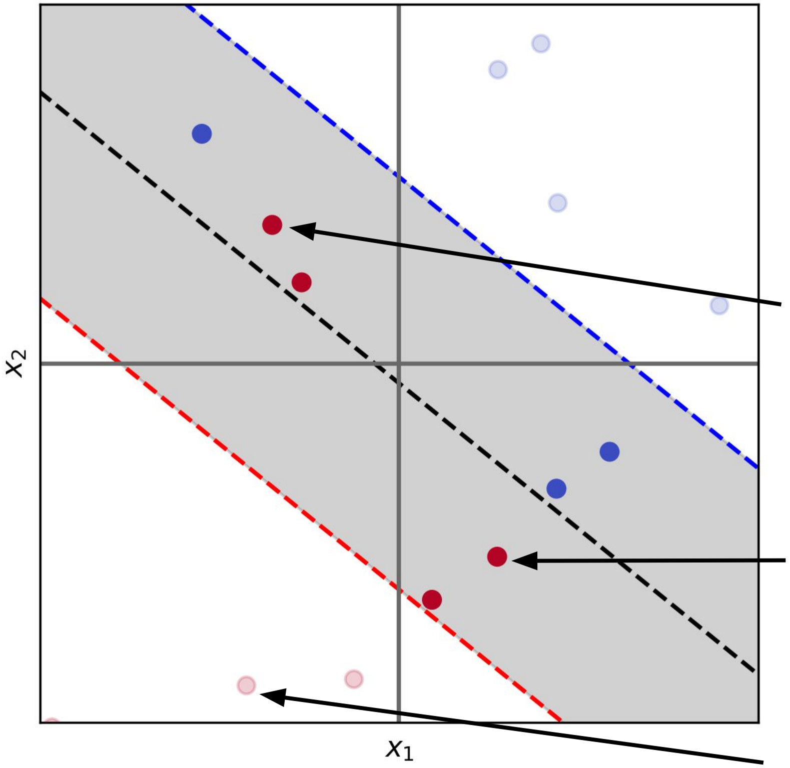
$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 \ \text{ subject to: } y_i(w^T x_i + b) \geq 1 \ \ \forall i$$

Soft Margin:

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$



Geometry of points near the Soft-SVM decision boundary



$$y_i(w^T x_i + b) < 0 \quad : \text{Incorrectly Classified}$$

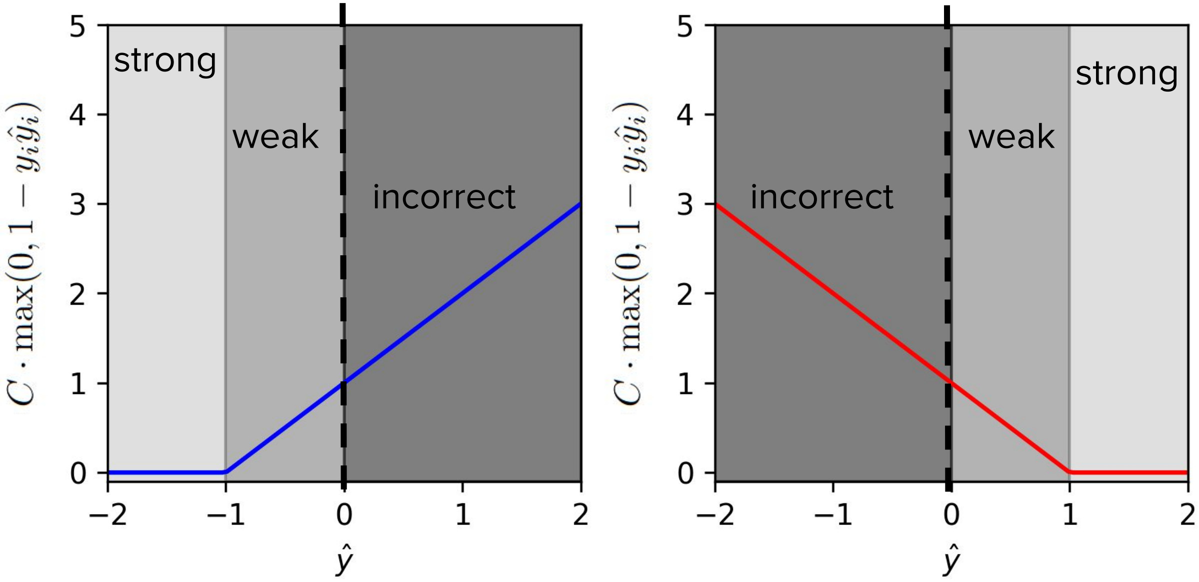
$$0 \leq y_i(w^T x_i + b) < 1 \quad : \text{Weakly Correctly Classified}$$

$$1 \leq y_i(w^T x_i + b) \quad : \text{Strongly Correctly Classified}$$

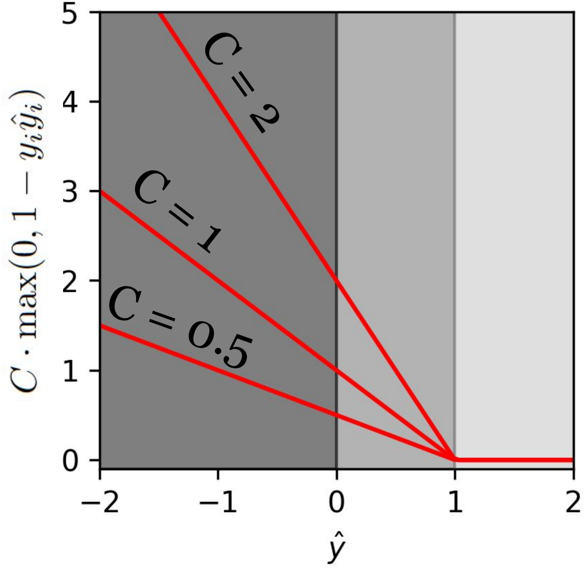
Building intuition for the margin/boundary violation penalty "Hinge Loss"

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C \underline{\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))}$$

$$\hat{y} = w^T x_i + b$$



Hyperparameter C sets the penalty on margin violations



$$\hat{w}, \hat{b} = \underset{w,b}{\text{argmin}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$

$C = 0$:

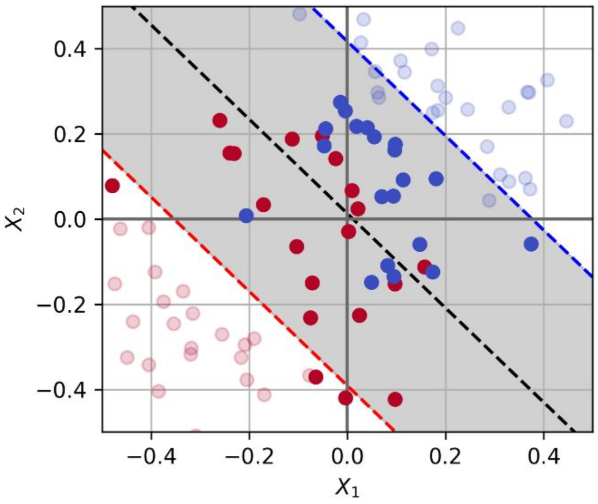
Drives to uninteresting solution, $\vec{w} = 0$

$C \to \infty$:

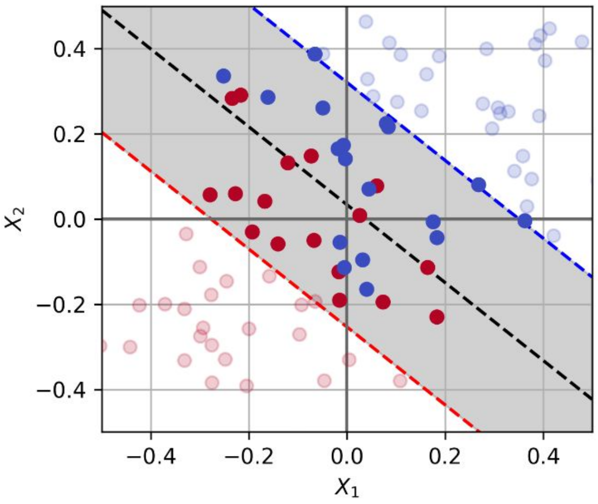
Reduces to Hard-margin SVM

Playing with hyperparameter C

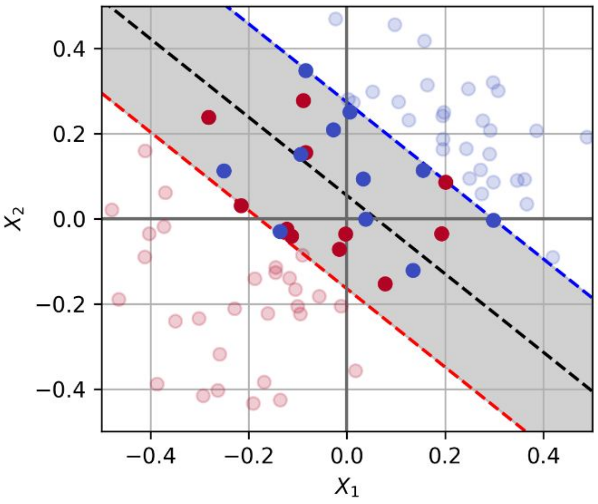
C = 1.0



C = 2.0



C = 10.0



Soft SVM as regularized regression

Ridge Regression:

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - (w^\top x_i + b))^2 + \lambda \|w\|^2 \right), \text{where}$$

$$l_w(x, y) = (y_i - (w^\top x_i + b))^2$$

$$\implies \hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \left(E[l(x,y)] + \lambda \|w\|^2 \right)$$

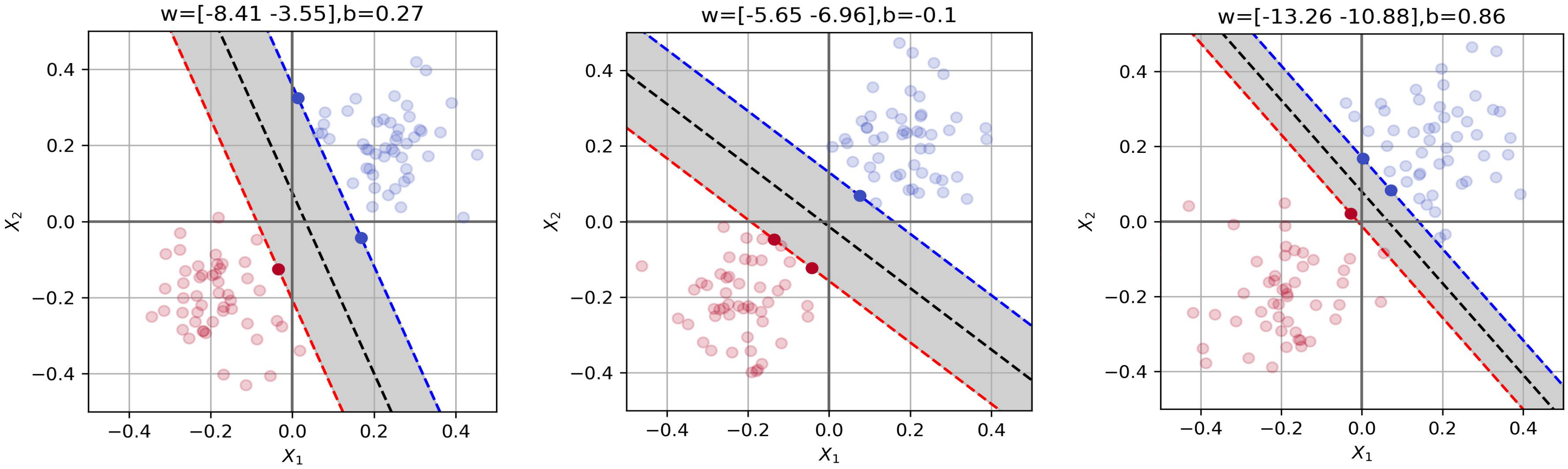
Soft SVM:

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \left(C \sum_{i=1}^{n} \max(0, 1 - y_i(w^\top x_i + b)) + \frac{1}{2} \|w\|^2 \right), \text{letting}$$

$$l_w(x, y) = \max(0, 1 - y_i(w^\top x_i + b))$$

$$\implies \hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \left(E[l(x,y)] + \frac{1}{2nC} \|w\|^2 \right)$$

The maximum margin classifier has high-variance

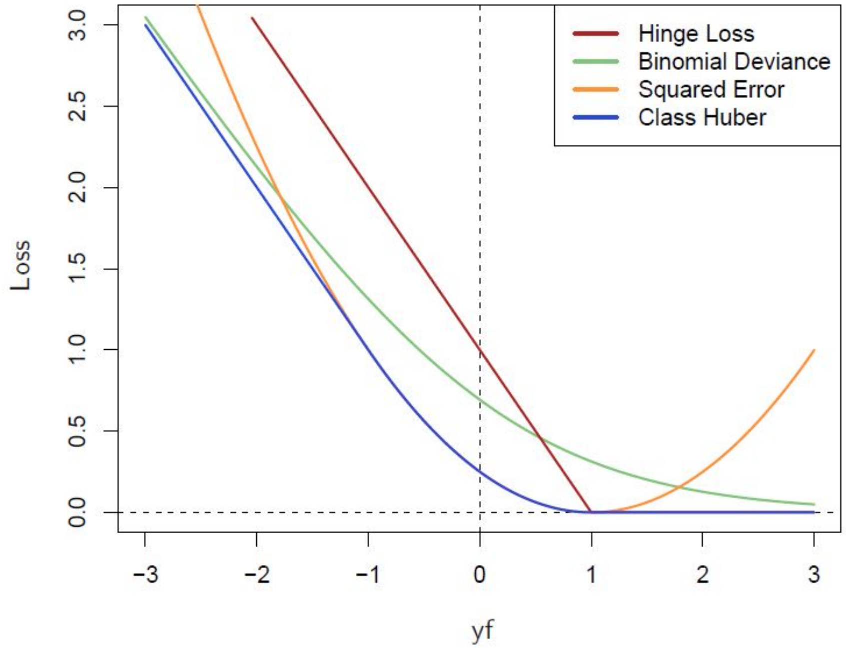


seed = 1

seed = 8

seed = 42

Comparing Hinge loss to other loss functions



Loss Function	$L[y, f(x)]$
Binomial Deviance	$\log[1 + e^{-yf(x)}]$
SVM Hinge Loss	$[1 - yf(x)]_+$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$

Elements of Statistical Learning, Section 12.3

Key Questions

I. How can we relax the hard-margin constraints?

II. **Can we gain any insights deriving the dual?**

III. How do we optimize?

Dual form of Hard SVM problem yielded substantial insights

Soft Margin: $\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$

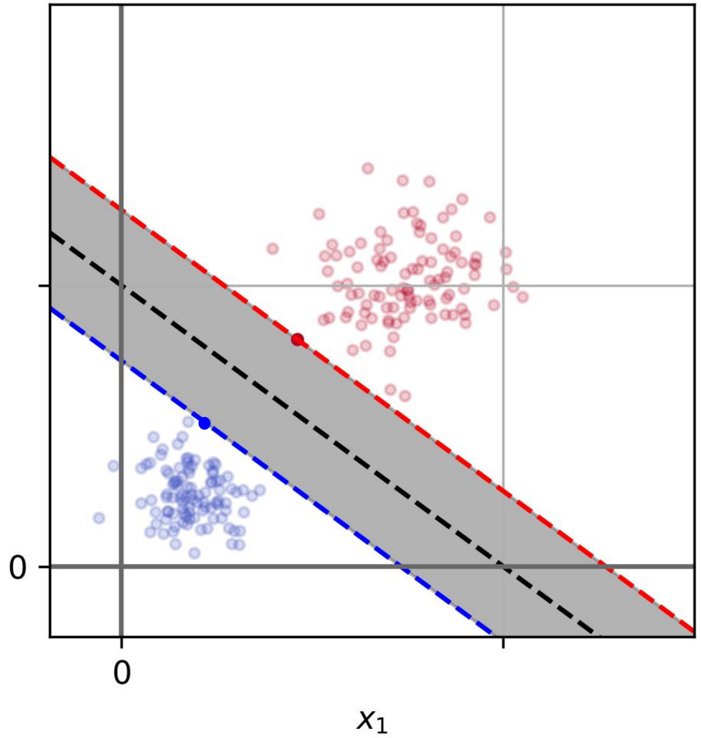
Hard Margin: $\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2$ subject to: $y_i(w^T x_i + b) \geq 1 \;\; \forall i$

$$\mathcal{L}(w, b, \lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

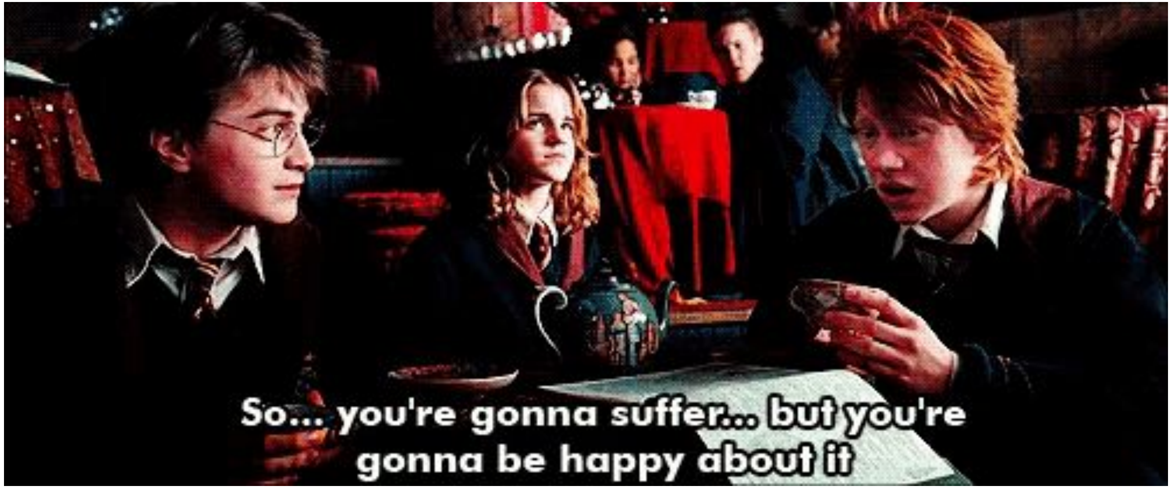
$\lambda_i > 0 \implies y_i(w^T x_i + b) = 1$
constraint is active, x_i defines the margin

$y_i(w^T x_i + b) > 1 \implies \lambda_i = 0$
constraint is inactive, x_i is far from the margin

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$



Deriving the dual of the Soft SVM problem



Deriving the dual of the Soft SVM problem: Slack variables

For a dataset of n samples $x_i \in \mathbb{R}^d$, with labels $y_i \in \{\pm 1\}$, linearly separated by a hyperplane parametrized by w and b; $w, x \in \mathbb{R}^d$, and $b \in \mathbb{R}$, the objective of the Soft SVM problem is given by

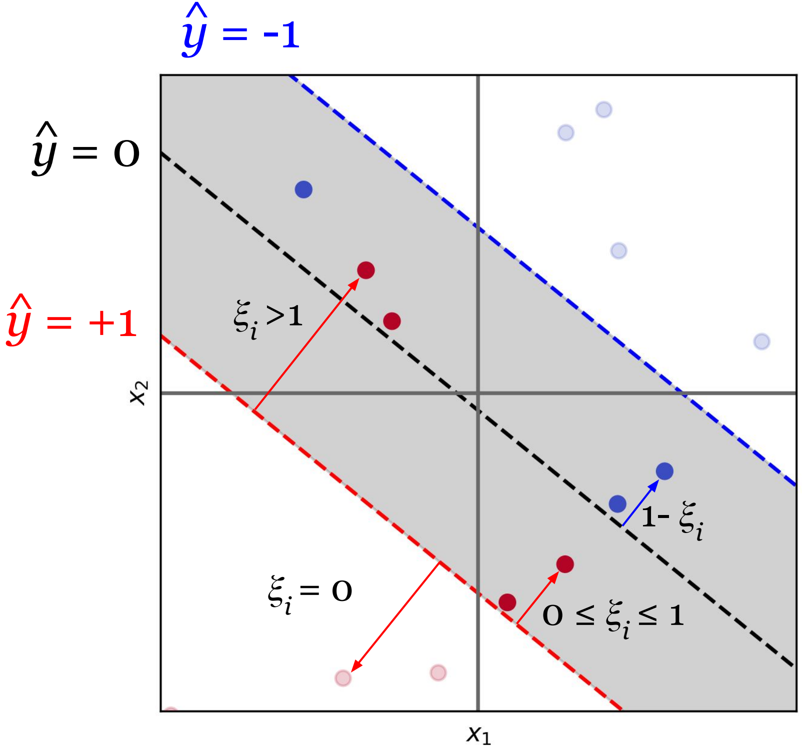
$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i \hat{y}_i)$$

where $\hat{y}_i = w^T x_i + b$

Towards deriving the dual form, we first introduce a constraint for each point in the dataset, using non-negative slack variables, ξ_i:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i \hat{y}_i \geq 1 - \xi_i \quad ; \xi_i \geq 0 \ \ \forall i$$



Deriving the dual of the Soft SVM problem: The Lagrangian

We next incorporate the constraints of

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t. } \begin{cases} y_i\hat{y}_i \geq 1-\xi_i & \forall i \\ \xi_i \geq 0 & \forall i \end{cases}$$

directly into the objective by introducing a set of Lagrange multipliers for each set of constraints to obtain the Lagrangian

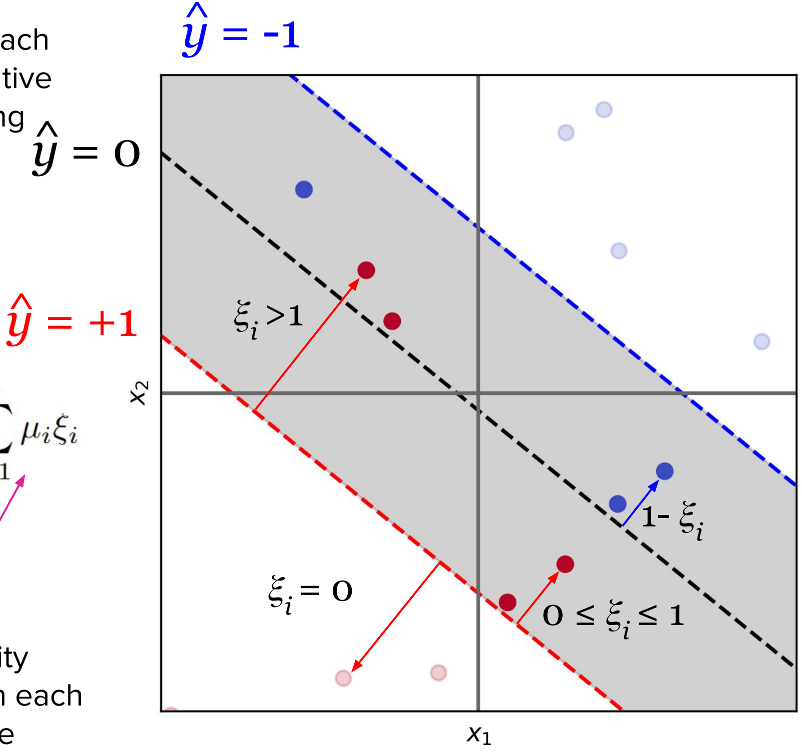
Constraint on each predictions relative to corresponding slack variable

$$L(w,b,\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\mu}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i[y_i\hat{y}_i - (1-\xi_i)] - \sum_{i=1}^{n}\mu_i\xi_i$$

Non-negativity constraint on each slack variable

Where $\boldsymbol{\lambda} = \{\lambda_i \geq 0\}$ and $\boldsymbol{\mu} = \{\mu_i \geq 0\}$ are sets of Lagrange multipliers. The corresponding objective is then

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}} L(w,b,\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\mu})$$



Expanding the Lagrangian

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i[y_i\hat{y}_i - (1-\xi_i)] - \sum_{i=1}^{n}\mu_i\xi_i\right)$$

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i y_i\hat{y}_i + \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i\right)$$

Substituting $\hat{y}_i = w^T x_i + b$

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i y_i(w^T x_i + b) + \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i\right)$$

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i y_i w^T x_i - \sum_{i=1}^{n}\lambda_i y_i b + \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i\right)$$

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\lambda_i y_i w^T x_i - \sum_{i=1}^{n}\lambda_i y_i b + \sum_{i=1}^{n}\lambda_i + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i\right)$$

Deriving the dual of the Soft SVM problem: Stationarity

$$\max_{\lambda,\mu\geq 0}\min_{w,b,\boldsymbol{\xi}}\left(\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\lambda_i y_i w^T x_i-\sum_{i=1}^{n}\lambda_i y_i b+\sum_{i=1}^{n}\lambda_i+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\lambda_i\xi_i-\sum_{i=1}^{n}\mu_i\xi_i\right)$$

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}\lambda_i y_i x_i=0$$

$$\implies w=\sum_{i=1}^{n}\lambda_i y_i x_i$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\lambda_i y_i=0$$

$$\implies \sum_{i=1}^{n}\lambda_i y_i=0$$

$$\frac{\partial L}{\partial \xi_i}=C-\lambda_i-\mu_i=0$$

$$\implies \mu_i=C-\lambda_i$$

Deriving the dual of the Soft SVM problem: Stationarity

$$\mathcal{L}_D = \max_{\lambda,\mu\geq 0} \min_{w,b,\boldsymbol{\xi}} \left(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i y_i w^T x_i - \sum_{i=1}^{n} \lambda_i y_i b + \sum_{i=1}^{n} \lambda_i + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i \xi_i - \sum_{i=1}^{n} \mu_i \xi_i \right)$$

$$\text{subject to:} \quad w = \sum_{i=1}^{n} \lambda_i y_i x_i, \quad \sum_{i=1}^{n} \lambda_i y_i = 0, \text{and} \quad \mu_i = C - \lambda_i$$

$$= \frac{1}{2}\Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big)^T \Big(\sum_{j=1}^{n} \lambda_j y_j x_j\Big) - \sum_{i=1}^{n} \lambda_i y_i \Big(\sum_{j=1}^{n} \lambda_j y_j x_j\Big) x_i - \cancel{\sum_{i=1}^{n} \lambda_i y_i b} + \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \xi_i \Big(C - \lambda_i - \mu_i\Big)$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j + \sum_{i=1}^{n} \lambda_i + \cancel{\sum_{i=1}^{n} \xi_i \Big(C - \lambda_i - (C - \lambda_i)\Big)}$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j + \sum_{i=1}^{n} \lambda_i$$

$$= \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

Wait… why?

Hard-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

Soft-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

Soft and Hard Margin SVM have different feasible regions

Hard-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

Soft-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

$$\mu_i \geq 0$$

$$\mu_i = C - \lambda_i$$

$$\implies C - \lambda_i \geq 0$$

$$\implies \lambda_i \leq C$$

$$\implies 0 \leq \lambda_i \leq C$$

Soft and Hard Margin SVM have different feasible regions

Hard-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

Soft-margin SVM:

$$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i x_j$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$$0 \leq \lambda_i \leq C$$

Geometry of support vectors in Soft-SVM

Complementary Slackness:

$$\lambda_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0$$

Either:

$\lambda_i = 0:$

$$\implies y_i(w^T x_i + b) - 1 + \xi_i > 0$$

$$\implies y_i(w^T x_i + b) > 1$$

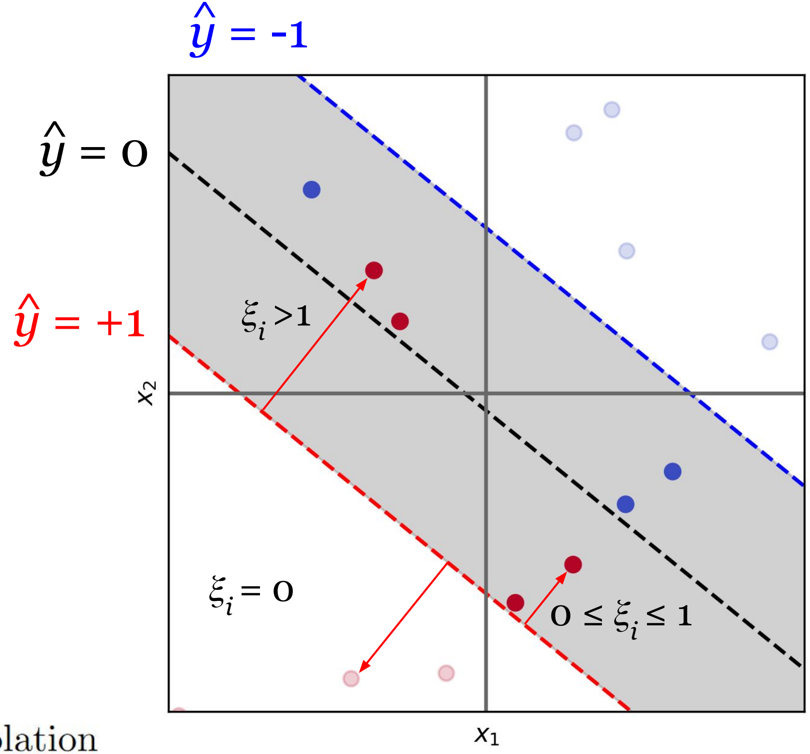
Or:

$\lambda_i > 0:$

$$\implies y_i(w^T x_i + b) - 1 + \xi_i = 0$$

$\xi = 0 \implies y_i(w^T x_i + b) = 1$ point is on the margin

$\xi > 0 \implies y_i(w^T x_i + b) < 1$ margin or boundary violation



The dual exposes a relationship between regularization and the size of the weights

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$0 \leq \lambda_i \leq C$$

Soft Margin:

$$\hat{w}, \hat{b} = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$

Key Questions

I. How can we relax the hard-margin constraints?

II. Can we gain any insights deriving the dual?

III. **How do we optimize?**

Solve for λ_i

λ_i

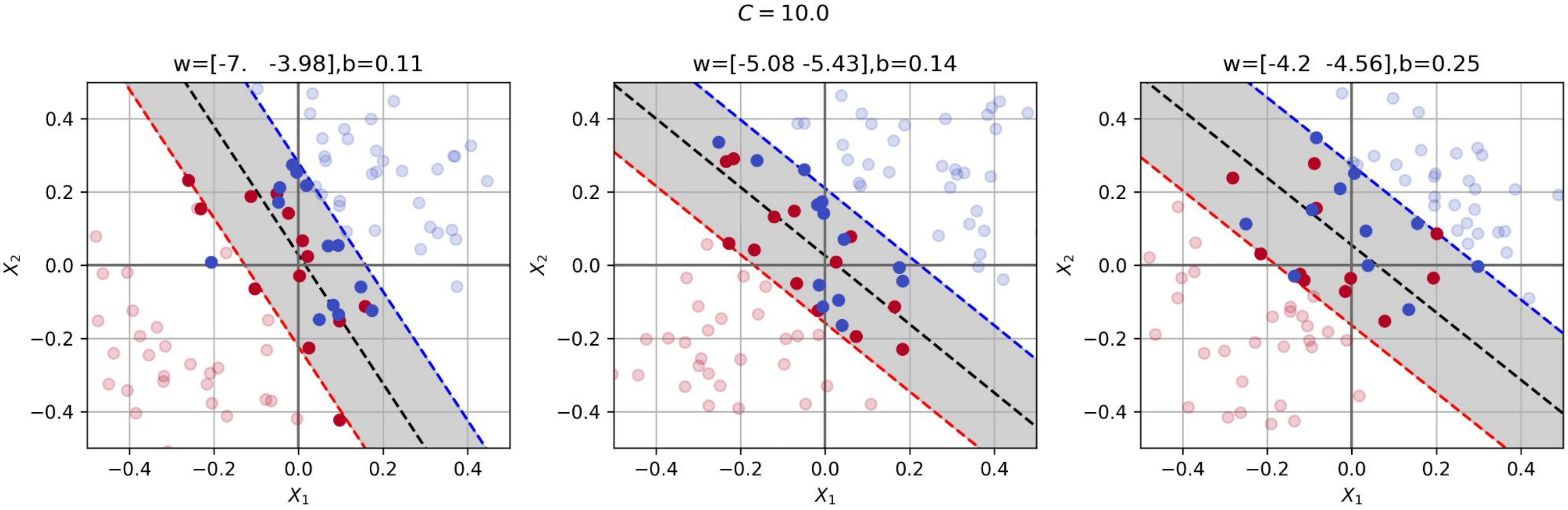
$$\sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j x_i x_j$$

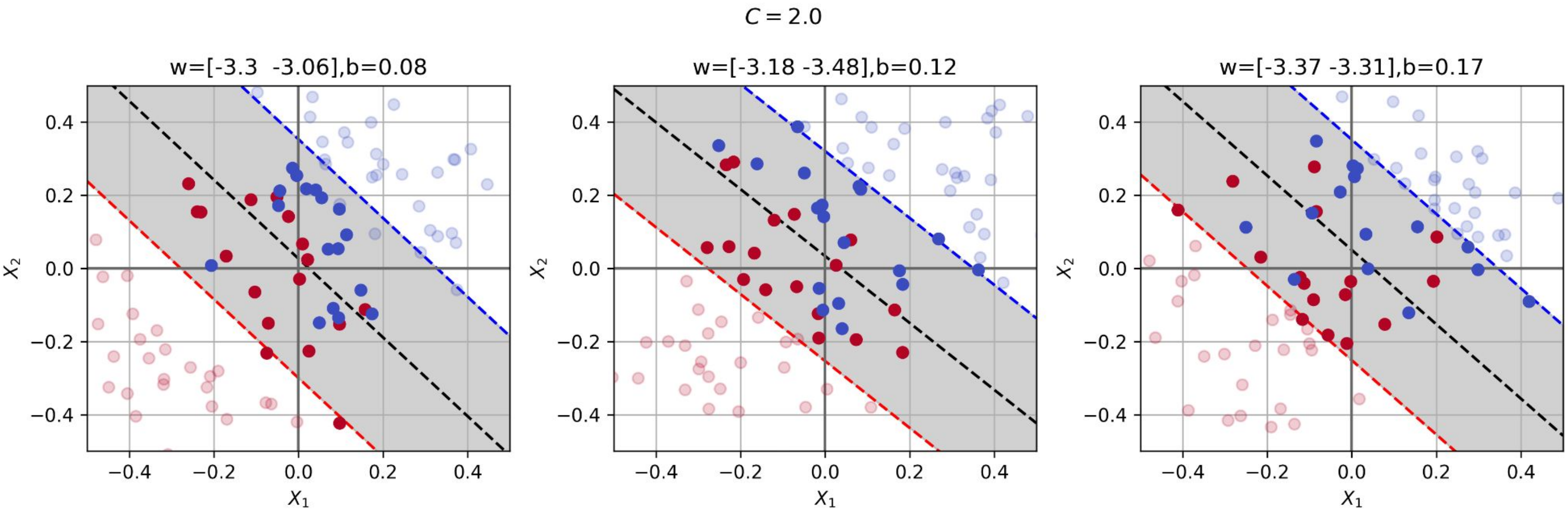
$$0 \leq \lambda_i \leq C$$

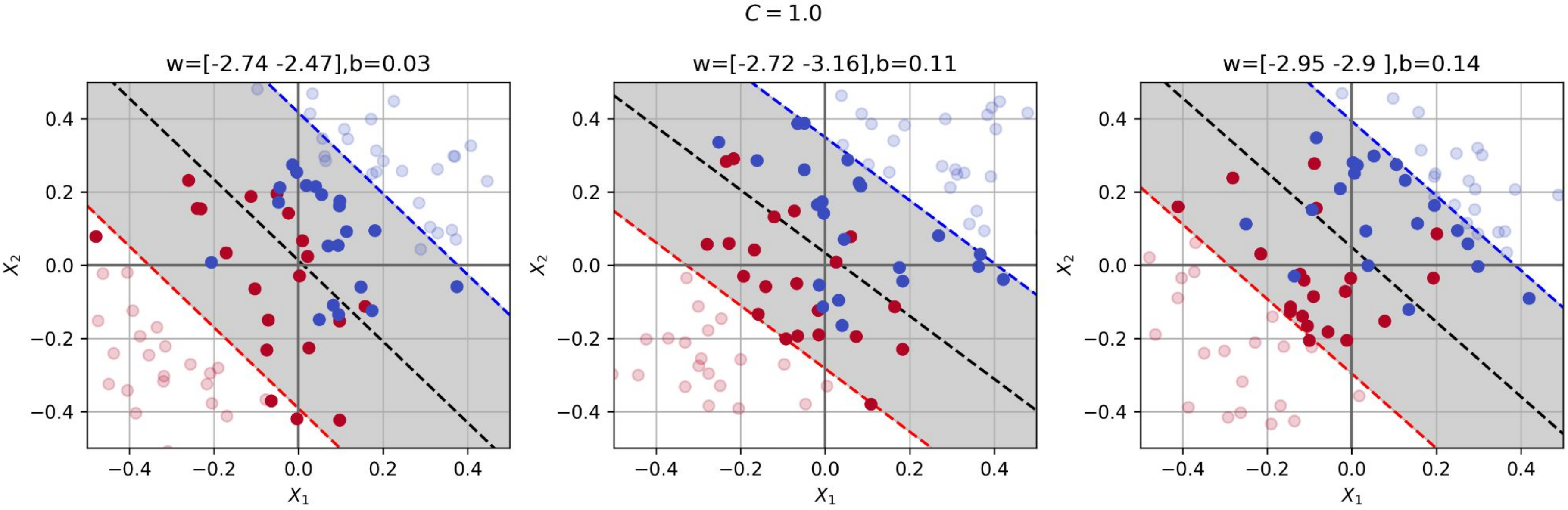
$$\sum_{i=1}^{n}\lambda_i y_i = 0$$

https://www.cvxpy.org/

```
import cvxpy as cp
import numpy as np

# Problem data.
m = 30
n = 20
np.random.seed(1)
A = np.random.randn(m, n)
b = np.random.randn(m)

# Construct the problem.
x = cp.Variable(n)
objective = cp.Minimize(cp.sum_squares(A @ x - b))
constraints = [0 <= x, x <= 1]
prob = cp.Problem(objective, constraints)

# The optimal objective value is returned by `prob.solve()`.
result = prob.solve()
# The optimal value for x is stored in `x.value`.
print(x.value)
# The optimal Lagrange multiplier for a constraint is stored in
# `constraint.dual_value`.
print(constraints[0].dual_value)
```


$C = 10.0$
w=[-7.   -3.98],b=0.11
w=[-5.08 -5.43],b=0.14
w=[-4.2 -4.56],b=0.25
X_1
X_2
0.4
0.2
0.0
−0.2
−0.4


$C = 2.0$
w=[-3.3 -3.06],b=0.08
w=[-3.18 -3.48],b=0.12
w=[-3.37 -3.31],b=0.17
X_1
X_2
0.4
0.2
0.0
−0.2
−0.4

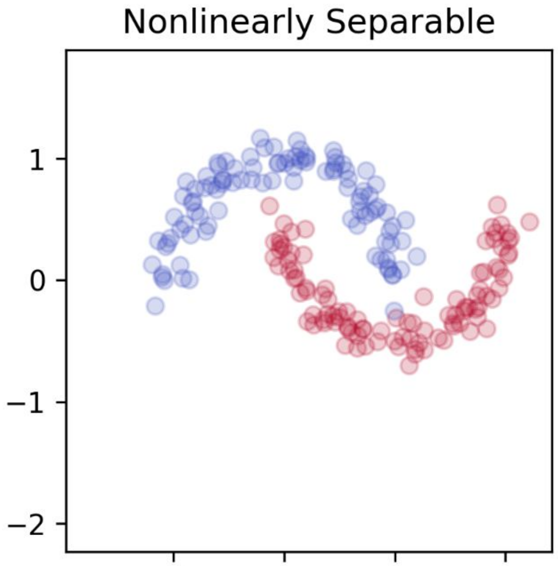

$C = 1.0$
w=[-2.74 -2.47],b=0.03
w=[-2.72 -3.16],b=0.11
w=[-2.95 -2.9],b=0.14
X_1
X_2
−0.4 −0.2 0.0 0.2 0.4

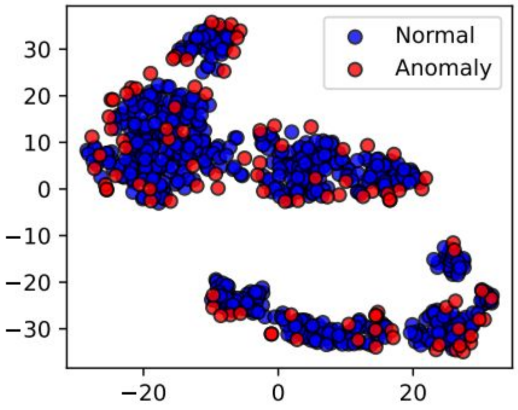
Can you derive the SGD update on the Primal objective?

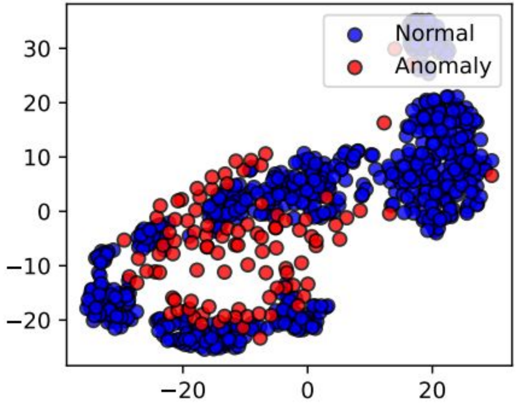
Now that we're at the end of the lecture, you should be able to…

★ Recognize the **hinge loss** for the soft-SVM problem
★ Interpret the **geometric properties of the soft-SVM decision boundary**, including **margin, support vectors,** and **slack variables**.
★ Describe the effect of **hyperparameter C** on model performance, generalization, and the norm of the weights.
★ Perform hyperparameter tuning using techniques like cross-validation to optimize soft-margin SVM parameter.
★ List the strengths and limitations of **algorithms for solving soft-margin SVMs**.

	Hard-Margin	Soft-Margin
Dual Objective	$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j x_i x_j$	$\mathcal{L}_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j x_i x_j$
Assumptions	Classes are linearly separable	Classes have some overlap
Misclassification	Not allowed	Allowed
Slack variables	N/A	ξ_i: One for each data point, measures margin violation
Dual variables	λ_i: enforces correct classification with some margin	λ_i: margin violation is no more than allowed slack μ_i: slack variables must be non-negative to enforce a penalty
KKT Conditions		
Stationarity	$w = \sum_{i=1}^{n} \lambda_i y_i x_i$ $\sum_{i=1}^{n} \lambda_i y_i = 0$	$w = \sum_{i=1}^{n} \lambda_i y_i x_i$ $\sum_{i=1}^{n} \lambda_i y_i = 0$ $\mu_i = C - \lambda_i$
Primal feasibility	$y_i(w^T x_i + b) - 1 \geq 0$	$y_i(w^T x_i + b) - 1 + \xi_i \geq 0$ $\xi_i \geq 0$
Dual feasibility	$\lambda_i \geq 0$	$\mu_i \geq 0$ $\lambda_i \geq 0$
Compl. slackness	$\lambda_i \left(y_i(w^T x_i + b) - 1\right) = 0$	$\lambda_i \left(y_i(w^T x_i + b) - 1 + \xi_i)\right) = 0$
Insights	Either: $\lambda_i > 0$ and point is on the margin, or $\lambda_i = 0$ and point is not on margin, far from the decision boundary and is correctly classified in such a way that the constraint does not directly influence the solution. Only points on the margin "support vectors" contribute to defining the separating hyperplane.	$0 \leq \lambda_i \leq C$


Nonlinearly Separable
1
0
−1
−2


Normal
Anomaly
30
20
10
0
−10
−20
−30
−20
0
20


Normal
Anomaly
30
20
10
0
−10
−20
−20
0
20