Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- Homework 3 is due Friday
- Slido poll questions today
  - Event code: 3183558
  - Go to slido.com
  - Join as a participant
  - When asked, fill in your name and usc email address
  - Leave your browser window open during lecture

Today's lecture

- Mean-squared-error (MSE) techniques for regression
  - Criterion function
  - Least squares
  - LMS
- MSE techniques for classification (1)
  - Criterion function and interpretation

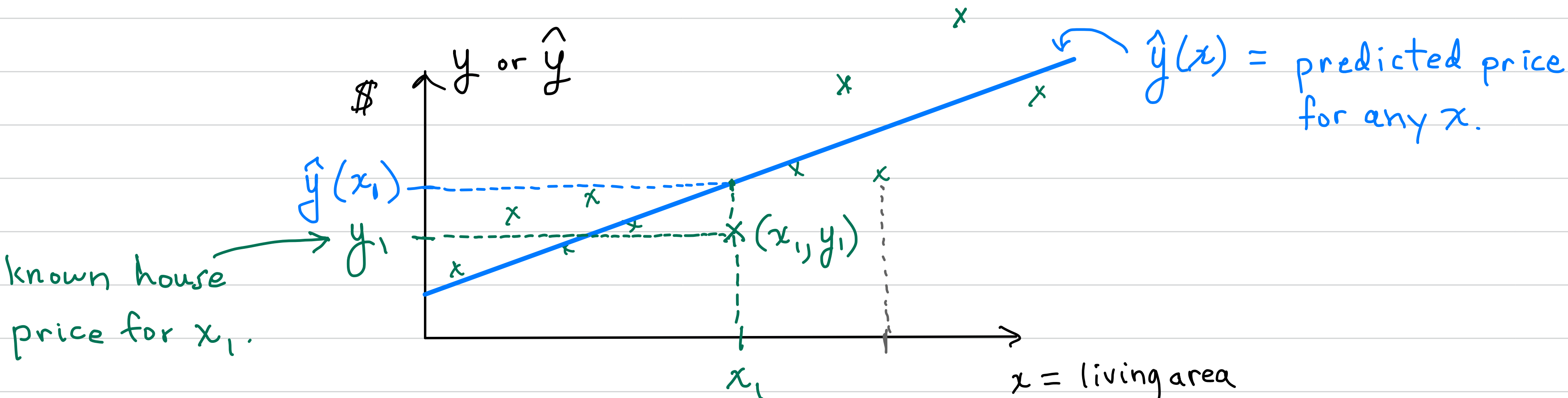
S shows locations of Slido questions.

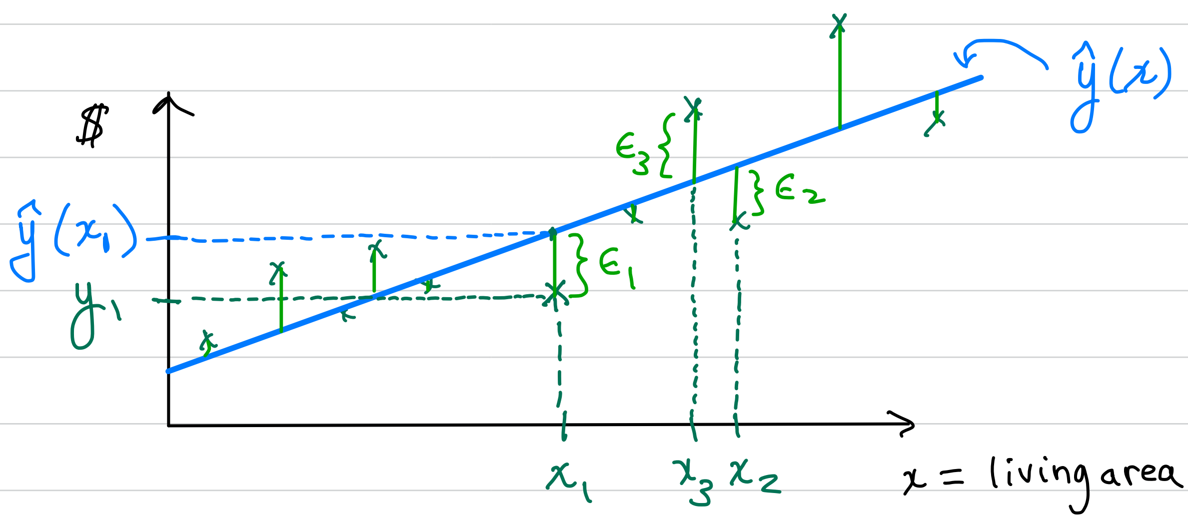
Mean-squared-error techniques: Regression (augmented notation)

Ex: predict house prices: $\hat{y}(\underline{x})$ = price prediction

y_i = actual price for data point $\underline{x}_i$.

Assume: linear model





slido notation: y-hat (x_i) means $\hat{y}(\underline{x}_i)$.

w^T x_i means $\underline{w}^T \underline{x}_i$

w_0 means w_0.

Criterion function: $J(\underline{w}) = \frac{1}{N} \sum_{i=1}^{N} \epsilon_i^2$

$$\boxed{J(\underline{w}) = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{y}(x_i) - y_i\right]^2 = MSE}$$ = mean-squared error

S Linear model: $\hat{y}(\underline{x}_i) = \underline{w}^T \underline{x}_i$

$$\Rightarrow J(\underline{w}) = \frac{1}{N} \sum_{i=1}^{N} \Big[\underbrace{\underline{w}^T \underline{x}_i}_{\text{house price prediction}} - y_i\Big]^2$$

known house price

Let $N = N_{Tr}$

Let $\underline{\underline{X}}_{Tr} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}$ = data matrix, (training data) $\quad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ = data-output (or target-output) vector

$$\Rightarrow \underline{\underline{X}}_{Tr} \underline{w} = \begin{bmatrix} \underline{x}_1^T \underline{w} \\ \underline{x}_2^T \underline{w} \\ \vdots \\ \underline{x}_N^T \underline{w} \end{bmatrix} = \begin{bmatrix} \hat{y}(\underline{x}_1) \\ \hat{y}(\underline{x}_2) \\ \vdots \\ \hat{y}(\underline{x}_N) \end{bmatrix} \triangleq \hat{\underline{y}}$$ = vector of output predictions

We can write $J(\underline{w})$ as:

$$\boxed{J(\underline{w}) = \frac{1}{N} \left\| \underline{\underline{X}}\, \underline{w} - \underline{y} \right\|_2^2 = \frac{1}{N} \left(\underline{\underline{X}}\, \underline{w} - \underline{y} \right)^T \left(\underline{\underline{X}}\, \underline{w} - \underline{y} \right)}$$

For learning (training), choose $\underline{\underline{X}} = \underline{\underline{X}}_{Tr}$ and $N = N_{Tr}$.

Minimization of $J(\underline{w})$:

Is $J(\underline{w})$ differentiable?

Is $J(\underline{w})$ convex?

Possible approaches:

1. $\nabla_{\underline{w}} J(\underline{w}) = \underline{0}$

   Solve algebraically

   $\Rightarrow$ least squares

   or ordinary least squares

   or pseudoinverse

2. Gradient descent

   Sequential or stochastic GD

   $\Rightarrow$ least mean squares (LMS)

Minimize $J(\underline{w})$ algebraically:

$$J(\underline{w}) = \frac{1}{N}\left(\underline{\underline{X}}\,\underline{w}-\underline{y}\right)^T\left(\underline{\underline{X}}\,\underline{w}-\underline{y}\right) = \frac{1}{N}\left[\underline{w}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - \underline{w}^T\underline{\underline{X}}^T\underline{y} - \underline{y}^T\underline{\underline{X}}\,\underline{w} + \underline{y}^T\underline{y}\right]$$

$$= \frac{1}{N}\left[\underline{w}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{w}^T\underline{\underline{X}}^T\underline{y} + \underline{y}^T\underline{y}\right]$$

$$\nabla_{\underline{w}} J(\underline{w}) = \frac{1}{N}\left[2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{\underline{X}}^T\underline{y}\right]$$

$$\nabla_{\underline{w}} J(\underline{\hat{w}}) = \underline{0} = \frac{1}{N}\left[2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\hat{w}} - 2\underline{\underline{X}}^T\underline{y}\right]$$

$$\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\hat{w}} = \underline{\underline{X}}^T\underline{y}$$

If $\underline{\underline{X}}^T\underline{\underline{X}}$ is nonsingular, then:

$$\underline{\hat{w}} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{y}$$

$$\boxed{\begin{array}{l} \text{Let } \underline{\underline{X}}^{-} \triangleq \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T = \text{Moore-Penrose (left) pseudoinverse of } \underline{\underline{X}} \\ \text{Then } \underline{\hat{w}} = \underline{\underline{X}}^{-}\underline{y} \qquad (\underline{\underline{X}} = \underline{\underline{X}}_{Tr}\text{, augmented}) \end{array}}$$

Solution $\underline{\hat{w}}$ for least-squares regression

Comments on $\underline{\underline{X}}^-$

1. $\underline{\underline{X}}^- \underline{\underline{X}} = \left[(\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \right] \underline{\underline{X}} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} (\underline{\underline{X}}^T \underline{\underline{X}}) = \underline{\underline{I}}$

2. $\underline{\underline{X}}\,\underline{\underline{X}}^- \neq \underline{\underline{I}}$ in general.

Comments on least-squares solution:

S

1. Predictions on dataset $\mathcal{D}$: $\hat{\underline{y}} = \underline{\underline{X}}_{\mathcal{D}} \hat{\underline{w}} = \underline{\underline{X}}_{\mathcal{D}} (\underline{\underline{X}}^- \underline{y})$

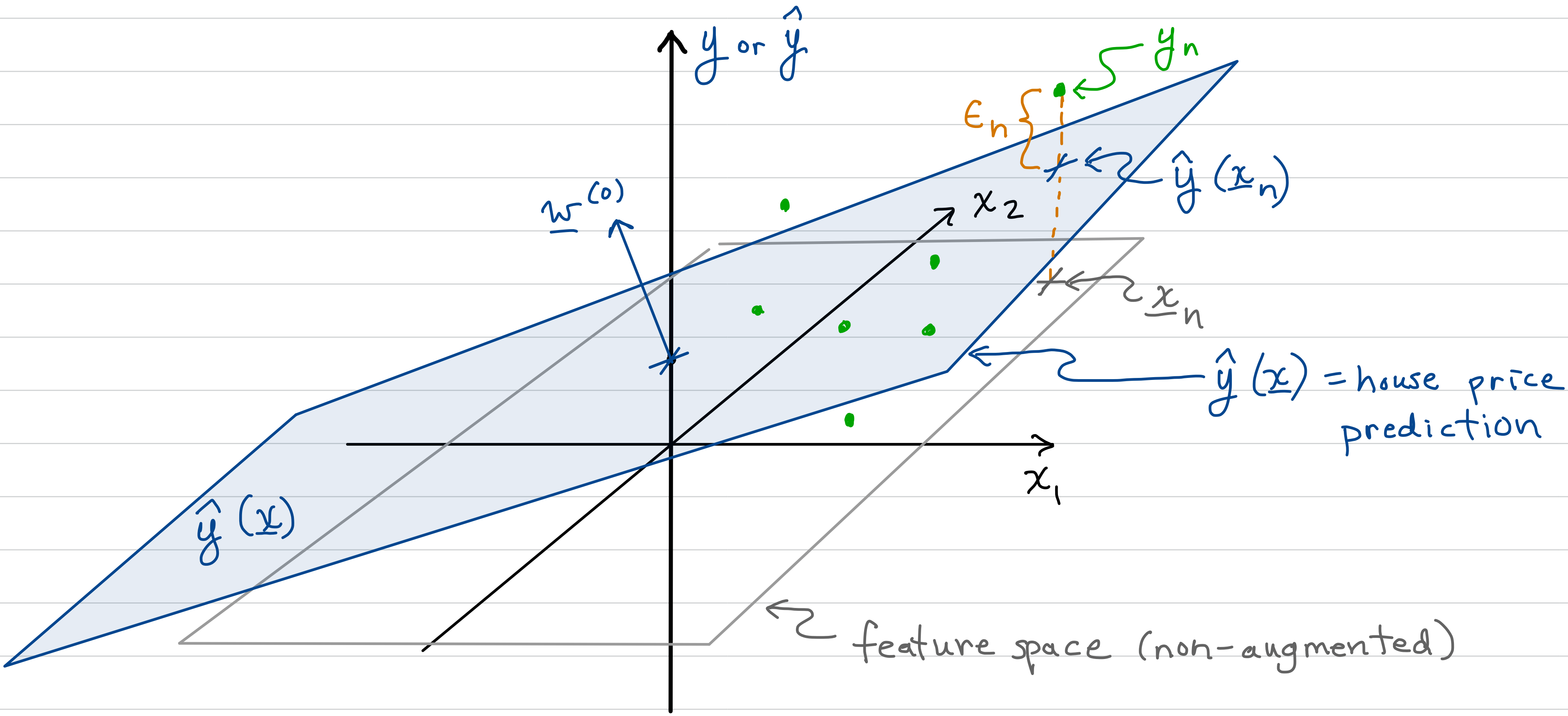
2. $\therefore$ Mean-squared error on dataset $\mathcal{D}$:

$$J(\hat{\underline{w}}) = \frac{1}{N} \| \underline{y} - \hat{\underline{y}} \|_2^2 = \frac{1}{N} \| \underline{y} - \underline{\underline{X}}_{\mathcal{D}} \underline{\underline{X}}^- \underline{y} \|_2^2 = \frac{1}{N} \| (\underline{\underline{I}} - \underline{\underline{X}}_{\mathcal{D}} \underline{\underline{X}}^-) \underline{y} \|_2^2$$

MSE Regression: plot in feature space

Criterion function: $J(\underline{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{y}(x_i) - y_i\right]^2 = MSE$

Linear model: $\hat{y}(x_i) = \underline{w}^T \underline{x}_i \quad \Rightarrow J(\underline{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[\underline{w}^T\underline{x}_i - y_i\right]^2$



$$J_n(\underline{w}) = \frac{1}{N}\epsilon_n^2, \qquad J(\underline{w}) = \sum_{n=1}^{N} J_n(\underline{w})$$

MSE Regression: Least-Mean-Squares (LMS)

→ Sequential GD (or stochastic GD) optimization of MSE criterion.

$$J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}\left[\hat{y}(\underline{x}_n) - y_n\right]^2 = \frac{1}{N}\sum_{n=1}^{N}\left[\underline{w}^T\underline{x}_n - y_n\right]^2 = \sum_{n=1}^{N} J_n(\underline{w})$$

$$J_n(\underline{w}) = \frac{1}{N}\left[\underline{w}^T\underline{x}_n - y_n\right]^2$$

$$\nabla_{\underline{w}} J_n(\underline{w}) = \frac{2}{N}\left[\underline{w}^T\underline{x}_n - y_n\right]\underline{x}_n$$

Sequential GD:

$$\underline{w}(i+1) = \underline{w}(i) - \eta'(i)\frac{2}{N}\left(\underline{w}(i)^T\underline{x}_n - y_n\right)\underline{x}_n, \quad n = (i \bmod N) + 1$$

$(i: 0, 1, 2, \cdots;\ n: 1, 2, \cdots, N)$

Let $\eta(i) = \frac{2}{N}\eta'(i)$

$$\boxed{\underline{w}(i+1) = \underline{w}(i) - \eta(i)\left(\underline{w}(i)^T\underline{x}_n - y_n\right)\underline{x}_n, \quad n = (i \bmod N) + 1}$$

$(i: 0, 1, 2, \cdots;\ n: 1, 2, \cdots, N)$

LMS algorithm weight update

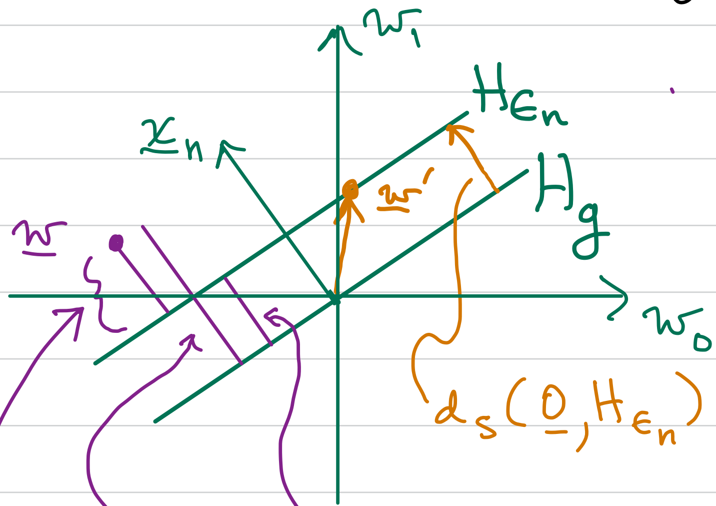
$(\underline{w}^T\underline{x}_n - y_n)$ is a signed error measure: $(\hat{y}(\underline{x}_n) - y_n)$; $\underline{x}_n$ gives its direction.

Plot in weight space: $\epsilon_n = (\underline{w}^T \underline{x}_n - y_n) = 0 \Rightarrow 0$ error for data point $\underline{x}_n$.

S H_g: $\underline{w}^T \underline{x}_n = 0$ is hyperplane: $H_g \perp \underline{x}_n$? H_g passes through origin?

What set of points is $\epsilon_n = 0$?

H_{ϵ_n}: $\epsilon_n = \underline{w}^T \underline{x}_n - y_n = 0$ is hyperplane $\perp$ $\underline{x}_n$, offset from origin by $\dfrac{y_n}{\|\underline{x}_n\|}$



Let $\underline{w}'$ be any $\underline{w}$ on H_{ϵ_n}

Let $d_s(\underline{0}, H_{\epsilon_n}) \triangleq$ signed distance between $\underline{0}$ and H_{ϵ_n}.

$$= \underline{w}'^T \frac{\underline{x}_n}{\|\underline{x}_n\|} = \underbrace{\frac{y_n}{\|\underline{x}_n\|} = d_s(\underline{0}, H_{\epsilon_n})}$$

and $d_s(\underline{0}, H_{\epsilon_n}) > 0$ iff $\underline{w}'^T \underline{x}_n > 0$

$$d_s(H_{\epsilon_n}, \underline{w}) = d_s(H_g, \underline{w}) - d_s(H_g, H_{\epsilon_n})$$

$$d_s(H_{\epsilon_n}, \underline{w}) = \underline{w}^T \frac{\underline{x}_n}{\|\underline{x}_n\|} - \frac{y_n}{\|\underline{x}_n\|} = \frac{\epsilon_n}{\|\underline{x}_n\|}$$

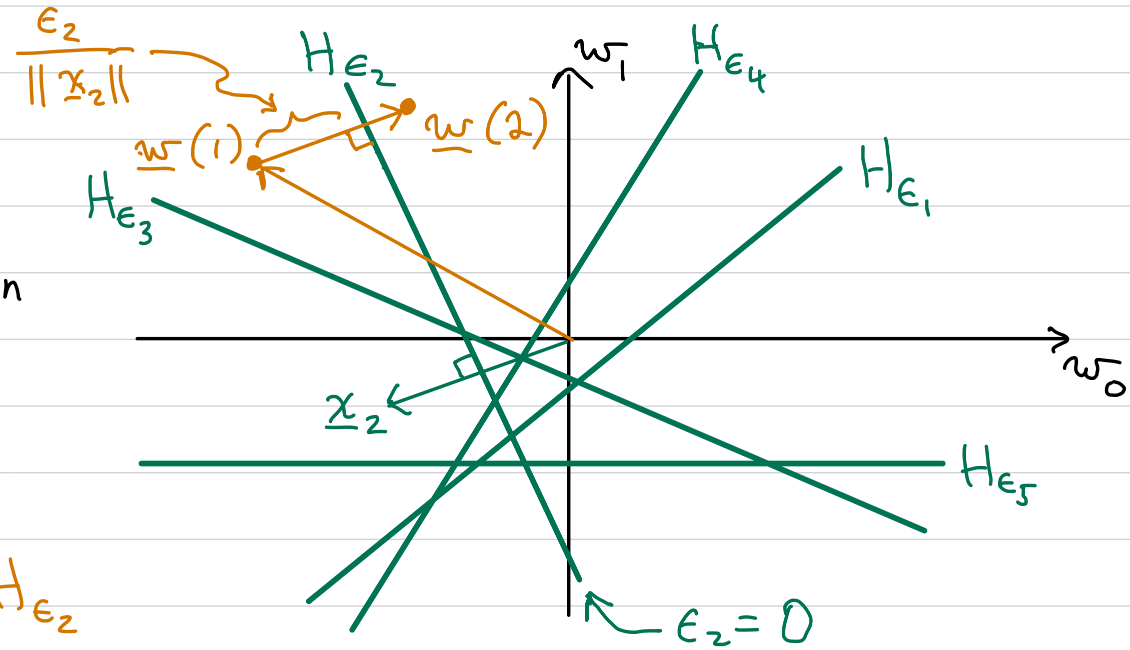
$\boxed{H_{\epsilon_n} \text{ is target hyperplane for data point } \underline{x}_n.}$

LMS example (1 weight update)

$N = 5$

$$\underline{w}(i+1) = \underline{w}(i) - \eta(i)\left(\underline{w}(i)^T \underline{x}_n - y_n\right)\underline{x}_n$$

$$\underline{w}(2) = \underline{w}(1) - \eta(1)\underbrace{\left(\underline{w}(1)^T\underline{x}_2 - y_2\right)}_{\epsilon_2}\underbrace{\underline{x}_2}_{\underline{x}_2 \perp H_{\epsilon_2}}$$



Note: if instead we had:

$$\underline{w}(2)-\underline{w}(1) = -\underbrace{\frac{\left(\underline{w}(1)^T\underline{x}_2 - y_2\right)}{\|\underline{x}_2\|}}_{\frac{\epsilon_2}{\|\underline{x}_2\|} = d_s(H_{\epsilon_2}, \underline{w}(1))}\frac{\underline{x}_2}{\|\underline{x}_2\|}$$

then $\underline{w}(2) \Rightarrow 0$ error on $\underline{x}_2$

Mean-squared-error techniques: Classification (augmented notation)

(2-class problems, reflected data points)

Can use above (MSE regression) techniques, except need a target output y_n.

Could choose: $y_n = z_n = \begin{cases} +1, & \underline{x}_n \in S_1 \\ -1, & \underline{x}_n \in S_2 \end{cases}$ ($\Leftarrow \underline{x}_n, y_n$ unreflected)

but can be more general: use b_n as a "target value", specified by the user.

(1) $\therefore$ $$\boxed{J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}\left[g(z_n \underline{x}_n) - b_n\right]^2 = \frac{1}{N}\sum_{n=1}^{N}\left[\underline{w}^T z_n \underline{x}_n - b_n\right]^2, \quad b_n > 0 \ \forall n.}$$

Let H_B = decision boundary:

$$g(z\underline{x}) = \underline{w}^T z \underline{x} = 0$$

Let $b_n = b \ \forall n$ for visualization

Let H_T = target hyperplane

$\underline{w}^T z_n \underline{x}_n - b = 0$