PREPARATORY WORK

DEFINITIONS

- PRIOR PROBABILITY DISTRIBUTION (often called PRIOR) of a certain quantity is the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account.

- LIKELIHOOD FUNCTION measures the goodness fit of a statistical model to a sample of data for a given values of unknown parameters. $\mathbb{P}(X|\theta)$ (prob. of evidence given parameters)

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|x) = \mathbb{P}_\theta(x) = \mathbb{P}_\theta(X=x) \quad \text{DISCRETE}$$

(notation)

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|x) = f_\theta(x) \quad \text{CONTINUOUS}$$

(f_θ: pdf)

- POSTERIOR PROBABILITY DISTRIBUTION is the probability distribution of an unknown quantity, treated as random variable, conditional on the evidence obtained from an experiment or survey.

$\mathbb{P}(\theta|X)$ (prob. of parameters given the evidence)

Given a prior belief that a probability distribution is $\mathbb{P}(\theta)$ (cdf(θ)) or $p(\theta)$ (pdf(θ)) and that the observation x have a likelihood $p(x|\theta)$, then the posterior probability is defined as:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(X)}$$

where $\mathbb{P}(X)$ is the normalizing const.

(cont.: $= \int \mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta)\, d\theta$; discr.: $\sum_{\theta \in \Theta} \mathbb{P}(X|\theta)\,\mathbb{P}(\theta)$)

Obs: the posterior probability is proportional to likelihood · prior prob.

- **MAXIMUM LIKELIHOOD ESTIMATION (ML)**: method of estimating the parameters of an assumed probability distribution given some observed data. This is achieved by maximizing the likelihood function.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta \mid y)$$

(← notation; observed data maybe: $\mathcal{L}(\theta \mid x)$)

$$= \underset{\theta}{\operatorname{argmax}}\ \mathbb{P}(x \mid \theta)$$

- **MAXIMUM A POSTERIORI ESTIMATION (MAP)**: method of estimating the parameters of a statistical model given some data. It's done by maximizing posterior prob. distribution. It's closely related to ML estimation, but employs an augmented optimization objective, which incorporates the prior distribution.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{P}(\theta \mid x) = \underset{\theta}{\operatorname{argmax}}\ \frac{\mathbb{P}(x \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(x)} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{P}(x \mid \theta) \cdot \mathbb{P}(\theta)$$

($\mathbb{P}(x)$: const.)

1.

a) done

b) $p(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$

$$p(\{x_n\}_{n=1}^N|\mu) = \prod_{n=1}^N p(x_n|\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-\frac{1}{2}\sum_{n=1}^N (x_n-\mu)^2} =$$

$$= (2\pi)^{-\frac{N}{2}} \cdot e^{-\frac{1}{2}\sum_{n=1}^N (x_n^2 - 2x_n\mu + \mu^2)} =$$

$$= (2\pi)^{-N/2} \cdot e^{-\frac{1}{2}N(\overline{x_n^2} - 2\mu\overline{x_n} + \mu^2)}$$

On the other hand:

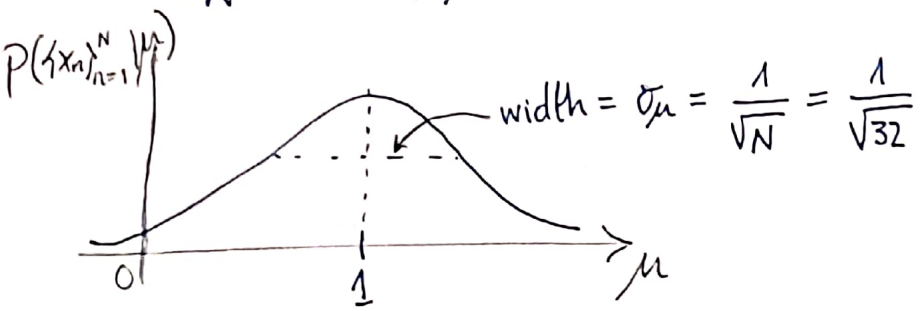
$$z\,\mathcal{N}(\mu;\mu_\mu,\sigma_\mu^2) = \frac{z}{\sqrt{2\pi\,\sigma_\mu^2}} e^{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_\mu)^2} =$$

$$= \frac{z}{\sqrt{2\pi\,\sigma_\mu^2}} e^{\frac{-1}{2\sigma_\mu^2}(\mu^2 - 2\mu\mu_n + \mu_\mu^2)}$$

compare coefficients

$\Rightarrow \sigma_\mu^2 = 1/N$ $\quad \mu_\mu = \overline{x_n} = 1$ ← info.

(we could also find z, but it is not asked)



NB: when fitting a Gaussian, the likelihood only depend on the data's 1st and 2nd moments, so even though the data appear to come from a uniform density here, we only need the two moments.

2. Info: $p(d) = \mathcal{N}(d; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2}$

$p(y|d, \sigma_y^2) = \mathcal{N}(y; d, \sigma_y^2) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{1}{2\sigma_y^2}(y-d)^2}$

a) Posterior: $p(d|y) \propto \overbrace{p(y|d, \sigma_y^2)}^{\text{likelihood}} \cdot \overbrace{p(d)}^{\text{prior}} \propto$

$\propto e^{-\frac{1}{2\sigma_y^2}(y-d)^2 - \frac{1}{2}d^2} = e^{-\frac{1}{2\sigma_y^2}(y^2 - 2yd + d^2) - \frac{1}{2}d^2} =$

$= e^{-\frac{1}{2}\left[d^2\left(\frac{1}{\sigma_y^2} + 1\right) - \frac{2dy}{\sigma_y^2} + \frac{y^2}{\sigma_y^2}\right]}$ ← compare coefficients

If we suppose that: $p(d|y) \sim \mathcal{N}(d; \mu_{d|y}, \sigma_{d|y})$ ↓

$e^{-\frac{1}{2}\left[\frac{d^2}{\sigma_{d|y}^2} - 2\frac{d\mu_{d|y}}{\sigma_{d|y}^2} + \frac{\mu_{d|y}^2}{\sigma_{d|y}^2}\right]}$

$\Rightarrow \frac{1}{\sigma_y^2} + 1 = \frac{1}{\sigma_{d|y}^2} \Rightarrow \sigma_{d|y}^2 = \frac{\sigma_y^2}{1 + \sigma_y^2}$

$\Rightarrow \frac{y}{\sigma_y^2} = \frac{\mu_{d|y}}{\sigma_{d|y}^2} \Rightarrow \mu_{d|y} = \sigma_{d|y}^2 \cdot \frac{y}{\sigma_y^2} = \frac{y}{1 + \sigma_y^2}$

b) If $\sigma_y^2 \to \infty \Rightarrow \sigma_{d|y}^2 \to 1$

$\Rightarrow \mu_{d|y} \to 0$

collapses to the prior, so the posterior doesn't give information

this is the prior variance as it should be as now the sensor is so noisy it doesn't tell us anything → we just have a priori beliefs

NB: note also that when $\sigma_y^2 \to 0$ the sensor gives us perfect information about y, so $\sigma_{d|y}^2 \to 0$ and $\mu_{d|y} \to y$ as expected.

3. $x_{1:N} = \underbrace{\{0,1,1,1,0,1,0,0,\dots\}}_{N \text{ coinflips}}$

$x_n = 1$ heads
$x_n = 0$ tails
$\rho :=$ prob. of landing heads

$$\mathbb{P}(x_n|\rho) = \rho^{x_n}(1-\rho)^{1-x_n}$$

$$\mathbb{P}(\rho|n_0,N_0) = \frac{1}{Z(n_0,N_0)}\rho^{n_0}(1-\rho)^{N_0-n_0} \quad ; \quad Z(n_0,N_0) := \text{normalizing constant}$$

a) $\underbrace{\mathbb{P}(\rho|x_{1:N},n_0,N_0)}_{\text{posterior}} \propto \underbrace{\mathbb{P}(x_{1:N}|\rho,n_0,N_0)}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}(\rho|n_0,N_0)}_{\text{prior}} =$

$$= \mathbb{P}(\rho|n_0,N_0) \cdot \prod_{i=1}^{N} \mathbb{P}(x_i|\rho,n_0,N_0) =$$

$$= \frac{1}{Z(n_0,N_0)}\rho^{n_0}(1-\rho)^{N_0-n_0} \cdot \prod_{i=1}^{N} \rho^{x_i}(1-\rho)^{1-x_i} =$$

$$= \frac{1}{Z(n_0,N_0)}\rho^{n_0}(1-\rho)^{N_0-n_0} \cdot \rho^{\sum x_i}\cdot(1-\rho)^{N-\sum x_i} =$$

($\sum x_i$ → number of 1's, $n = \sum x_i$)

$$= \frac{1}{Z_1(n_0,N_0)}\rho^{n_0+n}(1-\rho)^{N_0+N-n_0-n} \equiv \left[\frac{1}{Z(n',N')}\rho^{n'}(1-\rho)^{N'-n'}\right]$$

where $\begin{cases} n' = n_0 + n \\ N' = N_0 + N \end{cases}$

(this is a Beta distribution and it is conjugate to the likelihood, meaning that the posterior has the same form)

b) $\log \mathbb{P}(\rho|x_{1:N},n_0,N_0) = -\log Z(n_0,N_0) + n'\log(\rho) + (N'-n')\log(1-\rho)$

$$\Rightarrow \frac{d}{d\rho}\log \mathbb{P}(\rho|x_{1:N},n_0,N_0) = \frac{n'}{\rho_{MAP}} - \frac{N'-n'}{1-\rho_{MAP}} = 0 \Rightarrow \boxed{\rho_{MAP} = \frac{n'}{N'}}$$

c) N_0 = number of pseudo data points, i.e., points seen before real data
n_0 = number of 1's in pseudo data
ML estimate = MAP estimate when prior is constant, so
ML estimate is recovered when $N_0 = n_0 = 0$ (flat prior distrib.)
i.e. no pseudo data

4. WLG let door 1 be the one selected by the contestant

let S = position of prize $\quad S \in \{1,2,3,4\}$

A priori we assume $\mathbb{P}(S=k) = \frac{1}{4} \quad k=1,\dots,4.$

The datum we receive after choosing door 1 is either $D=2$, $D=3$, $D=4$, i.e., the gameshow host opens door 2, 3 or 4.

We assume that when the host has a choice about which door to open he selects between those doors without prize.

$\mathbb{P}(D=2|S=1) = \frac{1}{3}$ $\quad \mathbb{P}(D=2|S=2) = 0$ $\quad \mathbb{P}(D=2|S=3) = 1/2$ $\quad \mathbb{P}(D=2|S=4) = 1/2$

$\mathbb{P}(D=3|S=1) = 1/3$ $\quad \mathbb{P}(D=3|S=2) = \frac{1}{2}$ $\quad \mathbb{P}(D=3|S=3) = 0$ $\quad \mathbb{P}(D=3|S=4) = 1/2$

$\mathbb{P}(D=4|S=1) = 1/3$ $\quad \mathbb{P}(D=4|S=2) = \frac{1}{2}$ $\quad \mathbb{P}(D=4|S=3) = 1/2$ $\quad \mathbb{P}(D=4|S=4) = 0$

Applying Bayes' theorem:

$$\mathbb{P}(S=K|D=4) = \frac{\mathbb{P}(D=4|S=K)\cdot \overbrace{\mathbb{P}(S=K)}^{=\frac{1}{4}\ \text{(prior)}}}{\underbrace{\mathbb{P}(D=4)}_{=1/3}}$$

$$\mathbb{P}(S=1|D=4) = \frac{\mathbb{P}(D=4|S=1)\,\mathbb{P}(S=1)}{\mathbb{P}(D=4)} = \frac{1/3 \cdot 1/4}{1/3} = \frac{1}{4} \quad (\text{same as prior})$$

$$\mathbb{P}(S=2|D=4) = \frac{1/2 \cdot 1/4}{1/3} = \frac{3}{8} \quad ; \quad \mathbb{P}(S=3|D=4) = \frac{1/2 \cdot 1/4}{1/3} = \frac{3}{8}$$

$$\mathbb{P}(S=4|D=4) = 0$$

greater than prior

$\Rightarrow$ So if we switch to doors 2 or 3 we'll increase our chances of winning from 1/4 to 3/8 (i.e. 1'5x)

NB: To get intuition for the fact that the opening of the door by the host provides information, consider 100 doors and the host opening 98 of them.

NB2: Version of 'Monty Hall problem' (page 97, David McKay, Information Theory, Inference book)

5.

a) $\hat{x}_* = \underset{\hat{x}}{\operatorname{argmax}} \int R(\hat{x}, x)\, \mathbb{P}(x|y)\, dx$

b) Find optimum: $R(\hat{x}, x) = -(\hat{x}-x)^2$

$$\frac{d}{d\hat{x}} \int -(\hat{x}-x)^2\, \mathbb{P}(x|y)\, dx = 0 \iff 2\int (x-\hat{x})\, \mathbb{P}(x|y)\, dx = 0$$

$$\iff \int \hat{x}\, \mathbb{P}(x|y)\, dx = \hat{x} \underbrace{\int \mathbb{P}(x|y)\, dx}_{=1} = \int x\, \mathbb{P}(x|y)\, dx \implies \boxed{\hat{x}_* = \int x\, \mathbb{P}(x|y)\, dx}$$

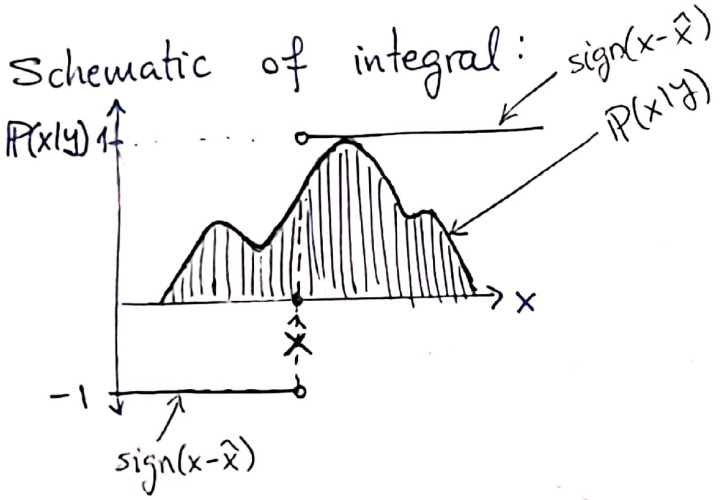
i.e. the posterior mean minimizes the expected squared error.

c) Find optimum when $R(\hat{x}, x) = -|\hat{x}-x|$

$$\frac{d}{d\hat{x}} \int -|\hat{x}-x|\, \mathbb{P}(x|y)\, dx = 0 \iff \frac{-d}{d\hat{x}} \int \sqrt{(\hat{x}-x)^2}\, \mathbb{P}(x|y)\, dx = 0$$

$$\iff \frac{1}{2} \cdot 2 \int \underbrace{\frac{(x-\hat{x})}{\sqrt{(x-\hat{x})^2}}}_{\text{sign}(x-\hat{x})} \mathbb{P}(x|y)\, dx = 0 \iff \int \text{sign}(x-\hat{x})\, \mathbb{P}(x|y)\, dx = 0$$

Schematic of integral:



$\implies$ need to find the point where the green and blue areas are equal

It is called the median of the distribution, $\hat{x}_*$, i.e. it has half the density above it and half below.

6.

a) LINEAR REGRESSION MODEL : gradient ≈ 5 and intercept $(0,0)$

The noise appears to be Gaussian but its standard dev. appears to grow with x. Therefore an appropiate model (including noise) would be: $y_n(x) = 5x + \sigma(x)\,\varepsilon_n$

where $\varepsilon_n \sim \mathcal{N}(0,1)$ and $\sigma(x) = |x|$

Note that there are many reasonable choices that could be made to model this data.

b) SINUSOIDAL MODEL : period ≈ 25 time steps

mean $= 2$

amplitude ≈ 1

Since there are outliers present, an appropiate noise model would be some heavy tailed model, such as a Student-t distribution: $y_n(x) = 2 + \sin\left(\frac{2\pi}{25}x\right) + \varepsilon_n$

with $\varepsilon_n \sim$ Student-t

(2 ↑ mean; sin ↑ amplitude is ≈ 1; $\frac{2\pi}{25}x$ ↖ time steps)

The Student-t parameters are hard to guess

- → mean (maybe 0)
- → variance (maybe 1)
- → degrees of freedom (maybe 2) ← even harder to guess!

NB: Important to note that outliers are present and that a Gaussian noise model may not be the most appropiate choice here.

7.

a) $$\mathbb{P}(\{y_n\}_{n=1}^N \mid a, \{x_n\}_{n=1}^N) = \prod_{n=1}^N \mathbb{P}(y_n \mid a, x_n) \implies$$

$$\implies \log \mathbb{P}(\{y_n\}_{n=1}^N \mid a, \{x_n\}_{n=1}^N) = \sum_{n=1}^N \log \mathbb{P}(y_n \mid a, x_n) =$$

$$= \sum_{n=1}^N \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_n - a x_n)^2 \right] =$$

$$= \frac{-N}{2}\log(2\pi) - \frac{1}{2}\sum_{n=1}^N (y_n - a x_n)^2 := \mathcal{L}(a)$$

b) $$\left.\frac{d\mathcal{L}(a)}{da}\right|_{a_{ML}} = \sum_{n=1}^N x_n(y_n - a_{ML} x_n) = 0 \iff a_{ML} = \frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2}$$

8.

a) $$\mathcal{L}(a) = \log \mathbb{P}(\{y_n\}_{n=1}^N, \{z_n\}_{n=1}^N \mid \{x_n\}_{n=1}^N, a) = \sum_{n=1}^N \log \mathbb{P}(y_n, z_n \mid x_n, a)$$

where $$\mathbb{P}(y_n, z_n \mid a, x_n) = \mathcal{N}\left(\begin{bmatrix} y_n \\ z_n \end{bmatrix}; \begin{bmatrix} a x_n \\ x_n \end{bmatrix}, \Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}^{-1} \right)$$

$$\implies \mathcal{L}(a) = \sum_{n=1}^N -\frac{1}{2}\left(\begin{bmatrix} y_n \\ z_n \end{bmatrix} - \begin{bmatrix} a x_n \\ x_n \end{bmatrix} \right)^T \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \left(\begin{bmatrix} y_n \\ z_n \end{bmatrix} - \begin{bmatrix} a x_n \\ x_n \end{bmatrix} \right) - \frac{N}{2}\log(\det(2\pi\Sigma))$$

$$= \frac{-1}{2}\left(\sum_n (y_n - a x_n)^2 + \sum_n (y_n - a x_n)(z_n - x_n) + \sum_n (z_n - x_n)^2 + N\log(\det(2\pi\Sigma)) \right)$$

a bit of info from just observing y_i ; extra bit from observing z's

b) $$\frac{d\mathcal{L}(a)}{da} = \sum_n (y_n - a x_n) x_n + \frac{1}{2}\sum_n (z_n - x_n) x_n = 0 \iff$$

$$\iff \sum_n y_n x_n + \frac{1}{2}\sum_n z_n x_n - \frac{1}{2}\sum_n x_n^2 - a\sum_n x_n^2 = 0 \iff$$

$$\iff a = \left(\sum_n y_n x_n + \frac{1}{2}\underbrace{\sum_n (z_n - x_n) x_n}_{:= z_n'} \right) \Big/ \sum_n x_n^2 \quad \text{(max likelihood estimate)}$$

new contribution from observing z's

c) The additional outputs change the ML estimate of a. This means that they must provide useful information about a. They do this because the noise in z_n is correlated with the noise in y_n, and so observing z_n reveals information about the noise ϵ_n and allows more accurate identification of a.

9. Info: $P(y|m,c,x) = \mathcal{N}(y; mx+c, 1)$ $\quad P(m) = \mathcal{N}(m; 0,1)$ $\quad P(c) = \mathcal{N}(c; 0,1)$

a) $P(m,c|y,x) = \dfrac{P(y|m,c,x)\cdot P(m)\cdot P(c)}{P(y|x)} \propto P(y|m,c,x)\,P(m)\cdot P(c) =$ ← product of Gaussian functions in m & c

$= \mathcal{N}\left(\begin{bmatrix} m \\ c \end{bmatrix}; \mu^{POST}, \Sigma^{POST}\right)$ ⇓ posterior must be Gaussian: just need to compute mean and covariance

Tactic:
1) Substitute in for these densities.
2) Identify mean and covariance by comparing coefficients.

1) $P(m,c|y,x) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-mx-c)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}m^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}c^2} \propto$

$\propto \exp\left(-\frac{1}{2}y^2 - \frac{1}{2}m^2(x^2+1) - \frac{1}{2}c^2 \cdot 2 - mcx + y(mx+c)\right) \propto$

$\propto \exp\left(-\frac{1}{2}\left(\begin{bmatrix} m \\ c \end{bmatrix} - \mu^{POST}\right)\left(\Sigma^{POST}\right)^{-1}\left(\begin{bmatrix} m \\ c \end{bmatrix} - \mu^{POST}\right)\right) \propto$

$\propto \exp\left(-\frac{1}{2}\left(\begin{bmatrix} m \\ c \end{bmatrix}^T \left(\Sigma^{POST}\right)^{-1}\begin{bmatrix} m \\ c \end{bmatrix} + \begin{bmatrix} m \\ c \end{bmatrix}^T \left(\Sigma^{POST}\right)^{-1}\mu^{POST}\right)\right)$

← inverse covariance times mean $\begin{bmatrix} m \\ c \end{bmatrix}^T \begin{bmatrix} yx \\ y \end{bmatrix}$

inverse covariance

$\Rightarrow \Sigma^{POST} = \begin{bmatrix} x^2+1 & x \\ x & 2 \end{bmatrix}^{-1} = \frac{1}{2(x^2+1) - x^2}\begin{pmatrix} 2 & -x \\ -x & x^2+1 \end{pmatrix} = \frac{1}{x^2+2}\begin{pmatrix} 2 & -x \\ -x & x^2+1 \end{pmatrix}$

$\Rightarrow$ Since $\left(\Sigma^{POST}\right)^{-1}\mu^{POST} = \begin{bmatrix} yx \\ y \end{bmatrix} \Rightarrow \mu^{POST} = \frac{1}{x^2+2}\begin{pmatrix} 2 & -x \\ -x & x^2+1 \end{pmatrix}\begin{bmatrix} yx \\ y \end{bmatrix} =$

$= \frac{1}{x^2+2}\begin{bmatrix} 2yx - yx \\ -yx^2 + (x^2+1)y \end{bmatrix} = \frac{1}{x^2+2}\begin{bmatrix} yx \\ y \end{bmatrix}$

b) Check these expressions match with those in lectures:

$$\underline{y} = \underline{\underline{\tilde{X}}}\,\underline{w} + \underline{\epsilon}\,, \quad \text{where } \underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 \underline{\underline{I}}), \quad \underline{w} \sim \mathcal{N}(\underline{0}, \lambda^{-1}\underline{\underline{I}})$$

$$P(\underline{w} \mid \text{Data}, \sigma^2, \lambda) = \mathcal{N}(\underline{w}; \underline{\mu}_{wID}, \underline{\underline{\Sigma}}_{wID})$$

where $\underline{\underline{\Sigma}}_{wID} = \left(\lambda \underline{\underline{I}} + \frac{1}{\sigma^2}\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}\right)^{-1}$

$$\underline{\mu}_{wID} = \underline{\underline{\Sigma}}_{wID} \frac{1}{\sigma^2} \underline{\underline{\tilde{X}}}^T \underline{y}$$

In this exercise $\sigma^2 = \lambda = 1$ (prior and observation noise are unit variance)

$$\underline{\underline{\tilde{X}}} = [x, 1], \quad \underline{w} = \begin{bmatrix} m \\ c \end{bmatrix} \implies \underline{\underline{\Sigma}}_{wID} = \left(\underline{\underline{I}} + \begin{bmatrix} x \\ 1 \end{bmatrix}[x,1]\right)^{-1} = \begin{pmatrix} x^2+1 & x \\ x & 2 \end{pmatrix}^{-1} \checkmark$$

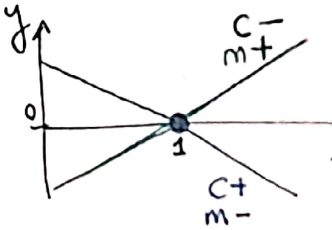
$$\underline{\mu}_{wID} = \begin{pmatrix} x^2+1 & x \\ x & 2 \end{pmatrix}^{-1} \begin{bmatrix} x \\ 1 \end{bmatrix} y \checkmark$$

c) we just need to plug this settings i) $x=0, y=0$ ii) $x=1, y=0$ iii) $x=100, y=100$ into Σ^{POST} and μ^{POST} calculated in a).

i) $x=0, y=0 \implies \Sigma^{POST} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, $\mu^{POST} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

data doesn't tell us anything about m: $y = mx + c$ and $x = 0$ This is why posterior mean and variance of m stays at prior (mean 0, variance 1). Data tells us that c is likely to be small (≈ 0), but the observation noise is quite large so only weak evidence for this. This is why posterior mean is 0 and posterior variance is $1/2$ (prior variance is 1).

ii) $x=1, y=0 \implies \Sigma^{POST} = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$, $\mu^{POST} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

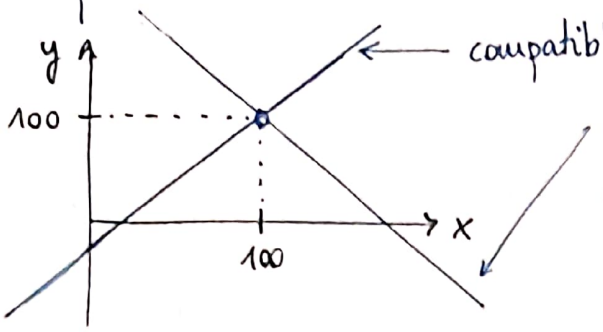


The data can be explained by either positive c and negative m or negative c & positive m: hence posterior covariance between m & c is negative ($-1/3$).

Dataset provides weak evidence that m and c are both small in magnitude and due to symmetry both still have posterior mean 0 (uncertainty has reduced from variance 1 to $2/3$).

iii) $x = 100$, $y = 100$ $\Rightarrow \Sigma^{POST} = \begin{bmatrix} 2/100^2 & -1/100 \\ -1/100 & 1 \end{bmatrix}$, $\mu^{POST} = \begin{bmatrix} 1 \\ 1/100 \end{bmatrix}$

The data $y=100$, $x=100$ rules out negative m:
We are therefore quite certain that m is close to 1
(posterior mean very close to this value and variance $\approx 2/100^2$)
The data give almost no information about c though with posterior mean and variance close to the prior.



← compatible with data

ruled out by prior on c as c would have to be massive in this case.

c.2) Predictive mean: use $y^* = mx^* + c + \varepsilon^*$

now average w.r.t. $\mathbb{P}(m, c, \varepsilon^* | x, y)$:

$$\mathbb{E}_{\mathbb{P}(m,c,\varepsilon^*|x,y)}\left[mx^* + c + \varepsilon^*\right] = \mu_m^{POST} x^* + \mu_c^{POST} + 0$$

↑ independent from x and y

i) predictive mean $= 0$ (horizontal line)

ii) predictive mean $= 0$ (horizontal line)

iii) predictive mean $= \frac{1}{100^2+2}\left(100^2 x^* + 100\right)$

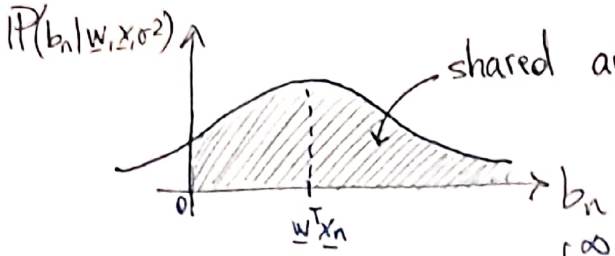
line with gradient very close to 1
and a small positive intercept.

10.

a) Info: $y_n = H(\underline{w}^T \underline{x}_n + \varepsilon_n)$ where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$

Let $b_n = \underline{w}^T \underline{x}_n + \varepsilon_n$, $b_n \sim \mathcal{N}(\underline{w}^T \underline{x}_n, \sigma^2)$

$\Rightarrow \mathbb{P}(y_n = 1 \mid \underline{w}, \underline{x}, \sigma^2) = \mathbb{P}(b_n > 0 \mid \underline{w}, \underline{x}, \sigma^2)$



$$\Rightarrow \mathbb{P}(y_n = 1 \mid \underline{w}, \underline{x}, \sigma^2) = \int_0^\infty \mathcal{N}(b_n; \underline{w}^T \underline{x}_n, \sigma^2)\, db_n$$

Let $v_n = -\dfrac{b_n - \underline{w}^T \underline{x}_n}{\sigma^2}$, and transform to integral over this quantity (v_n will be distributed to a standard normal)

- when $b_n = 0 \Rightarrow v_n = \dfrac{\underline{w}^T \underline{x}_n}{\sigma^2}$
- when $b_n = \infty \Rightarrow v_n = -\infty$

$$\Rightarrow \mathbb{P}(y_n = 1 \mid \underline{w}, \underline{x}_n, \sigma^2) = \int_{-\infty}^{\frac{\underline{w}^T \underline{x}_n}{\sigma^2}} \mathcal{N}(v_n; 0, 1)\, dv_n = \text{CDF}\left(\frac{\underline{w}^T \underline{x}_n}{\sigma^2}\right)$$

b) If $\sigma^2 \to \infty \Rightarrow \mathbb{P}(y_n = 1 \mid \underline{x}_n, \underline{w}, \sigma^2) = \text{CDF}(0) = 1/2$

The noise swamps $\underline{w}^T \underline{x}_n$ resulting in an output which is a coin toss.

11.

a) Let $\underline{w}_i = \beta \hat{\underline{w}}_i$ (β: magnitude, $\hat{\underline{w}}_i$: unit vector in direction of $\underline{w}_i$)

$$y_i(\underline{x}; \underline{\underline{W}}) = \frac{e^{\beta \hat{\underline{w}}_i^T \underline{x}}}{\sum_{k=1}^{K} e^{\beta \hat{\underline{w}}_k^T \underline{x}}} = \frac{\exp\left(\beta(\hat{\underline{w}}_i - \hat{\underline{w}}_{max})^T \underline{x}\right)}{\sum_{k=1}^{K} \exp\left(\beta(\hat{\underline{w}}_k - \hat{\underline{w}}_{max})^T \underline{x}\right)}$$

$\hat{\underline{w}}_{max} = \text{argmax}\, \hat{\underline{w}}_i^T \underline{x}$, $\hat{\underline{w}}_i \in \hat{\underline{w}}_1, \ldots, \hat{\underline{w}}_K$

$i \in \{1, \ldots, K\}$

$(\hat{\underline{w}}_i - \hat{\underline{w}}_{max})^T \underline{x}$ are a set of 'K scalars', one of which is zero and the others are negative

$\Rightarrow$ As $\beta \to \infty$, $e^{\beta(\hat{\underline{w}}_i - \hat{\underline{w}}_{max})} \longrightarrow \begin{cases} 0 & \text{if } \hat{\underline{w}}_i \neq \hat{\underline{w}}_{max} \\ \infty & \text{if } \hat{\underline{w}}_i = \hat{\underline{w}}_{max} \end{cases}$

(should it be 1? not ∞?)

$\Rightarrow y_i(\underline{x}, \underline{\underline{W}}) \longrightarrow \begin{cases} 0 & \text{if } i \neq i_{max} \\ 1 & \text{if } i = i_{max} \end{cases}$

This is a zero/one encoding of the argmax function, hence the general name "softmax":

i.e. as $\beta \to \infty$ $\quad y_i(\underline{x}, \underline{\underline{W}}) \longrightarrow \mathbb{1}\left(\underset{k}{\text{argmax}}\, \hat{\underline{w}}_k^T \underline{x} = i\right)$

$\mathbb{1}$: indicator function that takes the value 1 if the argument is true and is otherwise 0.

b) $y_i = P(t_i = 1 \mid \underline{\underline{W}}, \underline{x})$

$$P\left(\{t^{(n)}\}_{n=1}^{N} \mid \underline{\underline{W}}, \{x^{(n)}\}_{n=1}^{N}\right) = \prod_{n=1}^{N} P(t \mid \underline{\underline{W}}, x^{(n)}) =$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} y_k(\underline{x}^{(n)}, \underline{\underline{W}})^{t_k^{(n)}} = \exp\left(\sum_{n}^{N} \sum_{k}^{K} t_k^{(n)} \log y_k(\underline{x}^{(n)}, \underline{\underline{W}})\right)$$

$\Rightarrow$ cost function could be $\underset{\underline{\underline{W}}}{\text{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} t_k^{(n)} \log y_k(\underline{x}^{(n)}, \underline{\underline{W}})$

c) Consider $K=2$ (two classes)

$$y_1(\underline{x}, \underline{\underline{W}}) = \frac{e^{\underline{w}_1^T \underline{x}}}{e^{\underline{w}_1^T \underline{x}} + e^{\underline{w}_2^T \underline{x}}} = \frac{1}{1 + e^{(\underline{w}_2 - \underline{w}_1)^T x}} = \frac{1}{1 + e^{-\underline{v}^T \underline{x}}}$$

where $\underline{v} = \underline{w}_2 - \underline{w}_1$. This is equivalent to the expression for logistic regression applied to a two-class classification problem. However, notice that although the softmax and logistic functions are identical for $K=2$ classes, the parameterisation is different. The softmax version is over-parameterised having two sets of parameters whose difference affects the input-output function. For this reason the parameters of the softmax are not identifiable: adding the same vector to each weight $w_k \leftarrow w_k + b$ causes no change in the input-output function.