Topics: Model Calibration
- Prior-predictive checks
- Posterior-predictive checks

Predictive Accuracy
- log posterior predictive density (lppd)
- Cross validation

Approximate inference
- What to do when we can't compute posteriors?
- Parametric approximations
- Monte Carlo methods

Assignments: Quiz 4 (credit for attendance)
- use HW 4 to help complete

Model calibration: To check if the model has fit well to the observed data

- we introduced the posterior-predictive distribution

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

$\tilde{x}$: new samples; $x_{1:n}$: observed historical data

(chain rule)

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

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

$\tilde{x} \perp\!\!\!\perp x_{1:n}$ given $\theta$

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

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

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

observe $x_{1:n}$

$\hat{\theta} = g(x_1, \ldots, x_n)$

↓ plug in

only one realization of the parameter

$$x \sim p(x \mid \hat{\theta})$$

Bayesian estimation

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

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

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

more than 1 realization of the parameter

1. Prior-predictive check
2. Posterior-predictive check

→ used as means to verify that our Bayesian model can simulate the data

Prior-predictive: $p(\tilde{x}) = \int p(\tilde{x}, \theta)\, d\theta$

$= \int p(\tilde{x} \mid \theta)\, p(\theta)\, d\theta$

$= \mathbb{E}\left[ p(\tilde{x} \mid \theta) \right]$

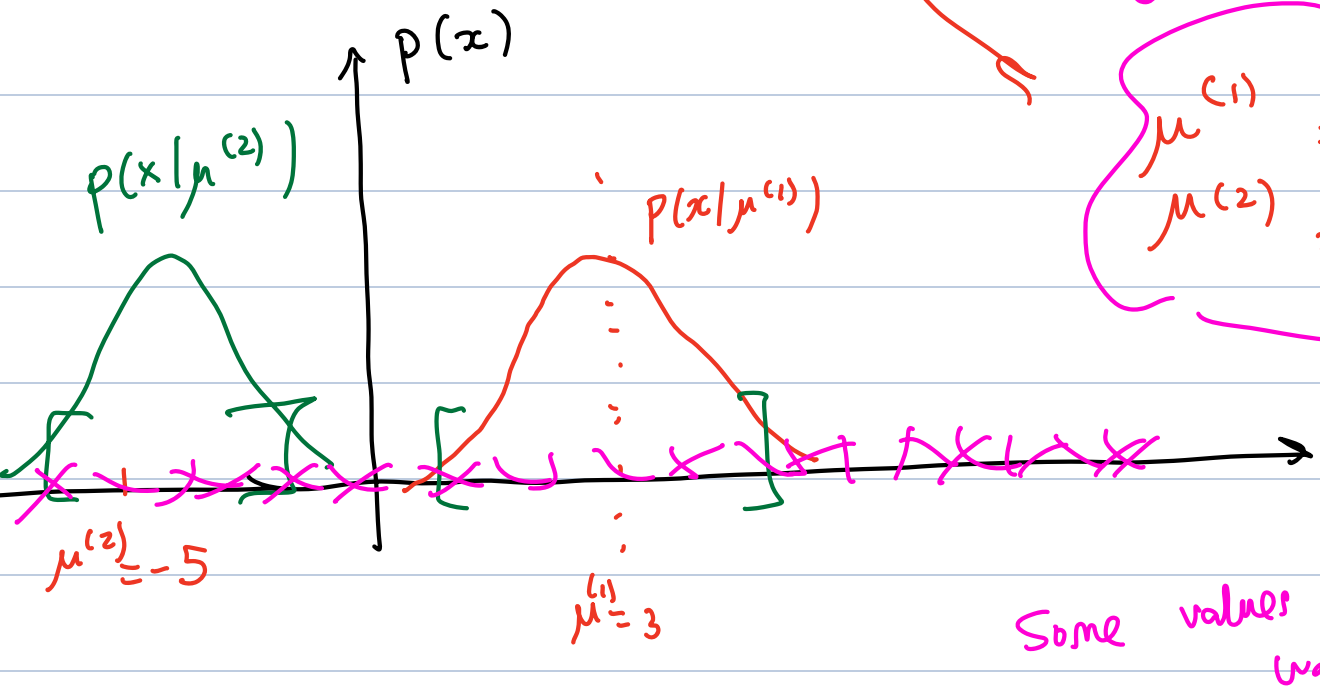
This expectation is taken w.r.t. the prior dist. of $\theta$

$x \sim N(\mu, 1)$

$\mu \sim N(5, 100)$ ($sd = 10$)

$\mu^{(1)} = 3$

$\mu^{(2)} = -5$



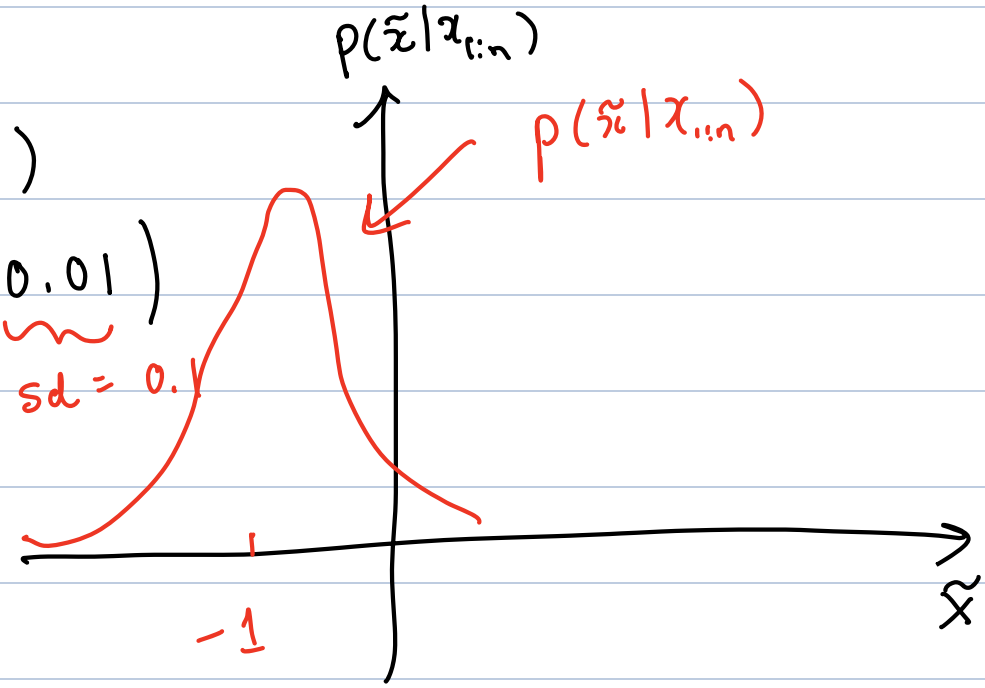

Some values of $\mu$ we might be

sampled with high probability

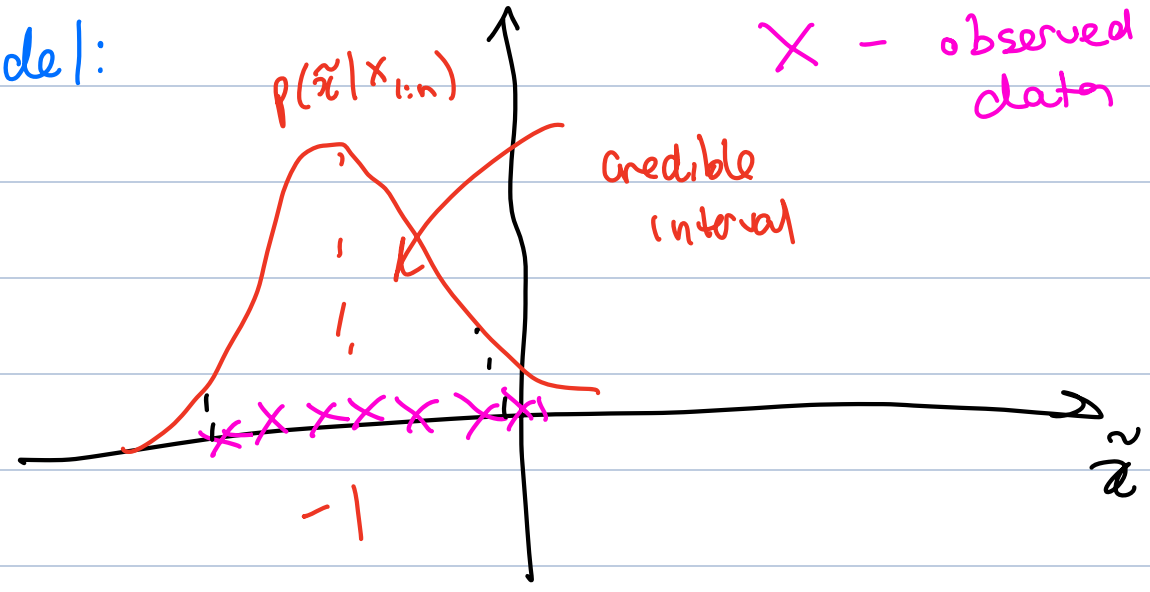
## Posterior-predictive check

- A way for us to measure how well a model is calibrated after we have done inference

$$x \sim N(\mu, 1)$$
$$\mu | x_{1:n} \sim N(-1, 0.01)$$




Good model:




X - observed data

Bad model:


## Credible interval:

↳ a metric/estimator Bayesian statisticians use to assert the probability of something happening in a Bayesian model

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

Where does the 95% of the probability mass of the posterior lie?

$$I = [a - \mu, a + \mu] \quad \text{s.t.}$$

$$\mathbb{P}(\theta \in I \mid x_{1:n}) = 0.95$$

95% CI

→ Ways to get this:

→ Approximate posterior with a Gaussian and use the Z-scores

→ Simulate samples and empirically evaluate the credible intervals

## Predictive Accuracy

- Before, we were using the posterior-predictive distribution $p(\tilde{x} \mid x_{1:n})$ to check the calibration of our model
- We may also want to see how well our model generalizes to unseen observations

Metric: log posterior predictive density (lppd)

↳ Observe new data $\tilde{x}_*$

↳ Evaluate the predictive distribution

one $\tilde{x}_*$

$$\text{lppd} = \log p(\tilde{x}_* \mid x_{1:n})$$

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

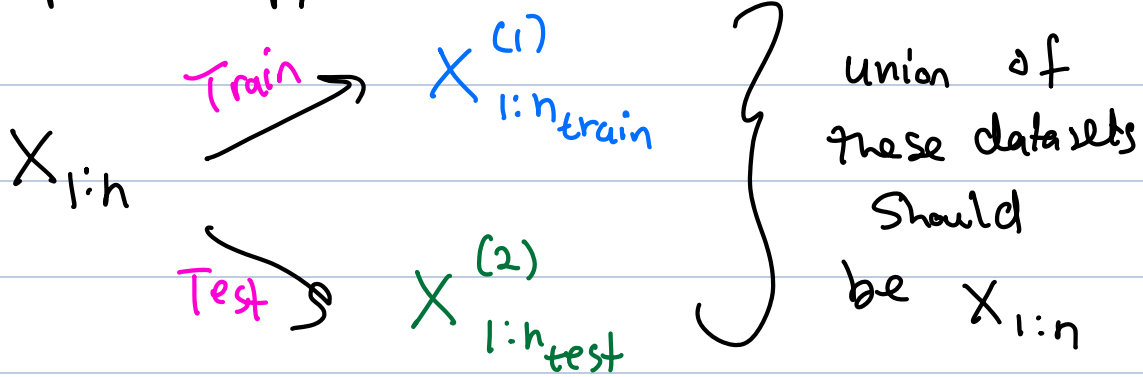
$\theta^{(s)} \sim p(\theta \mid x_{1:n})$

→ the higher this metric, the better the performance of the model

→ evaluate this always on out-of-sample data

Question: In practice, we don't want to assess this metric after observing new data. We want to know it after we train the model!

1. Simplest approach



↳ train model on $X^{(1)}_{1:n_{train}}$

↳ evaluate lppd on $X^{(2)}_{1:n_{test}}$

Problems with this approach:

→ depends on the size of the split

→ if $n_{train}$ is small, then we don't have enough data to train the model

→ if $n_{test}$ is small, the lppd estimator will have high variance

→ depends on the way we split the data

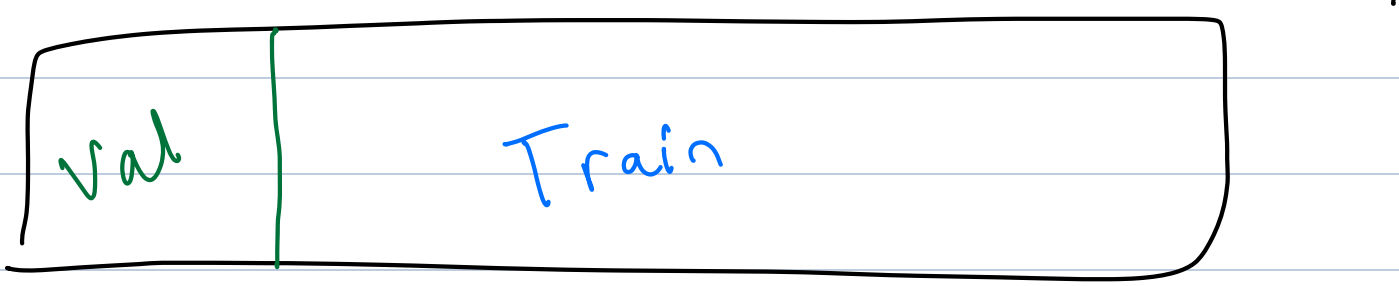
In practice, we evaluate the average performance

Average lppd: $\mathbb{E}_{p(x)}\left[\text{lppd}\right]$ ≈ $\frac{1}{s}$

$\{x_1^{(2)}, \ldots, x_{n_{test}}^{(2)}\}$ small number of samples

$$\text{Average lppd} \simeq \frac{1}{n_{test}} \sum_{n=1}^{n_{test}} \text{lppd}(x_2^{(n)})$$

## Cross-validation




$$p(\theta \mid x_{train}^{(1)})$$

Evaluate average $\text{lppd}^{(1)} = \frac{1}{n_{test}^{(1)}} \sum_{n=1}^{n_{test}^{(1)}} \text{lppd}(x_{val,n}^{(1)})$

↓ splitting and evaluation is repeated for K folds

$$\text{Average lppd} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_{test}^{(k)}} \sum_{n=1}^{n_{test}^{(k)}} \text{lppd}(x_{val,n}^{(k)})$$

↳ averaging the result over many possible train/validation splits

→ When $n_{test}^{(k)} = 1$ for all $k$, this is called Leave-one-out cross validation (LOOCV)

## Approximate (Bayesian) Inference

- We have seen that Bayesian methods are capable of:
  1. Incorporating domain knowledge into the inference procedure via the prior $p(\theta)$
  2. Capture uncertainty about the parameter and about predictions using Bayesian inference

We have only dealt with models where the posterior is analytically tractable:

$$p(\theta | x_{1:n}) \propto p(x_{1:n} | \theta)\, p(\theta) \propto \tilde{\pi}(\theta)$$

The $\tilde{\pi}(\theta)$ usually has had the form of a dist. we already know. For more general choices of $g(\theta)$, we cannot analytically determine the posterior.

↳ Restricted to only using certain priors if we want to obtain an analytical form to the posterior

Approximate Bayesian inference aims to obtain an approximate posterior (with nice properties) via the following two approaches:

1. Parametric approximations
   → choose a distribution that we know and estimate its parameters so that it is close enough to the true posterior
   → Laplace approximation
   → Variational inference

# Laplace Approximation

→ Assume the posterior is a Gaussian

$$p(\theta \mid x_{1:n}) \approx N(\theta \mid \hat{\mu}_\theta, \hat{\sigma}^2_\theta)$$

→ Want to choose $\hat{\mu}_\theta$ and $\hat{\sigma}^2_\theta$ such that as $n \to \infty$, the approximation is perfect

→ Bernstein von-Mises theorem

$$\theta \mid X_{1:n} \xrightarrow{d} N\left(\underbrace{\hat{\theta}_{MAP}}_{\downarrow n\to\infty \;\; \hat{\theta}_{MLE}}, \underbrace{I^{-1}_{Bayes}(\hat{\theta}_{MAP})}_{\downarrow n\to\infty \;\; 0}\right)$$

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

Bayesian information

$$I_{Bayes}(\hat{\theta}_{MAP}) = -\mathbb{E}\left[\frac{\partial^2 \log p(\theta, X_{1:n})}{\partial \theta^2} \,\middle|\, \theta = \hat{\theta}_{MAP}\right]$$

$$I(\hat{\theta}_{MAP}) + J(\hat{\theta}_{MAP})$$

Fisher information

information of the prior

→ In practice, we evaluate the MAP and then approximate the Bayesian information to form the Laplace approximation

→ Since posteriors are asymptotically normal, we expect this to be a good approximation, when $n$ is large.

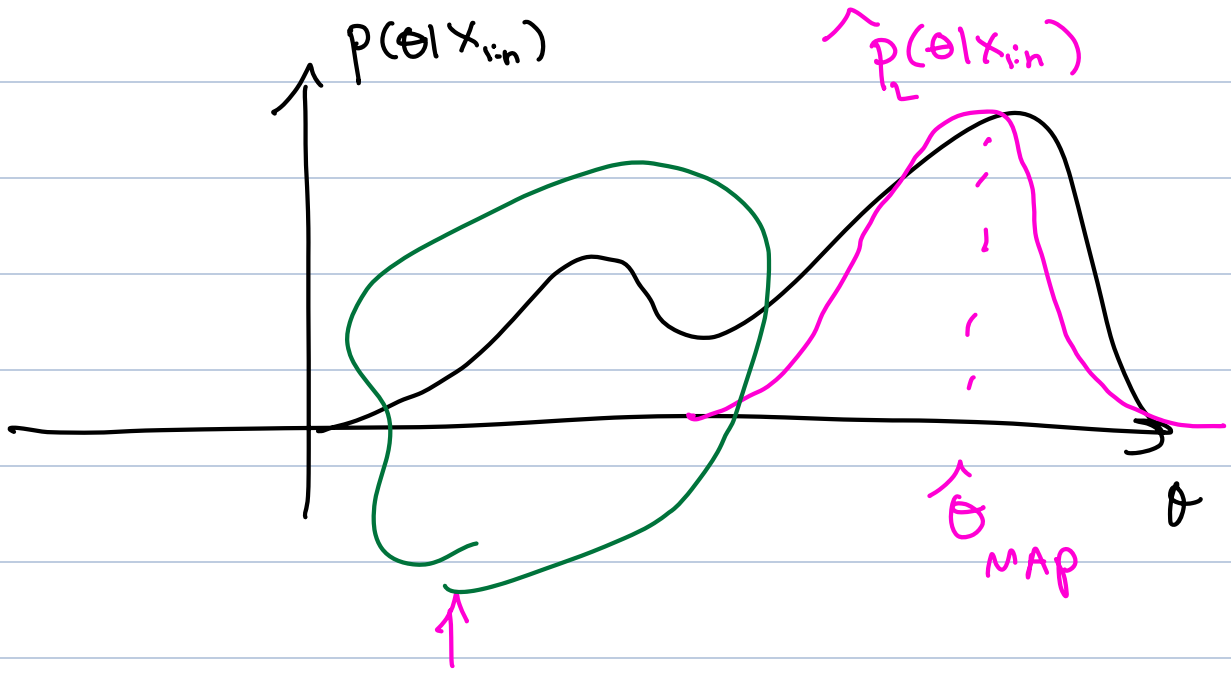
## Limitations

1. No convergence guarantees for derived estimators:

$$\int \theta \, p(\theta \mid X_{1:n}) \, d\theta \neq \mathbb{E}[\theta \mid X_{1:n}]$$

replaced with Laplace approximation

2. Approximation is bad when $n$ is small and the likelihood is non-Gaussian




→ We seek a more theoretically rigorous approach to obtaining the posterior distribution

→ Monte Carlo methods!

## (2.) Monte Carlo (MC) methods

- Let $\pi(\theta) \triangleq p(\theta|X_{1:n})$ denote the distribution of interest.

- MC methods aim to approximate expectations taken w.r.t. to the target density $\pi(\theta)$

- Let's say we are interested in computing:

$$\mathbb{E}_{\theta \sim \pi(\theta)}\left[h(\theta)\right] = \int h(\theta)\,\pi(\theta)\,d\theta$$

- Standard Monte Carlo:

  - Simulate $M$ samples from $\pi(\theta)$:

$$\theta^{(m)} \overset{\text{i.i.d.}}{\sim} \pi(\theta)$$

  - Approximate the integral using a sample average:

$$\mathbb{E}_{\theta \sim \pi(\theta)}\left[h(\theta)\right] \approx \underbrace{\frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})}_{\hat{I}_{MC}(h)}$$

$$\hat{I}_{MC}(h) = \frac{1}{M}\sum_{m=1}^{M} h(\overbrace{\theta^{(m)}}^{\text{random sample}})$$

random sample

$$\xrightarrow{a.s.} \mathbb{E}_{\theta \sim \pi(\theta)}\left[h(\theta)\right]$$

if $h(\theta)$ is integrable w.r.t. $\pi(\theta)$ :

$$\int |h(\theta)|\,\pi(\theta)\,d\theta < \infty$$

So, if we can draw samples from $\pi(\theta)$, we can approximate quantity related to $\pi(\theta)$

PROBLEM : We cannot draw samples from $\pi(\theta)$

We utilize more clever sampling techniques to approximately generate samples from $\pi(\theta)$

1. Rejection sampling
2. Importance sampling
3. MCMC

# Rejection Sampling:

- **Setup:** Target distribution: $\pi(\theta)$
  Proposal distribution: $q(\theta)$
  $M \geq \sup_{\theta \in \Theta} \frac{\pi(\theta)}{q(\theta)}$

Supremum

- **Algorithm:**
  1. Sample $\theta_* \sim q(\theta)$
  2. Sample $u \sim \mathcal{U}(0,1)$
  3. Check if $u < \frac{\pi(\theta)}{M q(\theta)}$:

Accept the sample
$\theta = \theta_*$

Otherwise:

reject the sample

**Conditions:** $M \geq \sup_{\theta \in \Theta} \frac{\pi(\theta)}{q(\theta)}$

The support of $\pi(\theta)$ should be a subset of the support of $q(\theta)$

$q(\theta) > 0$ whenever $\pi(\theta) > 0$

Prove that the samples that are accepted are samples from $\pi(\theta)$:

not a RV

$$F_\theta(\tilde{\theta}) = P(\theta \le \tilde{\theta})$$

$$= P(\theta_* \le \tilde{\theta} \mid \text{accepted } \theta_*)$$

$\theta_*$ = accepted $\theta_*$

$$U \le \frac{\pi(\theta_*)}{Mq(\theta_*)}$$

$$\overset{①}{=} \frac{P(\theta_* \le \tilde{\theta},\ U \le \frac{\pi(\theta_*)}{Mq(\theta_*)})}{\overset{②}{P(U \le \frac{\pi(\theta_*)}{Mq(\theta_*)})}}$$

Our goal is to show that this is the CDF of the target distribution!

① $P(\theta_* \le \tilde{\theta},\ U \le \frac{\pi(\theta_*)}{Mq(\theta_*)})$

$$= \int P(\theta_* \le \tilde{\theta},\ U \le \frac{\pi(\theta_*)}{Mq(\theta_*)} \mid \theta_* = \gamma)$$

$$q(\gamma)\, d\gamma$$

$$= \int I(\gamma \le \tilde{\theta}) \underbrace{P\left(U \le \frac{\pi(\gamma)}{M q(\gamma)}\right)} q(\gamma)\, d\gamma$$

$$= \int_{-\infty}^{\tilde{\theta}} \underbrace{P\left(U \le \frac{\pi(\gamma)}{M q(\gamma)}\right)} q(\gamma)\, d\gamma$$

$$= \int_{-\infty}^{\tilde{\theta}} \frac{\pi(\gamma)}{M q(\gamma)}\, q(\gamma)\, d\gamma$$

$$= \frac{1}{M} \underbrace{\int_{-\infty}^{\tilde{\theta}} \pi(\gamma)\, d\gamma}$$

$$= \frac{1}{M} P_{\pi}(\theta \le \tilde{\theta})$$

This is the CDF of $\pi(\theta)$

(2.) $$\mathbb{P}\left(U \le \frac{\pi(\theta_*)}{M q(\theta_*)}\right)$$

$$= \int \mathbb{P}\left(U \le \frac{\pi(\theta_*)}{M q(\theta_*)} \,\middle|\, \theta_* = \gamma\right) q(\gamma)\, d\gamma$$

$$= \int \frac{\pi(\gamma)}{M q(\gamma)} q(\gamma)\, d\gamma$$

$$= \frac{1}{M} \underbrace{\int \pi(\gamma)\, d\gamma}_{=1} = \frac{1}{M}$$

Plugging (1.) and (2.) together, we find that

$$\underbrace{F(\theta \le \tilde{\theta})}_{\text{distribution of our generated samples}} = F_{\pi}(\theta \le \tilde{\theta})$$

The idea is to generate $M$ samples from rejection sampling and then apply MC to obtain estimators of expected values (taken w.r.t. $\pi(\theta)$)

## Limitations

1. The probability of accepting a sample:

$$\mathbb{P}(\text{accept}) = \mathbb{P}\left(U \leq \frac{\pi(\theta_*)}{M\, q(\theta_*)}\right)$$

$$= \frac{1}{M}$$

↳ we must choose $M \geq \sup_{\theta \in \Theta} \left(\frac{\pi(\theta)}{M q(\theta)}\right)$

↳ if we choose $M$ too large, we will reject a lot of samples and rejection sampling becomes inefficient

$$\pi(\theta) \triangleq p(\theta | X_{1:n}) \propto \underbrace{p(X_{1:n}|\theta)\, p(\theta)}_{\tilde{\pi}(\theta)/Z}$$

↳ Best $M$ will also depend on $q(\theta)$

# Importance Sampling

$$\mathbb{E}_{\theta \sim \pi(\theta)}\left[h(\theta)\right] = \int h(\theta)\,\pi(\theta)\,d\theta$$

Idea: Rewrite this expectation as an expectation taken w.r.t. another distribution $q(\theta)$

$$= \int h(\theta)\,\pi(\theta)\,\frac{q(\theta)}{q(\theta)}\,d\theta$$

Importance weight: $w(\theta) = \frac{\pi(\theta)}{q(\theta)}$

$$= \int h(\theta)\left(\frac{\pi(\theta)}{q(\theta)}\right) q(\theta)\,d\theta$$

integrand

$$= \mathbb{E}_{\theta \sim q(\theta)}\left[h(\theta)\,w(\theta)\right]$$

# Algorithm (Importance Sampling)

$$\theta^{(1)}, \dots, \theta^{(M)} \overset{\text{i.i.d.}}{\sim} q(\theta)$$

$$\hat{I}_{IS}(h) \simeq \frac{1}{M} \sum_{m=1}^{M} \underbrace{h(\theta^{(m)}) w(\theta^{(m)})}_{g(\theta^{(m)})}$$

$$\xrightarrow{\text{a.s.}} \mathbb{E}_{\theta \sim \pi(\theta)}\left[h(\theta)\right] \quad \text{by SLLN}$$

## Condition for convergence:

$$\int |h(\theta)| w(\theta) q(\theta) d\theta < \infty$$

**Limitations:** ① $\mathbb{V}\left[\hat{I}_{IS}(h)\right] \propto \mathbb{V}\left[\frac{h(x)\pi(x)}{q(x)}\right]$

If $q$ is poorly chosen, this variance can be large

The optimal choice:

$$q(\theta) \propto |h(\theta)| \pi(\theta)$$

② Bayesian inference:

$$\pi(\theta) = \frac{\tilde{\pi}(\theta)}{Z}$$

we don't know $Z$

plug back into our estimator

$$\tilde{w}(\theta^{(m)})$$

$$\hat{I}_{IS}(h) = \frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)}) \frac{\tilde{\pi}(\theta^{(m)})}{q(\theta^{(m)})} Z$$

$$= \frac{1}{MZ} \sum_{m=1}^{M} h(\theta^{(m)}) \tilde{w}(\theta^{(m)})$$

$$Z = \int \tilde{\pi}(\theta) d\theta \approx \mathbb{E}_{\theta \sim q(\theta)} \left[ \tilde{w}(\theta) \right]$$

$\xrightarrow{a.s.} Z$

$$= \frac{1}{M\left(\frac{1}{M}\sum_{m=1}^{M}\tilde{w}(\theta^{(m)})\right)} \sum_{m=1}^{M} h(\theta^{(m)})\tilde{w}(\theta^{(m)})$$

$\xrightarrow{a.s.} Z$ $\qquad \xrightarrow{a.s.} \mathbb{E}_{\theta\sim q(\theta)}\left[h(\theta)\tilde{w}(\theta)\right]$

$X_n \xrightarrow{a.s.} \mu_X$

$Y_n \xrightarrow{a.s.} \mu_Y$

$$= \frac{1}{\sum_{m=1}^{M}\tilde{w}(\theta^{(m)})} \sum_{m=1}^{M} h(\theta^{(m)})\tilde{w}(\theta^{(m)})$$

$\frac{X_n}{Y_n} \xrightarrow{P} \frac{\mu_X}{\mu_Y}$

a.s.

BY Slutsky's theorem $\xrightarrow{P.}$ $\mathbb{E}_{\theta\sim\pi(\theta)}\left[h(\theta)\right]$

even if the constant $Z$ is unknown

Next time,
- → MCMC
- → Linear models (Next week)
- → Hypothesis test (Rest of course)

$$\hat{I}_{SNIS}(h) = \left(\frac{1}{M}\sum_{m=1}^{M}\tilde{w}(\theta^{(m)})\right)\frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})\,\tilde{w}(\theta^{(m)})$$

$\downarrow$ a.s. $\qquad\qquad$ $\downarrow$ a.s.

$Z \qquad\qquad \mathbb{E}_{\theta\sim q(\theta)}\left[h(\theta)\,\tilde{w}(\theta)\right]$

$$\xrightarrow{P} \frac{1}{Z}\,\mathbb{E}_{\theta\sim q(\theta)}\left[h(\theta)\,\tilde{w}(\theta)\right] \qquad \tilde{w}(\theta) \to \frac{\pi(\theta)}{q(\theta)}$$

$$= \mathbb{E}_{\theta\sim q(\theta)}\left[h(\theta)\,\frac{\tilde{w}(\theta)}{Z}\right]$$

$$= \int h(\theta)\,\frac{\pi(\theta)}{q(\theta)}\,q(\theta)\,d\theta$$

$$= \mathbb{E}_{\theta\sim\pi(\theta)}\left[h(\theta)\right]$$