Introduction to Bayesian Inference

March 15, 2022

1 Principles of Probability

1.1 Stochastic processes (chance regularities)

Example 1. Sum of two dices.
Suppose we have $T = X_1 + X_2$. We know this is Stochastic. $T \asymp \mathcal{N}(\mu, \sigma)$.

Definition 2. Event space ($\mathcal{F}$) mush conform to a series of conditions:
1. The event space contains sample space $\mathcal{S} \in \mathcal{F}$.
2. The event space is closed under compliments.
3. The event space is closed under countable unions, $E_i \in \mathcal{F} \rightarrow (\cup_{i=1}^{\infty} E_i) \in \mathcal{F}$.

1.2 Factorizing joint probabilities

$$p(A, B) = p(A \mid B)p(B)$$

$$p(A, B, C) = p(A \mid B, C)p(B, C) = p(A \mid B, C)p(B \mid C)p(C)$$

1.3 Joint distributions

$$P_{XY}(x, y) = P(X = x, Y = y)$$

Consider the expectation

$$\mathbb{E}[f(x, y)] = \int_{x,y} f(x, y)p(x, y)dxdy$$

1.4 Marginal distribution

$$P_X(x) = \sum_{all\ y_i} P_{XY}(x, y_i)$$

1.5 Conditional distribution

$$P_{X|Y}(x_i \mid y_i) = \frac{P_{XY}(x_i, y_i)}{P_Y(y_i)}$$

The conditional expectation is given by

$$\mathbb{E}[X \mid Y = y_i] = \sum_{x_i \in X} x_i P_{X|Y}(x_i \mid y_i)$$

In continuous setting,

$$\mathbb{E}[g(Y) \mid x] = \int_{\mathbb{R}} g(y) p(y \mid x) dy$$

1.6 Independence

It can be shown that if X and Y are independent, there exists some functons $g(x)$ and $h(y)$ such that:

$$f(x, y) = g(x)h(y), \quad \forall x, y$$

1.7 Covariance

$$cov[x, y] = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}_x[x]\mathbb{E}_y[y]$$

2 Parameter Estimation

2.1 Properties of estimators

Consistency

$$P(|\theta_n - \theta| > 0) \to 0 \; as \; n \to \infty$$

Bias θ_n is unbiased if

$$\mathbb{E}[\theta_n] = \theta$$

Efficiency A UMVUE is considered to be efficient.

MSE

$$MSE = variance + bias^2$$

2.2 Method of moments

The method of moments amounts to matching population moments to sample moments. Using $X_i \sim Ber(\theta)$ as an example,

$$\mathbb{E}[X_i] = \theta$$

Therefore,

$$\hat{\theta} = \frac{1}{N} \sum x_i$$

MoM is consistent, but might not efficient.

2.3 Maximum Likelihood Estimation

First write out the likelihood function

$$\mathcal{L}(\theta \mid x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Our job is to solve

$$\frac{d}{d\theta} \mathcal{L}(\theta \mid x) = 0$$

2.4 Maximum a posteriori estimate

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta \mid x_1, ..., x_n)\pi(\theta)$$

3 Bayes Theorem

3.1 Bayes' Rule

$$P(A, B) = P(A \mid B)P(B)$$

Samely we have

$$P(A, B) = P(B \mid A)P(A)$$

Therefore, we have

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Posterior

$$P(A \mid B)$$

Likelihood

$$P(B \mid A)$$

Prior

$$P(A)$$

Marginal

$$P(B)$$

Samely,

$$posterior = \frac{likelihood \times prior}{marginal}$$

3.2 Inference and Decisions

3.2.1 Classification

$$\arg\min p(mistake) = \sum_{i=1}^{k} p(x_{k \notin j}, C_k)$$

3.2.2 Loss minimization

3.3 Prior introduction

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

The priors can be proper or improper.

3.3.1 Conjugate priors

Conjugate prior are priors that induce a known distribution in the posterior. When computing the posterior probability, if we have a justifiable reason for using pairing the likelihood with a conjugate prior, we will find the posterior paobability is a known distribution. For example, consider

$$X \sim Ber(\theta)$$

The likelihood takes the form

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^k(1-\theta)^{n-k}, \quad k = \sum x_i$$

If we assume the prior take the form of a Beta distribution,

$$f(x|\theta)p(\theta) \propto \theta^k(1-\theta)^{n-k} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

which we recognize as $Beta(\alpha + k, \beta + n - k)$ in θ.

3.3.2 Informative vs non-informative priors

Noninformative priors are priors that suggest ignorance as to the parameters. These are sometimes called vague or diffuse priors. The priors generally cover the region of the parameter space relatively smoothly.

Common noninformative priors include $Unif(-1000, 1000), \mathcal{N}(0, 10000)$.

Note that seemingly vague priors can actually be strongly informative.

Consider the case of modeling a binary model for y following Bernoulli. A common modeling technique would be to transform the problem using the logit function. For instance:

$$y \sim Ber(p)$$

$$p = logit^{-1}(\beta_0 + \beta_1)$$

Placing priors $[\beta_0, \beta_1] \sim \mathcal{N}(0, 100)$ places under weight on 0 through the transform while using a weakly informative prior $[\beta_0, \beta_1] \sim \mathcal{N}(0, 2^2)$ gives a more diffuse effect on the parameter posterior.

3.3.3 Jeffrey's prior

This prior is non-informative in that we don't specify prior information, but it is informative in that we use the data to information to shape the prior.

The fisher information tells us how much information about θ is included in the data.

Formally, Jeffrey's prior is derived by:
$p(\theta) \propto \sqrt{I_n(\theta)}$, where

$$I_n(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ln f(\theta)}{\partial \theta} \right] = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right]$$

4 Common distributions

4.1 Parametric vs. Non-parametric models

4.1.1 Student's t-distribution

With univariate Normal distribution $\mathcal{N}(x|\mu, \tau^{-1})$. The conjugate prior for the precision τ (inverse of variance) is given by a Gamma distribution $\Gamma(\tau|a, b)$.

We can compute the marginal distribution for x by using the prior of the precision and integrating our the dependence of the normal distribution on its precision over all values of precision from 0 to ∞.

$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) G(\tau|a, b) d\tau$$

Sometimes, people call this mixture of individual normal distributions with different variances.

4.2 MLE with Beta distribution

4.3 Gaussian Mixture Model

Arbitraty distribution can be approximated by Gaussian mixture model.

$$p(x) = \sum_k w_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where w_k is the weight of each Gaussian.

It is also possible to create mixture of Bernoulli with $\sum w_k = 1$. The term w_k can be interpreted as the prior probability of picking the "k"th term. The term $p(x)$ can be written using Bayes Theorem

$$p(x) = \sum_k p(k)p(x|k)$$

This can give rise to $p(k|x)$ as responsibilities. The name arises from the fact that the term $p(k|x)$ explains the responsibility that the 'k'th component has in explaining the observation x.

4.4 Non-parametric methods: Kernel Density Estimation

Histogram, a non-parametric method.

Kernel Density Estimation Gaussian kernel.

5 Sampling Algorithms

How we can sample from distributions.

Why don't we just do uniform sampling??

In high-dimensional settting, we have to sample n^d samples, while most of the things may be very concentrated.

5.1 Basic of everything: Sampling $Unif(0, 1)$

Generate a random number from 0 to 1.

5.2 Sampling discrete $Unif(0, n)$

We can use $int(n \cdot c),\ c \sim Unif(0, 1)$.

5.3 Inverse transform sampling

If a RV Y is generated by applying function F to X we get

$$Y = F(X)$$

which implies that we can apply an inverse transformation to Y (if exists) to obtain X.

$$F^{-1}(Y) = X$$

5.3.1 Derivation

1. PDF of Y is $p(y)$
   2. PDF of Z is $p(z)$

$$y = f(z)$$

The distribution of Y

$$p(y) = p(z)|\frac{dz}{dy}|$$

If $Z \sim Unif(0, 1)$,

$$p(y) = |\frac{dz}{dy}|$$

Therefore,

$$z = \int_{-\infty}^{y} p(y)dy = h(y)$$

Generate y by

$$y = h^{-1}(z)$$

5.3.2 Algorithm

1. Calculate the inverse of CDF, given by F^{-1}
   2. Sample from $Z \sim Unif(0, 1)$
   3. Use sampled Z_i to obtain

$$Y \sim F^{-1}(Unif(0, 1))$$

5.3.3 Example

An exponential distribution is given by the PDF

$$p(y) = \lambda e^{-\lambda y}$$

The CDF is given by

$$F(y) = 1 - e^{-\lambda y}$$

Consider the inverse function

$$F(y) = h(y) = z = 1 - e^{-\lambda y}$$

We then have

$$y = -\log(1 - z)/\lambda$$

5.4 Rejection sampling

1. Draw z_0 from $q(z)$ and compute $kq(z_0)$
2. Draw a uniform number u_0 from $[0, kq(z_0)]$
3. If $u_0 > p(z_0)$, the sample is rejected otherwise save u_0.
4. Continue with (1) - (3) until enough samples are drawn.

The samples are accepted with probability $\frac{p(z)}{kq(z)}$. For effective rejection sampling, we want the number of samples that are rejected to be minimal. This is possible only when the envelope distribution is close to the desired distribution. It is also inefficient to use in high dimensional spaces for the following reasons:

- The ideal value of k in a 'D' dimensional space is given by $(\frac{\sigma^q}{\sigma_p})^D$, which may be very large.

- The acceptance ratio for two normalized fistributions with densities $p(x)$ and $q(x)$ is simply $\frac{1}{k}$

- This would be extremely inefficient

5.5 Importance sampling

Importance sampling is useful for computing terms such as the expectation of a function $f(x)$ with distribution $p(x)$.

Ideally, we want to sample in space where the product $f(x)p(x)$ is high since the expected value is computed for a discrete distribution as

$$\mathbb{E}[f] = \sum_i p(x_i) f(x_i)$$

Or, for continuous cases

$$\mathbb{E}[f] = \int p(x) f(x) dx$$

Importance sampling is also based on the idea of using another proposal distribution that is easier to sample from, compared to the original distribution $p(x)$.

$$\mathbb{E}[f] = \int f(x)p(x)dx = \int \frac{p(x)}{q(x)} q(x) f(x) dx \approx \frac{1}{L} \overset{drawn\ from\ Q(x)}{\sum_l} \frac{p(x)}{q(x)} f(x)$$

The term $\frac{p(x)}{q(x)}$ is known as importance weights. Similar to rejection sampling, the envelope distribution should be close to the desired distribution for efficient sampling.

6 Bayesian vs. Frequentists

6.1 Likelihood

Likelihood is type of probability that has already been observed given a certain hypothesis parameters.

$$Likelihood(\theta \mid data) = P(x \mid \theta)$$

Likelihood is not a probability (not integrate to 1). Therefore we call it likelihood (since not integrating to 1).

6.2 Inference

Inference refers to the process of identifying the distribution of the parameters that represent out hypothesis. We can denote posterior as

$$P(\theta \mid X)$$

6.3 Features of Bayesian Inference

1. Assign a probability to both hypothesis (Posterior) and data (Likelihood)

2. Utilize expert knowledge through the formulation of 'subjective' priors. The use of priors has been a source of debate.

However, when this is clearly stated it allows everyone to understand and challenge the assumption behind the results possibly allowing for refinement of the priors.

3. Can be computationally expensive to compute the posterior (need to integrate over several parameters).

A lot of times, we have to resort to approximate techniques since the integrals associated with the posterior calculation in Bayesian statistics cannot be computed analytically.

A number of approximation techniques are employed.

3.1 Laplacian approximation

3.2 Variational approximation

3.3 Monte Carlo techniques (*)

3.4 Message passing algorithms

7 Model Performance

7.1 Overfitting vs. Underfitting

7.2 R^2 and Explained Variance

Derivation of R^2.

If we observe data given by y_i, such that the fitted model predicts f_i for each point i, we can write the mean of all the observed data, given by y_{mean} as

$$y_{mean} = \frac{1}{n}\sum y_i$$

Total sum of squares, which is proportional to the variance of the data, is

$$SS_{tot} = \sum (y_i - y_{mean})^2$$

The residual sun of squares (also called the error)

$$SS_{res} = \sum (y_i - f_i)^2$$

Now, R^2 is defined as

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ &= 1 - UnexplainedVariance \\ &= ExplainedVariance \end{aligned}$$

7.3 Cross-Validation

k-fold validation.

7.4 Information Criteria

1. Log-likelihood
2. Akaike Information Criterion (AIC)
3. Widely Applicable Information Criterion (WAIC)
4. Deviance Information Criterion (DIC)
5. Bayesian Information Criterion (BIC)

For (2) to (5),

- They have similar form like:

$$metric = model\ fit + penalization$$

- The model fit is measured using log-likelihood, penalization normally represents uncertainty of parameter
- Lower values imply a better fit

AIC, BIC, DIC use the joint probability of the data, wheras WAIC computes the pointwise probability of the data.

In the following, we assume the model params are independent, thereby the joint probability is the same as the product of the pointwise estimates.

7.5 Log-likelihood and Deviance

In a normal sense, the Mean Squared Error is given by

$$MSE = \sum (y_{true} - y_{pred})^2 / n$$

7.5.1 Log-likelihood

$$Loglikelihood = \sum \log p(y_i|\theta)$$

If the likelihood function is Normal, we have log-likelihood porpotional to MSE.

7.5.2 Deviance

$$Deviance = -2 \sum (\log p(y_i|\theta) - \log p(y_i|\theta_s))$$

7.5.3 A note on MLE

Find θ that maximize $\sum p(y_i|\theta)$.

7.6 Posterior Predictive Distribution

Bayesian estimate of posterior predicrive distribution. This allows us to measure the model's probability of generating the new data i.e. $p(y_{new}|y)$.

This can be interpreted as asking "What is the probability of seeing the new out-of-sample data, given the model that was trained on the in-sample data?". The predictive accuracy can be written as

$$accuracy = p(y_{new}|y) = \int p(y_{new}|\theta) P(\theta|y) d\theta,$$

where $p(\theta|y)$ is the posterior distribution for θ and we integrate over the entire distribution of θ.

Now this is simply the expectation of $p(y_{new}|\theta)$ over the posterior distribution of θ.

In simple terms, it is the average of all the probabilities of seeing y_{new} calculated over all possible values of θ.

$$accuracy = \mathbb{E}[p(y_{new}|\theta)]$$

This has the following steps

1. Draw a θ_i from the posterior distribution for θ.
2. Given the θ_i, how likely we can observe y_{new} (compute $p(y_{new}|\theta)$)
3. Repeat (1) and (2) several times to compute this expectation.

This is also computed using the log frequently as

$$accuracy = \log(\mathbb{E}[p(y_{new}|\theta)])$$

7.7 AIC, BIC, DIC, and WAIC

7.7.1 Akaike Information Criterion (AIC)

The AIC is derived from the world of Frequentist statistics and does not use the posterior distribution.

Therefore, instead of integrating over the posterior, it uses the MLE estimate of θ. The term $\mathbb{E}[p(y_{new}|\theta)]$ is now replaced by $p(y_{new}|\theta_{MLE})$.

$$AIC = -2\sum \log p(y_i|\theta_{MLE}) + 2n_{params}$$

Here n_{params} refers to the number of parameters in the model and θ_{MLE} is the MLE estimate of θ. We want a model with a lower AIC and the second term is intended to penalize complex models.

7.7.2 Bayesian Information Criterion

$$BIC = -2\sum \log p(y_i|\theta_{MLE}) + n_{param} \log n_{samples}$$

7.7.3 Deviance Information Criterion

We use θ_{Bayes} as the posterior mean for DIC.

$$DIC = -2\sum \log p(y_i|\theta_{Bayes}) + 2Var_{\theta \sim posterior}[\log p(y_i|\theta)]$$

7.7.4 Widely Applicable Information Criterion (WAIC)

WAIC is a Bayesian extension to AIC. The derivation for the log pointwise predictive density is similar to what we covered above, but it replicated here to keep it consistent with referred paper.

$$WAIC = -2\sum \log \frac{1}{S}\sum_S p(y_{new_i}|\theta_S) + 2\sum Var_S(\log p(y_{new_i}|\theta_S))$$

7.7.5 Qualitative Discussion

7.8 Entropy and KL Divergence

Entropy is defined as

$$H(x) = -\sum_x p(x) \log p(x)$$

KL divergence is defined as

$$\begin{aligned} KL(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \end{aligned}$$

8 Foundation of Bayesian Inference

8.1 Markov Chains

8.1.1 Stationary distributions

$$X_{t+1} = X_t$$

8.1.2 Ergodicity

1. If we sample s space long enough we will cover almost every point in that space (theoretically)

2. If we obtain a statistics from a sequence such as the mean, this statistics should be the same if we recompute it using a different sequence drawn from the same set of events. The implication here is that there is only one distribution unlike a non-stationary distribution which has an infinite set of PDFs.

8.1.3 Why does this work?

Consider a simple case that

$$A \leftrightarrow B \leftrightarrow C$$

Consider from B to C ($B \to C$).
The transition probability can be written as

$$P_{BC} = 0.5 \cdot \min(\frac{P_C}{P_B}, 1)$$

The transition probability from C to B is

$$P_{CB} = 0.5 \cdot \min(\frac{P_B}{P_C}, 1)$$

Say the ratio is

$$\frac{P_{BC}}{P_{CB}} = \frac{\min(\frac{P_C}{P_B}, 1)}{\min(\frac{P_B}{P_C}, 1)}$$

If $P_C > P_B$

$$\frac{P_{BC}}{P_{CB}} = \frac{P_C}{P_B}$$

If $P_C < P_B$

$$\frac{P_{BC}}{P_{CB}} = \frac{P_C}{P_B}$$

With long enough runs, the amount of samples will be equal to its ratio within the entire PDF. Therefore we can generate such empirical distribution via MCMC.

8.1.4 Proposal distribution

An easy to sample distribution such as Gaussian $q(x)$ such that

$$q(x_{i+1}|x_i) \sim \mathcal{N}(\mu, \sigma)$$

8.2 The Bayesian Inference Process

1. Obtain the data and inspect it for a high-level understanding of dist. and outliers
2. Define a prior for the data based on the understanding of the problem
3. Define a likelihood distribution for the data and obtain the likelihood of the data given this likelihood distribution
4. Obtain posterior distribution using prior (2) and likelihood (3) by applying Bayes Theorem

9 Metropolis Algorithm for Sampling

9.1 Problem statement

We start off by modeling discrete events via Poisson

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}$$

9.2 Outline of Metropolis Algorithm (MA)

What do we want to compute? Estimate distribution of μ.

What do we have available? Observed data!

1. Start with a parameter sample $\mu_{current}$ that is drawn from a distribution
2. Draw a second parameter $\mu_{proposed}$ from a proposal distribution
3. Compute the likelihood for both params
4. Compute the prior probability density for both params
5. Compute posterior probability density of both params by multiplying prior and likelihood from (3) and (4)
6. Select one param from the posterior probability density computed above using a rule and save the selected on as $\mu_{current}$
7. Repeats (2) to (7) till a large number of parameters have been drawn
8. Compute the distribution of the parameter μ by plotting histogram and estimate via MAP (or anything)

9.3 Detailed example with calculation

1. Propose a single value for param $\mu_{current} = 7.5$
2. Compute prior of $\mu = 7.5$. Using Gamma prior

$$Gamma(\mu = 7.5|\alpha, \beta) = \beta^{\alpha} 7.5^{\alpha-1} e^{-7.5\beta}/\Gamma(\alpha)$$

3. Compute the likelihood of single point data 'x'. Given the param of 7.5. The likelihood distribution was a Poisson distribution given by

$$Poisson(x|\mu = 7.5) = e^{-\mu}\mu^{x}/x! = e^{-7.5}7.5^{x}/x!$$

4. Compute the posterior density via

$$Posterior \propto Prior \cdot Likelihood$$

By Gamma-Poisson, we know

$$\alpha_{posterior} = \alpha_{prior} + \sum_{i=0}^{n} x_i$$
$$\beta_{posterior} = \beta_{prior} + n$$

5. Propose a second value $\mu_{proposed}$, which is drawn from a distribution called a proposal distribution centered on $\mu_{current}$. The Standard deviation is a tunable parameter. We assume we drawn 8.5.

6. Select one value from the current and the proposed value with the following two steps
   a. Compute the probability of moving to the proposed value as

$$p_{move} = \min(\frac{P(\mu_{proposed}|data)}{P(\mu_{current}|data)}, 1)$$

7. If we moved to the proposed value, save the proposed value.

Traceplot

10 Gibbs sampling

Let's consider $\mathcal{N}(\mu, \tau)$ we will need the conjugate solution for computing posterior.

If

$$\mu \sim \mathcal{N}(\mu_{prior}, \tau_{prior})$$

Select $\mu_{prior} = 12$ and $\tau_{prior} = 0.625$ which corresponds to a $\sigma = 4$.
Select the shape parameter $\alpha_{prior} = 25$ and the rate parameter $\beta_{prior} = 0.5$.
The conjugate solution is

$$\mu_{posterior} = (\tau_{prior}\mu_{prior} + \tau_0 \sum x_i)/(\tau_{prior} + n\tau_0)$$

$$\tau_{posterior} = \tau_{prior} + n \times \tau_0$$

Conjugate solution for τ with a Gamma prior

$$\alpha_{posterior} = \alpha_{prior} + n/2$$

$$\beta_{posterior} = \beta_{prior} + \sum (x_i - \mu_1)^2/2$$

10.1 Outline of Algorithm

1. Specify priors for μ, τ
2. Choose τ to start and select τ_0 from the Gamma prior distribution.
3. Start first trial. We obtain a sample for μ from the posterior distribution of μ given the value of τ_0.
4. We continue trail 1 since we need to obtain a value for τ_1 conditional on the value μ_1. Similar to step (3), we use conjugate solution to obtain posterior with sampled μ_1. Deaw τ_1 from posterior.

5. Accept both values in (3) and (4).

6. Repeat (3) to (5) till we have sufficient number of samples. Iteratively update the params to coordinate ascent the optimization.

11 Hamiltonian Monte Carlo

Consider position x, momentum m and Hamiltonian H through

$$\frac{dx}{dt} = \frac{dH}{dm}$$

$$\frac{dm}{dt} = -\frac{dH}{dx}$$

These differential equations depend on the probability distributions we trying to learn.

We navigate these distributions by moving around them in a trajectory using steps that are defined by the position and momentum at that position.

HMC is based on conservation of energy. The sum of kinectic and potential energy of particle. Or just saying the total energy of the system.

$$H(x, m) = U(x) + KE(m)$$

where $U(x)$ is potential energy and $KE(m)$ is the kenetic energy.

The potential energy is measured using the negative log density of posterior distribution. When the sampler is far away from the probability mass center, it has high potentital energy but low kinetic energy.

When the trajectory is closer to the center, it will have high kinetic energy but low potential energy. The KE involves a mass matrix Σ that is also the covariance of normal distribution from which we randomly draw a momentum value m in our Monte Carlo process. An outline of the steps involved in this algorithm is given below.

11.1 Outline of Algorithm

- We start from initial position x_0
- Each step, we select a random value for momentun from a proposal distribution. This is usually a normal distribution such that

$$m \sim \mathcal{N}(\mu, \Sigma)$$

- From the current position and using the sampled value for momentum, we run the particle for time $L \cdot \Delta t$ using leapfrog integrator which is a numerical integration scheme to march forward in time. This terms Δt refers to the time step taken for the integrator, and L refers to the total number of steps taken. L is a hyperparam that needs to be tuned carefully. If we are at a spatial location indicated by step n, we start from time 0 (integration

time) and integrate till time t to get the following

$$x_n(0) \rightarrow x_n(L\Delta t)$$

$$m_n(0) \rightarrow m_n(L\Delta t)$$

- The leapfrog can introduce error. We can correct the error using MH step that probabilistically accepts new values of x_{n+1} as $x_n(L\Delta t)$ or $x_n(0)$. The acceptance probability here is given below.

$$acceptance = \frac{p(x_n(L\Delta t))}{p(x_n(0))} \times \frac{q(m(L\Delta t))}{q(m(0))}$$

  Here, $p(x_n(L\Delta t))$ corresponds to the posterior probability density at the end of the integration scheme and $p(x_n(0))$ corresponds to the posterior probability density at the beginning of the integration scheme. Also, $q(m)$ is the probability density of the proposal distribution for the momentum.

- Draw a random value u from a uniform distribution $Uni[0, 1]$ and perform MH acceptance.

- Record new position x_{n+1}. We repeat this step for n times.

11.2 Impact of $T = L\Delta t$

When there are divergences the sampling process that happens in regions of high curvature, we might have to resort to smaller values of Δt.

The use of larger than desirable values for $T = L\Delta t$ results in the sampler making U-turns at high curvature positions.

The No U-Turn sampler is a HMC that L is automatically tuned.

12 Properties of MCMC

12.1 Representativeness

The samplers from the MCMC process should be representative of the posterior distribution, it should cover the distribution space thoroughly. The final state of the inferred distribution should be independent of the initial value.

There are two ways to measure if your inferred distirbution is representative of true distribution.

1. Visualize inspection using a trace of convergence.
2. Numerically measures for convergence.

12.2 Efficiency

Choose the correct sampling method.

13 MCMC

13.1 Monte Carlo estimation

We know that

$$\begin{aligned}\frac{\pi}{4} &= \mathbb{E}[x^2 + y^2 \leq 1] \\ &\approx \frac{1}{M}\sum_{s=1}^{M}[x_s^2 + y_s^2 \leq 1] \\ x_s, y_s &\sim \mathcal{U}(0,1)\end{aligned}$$

In general, we wish to estimate

$$\begin{aligned}\mathbb{E}_{x\sim p(x)}[f(x)] &\approx \frac{1}{M}\sum_{s=1}^{M} f(x_s) \\ x_s &\sim p(x)\end{aligned}$$

Why do we need to estimate expected value?
Example 1. Full Bayesian inference

$$\begin{aligned}p(y|x, Y_{train}, X_{train}) &= \int \underbrace{p(y|x,w)}_{neural\ net}\underbrace{p(w|Y_{train}, X_{train})}_{All\ possible\ networks} dw \\ &= \mathbb{E}_{p(w|Y_{train},X_{train})}[p(y|x,w)]\end{aligned}$$

We can have the posterior via

$$p(w|Y_{train}, X_{train}) = \frac{\overbrace{p(Y_{train}|X_{train}, w)}^{loss}\overbrace{p(w)}^{prior}}{Z}$$

Example 2. M-step of EM-algorithm

$$\max_{\theta} \mathbb{E}_q \log p(X, T|\theta)$$

13.2 Sampling from 1-d distributions

13.2.1 Discrete distributions

$$r \sim \mathcal{U}[0,1]$$

Suppose we have the case that

$$p(A = a_1) = 0.6, p(A = a_2) = 0.1, p(A = a_3) = 0.3$$

Basically, 1d discrete distribution with finite number of values are easy.
At least then number of values are $< 100{,}000$.

13.2.2 Continuous sampling (rejection)

How can we generate samples from $\mathcal{N}(0,1)$?

Use CLT.

$$z = \sum_{i=1}^{12} x_i - 6, x_i \sim \mathcal{U}[0,1]$$

Then,

$$p(z) \approx \mathcal{N}(0,1)$$

What if we have a mixture of Gaussian, something non-convex?

Let's upper-bound this distribution by $q(x) = \mathcal{N}(1, 3^2)$, and have $p(x) \leq 2q(x)$.

First, we have

$$\tilde{x} \sim q(x), y \sim \mathcal{U}[0, 2q(\tilde{x})]$$

We have to reject some of the points.

Accept $\tilde{x}$ with probability $\frac{p(x)}{2q(x)}$: if $y \leq p(x)$.

How efficient is it?

$$p(x) \leq Mq(x)$$

Accepts $\frac{1}{M}$ points on average.

$$\hat{p}(x) \leq \underbrace{ZM}_{\tilde{M}} q(x)$$

Pros:

Work for most distributions

Cons:

If p and q are too different, we will have large M, then the sampling will be very slow.

M will also be very large for d-dimensional distributions.

13.3 Markov Chain Monte Carlo

13.3.1 Markov Chains

$$T(L \to L) = 0.3, T(L \to R) = 0.7, T(R \to L) = 0.5, T(R \to R) = 0.5.$$

13.3.2 Using Markov Chain

- We want to sample from $p(x)$
- Build a Markov Chain that converge to $p(x)$
- Start from any x^0
- For $k = 0, 1, ...$

$$x^{k+1} \sim T(x^k \to x^{k+1})$$

- Eventually x^k will look like samples from $p(x)$.

13.3.3 Do Markov Chains always converge?

If we have such system that

$$T(L \to L) = 0, T(L \to R) = 1, T(R \to L) = 1, T(R \to R) = 0.$$

This never converge.

Definition 3. A distribution π is called stationary if

$$\pi(x') = \sum_x T(x \to x')\pi(x)$$

Theorem 4. *If $T(x \to x') > 0$ for all x, x', then exists unique π:*

$$\pi(x') = \sum_x T(x \to x')\pi(x)$$