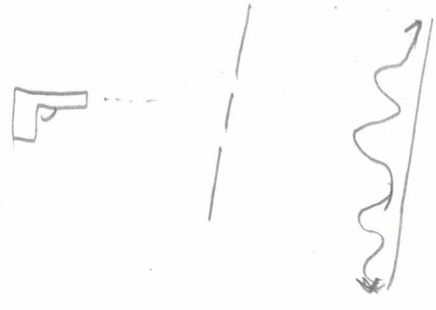
MATH 341/641 LEC 16

$$\overbrace{P(\theta|\vec{x})}^{\text{posterior}} = \frac{\overbrace{P(\vec{x}|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\vec{x})}_{\text{prior predictive}}}$$

In order to make the posterior non-degenerate, the prior must be non-degenerate. $P(\theta)$ is the distr of θ without ever seeing data. This is the big pill to swallow. Frequentists scream "it's not objective"! You can't just use your own ideas. If my $P(\theta)$ differs from your $P(\theta)$, we get different answers!!! Also is θ still considered a fixed value? One camp says yes, where $P(\theta|\vec{x})$ just represents uncertainty in θ. The other camp says no. Different θ's exists in quantum states like an electron cloud around an atom.



electron diffraction from double slit experiment with no immediate observer

$X_1, \ldots, X_n$

DGP: $\overset{iid}{\sim}$ Bern(θ) $\quad \vec{x} = \langle 0,1,1 \rangle$

let $\Theta_0 = \{0.5, 0.75\} \neq (0,1)$ we will consider $(0,1)$ later

$$P(\theta = 0.75 \mid \vec{x}) = \frac{P(\vec{x} \mid \theta = 0.75)\, P(\theta = 0.75)}{P(\vec{x} \mid \theta = 0.75)\, P(\theta = 0.75) + P(\vec{x} \mid \theta = 0.5)\, P(\theta = 0.5)}$$

we need $P(\Theta)$ to calculate the posterior prob. How about...

$$P(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.75 \\ 0.5 & \text{if } \theta = 0.5 \end{cases}$$

Generally...

$$P(\theta) = \frac{1}{|\Theta_0|}$$

This is called the "principle of indifference" or the "Laplace prior" (as it was his idea). This only works if Θ_0 is a finitely-sized set.

Now we can compute

If θ is a single value.. which is more likely θ?

$$P(\theta = 0.75 \mid \vec{x}) = \frac{P(\vec{x} \mid \theta = 0.75)\, \cancel{P(\theta = .75)}}{P(\vec{x} \mid \theta = .75)\cancel{P(\theta = .75)} + P(\vec{x} \mid \theta = 0.5)\cancel{P(\theta = .5)}} = \frac{0.25 \cdot 0.75^2}{0.25 \cdot 0.75^2 + 0.5^3} = 0.53 \leftarrow$$

But if θ is not a single value.. they exist in quantum state

$$P(\theta = 0.5 \mid \vec{x}) = \frac{P(\vec{x} \mid \theta = .5)\, \cancel{P(\theta = .5)}}{P(\vec{x} \mid \theta = .75)\cancel{P(\theta = .75)} + P(\vec{x} \mid \theta = 0.5)\cancel{P(\theta = .5)}} = \frac{0.5^3}{0.25 \cdot 0.75^2 + 0.5^3} = 0.47$$

Let's redo this with kernels. First of all...

$$p(\theta|\vec{x}) \propto p(\vec{x}|\theta)\,p(\theta)$$

$$\propto k(\vec{x}|\theta)\,k(\theta) = k(\theta|\vec{x})$$

Why? posterior is a distr in θ, anything else belongs to $c \Rightarrow \frac{1}{p(x)}$ is part of c

$$p(\theta) = \tfrac{1}{2}\mathbb{1}_{\theta=.75} + \tfrac{1}{2}\mathbb{1}_{\theta=0.5} \propto \mathbb{1}_{\theta=.75} + \mathbb{1}_{\theta=.5}$$

$$p(\vec{x}|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$k(\theta|\vec{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\left(\mathbb{1}_{\theta=.75} + \mathbb{1}_{\theta=.5}\right)$$

$\sum x_i = 2$, $\Theta = \{0.5, 0.75\}$

$k(.75|\vec{x}) = .75^2\,.25^1 = .141$

$k(.5|\vec{x}) = .5^3 = .125$

which is more likely value of θ? Note: Don't need $p(\theta|\vec{x})$ to answer, just need $k(\theta|\vec{x})$!

$$\Rightarrow c = \frac{1}{\sum_{x \in \Theta} k(\theta|x)} = \frac{1}{.141+.125} = 3.76 \Rightarrow p(.75|\vec{x}) = 3.76 \cdot .141 = .53,\quad p(.5|\vec{x}) = 3.76 \cdot .125 = .47$$

Point Estimation: provide best guess of θ

Define $\hat{\theta}^{MAP} := \underset{\theta \in \Theta}{\operatorname{argmax}}\{p(\theta|x)\} = \operatorname{argmax}\{p(x|\theta)\,p(\theta)\} = \operatorname{argmax}\{k(x|\theta)\,k(\theta)\}$

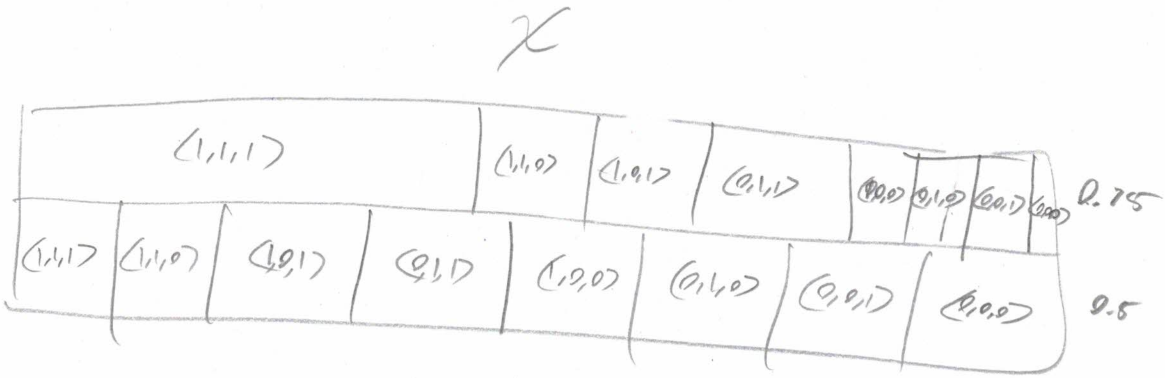
maximum a posteriori

if $p(\theta)$ is principle of indifference

$= \operatorname{argmax}\{k(x|\theta)\}$

Let's see the whole problem. $\vec{x} \in \mathcal{X}$, $\theta \in \Theta$. Illustrate $P(\mathcal{X}, \Theta)$

$\mathcal{X} = \{\langle 0,0,0\rangle, \langle 1,0,0\rangle, \langle 0,1,0\rangle, \langle 0,0,1\rangle, \langle 1,1,0\rangle, \langle 1,0,1\rangle, \langle 0,1,1\rangle, \langle 1,1,1\rangle\}$



$P(\vec{x} = \langle 1,1,1\rangle \mid \theta = .75) = .75^3 = .422$

$P(\vec{x} = \langle 1,1,0\rangle \mid \vec{\theta} = .75) = .75^2 \cdot .25 = .141$

$P(\vec{x} = \langle 1,0,0\rangle \mid \vec{\theta} = .75) = .75 \cdot .25^2 = .047$

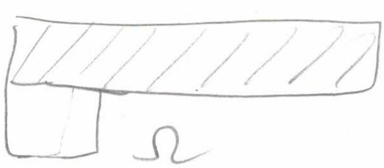
$P(\vec{x} = \langle 0,0,0\rangle \mid \vec{\theta} = .75) = .25^3 = .016$

$P(\vec{x} = \langle 1,1,1\rangle \mid \theta = .5) = .125$

How to visualize

$P(\theta = .75 \mid \vec{x} = \langle 1,1,1\rangle)$?

$$\frac{\{\langle 1,1,1\rangle, .75\}}{\{\langle 1,1,1\rangle, .75\} + \{\langle 1,1,1\rangle, .5\}}$$



$\Theta_0 = \{0.5, 0.75\} \neq \Theta = (0,1)$ Not the full parameter space.

and $\hat{\theta}^{MLE} \notin \Theta_0$! Ridiculous...

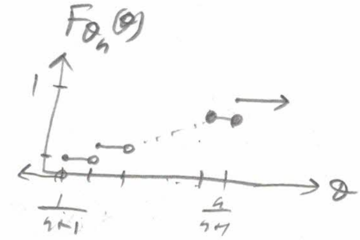
What does Laplace's Prior of Indifference look like?

Let's use our understanding of 340 to derive this...

$\Theta_0(3) = \left\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\right\} \Rightarrow P_3(\theta) = U\left(\left\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\right\}\right)$

$\Theta_0(4) = \left\{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\right\} \Rightarrow P_4(\theta) = U\left(\left\{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\right\}\right)$

$\Theta_0(n) = \left\{\frac{1}{n+1}, \frac{2}{n+1}, \dots, \frac{n}{n+1}\right\} \Rightarrow P_n(\theta) = U\left(\left\{\frac{1}{n+1}, \frac{2}{n+1}, \dots, \frac{n}{n+1}\right\}\right)$



$$F_{\theta_n}(\theta) = \begin{cases} 0 & \text{if } \theta < \frac{1}{n+1} \\ \frac{1}{n}\lfloor (n+1)\theta \rfloor & \text{if } \theta \in \left[\frac{1}{n+1}, \frac{n}{n+1}\right] \\ 1 & \text{if } \theta > \frac{n}{n+1} \end{cases}$$

$n = 9 \Rightarrow n+1 = 10$

$\theta = 0.05 \quad 10\theta = .5$

$\theta = .15 \quad \lfloor 10\theta \rfloor = \lfloor 1.5 \rfloor = 1$

$$\lim F_{\theta_n}(\theta) = \begin{cases} 0 & \text{if } \theta < 0 \\ \theta & \text{if } \theta \in (0,1) \\ 1 & \text{if } \theta > 1 \end{cases}$$

$\Rightarrow \theta_n \sim U(0,1)$

$\theta = \lim \Theta_0(n) \Rightarrow f(\theta) = U(0,1)$

Switched from discrete $\to$ continuous

DGP: $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$$\Rightarrow f(\theta \mid x) = \frac{\underset{\text{cont.}}{p(x \mid \theta)}\, \underset{\text{cont.}}{f(\theta)}}{p(x)}$$

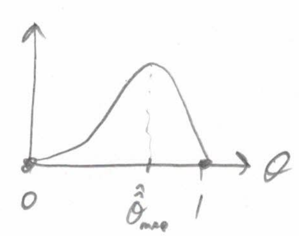
$\propto k(x \mid \theta)\, k(\theta)$

$\propto \theta^{\sum x_i}(1-\theta)^{n - \sum x_i} \mathbb{1}_{\theta \in (0,1)}$

$= \theta^{\sum x_i + 1 - 1}(1-\theta)^{n - \sum x_i + 1 - 1} \mathbb{1}_{\theta \in (0,1)}$

$\propto \text{Beta}\left(\sum x_i + 1,\ n - \sum x_i + 1\right)$

eg $\vec{x} = \langle 1, 1, 0 \rangle$, $\theta \sim U(0,1) \Rightarrow \theta \mid \vec{x} \sim \text{Beta}(3, 2)$



What is $\hat{\theta}^{MAP}$ if $\theta | X \sim \text{Beta}(\alpha, \beta)$

$$\underset{\theta \in [0,1]}{\text{argmax}} \left\{ \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right\} = \text{argmax} \left\{ \theta^{\alpha-1}(1-\theta)^{\beta-1} \right\}$$

$$\frac{d}{d\theta}\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] = -(\beta-1)\theta^{\alpha-1}(1-\theta)^{\beta-2} + (\alpha-1)(1-\theta)^{\beta-1}\theta^{\alpha-2} \overset{\text{set}}{=} 0$$

$$= \cancel{(1-\theta)^{\beta-2}\theta^{\alpha-2}}\left((\alpha-1)(1-\theta) - (\beta-1)\theta\right) = 0$$

$$\Rightarrow (\alpha-1)(1-\theta) = (\beta-1)\theta \Rightarrow \frac{1-\theta}{\theta} = \frac{\beta-1}{\alpha-1} \Rightarrow \frac{1}{\theta} - 1 = \frac{\beta-1}{\alpha-1} \Rightarrow \frac{1}{\theta} = \frac{\beta-1}{\alpha-1} + \frac{\alpha-1}{\alpha-1}$$

$$\Rightarrow \frac{1}{\theta} = \frac{\beta+\alpha-2}{\alpha-1}$$

$$\Rightarrow \theta_* = \frac{\alpha-1}{\alpha+\beta-2}$$

Second deriv test confirms $\alpha, \beta \geq 1$ for mode to exist otherwise it's min.

$\Rightarrow \vec{x} = \langle 1,1,0 \rangle$, $\theta \sim U(0,1) \Rightarrow \theta | \vec{x} \sim \text{Beta}(3,2) \Rightarrow \hat{\theta}^{MAP} = \frac{3-1}{3+2-2} = \frac{2}{3} = \hat{\theta}^{MLE}$

Is there another point estimate?

Consider the following prob. problem. Let X be a r.v.

Let $g(t) := E[(X-t)^2]$. What is $\text{argmin}\{g(t)\}$?

$$g'(t) = \frac{d}{dt}\left[E[X^2 - 2tX + t^2]\right] = \frac{d}{dt}\left[-2t\mu + 2t\right] = -2\mu + 2t \overset{\text{set}}{=} 0 \Rightarrow t = \mu.$$

Where we going with this?

Let $\hat{\theta}^{MMSE} := E[\theta | \vec{x}]$ which minimizes $E[(\theta - \hat{\theta})^2 | \vec{x}] = MSE[\hat{\theta} | \vec{x}]$

(RISK under L^2 loss)

MMSE = "min. mean sq. error"

$\vec{x} = \langle 1,1,0 \rangle$, $\theta \sim U(0,1) \Rightarrow \hat{\theta}^{MMSE} = \frac{\alpha}{\alpha+\beta} = \frac{3}{3+2} = \frac{3}{5}$

$\vec{x} = \langle 0,0,0 \rangle$, $\theta \sim U(0,1)$

$\Rightarrow \theta | X \sim \text{Beta}(1,4) \Rightarrow \hat{\theta}^{MAP} = 0$, $\hat{\theta}^{MMSE} = \frac{1}{1+4} = 0.2$

Gets around problem with MLE!! useful!