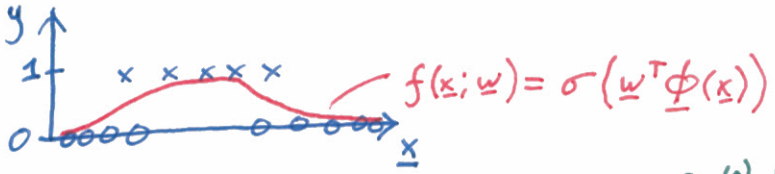
Logistic Regression

for $y \in \{0, 1\}$

Model the outputs: $P(y=1 \mid \underline{x}, \underline{w}) = f(\underline{x}; \underline{w}) = \sigma(\underline{w}^T \underline{x})$



Maximum Likelihood (minimize −ve log likelihood)

$$NLL = -\sum_n \log p(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}) = -\sum_n \log \underbrace{\sigma\left(z^{(n)} \underline{w}^T \underline{x}^{(n)}\right)}_{\sigma_n}$$

where $z^{(n)} = 2y^{(n)} - 1$

$$\nabla_{\underline{w}} NLL = -\sum_n \nabla_{\underline{w}} \log \sigma_n$$

$\frac{d \log a}{da} = \frac{1}{a}$ + chain rule

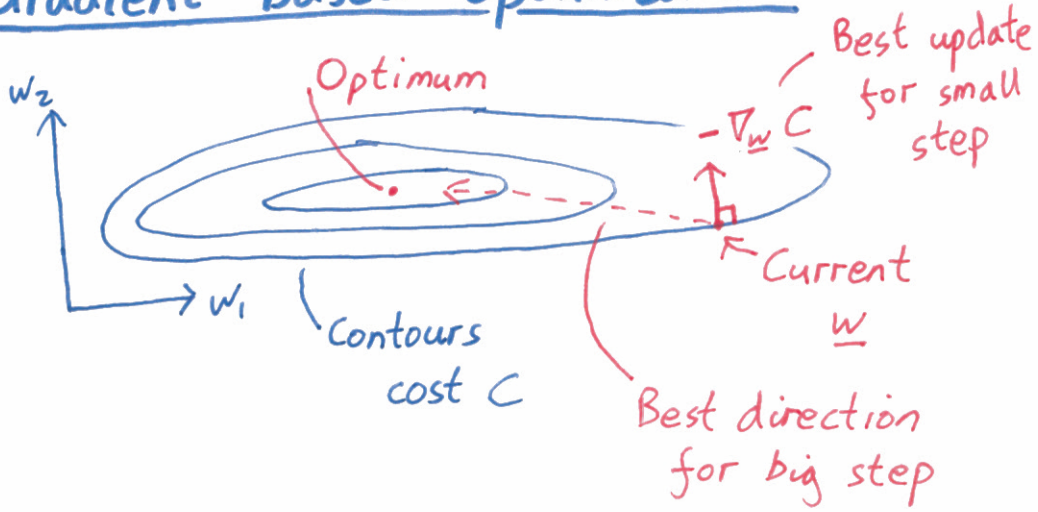
$$= -\sum_n \frac{1}{\sigma_n} \nabla_{\underline{w}} \sigma_n$$

$\frac{d\sigma(a)}{d(a)} = \sigma(a)(1 - \sigma(a))$

$$= -\sum_n \frac{1}{\sigma_n} \sigma_n (1 - \sigma_n) \nabla_{\underline{w}} z^{(n)} \underline{w}^T \underline{x}^{(n)}$$

$$= \underbrace{-\sum_n (1 - \sigma_n) z^{(n)} \underline{x}^{(n)}}_{\sigma(-z^{(n)} \underline{w}^T \underline{x}^{(n)})} = -\sum_n \underbrace{(y^{(n)} - f^{(n)})}_{\sigma(\underline{w}^T \underline{x}^{(n)})} \underline{x}^{(n)}$$

Gradient-based optimization



Finding better directions

- Non-linear conjugate gradients
- L-BFGS
- Newton's method ...

Often "batch" methods:

C and $\nabla_{\underline{w}} C$ use whole data set

Stochastic / Online or "Minibatch" use data subsets

Stochastic Gradient Descent

$-(1-\sigma_n) z^{(n)} \underline{x}^{(n)}$

Average Gradient over examples:

$$\frac{1}{N} \nabla_{\underline{w}} C = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\underline{w}} C_n$$

C_n: Cost for n^{th} example.

Monte Carlo Approximation:

Sample mini-batch of B examples:

$$\approx \frac{1}{B} \sum_{b=1}^{B} \nabla_{\underline{w}} C_b = \hat{\underline{g}}$$

SGD

Initialize $\underline{w} \leftarrow \underline{w}^{(0)}$ (maybe $\underline{0}$)

for $t = 1 \ldots T$:

$$\underline{w} \leftarrow \underline{w} - \eta_t \hat{\underline{g}}^{(t)}$$

η_t: step-size

Softmax Regression

Multi-class classification

Target $\underline{y}^{(n)} = [0\ 0\ 0 \ldots 0\ 1\ 0 \cdots 0]^T$

c^{th} location

If example n has label "$y=c$"

Fit $\underline{f}$, predictions from our model:

$$P(\text{class} = c \mid \underline{x}, W) = f_c(\underline{x}; W)$$

Positive score for each class:

$$s_k = e^{\underline{w}^{(k)T}\underline{x}}$$

Want $\underline{f}$ to be normalized: $\sum_k f_k = 1$

$$f_k = \frac{s_k}{\sum_{k'} s_{k'}} = \frac{e^{\underline{w}^{(k)T}\underline{x}}}{\sum_{k'} e^{\underline{w}^{(k')T}\underline{x}}}$$

Softmax $(W\underline{x})$

Model has parameters W, $K \times D$ ← # features

$$W = \begin{pmatrix} — & \underline{w}^{(1)} & — \\ & \vdots & \\ — & \underline{w}^{(K)} & — \end{pmatrix}$$

Maximize <u>likelihood of $\underline{w}$</u>

For SGD we use one example at $\underline{x}$ with class label c

- log prob. of this example given W

$$-\log f_c = -\underline{w}^{(c)T}\underline{x} + \log \sum_{k'} e^{\underline{w}^{(k')T}\underline{x}}$$

$$-\nabla_{\underline{w}^{(k)}} \log f_c = \underbrace{-\delta_{kc}}_{\text{Kronecker Delta}}\underline{x} + \underbrace{\frac{1}{\sum_{k'} \cdots} e^{\underline{w}^{(k)T}\underline{x}}}_{f_k}\underline{x}$$

$$= -(y_k - f_k)\underline{x}$$

2018 L10 (5)

Logistic Regression

Two classes

$$P(y=1 \mid \underline{x}, W) = \frac{e^{\underline{w}^{(1)T}\underline{x}}}{e^{\underline{w}^{(1)T}\underline{x}} + e^{\underline{w}^{(0)T}\underline{x}}}$$

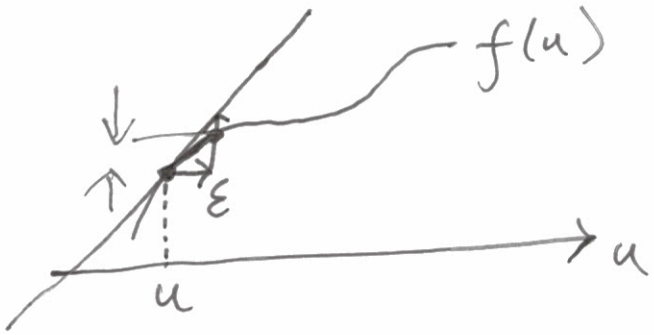
$$= \frac{1}{1 + e^{\underbrace{\underline{w}^{(0)T}\underline{x} - \underline{w}^{(1)T}\underline{x}}_{\text{"}-a\text{"}}}}$$

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

$$= \sigma\Big(\underbrace{\left(\underline{w}^{(1)} - \underline{w}^{(0)}\right)^T}_{\text{"}\underline{w}^T\text{"}}\underline{x}\Big)$$

Parameters are redundant.

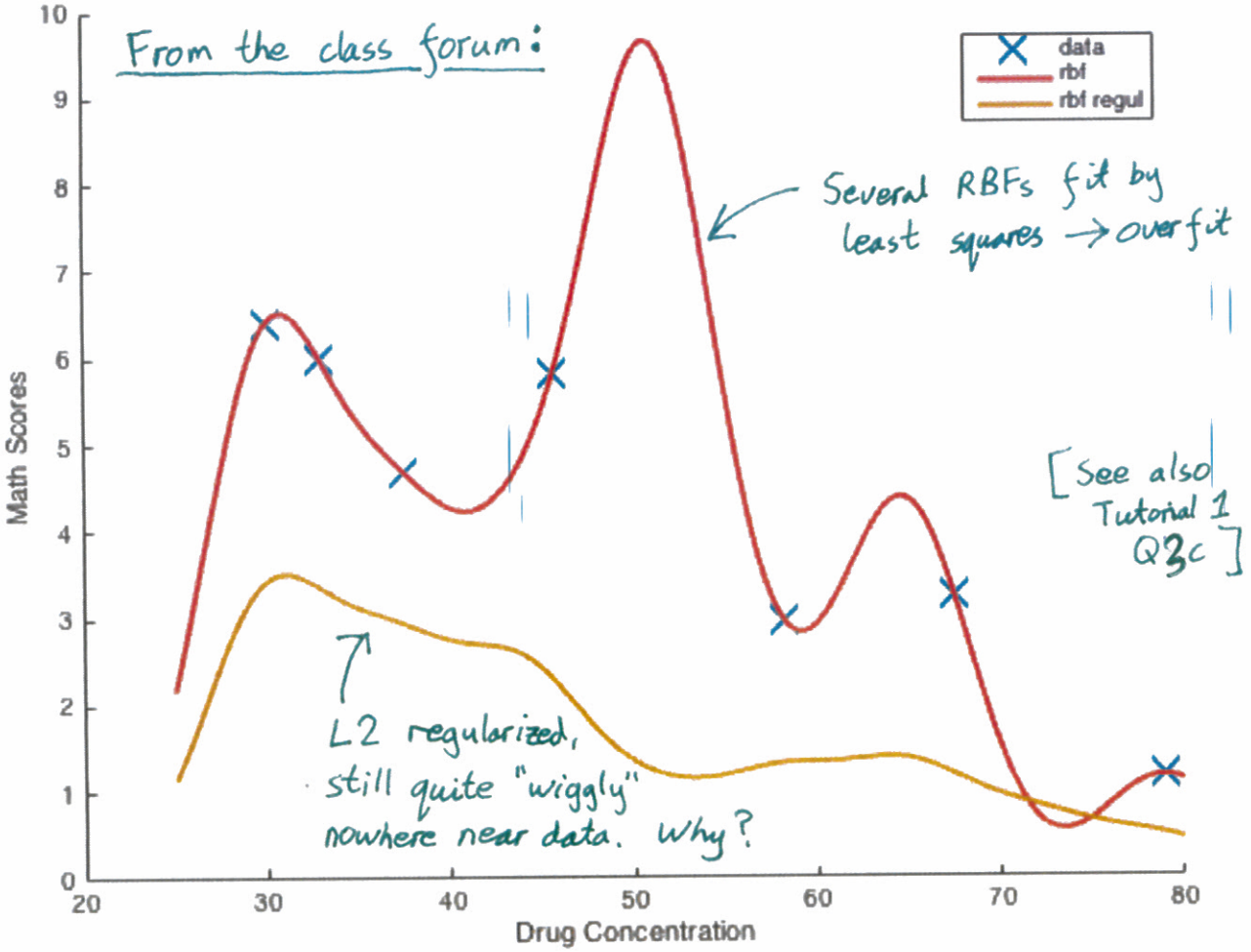
Check your derivatives



$$\left.\frac{df}{du}\right|_u = f'(u) \approx \frac{f(u+\varepsilon) - f(u)}{\varepsilon}$$

correct $\lim \varepsilon \to 0$

approx $\varepsilon = 10^{-5}$


From the class forum:
data
rbf
rbf regul
Several RBFs fit by least squares → overfit
[See also Tutorial 1 Q3c]
L2 regularized, still quite "wiggly" nowhere near data. Why?
Math Scores
Drug Concentration
0
1
2
3
4
5
6
7
8
9
10
20
30
40
50
60
70
80