COS234:

Introduction To Machine Learning

Prof. Yoram Singer

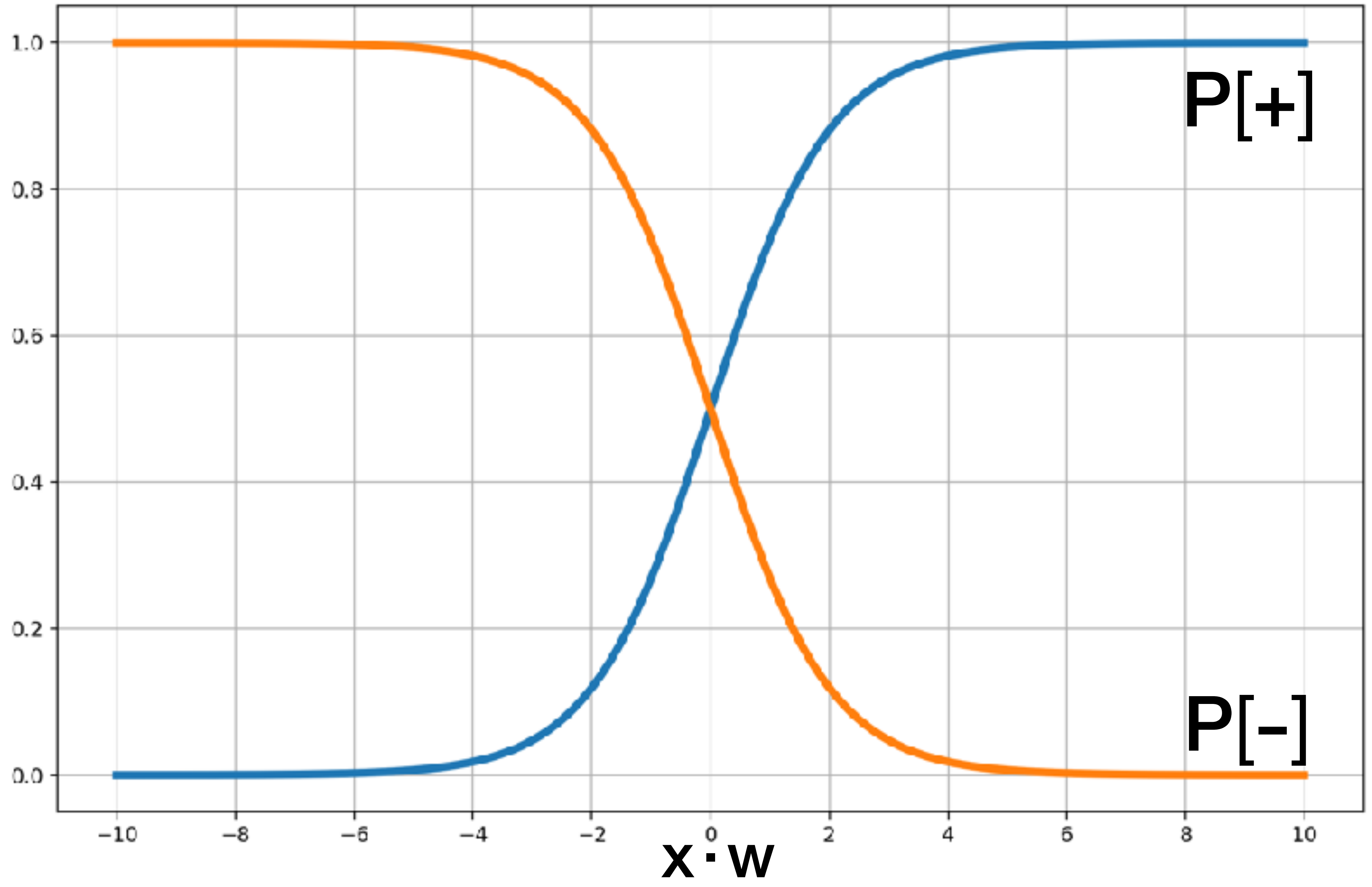
PRINCETON UNIVERSITY

Topic: Logistic Regression

Logistic Regression

- Given **x** "probability" of y to be +1: $\mathbf{P}\left[+1 \mid \mathbf{x}; \mathbf{w}\right] = \dfrac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}$

- Probability of y to be -1: $\mathbf{P}\left[-1 \mid \mathbf{x}; \mathbf{w}\right] = 1 - \dfrac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}} = \dfrac{1}{1 + e^{\mathbf{w}\cdot\mathbf{x}}}$



Logistic Regression

- Given **x** "probability" of y to be +1: $\mathbf{P}\left[+\,|\,\mathbf{x};\mathbf{w}\right] = \dfrac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}$

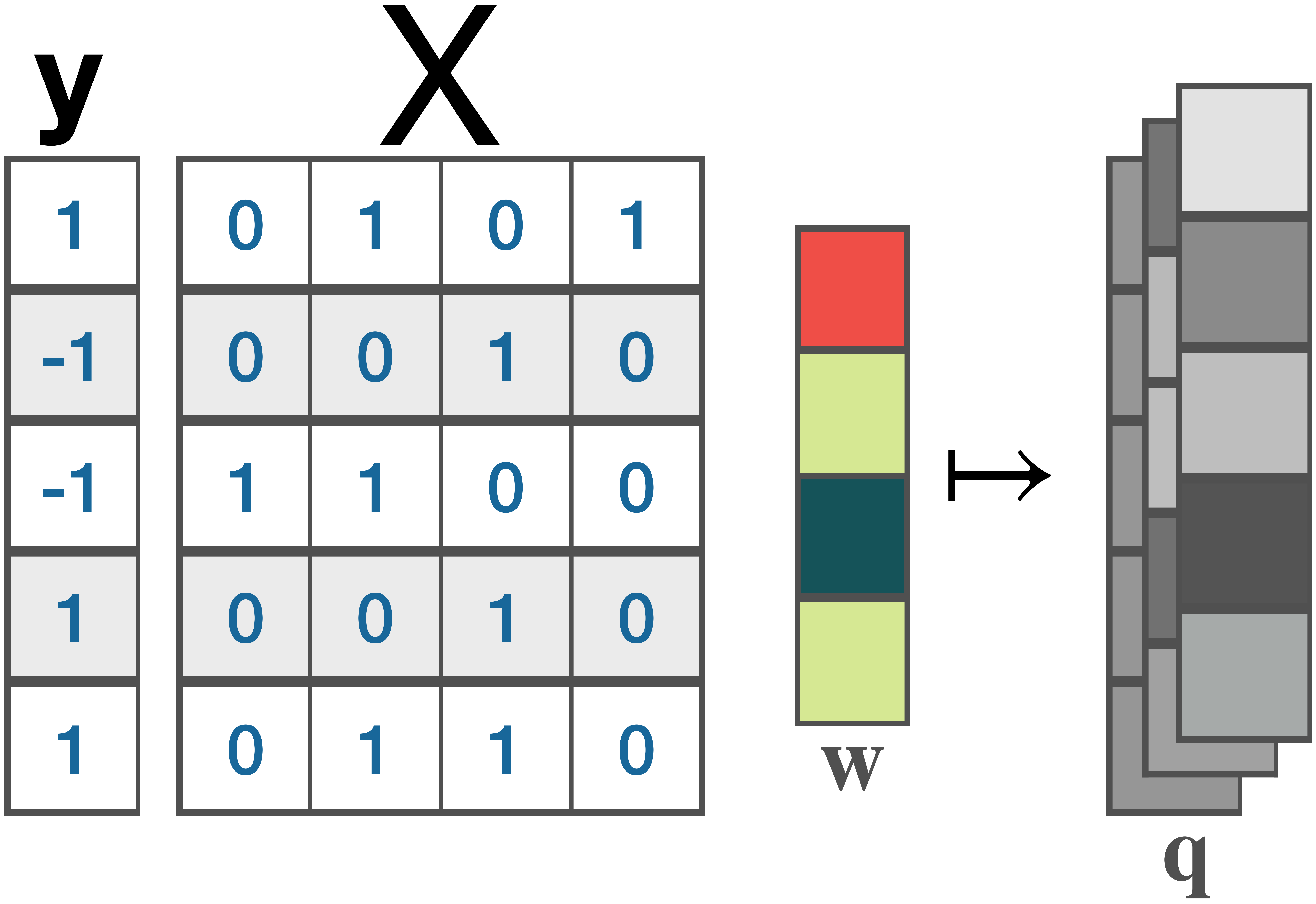
- Probability of y to be -1: $\mathbf{P}\left[-\,|\,\mathbf{x};\mathbf{w}\right] = 1 - \dfrac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} = \dfrac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}$

- Combine two cases: $\mathbf{P}\left[\mathrm{y}\,|\,\mathbf{x};\mathbf{w}\right] = \dfrac{1}{1+e^{-\mathrm{y}\,\mathbf{w}\cdot\mathbf{x}}} = \dfrac{1}{1+e^{-\mathrm{z}}}$

  Predict +1 w.p. P[+|**x**] and -1 w.p. P[-|**x**]

- Define loss to be negative of log-probability (log-likelihood):

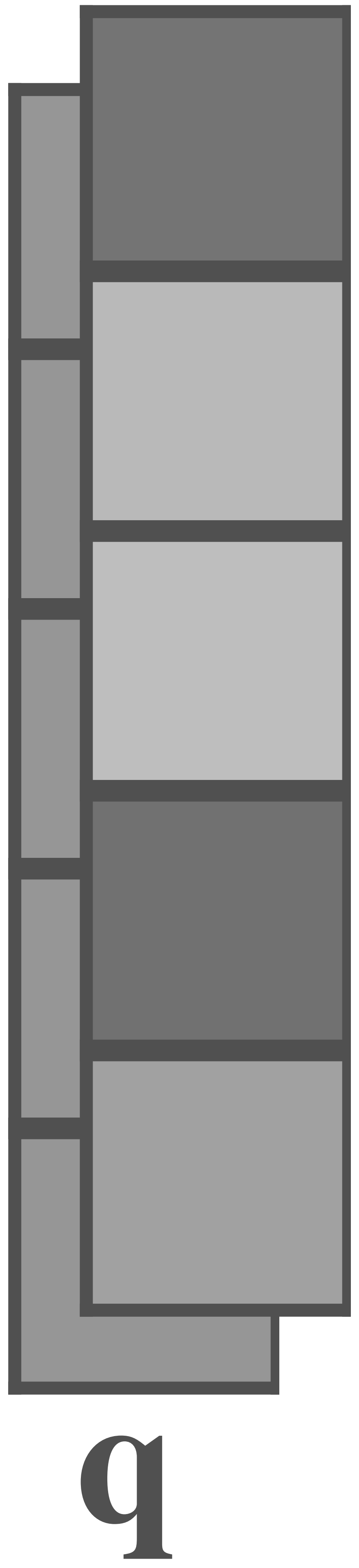
$$-\log\left(\mathbf{P}\left[\mathrm{y}\,|\,\mathbf{x};\mathbf{w}\right]\right) = \log\left(1+\mathrm{e}^{-\mathrm{z}}\right)$$


y
X
1 0 1 0 1
-1 0 0 1 0
-1 1 1 0 0
1 0 0 1 0
1 0 1 1 0
w
↦
q

- We want to find **w** so as to (approximately) minimize

$$\mathscr{L}(\mathbf{w}) = \sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i(\mathbf{w} \cdot X_{i*})\big)\Big)$$

- Initialize **w** = **0** = (0,0, … ,0)
- Initialize **q** = (0.5, 0.5, … , 0.5)
- Loop:
  - Pick an index j at random from {1,2, … ,d}
  - Replace $\mathbf{w}_j$ with a <u>better</u> estimate $\mathbf{w}_j^{new}$
  - Update **q**



Importance Weights: q

- **q** reflects how well **w** predicts the labels (y_1, ..., y_n)
  - $q_i \in (0, 1)$ large: poor prediction by **w** of y_i
  - $q_i \in (0, 1)$ small: good prediction by **w** of y_i
- after update of **w** $\mapsto$ **q** is updated

$$w_j \mapsto w_j^{new}$$

Picked index j from {1,2, … ,d}

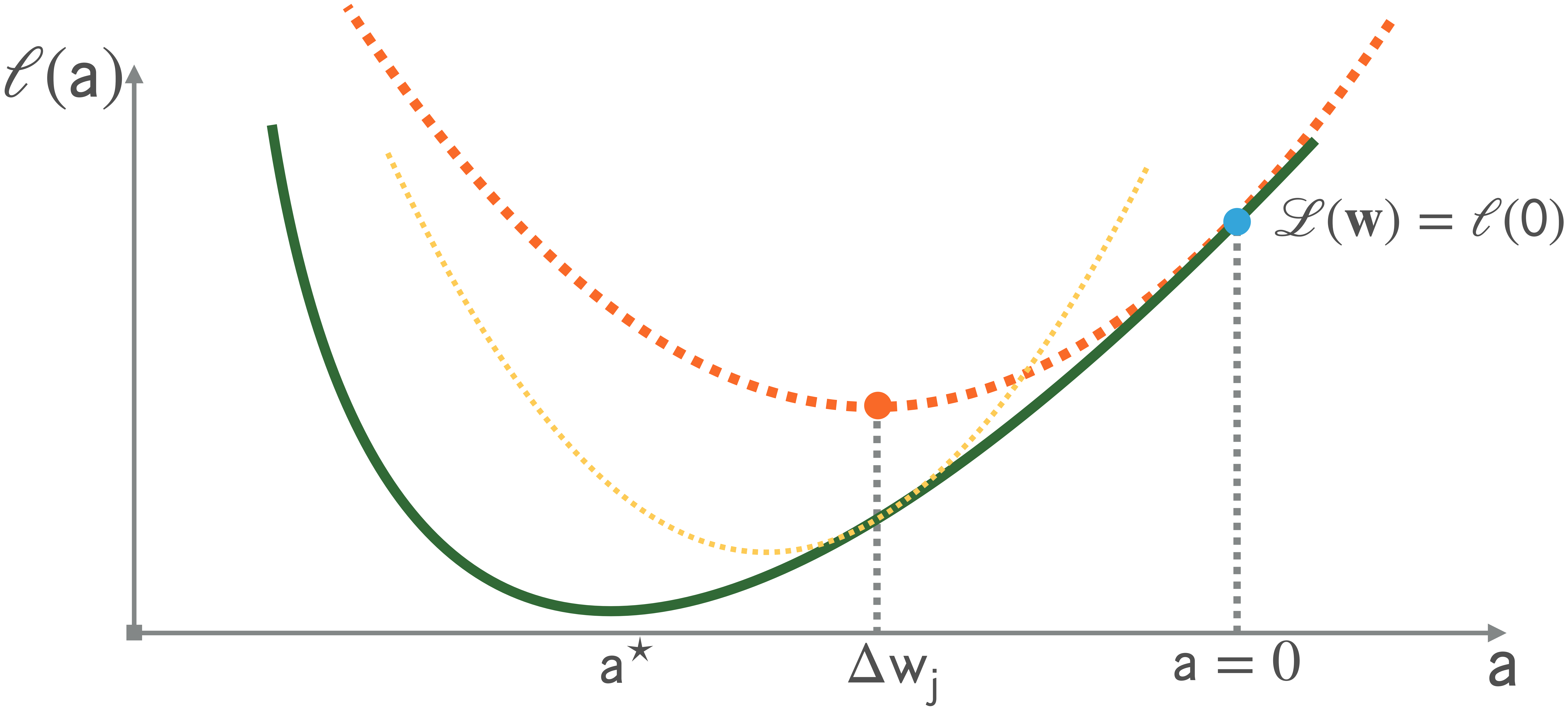
Calculate: $r_+ = \sum_{y_i X_{i,j}=+1} q_i \qquad r_- = \sum_{y_i X_{i,j}=-1} q_i \qquad n_j = \sum_i X_{ij}$

Calculate & update **w**: $\Delta w_j = 4\left(\frac{r_+ - r_-}{n_j}\right) \quad \mapsto \quad w_j = w_j + \Delta w_j$

For all i s.t. $X_{ij}=1$ $\qquad q_i \leftarrow \frac{q_i}{q_i + (1 - q_i)\, e^{y_i \Delta w_j}}$

Steps

- Write loss as a function of $\mathbf{a} = \Delta \mathrm{w}_j$
- Simplify $\ell(\mathrm{a})$ using one auxiliary variable per example
- Calculate $\dfrac{d\,\ell(\mathrm{a})}{d\mathrm{a}}$ and $\dfrac{d^2\ell(\mathrm{a})}{d\mathrm{a}^2}$
- Use MVT: $\ell(\mathrm{a}) = \ell(0) + \ell'(0)\,\mathrm{a} + \dfrac{1}{2}\ell''(\alpha)\,\mathrm{a}^2$ with $\alpha \in [0, \mathrm{a}]$
- Bound $\ell''(\alpha) \leq \mathrm{val}(X_{\star j}) \;\Rightarrow\; \ell(\mathrm{a}) \leq \ell(0) + \ell'(0)\,\mathrm{a} + \mathrm{val}(X_{\star j})\,\mathrm{a}^2$
- Set $\mathbf{a} \sim \dfrac{\mathrm{val}(X_{\star j})}{\ell'(0)}$



$\ell(a)$
$\mathcal{L}(\mathbf{w}) = \ell(0)$
$a^\star$
Δw_j
$a = 0$
a

Derivation

- Fix all but coordinate j of w
- Write $\ell(a) = \mathscr{L}\left(w_1, \ldots, w_{j-1}, w_j + a, w_{j+1}, \ldots, w_d\right)$
- Omitting 1/n

$$\ell(a) = \sum_i \log\left(1 + \exp\left(-y_i\left(\mathbf{w} \cdot X_{i\star} + aX_{ij}\right)\right)\right)$$

- Denote

$$b_i = -y_i\left(\mathbf{w} \cdot X_{i\star} + aX_{ij}\right)$$

- Hence

$$\ell(a) = \sum_i \log\left(1 + \exp\left(b_i\right)\right)$$

- Chain rule:

$$\frac{d}{db}\left[\log\left(1+\exp(b)\right)\right] = \frac{\exp(b)}{1+\exp(b)} = \frac{1}{1+\exp(-b)}$$

- For $b_i = -\,y_i\left(\mathbf{w}\cdot X_{i\star} + aX_{ij}\right)$

$$\frac{d\,\ell(a)}{da} = \sum_i \frac{1}{1+\exp\left(-\,b_i\right)}\frac{db_i}{da}$$

- Since $\dfrac{db_i}{da} = -\,y_i X_{ij}$

$$\frac{d\,\ell(a)}{da} = -\sum_i \frac{y_i X_{ij}}{1+\exp\left(-\,b_i\right)}$$

- Apply chain rule to

$$\frac{d^2\,\ell(a)}{da^2} = \frac{d}{da}\left[\frac{d\,\ell(a)}{da}\right]$$

- Using again $b_i = -\,y_i\left(\mathbf{w}\cdot X_{i\star} + aX_{ij}\right)$

$$\frac{d^2\,\ell(a)}{da^2} = \sum_i \frac{y_i^2\,X_{ij}^2\,\exp(-b_i)}{\left(1+\exp(-b_i)\right)^2} = \sum_i \frac{X_{ij}\,\exp(-b_i)}{\left(1+\exp(-b_i)\right)^2}$$

- Or alternatively

$$\frac{d^2\,\ell(a)}{da^2} = \sum_i \frac{X_{ij}}{\left(1+\exp(-b_i)\right)\left(1+\exp(b_i)\right)}$$

Mean Value Theorem

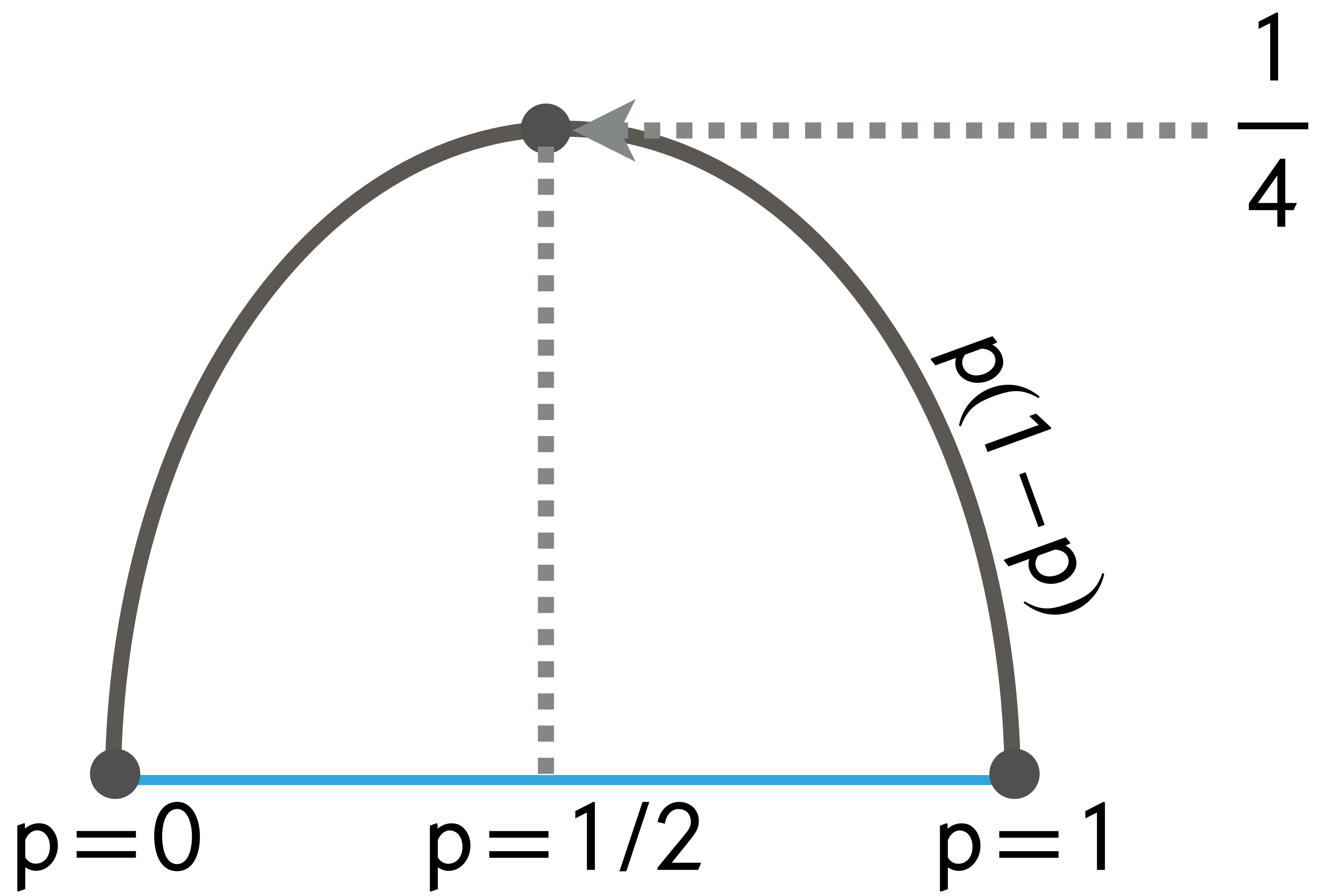
- There exists $\alpha \in [0, a]$ such that $\ell(a) = \ell(0) + \ell'(0)a + \frac{1}{2}\ell''(\alpha)\, a^2$

- Denote $z_i = y_i \left(\mathbf{w} \cdot X_{i\star}\right)$ and $c_i = y_i \left(\mathbf{w} \cdot X_{i\star} + \alpha X_{ij}\right)$

$$\ell(a) = \ell(0) - a \sum_i \frac{y_i X_{ij}}{1 + \exp\left(z_i\right)} + \frac{a^2}{2} \sum_i \frac{X_{ij}}{\left(1 + \exp(-c_i)\right)\left(1 + \exp(c_i)\right)}$$

- Recall that $\dfrac{1}{1+\exp(-c_i)} + \dfrac{1}{1+\exp(c_i)} = 1$

- Call $p = \dfrac{1}{1+\exp(-c_i)}$ and note that $p\,(1-p) \leq \dfrac{1}{4}$



Putting It All Together

- Using MVT and the expression for the loss

$$\ell(a) \leq \ell(0) - a \sum_i \frac{y_i X_{ij}}{1 + \exp(z_i)} + \frac{a^2}{8} \sum_i X_{ij}$$

- Define

$$q_i = \frac{1}{1 + \exp(z_i)} \quad r_+ = \sum_{i:y_i=1} q_i X_{ij} \quad r_- = \sum_{i:y_i=-1} q_i X_{ij} \quad n_j = \sum_i X_{ij}$$

- Use definitions

$$\ell(a) \leq \ell(0) - a\,(r_+ - r_-) + \frac{a^2}{8} n_j$$

- Use the upper bound on $\ell(a)$

$$\frac{d}{da}\left[\ell(0) - a\,(r_+ - r_-) + \frac{a^2}{8}\,n_j\right] = 0$$

- Finally we get

$$a = 4\left(\frac{r_+ - r_-}{n_j}\right)$$

- Update

$$w_j \leftarrow w_j + a \qquad q_i \leftarrow \frac{1}{1 + \exp(z_i + a y_i X_{ij})}$$

- Initialize: $\mathbf{w} = \mathbf{0} = (0, 0, \ldots, 0)$ $\quad \mathbf{q} = \left(1/2, 1/2, \ldots, 1/2\right)$
- Calculate for all j: $n_j = \sum_i X_{ij}$
- Loop:
  - Pick j at random from $\{1, 2, \ldots, d\}$
  - Calculate $\quad r_+ = \sum_{y_i X_{i,j}=+1} q_i \qquad r_- = \sum_{y_i X_{i,j}=-1} q_i$
  - Update $\quad w_j \leftarrow w_j + 4\left(\frac{r_+ - r_-}{n_j}\right)$
  - Update for all i s.t. $X_{ij}=1$: $\quad q_i \leftarrow \dfrac{q_i}{q_i + (1 - q_i)\, e^{y_i \Delta w_j}}$