COMPSCI 689
Lecture 21: Exact and Approximate Bayesian Inference

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

Bayesian Probabilistic Modeling

- The fundamental idea of the Bayesian approach to probabilistic modeling is to treat unknown model parameters as latent variables.
- The term *Bayesian inference* refers to the problem of computing the parameter posterior $P(\theta|\mathcal{D})$, which takes the place of learning.
- Predictions are made by integrating over the full parameter posterior instead of plugging in a summary like the posterior mode (ie: MAP estimation).

Definitions I

- Parameters: $\theta \in \mathbb{R}^K$
- Data (Unsupervised): $\mathcal{D} = \{\mathbf{x}_n\}_{1:N}$
- Likelihood (Unsupervised): $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} P(\mathbf{X}_n = \mathbf{x}_n|\theta)$
- Data (Supervised): $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{1:N}$
- Likelihood (Supervised): $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} P(Y_n = y_n|\mathbf{X}_n = \mathbf{x}_n, \theta)$

Definitions II

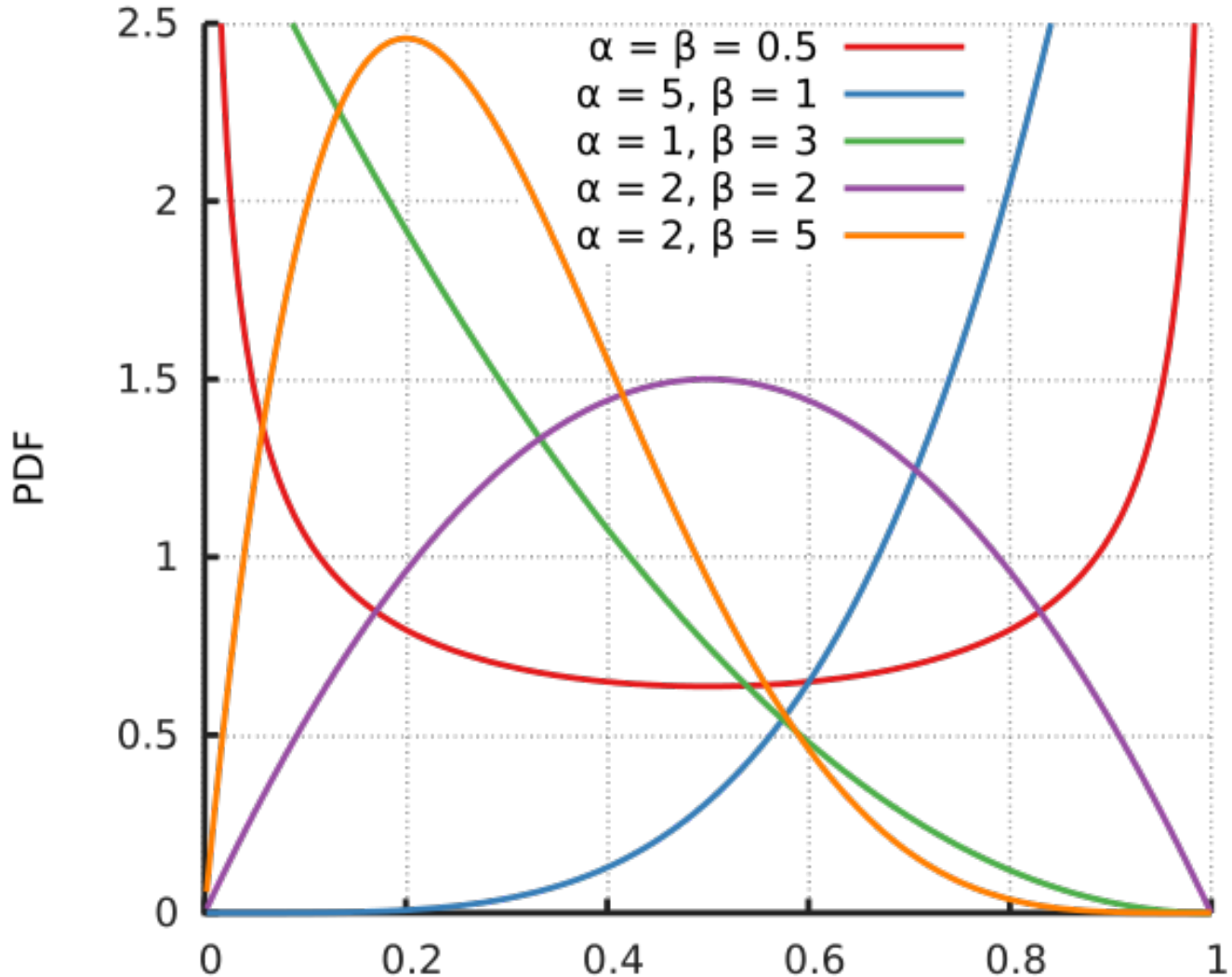
- Prior: $P(\theta)$
- Joint: $P(\mathcal{D}, \theta) = P(\mathcal{D}|\theta)P(\theta)$
- Evidence: $P(\mathcal{D}) = \int P(\mathcal{D}, \theta)d\theta = \int P(\mathcal{D}|\theta)P(\theta)d\theta$
- Parameter Posterior: $P(\theta|\mathcal{D}) = \frac{P(\mathcal{D},\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int P(\mathcal{D}|\theta)P(\theta)d\theta}$
- Posterior Predictive Distribution (Unsupervised): $P(\mathbf{x}|\mathcal{D}) = \int P(\mathbf{x}|\theta)P(\theta|\mathcal{D})d\theta$
- Posterior Predictive Distribution (Supervised): $P(y|\mathbf{x}, \mathcal{D}) = \int P(y|\mathbf{x}, \theta)P(\theta|\mathcal{D})d\theta$

The Beta-Bernoulli Model

- The Beta-Bernoulli model is a classical application of the ideas of Bayesian inference.
- This model has binary data $x \in \{0, 1\}$ with a Bernoulli likelihood $P(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)}$.
- The prior distribution on the unknown parameter $\theta \in [0, 1]$ is given by the Beta distribution:

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

The Beta Distribution



The Beta Distribution

- The Beta distribution is the *conjugate prior* to the Bernoulli likelihood. A conjugate prior has the same functional form as the terms in the likelihood that involve the parameters.
- The use of a conjugate prior results in a posterior that belongs to the same family of distributions as the prior.
- The function $\Gamma(x)$ in the definition of the Beta distribution is called the *gamma function* and is a generalization of the factorial function to the real numbers. For positive integers, it satisfies the property $\Gamma(x+1) = x\Gamma(x)$.
- Note that since we know that $P(\theta)$ is a normalized probability density, we have that $\int P(\theta)d\theta = 1$, and thus:

$$\int \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Beta-Bernoulli Inference: Joint

To start, note that we can simplify the likelihood using N_1 as the number of 1's in the data set and N_0 as the number of 0's:

$$P(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{N_1}(1-\theta)^{N_0}$$

The joint distribution is then given by:

$$\begin{aligned} P(\mathcal{D},\theta) &= \theta^{N_1}(1-\theta)^{N_0} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \end{aligned}$$

Beta-Bernoulli Inference: Evidence

To get the evidence, we marginalize the joint over the parameters. This uses the unnormalized Beta integral result we noted above:

$$P(\mathcal{D}) = \int P(\mathcal{D}, \theta) d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} d\theta$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)}$$

Beta-Bernoulli Inference: Parameter Posterior

Now, we can get the parameter posterior as shown below:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D},\theta)}{P(\mathcal{D})}$$

$$= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}}{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)}}$$

$$= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$= \text{Beta}(N_1+a, N_0+b)$$

Beta-Bernoulli Inference: Parameter Posterior

Alternate derivation:

$$
\begin{aligned}
P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D},\theta)}{P(\mathcal{D})} \\
&\propto P(\mathcal{D},\theta) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \\
&\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \\
&\rightarrow \text{Beta}(N_1+a, N_0+b)
\end{aligned}
$$

Beta-Bernoulli Inference: Posterior Predictive

Finally, we can obtain the posterior predictive distribution:

$$
\begin{aligned}
P(X=1|\mathcal{D}) &= \int P(X=1|\theta)\cdot P(\theta|\mathcal{D})d\theta \\
&= \int \theta \cdot P(\theta|\mathcal{D})d\theta \\
&= \int \theta \cdot \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}d\theta \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\int \theta^{N_1+a+1-1}(1-\theta)^{N_0+b-1}d\theta \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)}\frac{\Gamma(N_1+a+1)\Gamma(N_0+b)}{\Gamma(N+a+b+1)}
\end{aligned}
$$

Beta-Bernoulli Inference: Posterior Predictive

$$
\begin{aligned}
P(X=1|\mathcal{D}) &= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)} \frac{\Gamma(N_1+a+1)\Gamma(N_0+b)}{\Gamma(N+a+b+1)} \\
&= \frac{\Gamma(N+a+b)}{\Gamma(N_1+a)\Gamma(N_0+b)} \frac{\Gamma(N_1+a)(N_1+a)\Gamma(N_0+b)}{\Gamma(N+a+b)(N+a+b)} \\
&= \frac{N_1+a}{N+a+b} \\
&= \mathbb{E}_{Beta(N_1+a,N_0+b)}[\theta]
\end{aligned}
$$

Analysis

- Given N_1 ones and N_0 zeros in the data set, the posterior predictive distribution gives the probability that the next observation will be a one as $\frac{N_1+a}{N+a+b}$.
- If we use the MLE as a pug-in estimate, we obtain $\theta_{MLE} = N_1/N$.
- If we compute the posterior mode, we would find $\theta_{MAP} = \frac{N_1+a-1}{N+a+b-2}$.

Analysis

- Further, we can see that as N goes to infinity, the effect of the prior will go to zero as expected.
- However, when N is small, the three predictive probabilities can differ substantially.
- In general, the MLE will be the worst conditioned. With only one observation in the data set, the MLE predicts that all future data cases will match the first observation with probability 1. This is nonsense (almost always).
- The MAP will be better conditioned, but only if a and b are strictly greater than one.
- The Bayesian posterior predictive distribution makes a non-degenerate prediction when a and b are any non-zero values.

Problems with Bayesian Inference

- The primary theoretical issue with Bayesian inference is the subjectivity of the choice of prior.
- The primary technical issue with Bayesian inference is that the required integrals are not always analytically tractable.
- This leads to the need for approximate Bayesian computations.
- The most common framework for approximate Bayesian computation is the Monte Carlo framework, which we will discuss next.

Monte Carlo Methods

- Monte Carlo methods are an important class of solutions to the Bayesian computation problem.
- The basic Monte Carlo approximation result is the following:

$$\mathbb{E}_{P(X)}[f(X)] = \int P(X = x) f(x) dx \approx \frac{1}{S} \sum_{s=1}^{S} f(x_s), \quad x_s \sim P(X)$$

- This result says that we can approximate an expectation with respect to a distribution $P(X)$ by drawing independent samples from $P(X)$ and computing an average over the samples.
- Such an approximation is unbiased for any S and the weak law of large numbers states that such an average will converge to the corresponding expectation as S increases.

Monte Carlo Methods for Prediction

- An important application of Monte Carlo approximation is approximating the posterior predictive distribution.
- Recall that the general posterior predictive distribution in the supervised case is given by:

$$P(Y = y|\mathbf{x}, \mathcal{D}) = \int P(Y = y|\mathbf{x}, \theta)P(\theta|\mathcal{D})d\theta$$

- If the integral is intractable, but we can get samples θ_s from $P(\theta|\mathcal{D})$, we can approximate the posterior predictive distribution using:

$$P(Y = y|\mathbf{x}, \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} P(Y = y|\mathbf{x}, \theta_s)$$

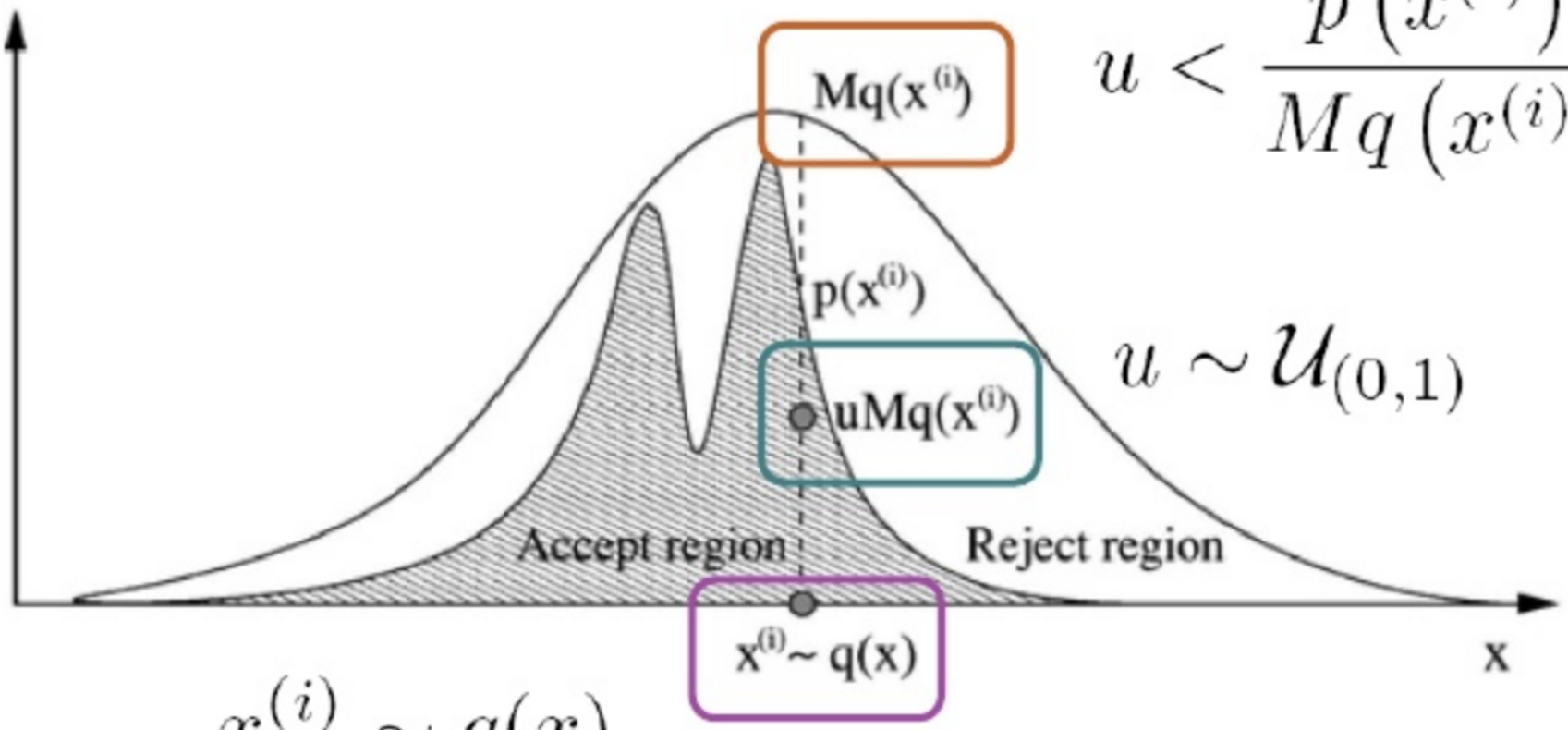
Samplers

- The core of Monte Carlo methods are algorithms for sampling from complex probability distributions.
- There are many, many samplers with different kinds of properties. Some methods allow for valid sampling from $P(\theta|\mathcal{D})$ even if the normalizing constant is intractable.
- One of the most basic methods is rejection sampling.

Rejection Sampling

- To sample from a distribution $P(X)$ using a rejection sampler, we propose candidate samples from a simpler distribution $Q(X)$.
- $Q(X)$ must have the property that $P(X = x) \leq MQ(X = x)$ for all x and a fixed constant M.
- We sample $x \sim Q(X)$ and then draw a uniform random number $u \sim U(0, 1)$. We accept x as a sample if $\frac{P(X=x)}{M \cdot Q(X=x)} \geq u$.
- Interestingly, this algorithm doesn't require the normalization constant of $P(X)$ to be known. If $P(X) = \tilde{P}(X)/Z$ and we have an M such that $\tilde{P}(X = x) \leq MQ(X = x)$ for all x, then the method still works.

Rejection Sampling

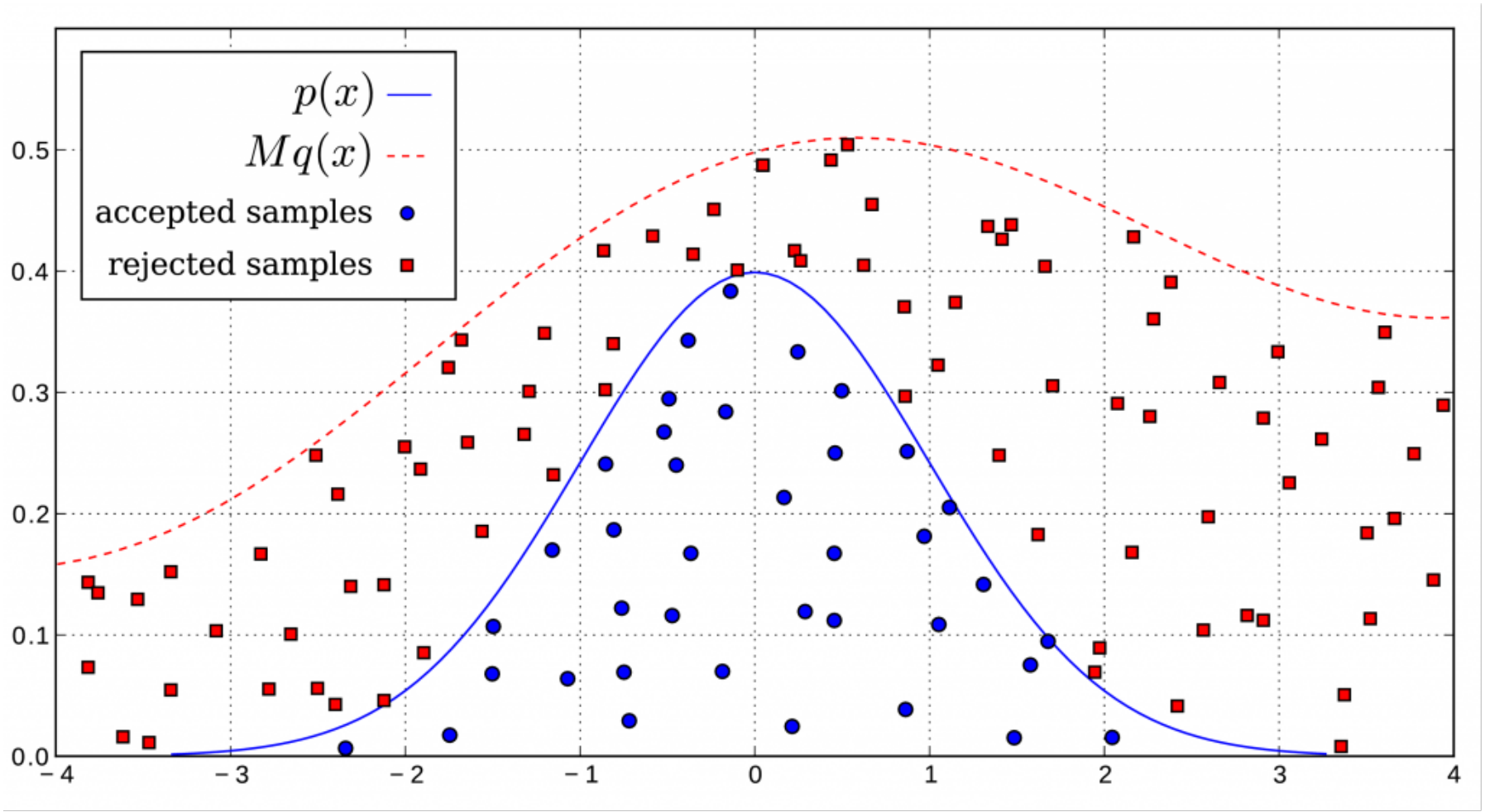


$$u < \frac{p\left(x^{(i)}\right)}{Mq\left(x^{(i)}\right)}$$

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

$$x^{(i)} \sim q(x)$$

Rejection Sampling



Rejection Sampling: Proof of Correctness

- Let A be a binary random variable representing the fact that a sample x is accepted.
- The assertion is that the distribution of accepted samples $Q(X = x|A = 1)$ is equal to the target distribution $P(X)$.
- $Q(X = x|A = 1) = \frac{Q(X=x,A=1)}{Q(A=1)}$
- $Q(X = x, A = 1) = Q(X = x)Q(A = 1|X = x) = ?$
- $... = Q(X = x)\frac{P(X=x)}{MQ(X=x)} = P(X = x)/M$
- $Q(A = 1) = \int Q(X = x, A = 1)dx = \int \frac{P(X=x)}{M} = 1/M$
- $Q(X = x|A = 1) = \frac{P(X=x)/M}{1/M} = P(X = x)$

Rejection Sampling is Inefficient

- An important question about the rejection sampler is its efficiency in terms of what fraction of generated samples are actually accepted. This is the sampler's *acceptance rate*.
- For the rejection sampler, we propose point $\mathbf{x}$ with probability $Q(\mathbf{x})$ and accept the point with probability $\tilde{P}(\mathbf{x})/MQ(\mathbf{x})$. This means the expected acceptance rate is:

$$A_{RS} = \int_{\mathcal{X}} Q(\mathbf{x}) \cdot \frac{\tilde{P}(\mathbf{x})}{MQ(\mathbf{x})} d\mathbf{x} = \frac{1}{M} \int_{\mathcal{X}} \tilde{P}(\mathbf{x}) d\mathbf{x}$$

- The rejection sampler's efficiency is inversely proportional to M.
- Research on Monte Carlo methods focuses on more efficient methods for generating samples from complex distributions.