# Importance Sampling with Applications to Bayesian Logistic Regression

Ádám Jung

Supervisor: Balázs Csanád Csáji

December 2023

## 1 Introduction

In this document we will provide an introduction to *importance sampling*, that is a sampling technique designed to estimate expectations of rare events. Afterwards we will provide a brief introduction to Bayesian Logistic Regression (BLR), and present a case study on the usage of importance sampling for the parameters of a BLR model.

## 2 Importance Sampling

Lets consider a random variable $X$ and a given function $f$, for which we would like to estimate the expectation $\mu := \mathbb{E}[f(X)]$, using sampling. In the case, when $f$ takes its nonzero values mainly from low probability regions of $X$, the mean from i.i.d. samples $\hat{\mu} = 1/n \sum_{i=1}^{n} f(x_i)$ is likely to over-sample regions for which $f$ is always zero, and take only a few samples from the *important* region, where $f$ takes its values defining $\mathbb{E}[f(X)]$.

Therefore the concept of importance sampling[4] is to choose an alternative distribution $q$ to generate the $x_1, \ldots, x_n$ samples from, and use a weighted average to estimate $\mu$:

$$\hat{\mu}_q = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \frac{p(x_i)}{q(x_i)}, \tag{1}$$

where $p$ is the probability density function (pdf) of $X$, and $q$ is the pdf of the sampling distribution. Note that if $f(x)p(x) \neq 0 \implies q(x) > 0$ holds, then $\hat{\mu}_q$ is an unbiased estimate for $\mu$, since

$$\mathbb{E}_q[\hat{\mu}_q] = \mathbb{E}_q\left[f(X)\frac{p(X)}{q(X)}\right] = \int_Q f(x)\frac{p(x)}{q(x)} q(x)dx = \mu,$$

with the notations $Q = supp(q)$ and $\mathbb{E}_q$ for expectation under $q$.

The goal of importance sampling is to reduce the variance $\mathbb{D}^2_q[\hat{\mu}_q]$ of the estimation, that is

$$\mathbb{E}_q[(\hat{\mu}_q - \mu)^2] = \frac{1}{n}\left[\int_Q \frac{f^2(x)p^2(x)}{q(x)} dx - \mu^2\right]. \tag{2}$$

We can gain insights about what are the good sampling distributions - for which this variance is small - by analyzing the integrand on the rhs of Eq (2). First lets suppose that $f \geq 0$. In this case if we choose $q \propto fp$, we can conclude that the the normalizing constant of $q$ is $1/\mu$. This is unfortunate in the sense, that $\mu$ is what we are looking for in the first place, however substituting $fp/\mu$ into Eq. (2) quite surprisingly yields zero variance. Using this $q$ clearly doesn't solve our original problem, but it shades light on the fact that choosing an appropriate $q$ can reduce the required sample size in Eq. (1) vastly.

In the general case, when $f \geq 0$ not necessarily holds, the optimal sampling distribution[2] is

$$q^*(x) = \frac{|f(x)|p(x)}{\mathbb{E}[|f(X)|]}, \tag{3}$$

which can be shown using Jensen's inequality:

$$\begin{aligned} \mu^2 + n\mathbb{D}^2_{q^*}[\hat{\mu}_{q^*}] &= \int_Q \frac{f^2(x)p^2(x)}{|f(x)|p(x)/\mathbb{E}[|f(X)|]} dx \qquad (4)\\ &= \mathbb{E}^2[|f(X)|] = \mathbb{E}^2_q[|f(X)|p(X)/q(X)] \\ &\leq \mathbb{E}_q[f^2(X)p^2(X)/q^2(X)] = \mu^2 + n\mathbb{D}^2_q[\hat{\mu}_q], \end{aligned}$$

so $\mathbb{D}^2_{q^*}[\hat{\mu}_{q^*}] \leq \mathbb{D}^2_q[\hat{\mu}_q]$ holds. It follows that the value of the optimal variance is $\mathbb{D}^2_{q^*}[\hat{\mu}_{q^*}] = 1/n\left[\mathbb{E}^2[|f(X)|] - \mu^2\right]$.

It is useful to keep in mind that a miss-chosen $q$ could lead to infinite variances of $\hat{\mu}_q$, while $\mathbb{D}^2[\hat{\mu}]$ could still be finite. For example, if $q$ puts small weights on regions, where $p(x)f(x)$ is large, then the integrand in Eq. (2) could diverge to $\infty$.

A possible approach for choosing an appropriate $q$ in practice will be presented in the following case study, while in general, $q$ could be found using domain knowledge and educated guessing.

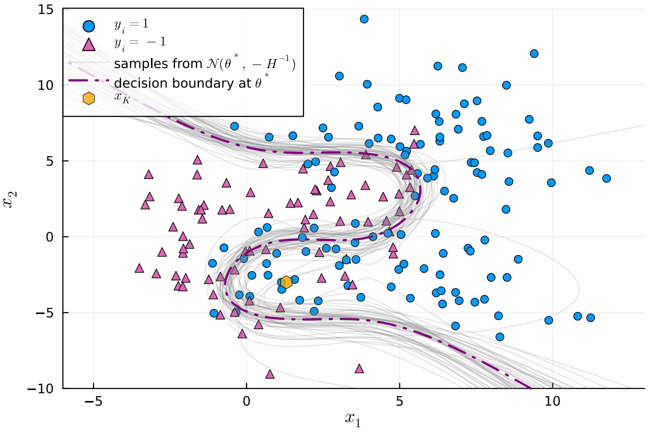

Figure 1: Decision boundaries sampled from the approximation of the posterior.

## 3 Bayesian Logistic Regression

Let us consider Logistic Regression[1] in a Bayesian setting, used for binary classification. Logistic Regression is a special case of a more general model class, the Generalized Linear Models.[3,5]

Let the observation pairs be $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$. The $x_i, i \in [n]$ input features are interpreted as constants, and the corresponding response variables assumed to follow the parametric model

$$\mathbb{P}(Y_i = 1 \mid \theta) = \sigma(\theta^T \Phi(x_i)), \tag{5}$$

where $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid function, and $\Phi : \mathbb{R}^d \to \mathbb{R}^l$ is a basis function, used to enable nonlinear decision boundaries.

In Bayesian modeling the model parameter $\theta$ is a random variable itself, and has a prior distribution $\pi(\theta)$.

Let $\mathcal{Y}$ denote $(Y_1, \ldots, Y_n)^T$. Then the posterior distribution of $\theta$ is proportional to

$$\pi(\theta \mid \mathcal{Y}) = \pi(\theta) \prod_{i=1}^{n} \sigma(Y_i \theta^T \Phi(x_i)). \tag{6}$$

To find the maximum likelihood estimate $\theta^* = \arg\max_\theta \pi(\theta \mid \mathcal{Y} = (y_1, \ldots, y_n)^T)$, one can use numerical methods like the Newton method or Stochastic Gradient descent. Then if we approximate the log posterior with its second order Taylor expansion around $\theta^*$

$$\log(\pi(\theta \mid \mathcal{Y})) \approx \log(\pi(\theta^* \mid \mathcal{Y})) + \frac{1}{2}(\theta - \theta^*)^T H (\theta - \theta^*), \tag{7}$$

it can be interpreted as $\theta \mid \mathcal{Y} \sim \mathcal{N}(\theta^*, -H^{-1})$. Note that in Eq. (7) the first gradient should vanish at $\theta^*$ under regularity assumptions and $H$ denotes the Hessian of $\pi(\theta \mid \mathcal{Y} = (y_1, \ldots, y_n)^T)$ at $\theta^*$.

# 4 Case study

Suppose we would like to compute $\mathbb{E}[f(\theta) \mid \mathcal{Y}]$ for some $f$, which takes its nonzero values at unlikely parameter settings of $\theta \mid \mathcal{Y}$. When the important region of $f$ can't be sufficently sampled from $\mathcal{N}(\theta^*, -H^{-1})$, it is a general approach to use a multivariate $t$ distribution $t(\nu, \theta^*, -H^{-1})$ as $q$ to generate samples from the tail of $\theta \mid \mathcal{Y}$ more often. It can be fine tuned with $\nu$ how spread out the samples are.

For a concrete example lets consider a fixed point $x_K \in \mathbb{R}^2$ and let $f(\theta) = \mathbb{I}(\sigma(\theta^T \Phi(x_K)) < 1/2)$. Then $\mathbb{E}[f(\theta)]$ is the probability of classifying $x_K$ to the negative class using a hard decision boundary distributed as $\pi(\theta \mid \mathcal{Y})$.

Samples from the Gaussian approximation of the posterior distribution can be seen in Fig. (1), and since $x_K$ lies quite far from the expected boundary defined by $\theta^*$, we can conclude that classifying $x_K$ to the negative class is indeed a low probability event.

The basis function used in the experiment was $\Phi(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1^3, x_2^3)$, and the prior distribution was a result of an earlier experiment classifying a linearly separable dataset with the same model. The slope of the decision boundaries from the prior distribution are somewhat similar to the trend of the actual data, but it lacks the separation of the "S shaped" middle part of it (see Fig. 2).

The true expectation $\mu$ was approximated from $100,000,000$ samples, generated from $\mathcal{N}(\theta^*, -H^{-1})$ and is denoted by $\tilde{\mu}$. Fig. (3) shows empirical variances of $\mathbb{E}[(\hat{\mu}_p - \tilde{\mu})^2]$ and $\mathbb{E}[(\hat{\mu}_q - \tilde{\mu})^2]$ for different sample sizes, where $\hat{\mu}_p$ was sampled from $\mathcal{N}(\theta^*, -H^{-1})$ as well. It can be concluded that sampling from $q$ did improved on the variances of the estimates, and the improvement is slightly more significant at low sample sizes.

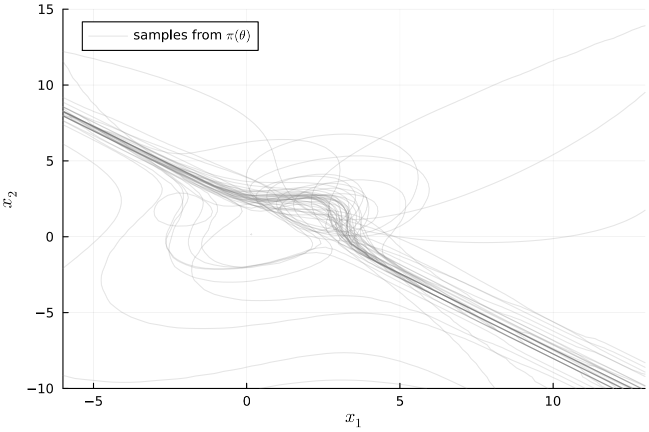

Figure 2: Decision boundaries, sampled form the prior distribution $\pi(\theta)$.

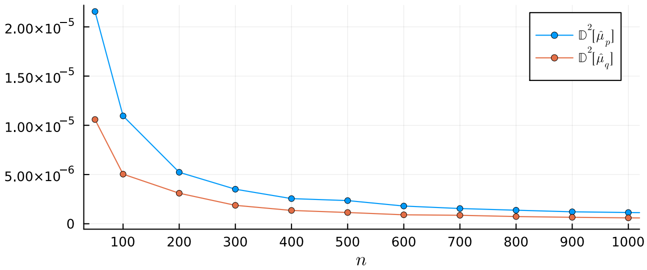

Figure 3: Empirical variances computed from 3000 repeated estimations at each sample size $n$.

# References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. 02 2009.

[2] H. Kahn and A. W. Marshall. Methods of Reducing Sample Size in Monte Carlo Computations. *Operations Research*, 1(5):263–278, November 1953.

[3] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[4] Art B. Owen. *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`, 2013.

[5] Andrew Ng Tengyu Ma. Cs229 lecture notes, stanford university. `https://cs229.stanford.edu/notes2022fall/main_notes.pdf`.