Recipe: fit parameters to data

Loss and function: $L(y^{(n)}, f(\underline{x}^{(n)}, \underline{w})) = L_n$

Square loss: $(y^{(n)} - f_n)^2$

Negative log probability: $-\log p(y^{(n)} | \underline{x}^{(n)}, \underline{w})$

Cost

$C = \sum_n L_n$ sum over training set

Learning direction

$$-\frac{1}{N} \nabla_{\underline{w}} C = \underbrace{-\frac{1}{N} \sum_n \nabla_{\underline{w}} L_n}$$

Monte Carlo approx:

$-\nabla_{\underline{w}} L_n$ for random n

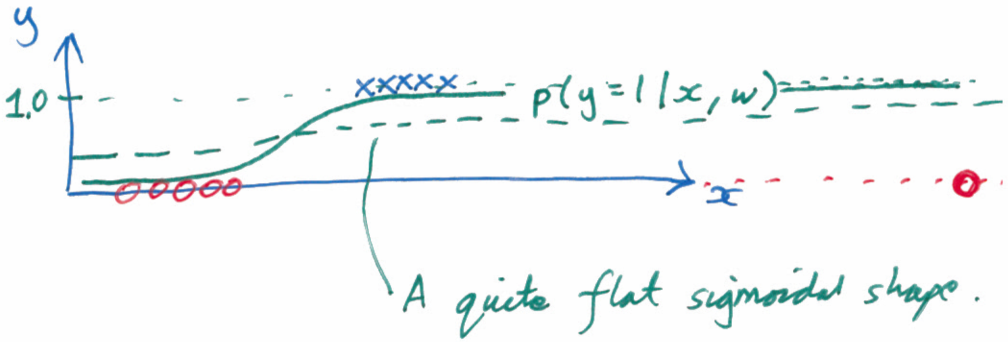
$\Rightarrow$ Identify probabilistic model of y, $p(y | \underline{x}, \underline{w})$

$\Rightarrow$ Update parameters:

$$\underline{w} \leftarrow \underline{w} + \eta \nabla_{\underline{w}} \log p(y^{(n)} | \underline{x}^{(n)}, \underline{w})$$

S.G.D. on loss. S.G. <u>Ascent</u> on log likelihood

Robust Logistic Regression



Each example has binary variable

$$m^{(n)} \in \{0, 1\}$$ hidden variable / latent

I will assume:

$$p(m|\varepsilon) = \begin{cases} 1-\varepsilon & m=1 \\ \varepsilon & m=0 \end{cases}$$

eg. $\varepsilon = 0.01$

Model for labels:

$$p(y=1|\underline{x},\underline{w},m) = \begin{cases} \sigma(\underline{w}^T\underline{x}) & m=1 \\ 1/2 & m=0 \end{cases}$$

Need Likelihood of $\underline{w}$, ε

$$p(y=1 \mid \underline{x}, \underline{w}, \varepsilon) = \sum_{m \in \{0,1\}} P(y=1, m \mid \underline{x}, \underline{w}, \varepsilon)$$

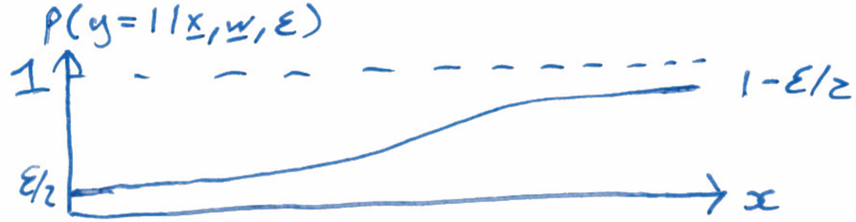
(Sum Rule)

$$= \sum_{m \in \{0,1\}} P(y=1 \mid \underline{x}, w, \cancel{\varepsilon}, m)\, P(m \mid \cancel{\underline{x}}, \cancel{\underline{w}}, \varepsilon)$$

(Product rule)

For this model

$$= \underbrace{(1-\varepsilon)\,\sigma(\underline{w}^{T}\underline{x})}_{m=1} + \underbrace{\varepsilon\ 1/2}_{m=0}$$



$$\nabla_{\underline{w}} \log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}, \varepsilon) = \ldots \text{calculus / algebra}$$

$$= \frac{1}{1 + \frac{1}{2}\left(\frac{\varepsilon}{1-\varepsilon}\right)\frac{1}{\sigma_n}}\ \nabla_{\underline{w}} \log \underline{\sigma_n}$$

$p(y=1 \mid \underline{w}, \underline{x})$ for logistic regression

How do we fit ε?

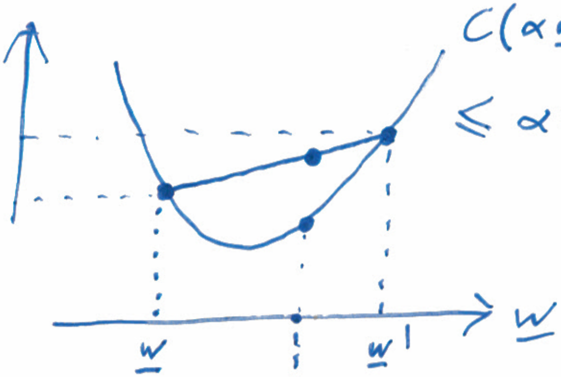
Set it by hand?

Grid of settings $\varepsilon \in [0, 1]$

Maybe on a log scale $0.1, 0.01, 0.001, \ldots$

For fixed ε the cost function is convex.

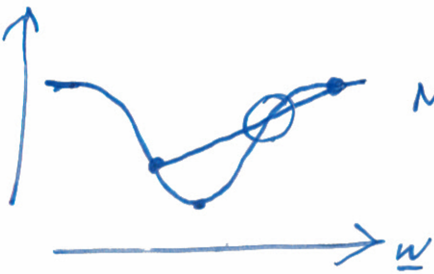
Cost
$C(\underline{w})$



$C(\alpha \underline{w} + (1-\alpha)\underline{w}')$
$\leq \alpha C(\underline{w}) + (1-\alpha) C(\underline{w}')$

$\alpha \underline{w} + (1-\alpha)\underline{w}', \quad 0 \leq \alpha \leq 1$

$C(\underline{w})$



Not convex

Use gradients to fit ε

Jointly fit $\theta = \begin{bmatrix} \underline{w} \\ \varepsilon \end{bmatrix}$, $\nabla_\theta C$

1 Problem $C(\varepsilon)$ is not convex.

→ Don't worry, do it anyway.

A real problem

$\varepsilon \in [0,1]$, ε is constrained.

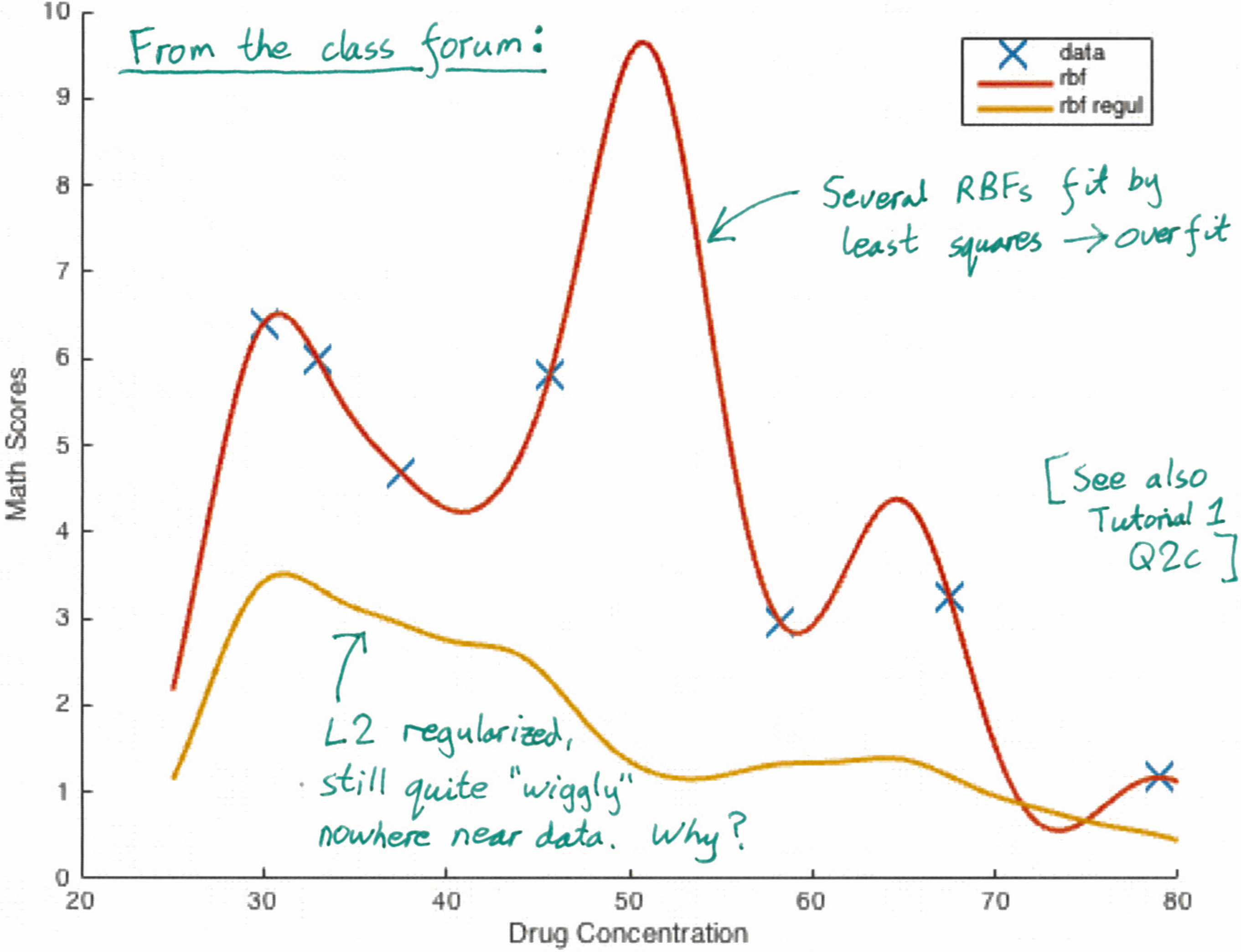
Trick: reparameterize model

$\varepsilon = \sigma(b)$, $\quad b = \log\left(\frac{\varepsilon}{1-\varepsilon}\right) = \text{logit}(\varepsilon)$

$-\infty < b < \infty$

logistic sigmoid.

Derive $\frac{\partial C}{\partial b}$ and optimize $\begin{bmatrix} \underline{w} \\ b \end{bmatrix}$



From the class forum:
data
rbf
rbf regul
Several RBFs fit by least squares → overfit
[See also Tutorial 1 Q2c]
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