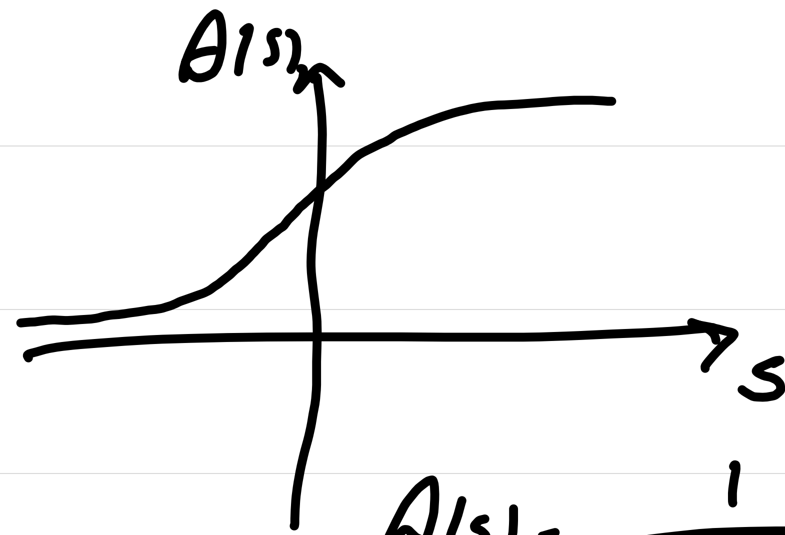
Recap : Logistic Regression



$$\theta(s) = \frac{1}{1+e^{-s}}$$

$$\hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) = \theta(y_n \underline{w}^T \underline{x}_n)$$

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) = \log\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$$

Today: We will see why we use this loss function ?
(Mathematically speaking)

We want to predict with randomness

$\mathcal{H}: \hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

We saw why it makes sense

We need a $E_{in}(\underline{w})$

We use log-loss

$e_n(w) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

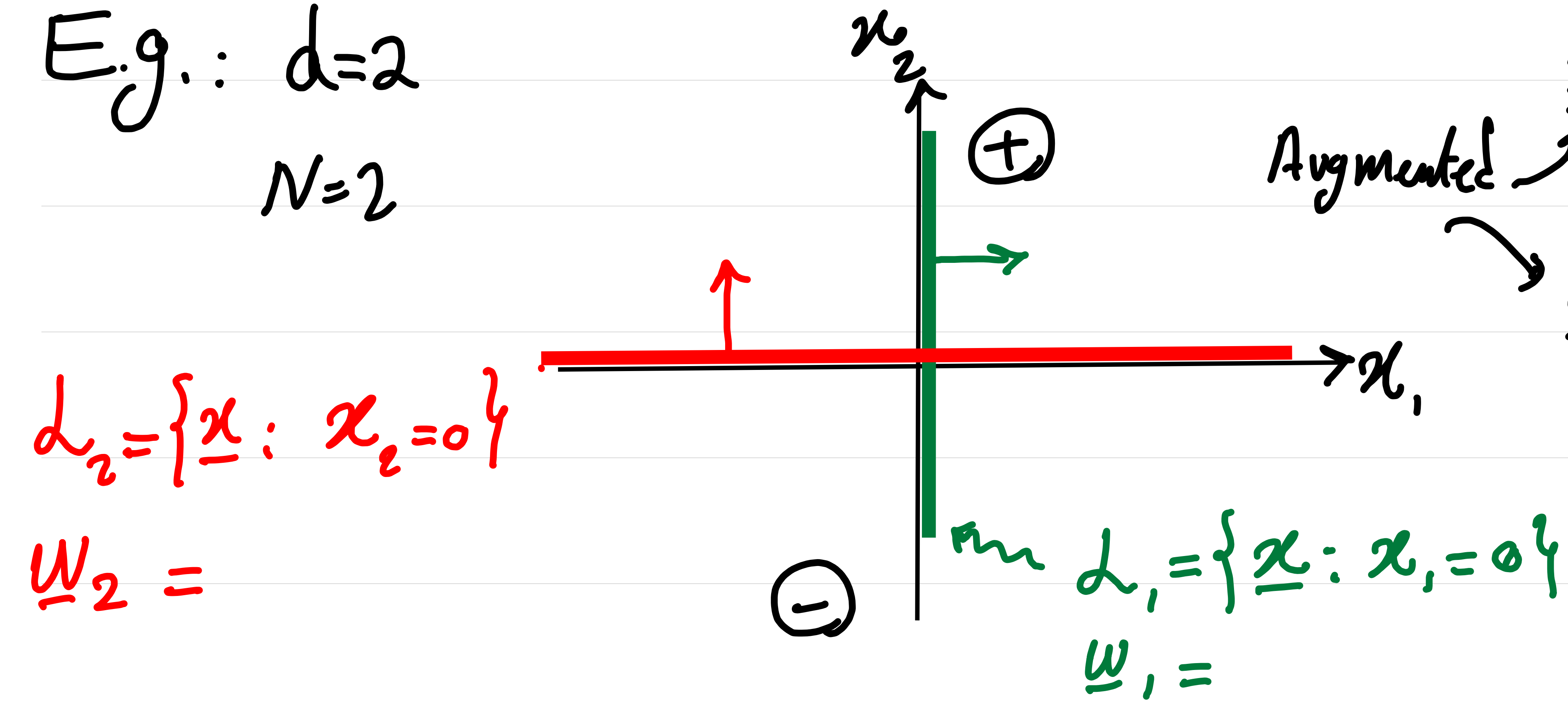
Today: why does log-loss make sense?

It can distinguish better different hyperplanes: Numerical examples

Benefits Over Linear Classification

E.g.: $d=2$

$N=2$



Augmented

$\underline{x}_1 = (1, 0.001, 10), \; y_1 = +1$

$\underline{x}_2 = (1, -0.001, -10), \; y_2 = -1$

- For linear classification, which line is better?

  $\left(e_n(\underline{w}) = \mathbb{1}(y_n \neq \text{sign}(\underline{w}^T \underline{x}_n)) \right)$: They are the same

  $E_{in}(\underline{w}_1) = E_{in}(\underline{w}_2) = 0$

- Intuitively, which line is better?

For logistic regression, which line is better?

$$E_{in}(\underline{w}_1) = \frac{1}{2}\left[\log\left(1+e^{-y_1 \underline{w}_1^T \underline{x}_1}\right) + \log\left(1+e^{-y_2 \underline{w}_1^T \underline{x}_2}\right)\right]$$

$$= \frac{1}{2}\left[\log\left(\quad \right) + \log\left(\quad \right)\right] \approx$$

$$E_{in}(\underline{w}_2) = \frac{1}{2}\left[\log\left(1+e^{-y_1 \underline{w}_2^T \underline{x}_1}\right) + \log\left(1+e^{-y_2 \underline{w}_2^T \underline{x}_2}\right)\right]$$

$$= \frac{1}{2}\left[\log\left(\quad \right) + \log\left(\quad \right)\right] \approx$$

So,

- ☐ ℓ_1 is preferred
- ☐ ℓ_2 is preferred

- What about $\underline{w}_3 = (0, 0, 100)$?
  - This is the same line as ℓ_2.
  - What about $E_{in}(\underline{w}_3)$? $\frac{1}{2}\left(\log(1+e^{-100}) + \log(1+e^{-100})\right)$
    - $E_{in}(\underline{w}_3)$ is much lower!
    - But $\underline{w}_3$ and $\underline{w}_2$ are the same lines.

- We can fix the norm of $\|w\|_2^2$ to be 1.
  - But this constraint makes the optimization more difficult challenging $\rightsquigarrow$ $\min_{\underline{w}} E_{in}(\underline{w})$ s.t. $\|\underline{w}\| = 1$

- Typically, we regularize logistic regression

$$\min_{\underline{w}} E_{in}(\underline{w}) + \lambda \|w\|_2^2$$

- So far, we have seen numerical demostration why log-loss makes sense.
- Time for more rigorous math.
  - There are two mathematical explanation for log-loss
    1- When we minimize log-loss, we maximize the likelihood
    2- when we minimize log-loss, we minimize cross-entropy

We want to predict with randomness

$\mathcal{H}: \hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

We saw why it makes sense

We need a $E_{in}(\underline{w})$

We use log-loss

$e_n(w) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Today: why does log-loss make sense?

It can distinguish better different hyperplanes: Numerical examples

Maximum likelihood interpretation: when we minimize log-loss, we maximize likelihood

Maximum Likelihood Interpretation:

- Let $\mathcal{D} = \{(\underline{x}_1, y_1), \ldots, (\underline{x}_N, y_N)\}$ be the observed datapoint.

- Consider $P(y_1, y_2, \ldots, y_N \mid \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N)$

$= \mathbb{P}[1^{st}$ label is y_1, 2^{nd} label is $y_2, \ldots \mid 1^{st}$ example $= \underline{x}_1$, 2^{nd} example $= \underline{x}_2, \ldots]$

$=$ "likelihood of observing this particular labels $y_1, \ldots, y_N$ given datapoint $\underline{x}_1, \ldots, \underline{x}_N$"

- This is the joint distribution.
- Assuming having I.I.D. examples, independent and identically distributed

$$P(y_1, y_2, \dots, y_N \mid \underline{x}_1, \underline{x}_2, \dots, \underline{x}_N) = \prod_{n=1}^{N} \underbrace{P}_{\text{unknow target probability distribution}}(y_n \mid \underline{x}_N)$$

(we want a $\hat{P}_{\underline{w}} \in \mathcal{H}$ s.t.)

$$\approx \prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

We want to find $\underline{w}$ that maximizes $\prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$ likelihood

$$\Longleftrightarrow \underset{\underline{w}}{\text{maximize}} \ \frac{1}{N} \log \prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

$$\Longleftrightarrow \max_{\underline{w}} \ \frac{1}{N} \sum_{n=1}^{N} \log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) \Longleftrightarrow \min_{\underline{w}} \ \frac{1}{N} \sum_{n=1}^{N} \overbrace{-\log (y_n \mid \underline{x}_n)}^{e_n(\underline{w})}$$

$$\Longleftrightarrow \min_{\underline{w}} \ E_{in}(\underline{w})$$

We want to predict with randomness

$\mathcal{H}$: $\hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

We saw why it makes sense

We need a $E_{in}(\underline{w})$

We use log-loss

$e_n(w) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Today: why does log-loss make sense?

It can distinguish better different hyperplanes: Numerical examples

Maximum likelihood interpretation: when we minimize log-loss, we maximize likelihood

Cross-entropy interpretation: When we minimize log-loss, we minimize cross-entropy

Cross-Entropy Interpretation

Defn: Suppose P and Q are to distribution over the same sample space S

Sample space S: The set of all values a r.v. can take

E.g.: Poisson R.V. with mean:
$S = \{0, 1, \dots\}$
distribution: $P(K) = \frac{\lambda^K}{K!} e^{-\lambda}$

The cross-entropy b/w P and Q is

$$CE(P, Q) = -\sum_{K \in S} P(K) \log Q(K)$$

(measures the difference b/w P and Q)

- For the n^{th} example, Consider the "true" distribution

$$P_n = \left(\mathbb{P}[y_n = +1] ,\ \mathbb{P}[y_n = -1] \right)$$ (distribution of the true labels)

$$= \begin{cases} (1, 0), & \text{if } y_n = 1 \\ (0, 1), & \text{if } y_n = -1 \end{cases}$$

$$= \left(\mathbb{1}(y_n = +1) ,\ \mathbb{1}(y_n = -1) \right)$$

- For n^{th} example, Consider the estimated distribution

$$Q_n = \left(\hat{P}_{\underline{w}}(1 \mid \underline{x}_n),\ \hat{P}_{\underline{w}}(-1 \mid \underline{x}_n) \right)$$

"estimated distribution of y_n given example $\underline{x}_n$"

- The "closer" Q_n's are to P_n's, the better it is.
- what does "closeness" mean? Cross-Entropy
  - $$CE(P_n, Q_n) = -\left(\mathbb{1}(y_n=+1)\, \hat{P}_{\underline{w}}(1 \mid \underline{x}_n) + \mathbb{1}(y_n=-1)\, \hat{P}_{\underline{w}}(-1 \mid \underline{x}_n) \right)$$

$$= -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) = e_n(\underline{w})$$

min $E_{in}(\underline{w}) \Longleftrightarrow$ min average "distance" b/w P_n's and Q_n's.

$$\frac{1}{N} \sum_{n=1}^{N} CE(P_n, Q_n)$$

- Hopefully we are all convinced that log-loss is a reasonable loss function

- But, what is our learning algorithm to find the $\underline{w}$ that minimizes $E_{in}(\underline{w})$

    ↳ Next time. Gradient descent.