Lec 10 MATH 341/641

We say the Beta distr is the "conjugate prior" for the iid Bernoulli DGP. Conjugacy means prior and posterior are same rv (but different parameter values updated by the data).

$$\text{Beta}(\alpha,\beta) \xrightarrow{\vec{x}} \text{Beta}\left(\alpha + \underbrace{\textstyle\sum x_i}_{\#\text{successes}},\ \beta + n - \underbrace{\textstyle\sum x_i}_{\#\text{failures}}\right)$$

(# pseudosuccesses, # pseudofailures)

What is the interpretation of the values of the hyperparameters?

The prior's hyperparameters are like observing fake data, where $n_0 = \alpha+\beta$ # pseudotrials

If we employ Laplace's prior of indifference $U(0,1) = \text{Beta}(\alpha=1,\beta=1)$

$\Rightarrow n_0 = 2$ pseudotrials. This is weird. Further, consider

$$\hat{\theta}^{\text{MMSE}} = \frac{\alpha+\sum x_i}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n}\cdot\frac{\alpha+\beta}{\alpha+\beta} + \frac{\sum x_i}{\alpha+\beta+n}\cdot\frac{n}{n}$$

$$= \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\,\frac{\sum x_i}{n}$$

$$= \varrho\, E[\theta] + (1-\varrho)\,\hat{\theta}^{\text{MLE}}$$

($E[\theta]$: prior guess; $\hat{\theta}^{\text{MLE}}$: pure data estimate)

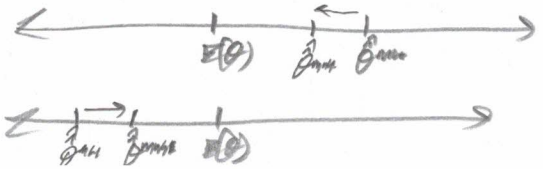
Let $\varrho := \frac{\alpha+\beta}{\alpha+\beta+n}$, the shrinkage metric, measures strength of prior $\in [0,1]$ on the most popular point estimate.

e.g. $\alpha=\beta=1$, $n=3 \Rightarrow \varrho = \frac{2}{5}$

$\Rightarrow$ 40% weight on prior, 60% weight on data

$\lim_{n\to\infty} \hat{\theta}^{\text{MMSE}} = \hat{\theta}^{\text{MLE}}$ since $\lim_{n\to\infty} \varrho = 0$

$\hat{\theta}^{\text{MMSE}}$ is called a "shrinkage estimator" since it shrinks toward $E(\theta)$



Prop: $X \sim \text{Bin}(n, \theta)$, $f(\theta) = \text{Beta}(\alpha, \beta)$

$$f(\theta|x) \propto p(x|\theta) f(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} 1_{\theta \in (0,1)}$$

$$\propto \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1} 1_{\theta \in (0,1)}$$

$$\propto \text{Beta}(\alpha+x, \beta+n-x)$$

Same as before although $x = \sum x_i$ what is exactly what it is

Since $\text{Binomial}(n, \theta)$ = sum of n iid $\text{Bern}(\theta)$

Let's now ask a different question: You see the n trials and now you want to predict how the future n_* trials are distributed.

$$\underbrace{\frac{e}{1}\ \frac{1}{2}\ \cdots\ \frac{1}{n}}_{\text{1 2 } \cdots \text{ } n} \Bigg| \underbrace{\frac{}{1}\ \frac{}{2}\ \cdots\ \frac{}{n_*}}_{\text{1 2 } \cdots \text{ } n_*}$$

How is X_* distributed?

This is called "prediction" or "forecasting" and is mainly the subject of MATH 342/642. Statistical Inference focuses on explanation of models, not prediction. But Bayesian inference gives you it for free!

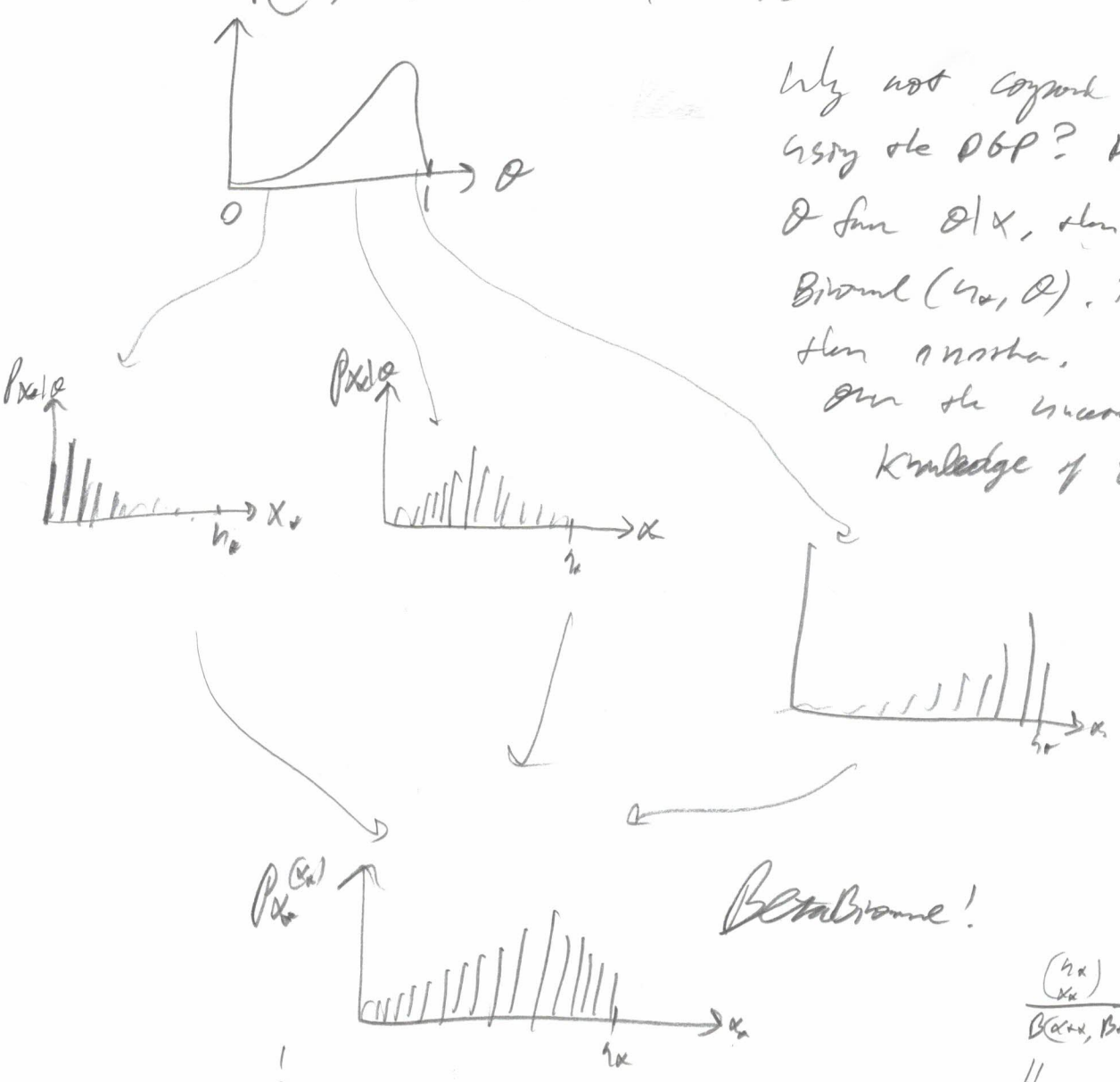
Obviously if θ is known, $X_* \sim \text{Bin}(n_*, \theta)$.

So what's the ~~next~~ last step $X_* \sim \text{Bin}(n_*, \hat{\theta}_{MLE})$?

Yes, in frequentist statistics, this is the best you can do, but it ignores the uncertainty in the estimate $\hat{\theta}_{MLE}$! So it's not really correct.

In Bayesian, you have an idea of the uncertainty in θ after the data is seen. It's called the posterior!

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



Why not compute this distribution using the DGP? Draw a θ from $\theta|x$, then sample a $\text{Binom}(n_n, \theta)$. Then another, then another. This averages over the uncertainty in the knowledge of θ.

BetaBinomial!

$$\frac{\binom{n_n}{x_n}}{\frac{B(\alpha+x, \beta+n-x)}{B(\alpha+x, \beta+n-x)}} B(\alpha+x+x_n, \beta+n-x+n_n-x_n)$$

$$P(x_n|x) = \int_0^1 \underbrace{P(x_n|\theta)}_{\text{Bin}(n_n,\theta)} \underbrace{f(\theta|x)}_{\text{Beta}(\alpha+x,\beta+n-x)} d\theta = \text{BetaBinom}(n_n, \alpha+x, \beta+n-x)$$

Proof in MATH 340

Concrete Example. Baseball player has $n=10$ at bats. $x=6$ are hits.

What's the prob he will have $x_n=17$ hits or the next $n_n=37$ at bats?

Assume Uniform prior

$$P(x_n|x=6) = \text{BetaBinom}(37, 1+6, 1+4)$$

$$P(x_n=17|x=6) = \frac{\binom{37}{17}}{B(7,5)} B(24,25) = \text{dbetabinom}(17, 37, 7, 5) =$$

(x_n, n_n, $\alpha+x$, $\beta+n-x$)

↑ needed for next

What's the prob he gets 17 hits or less on the next 37 at bats?

$$P(X_\alpha \le 17, \alpha = 6) = p\,\text{betabinom}(17, 37, 7, 5)$$

(17 = max x_α; 37 = n_α; 7 = $\alpha+\alpha$; 5 = $\beta+n-\alpha$)

will realize to either value of

What if $n_\alpha = 1$? $X_\alpha \sim \text{Bernoulli}(\theta_\alpha)$ since it is 0 or 1. What is θ_α?

$$P(X_\alpha \mid \alpha) = \text{BetaBinom}(1, \alpha+\alpha, \beta+n-\alpha) = \frac{\binom{1}{x_\alpha}}{B(\alpha+\alpha, \beta+n-\alpha)} B(x_\alpha + \alpha + \alpha,\ 1 + n_\alpha - x_\alpha + \beta + n - \alpha)$$

($n_\alpha = 1$)

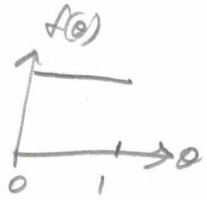
$$\theta_\alpha = P(X_\alpha = 1 \mid \alpha) = \frac{\binom{1}{1}}{B(\alpha+\alpha, \beta+n-\alpha)} B(1+\alpha+\alpha, \beta+n-\alpha)$$

$$= \frac{1}{\frac{\Gamma(\alpha+\alpha)\Gamma(\beta+n-\alpha)}{\Gamma(\alpha+\beta+n)}} \cdot \frac{\Gamma(1+\alpha+\alpha)\Gamma(\beta+n-\alpha)}{\frac{\Gamma(1+\alpha+\beta+n)}{(\alpha+\beta+n)\Gamma(\alpha+\beta+n)}} = \frac{\alpha+\alpha}{\alpha+\beta+n} = \hat{\theta}_{\text{mmse}}$$

The prob the next at bat we will a hit is $\frac{\alpha+\alpha}{\alpha+\beta+n}$

Priors are subjective. But can we make them "objective" to minimize their effect on the inference. Have cake and eat it too.

Laplace's prior is "flat" i.e. it doesn't give any special preference to any of the θ's.



But we also say it implies $E(\theta) = 0.5$ and $n_0 = 2$ one $x=1$, one $x=0$.

$f(\theta|x) = Beta(\alpha + x, b + n - x)$

This doesn't seem too objective! Further...

Let's say we care about inference for $\phi = g(\theta) = \frac{1-\theta}{\theta}$, odds against

1:1 transformation

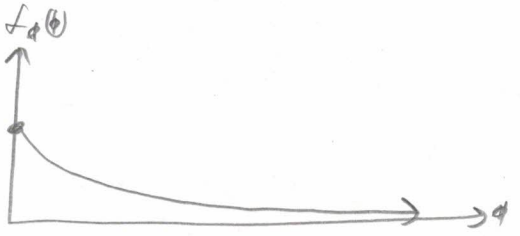
Where does Laplace's prior reign on this new scale?

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left|\frac{\partial}{\partial\phi}[g^{-1}(\phi)]\right| = \mathbb{1}_{\frac{1}{\phi+1} \in (0,1)} \frac{1}{(\phi+1)^2} = \frac{1}{(\phi+1)^2}\mathbb{1}_{\phi+1 \in (1,\infty)} = \frac{1}{(\phi+1)^2}\mathbb{1}_{\phi \in (0,\infty)}$$

$$\phi = \frac{1}{\theta} - 1 \Rightarrow \phi + 1 = \frac{1}{\theta} \Rightarrow \theta = \frac{1}{\phi+1} = g^{-1}(\phi) \qquad \left|\frac{d}{d\phi}\left[-(\phi+1)^{-2}\right]\right| = \frac{1}{(\phi+1)^2}$$

$\propto F_{2/2}$ dist

$= BetaPrime(1,1)$



which is not flat!! Further, there is no way to have a flat PDF on $(0,\infty)$!

Difference prior

Laplace's is final to a specific parameterization

You are saying small odds against values are more likely than large

What happens in this case??