# CS540 Introduction to Artificial Intelligence
# Lecture 5

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 28, 2021

# Guess the Percentage

## Admin

Q1

- Guess what percentage of the students (who are here) started P1?
- A: 0 to 20 percent.
- B: 20 to 40 percent.
- C: 40 to 60 percent.
- D: 60 to 80 percent.
- E: 80 to 100 percent.

# The Percentage

## Admin

Q2

- Did you start P1?
- A:
- B: Yes.
- C:
- D: No.
- E:

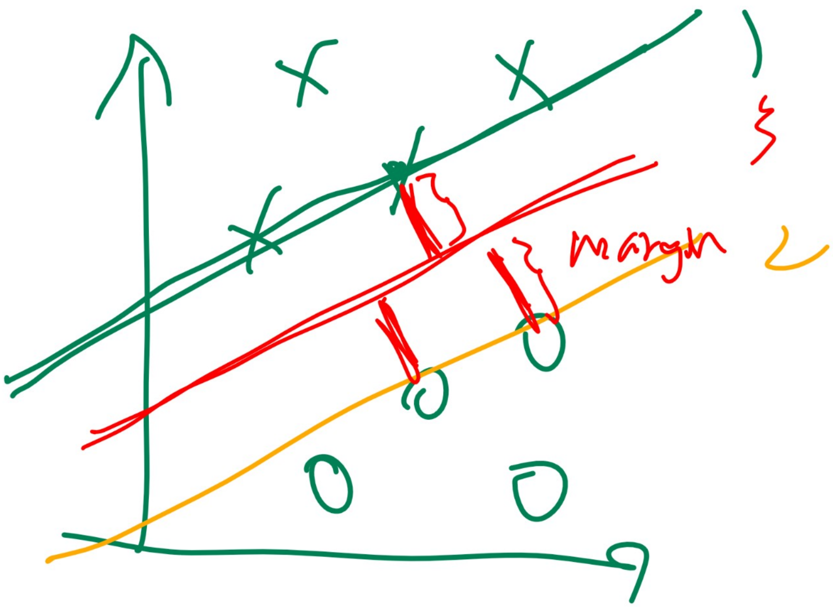
# Remind Me to Start Recording

## Admin

- The messages you send in chat will be recorded: you can change your Zoom name now before I start recording.
- There will be more smaller quiz questions during the lectures (not all at the end).
- No lecture next Monday.

# Maximum Margin Diagram

## Motivation



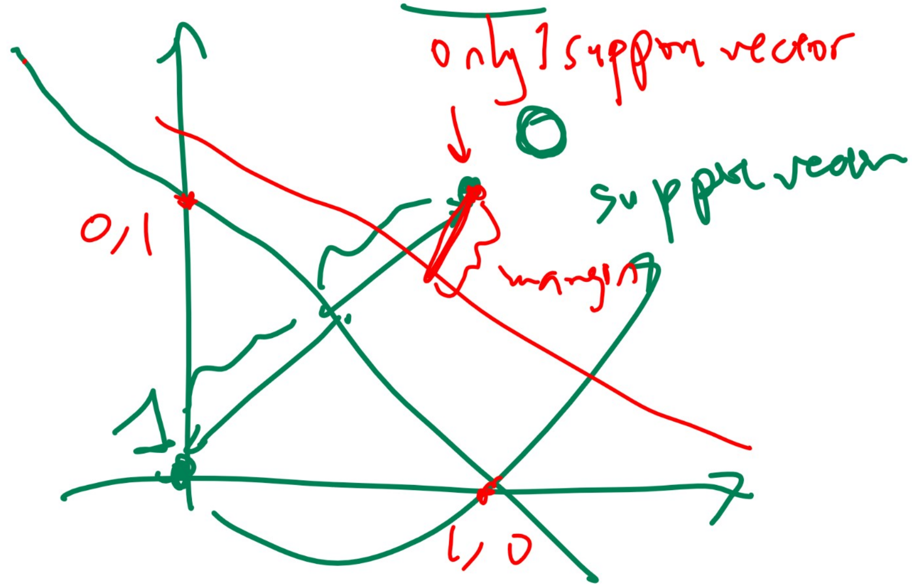

loss = 0

max margin

# SVM Weights

## Quiz

- Fall 2005 Final Q15 and Fall 2006 Final Q15 support vector
- Find the weights $w_1, w_2$ for the SVM classifier $\mathbb{1}_{\{\overset{-1}{w_1} x_{i1} + \overset{-1}{w_2} x_{i2} + 1 \geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = 0$.
- A: $w_1 = 0, w_2 = -2$
- B: $w_1 = -2, w_2 = 0$
- C: $w_1 = -1, w_2 = -1$
- D: $w_1 = -2, w_2 = -2$



# SVM Weights Diagram

## Quiz

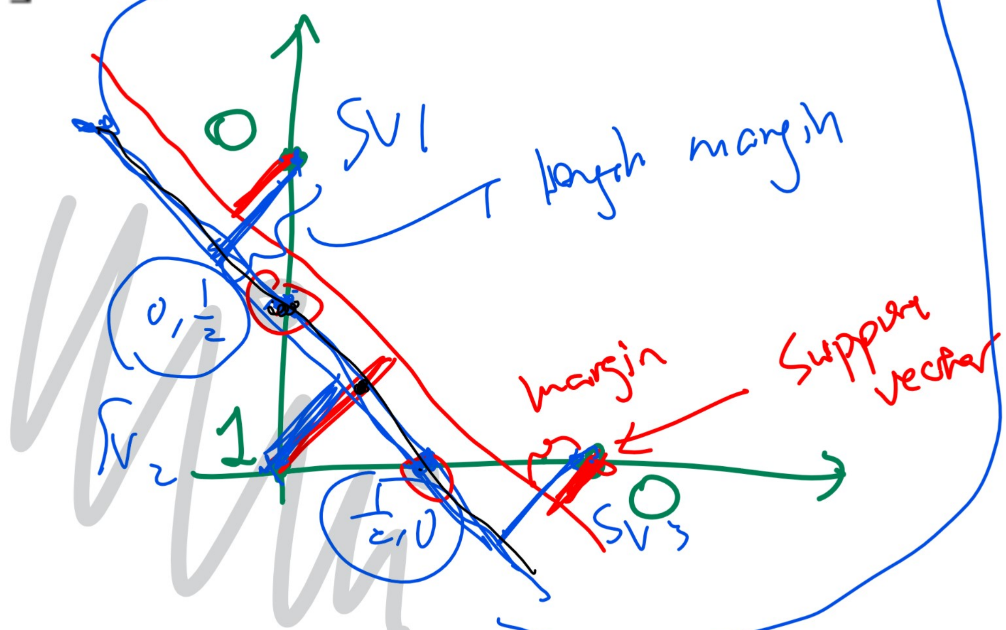
# SVM Weights

## Quiz

Q3

- Fall 2005 Final Q15 and Fall 2006 Final Q15
- Find the weights $w_1, w_2$ for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geq 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = y_3 = 0$.
- A: $w_1 = -1.5, w_2 = -1.5$
- B: $w_1 = -2, w_2 = -1.5$
- C: $w_1 = -1.5, w_2 = -2$
- D: $w_1 = -2, w_2 = -2$
- E: $w_1 = -4, w_2 = -4$

$w_1 x_1 + w_2 x_2 + 1 = 0$

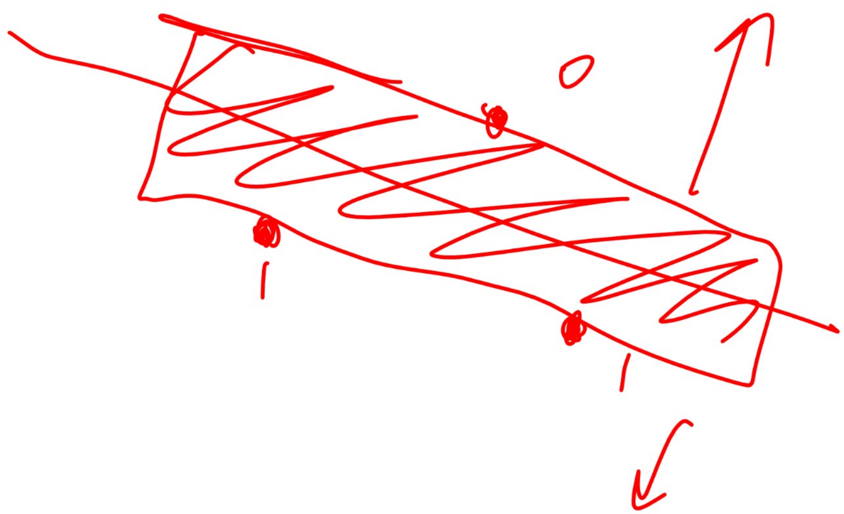



$w_1 \cdot 0 + w_2 \cdot \frac{1}{2} + 1 = 0$

AND $w_1 \cdot \frac{1}{4} + w_2 \cdot \frac{1}{4} + 1 = 0$

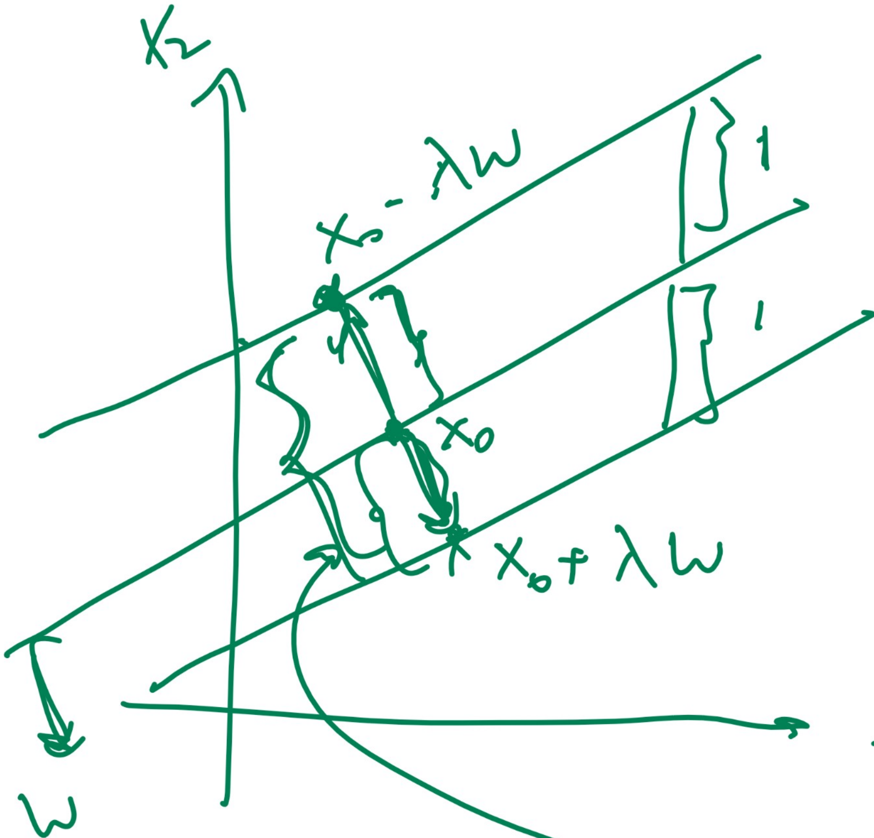
# SVM Weights Diagram

## Quiz



# Constrained Optimization Derivation

Definition




$$\boxed{w^T x + b = 1}$$

$$w^T x + b = 0$$

$$\boxed{w^T x + b = -1}$$

$$w^T(x_0 - \lambda w) + b = 1$$

$$w^T(x_0 + \lambda w) + b = -1$$

$$2\lambda w^T w = -2$$

$$\lambda = -\frac{1}{w^T w}$$

$$\|\lambda w\| = |\lambda| \|w\| = \frac{1}{w^T w}\sqrt{w^T w} = \frac{1}{\sqrt{w^T w}}$$

# Constrained Optimization

## Definition

- The goal is to maximize the margin subject to the constraint that the plus plane and the minus plane separates the instances with $y_i = 0$ and $y_i = 1$.

$$\max_w \frac{2}{\sqrt{w^T w}} \text{ such that } \begin{cases} \left(w^T x_i + b\right) \leqslant -1 & \text{if } y_i = 0 \\ \left(w^T x_i + b\right) \geqslant 1 & \text{if } y_i = 1 \end{cases}, i = 1, 2, ..., n$$

margin

- The two constraints can be combined.

$$\max_w \frac{2}{\sqrt{w^T w}} \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

$\{0, 1\}$

$\{-1, 1\}$

# Hard Margin SVM

## Definition

$$\max_{w} \frac{2}{\sqrt{w^T w}} \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, \ldots, n$$

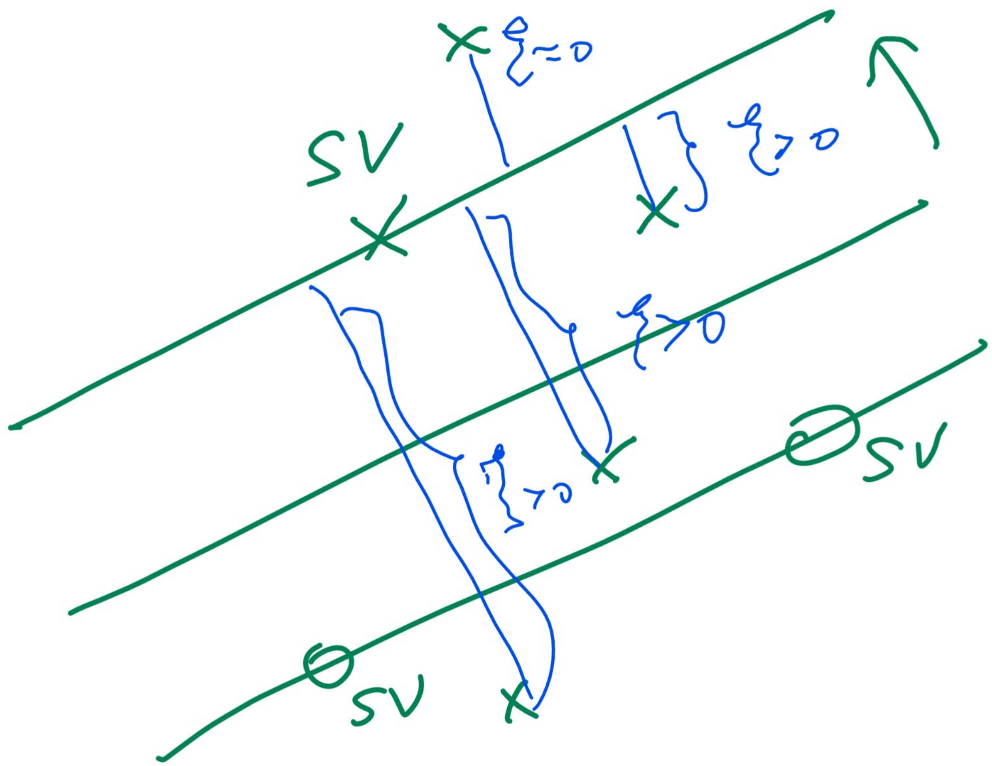
different objective value, same w.

- This is equivalent to the following minimization problem, called hard margin SVM.

$$\boxed{\min_{w} \frac{1}{2} w^T w} \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, \ldots, n$$

# Soft Margin Diagram

## Definition



# Soft Margin SVM

## Definition

max margin

$\xi$

$$\min_{w} \frac{1}{2} w^T w + \frac{1}{\lambda}\frac{1}{n}\sum_{i=1}^{n} \xi_i$$

min mistake

such that $\boxed{(2y_i - 1)\left(w^T x_i + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0,}\ i = 1, 2, ..., n$

trade off parameter of margin vs mistake.

- This is equivalent to the following minimization problem, called $\boxed{\text{soft margin SVM.}}$

$\xi \geq 1 - (2y_i - 1)(w^T x_i + b)$

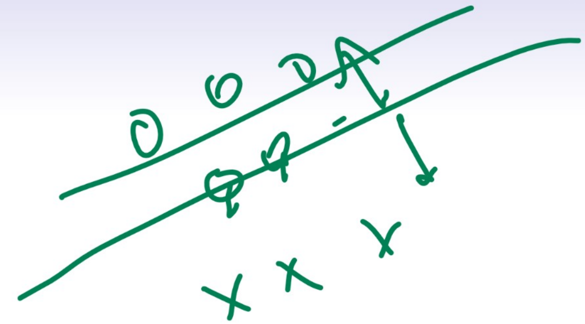
$\xi \geq 0$

$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n}\sum_{i=1}^{n} \boxed{\max\left\{0, 1 - (2y_i - 1)\left(w^T x_i + b\right)\right\}}$$

min $x$ $x \geq b$ $x = b$

# Soft Margin

## Quiz



- Fall 2011 Midterm Q8 and Fall 2009 Final Q1
- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$, $y = 0$, what is the smallest slack variable $\xi$ for it to satisfy the margin constraint?

18

$$(2y_i - 1)\left(w^T x_i + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0$$

0

$$-1\left((1,2)\begin{pmatrix} 4 \\ 5 \end{pmatrix} + 3\right) \geq 1 - \xi$$

$$-17 \geq 1 - \xi$$

$$\xi \geq \underline{18} \qquad \xi \geq \underline{0}$$

# Soft Margin 2

## Quiz

Q4

- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} -4 \\ -5 \end{bmatrix}$, $y = 0$, what is the smallest slack variable $\boxed{\xi}$ for it to satisfy the margin constraint?
- ~~A: -12~~
- ~~B: -10~~
- C: 0
- D: 10
- E: 12

$$\boxed{(2y_i - 1)(w^T x_i + b) \geq 1 - \xi_i}, \quad \boxed{\xi_i \geq 0}$$

$$-1\left((1,2)\begin{pmatrix}-4\\-5\end{pmatrix} + 3\right) \geq 1 - \xi$$

$$+11 \geq 1 - \xi$$

$$\boxed{\xi \geq -10}$$

# Subgradient Descent

## Definition

$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, 1 - (2y_i - 1) \left( w^T x_i + b \right) \right\}$$

- The gradient for the above expression is not defined at points with $1 - (2y_i - 1) \left( w^T x_i + b \right) = 0$.
- Subgradient can be used instead of a gradient.

# Subgradient

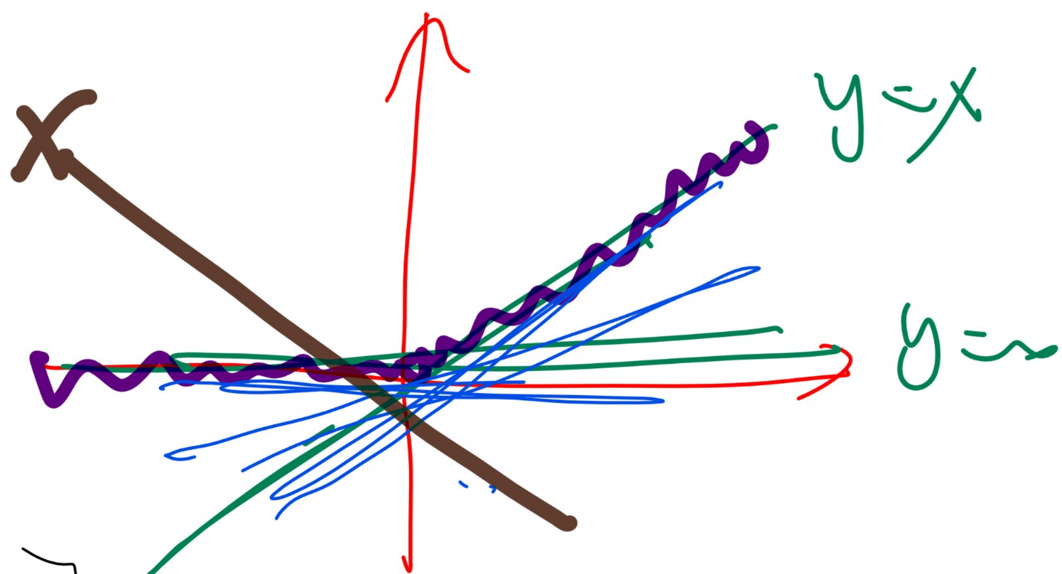
- The subderivative at a point of a convex function in one dimension is the set of slopes of the lines that are tangent to the function at that point.
- The subgradient is the version for higher dimensions.
- The subgradient $\partial f(x)$ is formally defined as the following set.

$$\partial f(x) = \left\{ v : f(x') \geqslant f(x) + v^T (x' - x) \ \forall x' \right\}$$

# Subgradient 1

## Quiz

- Which ones are subderivatives of $\max\{x, 0\}$ at $x = 0$?
- A: -1
- B: -0.5
- C: 0
- D: 0.5
- E: 1



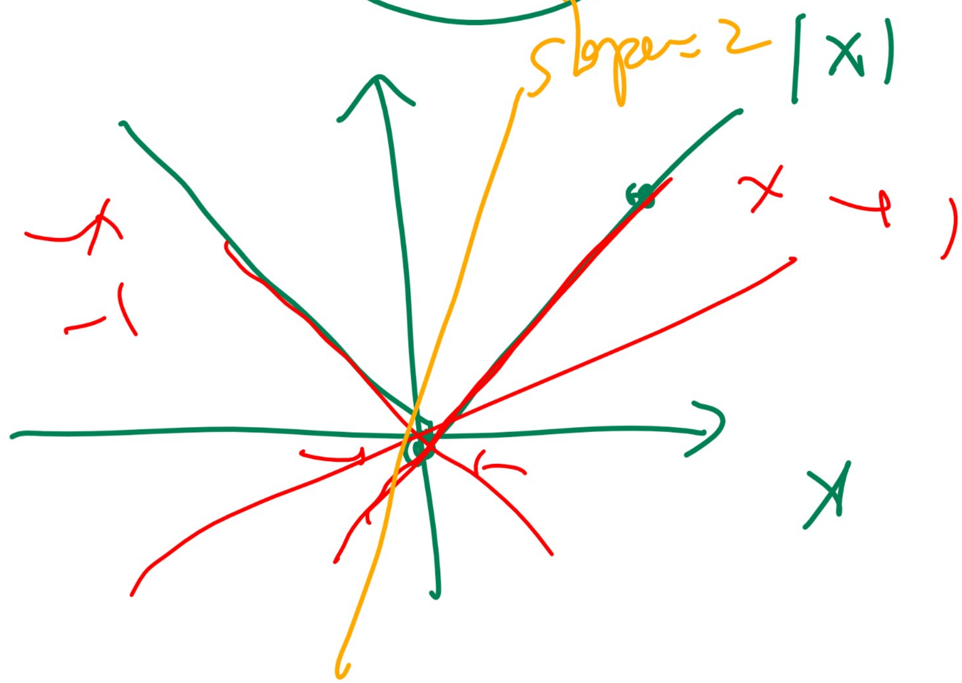

$$\partial_x \max = [0, 1]$$

# Subgradient 2

## Quiz

Q5

- Which ones are subderivatives of $|x|$ at $x = 0$?
- A: -1
- B: -0.5
- C: 0
- D: 0.5
- E: 1

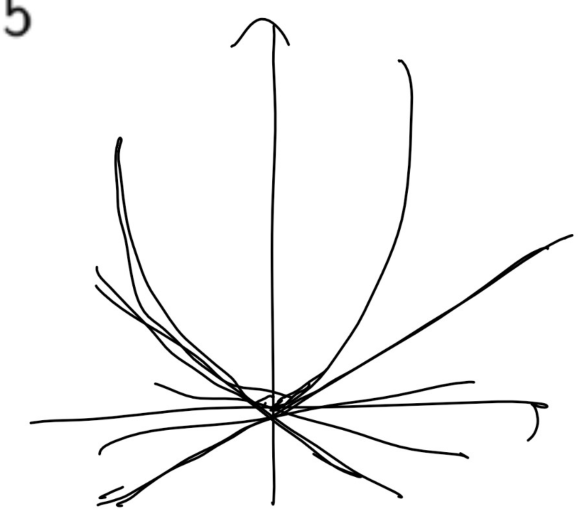


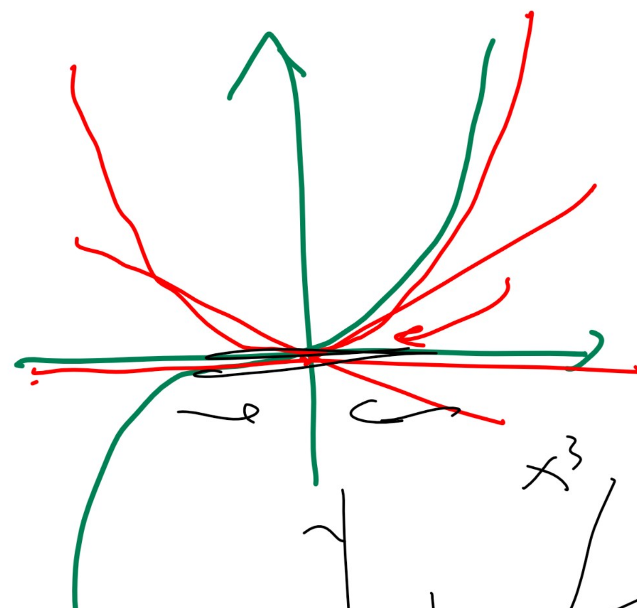
# Subgradient 3

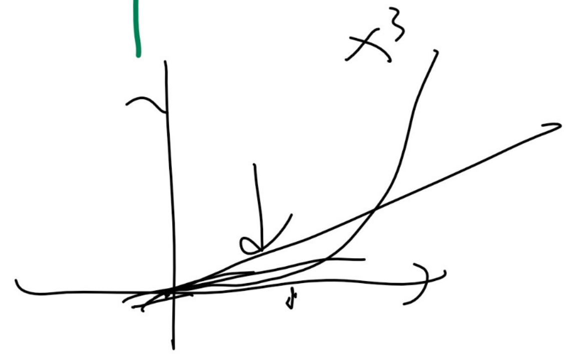
## Quiz

Q6

- Which ones are subderivatives of $|x^3|$ at $x = 0$?
- A: -1
- B: -0.5
- C: 0
- D: 0.5
- E: 1







# Subgradient Descent Step

## Definition

- One possible set of subgradients with respect to $w$ and $b$ are the following.

max{ , 0}

$\nabla_w C$

$$\partial_w C \ni \lambda w - \sum_{i=1}^{n} (2y_i - 1)\, x_i \mathbb{1}_{\{(2y_i-1)(w^T x_i + b) \geqslant 1\}} \in \partial_w C$$

[0, 1]

$$\partial_b C \ni - \sum_{i=1}^{n} (2y_i - 1)) \mathbb{1}_{\{(2y_i-1)(w^T x_i + b) \geqslant 1\}}$$

- The gradient descent step is the same as usual, using one of the subgradients in place of the gradient.

# PEGASOS Algorithm

Primal Estimated subGrAdient SOlver for Svm

## Algorithm

- Inputs: instances: $\{x_i\}_{i=1}^n$ and $\{z_i = 2y_i - 1\}_{i=1}^n$
- Outputs: weights: $\{w_j\}_{j=1}^m$
- Initialize the weights.

-1, 1

$$w_j \sim \text{ Unif } [\cancel{0}, 1]$$

- Randomly permute (shuffle) the training set and performance subgradient descent for each instance $i$.

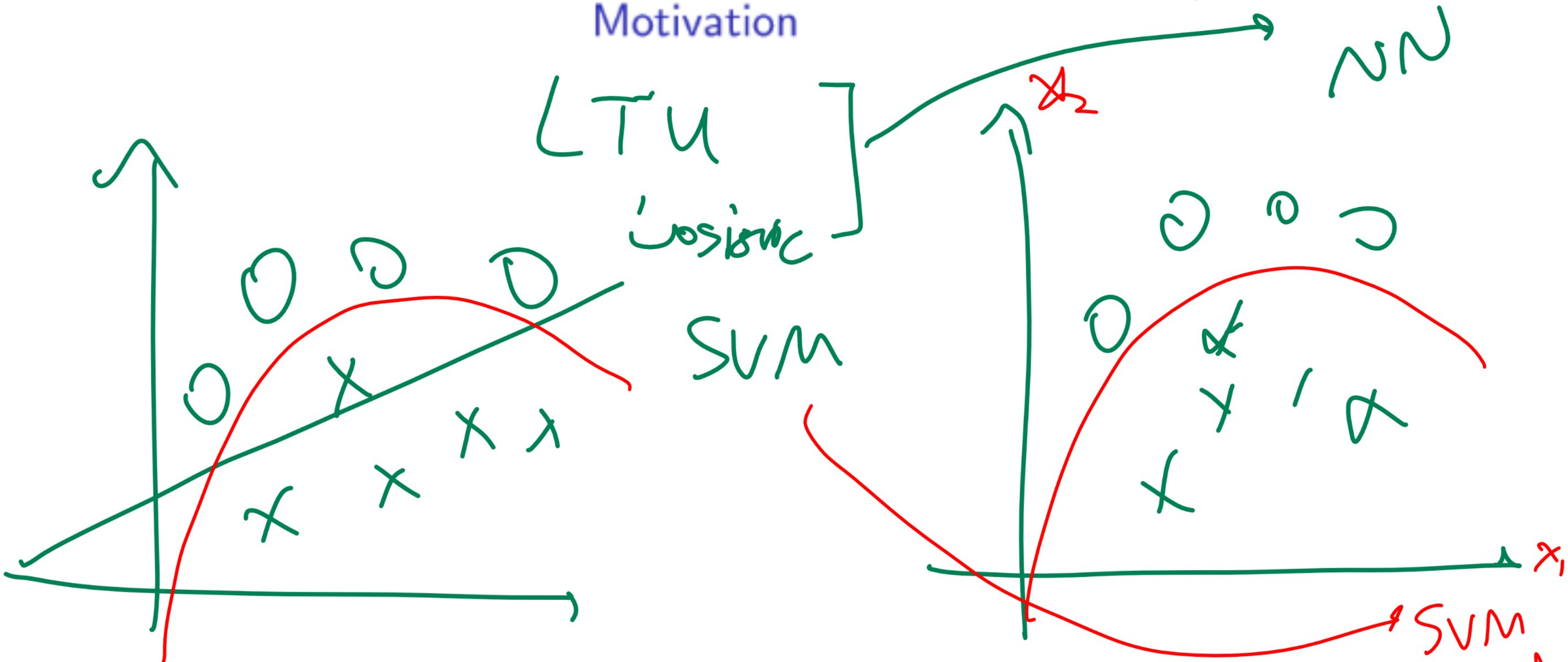
$$w = (1 - \lambda)\, w - \alpha z_i \mathbb{1}_{\{z_i w^T x_i \geq 1\}} x_i$$

- Repeat for a fixed number of iterations.

# Kernel Trick 1$D$ Diagram

$x_3 = x_2^2$

## Motivation



# Kernelized SVM

## Definition

- With a feature map $\varphi$, the SVM can be trained on new data points $\{(\varphi(x_1), y_1), (\varphi(x_2), y_2), ..., (\varphi(x_n), y_n)\}$.
- The weights $w$ correspond to the new features $\varphi(x_i)$.
- Therefore, test instances are transformed to have the same new features.

$$\hat{y}_i = \mathbb{1}_{\{w^T\varphi(x_i) \geqslant 0\}}$$

# Kernel Trick for XOR

## Quiz

- March 2018 Final Q17
- SVM with quadratic kernel $\varphi(x) = \left(x_1^2, \sqrt{2}x_1 x_2, x_2^2\right)$ can correctly classify the following training set?

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

# Kernel Trick for XOR

## Quiz

- SVM with kernel $\varphi(x) = (x_1, x_1 x_2, x_2)$ can correctly classify the following training set?

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

- A: True.
- B: False.

# Kernel Matrix

## Definition

- The feature map is usually represented by a $n \times n$ matrix $K$ called the Gram matrix (or kernel matrix).

$$K_{ii'} = \varphi\left(x_i\right)^T \varphi\left(x_{i'}\right)$$

# Examples of Kernel Matrix

## Definition

- For example, if $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, then the kernel matrix can be simplified.

$$K_{ii'} = \left(x_i^T x_{i'}\right)^2$$

- Another example is the quadratic kernel $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^2$. It can be factored to have the following feature representations.

$$\varphi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)$$

# Examples of Kernel Matrix Derivation

## Definition

# Popular Kernels

## Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ii'} = x_i^T x_{i'}$
2. Polynomial kernel: $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^d$
3. Radial Basis Function (Gaussian) kernel:
$$K_{ii'} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_{i'}\right)^T\left(x_i - x_{i'}\right)\right)$$

- Gaussian kernel has infinite-dimensional feature representations. There are dual optimization techniques to find $w$ and $b$ for these kernels.

# Kernel Matrix

## Quiz

- Fall 2009 Final Q2
- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = \exp(x_i + x_{i'}) + \sqrt{x_i x_{i'}} + 3$?
- A: $(\exp(x), \sqrt{x}, 3)$
- B: $\left(\exp(x), \sqrt{x}, \sqrt{3}\right)$
- C: $\left(\sqrt{\exp(x)}, \sqrt{x}, 3\right)$
- D: $\left(\sqrt{\exp(x)}, \sqrt{x}, \sqrt{3}\right)$
- E: None of the above

# Kernel Matrix Math

## Quiz