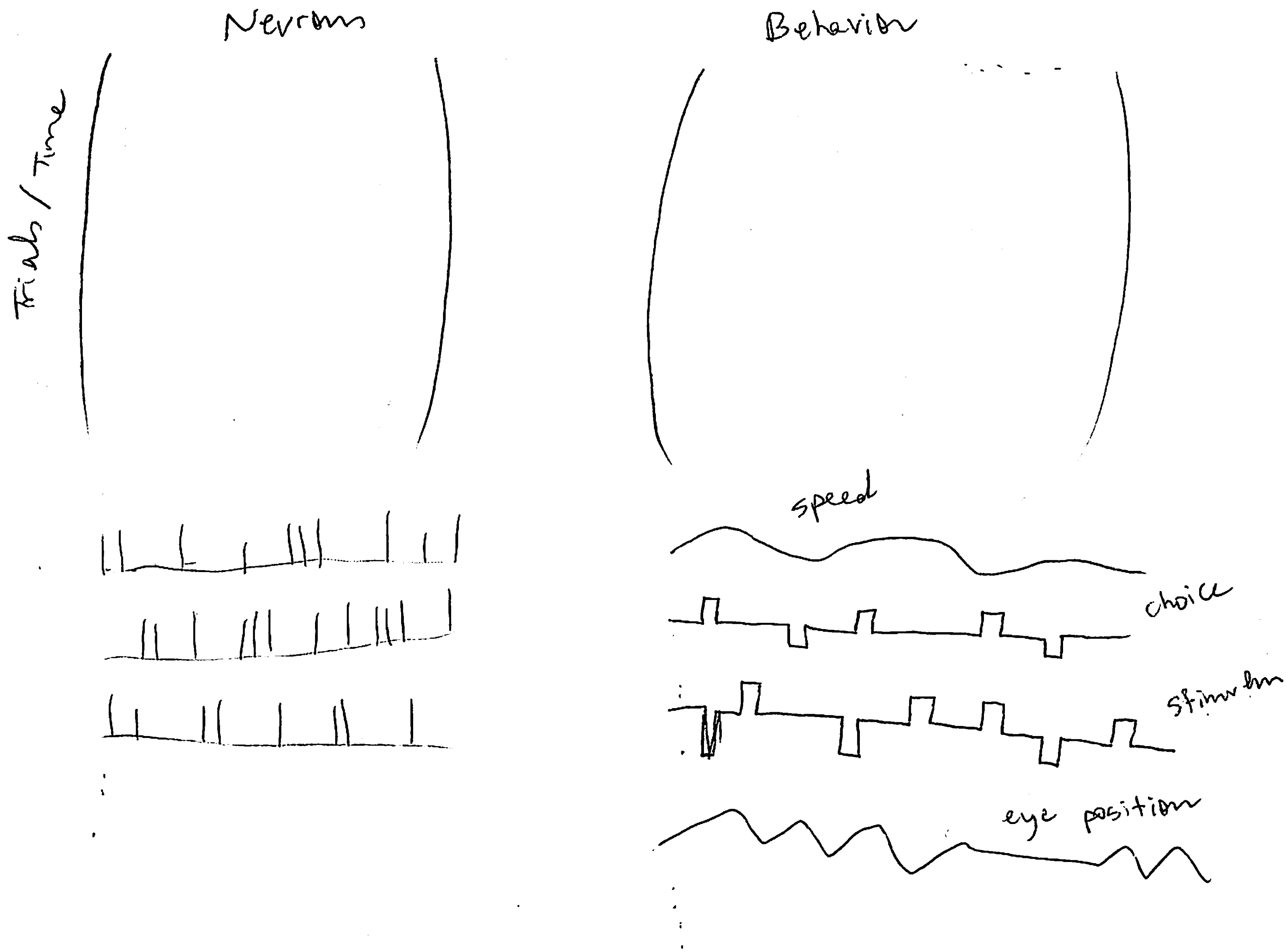
ENCODING & DECODING I

0. Introduction

- Let's first start describing a dataset that ~~would~~ represents a typical dataset we would encounter in neuroscience.



- In this and the upcoming ~~next~~ lecture we are going to see the two most important branches of supervised learning: regression and classification.
- Regression and classification can be defined as predicting a dependent variable Y from an independent D dimensional variable x.

- The difference between regression and classification relies on the fact that in regression the dependent variable is continuous whereas in classification is discrete.

In general we are going to assume that:

$$y = f(\vec{x})$$

- Restrict a bit more our assumption and say:

$$y = \vec{w}^T \vec{\phi}(\vec{x}) + w_0$$

where $\phi_i(\vec{x})$ are called the basis functions and represent in general an arbitrary function of x.
Here we are going to focus only on linear regression and therefore $\vec{\phi}(\vec{x}) = \vec{x}$.

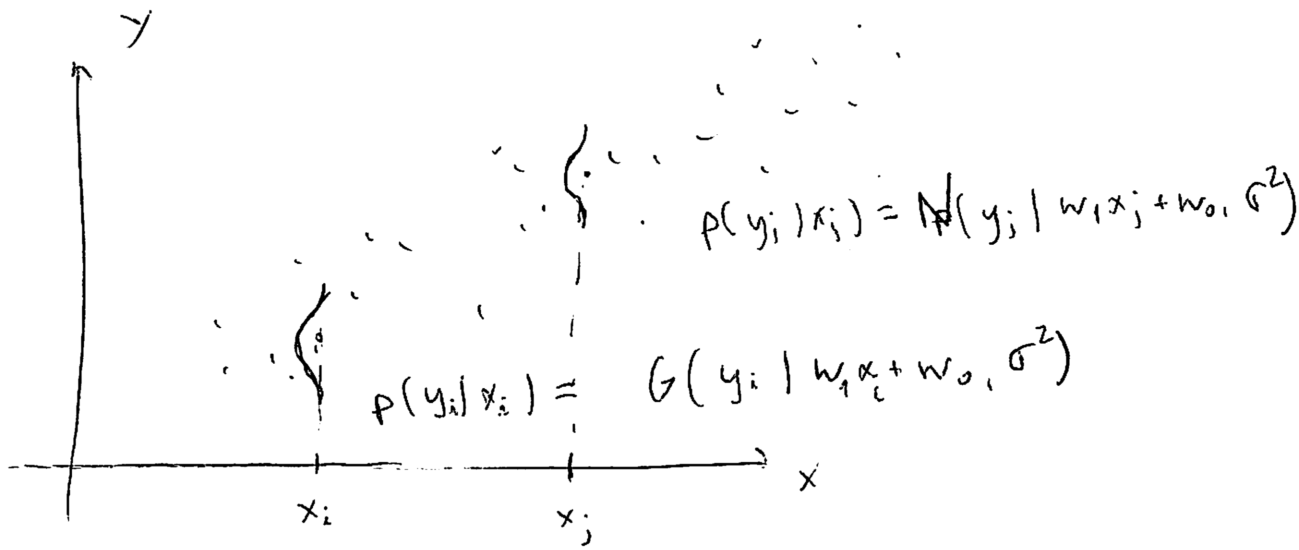
$$y = \vec{w}^T \vec{x} + w_0$$

Let's assume for instance

$$\rightarrow \boxed{r_i = \vec{w}^T \vec{B} + w_0}$$

---

1. <u>Maximum likelihood and least squares for linear regression:</u>



Our models assumes:

$$y = \vec{w}^T\vec{x} + w_0 + \sigma\xi = \vec{w}^T\vec{x} + \sigma\xi$$

where $\vec{x} = (\vec{x}, 1)$

$$\mathbb{E}[y|x] = \int y p(y|x) dy = \vec{w}^T\vec{x}$$

- Consider now a set of $\{y_1 - y_N\}$, $\{\vec{x}_1 -, \vec{x}_N\}$

The likelihood becomes:

$$p(\vec{y} | \vec{X}, \vec{w}) \underset{\text{iid}}{=} \prod_i p(y_i | \vec{x}_i, \vec{w}) = \prod \mathcal{N}(y_i | \vec{w}\vec{x}_i, \sigma^2)$$

$$\to \ln p(\vec{y} | X, \vec{w}, \sigma^2) = \sum_i \ln p(y_i | x_i, \vec{w}, \sigma^2) = -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_i (y_i - \vec{w}\vec{x})^2$$

$$\Rightarrow \boxed{E(\vec{w}) = \frac{1}{2}\sum_i (y_i - \vec{w}^T\vec{x})^2} \leftarrow \text{Only one depending on } \vec{w}$$

least squares cost function

- Maximize log prob. under Gaussian noise is equivalent to minimize least squares lost function.

$$\nabla_{\vec{w}} \ln p(\vec{y} | X, \vec{w}, \sigma^2) = \frac{1}{\sigma^2}\sum_i (y_i - \vec{w}^T\vec{x}_i)\vec{x}_i^T = 0$$

$$\sum_i y_i \vec{x}_i^T - \sum_i \vec{w}^T \vec{x}_i \vec{x}_i^T = 0 \Rightarrow \vec{w}^T \sum_i \vec{x}_i \vec{x}_i^T = \sum_i y_i \vec{x}^T$$

$$\boxed{\vec{w}_{ML} = (X^T X)^{-1} X^T \vec{y}}$$

let's have a look to the bias explicitly:

$$\frac{\partial}{\partial w_0} \ln p(\vec{y} | X, \vec{w}, \sigma^2) = \frac{\partial}{\partial w_0} \frac{1}{\sigma^2} \Sigma \left(y_i - (\vec{w}\vec{x}_i + w_0) \right)^2 = 0$$

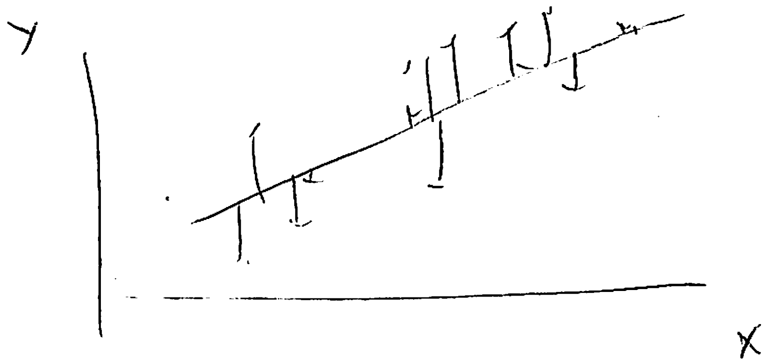
$$= \frac{1}{\sigma^2} \Sigma \left(y_i - \vec{w}^T \vec{x}_i + w_0 \right) = 0$$

$$\Sigma y_i - \Sigma \vec{w}^T \vec{x}_i - N w_0 = 0 \Rightarrow \boxed{w_{0\,ML} = \frac{\Sigma y_i}{N} - \vec{w}_{ML} \frac{\Sigma \vec{x}_i}{N}}$$

$$w_{0\,ML} = \bar{y} - \vec{w}_{ML} \bar{x}_i$$

Solution for σ^2

$$\sigma^2 = \frac{1}{N} \sum_i \left(y_i - \vec{w}_{ML} \vec{x}_i \right)^2$$ which is the sum of residuals square



2. Regularized least squares

- In the previous section we derived the likelihood function: $\mathcal{N}(\vec{y}|\vec{m}_0, \sigma^2 \mathbb{1})$, $\vec{m}_0 = X\vec{w}$

$$p(\vec{y}|\vec{x}, \vec{w}, \sigma^2) = \prod_i p(y_i|\vec{x}_i, \vec{w}, \sigma^2) = \prod_i \mathcal{N}(y_i|\vec{w}\vec{x}, \sigma^2)$$

To make things easier, let's skip the dependence on $\vec{x}$ and σ^2

$$p(\vec{y}|\vec{w}) = \mathcal{N}(\vec{y}|\vec{w})$$

Bayes Rule: $\quad p(a|b) = \dfrac{p(b|a)\,p(a)}{p(b)} \propto p(b|a)\,p(a)$

$$\Rightarrow \underbrace{p(\vec{w}|\vec{y})}_{\text{Gaussian}} \propto \underbrace{p(\vec{y}|\vec{w})}_{\text{Gaussian}}\,\underbrace{p(\vec{w})}_{\text{Gaussian}}$$

Gaussian is the conjugate prior of itself

- Let's assume $p(\vec{w}) = \mathcal{N}(\vec{w}|\vec{m}_0, S_0)$

Then $\boxed{\text{Exercise}}$

$$p(\vec{w}|\vec{y}) = \mathcal{N}(\vec{w}|\vec{m}_N, S_N)$$

where
$$\begin{cases} \vec{m}_N = S_N\left(S_0^{-1}\vec{m}_0 + \frac{1}{\sigma^2}X^T\vec{y}\right) \\ S_N^{-1} = S_0^{-1} + \frac{1}{\sigma^2}X^TX \end{cases}$$

Let's assume $S_0 = \alpha\mathbb{1}$, then:
$$\begin{cases} \vec{m}_N = S_N\left(\alpha\vec{m}_0 + \frac{1}{\sigma^2}X^T\vec{y}\right) \\ S_N^{-1} = \alpha^{-1}\mathbb{1} + \frac{1}{\sigma^2}X^TX \end{cases}$$

~~$p(\vec{w}|\alpha) = \mathcal{N}(\vec{w}|0$~~

If the prior in infinitely broad, $\alpha \to 0$, then:

$$\vec{m}_N = (X^T X)^{-1} X^T \vec{y} \Rightarrow \vec{w}_{ML}$$ & likelihood maximization.

- let's now assume $p(\vec{w}|\alpha) = \mathcal{N}(\vec{w}|\vec{0}, \alpha^{-1}\mathbb{I})$

$$\begin{cases} \vec{m}_N = \frac{1}{\sigma^2} S_N X^T \vec{y} \\ S_N^{-1} = \alpha \mathbb{I} + \frac{1}{\sigma^2} X^T X \end{cases} \quad \to p(\vec{w}|\vec{y}) = \mathcal{N}(\vec{w}|\vec{m}_N, S_N)$$

$$\vec{m}_N = \frac{1}{\sigma^2}\left(\alpha \mathbb{I} + \frac{1}{\sigma^2} X^T X\right)^{-1} X^T \vec{y} = \left(\sigma^2 \alpha \mathbb{I} + X^T X\right)^{-1} X^T \vec{y}$$

Therefore $$\boxed{\vec{w}_{MAP} = (\lambda \mathbb{I} + X^T X)^{-1} X^T \vec{y}}$$

where $\lambda = \alpha \sigma^2$

$\left(\underset{\vec{w}}{\operatorname{argmax}}\, \mathcal{N}(\vec{w}|\vec{m}_N, S_N) \Rightarrow \vec{w} = \vec{m}_N\right)$

We can also reason the following way:

$$p(\vec{w}|\vec{y}) = p(\vec{y}|\vec{w})\, p(\vec{w}) \Rightarrow \ln p(\vec{w}|\vec{y}) = \ln p(\vec{y}|\vec{w}) + \ln p(\vec{w})$$

$$\ln p(\vec{y}|\vec{w}) = -\frac{1}{2\sigma^2} \sum_i (y_i - \vec{w}\vec{x})^2 + \text{const}$$

$$\ln p(\vec{w}) = -\frac{1}{2} \alpha \vec{w}^T \vec{w}$$

$$\vec{w}_{MAP} = \underset{\vec{w}}{\operatorname{argmax}} \left[\ln p(\vec{w}|\vec{y})\right]$$

$$\nabla_w \ln p(\vec{w}|\vec{y}) = \frac{1}{\sigma^2} \sum_i (y_i - \vec{w}^T \vec{x}_i)\vec{x}_i^T - \alpha \vec{w}^T = 0$$

$$\vec{w}^T(\sigma^2 \alpha \mathbb{I} + X^T X) = X\vec{y} \to \boxed{\vec{w}^T = (\lambda \mathbb{I} + X^T X)^{-1} X^T \vec{y}}$$

3. Predictive distribution

- It is not only important to find $\vec{w}_{ML}$ or $p(\vec{w})$, but we also are interested in making predictions of y_i given a particular $\vec{x}_i$.

- We can do: $p(y_i \mid \vec{x}_i) = N(y_i \mid \vec{w}_{ML}^T \vec{x}_i, \sigma^2)$

or we can take a Bayesian approach that will ~~instead~~ furthermore prevent overfitting:

$\cancel{p(y_i \mid \vec{y}, X_i}$

$$p(y_i \mid \vec{x}_i, \vec{y}, X, \alpha, \sigma^2) = \int p(y \mid \vec{x}_i, \vec{w}, \sigma^2)\, p(\vec{w} \mid \vec{y}, X, \alpha, \sigma^2)\, d\vec{w} =$$

where
$$p(y \mid \vec{x}_i, \vec{w}, \sigma^2) = N(y \mid \vec{w}\vec{x}, \sigma^2)$$
$$p(\vec{w} \mid \vec{y}, X, \alpha, \sigma^2) = N(\vec{w} \mid \vec{m}_N, S_N)$$

$\cancel{\vec{m}_N = S_N(S}$

$$\begin{cases} \vec{m}_N = \dfrac{1}{\sigma^2} S_N X^T y \\ S_N^{-1} = \alpha I + \dfrac{1}{\sigma^2} X^T X \end{cases}$$

$\boxed{\text{Exercise}}$

$$\Rightarrow p(y \mid \vec{x}_i, \vec{y}, X, \alpha, \sigma^2) = N(y \mid \vec{m}_N \vec{x}, \sigma^2_{N(\vec{x})})$$

where $\sigma^2_{N(\vec{x})} = \sigma^2 + \vec{x}^T S_N \vec{x}$

$$\cancel{\sigma_N^2 = \sigma^2 + x^T\left(\alpha \mathbb{1} + \frac{1}{\sigma^2} X^T X\right)^{-1} x}$$

~~if $N \to \infty$~~

$$\cancel{\sigma_N^2 = \sigma^2 + \sigma^2 x^T (X^T X)^{-1} x}$$

$$\cancel{= \sigma^2 + \sigma^2 x^T X^{-1} X^{T-1} x}$$

$$\cancel{= \sigma^2 (\mathbb{1} + x^T X^{-1} X^{T-1} x)}$$

if $N \to \infty$

$$\sigma_N(\check{x}) = \sigma^2 + \sigma^2 \check{x}^T (X^T X)^{-1} \check{x}$$

$$= \sigma^2\left(1 + \cancel{\check{x}^T (X^T X)^{-1} \check{x}}\right) = \sigma^2$$

$\downarrow$

if $N \to \infty$

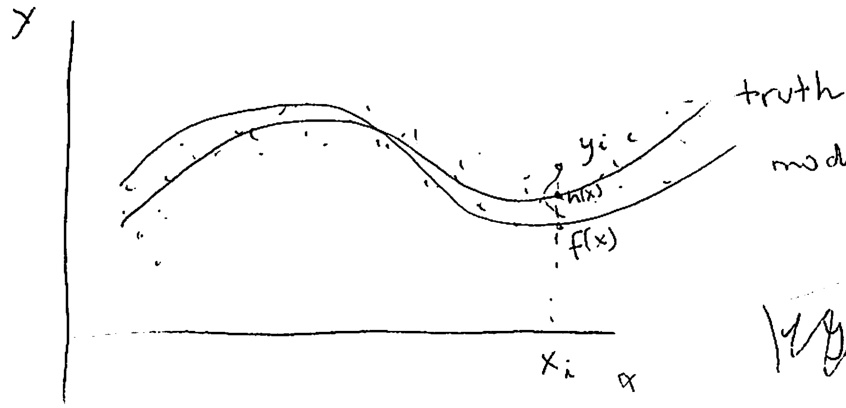
$(X^T X)^{-1} \to 0$

The more data the more precise in our estimate.

We can never break the ceiling imposed by the noise of the data itself.

4. Bias-Variance Trade-off

. Let's assume we want to use a more flexible model, not just linear regression:



$$\boxed{y = \vec{w}^T \vec{\phi}(\vec{x})}$$

$$y = w_0 + w_1 x + w_2 x^2 + \dots$$

. If we let $\vec{w}$ full flexibility we can end-up overfitting. What is the best strength for the regularization?

Let's define:

$$\begin{cases} h(x) = \int y \, P_t(y|x) \, dy = \mathbb{E}_t[y|x] \leftarrow \text{True} \\ f(x) = \int y \, P_m(y|x) \, dy = \mathbb{E}_m[y|x] \leftarrow \text{model} \end{cases}$$

$$\mathbb{E}[\mathcal{L}] = \mathbb{E}_{x,y}\left[(y - f(x))^2\right] = \mathbb{E}_{x,y}\left[(y - h(x) + h(x) - f(x))^2\right]$$

$$= \mathbb{E}_{x,y}\left[(y - h(x))^2\right] + \underbrace{\mathbb{E}_x\left[(h(x) - f(x))^2\right]}$$

$\boxed{\text{Exercise}}$

• Let's focus now on the term $\mathbb{E}_x[(f(x)-h(x))^2]$

Important to note $f(x) = f(x;D)$ where D is the dataset we used to learn our model

~~f(x) - h(x~~

$$(f(x;D)-h(x))^2 = (f(x;D) - \mathbb{E}_D[f(x;D)] + \mathbb{E}_D[f(x;D)] - h(x))^2$$

$$\Rightarrow \mathbb{E}_D[(f(x;D)-h(x))^2] = \{\mathbb{E}_D[f(x;D)] - h(x)\}^2 + \mathbb{E}_D[(f(x;D) - \mathbb{E}_D[f(x;D)])^2]$$

$\boxed{\text{Exercise}}$

Therefore: $\mathbb{E}_{x,y}[\mathcal{L}] = \mathbb{E}_{x,y}[(y - f(x))^2] =$

$$= \{\mathbb{E}_D[f(x;D)] - h(x)\}^2 \quad \text{Bias}^2$$

$$+ \mathbb{E}_D[(f(x;D) - \mathbb{E}_D[f(x;D)])^2] \quad \text{Variance}$$

$$+ \mathbb{E}_{x,y}[(h(x) - y)^2] \quad \text{Noise}$$