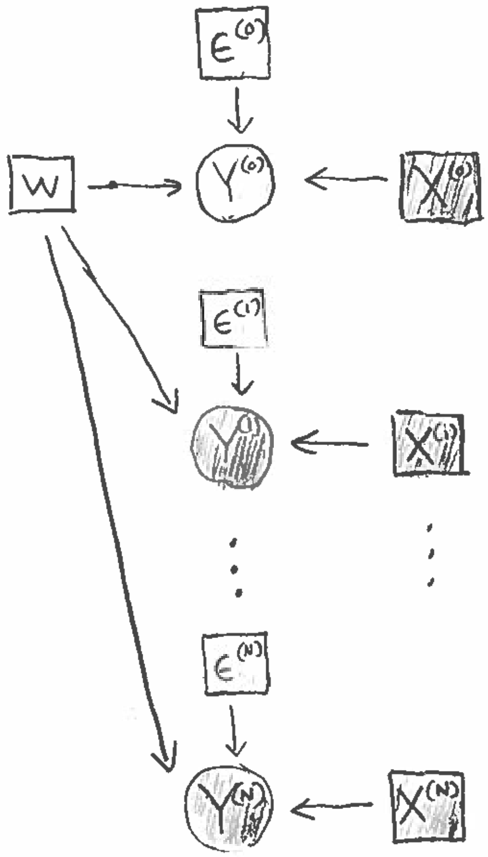
LOGISTIC REGRESSION: MLE

① Recall "logistic regression":



where: $\epsilon^{(n)} \sim \text{Constant}(0,1) \quad \forall n \in \{0, \dots, N\}$

$$y^{(n)} \leftarrow 1_{\epsilon^{(n)} < (1+\exp(-x^{(n)}\omega))^{-1}} = \begin{cases} 1 & \text{if } \epsilon^{(n)} < \text{sign}(x^{(n)}\omega) \\ 0 & \text{otherwise} \end{cases}$$

② Also recall that one way to estimate the value of the unobserved response variable $Y^{(0)}$ is through maximum likelihood estimation (MLE):

(a) compute $\hat{\omega} = \underset{\omega}{\text{argmax}} \prod_{n=1}^{N} P(y^{(n)} | \omega, x^{(n)})$

(b) compute $\hat{y}^{(0)} = \underset{y^{(0)}}{\text{argmax}}\ P(y^{(0)} | \hat{\omega}, x^{(0)})$

LOGISTIC REGRESSION: MLE

③ To compute the second step, observe:

$$P\left(Y^{(n)}=1 \mid w, x^{(n)}\right)$$

$$= \int_0^1 P\left(Y^{(n)}=1, \epsilon^{(n)} \mid w, x^{(n)}\right) d\epsilon^{(n)} \qquad \text{[total probability]}$$

$\epsilon^{(n)}$ must be between 0 and 1, b/c $\epsilon^{(n)} \sim \text{Constant}(0,1)$

$$= \int_0^1 P\left(Y^{(n)}=1 \mid w, x^{(n)}, \epsilon^{(n)}\right) P\left(\epsilon^{(n)} \mid w, x^{(n)}\right) d\epsilon^{(n)} \qquad \text{[Chain Rule]}$$

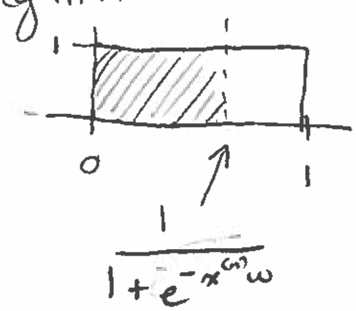
$$= \int_0^1 P\left(Y^{(n)}=1 \mid w, x^{(n)}, \epsilon^{(n)}\right) P\left(\epsilon^{(n)}\right) d\epsilon^{(n)} \qquad \text{[d-separation]}$$

$$= \int_0^1 P\left(\epsilon^{(n)} < \frac{1}{1+e^{-x^{(n)}w}}\right) P\left(\epsilon^{(n)}\right) d\epsilon^{(n)} \qquad \left[\text{b/c } y^{(n)} \leftarrow \mathbb{1}_{\epsilon^{(n)} < (1+\exp(-x^{(n)}w))^{-1}}\right]$$

$$= \int_0^{\frac{1}{1+e^{-x^{(n)}w}}} P\left(\epsilon^{(n)}\right) d\epsilon^{(n)} \qquad \left[\text{everywhere else, } P\left(\epsilon^{(n)} < \frac{1}{1+e^{-x^{(n)}w}}\right) = 0\right]$$

$$= \frac{1}{1+e^{-x^{(n)}w}}$$

[constant distribution integration



]

Thus:

$$P\left(Y^{(n)}=0 \mid w, x^{(n)}\right) = 1 - \frac{1}{1+e^{-x^{(n)}w}} = \frac{1+e^{-x^{(n)}w}}{1+e^{-x^{(n)}w}} - \frac{1}{1+e^{-x^{(n)}w}}$$

$$= \frac{e^{-x^{(n)}w}}{1+e^{-x^{(n)}w}}$$

LOGISTIC REGRESSION: MLE

④ Or, more compactly:

$$P(y^{(n)} \mid w, x^{(n)}) = \frac{e^{-(1-y^{(n)})x^{(n)}w}}{1+e^{-x^{(n)}w}}$$

$$= \begin{cases} \dfrac{e^{-x^{(n)}w}}{1+e^{-x^{(n)}w}} & \text{if } y^{(n)}=0 \\ \dfrac{1}{1+e^{-x^{(n)}w}} & \text{if } y^{(n)}=1 \end{cases}$$

⑤ This allows us to express ②(b) as:

$$\hat{y}^{(0)} = \underset{y^{(0)}}{\operatorname{argmax}} \; P(y^{(0)} \mid w, x^{(0)})$$

$$\boxed{= \underset{y^{(0)} \in \{0,1\}}{\operatorname{argmax}} \; \frac{e^{-(1-y^{(0)})x^{(0)}w}}{1+e^{-x^{(0)}w}}}$$

Logistic Regression: MLE

⑥ To compute 2(a), we start with some simplifications:

$$\hat{w} = \operatorname*{argmax}_w \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$

$$= \operatorname*{argmax}_w \log \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$

$$= \operatorname*{argmax}_w \sum_{n=1}^{N} \log P(y^{(n)} \mid w, x^{(n)})$$

$$= \operatorname*{argmax}_w \sum_{n=1}^{N} \log \frac{e^{-(1-y^{(n)})x^{(n)}w}}{1+e^{-x^{(n)}w}} \qquad \left[\text{from } ④\right]$$

$$= \operatorname*{argmax}_w \sum_{n=1}^{N} \log e^{-(1-y^{(n)})x^{(n)}w} - \log\left(1+e^{-x^{(n)}w}\right)$$

$$= \operatorname*{argmax}_w \sum_{n=1}^{N} -(1-y^{(n)})x^{(n)}w + \log \frac{1}{1+e^{-x^{(n)}w}}$$

$$= \operatorname*{argmin}_w \sum_{n=1}^{N} (1-y^{(n)})x^{(n)}w - \log \frac{1}{1+e^{-x^{(n)}w}}$$

$$= \operatorname*{argmin}_w \sum_{n=1}^{N} (1-y^{(n)})x^{(n)}w - \log \sigma(x^{(n)}w)$$

the logistic sigmoid function

So for logistic regression, our loss function is:

$$L_{\text{logistic}}(w) = \sum_{n=1}^{N} (1-y^{(n)})x^{(n)}w - \log \sigma(x^{(n)}w)$$

LOGISTIC REGRESSION: MLE

⑦ To compute the gradient of $L_{logistic}(w)$, we'll first prove the following lemma:

> **Lemma:** If $\sigma(a) = \dfrac{1}{1+e^{-a}}$, then:
>
> $$\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a))$$

Proof:

$$\frac{d}{da}\sigma(a) = \frac{-1}{(1+e^{-a})^2} \cdot e^{-a} \cdot -1$$

$$= \frac{e^{-a}}{(1+e^{-a})^2}$$

$$= \left(\frac{1}{1+e^{-a}}\right)\left(\frac{e^{-a}}{1+e^{-a}}\right)$$

$$= \sigma(a)\left(\frac{1+e^{-a}-1}{1+e^{-a}}\right)$$

$$= \sigma(a)\left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right)$$

$$= \sigma(a)(1-\sigma(a)) \quad \boxtimes$$

⑧ So the gradient of the loss function is:

$$\frac{d}{dw} L_{logistic}(w)$$

$$= \sum_{n=1}^{N} \frac{d}{dw} (1-y^{(n)}) x^{(n)} w - \frac{d}{dw} \log \sigma(x^{(n)} w)$$

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw} x^{(n)} w - \frac{1}{\sigma(x^{(n)} w)} \frac{d}{dw} \sigma(x^{(n)} w)$$

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw} (x^{(n)} w) - \frac{1}{\sigma(x^{(n)} w)} \sigma(x^{(n)} w)(1-\sigma(x^{(n)} w)) \frac{d}{dw} x^{(n)} w$$

[from Lemma]

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw} (x^{(n)} w) - (1-\sigma(x^{(n)} w)) \frac{d}{dw} (x^{(n)} w)$$

$$= \sum_{n=1}^{N} (1-y^{(n)})(x^{(n)})^T - (1-\sigma(x^{(n)} w))(x^{(n)})^T$$

[b/c $\frac{d}{dw} x \cdot w = x^T$]

$$= \sum_{n=1}^{N} (1-y^{(n)} - 1 + \sigma(x^{(n)} w))(x^{(n)})^T$$

$$\boxed{= \sum_{n=1}^{N} (\sigma(x^{(n)} w) - y^{(n)})(x^{(n)})^T}$$

LOGISTIC REGRESSION: MLE

⑨ This can be expressed even more compactly in terms of the evidence matrix X and response vector y:

$$\frac{d}{dw} L_{logistic}(w) = \sum_{n=1}^{N} \left(\sigma(x^{(n)} w) - y^{(n)}\right)\left(x^{(n)}\right)^T$$

$$= X^T \left(\sigma(Xw) - y\right)$$

Exercise: Show $X^T(\sigma(Xw)-y) = \sum_{n=1}^{N} \left(\sigma(x^{(n)} w) - y^{(n)}\right)\left(x^{(n)}\right)^T$

Logistic Regression: MLE

(10) As usual, there isn't a known way to solve directly for $\frac{d}{dw} L_{logistic}(w) = 0$, however we are free to use gradient descent.

Logistic Regression $(X, y, x^{(0)})$:

- Compute point estimate $\hat{w} = \text{GradDescent}(L_{logistic})$
- Compute prediction $\hat{y}^{(0)} = \underset{y^{(0)} \in \{0,1\}}{\text{argmax}} \dfrac{e^{-(1-y^{(0)})x^{(0)}w}}{1+e^{-x^{(0)}w}}$
- return $\hat{y}^{(0)}$