Linear Discriminant Models (Linear Models for Classification)

Definition: Linear Discriminant Model (LDM) is the following set:

$$\mathcal{F}_L = \left\{ f \;\middle|\; \exists\, \omega = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_n \end{bmatrix} \ni f(x) = \operatorname{sign}(\omega^T \varphi(x) - b) \;\forall x \in \mathcal{X} \right\}$$

→ functions in $\mathcal{F}_L$ can be identified by

model parameters $(\omega, b) \in \mathbb{R}^{n+1}$

→ Effective no. parameters = n

(One can be fixed arbitrarily)

→ Learning with this model can be posed as problem of obtaining the "best" parameters.

IDEAL OBJ.

$$\min_{\omega \in \mathbb{R}^n,\, b \in \mathbb{R}} P\left[Y \neq \operatorname{sign}(\omega^T \varphi(X) - b)\right]$$

s.t. $\|\omega\| = 1$

(I*)

$\approx$

EMPIRICAL OBJ.

$$\min_{\omega \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i \neq \operatorname{sign}(\omega^T \varphi(x_i) - b)\}}$$

s.t. $\|\omega\| = 1$

(I)

→ Minimizes of Training error will go to that of true error, as $m \to \infty$.

→ However, (I) is a non-convex problem with discontinuous objective!

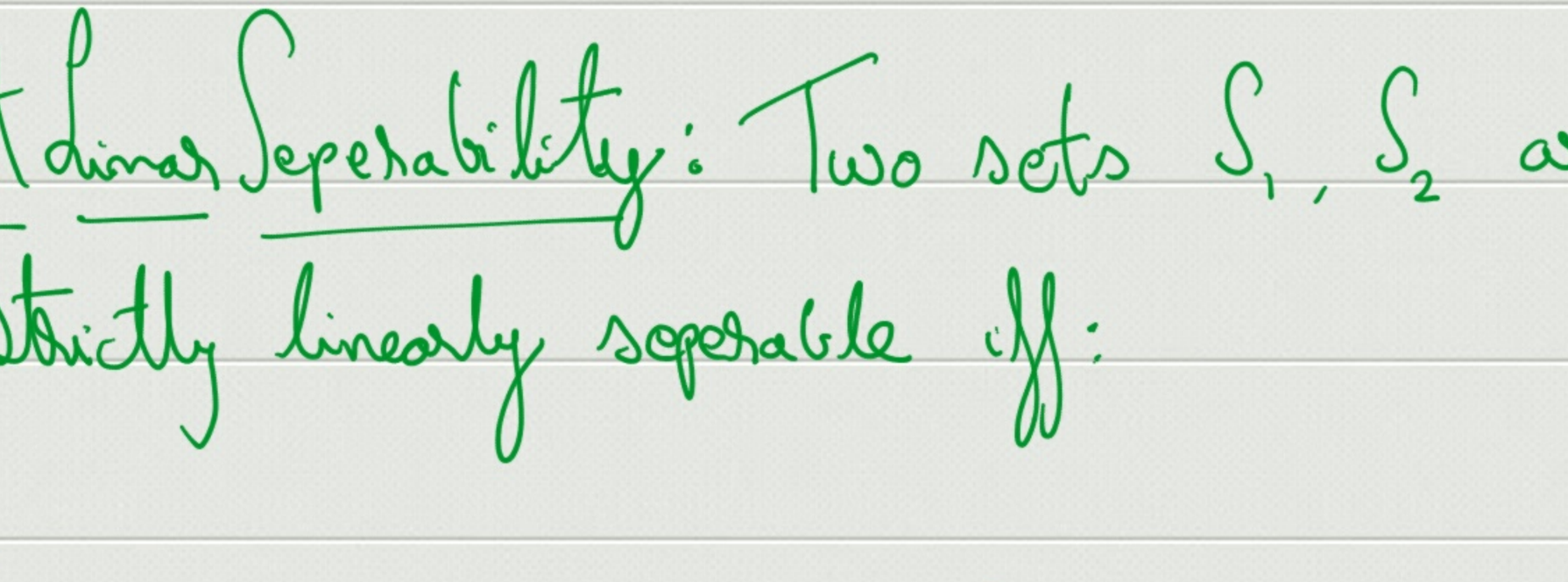
→ In general, known to be Computationally hard

→ If, the two classes are strictly linearly seperable, then polynomial time solver.

Strict Linear Seperability: Two sets S_1, S_2 are said to be strictly linearly seperable iff:

$\exists\ w \in \mathbb{R}^n,\ b \in \mathbb{R},\ \delta > 0 \ \ni$ (i) $w^T x - b \leq -\delta \ \forall\ x \in S_1$

(ii) $w^T x - b \geq \delta \ \forall\ x \in S_2$



With the assumption, (I) can be written as:

$$\min_{w \in \mathbb{R}^n,\ b \in \mathbb{R},\ \rho > 0} \quad -\rho$$

$$\text{s.t.} \quad y_i\left(w^T \varphi(x_i) - b\right) \geq \rho, \quad \forall\, i = 1 \text{ to } m,$$

$$\|w\| = 1.$$

The constraints $\|w\|=1$ & $\rho > 0$ are spoilers:

→ Option (i): Maximize margin.

→ Option (ii): Different normalization.

$$\text{Margin} = 2\rho / \|w\|$$

Option (i):

$$\min_{w \in \mathbb{R}^n,\ b \in \mathbb{R},\ \rho \geq 0} \quad -\rho$$

$$\text{s.t.} \quad y_i\left(w^T \varphi(x_i) - b\right) \geq \rho, \quad \forall\, i = 1 \text{ to } m$$

$$\|w\| \leq 1.$$

(II)

Option 2: Select $\xi = 1$, then Margin $= 2/\|w\|$.

$$\min \quad \frac{1}{2}\|w\|^2$$

$$w \in \mathbb{R}^n, \quad b \in \mathbb{R}$$

$$\text{s.t.} \quad y_i(w^T\varphi(x_i) - b) \geq 1 \quad \forall i = 1 \text{ to } m$$

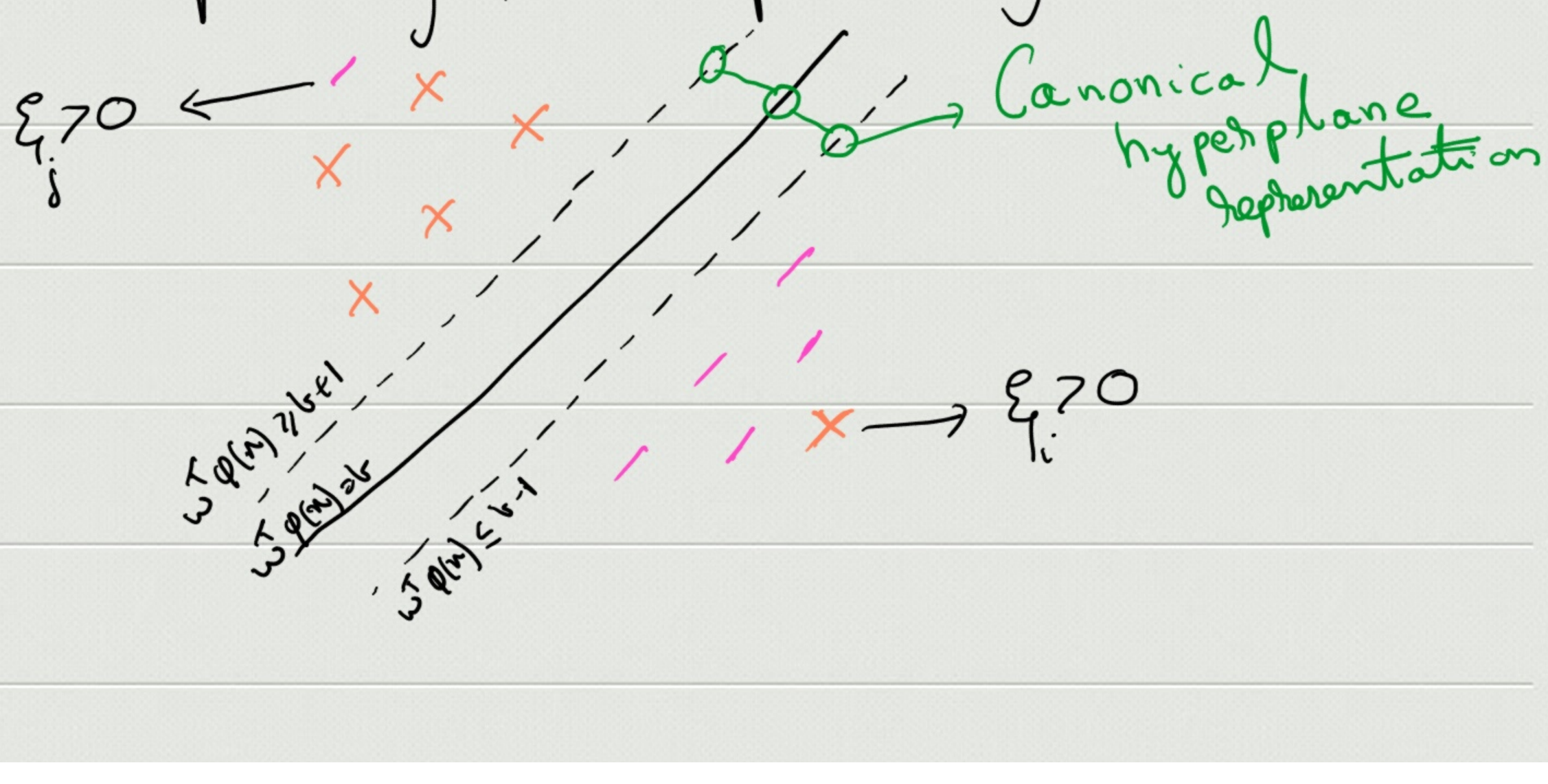
Option 1 meets Option 2:

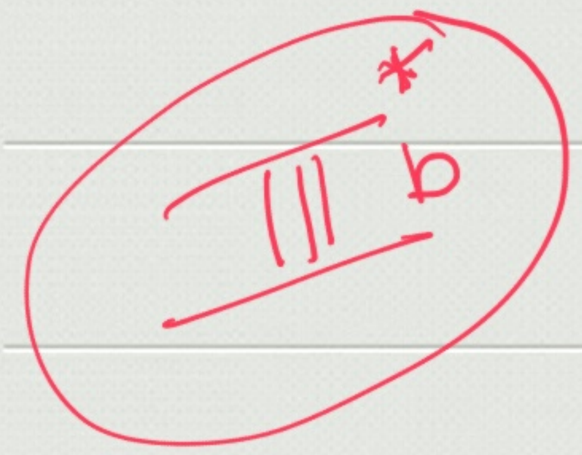
$$\min \quad \frac{1}{2}\|w\|^2$$

$$w \in \mathbb{R}^n, \quad b \in \mathbb{R}$$

$$\text{s.t.} \quad y_i(w^T\varphi(x_i) - b) \geq 1 \quad \forall i = 1 \text{ to } m$$

If assumption of strict separation fails, then:





$$\min_{\substack{w\in\mathbb{R}^n,\\ b\in\mathbb{R},\\ \xi\in\mathbb{R}^m}} \quad \sum_{i=1}^{m} 1_{\{\xi_i>0\}}$$

$$\text{s.t.} \quad y_i\left(w^T\varphi(x_i)-b\right)\geq 1-\xi_i,\ \xi_i\geq 0,$$

$$\forall i=1 \text{ to } m.$$

Combinatorial objective is reason for hardness.

→ Alternative perf. measure?

↓ usual trick

replace no. by sum. (tightest Convex relaxation)

$$\min_{\substack{w\in\mathbb{R}^n,\\ b\in\mathbb{R},\\ \xi\in\mathbb{R}^m}} \quad \sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad y_i\left(w^T\varphi(x_i)-b\right)\geq 1-\xi_i,\ \xi_i\geq 0 \quad \forall i=1 \text{ to } m$$

(III b) Can be re-written as:

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \max\left(0, 1 - y_i(w^T \varphi(x_i) - b)\right)$$

$$\downarrow$$

$$l(y_i\, g(x_i))$$

Equivalent to minimizing average loss $l(yg(x))$

→ Average 0-1 loss gives accuracy

→ Average hinge-loss gives obj. in (III b).

$$\downarrow m \to \infty$$

$$\mathbb{E}\left[\max(0, 1 - \gamma(w^T \varphi(x) - b))\right]$$

expected hinge-loss.

(upper bound on misclass. error).

Surrogate Loss Functions for Binary Classification

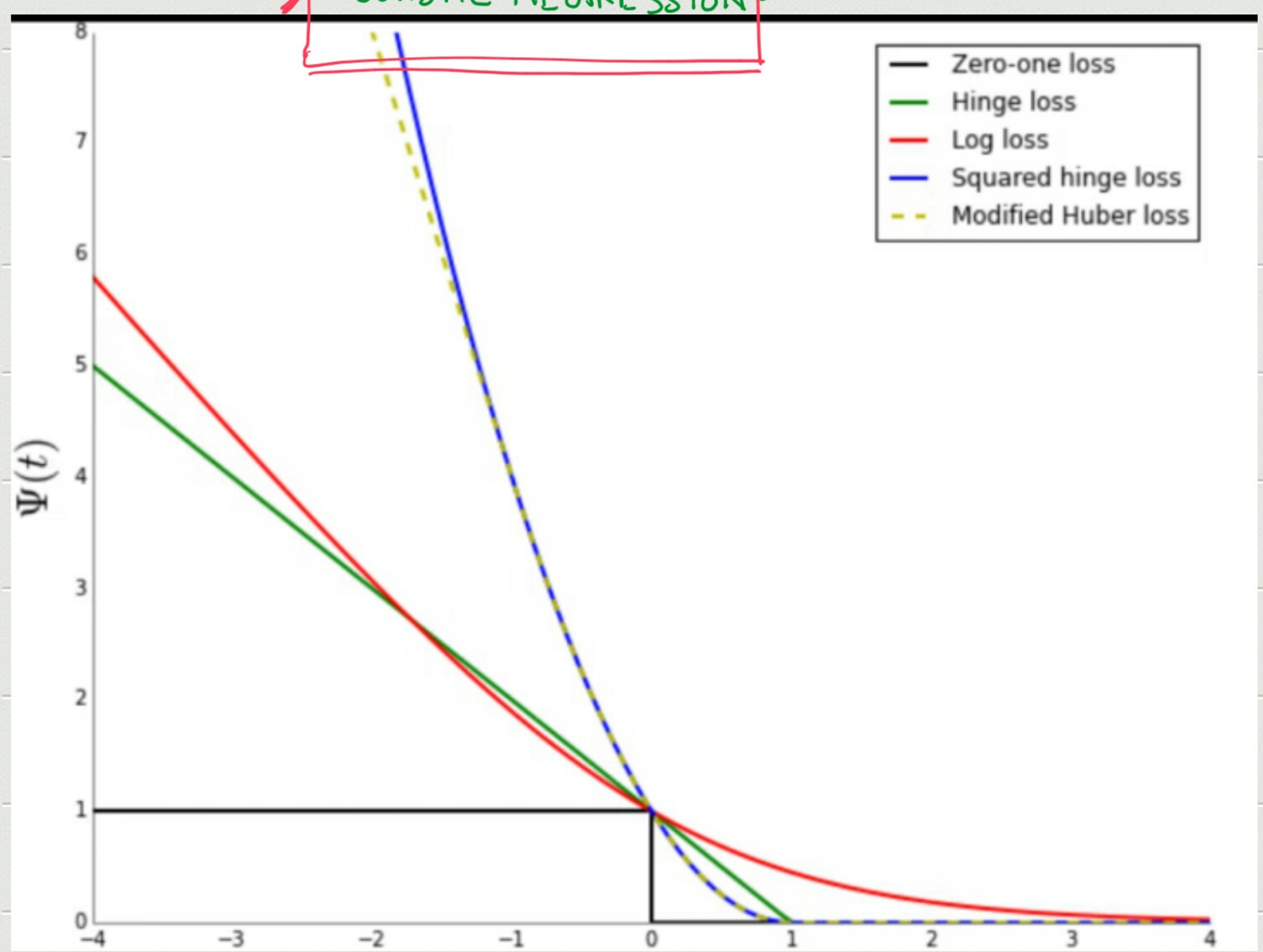
→ 0-1 loss : $l(z) = \begin{cases} 0 & \text{if } z \geq 0 \\ 1 & \text{if } z < 0 \end{cases}$

→ hinge-loss : $l(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ 1-z & \text{if } z < 1 \end{cases}$ **L1-Loss**

→ squared-hinge-loss : $l(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ (1-z)^2 & \text{if } z < 1 \end{cases}$ **L2-Loss**

→ logistic-loss : $l(z) = \log(1 + e^{-z})$

LOGISTIC REGRESSION



Introducing surrogate loss functions in (II) (IV)
will need a "regularization" parameter:

(IIIb)

$$\min_{\omega \in \mathbb{R}^n,\ b \in \mathbb{R},\ \xi \in \mathbb{R}^m} \ \frac{1}{2}\|\omega\|^2 + \frac{C}{m}\sum_{i=1} \xi_i$$

$$\text{s.t.} \quad y_i\left(\omega^T \varphi(x_i) - b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \ \ \forall i = 1 \text{ to } m$$

SVM (Support Vector Machine)

(IIa) is called hard-margin SVM.

C is reg. parameter

(IIb)

$$\min_{\omega \in \mathbb{R}^n,\ b \in \mathbb{R},\ \xi \in \mathbb{R}^m,\ \rho \geq 0} \ \sum_{i=1}^{m} \xi_i - \nu\rho$$

ν is the reg. parameter.

$$\text{s.t.} \quad y_i\left(\omega^T \varphi(x_i) - b\right) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1 \text{ to } m,$$

$$\|\omega\| \leq 1.$$

ν-SVM (ν-Support Vector Machine)

(IIa) is called hard-margin ν-SVM.

General form for maximum-margin lin. disc. :

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}^n} \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} l\left(y_i(w^T\varphi(x_i) - b)\right)$$

"C" is called "hyperparameter"

and for linear discriminators is :

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{m}\sum_{i=1}^{m} l\left(y_i(w^T\varphi(x_i) - b)\right)$$

max-margin lin.
+
sparse loss
= SUPPORT VECTOR METHOD/MODEL

Analysis of Linear Discriminant Models

① They are essentially nearest neighbour models

This is more popularly known as Representer Theorem.

② Above + "sparse" loss = "smart" nearest neighbours.

That is where name Support Vector comes in

Representer Theorem: Both lin. disc. and max-margin versions admit optimal solution w^* of the form:

$$w^* = \sum_{i=1}^{m} \alpha_i y_i \varphi(x_i) \quad \text{for some } \alpha_i \in \mathbb{R}.$$

Inference:

$$g(x) = \text{sign}(w^{*T}\varphi(x) - b)$$

$$= \begin{cases} 1 & \text{if } \sum_i \alpha_i y_i k(x, x_i) \geq b \\ 0 & \text{if } \sum_i \alpha_i y_i k(x, x_i) < b \end{cases}$$

(nearest neighbour)

where, $k(x, z) \equiv \varphi(x)^T \varphi(z) \quad \forall x, z \in \mathcal{X}$

With losses like hinge-loss

$\Rightarrow$ w^* does not depend on the "correct" examples

Further if training error is low,

$\Rightarrow$ many α_i are zero. Non-zero α_i examples are called Support Vectors

$\Rightarrow$ "Smart" nearest neighbours.

Representer result is useful in atleast two more ways.

(i) Training and inferring with arbitrary $\mathcal{X}$ as long as $\varphi(x)^T\varphi(z)$ is known for $x, z \in \mathcal{X}$.

$\longrightarrow \vec{w}^T\varphi(x) = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x)$

(ii) Exploit 'sparsity' in α's for better algorithms.

On efficiently training linear models

$\longrightarrow$ note that there may be multiple choices of α's such that

$$w = \sum_i \alpha_i y_i \varphi(x_i)$$

$\longrightarrow$ not all will be sparse.

HENCE, substituting for w in the optimization need not be the "best" way for obtaining α's

However, it is indeed a baseline:

$$\min_{\alpha \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{2}\alpha^T K \alpha + \frac{C}{m}\sum_{i=1}^{m} l_i\left((K\alpha)_i - b\right),$$

where, $(K)_{ij} = \varphi(x_i)^T \varphi(x_j)$

Here, $(K)_{ij}$ denotes the i,j^{th} entry in the matrix K.

$(K\alpha)_i$ denotes the i^{th} entry in the vector $K\alpha$

One may solve this using 'cvx' or gradient descent. However, since "α" has no special" property, there were observed to not scale well.

Whereas, we intuitively know that α could be chosen to be sparse atleast with losses like hinge-loss. Hence we aim at analyzing the optimal solution, more carefully.

→ So, analyze optimal soln. more carefully

→ Simple gradient cond. is not insightful:

$$w + \frac{c}{m}\sum_{i=1}^{m} l'\left(y_i w^T \varphi(x_i)\right)\varphi(x_i) = 0$$

Solve following instead ← if decoupled then better insight.

$$w = \sum_i \alpha_i \varphi(x_i),$$

$$\frac{c}{m} l'\left(y_i w^T \varphi(x_i)\right) = -\alpha_i$$

insightful, but how to solve?

Following is a generic technique that expores both there:

Rewrite optimization prob. as:

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R},\, z \in \mathbb{R}^m} \frac{1}{2}\|w\|^2 + \frac{c}{m}\sum_{i=1}^{m} l_i(z_i)$$

$$\text{s.t} \quad z_i = w^T \varphi(x_i) - b_i \quad \forall i = 1 \text{ to } m$$

$$l_i(z_i) \equiv l(y_i z_i)$$

$$= \min_{\substack{w \in \mathbb{R}^m, b \in \mathbb{R}, \\ z \in \mathbb{R}^m}} \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} l_i(z_i) + \max_{\alpha \in \mathbb{R}^m} \sum_i \alpha_i \left(z_i - w^T\varphi(x_i) - b\right)$$

$$= \max_{\alpha \in \mathbb{R}^m} -\frac{1}{2}\alpha^T G \alpha - \frac{C}{m}\sum_{i=1}^{m} l_i^*\left(\frac{m\alpha_i y_i}{C}\right),$$

s.t. $\sum_i \alpha_i y_i = 0,$

where, $G_{ij} = y_i y_j k(x_i, x_j)$,

l^* is known as
① Conjugate (or)
② Fenchel dual (or)
③ Legendre transform of l

$$l_i^*(\lambda) = \max_{z \in \mathbb{R}} \lambda z - l_i(z)$$

and more importantly, at optimality, [given for hinge loss as below]

$$\rightarrow w = \sum_{i=1}^{m} \alpha_i y_i \varphi(x_i),$$

$$\rightarrow \alpha_i = 0 \Rightarrow y_i\left(w^T\varphi(x_i) - b\right) \geq 1$$

$$\rightarrow \alpha_i \in \left(0, \frac{C}{m}\right) \Rightarrow y_i\left(w^T\varphi(x_i) - b\right) = 1$$

$$\rightarrow \alpha_i = \frac{C}{m} \Rightarrow y_i\left(w^T\varphi(x_i) - b\right) \leq 1$$

} Support Vectors

Exercise: Compute conjugate of all the surrogate losses you know.

Whenever, optimal soln. is known to be sparse, two algorithms come to mind:

(i) Co-ordinate descent

→ Optimize one of few variables iteratively.

(ii) Active-Set methods

→ Guess the right "active" variables.

→ In each case, the sub-problem at every iterate is small and no. iterations may not be high as solution is known to be sparse

→ Minimum no. variables to optimize is 2, ∵ $\sum_i \alpha_i y_i = 0$

→ Optimization wrt such two variables

→ has closed form solution

→ Selection of pair by heuristic:

→ Worst optimality violator.

→ If 'b' is set to zero (Say by $\phi(x) \rightarrow \begin{bmatrix} \phi(x) \\ 1 \end{bmatrix}$)

→ round robin variable by variable is ok.

→ For non-sparse losses, above will be slow.

→ logistic regression will be covered later.

LAGRANGE DUALITY

Given, $\min_{x \in X} f(x)$

s.t. $g_i(x) \leq 0 \quad \forall i = 1 \text{ to } m$

(I) PRIMAL

'x' is primal variable.

Assumptions: (i) All f, g_i are convex

(ii) Feasibility, Boundedness, Solvability

(iii) Slater's cond. if g's are non-linear.

Define: $L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \quad \forall x \in X, \lambda \geq 0$

LAGRANGIAN FUNCTION

$\overline{L}(x) \equiv \max_{\lambda \geq 0} L(x, \lambda)$ | $\underline{L}(\lambda) \equiv \min_{x \in X} L(x, \lambda)$

LAGRANGE DUALITY

$$\text{(I)} = \min_{x \in X} \overline{L}(x) = \min_{x \in X} \max_{\lambda \geq 0} L(x, \lambda)$$

$$= \max_{\lambda \geq 0} \min_{x \in X} L(x, \lambda) = \max_{\lambda \geq 0} \underline{L}(\lambda) \equiv \text{(II)}$$

(I) ↓ DUAL (II) ↓

x^* is an optimal soln. of (I)
λ^* is an optimal soln. of (II)
$\Big\}$ $\overline{L}(x^*) = \underline{L}(\lambda^*)$

TST: x^* is an optimal soln. of $\min_{x \in X} L(x, \lambda^*)$;
λ^* is an optimal soln. of $\max_{\lambda \geq 0} L(x^*, \lambda)$

Proof:

$$\overline{L}(x^*) = \max_{\lambda \geq 0} L(x^*, \lambda) \geq L(x^*, \lambda^*) \geq \min_{x \in X} L(x, \lambda^*) = \underline{L}(\lambda^*)$$

$$f(x^*) = f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) \iff \lambda_i^* g_i(x^*) = 0 \quad \forall i$$

Optimality Conditions

Given λ^*, necessary conds. on x^* are:

(i) $x^* \in X, \; g_i(x^*) \leq 0 \; \forall i$ (Feasibility)

(ii) $x^* = \underset{x \in X}{\operatorname{argmin}} L(x, \lambda^*) \iff \underset{L \text{ is diff.}}{\overset{X \text{ is open}}{\Longleftrightarrow}} \nabla_x L(x, \lambda^*)\Big|_{x=x^*} = 0$ (gradient)

(iii) $\lambda_i^* g_i(x^*) = 0 \quad \forall i = 1 \text{ to } m.$ (Complementary Slackness)

Interestingly, above are sufficient for optimality of x^*.

→ infact, sufficient if λ exists!

→ So (i), (ii), (iii) are nec. & suff. conds. for optimality of both x^* and λ^*.

Dual of MMLC + hinge-loss

$$l_i(z) \equiv \max(0, 1 - y_i z)$$

$$l_i^*(\alpha) = \max_{z \in \mathbb{R}} \alpha z - l_i(z)$$

$$= \max_{z \in \mathbb{R}} \alpha z - \max(0, 1 - y_i z)$$

$$= \max\Bigg(\underbrace{\max_{z \geq 1 - y_i z < 0} \alpha z}_{\text{if } \alpha y_i \leq 0 \;\; -|\alpha| \;\; ; \;\; \text{if } \alpha y_i > 0 \;\; \infty}, \underbrace{\max_{z \geq 1 - y_i z \geq 0} (\alpha + y_i) z - 1}_{\text{if } (\alpha + y_i) y_i \geq 0 \;\; \alpha y_i \;\; ; \;\; \text{if } (\alpha + y_i) y_i < 0 \;\; \infty} \Bigg)$$

$$l_i^*(\alpha) = \begin{cases} -|\alpha| & \text{if } -\alpha y_i \in [0, 1] \\ \infty & \text{else} \end{cases}$$

From above (& otherwise), we can get z^* such that:

$$\alpha z^* - \max(0, 1 - y_i z^*) = -|\alpha|$$

$(-\alpha y_i \in [0,1])$

$\longrightarrow \alpha = 0$, then $1 - y_i z^* \leq 0$

$\longrightarrow -\alpha y_i = 1$, then $-y_i z^* - \max(0, 1 - y_i z^*) = -1$

$\Leftrightarrow 1 - y_i z^* \geq 0$

$\longrightarrow \alpha \in (-1, 0)$, then $\alpha(z^* - 1) = \max(0, 1 - z^*)$

$\Leftrightarrow z^* = 1$

$\longrightarrow \alpha \in (0, 1)$, then $\alpha(z^* + 1) = \max(0, 1 + z^*)$

$\Leftrightarrow z^* = -1$

$\longrightarrow -\alpha y_i \in (0, 1)$, then $1 - y_i z^* = 0$

Our dual: $$\max_{\alpha \in \mathbb{R}^m} -\frac{1}{2}\alpha^T K \alpha - \frac{C}{m}\sum_{i=1}^{m} l_i^*\left(-\frac{m\alpha_i}{C}\right)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i = 0$$

$$(K)_{ij} = \varphi(x_i)^T \varphi(x_j)$$

Lets substitute $\alpha_i y_i \longrightarrow \alpha_i$

$$= \max_{\alpha \in \mathbb{R}^m} -\frac{1}{2}\alpha^T G \alpha - \frac{C}{m}\sum_{i=1}^{m} l_i^*\left(-\frac{m y_i \alpha_i}{C}\right)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\alpha_i \begin{cases} \frac{m\alpha_i}{C} & \alpha_i \in [0, \frac{C}{m}] \\ \infty & \text{else} \end{cases}$$

For hinge-loss

$$(G)_{ij} = y_i y_j \varphi(x_i)^T \varphi(x_j)$$

$$= \boxed{\begin{aligned} &\max_{\alpha \in \mathbb{R}^m} -\frac{1}{2}\alpha^T G \alpha - 1^T \alpha \\ &\text{s.t.} \quad 0 \le \alpha \le \frac{C}{m} 1 \,, \; y^T \alpha = 0 \end{aligned}}$$

Optimality conds. for hinge-loss:

(i) $w \in \mathbb{R}^n$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}$, $\xi_i = w^T \varphi(x_i) - b \quad \forall i$ (Feasibility)

(ii) $w = \sum_{i=1}^{m} \alpha_i y_i \varphi(x_i)$, $\sum_{i=1}^{m} \alpha_i y_i = 0$, (gradient)

$\alpha_i = 0 \Rightarrow y_i (w^T \varphi(x_i) - b) \geq 1$

$\alpha_i = \frac{C}{m} \Rightarrow y_i (w^T \varphi(x_i) - b) \leq 1 \rightarrow$ bounded Support Vectors

$\alpha_i \in (0, \frac{C}{m}) \Rightarrow y_i (w^T \varphi(x_i) - b) = 1 \rightarrow$ non-bound. Support Vectors

Co-ordinate descent to exploit sparsity

→ assume $b = 0$, so that $\sum_i \alpha_i y_i = 0$ does not appear!

→ Update sequentially one α_i at a time.

→ repeat till convergence.

↘ 1-d convex QP
⇒ closed form solution

↓

Use duality gap.

→ Get w^*, b^* from α

→ (I) − (II) = duality gap ↓ 0