Quiz :

$$p(x|\theta) = \frac{1}{\theta} \mathbb{1}(0 \le x \le \theta)$$

$$p(\theta) = \alpha \frac{\beta^{\alpha}}{\theta^{\alpha+1}} \mathbb{1}(\theta \ge \beta)$$

$= \text{Pareto}(\alpha, \beta)$ $\quad \mathbb{E}[\theta] = \frac{\alpha\beta}{\alpha - 1}$

Find the posterior

$$p(\theta | x_{1:n}) \propto p(x_{1:n} | \theta) \; p(\theta)$$

prior

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

$$\propto \alpha \left(\frac{\beta}{\theta}\right)^{\alpha} \mathbb{1}(\theta \ge \beta) \times \prod_{i=1}^{n} \left(\frac{1}{\theta}\right) \mathbb{1}(0 \le x_i \le \theta)$$

$$= \alpha \beta^{\alpha} \left(\frac{1}{\theta}\right)^{\alpha+1} \mathbb{1}(\theta \ge \beta) \left(\frac{1}{\theta}\right)^{n} \prod_{i=1}^{n} \mathbb{1}(0 \le x_i \le \theta)$$

does not depend on θ

$$\left(\frac{1}{\theta}\right)^{\alpha+n+1} \qquad \alpha + 1$$

new α parameter

$$\alpha \left(\frac{1}{\theta}\right)^{\alpha+n+1} \mathbb{1}(\theta \geq \beta) \prod_{i=1}^{n} \mathbb{1}(0 \leq x_i \leq \theta)$$

α'

$\frac{1}{\theta^{\alpha+1}}$ ~

$$\mathbb{1}(\theta \geq \beta, \theta \geq x_1, \theta \geq x_2, \ldots, \theta \geq x_n) \quad \mathbb{1}(\theta \geq \beta_*)$$

$\mathbb{1}(\theta \geq$ ~ $\beta)$

$\beta_* = \max(x_1, \ldots, x_n, \beta)$

$$p(\theta \mid x_{1:n}) = P_{\text{pareto}}(\alpha+n, \max(x_1, \ldots, x_n, \beta))$$

β_*

b.) MAP

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

$$= \beta_* = \max(x_1, \ldots, x_n, \beta)$$

c.) MMSE

$$\hat{\theta}_{MMSE} = \mathbb{E}[\theta | X_{1:n}] \Rightarrow$$ based on the parameters of the posterio.

$$= \frac{(\alpha+n)\ \max(X_1, \ldots, X_n, \beta)}{(\alpha+n) - 1}$$

$$p(\theta) \propto \mathbb{1}(\theta \geq \beta) \times \theta^{-(\alpha+1)}$$

$$= \text{Pareto}(\alpha, \beta)$$

$$\mathbb{E}[\theta] = \frac{\alpha\beta}{\alpha - 1}$$

$$p(\theta \mid X_{1:n}) = \left(\frac{1}{\theta}\right)^{n} \left(\frac{1}{\theta}\right)^{\alpha+1}$$

likelihood

$$= \left(\frac{1}{\theta}\right)^{n+\alpha+1} \mathbb{1}(\theta \geq \beta_*)$$

α

- Choice of Prior
- Calibration of Prior Distributions
- Theoretical Aspects
- Posterior - Predictive

Non-informative prior

Weakly informative prior

Informative prior

posterior $\propto$ likelihood

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

Requires that: $p(\theta) \propto$ constant

$$\int_{-\infty}^{\infty} p(\theta)\, d\theta \neq 1 \quad > \infty$$

For example: $X_i \sim N(\mu, \sigma^2)$

- μ is unknown
- σ^2 is known

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

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

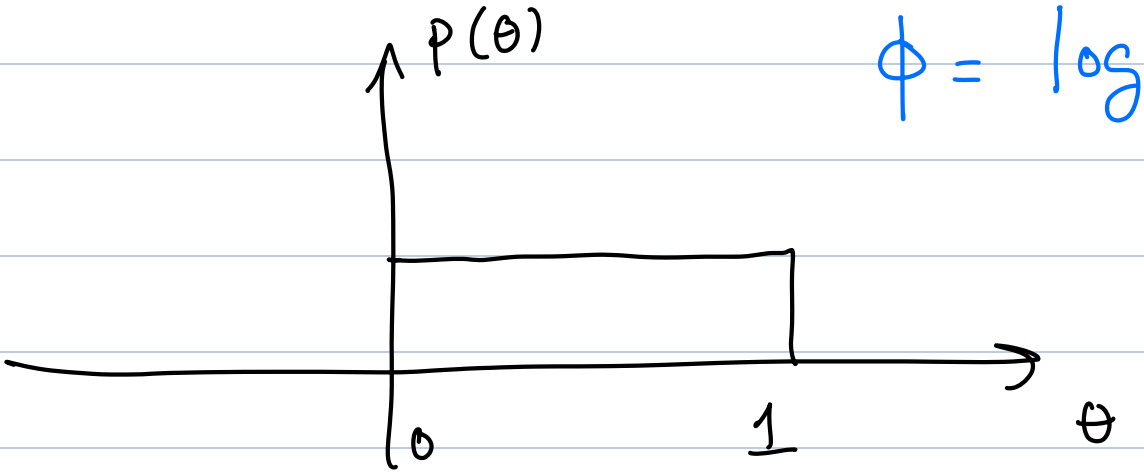
$$= N\left(\mu \mid \bar{X}_n, \frac{\sigma^2}{n}\right)$$

Improper prior can actually lead to a proper posterior distribution

$X_i \sim \text{Bernoulli}(\theta)$

θ: probability of success

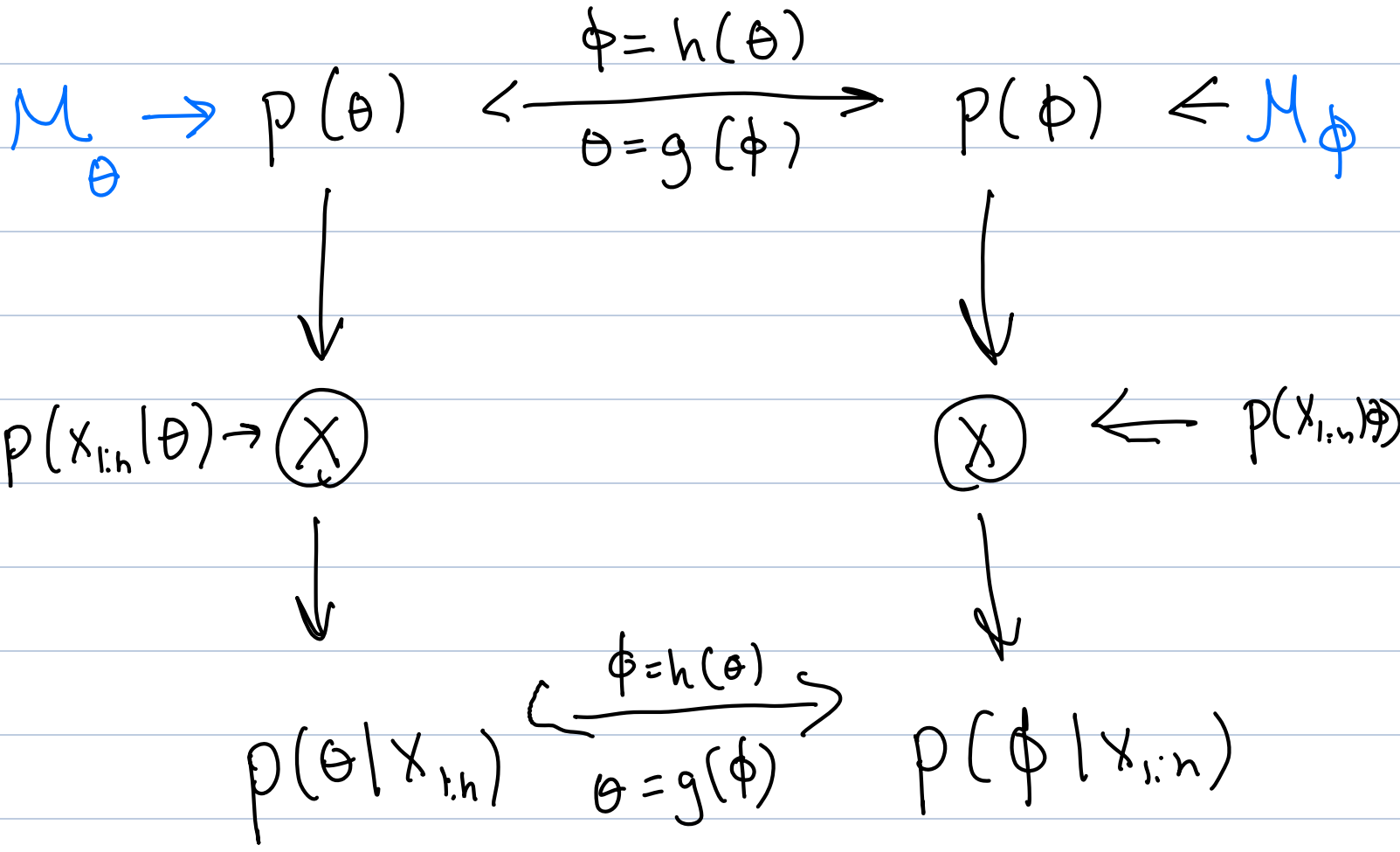
$\theta \sim \mathcal{U}(0, 1)$



$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

- The problem of using a flat prior is that although it is "flat" in one parameterization, it may be very informative in another

Jeffreys' Principle



Jeffreys' Prior: Let θ be a random variable that parameterizes another random variable $X_1, \ldots, X_n$, where $X_1, \ldots, X_n$ are iid. The Jeffreys' prior is given by:

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

$$I(\theta) = -\mathbb{E}\left[\left(\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right)\middle|\theta\right]$$

Claim: The Jeffreys' prior satisfies Jeffreys' invariance principle

Change of variables formulas:

$$p(\phi) = p(\theta)\left|\frac{dg}{d\phi}\right| \quad \text{where} \quad g(\phi) = h^{-1}(\phi)$$

$$p(\phi) = p(\theta)\left|\frac{dh}{d\theta}\right|^{-1}$$

$$p_*(\theta) = \sqrt{I(\theta)}$$

$$p_*(\phi) = \sqrt{I(\phi)}$$

Inference in θ-space: $p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\, p_*(\theta)$

Inference in ϕ-space: $p(\phi \mid x_{1:n}) \propto p(x_{1:n} \mid \phi)\, p_*(\phi)$

Start in θ-space:

Jeffreys prior

$$p(\theta \mid x_{1:n}) \propto p(x_{1:n} \mid \theta)\sqrt{I(\theta)}$$

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

Let's apply the change-of-variables formula:

Posterior of ϕ

$$p(\phi \mid x_{1:n}) \propto p\left(\theta = g(\phi) \mid X_{1:n}\right) \underbrace{\sqrt{I(g(\phi))}} \left|\frac{dh}{d\theta}\right|$$

What is this?

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

Substitute the transformation $\phi = h(\theta)$

$$= -\mathbb{E}\left[\left(\frac{\partial^2 \log p(x \mid \phi = h(\theta))}{\partial \phi^2}\right) \left|\frac{dh}{d\theta}\right|^2 \middle| \theta\right]$$

$$= I(\phi) \left|\frac{dh}{d\theta}\right|^2$$

$$p_\star(\theta) \propto \sqrt{I(\theta)} = \sqrt{I(\phi)} \left|\frac{dh}{d\theta}\right|$$

- plugging $I(\theta)$ back to our change of variables, we have

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

which is the desired result

Calibrate a Prior using Moment Matching

- The idea of moment matching is to calibrate a prior distributions theoretical moments to the moments provided by the domain expert
- Solving a system of equations
- Prior has its own parameters and we want to "estimate" them:

$$p(\theta ; \alpha)$$

↑ Bayesian model parameter

↑ hyperparameters of the prior

- Moment matching (for only two moments):

$$\mathbb{E}[\theta; \alpha] \approx \mu_{expert}$$

$$\mathbb{V}[\theta; \alpha] \approx \sigma^2_{expert}$$

- Let's do an example:

Suppose we have a Bayesian model with an unknown probability parameter θ (where $0 \leq \theta \leq 1$). We assume θ a priori follows a Beta distribution

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

$$\iff p(\theta) \propto \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}$$

Domain expert tells you that they believe with 95% confidence that

$$\theta \in [0.2, 0.6]$$

$$0.2 = \mu_{expert} - 2\sigma_{expert} \approx F(z \le 0.025)$$

$$0.6 = \mu_{expert} + 2\sigma_{expert} \approx F(z \le 0.975)$$

$z \sim N(\mu_e, \sigma_e^2)$

$$\mu_{expert} = 0.4$$

$$\sigma^2_{expert} = 0.01$$

For a Beta distribution, the mean and variance are given by:

$$0.4 = \mathbb{E}[\theta; \alpha_0, \beta_0] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

$$0.01 = \mathbb{V}[\theta; \alpha_0, \beta_0] = \frac{\alpha_0 \beta_0}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)}$$

↑ set equal to domain expert moments

We then solve this system of equations for α_0 and β_0 to obtain the hyperparameters:

$$\alpha_0 = \left(\frac{\mu_{expert}(1-\mu_{expert})}{\sigma^2_{expert}} - 1 \right) \mu_{expert}$$

$$\beta_0 = \left(\downarrow \right) (1-\mu_{expert})$$

Theoretical Aspects

Point estimation:

assume there is some ground truth θ_0

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

function of the random sample

$$\hat{\theta} \xrightarrow{a.s.} \theta_0$$

$$\xrightarrow{p} \theta_0$$

$$\xrightarrow{d} \theta_0$$

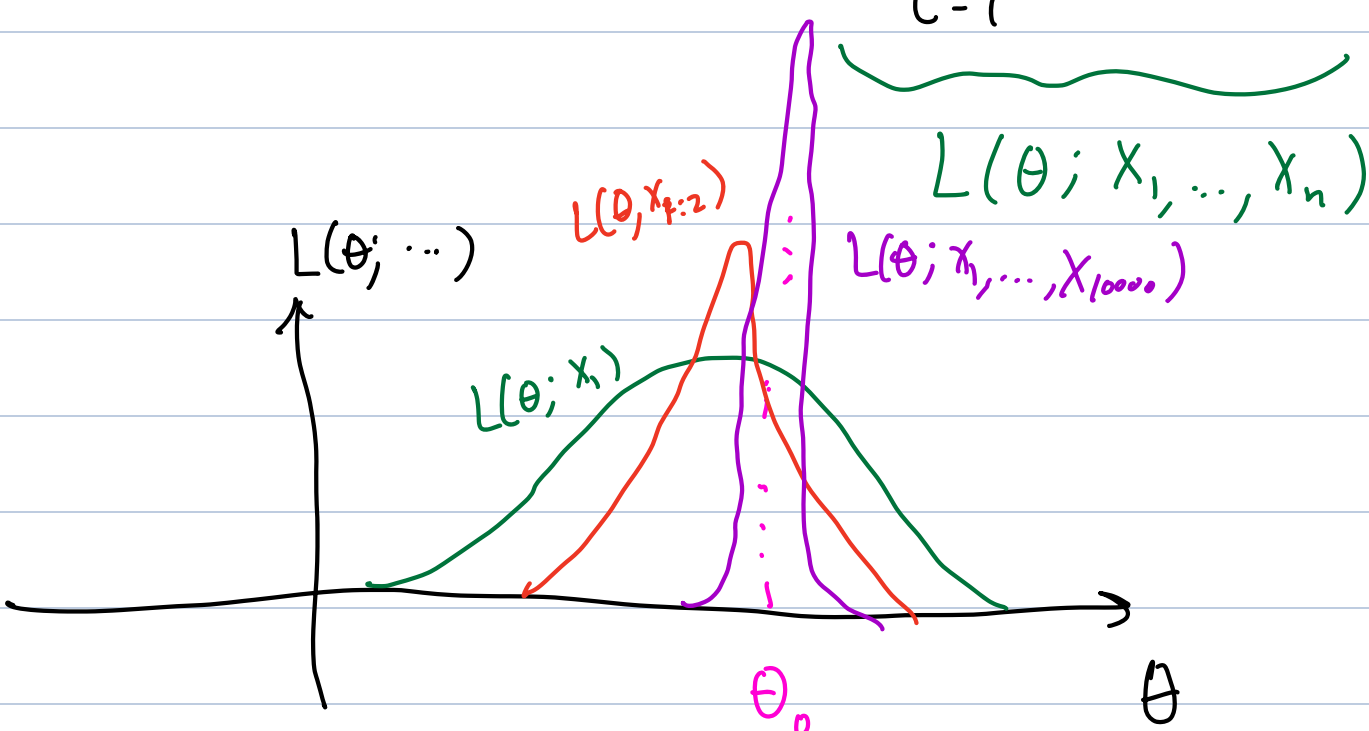
Limiting distribution of MLE

$$\hat{\theta}_{MLE} \overset{\cdot}{\sim} \mathcal{N}\left(\theta_0, I^{-1}(\theta)\right)$$

↑ approximately distributed according to...

Since in Bayesian estimation, the posterior is the product of the prior and the likelihood:

$$p(\theta | X_{1:n}) \propto p(\theta) \prod_{i=1}^{n} p(X_i | \theta)$$



Intuition: As $n \to \infty$, the likelihood dominates and approaches a $\delta(\theta - \theta_0)$

delta function centered at θ_0

Question: What happens to $p(\theta | X_{1:n})$ for large values of n?

- Start with a Taylor expansion of $\log p(\theta | X_{1:n})$ at $\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta | X_{1:n})$:

$$\log p(\theta|X_{1:n}) = \underbrace{\log p(\hat{\theta}_{MAP}|X_{1:n})}_{①} + \underbrace{(\theta - \hat{\theta}_{MAP})}_{②}\left[\frac{\partial \log p(\theta|X_{1:n})}{\partial \theta}\right]_{\theta=\hat{\theta}_{MAP}} + \frac{1}{2}(\theta - \hat{\theta}_{MAP})^2\left[\frac{\partial^2 \log p(\theta|X_{1:n})}{\partial \theta^2}\right]_{\theta=\hat{\theta}_{MAP}} + \underbrace{\text{higher order terms}}_{③}$$

$$\text{Higher-order terms} \propto O\left(|\theta - \hat{\theta}_{MAP}|^3\right)$$

Observations:

- The term ① is constant → does not effect the resulting posterior
- The term ② is 0. because $\hat{\theta}_{MAP}$ is the maximizer of $\log p(\theta | X_{1:n})$

- The higher-order terms (2.) will diminish in comparison to the second-order term as $n \to \infty$ because $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE} \xrightarrow{p} \theta_0$.

Conclusion: The second-order term becomes the dominant term in the Taylor series expansion and so by analysis

$$\theta \dot{\sim} N\left(\hat{\theta}_{MAP}, \frac{1}{I(\theta) + J(\theta)}\right)$$

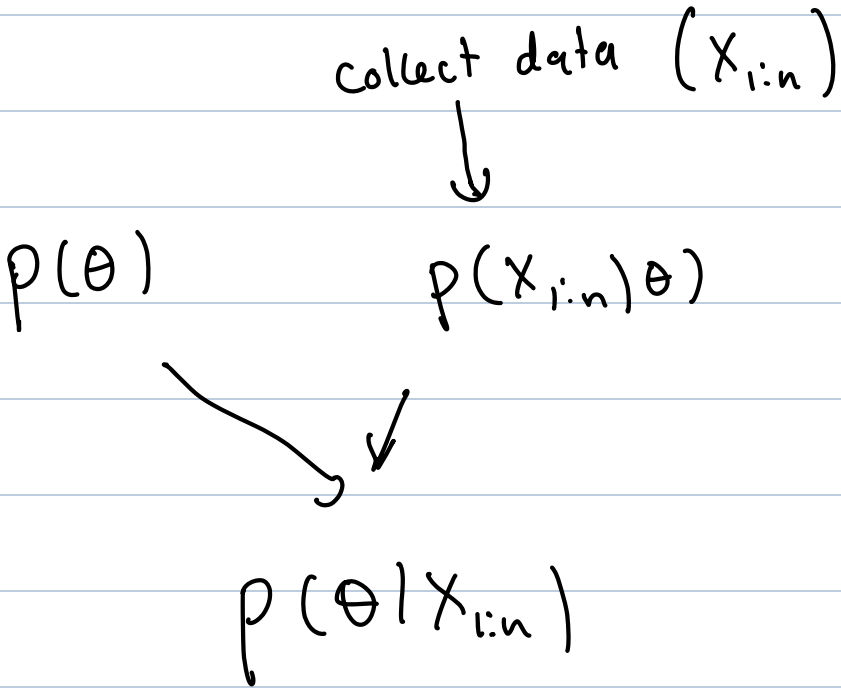
$I(\theta)$: Fisher information $-\mathbb{E}\left[\left(\frac{\partial^2 \log p(x_{1:n}|\theta)}{\partial \theta^2}\right)\right]$.

$J(\theta)$: Fisher information of the prior $-\mathbb{E}\left[\frac{\partial^2 \log p(\theta)}{\partial \theta^2}\right]$

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

Theorem: Bernstein von-Mises theorem

Posterior- Predictive Checks



- We want to know if our Bayesian is well-calibrated to the observed data after fitting the model

Predictive distribution

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

apply chain rule of probability

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

$\tilde{X} \perp\!\!\!\perp X_{1:n}$ when θ is given

$$p(\tilde{x} \mid \theta, x_{1:n}) \equiv p(\tilde{x} \mid \theta)$$

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

$$= \mathbb{E}\left[p(\tilde{x} \mid \theta) \,\middle|\, x_{1:n} \right]$$

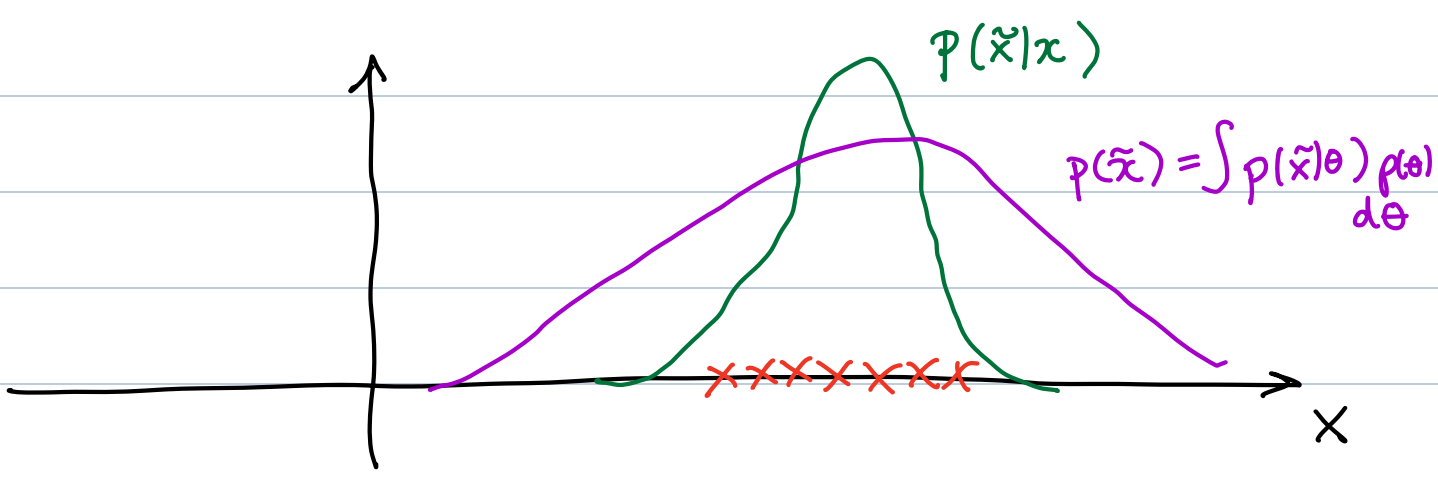
conditional expectation

S: # of samples

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta \mid x_{1:n})$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} p(\tilde{x} \mid \theta^{(s)})$$

Posterior-predictive check $\Rightarrow$ test to see if my historical data is contained in the support of the predictive distribution


$p(\tilde{x}|x)$
$p(\tilde{x}) = \int p(\tilde{x}|\theta)\, p(\theta|\mathcal{D})\, d\theta$
X