Probability Theory & Statistics

Monte Carlo Simulation

Prerequisites

- Inequalities (Law of Large Numbers & Central Limit Theorem)

Introduction

The solutions to many scientific problems involve intractable high-dimensional integrals. Standard deterministic numerical integration deteriorates rapidly with the dimension of the space. Monte Carlo methods are stochastic numerical methods that can be used to approximate high-dimensional integrals. They have a diverse range of applications: statistical/quantum physics, econometrics, ecology, epidemiology, finance, signal processing, weather forecasting, machine learning, and Bayesian statistics.

Approximation of Integrals

Riemann Sums

Many scientific problems require numerical computation of integrals, i.e.,

$$I = \int_{\mathbb{X}} f(x)dx \tag{1}$$

where $f : \mathbb{X} \to \mathbb{R}$. For simple choices of functions f and functional spaces $\mathbb{X}$, the integral can be computed exactly, but in general we require numerical approximations of I.

When $\mathbb{X} = [0, 1]$, then we can estimate I through,

$$\widehat{I}_n = \frac{1}{n}\sum_{i=0}^{n-1} f\left(\frac{i + 1/2}{n}\right), \tag{2}$$

called the *Riemann sum approximation*. This corresponds to an approximation of the area under the curve $y = f(x)$ by the sum of the areas of the rectangles (see figure 1).

When f is differentiable and $M = \sup_{x\in[0,1]} \left|f'(x)\right| < \infty$,[1] then the approximation error is $O\left(n^{-1}\right)$.

Proof. The error of the k-th rectangle, for $k \in \{0, \ldots, n-1\}$ is,

$$\varepsilon_k = \left|\int_{k/n}^{(k+1)/n} f(x)dx - \frac{1}{n} f\left(\frac{(k + 1/2)}{n}\right)\right| = \left|\int_{k/n}^{(k+1)/n} \left(f(x) - f\left(\frac{(k + 1/2)}{n}\right)\right)dx\right| \tag{3}$$

where we have used the fact that for all $x, y \in [a, b]$, there exists $c \in [a, b]$ such that,

$$f(x) - f(y) = (x - y)f'(c). \tag{4}$$

[1] There exists some finite maximum derivative on the interior of the domain $[0, 1]$.

Using the bound M on f', and the inequality

$$\left| \int f(x) \right| \leq \int |f(x)| \tag{5}$$

we obtain,

$$\varepsilon_k \leq \int_{\frac{k}{n}}^{\frac{k+1}{n}} \left| f(x) - f\left(\frac{(k+1/2)}{n} \right) \right| dx \leq M \int_{\frac{k}{n}}^{\frac{k+1}{n}} \left| x - \frac{(k+1/2)}{n} \right| dx \leq M \frac{1}{2n^2}. \tag{6}$$

Summing the errors over the n rectangles yield a total error $Mn \times \frac{1}{2n^2} = O\left(n^{-1}\right)$.

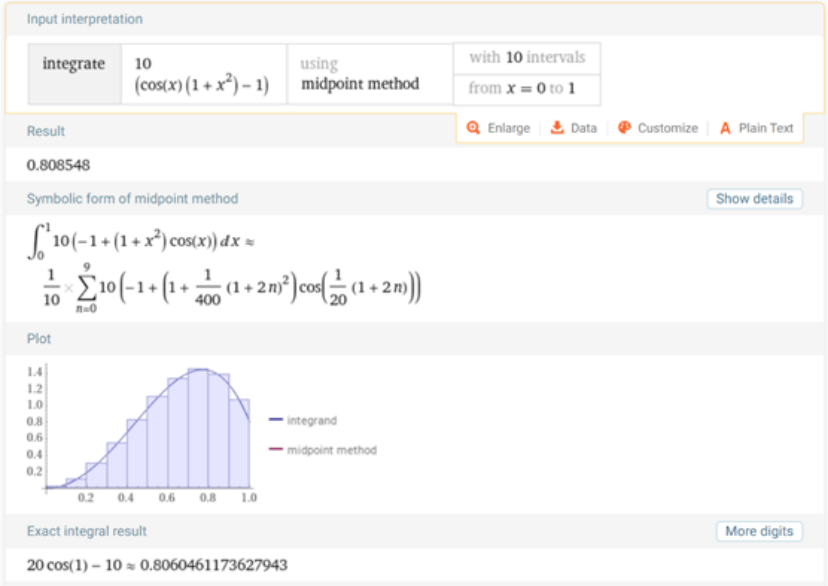

Figure 1: Numerical integration of $f : x \mapsto 10(\cos(x)\left(1 + x^2\right) - 1)$ from `https://www.wolframalpha.com/`.

The Curse of Dimensionality

For $\mathbb{X} = [0,1] \times [0,1]$ assuming,

$$\widehat{I}_n = \frac{1}{m^2} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} f\left(\frac{i+1/2}{m}, \frac{j+1/2}{m} \right) \tag{7}$$

and $n = m^2$, then the approximation error is $O\left(n^{-1/2}\right)$. For $\mathbb{X} = [0,1]^D$ therefore, the approximation error is in $O\left(n^{-1/D}\right)$ – the so-called *curse of dimensionality*.

This suggests that deterministic approximations aren't appropriate for computing high-dimensional integrals. More sophisticated deterministic approximations like the trapezoidal rule or Simpson's rule also suffer from the same degeneracy when the dimension increases.

Stochastic simulation methods, also known as Monte Carlo methods, are the most common tools for performing high-dimensional numerical integrations. Monte Carlo methods were introduced in the 1940s and have become extremely popular in statistics over the past 20 years because they make inferences possible in very complex statistical models. This session reviews Monte Carlo methods and some of their applications in Bayesian statistics and a few other examples.

Monte Carlo Integration

Definition

For $f : \mathbb{X} \to \mathbb{R}$, write,

$$I = \int_{\mathbb{X}} f(x)dx = \int_{\mathbb{X}} \varphi(x)p(x)dx \tag{8}$$

where $p(x)$ is a *probability density function* on $\mathbb{X}$, and

$$\varphi : x \mapsto f(x)/p(x). \tag{9}$$

A *Monte Carlo method* is defined as:

- sample $X_1, \ldots, X_n \sim p$,
- compute

$$\widehat{I}_n = \frac{1}{n}\sum_{i=0}^{n} \varphi(X_i) \tag{10}$$

The *strong law of large numbers* guarantees that $\widehat{I}_n \to I$ almost surely, and the *central limit theorem* tells us that the random approximation error is,

$$O\left(n^{-1/2}\right) \tag{11}$$

whatever the dimension of the state space $\mathbb{X}$!

Proof. Non-asymptotically we can prove this result using the mean-square error,

$$\begin{aligned}\left(I - \widehat{I}_n\right)^2 &= I^2 - 2I \times \widehat{I}_n + \widehat{I}_n^2 \\ &= I^2 - \frac{2I}{n}\sum_{i=0}^{n} \varphi\left(X_i\right) + \frac{1}{n^2}\sum_{i=0}^{n} \varphi\left(X_i\right)^2 + \frac{1}{n^2}\sum_{i\neq j}^{n} \varphi\left(X_i\right)\ \varphi(X_j)\end{aligned} \tag{12}$$

As the samples are IID and $I = E_p(\varphi(X))$, we have,

$$E_p\left(\left(I - \widehat{I}_n\right)^2\right) = I^2 - 2I^2 + \frac{1}{n}E_p\left(\varphi\left(X_1\right)^2\right) + \frac{1}{n^2}n(n-1)I^2 \tag{13}$$

$$= \frac{E_p\left(\varphi\left(X_1\right)^2\right) - I^2}{n} = \frac{Var_p\left(\varphi\left(X_1\right)\right)}{n} \tag{14}$$

Finally, $\sqrt{E_p\left(\left(I - \hat{I}_n\right)^2\right)} = \sqrt{\frac{Var_p(\varphi(X_1))}{n}} \leq \frac{1}{\sqrt{n}}$ if $|\varphi(x)| \leq 1 \ \forall x$. Because the r.h.s. is bounded by a constant, the convergence criterion is independent of the dimension.

According to the rate $\sqrt{n}$ of the CLT (Central Limit Theorem), the variance of the estimator I_n is of order $O\left(n^{-1}\right)$, hence the standard deviation is of order $O\left(n^{-1/2}\right)$. This means that if one wants to divide the standard deviation by 10 (to obtain "10 times" more precision), one needs to sample 100 times more draws from p, which typically corresponds to 100 times more computational effort. This rate of convergence can seem very slow in a single dimension. However, the rate of convergence does not depend on the dimension D_x of the sample space $\mathbb{X}$; it is always $\sqrt{n}$!

Thus, Monte Carlo methods converge slower than Riemann sums in one dimension, whose error was shown to decrease in $O(n^{-1})$; they are of the same accuracy as Riemann sums in two dimensions; and they are faster than Riemann sums for any dimension $D_x \geq 3$. Thus, Monte Carlo methods have become standard tools to approximate integrals of moderate to high dimensions.

In other words, Monte Carlo methods might seem slow, but they are still typically faster than alternative methods. Bakhvalov, Suldin, and other mathematicians have proven results on the minimum error that can be obtained by algorithms using n pointwise evaluations of f to approximate $\int_{\mathbb{X}} f(x)dx$, and the rate $n^{-1/2}$ is found to be optimal when the dimension of $\mathbb{X}$ is large and/or the "smoothness" of f is low, in some sense. For instance, the smoothness of f can be defined as the maximum integer k such that all k-th order partial derivatives of f are uniformly bounded on $\mathbb{X}$.

Note that the rate is $\sqrt{n}$ uniformly in D_x, but high-dimensional integrals are still harder to approximate than low-dimensional integrals, as one would expect. Typically, the error associated with Monte Carlo methods is in $f(D_x)/\sqrt{n}$, where $f(D_x)$ is a polynomial in D_x, or in the worst case, an exponential of D_x. Thus, the error might still be very large when D_x is large, as one might not have enough computational power to scale n with $f(D_x)$. In order to implement the above-described Monte Carlo method, one needs to obtain IID samples from p. Markov Chain Monte Carlo methods provide ways to obtain IID samples from generic distributions p.

In many statistical problems, the integral of interest is an expectation for the form,

$$I = \int_{\mathbb{X}} \varphi(x)\ p(x)dx = E_p\left(\varphi(X)\right), \tag{15}$$

for a specific function φ and density p. The density p is then often called the *target density*. The Monte Carlo method then relies on generating independent copies of,

$$X \sim p, \tag{16}$$

and replacing the expectations with empirical averages,

$$E_p\left(\varphi(X)\right) \approx \frac{1}{n}\sum_{i=0}^{n}\varphi(X_i) \tag{17}$$

Hence, the following relationship between integrals and sampling:

Monte Carlo method to approximate $E_p\left(\varphi(X)\right)$ equivalent to Simulation method to sample p

The Monte Carlo method is thus often referred to as a simulation method.

Applications

Volume of a Convex Body

Let $S \subset [0,\ 1]^D$ be a convex body. In numerous applications, we are interested in computing the volume of this body given by,

$$vol(S) = \int_{[0,1]^D} \mathbb{I}_S(x)dx \tag{18}$$

where $\mathbb{I}_S(x)$ is the *indicator function*, i.e., $\mathbb{I}_S(x) = 1$ if $x \in S$ and 0 otherwise. The "estimating the value of Pi" example is in this class of problems – calculating the volume of a 1-sphere embedded in a $2D$ space. More generally, any volume can be approximated by sampling from $\mathbb{I}_S(X_i)$ and calculating the empirical average over $\mathbb{X}$.

Statistical Mechanics

The Ising model is used to model the behaviour of a magnet and is the best known and most researched model in statistical physics. The magnetism of a material is modelled by the collective contribution of dipole moments of many atomic spins.

Consider a simple $2D$-Ising model on a finite lattice $\mathcal{G} = \{1,\ 2,\ \ldots,\ m\} \times \{1,\ 2,\ ...,\ m\}$ where each site $\sigma = (i,\ j)$ hosts a particle with a $+1$ or -1 spin, modelled as a random variable X_σ. The physical constraints of the system require the probability density of $X = \{X_\sigma\}_{\sigma\in\mathcal{G}}$ on $\mathbb{X} = \{-1,\ 1\}^{m^2}$ be given by the Gibbs distribution,

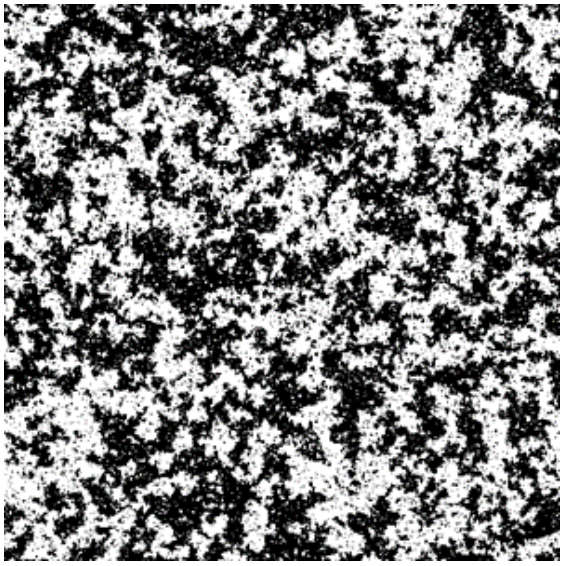

Figure 2: A 2D-Ising model sample.

$$\forall x \in \mathbb{X},\ \ p_\beta(x) = \frac{\exp\left(-\beta U(x)\right)}{Z_\beta} \tag{19}$$

where $p(x)$ is the *probability density*, $\beta > 0$ is the *inverse temperature*, and U is the *potential energy*,

$$\forall x \in \mathbb{X},\ \ U(x) = J \sum_{\sigma \sim \sigma'} x_\sigma x_{\sigma'}, \tag{20}$$

for some $J \in \mathbb{R}$, where $\sigma \sim \sigma'$ refers to the set of pairs of sites that are "neighbours" in some pre-defined sense.

For instance, we can define that two sites $\sigma = (i,\ j)$ and $\sigma' = (i',\ j')$ are neighbors if and only if $\left|i - i'\right| \leq 1$ and $\left|j - j'\right| \leq 1$. According to this form of potential energy, if $x_\sigma = x_{\sigma'}$ and $\sigma \sim \sigma'$, then the probability $p_\beta(x)$ includes a term $\exp(-J)$, otherwise it includes a term $\exp(J)$. Hence the sign of J tells us whether there is a preference for equal or opposite spins at sites σ and σ'.

The normalizing constant Z_β ensures that p_β is a probability distribution, that is, $\sum_{x \in \mathbb{X}} p_\beta(x) = 1$. Thus, it is defined as

$$Z_\beta = \sum_{x \in \mathbb{X}} \exp\left(-\beta U(x)\right). \tag{21}$$

Physicists are often interested in computing $E_{p_\beta}(U(X))$ and Z_β. However, analytic results for the Ising model are very difficult to obtain and physicists often use simulation methods in order to perform these calculations. Note that the problem of computing sums is equivalent to the problem of computing integrals and is formally unified by measure theory.

Although the Ising model was first formulated to solve problems in statistical mechanics it is a special case of the more general class of agent-based models and has applications in various fields including economics and finance.

Financial Mathematics

Let $S(t)$ denote the price of a stock at time t. A *call option* grants the holder the right to buy the stock at a fixed price K, at a fixed time T in the future; the current time being $t = 0$. The is the so-called *European option*.

If at time T the stock price $S(T)$ exceeds the strike price K, the holder exercises the option for a profit of $S(T) - K$. If $S(T) \leq K$, the option expires worthless. The payoff to the holder at time T is thus,

$$\max\left(S(T) - K\right)^+ = \max\left(0,\ S(T) - K\right). \tag{22}$$

The *present value* of the payoff[2] is calculated by discounting the expected future payoff,

$$V(0) = \exp(-rT) E\left(\max\left(S(T) - K\right)^+\right) \tag{23}$$

where $\exp\left(-rT\right)$ is the *discount factor*, r is the *compounded interest rate*, and E, is an expectation with respect to the distribution of the random variable $S(T)$.

If we knew the distribution of $S(T)$, then computing $E\left(\max\left(S(T) - K\right)^+\right)$ would be a low-dimensional integration problem. However, this distribution is typically not available, and we only have access to a stochastic model for $\{S(t)\}_{t \in N}$ of the form,

$$S(t+1) = g\left(S(t),\ W(t+1)\right) = g\left(g\left(S(t-1),\ W(t)\right),\ W(t+1)\right) = g^2\left(S(t-1),\ W(t),\ W(t+1)\right) = g^n\ (\tag{24}$$

[2]Money has an intrinsic time-value we need to consider.

where $W(t)$ is an IID sequence of random variables with probability density functions $\{p_W\}_{t\in\mathbb{N}}$ and g is a known nonlinear mapping – the *payoff function*.

We can rewrite the expectation in the present value equation as

$$E\left(\max\left(S(T)-K\right)^+\right) = \int \max\left(g^n\left(s(0),\ w(1),\ \ldots, w(T)\right)-K\right)^+ \cdot \prod_{t=1}^{T} p_W\left(w(t)\right)\, dw(1)\ldots dw(T) \tag{25}$$

which is a high-dimensional integral whenever T is large.

Bayesian Statistics

We will now consider several examples from Bayesian statistics. In statistics the *data* is usually a collection of n values $(y_1, \ldots,\ y_n) \in \mathcal{Y}^n$ in some space $\mathcal{Y}$, typically $\mathbb{R}^{D_y}$ for some D_y. A statistical model considers the data to be realisations of random variables $(Y_1,\ \ldots,\ Y_n)$ defined on the same space. Denote $Y_1,\ \ldots,\ Y_n$ by Y, and $y_1,\ \ldots,\ y_n$ by y. The distribution of these random variables, which is specified by the model, has a probability density, $p_Y(y;\theta)$, where θ is the parameter of the model living in some space Θ.[3]

The density of the observations, seen as a function of the parameter, is called the *likelihood*, and is denoted by $\mathcal{L}_n$:

$$\mathcal{L}_n\ :\ \theta\in\Theta \mapsto p_Y(y;\theta). \tag{26}$$

In the frequentist approach, θ is an unknown fixed value and inference is performed based on the likelihood function. The standard estimator is the maximum likelihood estimator $\widehat{\theta}_n$, that is the parameter θ maximizing $\mathcal{L}_n(\theta)$ for the dataset $(y_1, \ldots,\ y_n)$. Note, because θ is not random, we write a semi-colon ";" in $p_Y(y;\theta)$ instead of a vertical bar "|" to emphasize that this is not a conditional distribution.

On the contrary, in the Bayesian approach, the unknown parameter is regarded as the random variable ϑ, and we assign a prior probability distribution to it, of density $p_\vartheta(\theta)$.[4] The distribution of Y given $\vartheta = \theta$ can now be interpreted as a conditional distribution and we thus denote it by $p_{Y|\vartheta}(y \mid \theta)$.

Bayesian inference relies on the posterior density,

[3] Written with respect to some dominating measure. This is often required when defining conditional probabilities, here the posterior probability.

[4] Written with respect to a dominating measure denoted $d\theta$, say a Lebesgue measure if $\Theta = \mathbb{R}^{D_\theta}$ for some D_θ)

$$p_{\vartheta|Y}(\theta \mid y) = \frac{p_{Y|\vartheta}(y \mid \theta) \cdot p_{\vartheta}(\theta)}{p_Y(y)} \tag{27}$$

obtained using Bayes formula, where,

$$p_Y(y) = \int_{\Theta} p_{Y|\vartheta}(y \mid \theta) \cdot p_{\vartheta}(\theta) d\theta \tag{28}$$

is the so-called *marginal likelihood* or *evidence*.

Based on the posterior distribution, we can compute various point estimates such as the posterior mean of ϑ

$$E(\vartheta \mid y) = \int_{\Theta} \theta \cdot p_{\vartheta|Y}(\theta \mid y) d\theta \tag{29}$$

or the posterior variance. We can also compute credible intervals, an interval C such that,

$$P(\vartheta \in C \mid y) = 1 - \alpha. \tag{30}$$

The posterior distribution can be used in the prediction of new observations. Assume we want to predict the next observation y_{n+1} given that we already have $y = (y_1, \ldots, y_n)$. Then the predictive density of Y_{n+1} having observed $Y = y$ is

$$p_{Y_{n+1}|Y}(y_{n+1} \mid y) = \int_{\Theta} p_{Y_{n+1}|Y,\vartheta}(y_{n+1} \mid y, \theta) \cdot p_{\vartheta \mid Y}(\theta \mid y) d\theta \tag{31}$$

The above predictive density considers the uncertainty about the parameter θ. By contrast, if we had first estimated the parameter, say by some $\widehat{\theta}$, and then plugged the value into a predictive distribution of Y_{n+1} using $\widehat{\theta}$, then we would not have taken parameter uncertainty into account.

Remark: The above notation is precise but heavy. It is standard in the Bayesian literature not to use subscripts to index the densities of interest and to use a simpler notation, i.e., the first five equations of this section will be written in most of the literature as

$$p(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{p(y)}, \tag{32}$$

$$p(y) = \int_{\Theta} p(y \mid \theta) \cdot p(\theta) d\theta. \tag{33}$$

$$E(\vartheta \mid y) = \int_{\Theta} \theta \cdot p\ (\theta \mid y) d\theta, \tag{34}$$

and

$$p\,(y_{n+1} \mid\ y) = \int_{\Theta} p\,(y_{n+1} \mid\ y,\ \theta) \cdot p(\theta \mid y) d\theta \tag{35}$$

This is imprecise as arguments of the densities should only be dummy variables, whereas in this notation they define the densities we consider, i.e., $p(\theta)$ means $p_{\vartheta}(\theta)$ and $p(y)$ means $p_Y(y)$, $p(\theta \mid y)$ means $p_{\vartheta|Y}(\theta \mid y)$, etc. However, this is standard and will be used here whenever it does not lead to any confusion. Note that another way to improve this imprecise notation consists in using different letters for the densities, i.e., $\mu(\theta) = p_{\vartheta}(\theta)$, $g(y \mid \theta) = p_{Y|\vartheta}\ (y \mid \theta)$, and $p(\theta \mid y) = p_{\vartheta|Y}(\theta \mid\ y)$, etc..

Gaussian Data

Let $Y = (Y_1,\ ...,\ Y_n)$ be IID random variables with $Y_i \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known and θ unknown. To perform Bayesian inference, we assign a prior on θ by introducing the random variable $\vartheta \sim \mathcal{N}(\mu,\ \kappa^2)$, then posterior, $p(\theta \mid y)$, is Normal with parameters, ν, and ω^2

$$p(\theta \mid y) = \mathcal{N}\left(\theta; \nu,\ \omega^2\right) \tag{36}$$

where

$$\omega^2 = \frac{\kappa^2 \sigma^2}{n\kappa^2 + \sigma^2} \tag{37}$$

and

$$\nu = \frac{\omega^2}{\kappa^2}\mu + \frac{n\omega^2}{\sigma^2}\overline{y} = \frac{\sigma^2}{n\kappa^2 + \sigma^2}\mu + \frac{n\kappa^2}{n\kappa^2 + \sigma^2}\overline{y} \tag{38}$$

so that $E(\vartheta \mid y) = \nu$, and $Var(\vartheta \mid y) = E\left(\vartheta^2 \mid y\right) - E(\vartheta \mid y)^2 = \omega^2$.

If we set $C = \left[\nu - \omega \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),\ \nu + \omega \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$, where$\Phi^{-1}$ denotes the inverse of the cumulative distribution function of the Normal distribution, then the predictive density is also Normal

$$p\left(y_{n+1} \mid y\right) = \int_{\Theta} p\left(y_{n+1} \mid y,\ \theta\right) \cdot p(\theta \mid y) d\theta = \mathcal{N}\left(y_{n+1}; \nu,\ \omega^2 + \sigma^2\right) \tag{39}$$

In this simple example, all the calculations can be done analytically. This is because the Normal prior is "conjugate" with the Normal model with unknown mean and known variance, i.e., the posterior distribution is in the same family of distributions as the prior distribution (here, the family of Normal distributions). In general, the calculation of posterior quantities cannot be performed exactly. Indeed, one might want to use another prior distribution than the conjugate one, or the model might not admit any conjugate prior distribution. [ADD PROOF HERE]

Logistic Regression

Let $(x_i,\ Y_i) \in \mathbb{R}^D \times \{0, 1\}$ where $x_i \in \mathbb{R}^D$ is a given covariate and we assume that the data are independent with

$$P\left(Y_i = y_i \mid \theta\right) = \frac{\exp\left(-y_i x_i^T \theta\right)}{1 + \exp\left(-x_i^T \theta\right)} \tag{40}$$

To perform Bayesian inference, we assign a prior $p(\theta)$ on θ and Bayesian inference relies on

$$p(\theta \mid y_1,\ \ldots,\ y_n) = \frac{p(\theta) \prod_{i=1}^{n} P\left(Y_i = y_i \mid \theta\right)}{P(y_1,\ \ldots,\ y_n)} \tag{41}$$

which is not a standard distribution if $p(\theta)$ is chosen to be a Normal distribution. There exists a conjugate prior distribution for θ, but it is not standard itself. The denominator $P(y_1,\ ...,\ y_n)$ cannot be computed analytically.

In general, statistical models and the associated prior probability distributions should be chosen to represent a phenomenon and its uncertainties, and thus should not be chosen on the grounds of purely computational reasons, such as "to make the calculations easier". Thus, in many situations we will encounter posterior distributions such that we cannot analytically compute the integrals listed above, e.g., the posterior mean and so

on. Going back to the problem of computing integrals, in statistics the integrals will often be written

$$I = \int_{\Theta} \varphi(x)\ \pi(x)dx, \tag{42}$$

where π is a probability density function, φ is a "test" function and Θ a sample space; for instance, with $\varphi : \theta \mapsto \theta$ and $\pi(\theta) = p(\theta \mid y)$, the integral corresponds to the posterior mean. In the context of approximating I, the distribution π is often called the "target distribution". The integral can also be written

$$I = E_{\pi}\left(\varphi(\vartheta)\right) \tag{43}$$

where ϑ follows the distribution π. Monte Carlo methods generally consist in replacing such expectations by empirical averages.

Appendix

Maps

The map arrow-notation defines the rule of a function inline, without requiring a name to be given to the function. For example, $x \mapsto x + 1$ is the function that takes a real number as input, and outputs that number plus 1. An (input) domain and (output) codomain of $\mathbb{R}$ is implied.

The domain and codomain can also be explicitly stated, for example:

$$f : \mathbb{Z} \to \mathbb{Z} \tag{44}$$

$$x \mapsto x^2. \tag{45}$$

This defines a function f from the integers to the integers that returns the square of its input.