$$X_1, \ldots, X_n \overset{iid}{\sim} p(x; \theta)$$

$$\hat{\theta} = g(X_1, \ldots, X_n)$$

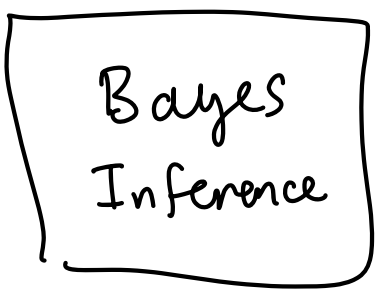
Bayesian statistics

- Point estimation: θ is fixed, but unknown constant

- Bayesian estimation: θ is assumed to be a random variable

Prior belief $p(\theta) \rightarrow$

Data $X_1, \ldots, X_n \rightarrow$



Bayes' Theorem

↳ $p(\theta | X_1, \ldots, X_n)$

Bayes theorem:

$$p(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)\, p(\theta)}{p(x_{1:n})}$$

? likelihood functio

prior distribution

posterior distribution

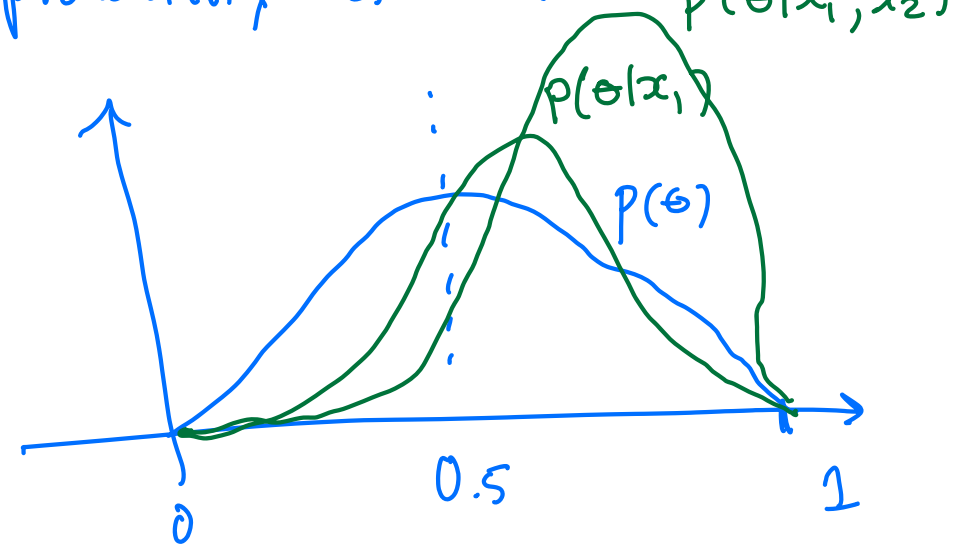
marginal likelihood

$$p(x_{1:n}) = \int p(x_{1:n} \mid \theta)\, p(\theta)\, d\theta$$

θ is the probability of heads

x_1 = outcome is heads

x_2 = outcome is heads



- How to obtain point estimators from a Bayesian perspective?
  - Maximum likelihood:

$$\hat{\theta} = \arg\max_{\theta} \; p(x_{1:n}; \theta) \qquad L(\theta; x_{1:n})$$

  - Maximum a posteriori (MAP):

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \; p(\theta \mid x_{1:n})$$

posterior mode

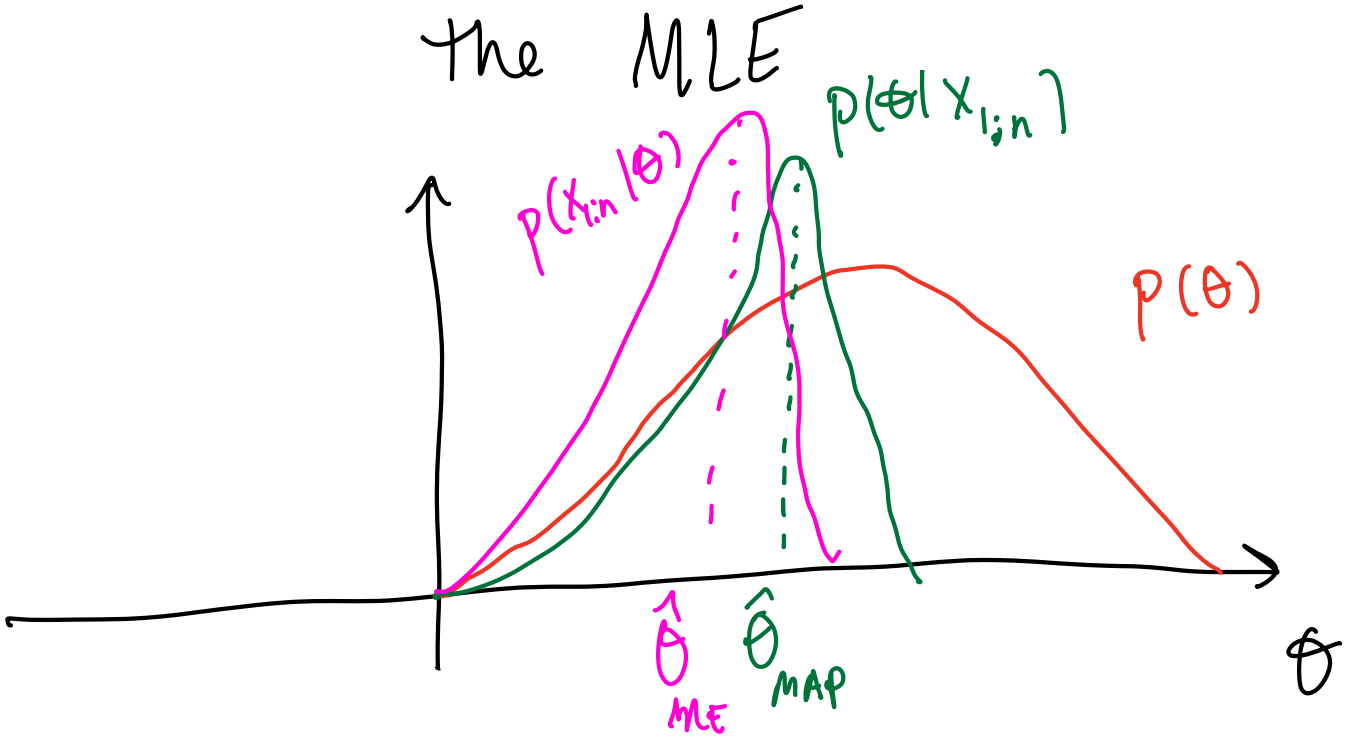
$$= \arg\max_{\theta} \; \frac{p(x_{1:n} \mid \theta)\, p(\theta)}{p(x_{1:n})}$$

$$= \arg\max_{\theta} \; p(x_{1:n} \mid \theta)\, p(\theta)$$

$$= \arg\max_{\theta} \log p(x_{1:n} | \theta) + \log p(\theta)$$

$$= \arg\max_{\theta} \left(\sum_{i=1}^{n} \log p(x_i | \theta) \right) + \log p(\theta)$$

- Insight: As $n \to \infty$, the MAP solution will coincide with the MLE



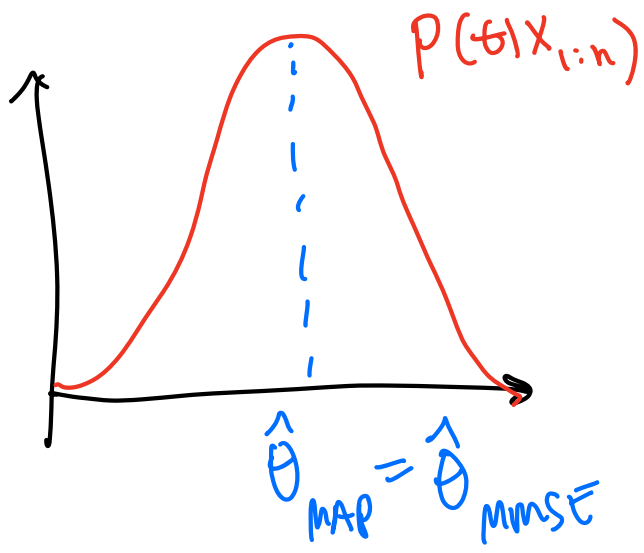
- $\hat{\theta}_{MAP} = g(X_1, \ldots, X_n)$ (It's still a point estimator)

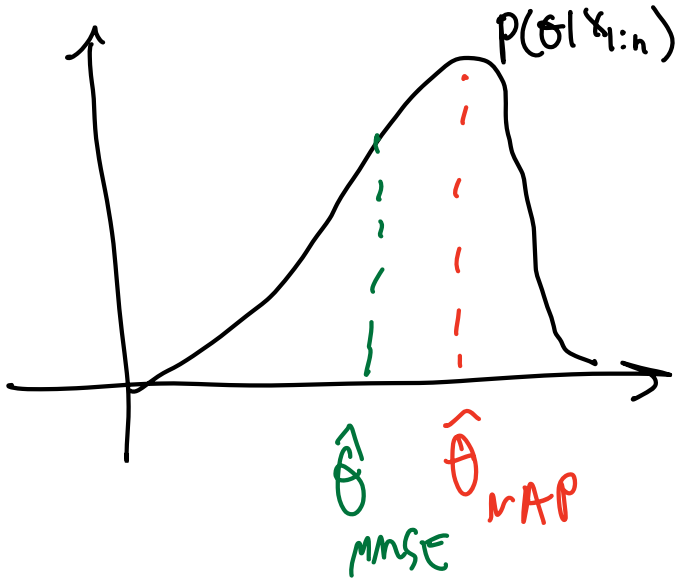
$$\mathbb{E}[\hat{\theta}] = \theta \quad \text{unbiased}$$
$$= \int \hat{\theta}\, p(x;\theta)\, dx$$

- Because the posterior uses a prior (will be or can be subjective), the estimator by design will be biased. Asymptotically, it will be unbiased

- Minimum Mean-Squared Error Estimator (MMSE)

$$\hat{\theta}_{MMSE} = \mathbb{E}[\theta \mid x_{1:n}]$$ (The posterior mean)
$$= \int \theta\, p(\theta \mid x_{1:n})\, d\theta$$





- If the posterior distribution is symmetric, then

$$\hat{\theta}_{MAP} = \hat{\theta}_{MMSE}$$

- If not, then in general,

$$\hat{\theta}_{MAP} \neq \hat{\theta}_{MMSE}$$

Why is it called MMSE? Because...

$$\hat{\theta}_{MMSE} = \arg\min_{\hat{\theta}} \mathbb{E}\left[(\theta - \hat{\theta})^2 \,\middle|\, X_1, \ldots, X_n\right]$$

Proof. Let $\mathcal{L}(\theta)$ denote the MSE loss

$$\mathcal{L}(\theta) = \mathbb{E}\left[(\theta - \hat{\theta})^2 \,\middle|\, X_{1:n}\right]$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\theta}} = \mathbb{E}\left[2(\theta - \hat{\theta}) \,\middle|\, X_{1:n}\right]$$

$$= 0 \Rightarrow \mathbb{E}[\theta \mid X_{1:n}] = \mathbb{E}[\hat{\theta} \mid X_{1:n}] = \hat{\theta}$$

$$\Rightarrow \hat{\theta}_{MMSE} = \mathbb{E}[\theta \mid X_{1:n}]$$

Example: Suppose we observe n coin flips $X_1, \ldots, X_n$ where

$$p(X_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

$$x_i \in \{0, 1\}$$

(0: tails, 1: heads)

and θ is unknown. Suppose a priori, θ follows a Beta distribution

$$\theta \sim \text{Beta}(\alpha_0, \beta_0)$$

$$p(\theta) = \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}$$

$$\theta \in [0, 1]$$

Beta function

a.) Find $p(\theta \mid x_{1:n})$.

Bayes theorem

likelihood, prior

$$p(\theta \mid x_{1:n}) = \frac{p(x_{1:n} \mid \theta)\, p(\theta)}{p(x_{1:n})}$$

$$= \frac{\left(\prod_{i=1}^{n} p(x_i|\theta)\right) p(\theta)}{p(x_{1:n})}$$

$$\propto \left(\prod_{i=1}^{n} p(x_i|\theta)\right) p(\theta)$$

$$= \left(\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\right)\left(\frac{1}{B(\alpha_0,\beta_0)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}\right)$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\,\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}$$

$$\Rightarrow p(\theta|x_{1:n}) \propto \theta^{\alpha_0+\sum_{i=1}^{n} x_i-1}(1-\theta)^{\beta_0+n-\sum_{i=1}^{n} x_i-1}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\alpha = \alpha_0 + \sum_{i=1}^{n} x_i$$

$$\beta = \beta_0 + n - \sum_{i=1}^{n} x_i$$

$$\theta \sim \text{Beta}(\alpha_0, \beta_0) \Rightarrow p(\theta) = \frac{\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}}{B(\alpha_0, \beta_0)}$$

$$\theta \mid x_{1:n} \sim \text{Beta}(\alpha, \beta)$$

$$\Rightarrow p(\theta \mid x_{1:n}) = \frac{\theta^{\alpha - 1}(1-\theta)^{\beta - 1}}{B(\alpha, \beta)}$$

$$\mathbb{E}[\theta] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

$$\mathbb{E}[\theta \mid x_{1:n}] = \frac{\alpha}{\alpha + \beta} = \frac{\alpha_0 + \sum_{i=1}^{n} x_i}{\alpha_0 + \beta_0 + n}$$

$$= \hat{\theta}_{MMSE}$$

[a lot of distributions will have property where

$\hat{\theta}_{MMSE}$ is a convex combination of MLE and prior mean]

$$E[\theta | X_{1:n}] = \frac{\alpha_0 + \sum_{i=1}^{n} X_i}{\alpha_0 + \beta_0 + n}$$

$$= \frac{\alpha_0}{\alpha_0 + \beta_0 + n} + \frac{\sum_{i=1}^{n} X_i}{\alpha_0 + \beta_0 + n}$$

$$= \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)\left(\frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + n}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{n}{\alpha_0 + \beta_0 + n}\right)$$

prior mean

MLE

$$p(\theta | x_{1:n}) \propto \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{c(\theta)}$$

$$\log c(\theta) = (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

$$\frac{\partial \log c}{\partial \theta} = \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\frac{\alpha-1}{\hat{\theta}_{MAP}} = \frac{\beta-1}{1-\hat{\theta}_{MAP}}$$

$$\frac{1}{\hat{\theta}_{MAP}} - 1 = \frac{\beta-1}{\alpha-1}$$

$$\frac{1}{\hat{\theta}_{MAP}} = \frac{\beta + \alpha - 2}{\alpha - 1}$$

$$\hat{\theta}_{MAP} = \frac{\alpha - 1}{\beta + \alpha - 2} \neq \hat{\theta}_{MMSE}$$

Example 2: Gaussian likelihood, Gaussian prior

Suppose we observe $X_1, \ldots, X_n$ and

$$p(X_i \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

MLE: $\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Suppose the μ is unknown, but σ^2 is known. If μ has the following prior

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

(a.) Find $p(\mu \mid x_{1:n})$

$$p(\mu \mid x_{1:n}) \propto p(x_{1:n} \mid \mu)\, p(\mu)$$

$$= p(\mu) \prod_{i=1}^{n} p(x_i \mid \mu)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma_0^2}}}_{\text{doesn't matter}} \exp\left(\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{doesn't matter}} \exp\left(\frac{-1}{2\sigma^2}(x_i-\mu)^2\right)$$

$$\propto \exp\left(\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \exp\left(\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \exp\left(\frac{-1}{2\sigma_0^2}\left(\mu^2-2\mu\mu_0+\mu_0^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i^2-2\mu x_i+\mu^2\right)\right)$$

$$= \exp\left(-\frac{1}{2\sigma_0^2}\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_0^2}\left(\mu^2 - 2\mu\mu_0\right) - \frac{1}{2\sigma^2}\left(-2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{\mu^2}{\sigma_0^2} - \frac{2\mu\mu_0}{\sigma_0^2} - \frac{2\mu \sum_{i=1}^{n} x_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2} \right) \right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_*^2}\left(\mu - \mu_*\right)^2 \right) \quad \text{we want this}$$

$1/\sigma_*^2$

$$= \exp\left(-\frac{1}{2}\left(\mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2} \right) \right) \right)$$

$$= \exp\left(\frac{-1}{2\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}} \left(\mu^2 - 2\mu \underbrace{\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2} \right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)}}_{\mu_*} \right) \right)$$

$$= N\left(\mu_*, \sigma_*^2 \right)$$

$$\mu_* = \sigma_*^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2} \right)$$

$$\sigma_*^2 = \underbrace{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}}$$

geometric mean .. or harmonic mean ? !! ☺

- If σ_0^2 is large ($\sigma_0^2 \to \infty$), the result only depends on the likelihood

$$p(\mu | x_{1:n}) \approx \mathcal{N}\left(\bar{X}_n, \frac{\sigma^2}{n} \right)$$

- If $\sigma_0^2 \to 0$,

$$p(\mu | x_{1:n}) = p(\mu)$$

$$\bar{X}_h = \frac{1}{n}\sum_{i=1}^{n} X_i$$ arithmetic mean

$$\left(\prod_{i=1}^{n} X_i\right)^{1/n}$$ ← geometric mean

$$\log(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

Prior Distributions

- Depending on the choice of the prior, we may not be able to analytically determine the posterior
- Conjugate priors always yield an analytical form for the posterior distribution

Def. (Conjugate Prior)

A prior $p(\theta)$ is said to be conjugate to a likelihood function $\ell(\theta|x_{1:n})$ if $p(\theta|x_{1:n})$ comes from the same family of distributions as the prior.

Examples:

prior	likelihood	post.
$\theta \sim \text{Beta}(\alpha_0, \beta_0)$	$X_i \sim \text{Bernoulli}(\theta)$	$\theta \mid X_{1:n} \sim \text{Beta}(\alpha, \beta)$
$\mu \sim N(\mu_0, \sigma_0^2)$	$X_i \sim N(\mu, \sigma^2)$	$\mu \mid X_{1:n} \sim N(\mu_*, \sigma_*^2)$

- Exponential Family Members (please see note below in *** on this)
  - We say a distribution belongs to

the exponential family if it can be expressed as:

$$p(x \mid \theta) = h(x) \exp\left(\eta(\theta)\, T(x) - A(\theta) \right)$$

sufficient statistic

$h(x)$: known function

$\eta(\theta)$: known function

$A(\theta)$: known function

We say it belongs to the canonical family if $\eta(\theta) = \theta$

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

$$= h(x) \exp\left(\theta \cdot T(x) - A(\theta) \right) p(\theta)$$

$$\propto \exp\left(\theta T(x) - A(\theta) \right) p(\theta)$$

$p(\theta)$: $\exp(\cdots)$

$$p(\theta) = g(\theta) \exp\left(\eta(\gamma)\, T(\theta) - A(\gamma) \right)$$

$$p(\theta|x) \propto p(x|\theta)\, p(\theta)$$

$$= \exp(\theta T(x) - A(\theta))\, g(\theta) \exp(\eta(\gamma)\, T(\theta))$$

$$= g(\theta) \exp(\theta T(x) - A(\theta) + \eta(\gamma) T(\theta))$$

$$= \underbrace{g(\theta)}\ \exp(\theta T(x)) \underbrace{\exp(\eta(\gamma) T(\theta) - A(\theta))}_{h(\theta)}$$

$$= g(\theta)\, h(\theta)\ \exp(\theta\, T(x))$$

$g(\theta)\, h(\theta)$: function of the data

$T(x)$: will follow up on this

Under some assumptions for exponential family, you will have a conjugate relationship.

Flat (Improper) Prior

Gaussian example

$$p(\mu \mid x_{1:n}) \propto \left(\prod_{i=1}^{n} p(x_i \mid \mu) \right) p(\mu)$$

$$p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$$

mean of μ ↓ , var. of μ ↓

μ: parameter of interest; μ_0, σ_0^2: hyperparameters

$$p(\mu) \propto C \quad \text{(flat prior)}$$

$$p(\mu \mid x_{1:n}) = \mathcal{N}\left(\mu \mid \frac{1}{n}\sum_{i=1}^{n} x_i, \frac{\sigma^2}{n}\right)$$

- An improper prior is a prior that does not integrate to 1.

→ flat prior means $p(\theta) \propto$ constant

$$\hat{\theta}_{MAP} = \hat{\theta}_{MLE} \quad \text{if} \quad p(\theta) \propto \text{constant}$$

Exponential family (Conjugate priors for Exp. Family) ***

If $p(x_i|\theta) = h(x_i) \exp\left(\theta\, T(x_i) - A(\theta) \right)$

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$\propto \exp\left(\theta \sum_{i=1}^{n} T(x_i) - n\, A(\theta) \right)$$

If we choose: $p(\theta) = h(\gamma, n_0) \exp\left(\gamma\, \theta - n_0 A(\theta) \right)$

Then, $p(\theta|x_{1:n}) \propto \exp\left(\theta \left(\gamma + \sum_{i=1}^{n} T(x_i) \right) - (n+n_0) A(\theta) \right)$

update via s.s.

Informative versus Noninformative Prior

- Informative prior - has heavy influence on the inference (n_0 is large)

example: Θ is the probability of heads

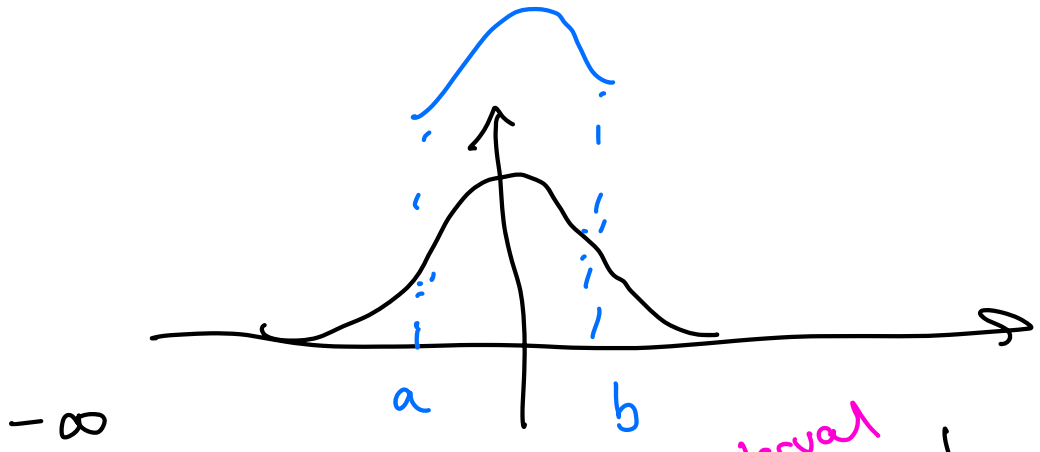
$$\Theta \sim N(0.5, 0.001)$$

Strong / informative

- Weakly informative prior: priors that include the constraint of the parameter but are not strong

example: Θ is the probability of heads

$$\Theta \sim N(0.5, 1{,}000{,}000)\, T(0,1)$$



Interval

$$TN(\mu, \sigma^2, I_{a,b}) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}{\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

$x \in [a, b]$

values x can take are constrained

↑ normalizes so that it integrates to 1

$$\theta \sim \text{Beta}(1, 1)$$ weakly informative

· Non-informative prior

MMSE ↙ $\mathbb{E}[\theta | X_{1:n}]$

$p(\theta | X_{1:n})$ ← posterior

↙ prediction (next time)

↘ MAP $\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta | X_{1:n})$

- We say a prior is non-informative if if it does not influence inference in different transformations of the parameters

$$\phi = h(\theta) \quad \text{given} \quad p(\theta)$$

$$p(\phi) = p(\theta) \left| h'(\theta) \right|^{-1}$$ h is invertible

$$p(\phi \mid x_{1:n}) \propto p(x_{1:n} \mid \phi)\, p(\phi)$$

$$= p(x_{1:n} \mid \phi)\, p\left(h^{-1}(\phi)\right) \left| h'\left(h^{-1}(\phi)\right) \right|^{-1}$$

Remark: Choices of the prior may appear to be non-informative in Θ space, but may be strongly informative in ϕ space

It is desirable to choose a prior that is "invariant" to one-to-one transformations

Jeffrey's prior: Let $I(\theta)$ denote the Fisher information

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log p(x_{1:n} \mid \theta)}{\partial \theta^2} \,\middle|\, \theta \right]$$

The Jeffrey's prior is given by

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

We will prove this next time.