3.5: Laplace prior on weights

Our assumed model is $y = wx + \varepsilon$

$\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the noise in the data

Prior on weights: $w \sim \text{Laplace}(0, b)$

$$P(w) = \frac{1}{2b} e^{-|w|/b}$$

Let's find $w^*_{MAP} = \operatorname{argmax} P(w \mid X, Y)$

$X = \{x_1, x_2 \dots x_N\}$, $Y = (y_1 \dots y_N)$.

$$w^*_{MAP} = \operatorname{argmax}_w \left(\frac{P(X, Y \mid w)\, P(w)}{P(X, Y)} \right)$$

$$= \operatorname{argmax}_w \left(\log P(X Y \mid w) + \log P(w) \right)$$

$$= \operatorname{argmax}_w \left(\sum_{n=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_n^2}{2\sigma^2}} \right) + \log \left(\frac{1}{2b} e^{-|w|/b} \right) \right)$$

$$= \operatorname{argmax}_w \left(-\log(\sqrt{2\pi\sigma^2}) - \sum_{n=1}^{N} \frac{(y_n - w x_n)^2}{2\sigma^2} + \log \frac{1}{2b} - \frac{|w|}{b} \right)$$

$\perp\!\!\!\perp$ from w $\qquad\qquad$ $\perp\!\!\!\perp$ from w

$$= \operatorname{argmax}_w \left(- \sum_{n=1}^{N} (y_n - w x_n)^2 - \frac{2\sigma^2}{b} |w| \right)$$

$$= \operatorname{argmin}_w \Big(\underbrace{\sum_{n=1}^{N} (y_n - w x_n)^2 + \frac{2\sigma^2}{b} |w|}_{\mathcal{L}_{laplace}} \Big)$$

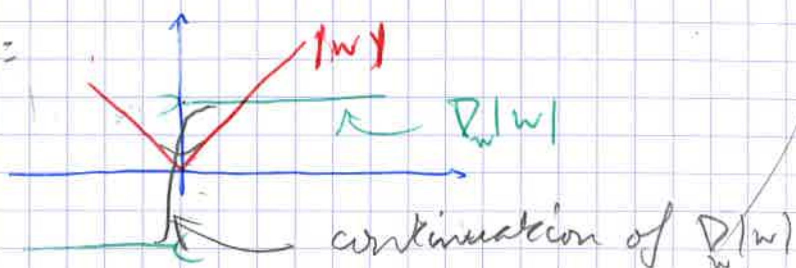
We can take a pseudo-gradient of this, keeping in mind that $|w|$ is not differentiable in 0, but the goal is just to find the minimum of $\mathcal{L}_{laplace}$

$$\nabla_w \mathcal{L} = 0 \iff -\lambda \sum_{n=1}^{N} (y_n - \vec{w}\vec{x}_n)\vec{x}_n + \frac{\lambda\sigma^2}{b}(\nabla_w |w|) = 0$$

$$\iff \begin{cases} \text{if } w \neq 0, & -\sum_{n=1}^{N} y_n \vec{x}_n + \sum_n (\vec{w}\vec{x}_n) x_n + \frac{\sigma^2}{b}\,\text{sign}(w) \\ \text{if } w = 0 & \underline{\qquad} - 0 + \frac{\sigma^2}{b}\, \text{"}\nabla_w |w|\text{"} \end{cases}$$

to be (interpreted...

(We can continue (prolongate) it as any number in $[-1, 1]$:



Rk: we should treat each w_d of $\vec{w} = (w_1 \dots w_D)$ separately.

for simplicity, assuming $D = 1$, $\vec{w} = w_1 \in \mathbb{R}^1$:

case 1: if $w = 0$, $-\sum_n y_n x_n + \frac{\sigma^2}{b} \text{"}\nabla_w |w|\text{"} = 0$

$$\frac{\sigma^2}{b} \cdot \text{"}\nabla_w |w|\text{"} = \sum_n^N y_n x_n$$

$\in [-1, 1] \leftarrow$ we can choose.

It's possible if $\frac{\sigma^2}{b} > \left| \sum_n^N y_n x_n \right|$.

case 2: if $w \neq 0$: $-\sum y_n x_n + \sum x_n (w x_n) + \frac{\sigma^2}{b} \text{sign}\, w = 0$

$$X^T (X w) = -\frac{\sigma^2}{b} \text{sign}\, \vec{w} + X^T Y \qquad (D > 1)$$

$$w_d = -\left((X^T X)^{-1} \frac{\sigma^2}{b} \text{sign}\, \vec{w}\right)_d + \left((X^T X)^{-1} X Y\right)_d$$ for a specific d.

$$= -\left(\vec{p} \cdot \text{sign}\, \vec{w}\right)_d + q_d$$

Where $p \hat{=} (X^T X)^{-1} \frac{\sigma^2}{b} \geq 0$ and $q \hat{=} (X^T X)^{-1} X Y$

(covariance matrix is positive semi-definite)

case 2-a: $w > 0$: $-p \,\text{sign}\, w + q = w > 0$

$\Rightarrow -p \times (+1) + q > 0 \Rightarrow q > p > 0$

$\Rightarrow \text{sign}(q) = +1$

case 2b: $w < 0$ $-p \times (-1) + q = w < 0 \Rightarrow q < -p < 0$

$\Rightarrow \text{sign}(q) = -1$

$\Big\}$ $|q| > p$ in both cases

In any sub-case 2a or 2b,
we have $|q| > p$ $\quad (q > p > 0$ or $+q < -p < 0)$

Thus $\left|(X^T X)^{-1} X Y\right| > (X^T X)^{-1} \frac{\sigma^2}{b}$

or $|XY| > \frac{\sigma^2}{b}$ (considering the two cases separately, and given $(X^T X)^{-1}$ is ≥ 0)
($(X^T X)^{-1}$ is positive semi-definite)

$$w \neq 0 \Rightarrow \left|\sum_{n=1}^{N} x_n y_n\right| > \frac{\sigma^2}{b}$$, and $\text{sign}(w) = \text{sign}\,(X^T X)^{-1} XY$

This is compatible with what we found in case 1 ($w = 0$).

We noticed that $\text{sign}\, w = \text{sign}\, q = \text{sign}\left((X^T X)^{-1} XY\right)$
$= \text{sign}\,(XY)$

So we simplify:

$$w = -p\,\text{sign}\, w + q \quad \text{into}$$

$$w = -(X^T X)^{-1} \frac{\sigma^2}{b} \text{sign}(XY) + (X^T X)^{-1} XY$$

which is a closed form.

Conclusion:

$$w^*_{MAP} = \begin{cases} 0 \text{ if } \left|\sum_n^N x_n y_n\right| < \frac{\sigma^2}{b} \\ (X^T X)^{-1}\left(XY - \frac{\sigma^2}{b}\text{sign}(XY)\right) \text{ if } \left|\sum_n^N x_n y_n\right| > \frac{\sigma^2}{b} \end{cases}$$

(If $|\Sigma x_n y_n| = \frac{\sigma^2}{b}$, the 2 expressions simply match)