M8. Classification: Support Vector Machines (SVMs)

Manikandan Narayanan

Week 13 (Oct 20- 2025)

PRML Jul-Nov 2025 (Grads Section)

Acknowledgment of Sources

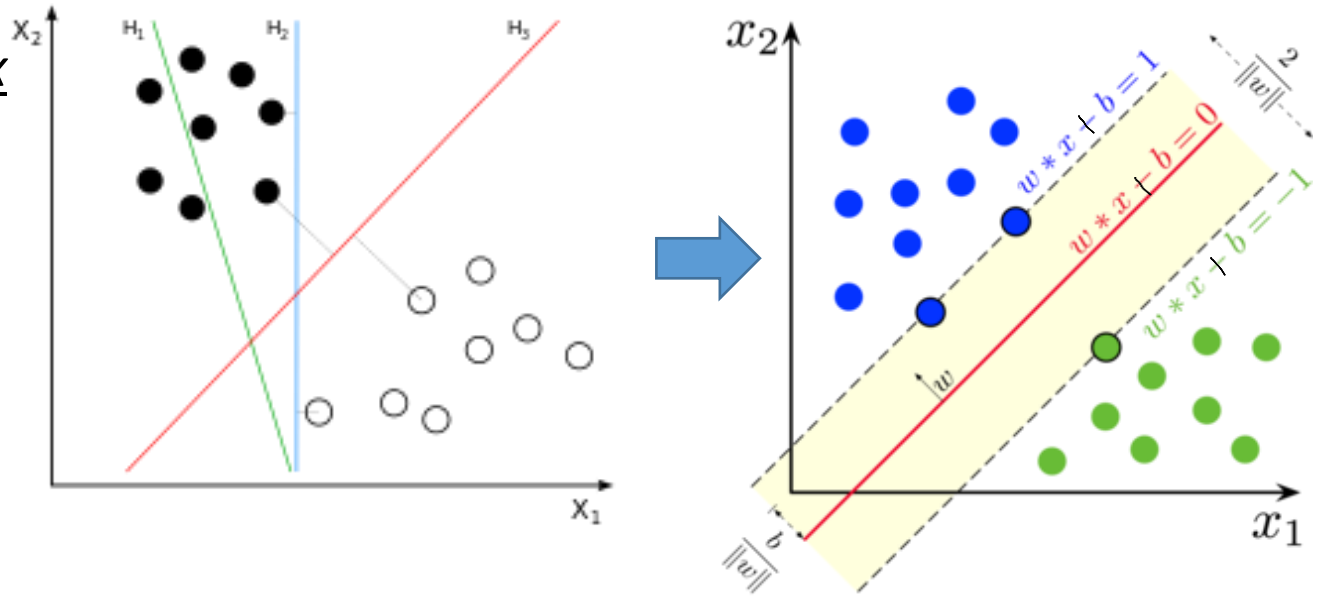
- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/Harish/Chandra's PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi's "Intro to ML" slides – cited respectively as [AR], **[HR]**, [CC], [BR] in the bottom right of a slide.
    - India – NPTEL PR course by IISc Prof. PS. Sastry (slides, etc.) – cited as [PSS] in the bottom right of a slide.
  - Books:
    - PRML by Bishop. (content, figures, slides, etc.) – cited as [CMB]
    - Pattern Classification by Duda, Hart and Stork. (content, figures, etc.) – [DHS]
    - Mathematics for ML by Deisenroth, Faisal and Ong. (content, figures, etc.) – [DFO]
    - Information Theory, Inference and Learning Algorithms by David JC MacKay – [DJM]

Outline for Module M8

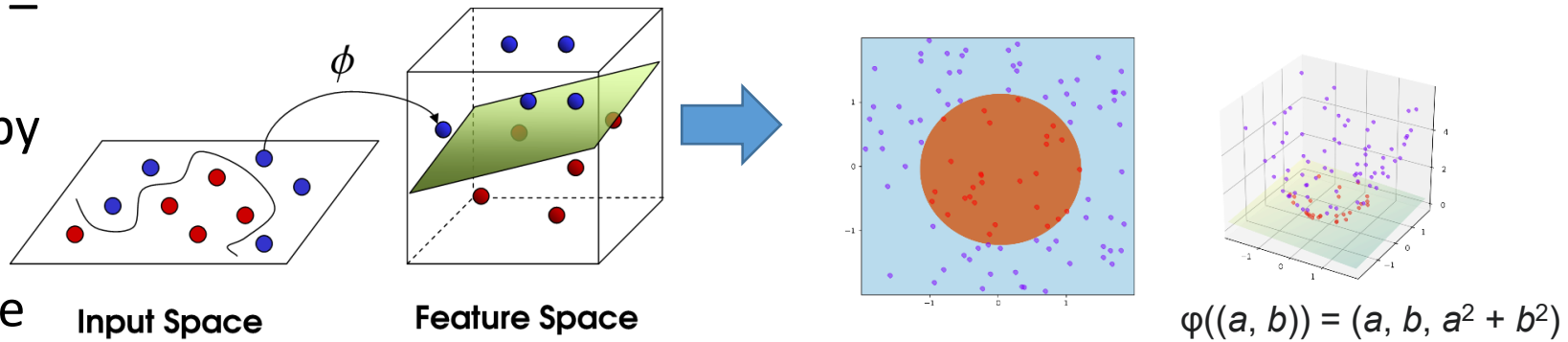
- M8. Classification (Support Vector Machines)
  - **M8.0 Introduction/Motivation**
    - **(concrete understanding of SVMs – beyond popular pictures & software)**
  - M8.1 SVM Problem Statement
    - (Hard/Soft-Margin SVM Problems)
  - M8.2 SVM Solution
    - (Background: Constrained optimization - KKT & Primal-Dual)
    - (SVM Dual Problem & Optimization algo. sketch)
  - M8.3 SVM Interpretations
    - (Support vectors, Kernels, Loss function view)
  - M8.4 Concluding thoughts

SVM hard-margin: popular pic. → geometry

- (Linear) SVM – *max margin, sparse support vectors*



- (Non-linear) SVM – uses *non-linear kernels* followed by applying SVM above in the feature map space

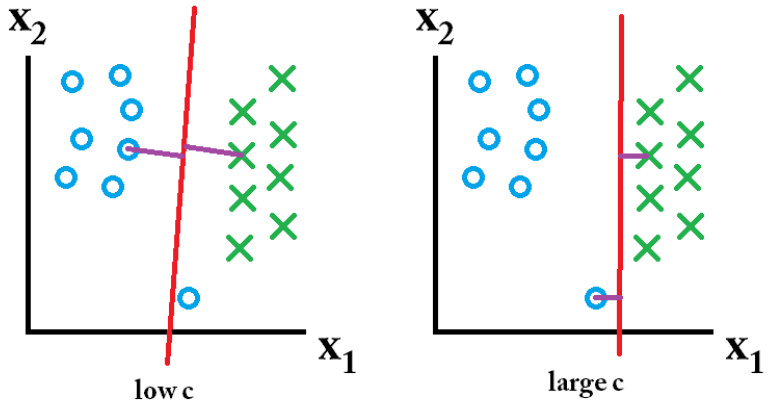


φ((*a*, *b*)) = (*a*, *b*, $a^2 + b^2$)

[Images source: https://en.wikipedia.org/wiki/Support-vector_machine]

SVM soft-margin (popular pic./software → …)

- Parameter C controls where you lie in the soft-hard margin spectrum.



[From: https://stats.stackexchange.com/a/159051]

- Software

sklearn.svm.SVC

class sklearn.svm. **SVC**(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=- 1, decision_function_shape='ovr', break_ties=False, random_state=None) ¶ [source]

SVM soft-margin: … → concrete understanding

Primal OP:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0, i = 1, \ldots, n$$

Dual OP:

$$\max_{\alpha} -\frac{1}{2} \alpha^T Q \alpha + e^T \alpha$$

$$\text{subject to } y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$.

Prediction for new point x:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b,$$

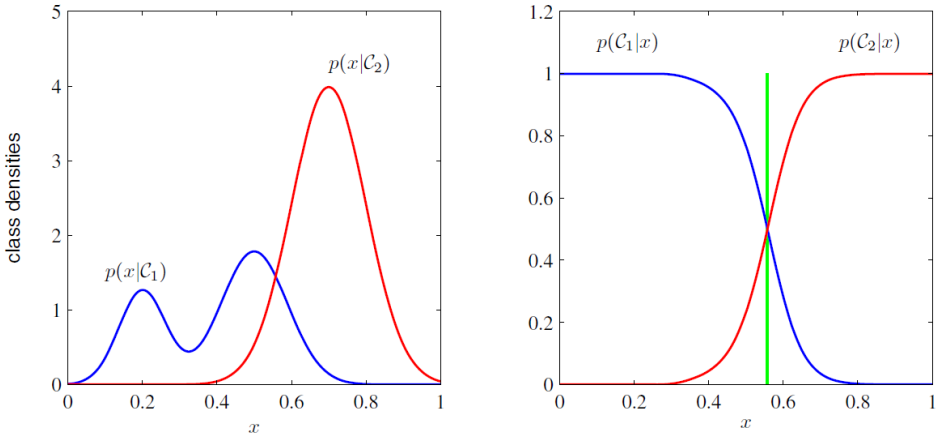
SVM aka *max-margin classifier* and is a type of *Sparse Kernel Machine (SKM)* method
(Relevance Vector Machine is another type of *SKM* method, specifically a probabilistic/Bayesian variant)

[Above formulas from sklearn help pages]

Recall: Inference and decision: three approaches for classification – SVM: ***discriminant approach*** initially motivated by ***Computational Learning Theory.***

- Generative model approach:
  - (I) Model $p(x, C_k) = p(x|C_k)p(C_k)$
  - (I) Use Bayes' theorem $p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$
  - (D) Apply optimal decision criteria

- Discriminative model approach:
  - (I) Model $p(\mathcal{C}_k|\mathbf{x})$ directly
  - (D) Apply optimal decision criteria



- ***Discriminant function approach:***
  - (D) Learn a function that maps each x to a class label directly from training data
  - Note: No posterior probabilities!

[CMB]

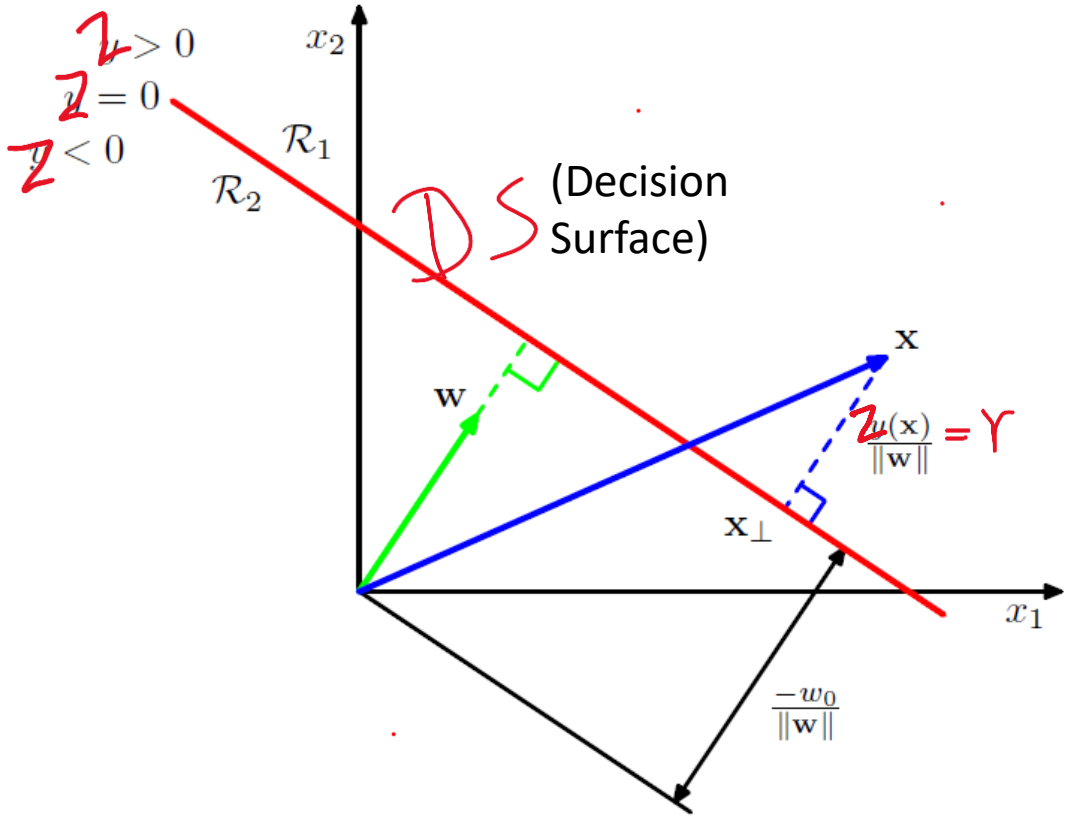
Recall: Discriminant

- Discriminant is a function that takes an input vector $x \in \mathbb{R}^d$ and assigns it to one of the K classes
  - we will assume K=2 henceforth for simplicity!

- We focus only on linear discriminants (i.e., those for which DB is a hyperplane wrt x (or $\phi(x)$)**.**
  - $z(x) = w^T x + w_0$ (or $w^T \phi(x)$)
  - DB: $z(x) = 0$ (hyperplane)
  - Prediction: $f\big(z(x)\big)$ = $\text{sign}(z(x))$ (i.e., Predict C_1 if $z(x) \geq 0$, & C_2 if $z(x) < 0$)

Recall defn. of hyperplane $\{x \in \mathbb{R}^d : w^T x = b\}$, which is a $(d-1)$-dimensional (affine) subspace of a d-dim. vector space.

Recall: Geometry of decision surfaces: *signed* distance from decision surface



DB is $w^T \mathrm{x} + w_0$ = *const.*, but *const.* can be absorbed into w_0 to get DB as $\boldsymbol{w^T x + w_0 = 0}$.
Let $z(w, x) = w^T \mathrm{x} + w_0$ with the *const.* absorbed.

Signed dist. r of x to DB:

Express $X = X_\perp + r \cdot \frac{w}{\|w\|}$

Mult. by W^T & add W_0 on both sides:

$\Rightarrow z(X) = 0 + r\|w\|$

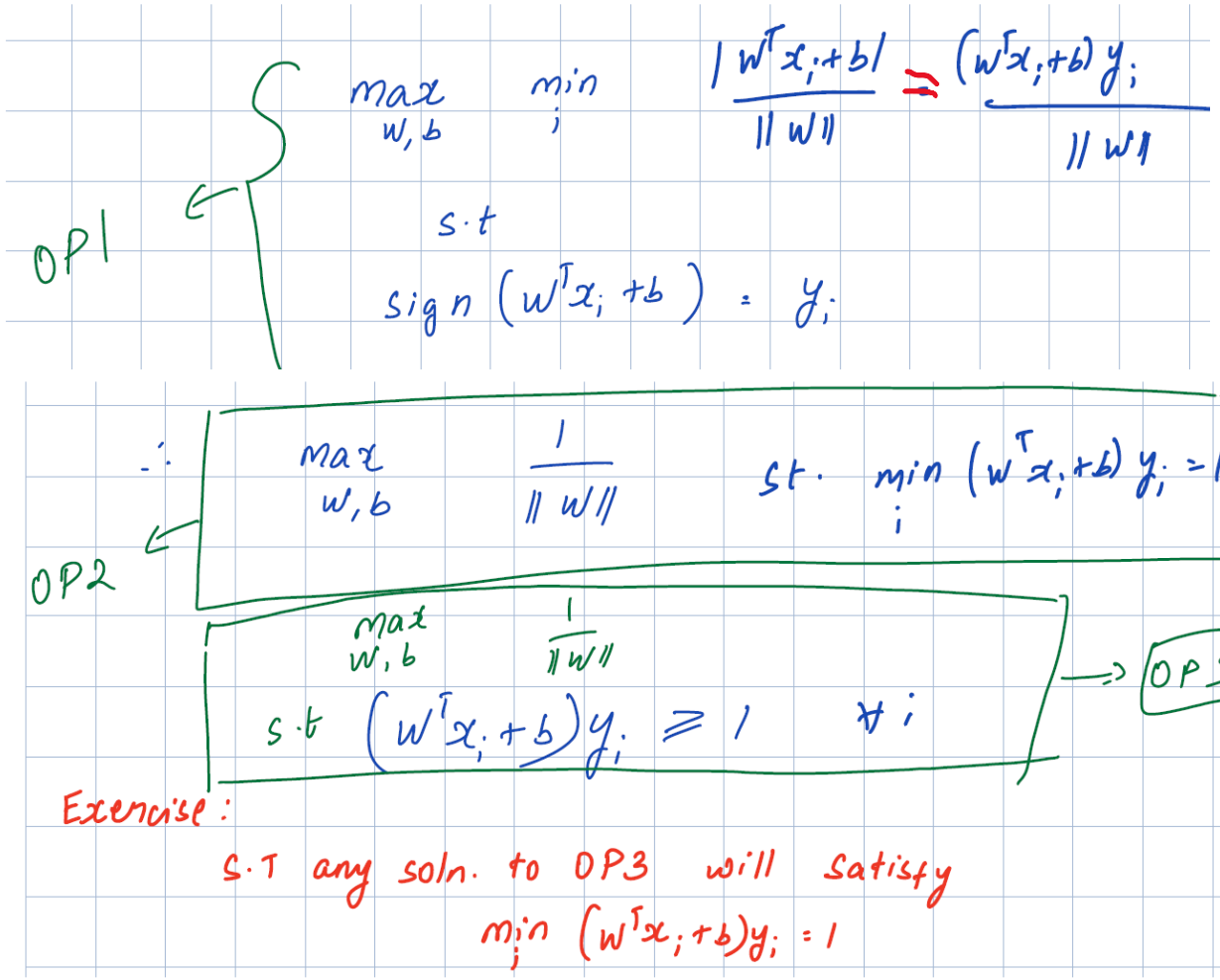
$\Rightarrow \boxed{r = \frac{z(X)}{\|w\|}}$

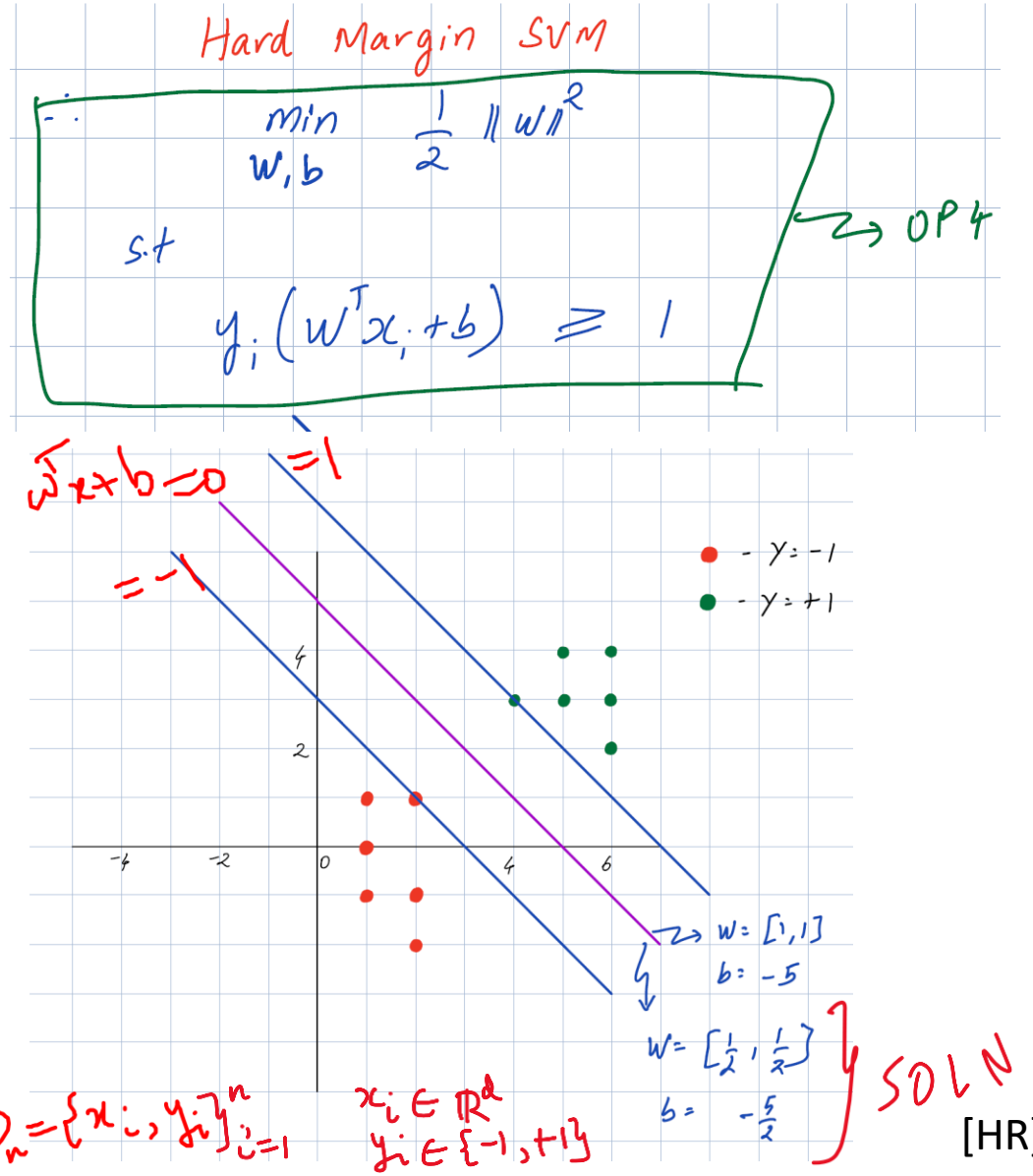
[CMB]

Outline for Module M8

- M8. Classification (Support Vector Machines)
  - M8.0 Introduction/Motivation
  - **M8.1 SVM Problem Statement**
    - **(Hard/Soft-Margin SVM Problems)**
  - M8.2 SVM Solution
  - M8.3 SVM Interpretations
  - M8.4 Concluding thoughts

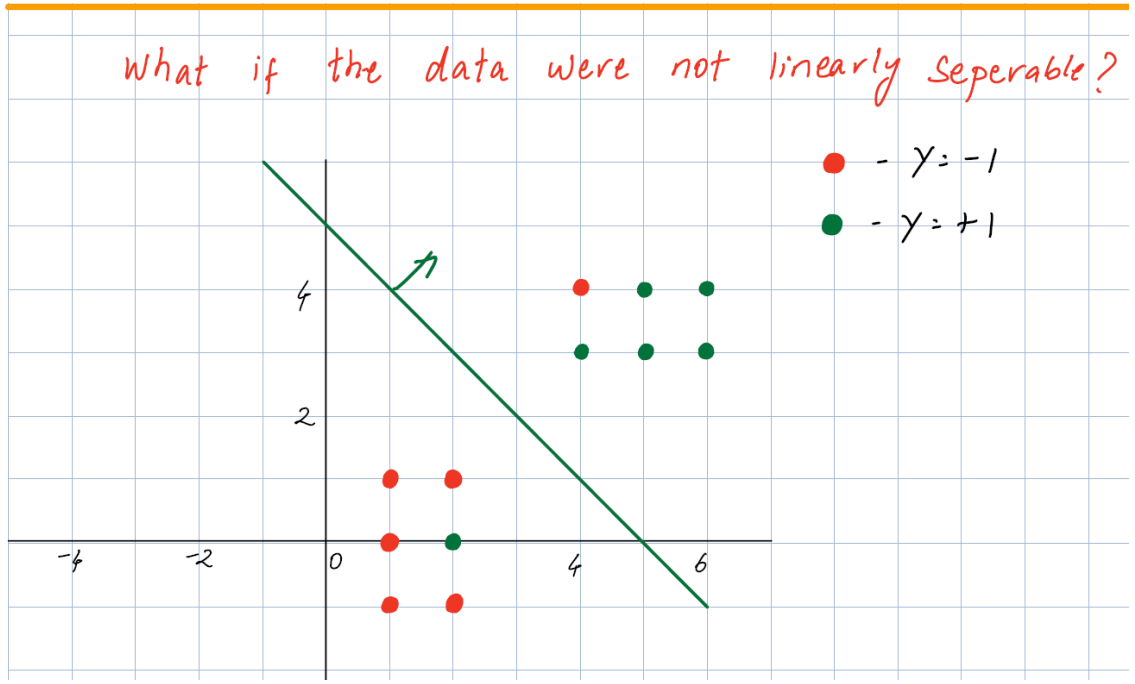
SVM hard-margin problem





[HR]

SVM soft-margin problem



(C-SVM) Soft Margin SVM

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

s.t

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

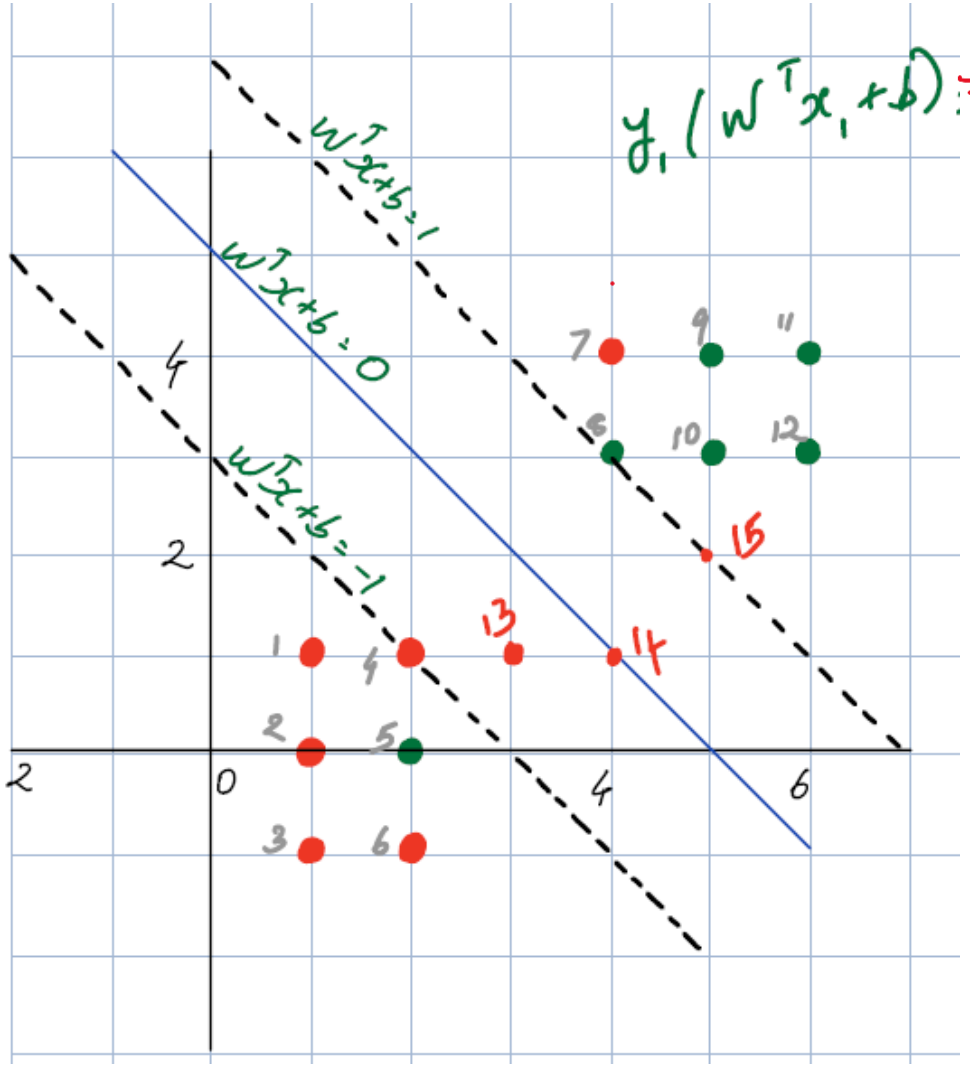
$$\xi_i \geq 0$$

In soft-margin SVM "any" w, b is feasible

Or $\forall w \in \mathbb{R}^d$, $b \in \mathbb{R}$, there exists $\xi \in \mathbb{R}_+^n$ s.t (w, b, ξ) is feasible.

[HR]

Example: what is ϵ_i for a given w, b?



$y_1(w^T x_1 + b) = \left(\begin{bmatrix} 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{5}{2}\right)(-1) = 1.5$

$\Rightarrow \xi_1 = 0$

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

$\therefore w = [\frac{1}{2}, \frac{1}{2}]$; $\|w\| = \frac{1}{\sqrt{2}}$

$b = -\frac{5}{2}$

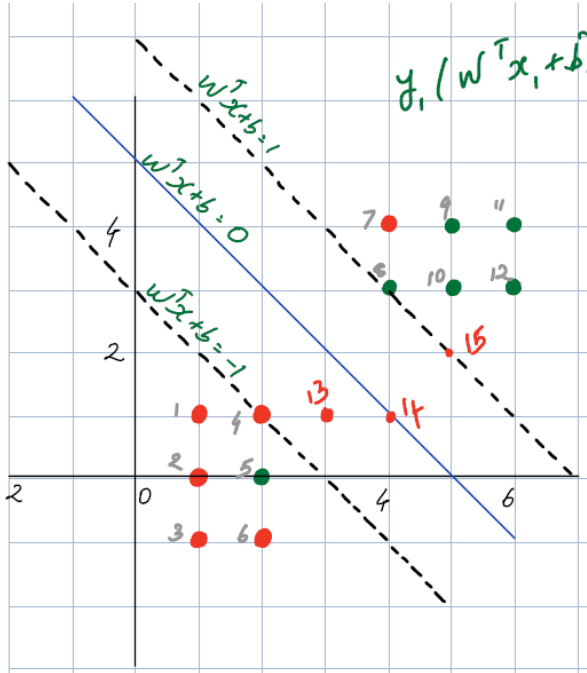
[HR]

Example: answer

(unnorm. signed dist)

$y_i(w^T x_i + b)$ $\quad \xi_i = 1 - \text{prev. col.}$

green x_5	-1.5	2.5
red x_3	$+0.5$	0.5
red x_4	$+0$	1
red x_5	-1	2



$y_i(w^T x_i + b) \geq 1 - \xi_i$, Let $C = 1$

Evaluate the following w, b:

(i) $w = [\frac{1}{2}, \frac{1}{2}]$, $b = -\frac{5}{2}$

$y_5 = +1$

$x_5 = (2, 0)$

$w^T x_5 + b = 1 - \frac{5}{2}$

$= -1.5$

$\xi_5 = 2.5$

$w^T x_5 + b \geq 1 - \xi_5$

If $w = [\frac{1}{2}, \frac{1}{2}]$, $b = -\frac{5}{2}$

$\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi_6 = 0$; $\quad \xi_5 = 5/2$

$\xi_8 = \xi_9 = \xi_{10} = \xi_{11} = \xi_{12} = 0$; $\quad \xi_7 = 5/2$

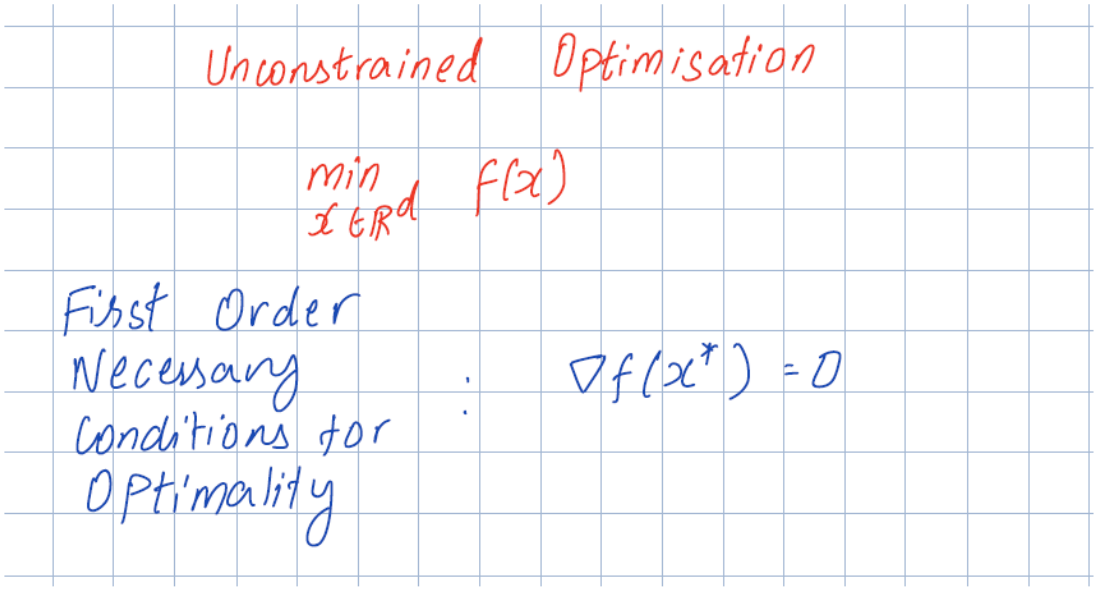
$\frac{1}{2}\|w\|^2 + C\sum \xi_i = \frac{1}{4} + 5 = \frac{21}{4}$

[HR]

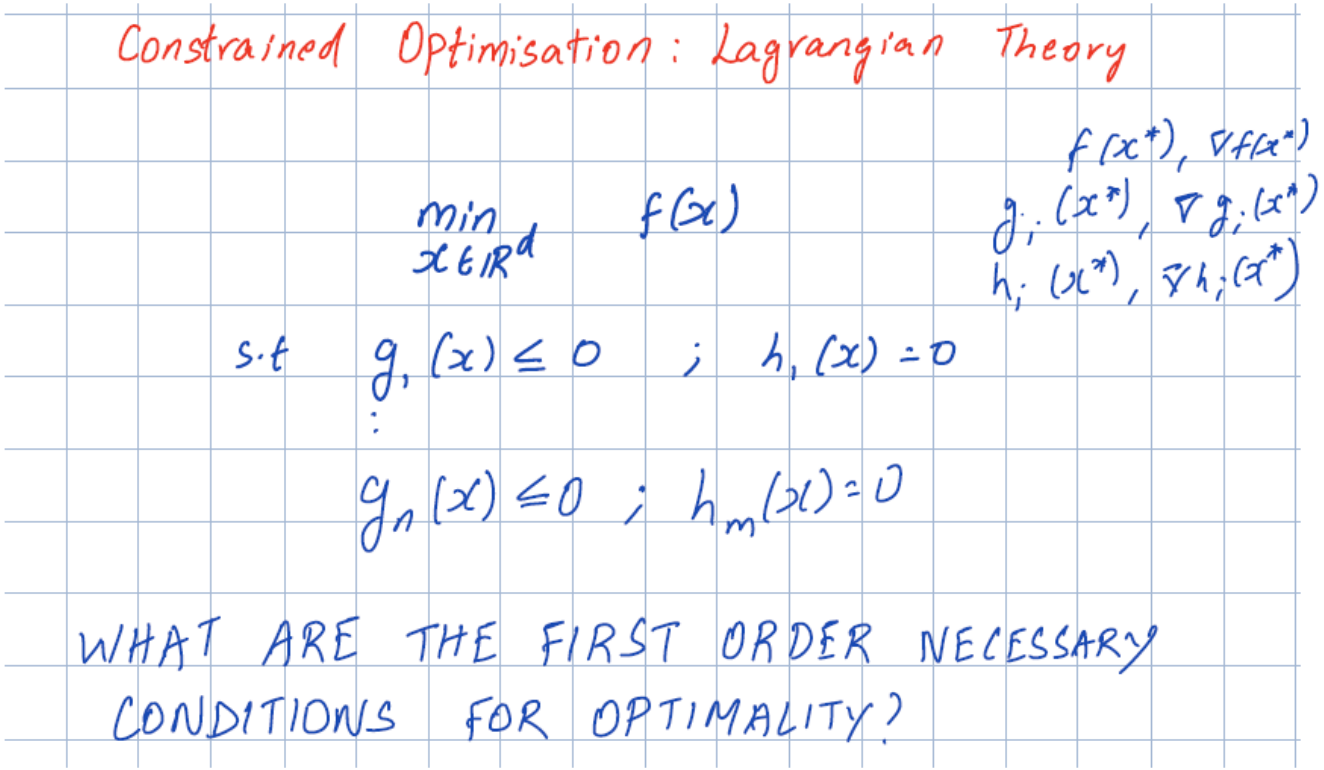
Outline for Module M8

- M8. Classification (Support Vector Machines)
  - M8.0 Introduction/Motivation
  - M8.1 SVM Problem Statement
  - **M8.2 SVM Solution**
    - **(Background: Constrained optimization - KKT & Primal-Dual)**
    - (SVM Dual Problem & Optimization algo. sketch)
  - M8.3 SVM Interpretations
  - M8.4 Concluding thoughts

From unconstrained to constrained opt. - FONC



FONC for x^* to be a local optima!



FONC for x^* to be a local feasible (constrained) optima?

[HR]

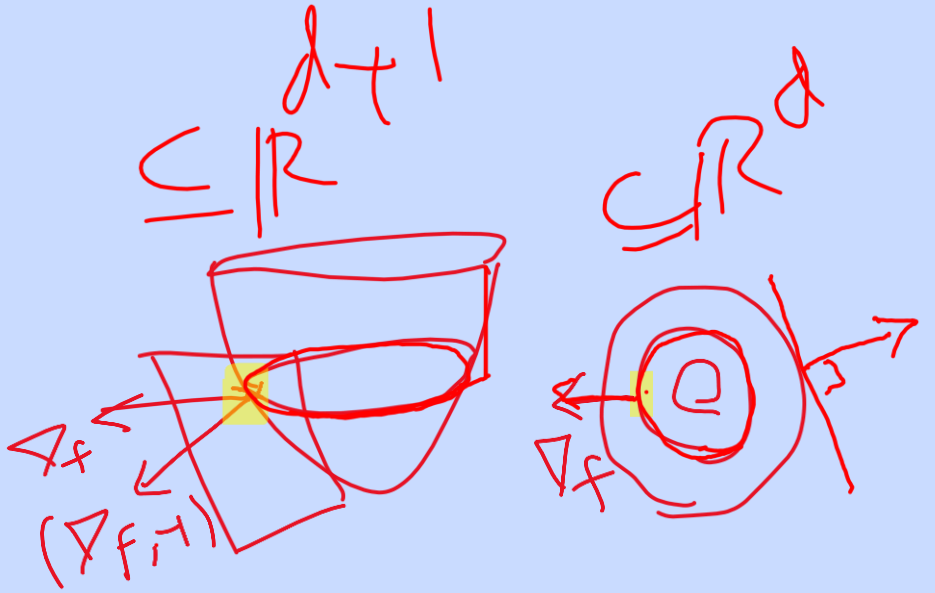
Recall: Linear approximation using gradient vector

Linear approximation around x $\quad L_x(f)(y)$

$$f(y) \approx f(x) + \nabla f(x)^T (y-x) \quad \text{— (1)}$$

The "graph" of linear approximation of f around x, is a "tangent plane" to the "graph" of f

[HR]



Facts:

At x_0: $\nabla f \perp$ level set $\{x : f(x) = f(x_0)\}$ — (2)

At $(x_0, f(x_0))$: $(\nabla f, -1) \perp$ tangent plane

Recall:

Spl case 1: One equality constraint:

$\min f(x)$

s.t

$h(x) = 0$

FONC for x^* to be a local minima:

(i) $h(x^*) = 0$

(ii) $\nabla f(x^*)$ is Parallel to $\nabla h(x^*)$

(Equivalently $\nabla f(x^*) + \mu \nabla h(x^*) = 0$)

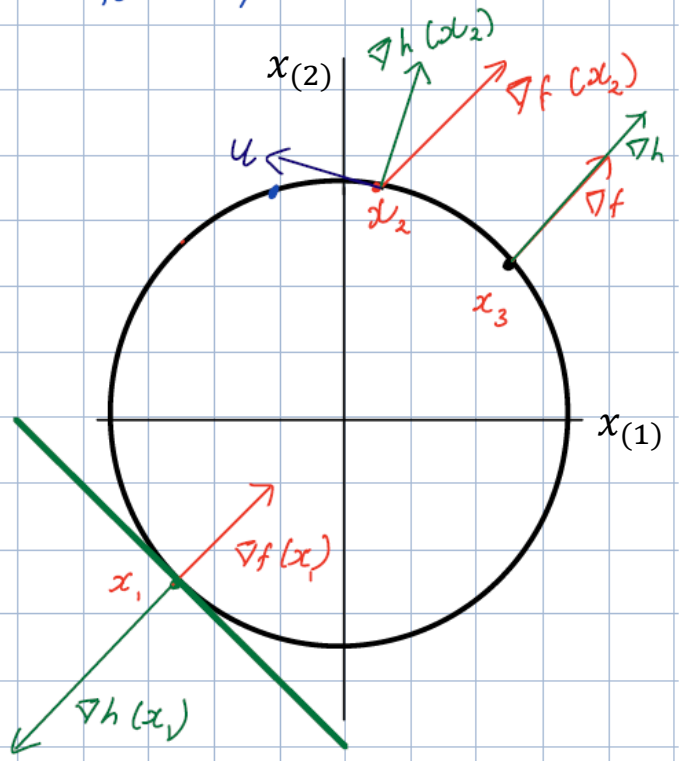
For some $\mu \in \mathbb{R}$

e.g

min $x_{(1)} + x_{(2)}$

s.t.

$x_{(1)}^2 + x_{(2)}^2 - 1 = 0$



[HR]

Proof sketch: Clearly x^* must satisfy $h(x^*) = 0$

F = Set of feasible directions $= \{u : h(x^* + \alpha u) = 0 \text{ for small } \alpha \geq 0\}$

$= \{u : h(x^* + \alpha u) = h(x^*) \text{ for small } \alpha > 0\}$

$= \{u : \nabla h(x^*)^T u = 0\}$ ⇐ Now, $h(x^* + \alpha u) - h(x^*) \approx \nabla h(x^*)^T (\alpha u)$

D = Set of Descent directions $= \{u : f(x^* + \alpha u) < f(x^*) \text{ for small } \alpha > 0\}$

$= \{u : \nabla f(x^*)^T u < 0\}$

When will $F \cap D$ be empty?

Only when $\nabla f(x^*) = \mu \nabla h(x^*)$ for some $\mu \in \mathbb{R}$.

→ This requires a Proof. (regularity conditions for KKT to hold)

[HR]

Special case 2 : One Inequality constraint.

$\min \; f(x)$

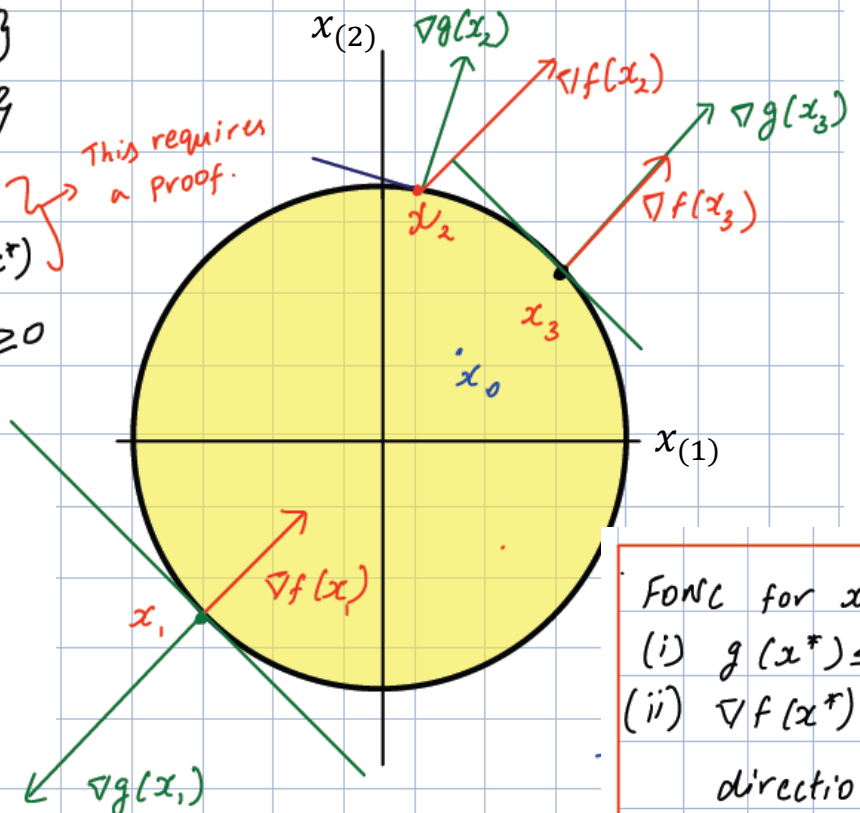
$s.t \quad g(x) \leq 0$

$D = \{u : \nabla f(x^*)^T u < 0\}$

$F = \{u : \nabla g(x^*)^T u \leq 0\}$

$F \cap D$ is empty only if $\nabla f(x^*) = -\lambda \nabla g(x^*)$ for some $\lambda \geq 0$

This requires a proof.



e.g.

$\min \; x_{(1)} + x_{(2)}$

s.t.

$x_{(1)}^2 + x_{(2)}^2 - 1 \leq 0$

FONC for x^* to be a local min

(i) $g(x^*) \leq 0$

(ii) $\nabla f(x^*)$ is parallel to $\nabla g(x^*)$ and pointing in opposite direction. (Equivalently $\nabla f(x^*) + \lambda \nabla g(x^*) = 0$, for some $\lambda \geq 0$)

(iii) If $g(x^*) < 0$ then $\nabla f(x^*) = 0$. (constraint g is inactive) $\lambda = 0$

(Equivalently $\lambda g(x^*) = 0$)

[HR]

KKT conditions (FONC) – General Case

General case : Multiple equality & inequality constraints

$$\min \; f(x)$$
$$\text{s.t}$$
$$g_1(x) \le 0 \;;\quad h_1(x) = 0$$
$$g_2(x) \le 0 \;;\quad h_2(x) = 0$$
$$\vdots \qquad\qquad \vdots$$
$$g_m(x) \le 0 \;;\quad h_n(x) = 0$$

Define Lagrangian $\qquad \lambda_i \ge 0$

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{n} \mu_i h_i(x)$$

Equivalently : KKT conditions.

FONC for x^* to be a local minima.

(i) $g_i(x^*) \le 0 \quad \forall\, i \in [m]$
$h_i(x^*) = 0 \quad \forall\, i \in [n]$ (Feasibility)

(ii) $\exists\, \lambda \in \mathbb{R}_+^m$, $\exists\, \mu \in \mathbb{R}^n$ such that

(a) $\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{i=1}^{n} \mu_i \nabla h_i(x^*) = 0$ (Stationarity)

(b) $\lambda_i = 0$ if $g_i(x^*) < 0$, or equivalently
$\lambda_i \cdot g_i(x^*) = 0, \; \forall\, i \in [m]$ → (Complementary Slack)

(or equivalently) $\lambda_i = 0 \quad \forall\, i \notin A(x^*)$

$A(x^*) = \{ i \in [m] : g_i(x^*) = 0 \}$ set of active ineq.

[HR]

KKT Conditions - Example

$$\min_{x \in \mathbb{R}} \quad x^2 + ax$$

$$\text{s.t} \quad 0 \leq x \leq 1$$

(For unconstrained problem minimiser is $x = \frac{-a}{2}$)

$$\mathcal{L}(x, \lambda_1, \lambda_2) = x^2 + ax + \lambda_1(-x) + \lambda_2(x-1)$$

$$\nabla_x \mathcal{L}(x, \lambda_1, \lambda_2) = 0 \Rightarrow 2x + a - \lambda_1 + \lambda_2 = 0 \quad \text{(Stationarity)}$$

$$x = \frac{\lambda_1 - \lambda_2 - a}{2}$$

[HR]

KKT Conditions – Example (checking FONC)

Case 1 Check if "$x = 0$" satisfies FONC

$\Rightarrow \lambda_1 \geq 0 \;,\; \lambda_2 = 0$ (complementary slack)

$x = \dfrac{\lambda_1 - a}{2} = 0$

$\Rightarrow \lambda_1 = a$

$\therefore$ if $a \geq 0$, $\exists\ \lambda \geq 0$ s.t (x, λ) satisfies FONC.

$x = 0$

If $a \geq 0$

$\hookrightarrow x = \frac{1}{2}$ (statn. cdtn not true)

Case 2 Check if "$x = 1$" satisfies FONC

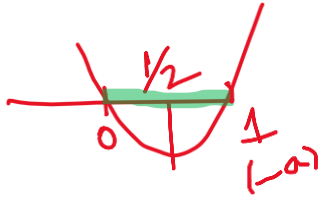
$\lambda_1 = 0 \;,\; \lambda_2 \geq 0$ (complementary slack)

$x = \dfrac{-\lambda_2 - a}{2} = 1 \Rightarrow \lambda_2 = -2 - a$

$\therefore$ if $a \leq -2$, $\exists\ \lambda \geq 0$ s.t (x, λ) satisfies FONC.

$x = 1$

Case 3: $x = \frac{1}{2}$ satisfies FONC if $a = -1$. Check!



[HR]

Optional: Bishop-AppxE is also a good read to get similar intuition about Lagrange multipliers and FONC.

$$L(\mathbf{x}, \lambda) \equiv f(\mathbf{x}) + \lambda g(\mathbf{x}).$$

FONC

$\exists \lambda \in \mathbb{R}$ s.t.

$\nabla_x f(x^*) + \lambda \nabla_x g(x^*) = 0$

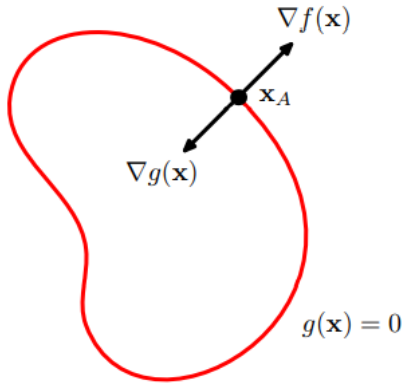
$g(x^*) = 0$

$\max_{x,\lambda} L = 0$

$\max f(x)$

s.t.

$g(x) = 0$



$\max f(x)$

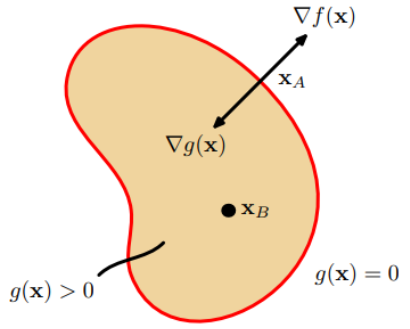
s.t.

$g(x) \geq 0$

$\exists \lambda \geq 0$ s.t. $\nabla_x f(x^*) + \lambda \nabla_x g(x^*) = 0$

$g(x^*) \geq 0$

$\lambda g(x^*) = 0$ (active & inactive constraint)



[CMB]

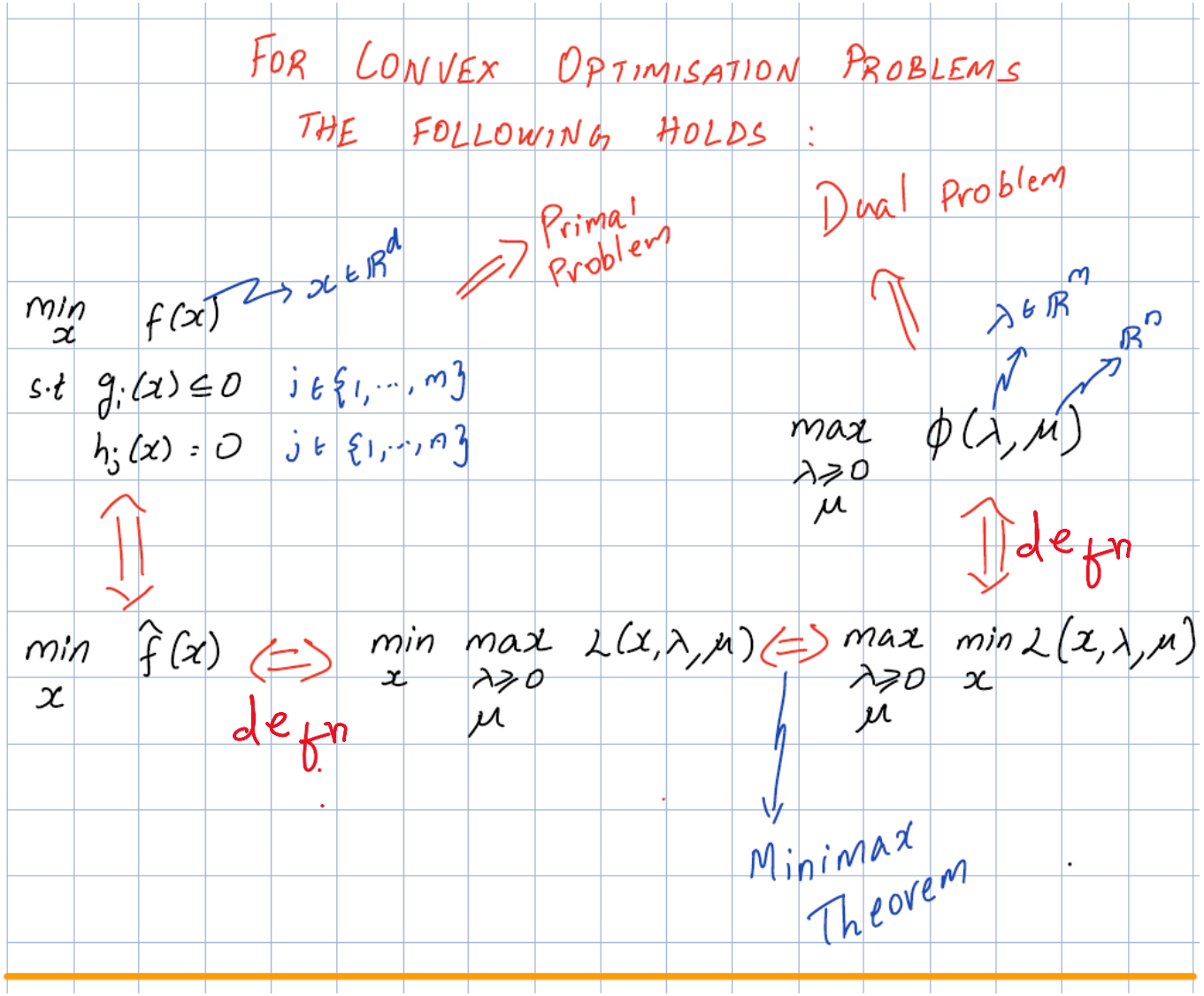
Exercises: Other constrained optimization problems!

- Prove that the KL-divergence $KL(p \,||\, q)$ is minimized when $q = p$.
- Prove that entropy $H(p)$ is maximized when p is uniform.
- What is the distance of a point $u \in \mathbb{R}^d$ to the closest point v in a hyperplane given by $\{x \in \mathbb{R}^d : w^T x + b = 0\}$?

[HR]

Having established KKT cdnts., can we actually **(constructively) find** the KKT multipliers $\lambda,^* \mu^*$ that satisfy the KKT cdtns.?

Primal-Dual Relation (via the Minimax Theorem)



Define Lagrangian $\lambda_i \ge 0$

$$\mathcal{L}(x,\lambda,\mu) = f(x) + \sum_{i=1}^{} \lambda_i g_i(x) + \sum_{i=1}^{m} \mu_i h_i(x)$$

[HR]

Function $\hat{f}(.)$

- Desired function:

$$\therefore \quad \hat{f}(x) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ \infty & \text{if } x \text{ is not feasible.} \end{cases}$$

- First attempt:

$$\hat{f}(x) = f(x) + \sum_i \infty \cdot \mathbb{1}_{g_i(x)>0} + \sum_j \infty \cdot \mathbb{1}_{h_j(x)\neq 0}$$

- Second attempt:

Define function $\hat{f}: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ as

$$\hat{f}(x) = \max_{\substack{\lambda \geq 0 \\ \mu}} \mathcal{L}(x, \lambda, \mu)$$

Define Lagrangian $\lambda_i \geq 0$

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{i=1}^{m} \mu_i h_i(x)$$

[HR]

Weak Duality :

$$\min_{x} \max_{\substack{\lambda \geq 0 \\ \mu}} \mathcal{L}(x, \lambda, \mu) \geq \max_{\substack{\lambda \geq 0 \\ \mu}} \min_{x} \mathcal{L}(x, \lambda, \mu)$$

If f, g are convex, h is Linear then

$$\min_{x} \max_{\substack{\lambda \geq 0 \\ \mu}} \mathcal{L}(x, \lambda, \mu) = \max_{\substack{\lambda \geq 0 \\ \mu}} \min_{x} \mathcal{L}(x, \lambda, \mu)$$

MINIMAX THEOREM !!

Key Result in Game Theory !!

Define Lagrangian $\lambda_i \geq 0$

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1} \lambda_i g_i(x) + \sum_{i=1}^{m} \mu_i h_i(x)$$

[HR]

How to get one solution from the other?

PRIMAL

$\min_x \; f(x)$

s.t $g_i(x) \le 0 \; ; \; i \in [m]$ $\iff$

$h_j(x) = 0 \; ; \; j \in [n]$

Let x^* be the soln to the Primal

DUAL

$\max_{\substack{\lambda \ge 0 \\ \mu}} \; \phi(\lambda, \mu)$

Let λ^*, μ^* be the soln. to the dual.

Fact : (Follows from Proof of min-max Theorem)

x^* satisfies KKT for Primal with the multipliers in the KKT condition given by λ^* & μ^* !!

Fact

Define $\phi(\lambda, \mu) = \min_x \; \mathcal{L}(x, \lambda, \mu)$

[HR]

How to get one solution from the other?

More precisely let x^* be a primal optimal soln.
let λ^*, μ^* be a dual optimal soln.

The above statement is equivalent to the below statement. (for convex opt.)

i) $g_i(x^*) \leq 0 \quad ; \quad \lambda_i^* \geq 0 \qquad \forall\ i \in [m]$
$h_j(x^*) = 0 \qquad \forall\ j \in [n]$

(ii) $\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{n} \mu_j^* \nabla h_j(x^*) = 0$

(iii) $\lambda_i^* g_i(x^*) = 0 \qquad \forall\ i \in [m]$

Fact ✡

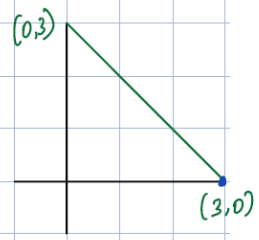
← solve these eqns.

[HR]

Duality Example: A linear program

$$\min_{x \in \mathbb{R}^2} \; x_1 + 2x_2 \quad \text{s.t.} \quad x_1 + x_2 = 3 \;(=), \quad x_1 \geq 0, \quad x_2 \geq 0$$

$$\min_{x} \; x_1 + 2x_2 \quad \text{s.t.} \quad x_1 + x_2 - 3 = 0, \quad -x_1 \leq 0, \quad -x_2 \leq 0$$



(standard form Primal)

$$\mathcal{L}(x, \lambda_1, \lambda_2, \mu) = x_1 + 2x_2 + \lambda_1(-x_1) + \lambda_2(-x_2) + \mu(x_1 + x_2 - 3)$$

$$= x_1(1 - \lambda_1 + \mu) + x_2(2 - \lambda_2 + \mu) - 3\mu$$

$$\phi(\lambda, \mu) = \min_{x \in \mathbb{R}^2} \mathcal{L}(x, \lambda, \mu)$$

$$= \begin{cases} -\infty & \text{if } 1 - \lambda_1 + \mu \neq 0 \\ -\infty & \text{if } 2 - \lambda_2 + \mu \neq 0 \\ -3\mu & \text{if } 1 - \lambda_1 + \mu = 0 \text{ and } 2 - \lambda_2 + \mu = 0 \end{cases}$$

[HR]

Duality Example (contd.)

$\therefore$ $\max_{\substack{\lambda \geq 0 \\ \mu}} \phi(\lambda, \mu)$

$\Updownarrow$

$\max \quad -3\mu$

$\lambda_1 \geq 0$

$\lambda_2 \geq 0$

$1 - \lambda_1 + \mu = 0$

$2 - \lambda_2 + \mu = 0$

$\Updownarrow$

$\max \quad -3\mu$

$\lambda_1 : 1 + \mu \geq 0$

$\lambda_2 : 2 + \mu \geq 0$

$\Updownarrow$

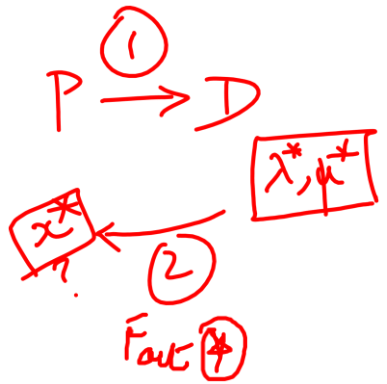
$\max \quad -3\mu$

$\mu \geq -1 \; ; \; \mu \geq -2$

$\Updownarrow$

$\max \quad -3\mu$

$\mu \geq -1$



[HR]

Duality Example (contd.)

$$\min_{x_1, x_2} \; x_1 + 2x_2 \qquad \max_{\mu, \lambda_1, \lambda_2} \; -3\mu$$

(Primal) s.t. $x_1 + x_2 = 3$ $\iff$ (Dual) s.t. $\mu \geq -1$

$-x_1 \leq 0$ $\quad \lambda_1 = 1 + \mu$

$-x_2 \leq 0$ $\quad \lambda_2 = 2 + \mu$

The Dual solution is clearly

$\mu^* = -1$

$\lambda_1^* = 0$

$\lambda_2^* = 1$

Now try to find a feasible x^*, satisfying the Stationarity & complementary slack conditions with these as the Lagrange multipliers:

(i) Feasibility $\Rightarrow$ $x_1^* + x_2^* = 3$ $\rightarrow$ (1)

$-x_1^* \leq 0$

$-x_2^* \leq 0$

(ii) Stationarity :

$\nabla f(x^*) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$; $\nabla g_1(x^*) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$; $\nabla g_2(x^*) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$; $\nabla h_1(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$\nabla f(x^*) + \lambda_1^* \nabla g_1(x^*) + \lambda_2^* \nabla g_2(x^*) + \mu_1^* \nabla h_1(x^*) = 0$$

$$\Rightarrow \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ -1 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

[This holds for all x, and hence useless for finding x^*.

(iii) Complementary slack :

$\lambda_1^* > 0 \Rightarrow g_1(x^*) = 0$

$\lambda_2^* > 0 \Rightarrow g_2(x^*) = 0$

only $\lambda_2^* > 0$ $\therefore$ $g_2(x^*) = 0$

$\Rightarrow -x_2^* = 0$ $\rightarrow$ (2)

Solving (1) and (2) we get $x^* = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$. ■

[HR]

Outline for Module M8

- M8. Classification (Support Vector Machines)
  - M8.0 Introduction/Motivation
  - M8.1 SVM Problem Statement
  - **M8.2 SVM Solution**
    - (Background: Constrained optimization - KKT & Primal-Dual)
    - **(SVM Dual Problem & Optimization algo. sketch)**
  - M8.3 SVM Interpretations
  - M8.4 Concluding thoughts

SVM: From Primal → Dual

Primal

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

s.t

$$y_i(w^T x_i + b) \geq 1 - \xi_i \qquad \forall i \in [n] \qquad \alpha_i$$

$$\xi_i \geq 0 \qquad \forall i \in [n] \qquad \beta_i$$

$g_i(\cdot) \leq 0$

Lagrangian

$$\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-\xi_i - y_i(w^T x_i + b)) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}^n$

$$= \frac{1}{2}\|w\|^2 + b\left(\sum_{i=1}^{n} -\alpha_i y_i\right) + \sum_{i=1}^{n}\xi_i(C - \alpha_i - \beta_i) + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i (w^T x_i)$$

Dual function

$$\phi(\alpha,\beta) = \min_{w,b,\xi} \mathcal{L}(w,b,\xi,\alpha,\beta)$$

$$\sum_{i=1}^{n} -\alpha_i y_i = 0 \rightarrow ① \qquad \left(\text{If not } \phi(\alpha,\beta) = -\infty\right)$$

$$C - \alpha_i - \beta_i = 0 \rightarrow ②$$

If Equations ① and ② are satisfied then:

$$\phi(\alpha,\beta) = \min_{w} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i w^T x_i + \sum_{i=1}^{n}\alpha_i$$

[HR]

SVM: From Primal → Dual

$$\phi(\alpha,\beta) = \min_{w} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i w^T x_i + \sum_{i=1}^{n} \alpha_i$$

Let $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$, $\alpha = [\alpha_1, \cdots, \alpha_n]^T$

Let $Y = \begin{bmatrix} y_1 & & 0 \\ & y_2 & \\ 0 & & \ddots \\ & & & y_n \end{bmatrix} \in \mathbb{R}^{n \times n}$ be a diagonal matrix.

$$\phi(\alpha,\beta) = \min_{w} \frac{1}{2}\|w\|^2 - \alpha^T Y X w + 1^T \alpha$$

Taking gradient w.r.t w and set to 0

$$\Rightarrow w - X^T Y \alpha = 0$$

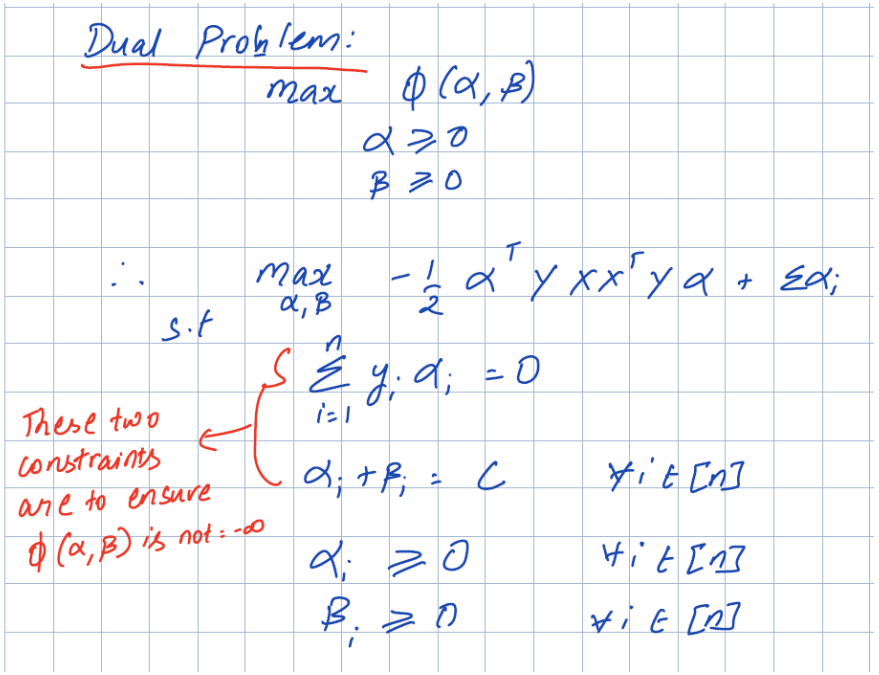
$$\therefore \boxed{w = X^T Y \alpha}$$

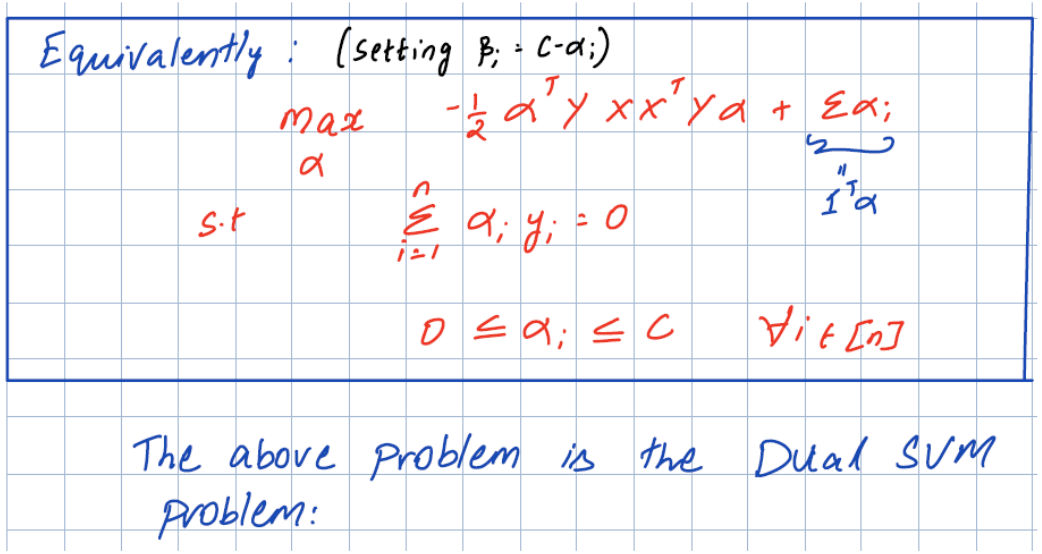
$$\therefore \phi(\alpha,\beta) = \frac{1}{2}\|X^T Y \alpha\|^2 - \alpha^T Y X X^T Y \alpha + \sum \alpha_i$$

$$= -\frac{1}{2} \alpha^T Y X X^T Y \alpha + \sum \alpha_i$$

[HR]

SVM Dual Problem





[HR]

Stop & Think: What have we achieved so far?

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

$$\Leftrightarrow$$

$$\max_{\alpha \in \mathbb{R}^n} \ -\frac{1}{2}\alpha^T Y X X^T Y \alpha + I^T\alpha \quad \text{s.t} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C$$

where $Y = \begin{bmatrix} y_1 & & 0 \\ & y_2 & \\ 0 & & y_n \end{bmatrix} \in \mathbb{R}^{n \times n}$; $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$

(To get the same for hard margin, just set $C = \infty$: Exercise: Why is this?)

Next steps:

- This Dual problem can be maximized using SMO or PGD methods much easier than the Primal problem!
- Then convert Dual to Primal solution using Fact ✡.

[HR]

Brief aside: Kernel-ize our Dual Problem!

Let $K \in \mathbb{R}^{n \times n}$ s.t

$$K = \Phi \Phi^T$$

$\therefore$ Dual SVM $\Rightarrow$

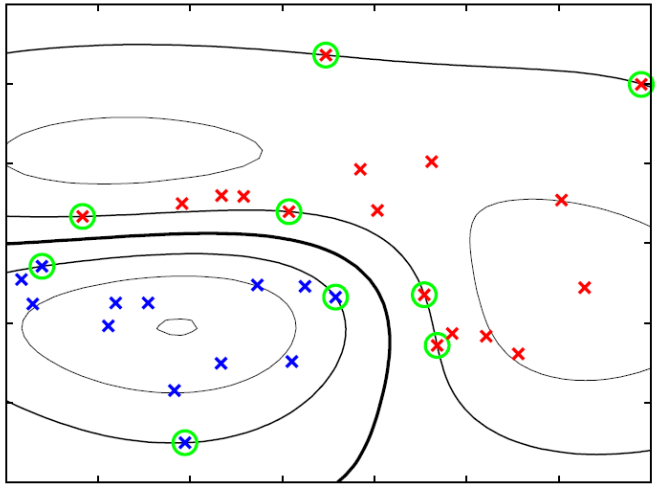
$$\max_{\alpha \in \mathbb{R}^n} \; -\frac{1}{2} \alpha^T Y K Y \alpha + I^T \alpha$$

$$\text{s.t} \quad \sum \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

Identity kernel gives back the original Dual Problem!

Recall: Why kernel-ize?



[CMB, HR]

Brief aside: Kernel SVM – Prediction for a new data point (using support vectors)

Need to go from one solution (dual: α^*) to the other (primal: $w^*, b^*(, \epsilon^*)$).

Fact ✡

We have that the primal soln

$$w^* = X^T Y \alpha^* \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

$$\therefore w^{*T} x = \alpha^{*T} Y X x$$

$$= \alpha^{*T} Y \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} x$$

$$= \alpha^{*T} Y \begin{bmatrix} x_1^T x \\ \vdots \\ x_n^T x \end{bmatrix}$$

$$= \alpha^{*T} Y \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{bmatrix}$$

(If we use feature vectors $\phi(x)$ instead)

$$\therefore w^{*T} \phi(x) = \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x)$$

Lagrangian

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i(w^T x_i + b)) + \sum_{i=1}^{n} \beta_i (-\xi_i)$$

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}^n$

[HR]

Having found w^* from α^* using KKT (stationarity); now find b^* using also KKT complementary slack

b^* can be found by the following:

If $\alpha_i^* > 0$ and $\alpha_i^* < C$, then the point x_i satisfies

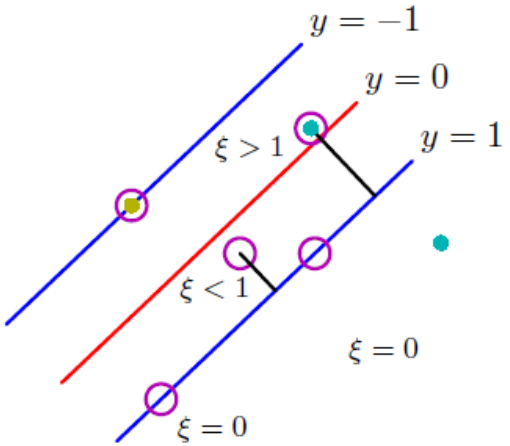
$$y_i (w^{*T} x_i + b^*) = 1$$ (Exercise: why?)

(Hint: what happens to α_i & β_i and use complementary slack)

$$\therefore \quad y_i \left(\alpha^{*T} Y \Phi x_i + b^* \right) = 1$$

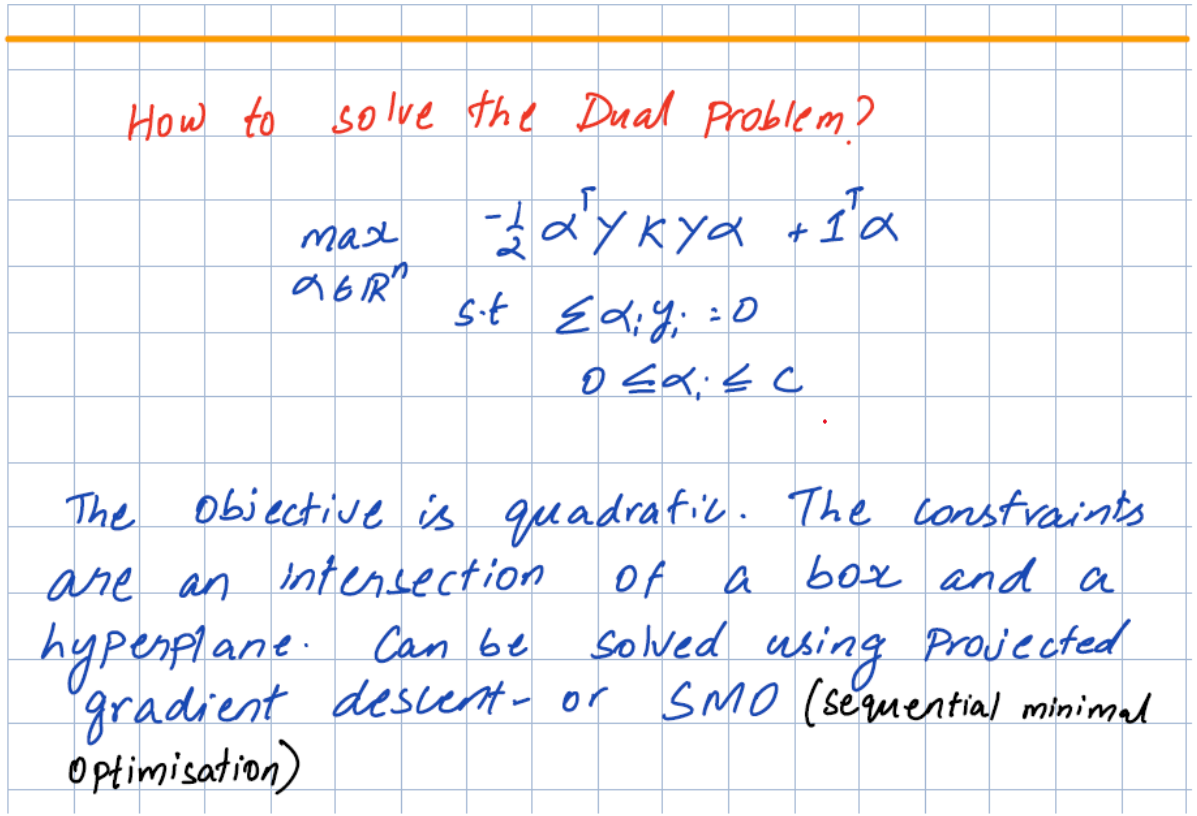
$$\therefore \quad y_i \left(\sum_{j=1}^{n} \alpha_j^* y_j K(x_j, x_i) + b^* \right) = 1$$

$$\therefore \quad b^* = y_i - \sum_{j=1}^{n} \alpha_j^* y_j K(x_j, x_i)$$



[CMB, HR]

How do we optimize the Dual Problem?

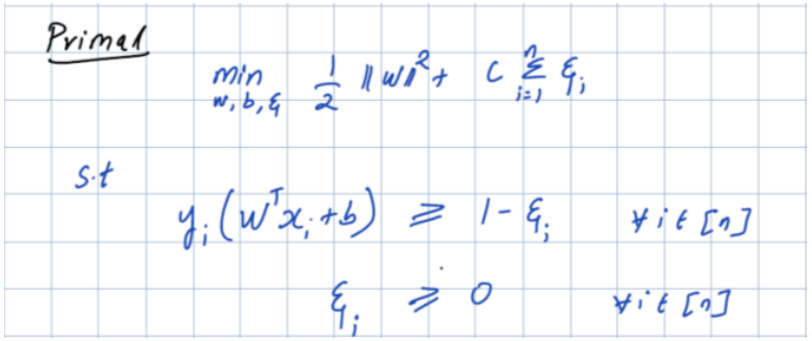


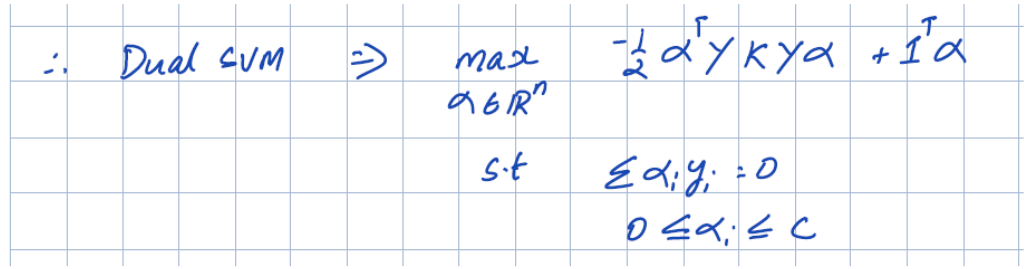
[HR]

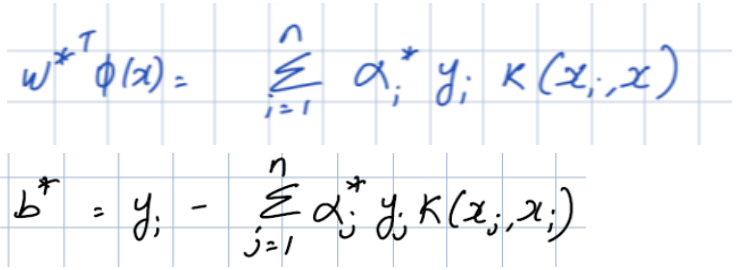
Outline for Module M8

- M8. Classification (Support Vector Machines)
  - M8.0 Introduction/Motivation
  - M8.1 SVM Problem Statement
  - M8.2 SVM Solution
  - **M8.3 SVM Interpretations**
    - **(Support vectors, Kernels, Loss function view)**
  - M8.4 Concluding thoughts

Summary so far, and support vectors/kernels







Optimize using PGD/SMO to find α^*

Support Vectors:

- Training data points for which $\alpha_i^* \neq 0$.
- Typically sparse. Why?

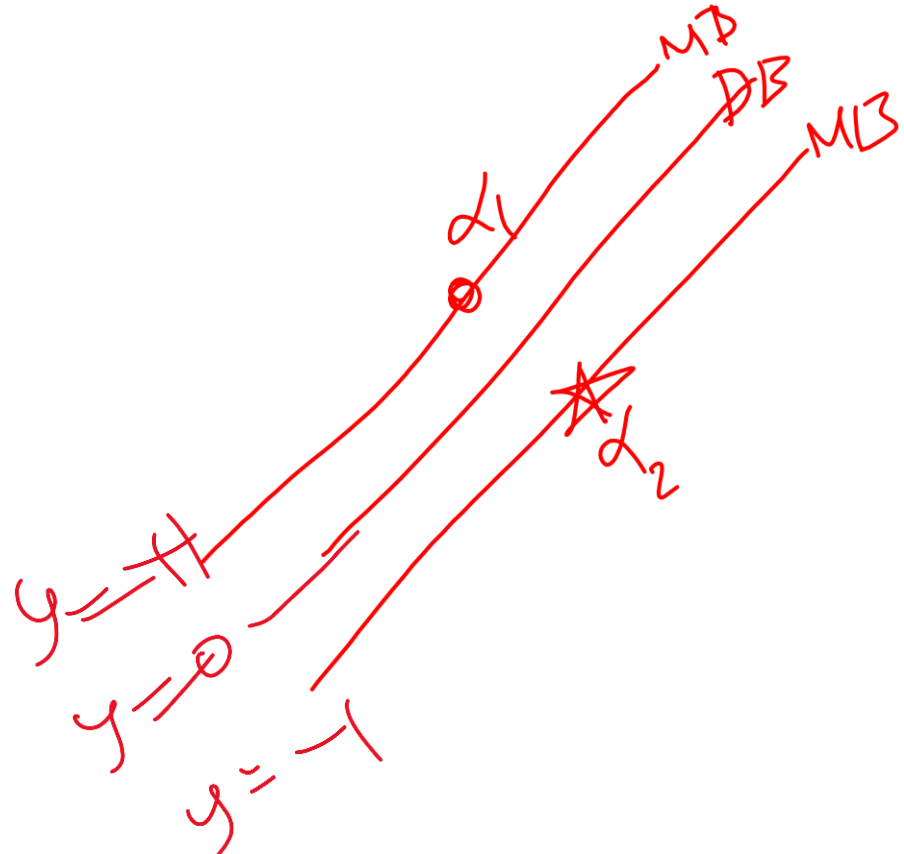
Exercise0: What are the possible values of $y_i(w^T x_i + b)$ (and hence the locn. of a training point x_i wrt DB/MBs) when:

(i) $\alpha_i^* = 0$,
(ii) $\alpha_i^* = C$,
(iii) $0 < \alpha_i^* < C$?

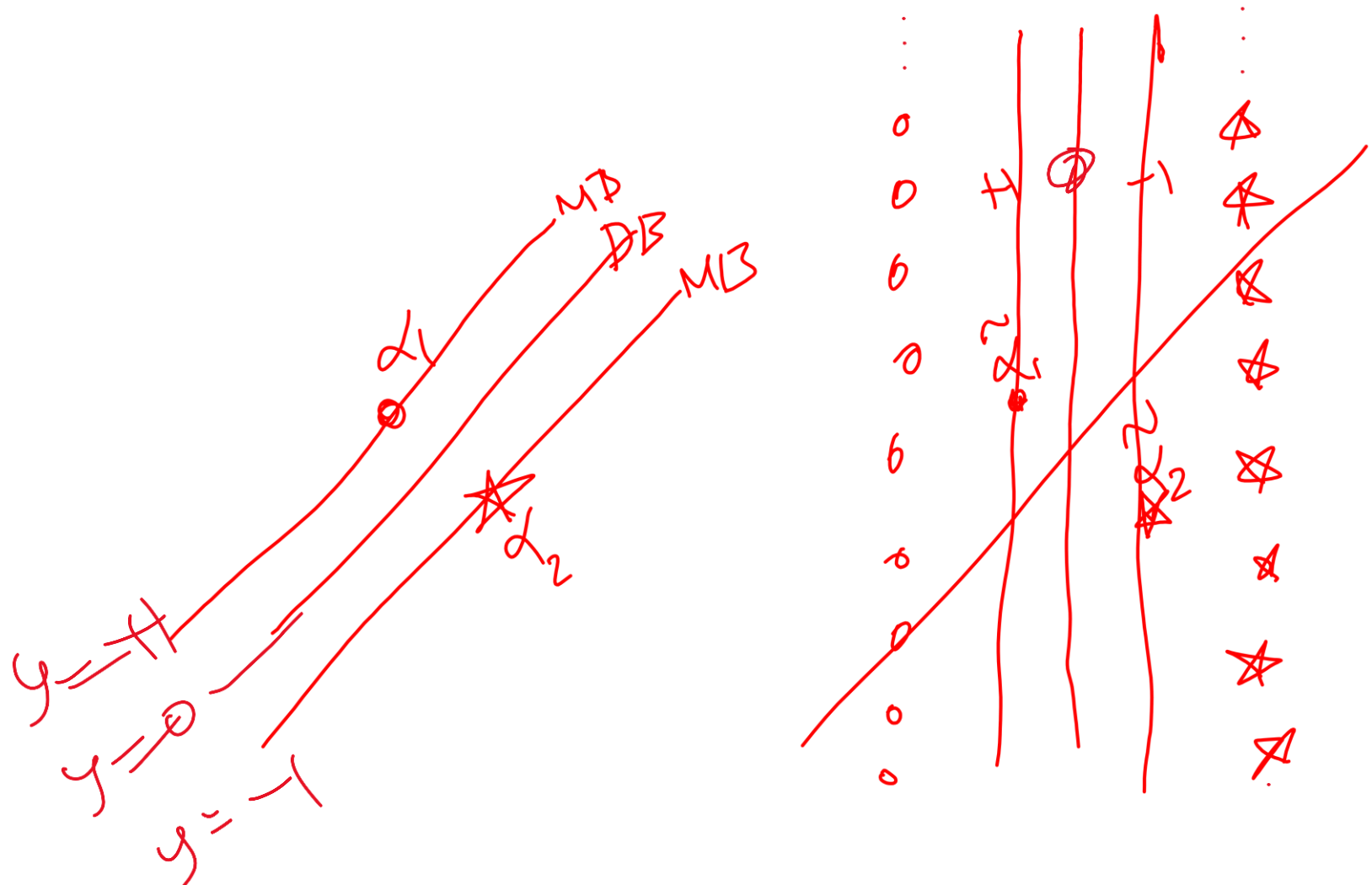
(Hint: use KKT complementary slack and $\beta_i^* = C - \alpha_i^*$))

[HR]

Example: the two SV case

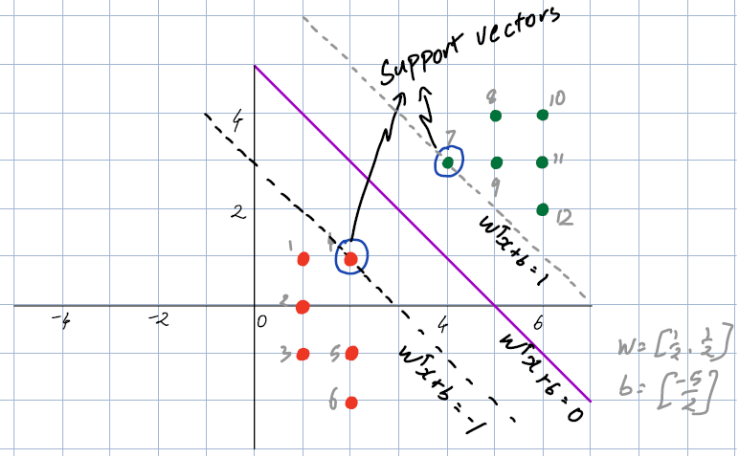


Example: the two SV case



Worked-out example:

At Optimality α is usually "Sparse"



The solution of hard Margin SVM for above problem with Linear Kernel :

$$\max_{\alpha} -\frac{1}{2} \alpha^T y x x^T y \alpha + I^T \alpha$$

s.t $\sum \alpha_i y_i = 0$ $\quad K(u,v) = u^T v$

$\alpha_i \geq 0$

is $\alpha_4^* = \alpha_7^* = \frac{1}{4}$, other $\alpha_i^* = 0$

(Exercise : Check Optimality of above point using KKT with just points 1, 4, 7, 6)

Calculate w^* and b^*:

$$\therefore \quad w^* = \frac{1}{4}(-1)\begin{bmatrix}2\\1\end{bmatrix} + \frac{1}{4}(1)\begin{bmatrix}4\\3\end{bmatrix} = \begin{bmatrix}1/2\\1/2\end{bmatrix}$$

Take $i=4$ for calculating b^* (as $0 < \alpha_4 < \infty$)

$$y_4 (w^{*T} x_4 + b^*) = 1$$

$$-1 \left(\begin{bmatrix}1/2 & 1/2\end{bmatrix} \begin{bmatrix}2\\1\end{bmatrix} + b^* \right) = 1$$

$$1.5 + b^* = -1$$

$$\therefore \quad b^* = -5/2$$

[HR]

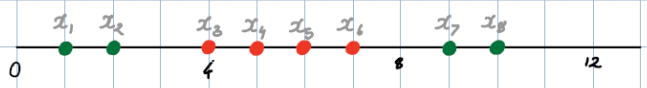
Exercise1:

Exercise :

$x^T = [1, 2, 4, 5, 6, 7, 9, 10]$

$y^T = [1, 1, -1, -1, -1, -1, 1, 1]$

(Use software to get intuition)



(Hard Margin)

(a) Solve Kernel SVM with (give optimal α^*)

$$K(u, v) = \exp\left(-\gamma (u - v)^2\right)$$

with $\gamma = 0.1$

Hint: $\alpha_2^* = \alpha_5^* = \alpha_3^* = \alpha_6^* > 0$ all other $\alpha_i^* = 0$

Find such an α^* and then use KKT conditions to show it is optimal.

(b) Give b^* for the α^* above.

(c) Give the decision function $w^{*T}\phi(x) + b^*$ for $x \in \mathbb{R}$.

[HR]

Exercise2:

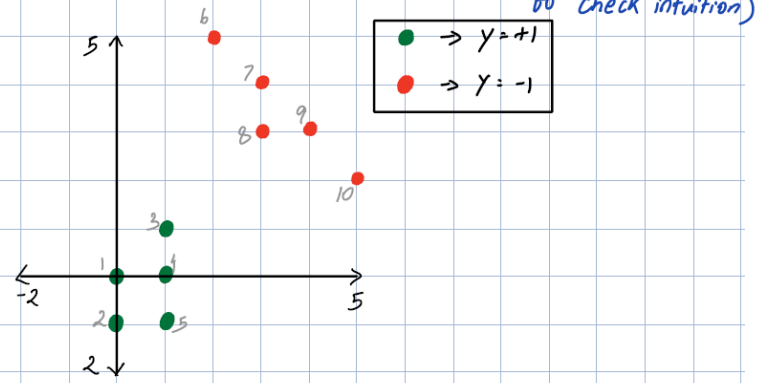
Exercise:

Consider the following hard margin SVM problem with both w & b. Kernel is the linear kernel.

(a) Argue what points are support vectors.

(b) Argue what would be the optimal hyperplane and give w^*, b^*.

(c) Also argue what α^* should be. (Use software to check intuition)



(d) Repeat all the above if data point (x_8, y_8) is removed.

(e) Repeat a, b, c with (x_8, y_8) (x_9, y_9) (x_7, y_7) removed.

Hint: Optimal α^* need not be unique even if w^* and b^* are.

[HR]

Loss function view: Hinge loss

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \Rightarrow \min_{w,b} C\sum_{i=1}^{n}\psi_H(y_i(w^T x_i + b)) + \frac{1}{2}\|w\|^2$$

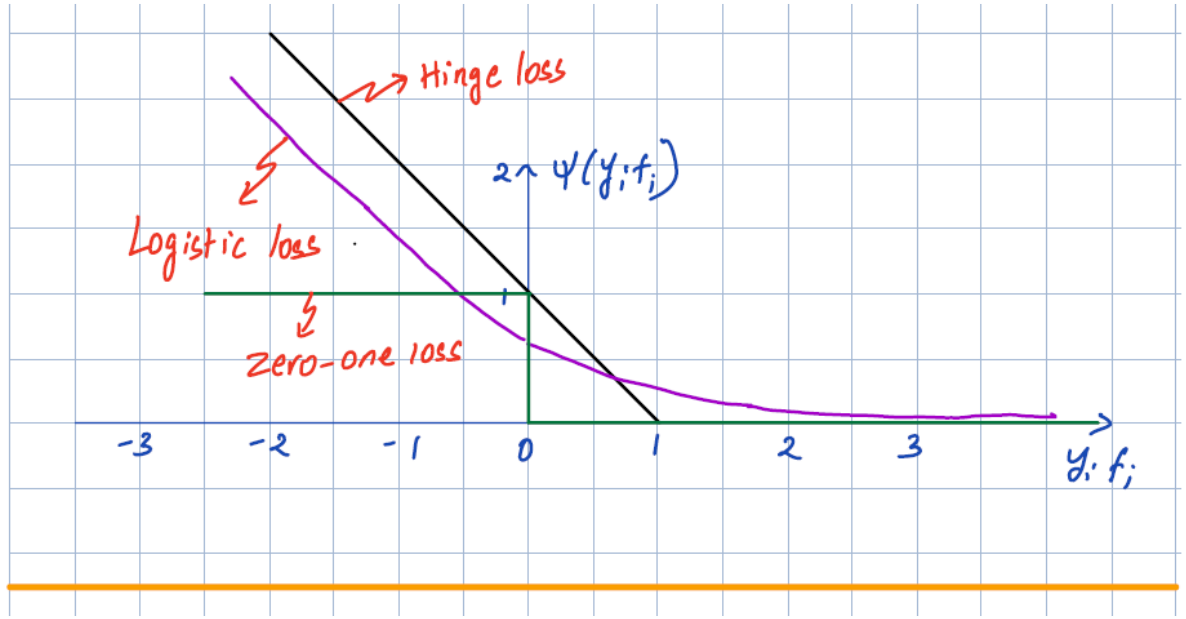
Express ξ_i as a function of w, b:

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

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

$= \max(1 - y_i f_i, 0)$ where $f_i = w^T x_i + b$

$= \psi_H(y_i f_i)$



$\psi_H(u) = \max(1-u, 0)$

[HR]

SVM Interpretations

- Support vectors
- Kernel machines
- Loss fn. view

Goal we set out: concrete understanding of SVM -- hope we reached it!

sklearn.svm.SVC

class sklearn.svm. **SVC**(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=- 1, decision_function_shape='ovr', break_ties=False, random_state=None) ¶ [source]

Primal OP:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0, i = 1, \ldots, n$$

Dual OP:

$$\max_{\alpha} -\frac{1}{2} \alpha^T Q \alpha + e^T \alpha$$

$$\text{subject to } y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$.

Prediction for new point x:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b,$$

[Above formulas from sklearn help pages]

Concluding thoughts

- SVM
  - Concept of max-margin classifn., sparse support vectors, & kernel machines.
  - Use of constrained optimization, and Primal - Dual problems.
  - Extensions: SVR (Support Vector Regression) and RVM (Relevance Vector Machines).

- Next steps: From linear to non-linear regression/classifn.

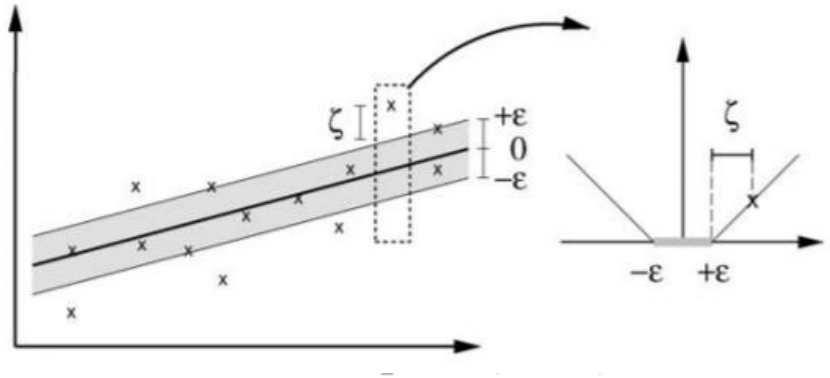
Non-linear method	**(Non-linear) Basis functions**	**Objective function / OP**
Vanilla extn. of linear models	**Fixed** – non-linear basis fns (feature map, $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$) fixed before seeing training data (manually via feature engineering); only weights of these basis fns learnt using training data.	Convex (unconstrained opt.)
SVM	**Selective** – center basis fns on training data points (dual/kernel view) and use training data to learn their weights and select a subset of them (non-zero weight support vectors) for eventual predictions.	Convex (constrained opt.)
Neural networks	**Adaptive** – Fix # of basis fns in advance, but allow them to be adaptive; parameterize basis fns and learn these parameters using training data.	Non-convex

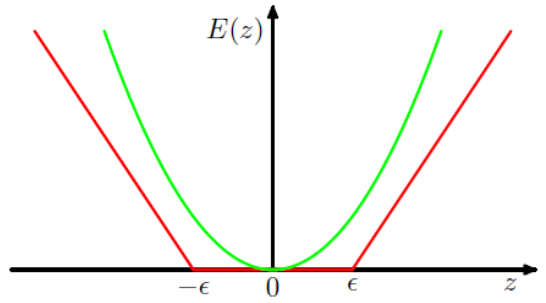
[CMB]

Thank you!

Backup

From classification to regression: Support Vector Regression or SVR (ϵ-insensitive "tube" and obj/loss fns.)

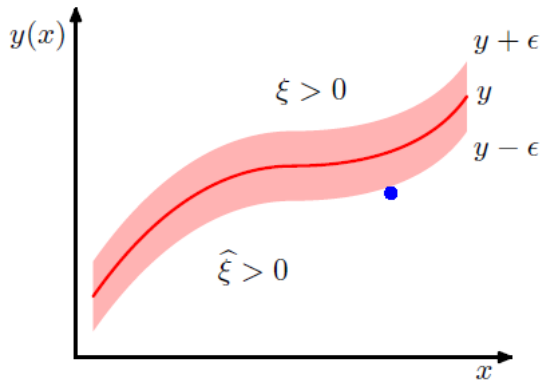




$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

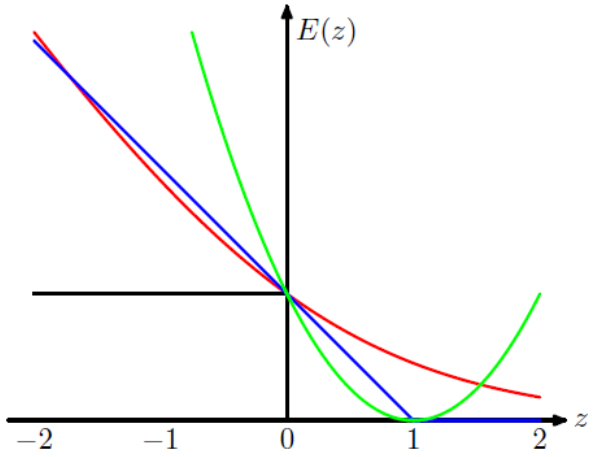
(Soft)



[From CMB; Smola and Scholkopf, 2004 tutorial]

Loss functions drawn to scale

Plot of the 'hinge' error function used in support vector machines, shown in blue, along with the error function for logistic regression, rescaled by a factor of $1/\ln(2)$ so that it passes through the point $(0, 1)$, shown in red. Also shown are the misclassification error in black and the squared error in green.



[CMB]