RECAP

- A point estimator $\hat{\Theta} = h(X_1, \ldots X_n)$
- $B(\hat{\Theta}) = E[\hat{\Theta}] - \theta^*$
- $MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta^*)^2]$. Furthermore, $MSE(\hat{\Theta}) = Var(\hat{\Theta}) + Bias(\hat{\Theta})^2$
- Consistent and Strongly consistent estimators.
- Esitmators for mean and Variance
- MLE Estimators

$$\begin{aligned} \hat{\Theta}_{ML} &= \arg\max_{\theta} L(x_1, \ldots, x_n; \theta) \\ &= \arg\max_{\theta} logL(x_1, \ldots, x_n; \theta) \end{aligned}$$

Bayesian Inference with posterior distribution

- In Bayesian Inference we aim to extract information about unknown quantity θ^* based on observing a collection $X = (x_1, x_2, \dots x_n)$ using Bayes rule.
- We model uncertainty about θ^* using a random variable Θ.
- The nature of Θ changes as we collect more data, reducing the uncertainty in θ^*
- Bayes rule: $\{$posterior on $\Theta\} \propto \{$liklihood of $X\}\times$ $\{$prior on $\theta\}$
- Θ and X each could be continuous or discrete variables, and vice versa case are analogously obtained.

Bayes rule revisited .... revisited

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)} \quad (X, \Theta \text{ continuous})$$

$$p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)} \quad (X, \Theta \text{ discrete})$$

$$p_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{f_X(x)} \quad (X \text{ cont}, \Theta \text{ discrete})$$

$$f_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{p_X(x)} \quad \Theta \text{ cont}, X \text{ discrete})$$

Example 1: Beta prior & Posterior, Binomial likelihood

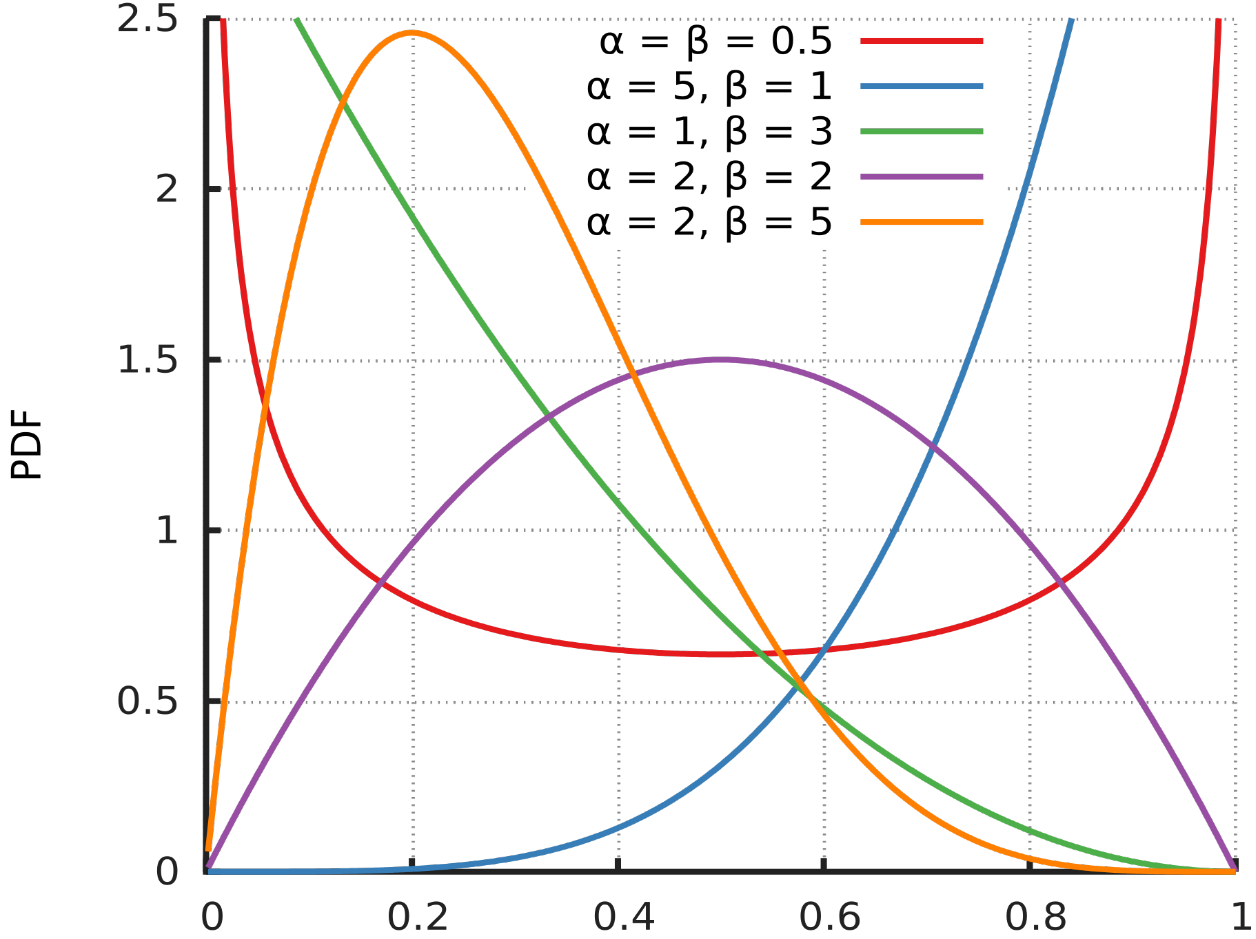
- Suppose I toss a biased coin with θ^* as the true probability of head which you want to estimate based on data $\mathcal{D}_n$ from n tosses.
- Let X denote the number of heads in $\mathcal{D}_n$.
- Suppose we assume a $Beta(\alpha, \beta)$ prior on θ^*,
- Then show that the posterior distribution $f_{\Theta|X}(\theta|k)$ has Beta distribution with parameters $\alpha' = \alpha + k$ and $\beta' = n - k + \beta$.

Beta distribution

- This is a continuous probability distribution on support $(0, 1)$ with two parameter (α, β).
- $\Theta \sim \text{Beta}(\alpha, \beta)$ implies

$$f_{\Theta}(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1.$$

- Here $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- $\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$. Note $\Gamma(n) = (n-1)!$
- `https://en.wikipedia.org/wiki/Beta_distribution`



α = β = 0.5
α = 5, β = 1
α = 1, β = 3
α = 2, β = 2
α = 2, β = 5
PDF
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

Example 1: Beta prior & Posterior, Binomial likelihood

- First note that the mean and variance for *Beta*(α, β) is given by $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- Also verify that when $\alpha = \beta = 1$, it corresponds top a uniform distribution.
- Now note that if we start with a uniform prior (or *Beta*$(1, 1)$), then the mean of the posterior distribution is given by $\frac{k+1}{n+2}$ and $\frac{(k+1)(n+1)}{(k+n+2)^2(k+n+2)}$.
- What happens as $n \to \infty$? The mean goes to θ^* almost surely using SLLN and the variance goes to zero.
- The posterior distribution therefore becomes a dirac-delta at θ^*.

Problem Setup: Beta Prior & Binomial Likelihood

- We observe n coin tosses with k heads. The goal is to find the posterior distribution of Θ, the probability of heads.

- Prior belief: $\Theta \sim \text{Beta}(\alpha, \beta)$,

$$f_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1.$$

- Likelihood of observing k heads given $\Theta = \theta$:

$$f_{X|\Theta}(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

- Bayes' Theorem:

$$f_{\Theta|X}(\theta|k) = \frac{f_{X|\Theta}(k|\theta) f_\Theta(\theta)}{f_X(k)}.$$

Substituting Likelihood and Prior

- Substitute the likelihood and prior into Bayes' formula:

$$f_{\Theta|X}(\theta|k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{f_X(k)}.$$

- Combine terms in the numerator:

$$f_{\Theta|X}(\theta|k) = \frac{\binom{n}{k}}{B(\alpha,\beta)} \cdot \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{f_X(k)}.$$

- Marginal likelihood ($f_X(k)$) ensures the posterior integrates to 1:

$$f_X(k) = \int_0^1 \binom{n}{k} \cdot \frac{1}{B(\alpha,\beta)} \cdot \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} d\theta.$$

Simplifying the Marginal Likelihood

- Factor out constants from the integral:

$$f_X(k) = \binom{n}{k} \cdot \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} d\theta.$$

- Recognize the integral as the Beta function:

$$\int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} d\theta = B(k+\alpha, n-k+\beta).$$

- Substitute back:

$$f_X(k) = \binom{n}{k} \cdot \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)}.$$

Deriving the Posterior

- Substitute the marginal likelihood $f_X(k)$ into the posterior formula:

$$f_{\Theta|X}(\theta|k) = \frac{\frac{\binom{n}{k}}{B(\alpha,\beta)} \cdot \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{\binom{n}{k} \cdot \frac{B(k+\alpha,n-k+\beta)}{B(\alpha,\beta)}}.$$

- Cancel $\binom{n}{k}$ and $\frac{1}{B(\alpha,\beta)}$:

$$f_{\Theta|X}(\theta|k) = \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}.$$

- Recognize this as the Beta distribution:

$$f_{\Theta|X}(\theta|k) \sim \text{Beta}(k+\alpha, n-k+\beta).$$

`https://mathlets.org/mathlets/beta-distribution/`

Example 2: Gaussain Pior, Likelihood & Posterior

- Suppose we observe realisation $x = (x_1, \dots, x_n)$ of $X = (X_1, \dots, X_n)$ where X_i are i.i.d with true mean θ^* and true variance σ^2. Suppose we know σ^2 but not θ^* and also know that X_i is Gaussian. How do we infer θ^*?
- Lets model $\theta*$ by a Gaussian random variable $\Theta \sim \mathcal{N}(\mu_0, \sigma^2)$.
- Since X_i are i.i.d, the likelihood are given by

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta)$$

- Now show that $f_{\Theta|X}(\theta|x)$ is Gaussian with mean $\frac{\sum_{i=1}^{n} x_i . + \mu_0}{n}$ and variance $\frac{\sigma^2}{n+1}$.
- What happens as $n \to \infty$?

Likelihood and Prior

- Likelihood of $X = (X_1, \ldots, X_n)$ given $\Theta = \theta$:

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i|\theta).$$

- Using the Gaussian form:

$$f_{X|\Theta}(x|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

- Prior on Θ:

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma^2}\right).$$

- Bayes' theorem for the posterior:

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_\Theta(\theta).$$

The Posterior Distribution

- After lots of simplification (HW) the posterior simplifies to:

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{(n+1)}{2\sigma^2}\left(\theta - \frac{\sum_{i=1}^n x_i + \mu_0}{n+1}\right)^2\right).$$

- This is a Gaussian distribution:

$$\Theta|X = x \sim \mathcal{N}\left(\frac{\sum_{i=1}^n x_i + \mu_0}{n+1}, \frac{\sigma^2}{n+1}\right).$$

Behavior as $n \to \infty$

- As $n \to \infty$:
  - Posterior mean: $\frac{\sum_{i=1}^{n} x_i + \mu_0}{n+1} \to \frac{1}{n}\sum_{i=1}^{n} x_i$, the sample mean.
  - Posterior variance: $\frac{\sigma^2}{n+1} \to 0$.
- Interpretation:
  - With more data ($n \to \infty$), the posterior concentrates around the sample mean.
  - The influence of prior μ_0 becomes negligible as n increases.

Conjugate Priors

- Clearly, there are occasions where the prior and posterior are of the same family of distributions.
- The prior and posterior are called conjugate distributions and the prior is called conjugate prior.
- This makes it very convenient as now you only need to keep track of the parameters of the distribution than the distribution itself.
- `https://en.wikipedia.org/wiki/Conjugate_prior`

Maximum aposteriori probability (MAP)

The MAP estimate $\hat{\theta}_{MAP}$ of θ^* given observation $X = x$ is the value of θ that maximizes $f_{\Theta|X}(\theta|x)$ (resp. $p_{\Theta|X}(\theta|x)$) when X is continuous (resp. discrete) random variable.

- From Bayes rule this is same as maximizing $f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)$ (ignoring the dinominator since it is independent of θ).
- How do you optimize this to obtain $\hat{\theta}_{MAP}$?
- $\hat{\theta}_{MAP} \in \left\{\theta : \frac{d}{d\theta}\left(f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)\right) = 0\right\}$
- Compare this with MLE

$$\hat{\theta}_{ML} = argmax_{\theta} f_{X|\Theta}(x|\theta)$$

MAP for Example 2

- Recall Example 2 where we saw that given Gaussian samples $(x_1, \ldots, x_n)$ but with unknown mean μ, we model the unknwon mean as a random variable Θ with a Gaussian prior.
- We then get a Gaussian posterior $f_{\Theta|X}(\theta|x)$ with mean $\frac{\sum_{i=1}^{n} x_i. + \mu_0}{n}$ and variance $\frac{\sigma^2}{n+1}$.
- What is $\hat{\theta}_{MAP}$?
- Gaussian is a unimodal function and hence $\hat{\theta}_{MAP} = \frac{\sum_{i=1}^{n} x_i. + \mu_0}{n}$
- Is it same as MLE? HW!

Conditional Expectation Estimator

- Yet another estimator for the unknown θ^* is the conditional expectation estimator given by

$$\theta_{CE} = E[\Theta | X = x] = \int_\theta \theta f_{\Theta|X}(\theta|x) d\theta$$

.

- Find θ_{CE} for all the previous examples.