Lecture 19: Support Vector Machines

Training data: $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} \text{—} & x_1^T & \text{—} \\ & \vdots & \\ \text{—} & x_n^T & \text{—} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

find ~~\$~~ weights w s.t. $Xw \approx y$

$$\hat{w} = \arg\min_w \underline{\|y - Xw\|_2^2} + \lambda r(w)$$

e.g. $r(w) = 0$ (Least squares)

$r(w) = \|w\|_2^2$ (Ridge regression)

$r(w) = \|w\|_1$ (Lasso estimation)

Prox gradient algorithm

$$\hat{w} = \arg\min_{w} \overbrace{\underbrace{l(w; X, y)}_{\substack{\text{loss function} \\ \text{measure of how} \\ \text{well } w \text{ fits} \\ \text{data} \\ \text{e.g. } \|y - Xw\|_2^2}}}^{l(w)} + \lambda \underbrace{r(w)}_{\text{regularizer}}$$

important for l and r to be convex functions for prox grad to find optimal w

alg

- compute initial $\hat{w}^{(0)}$
- for $k = 0, 1, 2, \ldots$

$$z^{(k)} = \hat{w}^{(k)} - \tau \nabla l(\hat{w}^{(k)})$$

if $l(w) = \|y - Xw\|_2^2$

$\nabla l(w) = -2X^T(y - Xw)$

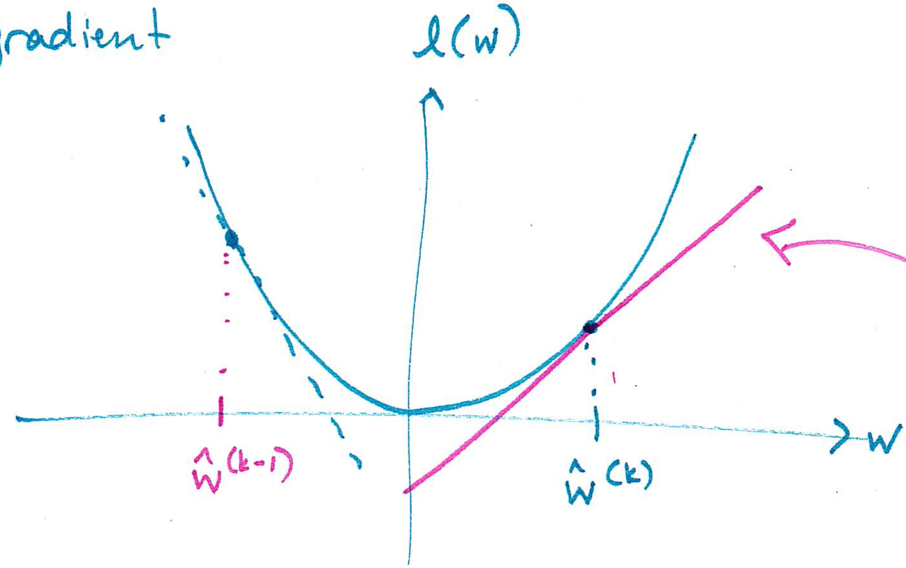
$$\left(\text{or } z^{(k)} = \hat{w}^{(k)} - \tau \underbrace{\partial l(\hat{w}^{(k)})}_{\text{subgradient}} \right)$$

$$\hat{w}^{(k+1)} = \arg\min_{w} \|z^{(k)} - w\|_2^2 + \lambda \tau r(w)$$

check for convergence: if $\|\hat{w}^{(k)} - \hat{w}^{(k+1)}\| < \varepsilon$, STOP

end

gradient



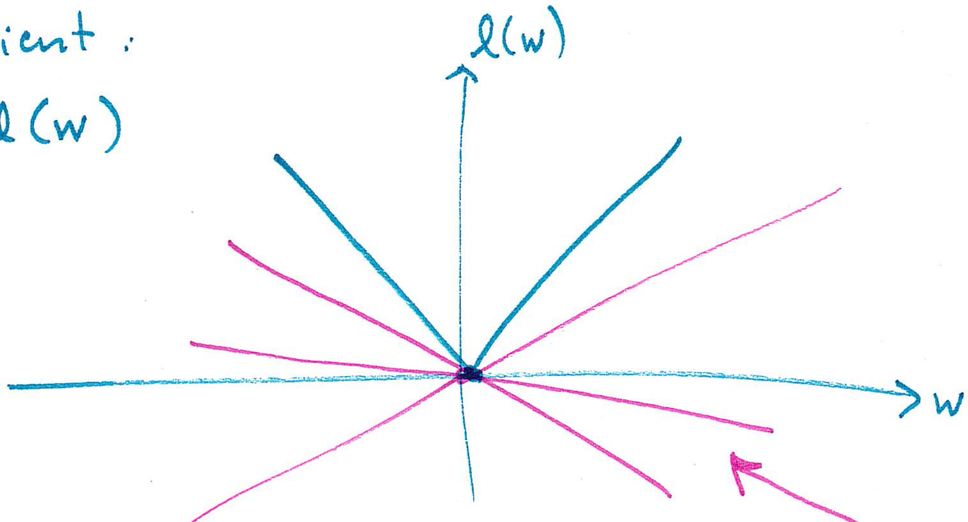
← here, $\ell(w)$ is convex

($\ell(w) \geq$ tangent / gradient @ w)

$$\ell(u) \geq \underbrace{\ell(\hat{w}^{(k)}) + (u - \hat{w}^{(k)})^T \underline{\underline{\nabla \ell(\hat{w}^{(k)})}}}_{\text{line}}$$

subgradient :

$\partial \ell(w)$

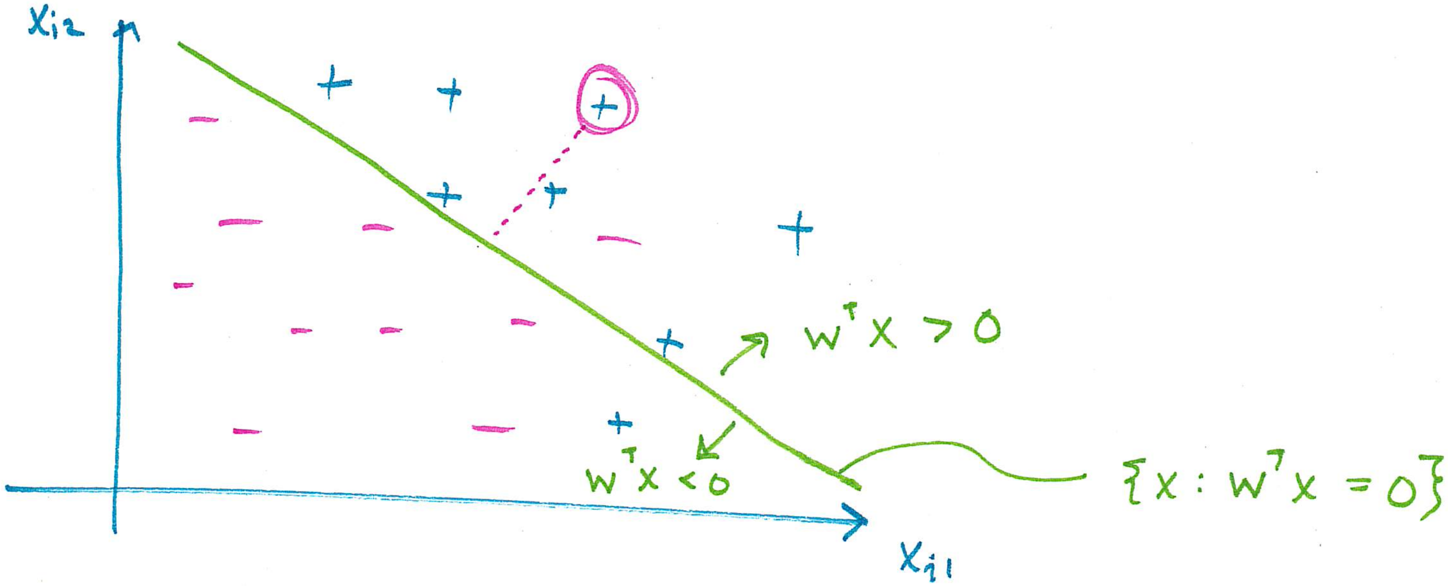


subgradient :

any vector v

$$\ell(u) \geq \ell(\hat{w}^{(k)}) + (u - \hat{w}^{(k)})^T \underline{\underline{v}}$$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ 1 \end{bmatrix}, \quad y_i \in \{-1, +1\}$$



Classification rule: $\hat{y} = \text{sign}(w^T x)$

Squared error loss

$$\|y - Xw\|_2^2 = \sum_{i=1}^{n} (y_i - x_i^T w)^2 = \sum_{i=1}^{n} (1 - \underbrace{y_i x_i^T w})^2$$

(b/c $y_i^2 = 1$)

if this $\neq 1$, then we incur loss even if point is correctly classified.

predict whether someone plays bball based on height :

$+1$ = plays bball

-1 = does not play bball

4 training samples / heights : 5'10, 5'11, 6'1, 6'10

labels : -1 -1 $+1$ $+1$

ideal

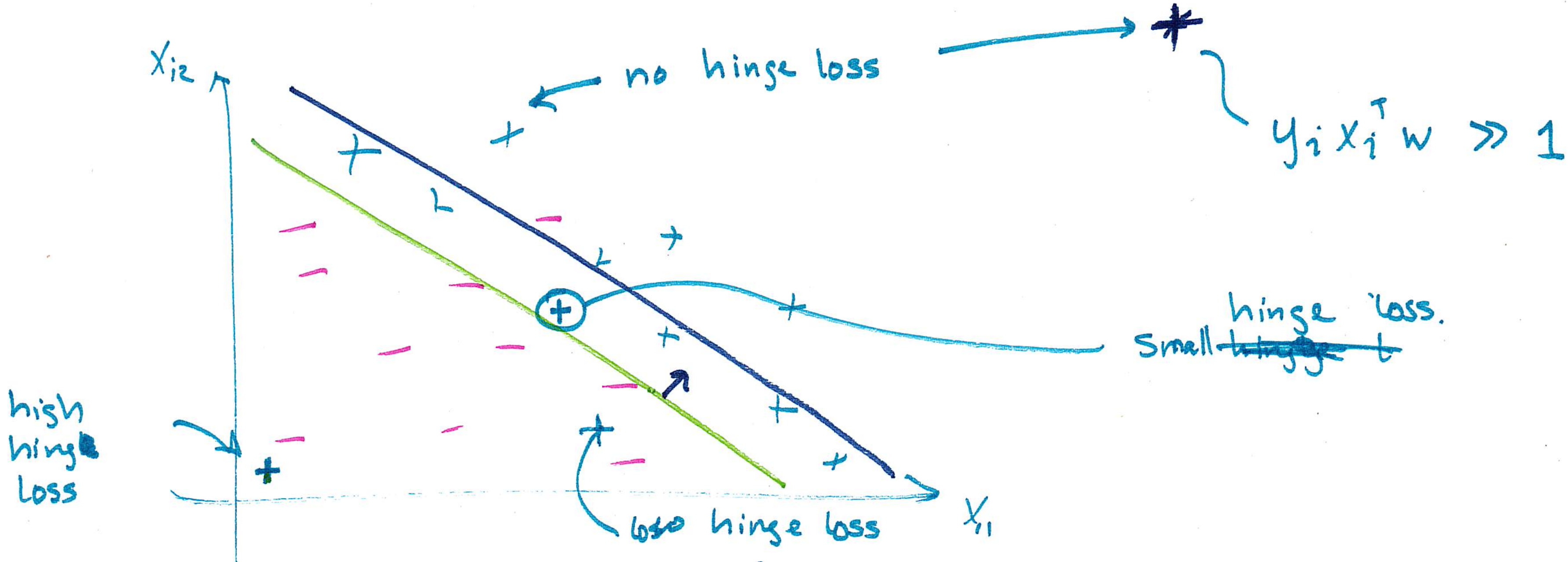
LS line

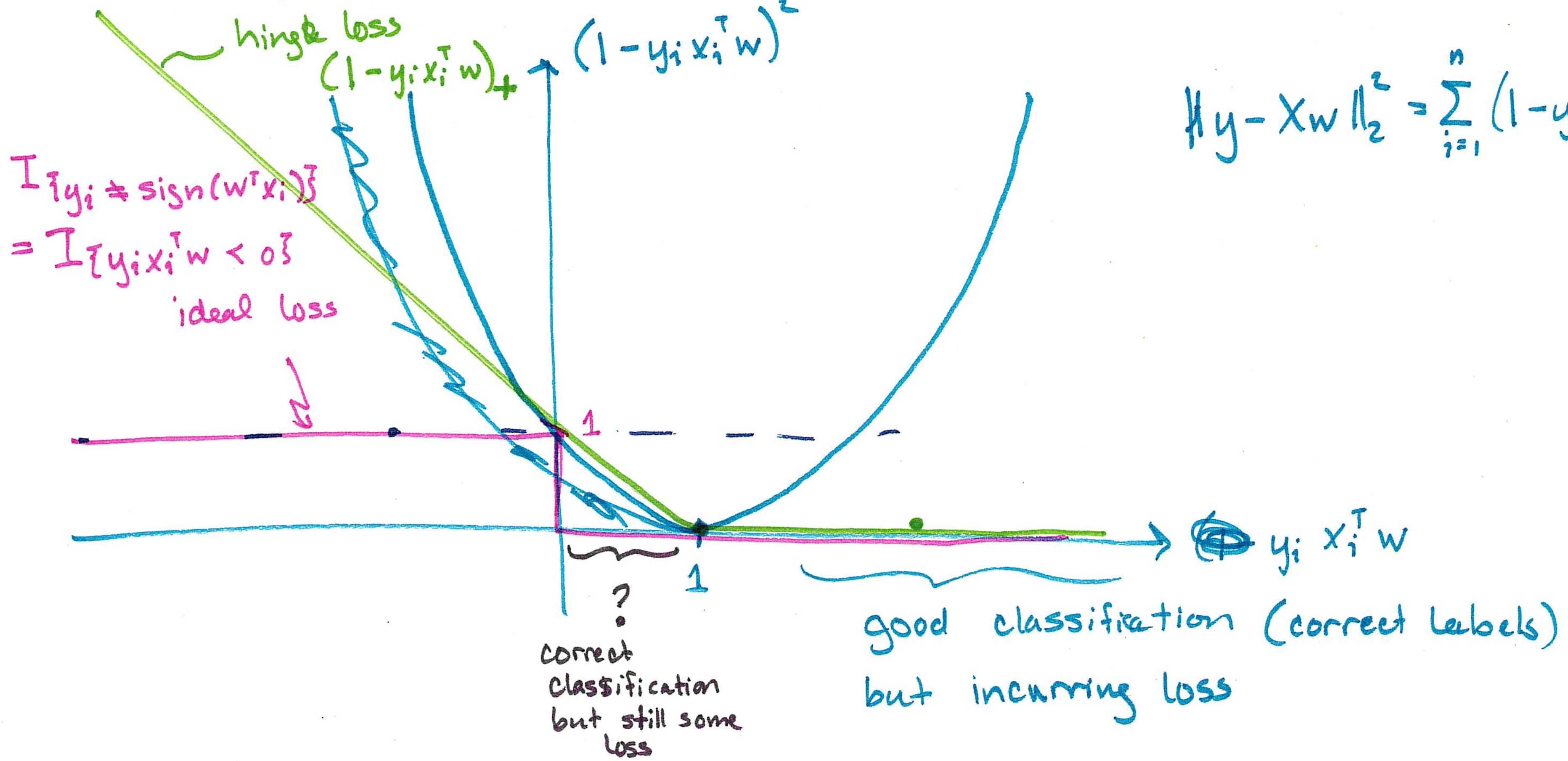
x_{i1} = height − 6'

$$x_i = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 10 \\ 1 \end{bmatrix}$$

$$X = \begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 10 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}, \quad \hat{w}_{LS} = (X^T X)^{-1} X^T y = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \cdot \frac{7}{45}$$

$x_i^T \hat{w}_{LS} > 0 \Rightarrow x_{i1} > 2 \Rightarrow$ classifiation rule says people 6'2" and taller play bball





$$\|y - Xw\|_2^2 = \sum_{i=1}^{n} (1 - y_i x_i^T w)^2$$

Goal: choose w s.t. $\text{sign}(w^T x_i) = y_i$ as often as possible

LS reality: instead of minimizing # mistakes

$$\sum_{i=1}^{n} I\{y_i \neq \text{sign}(w^T x_i)\} = \text{"ideal loss"}$$

we are minimizing

$$\sum_{i=1}^{n} (y_i - w^T x_i)^2$$

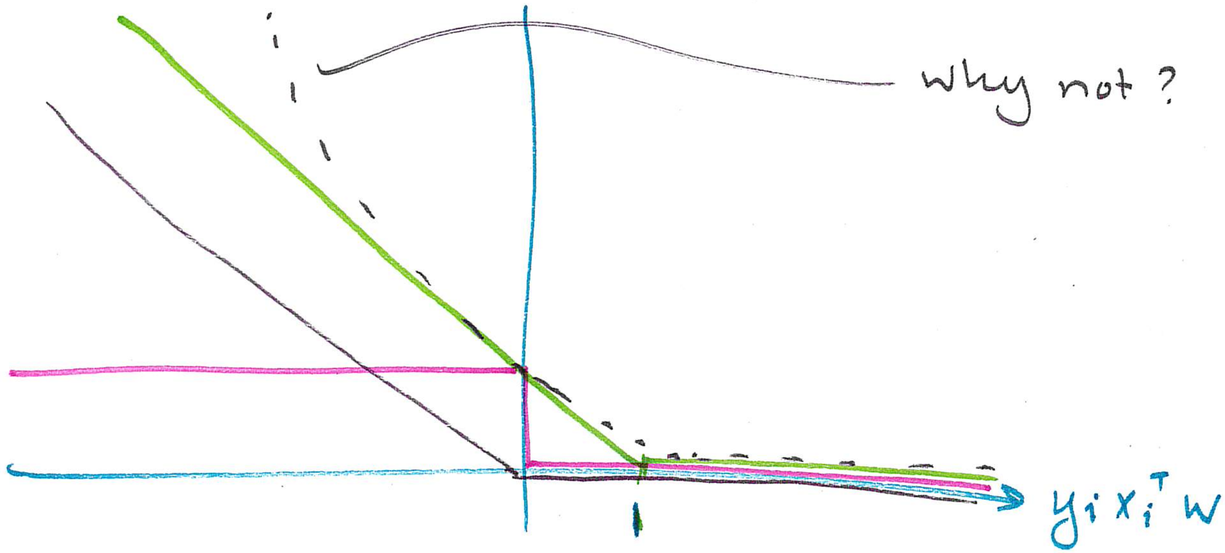
We need a <u>convex</u> loss function that "mimics" the ideal loss

$\boxed{\text{hinge loss}}$

$$l(w) = \sum_{i=1}^{n} (1 - y_i x_i^T w)_+$$

<u>convex</u>

$$(a)_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$



$$\nabla_w \left(1 - y_i x_i^T w \right) = - y_i x_i$$

$$\ell(w) = \sum_{i=1}^{n} \left(1 - y_i x_i^T w\right)_+$$

$$\nabla \ell(w) = \sum_{i=1}^{n} I_{\{y_i x_i^T w < 1\}} (-y_i x_i)$$

if we minimize $\sum_i \left(1 - y_i x_i^T w\right)_+ + \lambda \|w\|_2^2$

this is called as $\boxed{\text{support vector machine}}$