Lec 8 310-03-02 2/27/16 $\theta|X \sim \text{Beta}(\alpha+x, \beta+n-x)$ $\hat{\theta}_{MMSE} = \rho E(\theta) + (1-\rho)\hat{\theta}_{MLE}$ $\rho = \frac{\alpha+\beta}{\alpha+\beta+n}$ (1)

How do we really be uninformative?? How about $\alpha=\beta=0$? i.e. $\rho=0$

$$\rho = \frac{0+0}{0+0+n} = 0 \Rightarrow \hat{\theta}_{MMSE} = \frac{X}{n} = \hat{\theta}_{MLE}.$$ We get frequentist estimate with bayesian interpretation!

(HW3 #2e)

But $\theta \sim \text{Beta}(\alpha,\beta)$ Param space $\alpha, \beta \in (0,\infty)$ $\alpha, \beta \neq 0$!

Why? $$\theta \sim \text{Beta}(\overset{0}{\alpha}, \overset{0}{\beta}) = \frac{1}{B(0,0)}\theta^{0-1}(1-\theta)^{0-1} = \frac{\Gamma(0)\Gamma(0)}{\Gamma(0)}\frac{1}{\theta(1-\theta)}$$

$$\Gamma(t) := \int_0^\infty x^{t-1}e^{-x}\,dx \Rightarrow \Gamma(0) = \int_0^\infty \frac{e^{-x}}{x}\,dx = \infty$$

$\int u\,dv = uv - \int v\,du$

Further $\text{Beta}(0,0) \propto \frac{1}{\theta(1-\theta)}$ $$\int_0^1 \frac{1}{\theta(1-\theta)}\,d\theta = \left[\ln(\theta) - \ln(1-\theta)\right]_0^1 = \infty$$

Is an "improper density" Since it does not integrate to 1. $\theta \sim \text{Beta}(0,0)$ is called an "improper prior"

However, $\theta|X = \text{Beta}(\alpha+x, \beta+n-x) = \text{Beta}(x, n-x)$

is a "proper posterior density" as long as $x \neq 0$ & $x \neq n$ i.e. at least one success and at least one failure in the data.

Are improper priors "legal"?

Yes, but you need to ensure posterior is proper o/t inference not valid.

~~[illegible]~~ [2]

Why not just pick $\alpha \approx 0$ and $\beta \approx 0$ eg $\alpha = 1\times10^{-10}$, $\beta = 1\times10^{-10}$?

You can! Conjugacy is ugliest but it makes the theorists happy.

Beta $(0,0)$ is known as the Haldane prior (1932). This is the limit $\alpha, \beta \to 0$.

Downsides: if $x = 0$ or $x = n$ $\Rightarrow$ $P(\theta|x) = \text{Deg}(0)$ or $P(\theta|x) = \text{Deg}(1)$.

Bad! And ... anti-Bayesian in some way.

Also, it is equivalent of "seeing" -1 successes and -1 failures. In a way it's pushing you toward $\theta = 0$ or $\theta = 1$.

Jaynes "prior state of knowledge of complete ignorance i.e. you don't even know if successes or failures are even possible".

Bayes-Laplace-Objective prior: ignorance but you are confident both successes and failures are possible.

Not explicitly covered

Recall likelihood function

$$P(X;\theta) = \mathcal{L}(\theta;x) = \prod_{i=1}^{n} P(X_i;\theta)$$

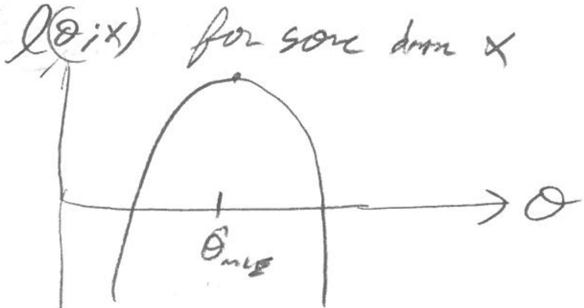
$\ell(\theta;x)$

log-likelihood

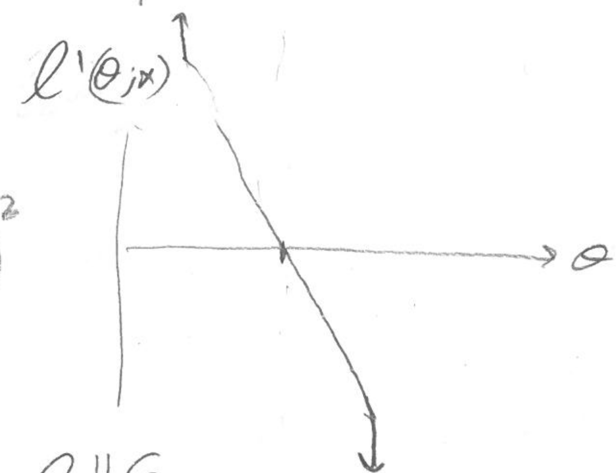
to get MLE

$$\hat{\theta}_{MLE} := \underset{\theta \in \Theta}{\text{argmax}} \left\{ \prod_{i=1}^{n} P(X_i;\theta) \right\} = \underset{\theta \in \Theta}{\text{argmax}} \left\{ \overbrace{\sum_{i=1}^{n} \ln\left(P(X_i;\theta)\right)} \right\}$$

Set $\underbrace{\frac{d}{d\theta}\left[\ell(\theta;x)\right]}_{\text{score function}} = 0$ and hope you find a global max



$\ell'(\theta;x)$



$I(\theta) := Var\left[\ell'(\theta;x)\right]$

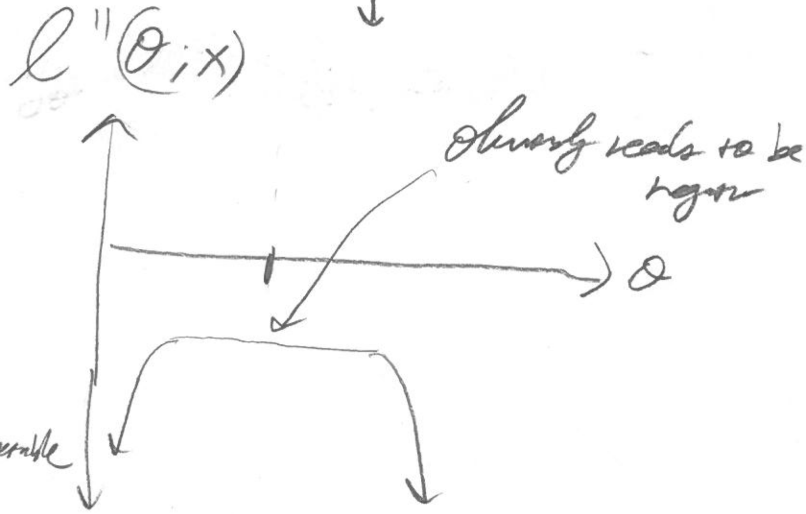
$= E\left[(\ell'(\theta;x))^2\right] - \underbrace{\left(E\left[\ell(\theta;x)\right]\right)^2}_{\text{shown to be zero}}$

$= E\left[\ell'(\theta;x)^2\right]$

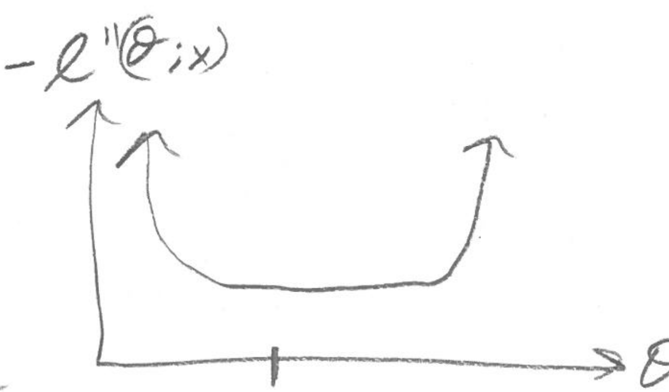
with regularity conditions

"Fisher information"

more interpretable



$I(\theta) := E\left[-\ell''(\theta;x)\right]$



On avg over all data sets, how large is this second derivative? If large, lots of information in data about MLE. If small, need lots of data to get to an MLE.

In fact $\hat{\theta}_{MLE} \sim N\left(\theta, \left(\sqrt{\frac{1}{I(\theta)}}\right)^2\right)$

If information large $\hat{\theta}_{MLE} \approx \theta$. If small, large variance...

e.g. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$$I(\theta) = E[-l''(\theta; x)]$$

$$l(\theta; x) = \sum_{i=1}^n \ln p(x_i; \theta) = \sum_{i=1}^n \ln\left(\theta^{x_i}(1-\theta)^{1-x_i}\right) = \sum_{i=1}^n x_i \ln(\theta) + (1-x_i)\ln(1-\theta)$$

$$l'(\theta; x) = \frac{\partial}{\partial\theta}[\;] = \sum_{i=1}^n \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}$$

$$l''(\theta; x) = \frac{\partial}{\partial\theta}[\;] = \sum_{i=1}^n -\frac{x_i}{\theta^2} - \frac{1-x_i}{(1-\theta)^2} = \frac{-n\bar{x}}{\theta^2} - \frac{n-n\bar{x}}{(1-\theta)^2}$$

$$-l''(\theta; x) = \frac{n\bar{x}}{\theta^2} - \frac{n-n\bar{x}}{(1-\theta)^2} = n\left(\frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2}\right)$$

$$I(\theta) := E[\;] = n\left(\frac{E(\bar{X})}{\theta^2} + \frac{1-E(\bar{X})}{(1-\theta)^2}\right) = n\left(\frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2}\right) = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) = n\left(\frac{1-\theta}{\theta(1-\theta)} + \frac{\theta}{\theta(1-\theta)}\right) = \boxed{\frac{n}{\theta(1-\theta)}}$$

θ is a const.

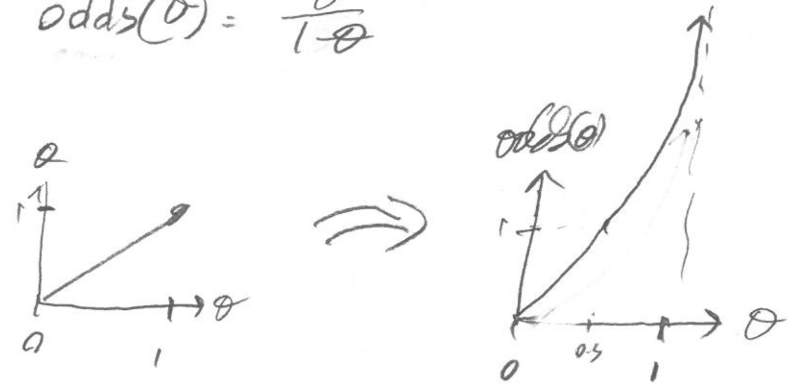
Why do we care?

Recall $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$, $\underbrace{\theta \sim U(0,1)}_{\text{principle of indifference}}$

Why is principle of indifference bad?

If someone works on a different scale

e.g. $\text{odds}(\theta) = \frac{\theta}{1-\theta}$



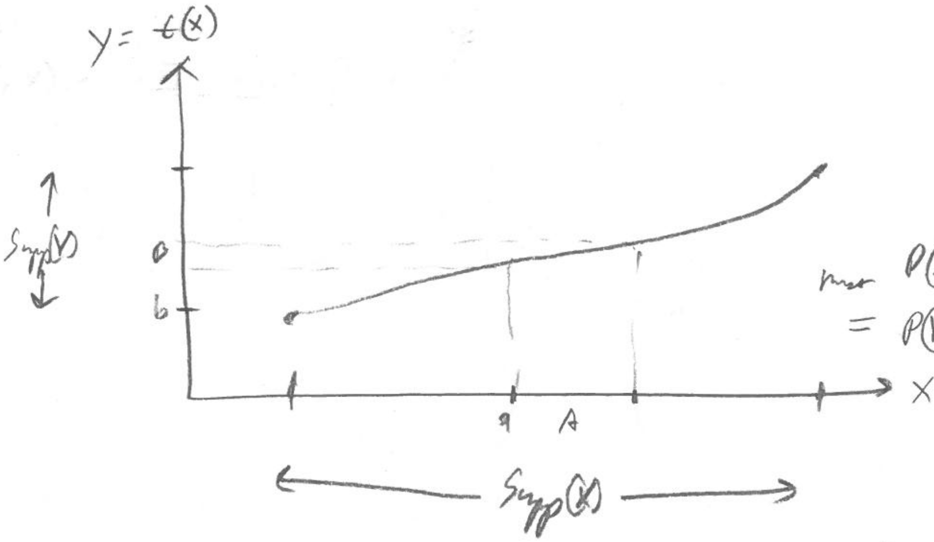
If I'm indifferent about θ, shouldn't I be indifferent about the odds(θ)?

$f(\theta) = 1 \ \mathbb{1}_{\theta \in (0,1)}$

$g(\theta) = \frac{\theta}{1-\theta}$ What is PDF of $g(\theta)$, a monotonic $\Rightarrow$ 1:1 function of θ?

Theory of transformation of variables. Given $Y = t(X)$ and $f_X(x)$

find $f_Y(y)$



must $P(X \in A) \approx f_X(a)\, A$
$= P(Y \in B) \approx f_Y(b)\, B$

$X = t^{-1}(Y)$

$$\Rightarrow f_X(x)\,|dx| = f_Y(y)\,|dy| \Rightarrow f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right|$$

$$\Rightarrow \quad = f_X(t^{-1}(y)) \left|\frac{dx}{dy}\right|$$

BAD if non-monotonic

$$= f_X(t^{-1}(y)) \left| \frac{d}{dy}\left[t^{-1}(y)\right] \right|$$

$$f_{odds}(o) = f_\theta\left(t^{-1}(o)\right)\left|\frac{d}{do}\left[t^{-1}(o)\right]\right| = (1)\left|\frac{1}{(o+1)^2}\right| = \frac{1}{(o+1)^2}\mathbb{1}_{o\in(0,\infty)}$$

$$o = \frac{\theta}{1-\theta} \Rightarrow o - o\theta = \theta \Rightarrow o = o\theta + \theta \Rightarrow o = \theta(o+1) \Rightarrow \theta = \frac{o}{o+1} = t^{-1}(o)$$

Is this a PDF? $\displaystyle\int_0^\infty \frac{1}{(o+1)^2}\,do \underset{=\left[\frac{o}{o+1}\right]_0^\infty = 1-0}{=} 1$ ✓

$O \not\sim U(0,1)$. Is this a contradiction??? If indifferent about $\theta \overset{?}{\Rightarrow}$ indifferent odds(θ)

Consider $l := \text{logit}(\theta) := \ln(\text{odds}(\theta)) = \ln\left(\frac{\theta}{1-\theta}\right)$ $\quad Supp(l) = \mathbb{R}$

which is a very well-used "link function" used in "logistic regression" or logit regression.

$$f_l(l) = f_X\Big(\overbrace{\frac{e^l}{1+e^l}}^{=1}\Big)\left|\frac{d}{dl}\left[\frac{e^l}{1+e^l}\right]\right| = \frac{e^l}{(1+e^l)^2} \neq U(0,1)!$$

$$l = \ln\frac{\theta}{1-\theta} \Rightarrow e^l = \frac{\theta}{1-\theta} \Rightarrow e^l - e^l\theta = \theta \Rightarrow e^l = \theta + e^l\theta \Rightarrow e^l = \theta(1+e^l)$$
$$\Rightarrow \theta = \frac{e^l}{1+e^l} = t^{-1}(l)$$
$$:= \text{expit}(l)$$

The second you go to another scale, principle of indifference goes bye-bye! It's as if you're weighting logits near 0 more heavily than logits elsewhere...

Note: $\theta \sim \text{Beta}(0,0) \propto \frac{1}{\theta(1-\theta)}$ Haldane prior

$l = \text{logit}(\theta) := \ln\left(\frac{\theta}{1-\theta}\right)$

$$f_l(l) = f_X\left(\frac{e^l}{1+e^l}\right)\frac{e^l}{(1+e^l)^2} = \frac{1}{\left(\frac{e^l}{1+e^l}\right)\left(1-\frac{e^l}{1+e^l}\right)}\,\frac{e^l}{(1+e^l)^2} = \frac{\cancel{e^l}}{\left(\frac{\cancel{e^l}}{\cancel{1+e^l}}\right)\left(\frac{1}{\cancel{1+e^l}}\right)\cancel{(1+e^l)^2}} = 1 \in \mathbb{R}$$

$\Rightarrow$ uniform prior on logit scale (but improper).

But is there $\theta \sim p(\theta)$ s.t. it is the same on all transformed scales? So you don't have to worry if it changes a little bit? <u>YES</u>

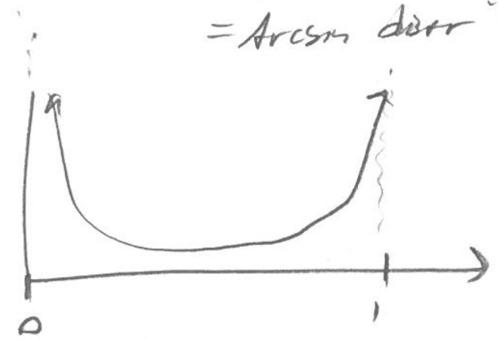
Let $p(\theta) \propto \sqrt{I(\theta)}$ ← Fisher Information of the <u>likelihood</u>

e.g in the <u>Binomial likelihood model</u>

$$p(\theta) \propto \sqrt{\frac{n}{\theta(1-\theta)}} = n^{\frac{1}{2}} \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}} \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}}$$

$$\propto \text{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

"= Arcsin distr"



Beta(0,0)	<	Beta$(\frac{1}{2},\frac{1}{2})$	<	Beta(1,1) = U(0,1)
improper Haldane		Jeffreys prior		objective on the θ scale
complete ignorance		"half" ignorance		ignorance

So is $\theta \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ really invariant to arbitrary monotone transform?

How about $t(\theta) = \text{odds}(\theta) = \frac{\theta}{1-\theta}$?

$$f_{odds}(\theta) = f_\theta\left(\frac{\theta}{\theta+1}\right) \left| \frac{d}{d\theta}\left[\frac{\theta}{\theta+1}\right] \right| = \frac{1}{B\left(\frac{1}{2},\frac{1}{2}\right)} \left(\frac{\theta}{\theta+1}\right)^{-\frac{1}{2}} \left(\frac{\theta}{\theta+1}\right)^{-\frac{1}{2}} \left| \frac{1}{(\theta+1)^2} \right|$$

$t^{-1}(\theta)$

$$\propto \frac{\theta+1}{\theta} \frac{1}{(\theta+1)^2} = \frac{1}{\theta(\theta+1)}$$

$$P(\theta) \propto \sqrt{f_{odds}(\theta)} = \sqrt{\frac{1}{\theta(\theta+1)}} = \theta^{-\frac{1}{2}}(\theta+1)^{-\frac{1}{2}} \propto \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

MAGIC!!!

Proof of Invariance of Jeffreys Prior

Consider $\theta \sim P(\theta)$ a prior s.t. $P(\theta) \propto \sqrt{I(\theta)}$

i.e. the Jeffreys prior.

WTS $P(\phi)) \propto \sqrt{I(\phi)}$ for arbitrary $\phi = t(\theta)$

We know
$$P_\phi(\phi) = P_\theta\left(t^{-1}(\phi)\right) \left| \frac{d\theta}{d\phi} \right|$$
$$= P_\theta(\theta) \left| \frac{d\theta}{d\phi} \right| \propto \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right|$$
$$= \sqrt{I(\theta) \left(\frac{d\theta}{d\phi}\right)^2}$$

$$= \sqrt{E\left[\left(\ell'(\theta;x)\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2}$$

$$= \sqrt{E\left[\left(\frac{d}{d\theta}\left[\ln(\mathcal{L})\right]\right)^2\left(\frac{d\theta}{d\phi}\right)^2\right]}$$

$$= \sqrt{E\left[\left(\frac{d\ln(\mathcal{L})}{d\theta}\frac{d\theta}{d\phi}\right)^2\right]} = \sqrt{E\left[\left(\frac{d\ln\mathcal{L}}{d\phi}\right)^2\right]} = \sqrt{I(\phi)}\checkmark$$

So Beta$\left(\frac{1}{2},\frac{1}{2}\right)$ is the Jeffreys / Invariant prior

and thus makes it a "natural" choice.

Another proof

$$p(\theta) \propto \sqrt{I(\theta)} \Rightarrow$$

WTS $p(\phi) \propto \sqrt{I(\phi)}$ for arbitrary $\phi := h(\theta)$

We know $p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|$

If this is $\propto \sqrt{I(\phi)}$ we're done

$$= \sqrt{E\left[-\left[\frac{d^2}{d\phi^2}\left[\underbrace{\ln p(x|\phi)}_{g(\phi)}\right]\right]\right]} = \sqrt{E\left[-\frac{d}{d\phi^2}\left[g(\phi)\right]\right]}$$

let $\theta = h^{-1}(\phi)$