
A study on approximation techniques for Bayesian Logistic Regression

By

MAURO CAMARA ESCUDERO



Department of Mathematics
UNIVERSITY OF SOUTHAMPTON

A dissertation submitted to the University of Southampton in
accordance with the requirements of the degree of BACHELOR
OF MATHEMATICS in the Faculty of Mathematics.

APRIL 2019

Word count: ten thousand and four

ABSTRACT

Write something here.

DEDICATION AND ACKNOWLEDGEMENTS

Thanks to someone.

TABLE OF CONTENTS

	Page
List of Tables	v
List of Figures	vi
1 Introduction	1
1.1 Bayesian Setup and Shortcomings	2
1.2 Overview of Approximate Inference	3
1.2.1 Stochastic Approximations	3
1.2.2 Deterministic Approximations	4
1.3 Bayesian Logistic Regression	4
2 Stochastic Approximations	6
2.1 Introduction to Sampling	6
2.2 Monte Carlo approximation	7
2.2.1 Uniform Sampling	8
2.2.2 Importance Sampling	9
2.2.3 Rejection Sampling	10
2.3 Markov Chains Monte Carlo	13
3 Deterministic Approximations	16
3.1 Laplace Approximation	16
3.1.1 A General Framework	16
3.1.2 Application to Bayesian Logistic Regression	17
3.2 Variational Methods	18
3.2.1 A General Framework	18
3.2.2 Local Variational Methods	20
4 Results	25
4.1 No Explanatory variables and No prior, only likelihood	25
4.1.1 Laplace Approximation	25
4.1.2 Variational Approximation	26

4.2	No explanatory variables, prior and posterior	27
4.2.1	Laplace with prior	28
4.2.2	Optimization of variational parameters	28
A	Appendix A - Statistics Background	29
A.1	Exponential Family of Distributions	29
A.1.1	Bernoulli Distribution	29
A.2	Generalized Linear Models	30
A.2.1	Explanatory Variables	30
A.2.2	Response Variables	31
A.3	Multivariate Normal Distribution	31
A.4	Information Theory	32
A.4.1	Non-negativity Property	32
A.5	Details of Local Variational Methods	33
A.5.1	Optimization of Parameters: EM Algorithm	33
B	Appendix B - Computing	38
B.1	some	38
	Bibliography	39

LIST OF TABLES

TABLE	Page
-------	------

LIST OF FIGURES

FIGURE	Page
2.1 Simulation of Probability Integral Transform	7
2.2 Estimation of a Normalization Constant via Uniform Sampling	8
2.3 Example of a proposal distribution and its comparison function	11
3.1 Example of Jaakkola and Jordan's lower bound	21
A.1 Illustration of Symmetry for variational bound	36

INTRODUCTION

Historically, the interpretation and definition of probability has engaged many philosophers, scientists and mathematicians. In the early days, the Frequentist interpretation was most prominent because it was in line with one of the core principles of the scientific method: reproducibility. In this interpretation, probabilities have meaning only if defined in the context of reproducible and repeatable experiments. They represent the limiting behavior of the ratio of the number of success over the number of attempts, usually called *trials*, as the number of trials goes to infinity. Although compatible with the working routines of many scientists, this interpretation has two major drawbacks: it does not allow probabilities to be defined for one-off, non-reproducible, experiments, and it often leads to seemingly unintuitive interpretations of the results.

The Bayesian interpretation promises to deliver a more intuitive framework that works also when no long-run frequency is involved. It does so by interpreting probabilities as a way of quantifying the uncertainty around a *statement*, for instance, Bayesian statisticians would be able to give meaning to the probability that the sun will rise tomorrow, whereas this statement would make no sense to a Frequentist statistician. The pitfall of the Bayesian framework is that while it allows greater flexibility and incorporation of subjective knowledge, it often leads to integrals that have no closed-form and that are numerically intractable, in practice.

In this introductory chapter we will see how these intractable expressions arise and an overview of the three main methods used to tackle the issue. Chapter 2 will explore sampling approximations, while Chapter 3 focuses on deterministic approximations instead. Finally, Chapter 4 presents a series of computational results comparing the three methods in the context of Bayesian Logistic Regression.

1.1 Bayesian Setup and Shortcomings

Throughout this thesis I will assume that we are interested in understanding some characteristics of a **population**, from which we have observed data $\mathbf{x} = (x_1, \dots, x_n)^\top$ as being the realization of a **sample** of n independent random variables $\mathbf{X} = (X_1, \dots, X_n)^\top$. We denote the *unknown* characteristics of the population by the vector of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ taking values in the p -dimensional **parameter space** Θ .

Often we have some *a priori* knowledge about the values of the parameters which we can express via a probability distribution $p(\boldsymbol{\theta})$ called the **prior distribution**. It communicates our subjective belief around the distribution of values of the parameters *before* we see the data. Observing the data will likely provide us with additional information that was not available before, it is therefore reasonable to imagine that our prior beliefs might need to be updated in some way. We call **posterior distribution** our *revised* beliefs about the characteristic of the population after seeing the data, and we denote it by $p(\boldsymbol{\theta} | \mathbf{x})$.

Bayes Theorem provides a direct and intuitive way of updating our prior beliefs:

$$(1.1) \quad p(\boldsymbol{\theta} | \mathbf{x}) = \frac{p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{x})}$$

Here $p(\mathbf{x} | \boldsymbol{\theta})$ is called **likelihood function** and it is sometimes also denoted by $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$. For a vector of values for $\boldsymbol{\theta}$, it gives the *likelihood* that such values represent the actual characteristic of the population, given that we have observed data \mathbf{x} . The denominator $p(\mathbf{x})$ is called **model evidence** and it acts as a *normalization constant*, making sure that the posterior distribution integrates to 1 when \mathbf{X} are continuous random variables, or sums up to 1 when they are discrete.

In other words, Bayes Theorem tells us that our uncertainty after seeing the data is proportional to the information about $\boldsymbol{\theta}$ contained in the data, and our prior knowledge about $\boldsymbol{\theta}$:

$$(1.2) \quad \underbrace{p(\boldsymbol{\theta} | \mathbf{x})}_{\text{posterior}} \propto \underbrace{p(\mathbf{x} | \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\text{prior}}$$

In practice, many of the problems with the Bayesian approach revolve around intractable integrals. Below we give an overview of some of the most common issues arising in Bayesian statistics [1, 9, 13]:

1. *Inference*: Obtaining the full posterior distribution requires us to compute the model evidence.

$$(1.3) \quad p(\mathbf{x}) = \int_{\Theta} p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

2. *Prediction*: Determining the distribution of new data point.

$$(1.4) \quad p(\mathbf{x}_{\text{new}} | \mathbf{x}) = \int_{\Theta} p(\mathbf{x}_{\text{new}} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}$$

3. *Posterior Features*: Several features of the posterior distribution are also of interest for Bayesian statisticians, such as quantiles and posterior moments, e.g.

$$(1.5) \quad \mu_n := \int (\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} | \mathbf{x}])^n p(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}$$

4. *Decision Theory*: Bayes estimator $\hat{\boldsymbol{\theta}}(\mathbf{X})$ is chosen to minimize the posterior expected loss

$$(1.6) \quad \rho(\hat{\boldsymbol{\theta}}(\mathbf{X}) | \mathbf{X} = \mathbf{x}) := \mathbb{E}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{X})) | \mathbf{X} = \mathbf{x}] = \int_{\Theta} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) p(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}$$

where $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is the loss incurred in estimating $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$, e.g. $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) := (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2$.

5. *Model Selection*: Two models \mathcal{M}_1 and \mathcal{M}_2 can be compared using Bayes Risk:

$$(1.7) \quad B(\mathbf{x}) := \frac{\int_{\Theta_1} p_1(\mathbf{x} | \boldsymbol{\theta}_1) p_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int_{\Theta_2} p_2(\mathbf{x} | \boldsymbol{\theta}_2) p_2(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}$$

6. *Credible Regions*: Credible regions with Highest Posterior Density given a chosen confidence level γ are found by

$$(1.8) \quad C(\mathbf{x}) := \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta} | \mathbf{x}) \geq k\} \quad \text{where} \quad \mathbb{P}(p(\boldsymbol{\theta} | \mathbf{x}) \geq k | \mathbf{x}) = \gamma$$

Notice equations (1.3)-(1.8) require us to compute integrals that can be interpreted as the expectation of some function $f(\boldsymbol{\theta})$ with respect to the posterior distribution $p(\boldsymbol{\theta} | \mathbf{x})$.

$$(1.9) \quad \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{x})}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}$$

For instance, one can choose $f(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} | \mathbf{x}])^n$ and $f(\boldsymbol{\theta}) = p(\mathbf{x})$ in (1.5) and (1.3) respectively.

In practice these integrals often have no closed-form solution or are numerically intractable. Developing theory and methodology to overcome these issues is the central focus of an area of Bayesian statistics called *Approximate Inference*.

1.2 Overview of Approximate Inference

1.2.1 Stochastic Approximations

Their drawback is that they require a large computational effort that does not scale well with dimensionality of the parameter space Θ , in other words, they suffer from the *curse of dimensionality*. On the other hand, they have the nice property of asymptotic convergence.

- *Monte Carlo methods*: Approximates expectations by averaging *independent* samples drawn directly¹ from the desired distribution.
- *Markov Chain Monte Carlo methods*: Runs a Markov Chain producing *dependent* samples that asymptotically mimic samples from the desired distribution. These can then be used to approximate any of the integral problems (1.3)-(1.8) using Monte Carlo approximations.

¹As we will see, sometimes we draw from a proposal distribution, but then correct such samples to make them samples from our desired distribution, see Section 2.2.2.

1.2.2 Deterministic Approximations

- *Laplace Approximation*: Approximates $p(\boldsymbol{\theta} | \mathbf{x})$ with a multivariate normal distribution centered at the mode of the posterior and having the inverse of the observed information matrix evaluated at the mode as its variance-covariance matrix.
- *Variational methods*: Approximates $p(\boldsymbol{\theta} | \mathbf{x})$ with the distribution $q(\boldsymbol{\theta}) \in \mathcal{D}$ that is closest to $p(\boldsymbol{\theta} | \mathbf{x})$ according to some objective function, among all the distributions in the family \mathcal{D} .

1.3 Bayesian Logistic Regression

To conclude this chapter, we will consider a key example of Bayesian modelling and inference: Bayesian Logistic Regression.

Consider a Generalized Linear Model, as described in Appendix A, where each response variable follows independently² and identically a Bernoulli distribution

$$p(y_i | \mathbf{X}_i = \mathbf{x}_i; \pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad \text{for } y_i \in \{0, 1\}$$

and where the link function is given by the **logit** function $g : [0, 1] \rightarrow \mathbb{R}$ given below.

$$(1.10) \quad g(\mathbb{E}[y_i | \mathbf{X}_i = \mathbf{x}_i]) = g(\pi_i) = \ln \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

Equivalently, one can use (1.10) to write the success probability π_i in terms of the explanatory variables and of our parameters of interest $\boldsymbol{\beta}$:

$$(1.11) \quad \pi_i(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$$

Leveraging the independence assumption we can write the joint distribution of the response variables (i.e. the likelihood) as a product of marginal probability mass functions

$$(1.12) \quad p(\mathbf{y} | \mathbf{X}_D^{\text{RV}} = \mathbf{X}_D^{\text{obs}}; \boldsymbol{\beta}) = \prod_{i=1}^n p(y_i; \mathbf{x}_i, \boldsymbol{\beta}) = \prod_{i=1}^n \pi_i(\boldsymbol{\beta})^{y_i} (1 - \pi_i(\boldsymbol{\beta}))^{1-y_i}$$

For ease of notation we will simply refer to the above expression as $p(\mathbf{y} | \boldsymbol{\beta})$ by neglecting the conditioning on the explanatory variables, which will be considered fixed.³

²The term "independent" here means *statistically* independent in the traditional sense. When we say "dependent variable" we mean that we expect \mathbf{y} to change due to a change in \mathbf{X}_D following the pattern of dependence, which includes random and structural components.

³The notation $p(\mathbf{y}; \boldsymbol{\beta})$ means that $\boldsymbol{\beta}$ are parameters and if we consider $p(\mathbf{y}; \boldsymbol{\beta})$ as a function of them, then it is called *likelihood*. On the other hand $p(\mathbf{y} | \boldsymbol{\pi})$ indicates that we are treating $\boldsymbol{\pi}$ as a random variable. This is the preferred way in Bayesian Inference.

Having found the likelihood, we just need to choose a prior that reflects our beliefs about the parameters $\boldsymbol{\beta}$. For simplicity, we will choose a multivariate normal distribution $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{p \times p}$, that is

$$(1.13) \quad p(\boldsymbol{\beta}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right)$$

Plugging (1.12) and (1.13) into (1.2) we obtain an expression proportional to the posterior distribution

$$p(\boldsymbol{\beta} | \mathbf{y}) \propto (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right) \prod_{i=1}^n \pi_i(\boldsymbol{\beta})^{y_i} (1 - \pi_i(\boldsymbol{\beta}))^{1-y_i}$$

We can simplify this expression by neglecting all those terms that do not depend on $\boldsymbol{\beta}$ and by taking the natural logarithm both sides; this will also have the advantage to contain potential overflows in the Python implementation shown in Chapter 4.

$$(1.14) \quad \ln(p(\boldsymbol{\beta} | \mathbf{y})) \propto -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \sum_{i=1}^n y_i \ln(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \ln(1 - \pi_i(\boldsymbol{\beta}))$$

This expression cannot be integrated analytically and it would be computationally intractable even using standard quadrature techniques.

STOCHASTIC APPROXIMATIONS

Markov Chain Monte Carlo (MCMC) are considered the gold standard for obtaining samples from an intractable posterior distribution which can be used to obtain summary statistics such as the arithmetic mean, or can be used to obtain non-parametric estimates. In this chapter we will first explore methods that generate independent samples, but that are inefficient in high dimensions, and then we will construct the theory behind MCMC techniques that produce dependent samples but are much more robust in high dimensional settings.

2.1 Introduction to Sampling

Successful application of the theory of Monte Carlo and Markov Chain Monte Carlo methods (MCMC) relies on the assumption that we have a way of sampling a uniform distribution or, in other words, that we are able to generate random numbers within a finite interval. Due to the deterministic nature of our computers, we cannot generate *truly* random numbers, however there exists several methods that produce m numbers within $[0, 1]$ such that, when compared through specific tests¹, they mimic the behavior of $V_1, \dots, V_m \sim \mathcal{U}(0, 1)$. Such numbers are called **pseudo random numbers** and, throughout this chapter, we will assume the existence of a generator of pseudo random numbers and consider its output as truly random.

Very often we are interested in obtaining samples from distributions other than the Uniform, for instance, the posterior distribution described in Section 1.1. Luckily, one can show [2] that every *continuous* random variable is a transformation of a uniform random variable.

¹One can use the Kolmogorov-Smirnov test or time series analysis methods.

Theorem 2.1 (Probability Integral Transform). *Let X be a continuous random variable with cumulative distribution function F . Let Y be a uniform random variable in the interval $[0, 1]$. Then $F^{-1}(Y)$ has cdf F . Equivalently, let X be defined as above. Then $U := F(X)$ follows a uniform distribution in the interval $[0, 1]$.*

For instance, suppose that our random variable X follows an exponential distribution with rate parameter λ

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad , \lambda > 0$$

then its cdf will be $F(x) = 1 - e^{-\lambda x}$, which is strictly increasing. We can therefore write down its inverse as $F^{-1}(y) = -\lambda \ln(1 - y)$. According to the Probability Integral Transform theorem, if $Y \sim \mathcal{U}(0, 1)$ then $F^{-1}(Y)$ should follow an exponential distribution with rate parameter λ . Figure 2.1 shows this in the case of $\lambda = 1$:

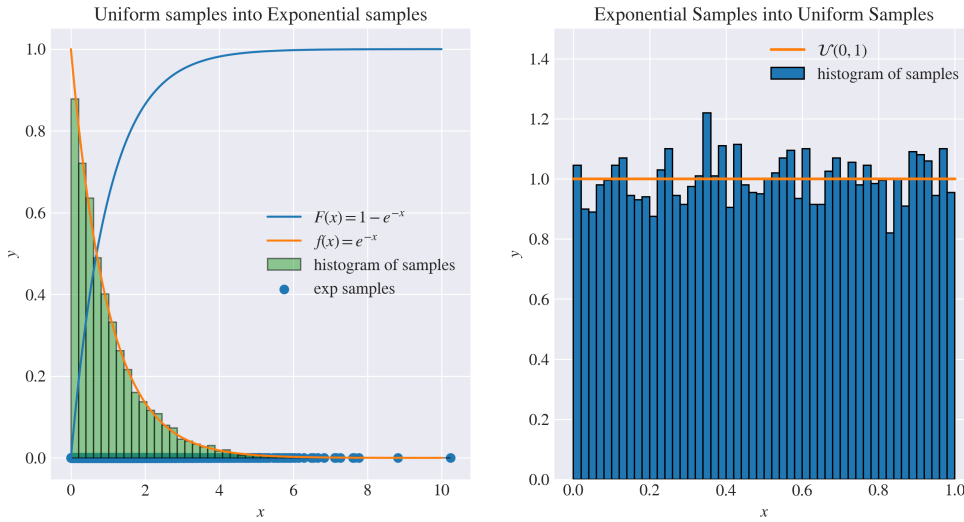


Figure 2.1: Simulation of Probability Integral Transform

2.2 Monte Carlo approximation

Suppose that we have obtained N independent samples $\theta_1, \dots, \theta_N$ from our target probability distribution $p(\theta | \mathbf{x})$. Expectation (1.9) can then be thought of the *population* mean of $f(\theta)$, and it can be approximate by a sample average:

$$(2.1) \quad \mathbb{E}_{p(\theta|\mathbf{x})} [f(\theta)] \approx \bar{f}(\theta) := \frac{1}{N} \sum_{i=1}^N f(\theta_i)$$

Then, by the strong law of large numbers we know that $\bar{f}(\theta) \xrightarrow{\text{a.s.}} \mathbb{E}_{p(\theta|\mathbf{x})} [f(\theta)]$ as $N \rightarrow \infty$.

2.2.1 Uniform Sampling

In equation (2.1) the samples $\theta_1, \dots, \theta_N$ are assumed to be drawn directly from $p(\theta | \mathbf{x})$, which guarantees that, if N is large enough, their (normalized) histogram will approximate the probability density. In other words, samples are drawn proportionally to $p(\theta | \mathbf{x})$.

A different way of obtaining the same estimate as in (2.1) is to draw $\theta_1, \dots, \theta_N$ *uniformly* in Θ , evaluate the density $p(\theta | \mathbf{x})$ at those samples $p(\theta_1 | \mathbf{x}), \dots, p(\theta_N | \mathbf{x})$ and use them as weights in a weighted average of $f(\theta_1), \dots, f(\theta_N)$:

$$(2.2) \quad \bar{f}(\theta) := \sum_{i=1}^N f(\theta_i) p(\theta_i | \mathbf{x})$$

For most densities, however, the majority of the probability distribution mass will be concentrated in a relatively small area of Θ called **typical set** [11]. It follows that if the samples $\theta_1, \dots, \theta_N$ are drawn uniformly in Θ , then most of them will be drawn from regions where the mass of the probability distribution is practically zero, which means that a very large number of samples will have a negligible contribution to the sum (2.2).

As an example, suppose that we are interested in solving problem (1.3) the one-dimensional case of a standard normal distribution. Namely, we know that our posterior distribution is proportional to a standard normal distribution

$$p(\theta | x) \propto e^{-\frac{x^2}{2}} \quad \text{or equivalently} \quad p(x | \theta) p(\theta) = e^{-\frac{x^2}{2}}$$

and so we can evaluate $p(\theta | x)$ for any $\theta \in \mathbb{R}$ up to a proportionality constant (which, in this example, is $\sqrt{2\pi} \approx 2.51$).

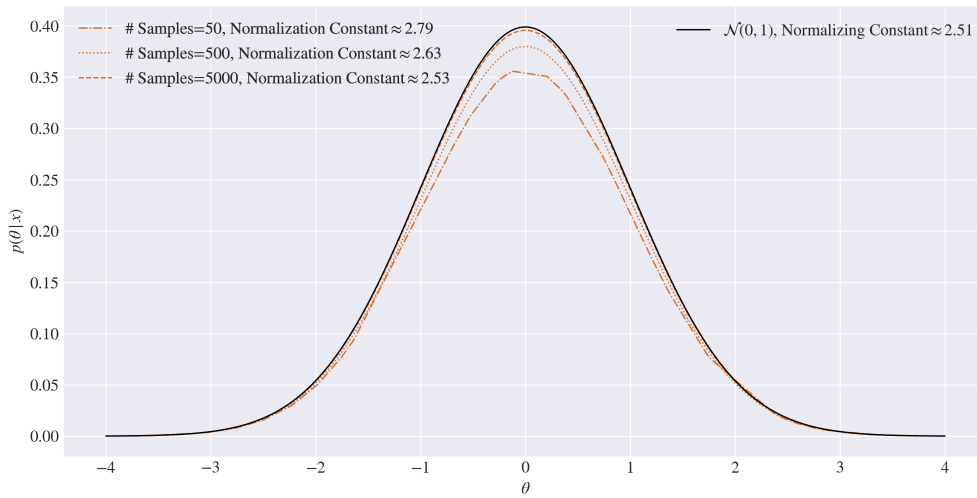


Figure 2.2: Estimation of a Normalization Constant via Uniform Sampling

Figure 2.2 shows how, even by restricting the sample space to $[-4, 4]$ rather than the whole of \mathbb{R} , uniform sampling requires a large number of samples to obtain a decent approximation for the normalization constant. In practice, we might not be able to restrict ourselves to such a narrow region of Θ and as the dimensionality grows we can expect the number of samples needed to increase drastically.

2.2.2 Importance Sampling

Rather than drawing our samples $\theta_1, \dots, \theta_N$ uniformly, we would like to have a mechanism to draw them from regions of Θ where we are confident that the density $p(\theta | \mathbf{x})$ has a non-negligible mass, to make the algorithm more efficient. One way in which we can do this, is to draw $\theta_1, \dots, \theta_N$ from a non-uniform distribution $q(\theta)$ that has a support Θ_q "similar" to Θ and then adjust such samples based on how different $q(\theta_i)$ is to $p(\theta_i | \mathbf{x})$. Consequently, we aim to find a **proposal distribution** $q(\theta)$ that is:

- "Biased" towards areas of Θ where $p(\theta | \mathbf{x})$ has more mass, and $q(\theta) > 0$ whenever $p(\theta | \mathbf{x}) > 0$.
- Easy to sample from, and such that we can compute $q(\theta)$ for every $\theta \in \Theta$ up to a proportionality constant, i.e. we can compute $\tilde{q}(\theta) = \mathcal{C}_q q(\theta)$ for $\mathcal{C}_q := \int_{\Theta} \tilde{q}(\theta) d\theta$.

In the following we show that even in the case in which we can only evaluate $p(\theta | \mathbf{x})$ up to a proportionality constant², the mechanism described above correctly approximates equation (1.9).

$$\begin{aligned}
 \mathbb{E}_{p(\theta|\mathbf{x})} [f(\theta)] &= \int_{\Theta} p(\theta | \mathbf{x}) f(\theta) d\theta && \text{LOTUS}^3 \\
 &= \int_{\Theta} q(\theta) \frac{p(\theta | \mathbf{x})}{q(\theta)} f(\theta) d\theta && q(\theta) \text{ is the proposal. Assume } q(\theta) > 0 \text{ for } \theta \in \Theta \\
 &= \frac{\mathcal{C}_q}{p(\mathbf{x})} \int_{\Theta} q(\theta) \frac{p(\mathbf{x} | \theta) p(\theta)}{\tilde{q}(\theta)} f(\theta) d\theta \\
 &= \frac{\mathcal{C}_q}{p(\mathbf{x})} \mathbb{E}_{q(\theta)} \left[\frac{p(\mathbf{x} | \theta) p(\theta)}{\tilde{q}(\theta)} f(\theta) \right] && \text{LOTUS} \\
 (2.3) \quad &\approx \frac{\mathcal{C}_q}{p(\mathbf{x})} \frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{x} | \theta_i) p(\theta_i)}{\tilde{q}(\theta_i)} f(\theta_i) && \text{where } \theta_i \sim q(\theta)
 \end{aligned}$$

In the last step we have used Monte Carlo approximation (2.1).

Our job is not done because the ratio of normalization constants is unknown. Luckily, we can estimate

²The proportionality constant is given by (1.3) as $p(\mathbf{x}) = \int_{\Theta} p(\mathbf{x} | \theta) p(\theta) d\theta$.

³Law of the Unconscious Statistician.

the inverse of such ratio by following the work of Bishop [3]

$$\begin{aligned}
 \frac{p(\mathbf{x})}{\mathcal{Z}_q} &= \frac{1}{\mathcal{Z}_q} \int_{\Theta} p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
 &= \frac{1}{\mathcal{Z}_q} \int_{\Theta} \frac{p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
 &= \int_{\Theta} \frac{p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\tilde{q}(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
 &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\frac{p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\tilde{q}(\boldsymbol{\theta})} \right] && \text{LOTUS} \\
 (2.4) \quad &\approx \frac{1}{N} \sum_{i=1}^N \frac{p(\mathbf{x} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)}{\tilde{q}(\boldsymbol{\theta}_i)} && \text{Monte Carlo approximation}
 \end{aligned}$$

Now plugging (2.4) into (2.3) having defined

$$(2.5) \quad \tilde{r}_i := \frac{p(\mathbf{x} | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)}{\tilde{q}(\boldsymbol{\theta}_i)} \quad \forall i \in \{1, \dots, N\}$$

for convenience of notation, we obtain the following estimate:

$$\begin{aligned}
 \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{x})} [f(\boldsymbol{\theta})] &\approx \left(\frac{1}{N} \sum_{j=1}^N \tilde{r}_j \right)^{-1} \frac{1}{N} \sum_{i=1}^N \tilde{r}_i f(\boldsymbol{\theta}_i) \\
 &= \sum_{i=1}^N \frac{\tilde{r}_i}{\frac{1}{N} \sum_{j=1}^N \tilde{r}_j} f(\boldsymbol{\theta}_i) \\
 (2.6) \quad &:= \sum_{i=1}^N w_i f(\boldsymbol{\theta}_i) && \text{define } w_i := \frac{\tilde{r}_i}{\frac{1}{N} \sum_{j=1}^N \tilde{r}_j}
 \end{aligned}$$

Where w_i are called **importance weights** and intuitively they compensate the error introduced by sampling from the distribution $q(\boldsymbol{\theta})$ rather than $p(\boldsymbol{\theta} | \mathbf{x})$. The introduction of importance weights should make it clear that importance sampling could perform badly if the proposal distribution $q(\boldsymbol{\theta})$ is close to zero in regions of the support Θ where $p(\boldsymbol{\theta} | \mathbf{x})$ is large, as this would invalidate the sum. Unfortunately, this tends to happen when we deal with high dimensional random variables, making Importance Sampling suffer from the *Curse of Dimensionality*, even though to a much lesser extent than uniform sampling. Indeed, one could think of Uniform sampling as begin a specific case of Importance sampling where the proposal distribution is just a uniform density⁴.

2.2.3 Rejection Sampling

Both Uniform and Importance sampling are methods that allow us to approximate expectations such as (1.9), but they are not methods that provide us with samples $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ from $p(\boldsymbol{\theta} | \mathbf{x})$ themselves. In theory, one could find the inverse cdf of $p(\boldsymbol{\theta} | \mathbf{x})$, apply theorem 2.1 and use a Monte Carlo approximation. However, this is often impractical, especially when we are working with complex

⁴Care should be taken in practice as one will need to restrict Θ to a finite interval to be able to draw samples from $q(\boldsymbol{\theta})$.

distributions, either because we might not have an explicit expression for F^{-1} , or because we can find a much more efficient method of sampling, such as Rejection Sampling. It is particularly suited when:

1. For any given θ in the support Θ , it is straightforward to compute $p(\theta | \mathbf{x})$ up to some normalizing constant, i.e. we can calculate $p(\mathbf{x} | \theta)p(\theta)$ but not $p(\theta | \mathbf{x})$.
2. It is difficult and computationally expensive to sample from $p(\theta | \mathbf{x})$.

The intuition behind Rejection Sampling is that we want to choose a distribution that it is easy to sample from, as in Importance sampling, and we want to rescale it so that the area of $p(\mathbf{x} | \theta)p(\theta)$ is underneath it. Then, we want to draw samples from this simpler distribution and have a criteria that allows us to "accept" some of those samples as coming from our target distribution $p(\theta | \mathbf{x})$.

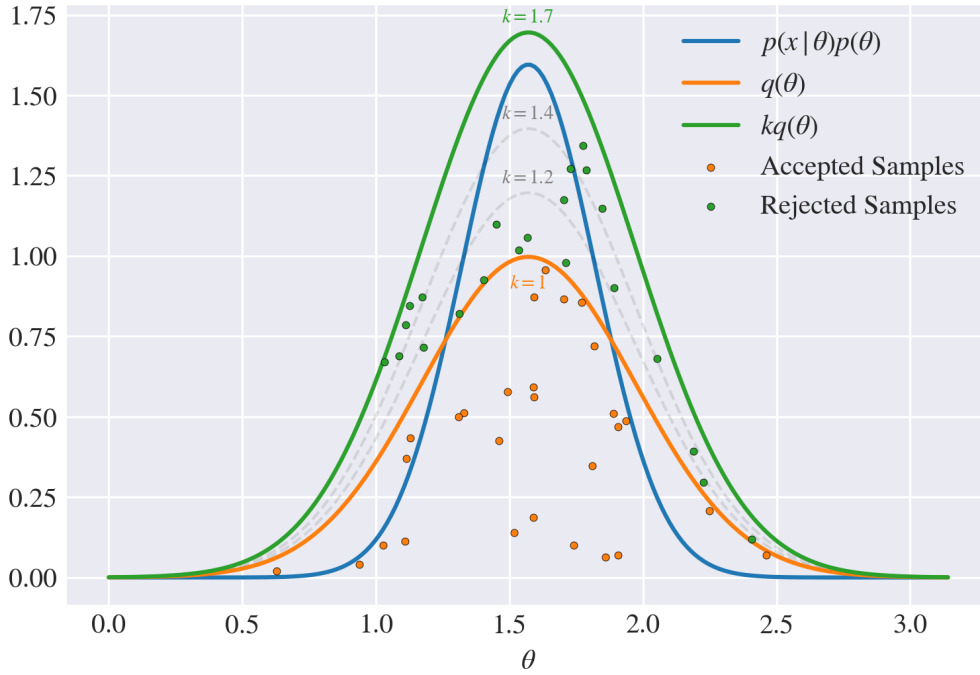


Figure 2.3: Example of a proposal distribution and its comparison function

More formally, let us choose a **proposal distribution** $q(\theta)$ that it is easy to sample from, i.e. we can effortlessly generate $\theta_i \sim q(\theta)$, and that has compatible support with Θ . Next, we want to rescale this distribution so that it is always above our unnormalized target distribution $p(\mathbf{x} | \theta)p(\theta)$. To do this, we look for a value $k \in \mathbb{R}$ such that

$$(2.7) \quad p(\mathbf{x} | \theta)p(\theta) \leq kq(\theta) \quad \forall \theta \in \Theta \quad k > 1$$

From Equation (2.7) and Figure 2.3 it should be clear that $kq(\boldsymbol{\theta})$ is not a probability distribution as it does not integrate to one, therefore it is often called **comparison function**.

The Rejection Sampling algorithm to generate samples $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ from $p(\boldsymbol{\theta} | \mathbf{x})$ then reads as follows:

Algorithm 1: Rejection Sampling

```

1 for  $i = 1, 2, \dots, N$  do:
2   Draw a sample from the proposal distribution  $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$ .
3   Draw a uniform sample underneath the comparison function  $u_i | \boldsymbol{\theta}_i \sim \mathcal{U}(0, kq(\boldsymbol{\theta}_i))$ .
4   if  $u_i > p(\mathbf{x} | \boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)$  do:
5     Reject  $\boldsymbol{\theta}_i$ .
6   else:
7     Accept  $\boldsymbol{\theta}_i$  as an independent sample from  $p(\boldsymbol{\theta} | \mathbf{x})$ .
8   end
9 end
    
```

Intuitively, we are generating *uniform* samples under the *graph* of the comparison function,

$$(\boldsymbol{\theta}_i, u_i) \sim \mathcal{U}(\{(\boldsymbol{\theta}, u) : 0 \leq u \leq kq(\boldsymbol{\theta})\})$$

and we are accepting them if they are *also* under the graph of $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$, as shown in Figure 2.3. To see this, denote $p(\boldsymbol{\theta}, u | \mathbf{x})$ the joint density. Then

$$p(\boldsymbol{\theta}, u | \mathbf{x}) = \begin{cases} p(u | \boldsymbol{\theta})p(\boldsymbol{\theta} | \mathbf{x}) = \mathcal{U}(0, kq(\boldsymbol{\theta})) \times q(\boldsymbol{\theta}) = \frac{1}{kq(\boldsymbol{\theta})} q(\boldsymbol{\theta}) = \frac{1}{k} & \text{if } (\boldsymbol{\theta}, u)^\top \in \{(\boldsymbol{\theta}, u) : 0 \leq u \leq kq(\boldsymbol{\theta})\} \\ 0 & \text{otherwise} \end{cases}$$

which proves the assertion above. A more detailed proof of this statement can be found in [6].

The **acceptance probability** can naturally be found by comparing the area of the unnormalized target distribution $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$ and the comparison function $kq(\boldsymbol{\theta})$:

$$(2.8) \quad \mathbb{P}(\text{accept}) := \int \frac{p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})}{kq(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{p(\mathbf{x})}{k} = \frac{\text{area under } p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})}{\text{area under } kq(\boldsymbol{\theta})}$$

Following the analysis of Devroye [7] one can model the number M of trials leading to a successful sample $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} | \mathbf{x})$ as a Geometric distribution with probability equal to $\frac{p(\mathbf{x})}{k}$ and with support $\{1, 2, \dots\}$

$$\mathbb{P}(M = j) = \underbrace{\left(1 - \frac{p(\mathbf{x})}{k}\right)^{j-1}}_{j-1 \text{ rejections}} \times \underbrace{\frac{p(\mathbf{x})}{k}}_{1 \text{ acceptance}} \quad \text{for } j \in \{1, 2, \dots\}$$

From this it is straightforward to find the expected number of draws leading to a successful sample

$$(2.9) \quad \mathbb{E}_{M \sim \text{Geom}\left(\frac{p(\mathbf{x})}{k}\right)}[M] = \frac{1}{\frac{p(\mathbf{x})}{k}} = \frac{k}{p(\mathbf{x})}$$

Equations (2.8) and (2.9) show that to minimize wasteful calculations we need to choose k to be as small as possible, i.e. we require the comparison function to envelope the unnormalized target

distribution as tightly as possible.

It is important to observe that throughout the algorithm we work with $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$ only, and yet we claim to obtain samples from $p(\boldsymbol{\theta} | \mathbf{x})$. The motivation is simple enough: the samples generated \mathbf{x}_i are proportional to the unnormalized target distribution $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$, but since $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x})p(\boldsymbol{\theta} | \mathbf{x})$, the samples $\boldsymbol{\theta}_i$ are also proportional to the normalized target distribution $p(\boldsymbol{\theta} | \mathbf{x})$ as the normalization constant $p(\mathbf{x})$ is common to all $\boldsymbol{\theta} \in \Theta$ [7].

2.3 Markov Chains Monte Carlo

In order to understand MCMC we need to introduce some concepts from the theory of Stochastic Processes and Markov Chains [8]. For all our purposes, we will consider only discrete-time Markov Chains defined on a discrete state-space⁵, even though the important results that will follow shortly, can be proved for more general settings [12, 13].

Definition 2.1 (Stochastic Process). A **stochastic process** is a collection of random variables

$$\{\mathbf{X}^{(t)} : t \in \mathcal{T}\}$$

that are indexed by an **index set** T and taking values in a **state space** S . If \mathcal{T} is discrete, we say that it is a *discrete-time* stochastic process⁶.

Definition 2.2 (Markov Chain). For our purposes, a **Markov Chain** is a discrete-time stochastic process with index \mathcal{T} , whose finite⁷ state-space S , with $|S| = r$, is discrete, and satisfying the **Markov Property**:

$$(2.10) \quad \mathbb{P}(\mathbf{X}^{(t+1)} = \mathbf{x}_{i_{t+1}} | \mathbf{X}^{(1)} = \mathbf{x}_{i_1}, \mathbf{X}^{(2)} = \mathbf{x}_{i_2}, \dots, \mathbf{X}^{(t)} = \mathbf{x}_{i_t}) = \mathbb{P}(\mathbf{X}^{(t+1)} = \mathbf{x}_{i_{t+1}} | \mathbf{X}^{(t)} = \mathbf{x}_{i_t}) \quad \forall \mathbf{x}_{i_j} \in S$$

The Markov property is also called the *memoryless* property of the chain. It ensures that the future state depends only on the current one, and not on the history of previous states. This makes the job of specifying a Markov Chain relatively easy, because we only need to specify a starting probability distribution over the states and how to go from one state to the next.

Definition 2.3 (Time-homogeneous Markov Chain). A Markov Chain is said to be **time-homogeneous** if its *transition probabilities*

$$(2.11) \quad \mathbb{P}(\mathbf{X}^{(t+1)} = \mathbf{x}_j | \mathbf{X}^{(t)} = \mathbf{x}_i) \equiv T(\mathbf{x}_i, \mathbf{x}_j) \quad \forall \mathbf{x}_i, \mathbf{x}_j \in S, t \in \mathcal{T}$$

are the same independently of t .

⁵To avoid sounding too tedious, I will refer to them as simply *Markov Chains*.

⁶Here we will focus on the case $\mathcal{T} = \mathbb{N}$

⁷We will often write $S = \{\mathbf{x}_1, \dots, \mathbf{x}_r\}$

Homogeneous Markov Chains are very convenient because we can use a common transition kernel⁸ $T(\mathbf{x}_i, \mathbf{x}_j)$ to define a probability distribution over the state-space S . It turns out that writing the transition probabilities in the form of a matrix \mathbf{T} where the $i - j^{\text{th}}$ element is the probability of going from state \mathbf{x}_i to state \mathbf{x}_j makes the job of studying the behavior of Markov Chains much easier.

$$\mathbf{T} = \begin{bmatrix} T(\mathbf{x}_1, \mathbf{x}_1) & \dots & T(\mathbf{x}_1, \mathbf{x}_r) \\ \vdots & \ddots & \vdots \\ T(\mathbf{x}_r, \mathbf{x}_1) & \dots & T(\mathbf{x}_r, \mathbf{x}_r) \end{bmatrix}$$

A consequence of Definition 2.3 is that one can find the probability of going from state \mathbf{x}_i to state \mathbf{x}_j in t steps by taking powers of the transition matrix \mathbf{T}^t . If we denote by \mathbf{p}_0 the *row* vector⁹ representing the probability distribution over the state-space S at time $t = 0$,

$$\mathbf{p}_0 := (p_0(\mathbf{x}_1), \dots, p_0(\mathbf{x}_r))$$

then the probability distribution at time $t > 0$ is given by:

$$\mathbf{p}_t = \mathbf{p}_0 \mathbf{T}^t = (p_t(\mathbf{x}_1), \dots, p_t(\mathbf{x}_r))$$

where each entry of \mathbf{p}_t can be found as

$$p_t(\mathbf{x}_j) = \sum_{i=1}^r p_0(\mathbf{x}_i) T^t(\mathbf{x}_i, \mathbf{x}_j)$$

It should be clear that generally \mathbf{p}_{t+1} will be a different vector from \mathbf{p}_t due to the arbitrary nature of \mathbf{p}_0 and \mathbf{T} . One can, however, imagine that under specific condition it could be possible to design a distribution \mathbf{p}_0 that remains unchanged when acted on by \mathbf{T}^t .

Definition 2.4 (Invariant Distribution). We say that a *row* vector $\mathbf{s} \in \mathbb{R}^{1 \times r}$ is a **stationary** or **invariant distribution** over the state-space S with respect to the transition matrix \mathbf{T} if

$$(2.12) \quad \mathbf{s} = \mathbf{s} \mathbf{T} \quad \text{or equivalently} \quad s(\mathbf{x}_j) = \sum_{i=1}^r s(\mathbf{x}_i) T(\mathbf{x}_i, \mathbf{x}_j) \quad \forall \mathbf{x}_j \in S$$

If any at step t the marginal distribution \mathbf{p}_t becomes equal to the stationary distribution \mathbf{s} , then \mathbf{s} will remain the probability distribution of the chain for all future steps. MCMC aim to design a transition matrix \mathbf{T} such that our target distribution of interest is invariant with respect to that matrix. In order to design such \mathbf{T} , we impose a further condition on the Markov Chain: it must be time-reversible. Reversibility amounts to having the same probability of going from state \mathbf{x}_i to state \mathbf{x}_j and vice-versa, under the stationary distribution.

⁸Technically, in the discrete case we have *transition probabilities* T_{ij} while in the continuous case we have *transition kernel*.

⁹In this discussion of Markov Chains we use a vector \mathbf{p}_0 because the state-space is finite and discrete. For continuous Markov chains we use a density. Notice that since this vector represents probabilities we must have $p_0(x_i) > 0$ for all $x_i \in S$, and $\sum_{i=1}^r p_0(\mathbf{x}_i) = 1$

Definition 2.5 (Reversible Markov Chain). A Markov Chain is said to be **reversible** or to satisfy the **detailed balance** equation if there exists a probability distribution \mathbf{s} on S such that

$$(2.13) \quad s(\mathbf{x}_i) T(\mathbf{x}_i, \mathbf{x}_j) = s(\mathbf{x}_j) T(\mathbf{x}_j, \mathbf{x}_i) \quad \forall \mathbf{x}_j, \mathbf{x}_i \in S$$

Algorithm 2: Metropolis-Hastings

```

1 for  $k = 1, 2, \dots, n$  do:
2   Specify a nonzero value for either  $\ell_{kk}$  (Doolittle) or  $u_{kk}$  (Crout) and compute the other
   from

$$\ell_{kk} u_{kk} = a_{kk} - \sum_{s=1}^{k-1} \ell_{ks} u_{sk}$$

   (Because the diagonal one is the last one before the zeros).
3   Build the  $k$ -th row of  $U$ :
4   for  $j = k + 1, k + 2, \dots, n$  do:

$$u_{kj} = \frac{1}{\ell_{kk}} \left( a_{kj} - \sum_{s=1}^{k-1} \ell_{is} u_{sk} \right)$$

5   end
6   Build the  $k$ -th column of  $L$ :
7   for  $i = k + 1, k + 2, \dots, n$  do:

$$\ell_{ik} = \frac{1}{u_{kk}} \left( a_{ik} - \sum_{s=1}^{k-1} \ell_{is} u_{sk} \right)$$

   end
8 end

```

DETERMINISTIC APPROXIMATIONS

Deterministic approximation methods aim to find a distribution that resamples the posterior distribution, rather than drawing samples from it. One of the main advantage of these methods over stochastic approximation methods seen in the previous chapter, is that we do not have to wait for the Markov Chain to converge to the equilibrium distribution, and for this reason deterministic approximation are more well-suited for large-scale problems. The drawback is that we will only be working with approximations and not with the exact posterior distribution, so we will not have any convergence guarantee.

3.1 Laplace Approximation

3.1.1 A General Framework

The idea of Laplace approximation is to approximate $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$ by a multivariate normal distribution centered at the mode. From basic results in optimization we know that the mode $\boldsymbol{\theta}_0 \in \Theta$ will satisfy:

- $\nabla_{\boldsymbol{\theta}} p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0$ (i.e. it will be a critical point)
- $\nabla^2 p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} < 0$ (i.e. it will be a maximum)

As usual, since the natural logarithm is a strictly increasing function, $\boldsymbol{\theta}_0$ will also be the maximum of $\ln(p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta}))$, which we can Taylor expand to find a second order approximation, using the first order optimality condition above:

$$\ln(p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})) \approx \ln(p(\mathbf{x} | \boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla^2 \ln(p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

Exponentiating both sides leads to

$$p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \approx p(\mathbf{x} | \boldsymbol{\theta}_0) p(\boldsymbol{\theta}_0) \exp \left\{ \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla^2 \ln(p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right\}$$

Comparing the expression above with (A.5) we notice that the exponential term can be seen as the exponential part of a multivariate normal distribution with

$$\boldsymbol{\mu} = \boldsymbol{\theta}_0 \in \mathbb{R}^p \quad \text{and} \quad \boldsymbol{\Sigma}^{-1} = -\nabla^2 \ln(p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \in \mathbb{R}^{p \times p}$$

With this in mind, we can define our Gaussian approximation to be

$$(3.1) \quad q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{\theta}_0, -H(\boldsymbol{\theta}_0)^{-1})$$

where $H(\boldsymbol{\theta}_0)$ is the Hessian matrix of $p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$ at the mode $\boldsymbol{\theta}_0$. Simply put, Laplace method approximates $p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$ with a multivariate normal centered at the mode whose variance-covariance matrix is the observed Fisher information matrix evaluated at the mode.

Notice that an important drawback of this method is that it bases its approximation solely on the behavior of $p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$ at a single point $\boldsymbol{\theta}_0$.

3.1.2 Application to Bayesian Logistic Regression

We can apply result (3.1) to the logistic regression log-posterior that was found in Section 1.3 to be

$$\ln(p(\boldsymbol{\beta} | \mathbf{y})) \propto -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \sum_{i=1}^n y_i \ln(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \ln(1 - \pi_i(\boldsymbol{\beta}))$$

First, let's find the variance-covariance matrix. The derivative of the first term of the log-likelihood can be found as follows:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} \left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right) &= -\frac{1}{2} \nabla_{\boldsymbol{\beta}} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta} + \boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0) \\ &= -\frac{1}{2} (2\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0) \\ &= -\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \end{aligned}$$

Differentiating again with respect to $\boldsymbol{\beta}$ and multiplying by -1 yields:

$$(3.2) \quad -\nabla_{\boldsymbol{\beta}}^2 \left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right) = \boldsymbol{\Sigma}_0^{-1}$$

In order to differentiate the second term with respect to $\boldsymbol{\beta}$, we need to use the following useful fact about the Sigmoid function:

$$(3.3) \quad \frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a)) \quad \text{where} \quad \sigma(a) := \frac{\exp(a)}{1 + \exp(a)}$$

It then follows from (1.11) that

$$\nabla_{\boldsymbol{\beta}} \pi_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})(1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})) \nabla_{\boldsymbol{\beta}} (\mathbf{x}_i^\top \boldsymbol{\beta}) = \pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta})) \mathbf{x}_i$$

We can then use this result to find the derivative of the second term of the log-posterior.

$$\begin{aligned}
 \nabla_{\boldsymbol{\beta}} \left(\sum_{i=1}^n y_i \ln(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \ln(1 - \pi_i(\boldsymbol{\beta})) \right) &= \sum_{i=1}^n \nabla_{\boldsymbol{\beta}} (y_i \ln(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \ln(1 - \pi_i(\boldsymbol{\beta}))) \\
 &= \sum_{i=1}^n \left[\frac{y_i}{\pi_i(\boldsymbol{\beta})} \nabla_{\boldsymbol{\beta}} \pi_i(\boldsymbol{\beta}) - \frac{1 - y_i}{1 - \pi_i(\boldsymbol{\beta})} \nabla_{\boldsymbol{\beta}} \pi_i(\boldsymbol{\beta}) \right] \\
 &= \sum_{i=1}^n [y_i(1 - \pi_i(\boldsymbol{\beta})) \mathbf{x}_i - (1 - y_i) \pi_i(\boldsymbol{\beta}) \mathbf{x}_i] \\
 &= \sum_{i=1}^n (y_i - \pi_i(\boldsymbol{\beta})) \mathbf{x}_i
 \end{aligned}$$

Differentiating again with respect to $\boldsymbol{\beta}$ and multiplying by -1 yields

$$(3.4) \quad -\nabla_{\boldsymbol{\beta}}^2 \left(\sum_{i=1}^n y_i \ln(\pi_i(\boldsymbol{\beta})) + (1 - y_i) \ln(1 - \pi_i(\boldsymbol{\beta})) \right) = \sum_{i=1}^n \pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta})) \mathbf{x}_i \mathbf{x}_i^\top$$

Combining (3.2) and (3.4) gives an expression for observed Fisher information matrix

$$-H(\boldsymbol{\beta}_0) = \boldsymbol{\Sigma}_0^{-1} + \sum_{i=1}^n \pi_i(\boldsymbol{\beta}_0)(1 - \pi_i(\boldsymbol{\beta}_0)) \mathbf{x}_i \mathbf{x}_i^\top$$

where $\boldsymbol{\beta}_0$ is the mode of the posterior, which in practice is often found using an optimization routine. The multivariate normal Laplace approximation to the posterior becomes:

$$(3.5) \quad q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, -H(\boldsymbol{\beta}_0)^{-1})$$

3.2 Variational Methods

3.2.1 A General Framework

As noted at the end of Section 3.1.1, Laplace approximation can look simplistic because it approximates the posterior using only information at the mode of the posterior distribution. On the other hand, Variational Inference aims to use more information to approximate $p(\boldsymbol{\theta} \mid \mathbf{x})p(\boldsymbol{\theta})$ with a similar distribution.

There are infinitely many distributions that we can choose from and considering all of them would render Variational Inference inefficient and ultimately useless. To overcome this issue, one has to specify a family of distributions \mathcal{D} that is flexible enough to allow distributions very similar to $p(\boldsymbol{\theta} \mid \mathbf{x})p(\boldsymbol{\theta})$ while being restrictive enough so that we don't have to consider every possible probability distribution available. Once the family of distributions \mathcal{D} has been defined, the next step is to choose the distribution $q(\boldsymbol{\theta}) \in \mathcal{D}$ that is closest to our target distribution, according to some objective function. Such an objective function should be able to compare how similar two distributions are, and a good candidate for comparing two continuous¹ distributions $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \mathbf{x})$ is the

¹Notice that the KL-divergence can equally be defined for discrete distributions by replacing integrals with summations.

Kullback-Leiber divergence, or KL-divergence, described in Appendix A.4.

$$(3.6) \quad \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) := \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ln \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathbf{x})} \right) \right] = \int_{\Theta} q(\boldsymbol{\theta}) \ln \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathbf{x})} \right) d\boldsymbol{\theta}$$

It is important to notice a few properties of this objective function:

- It is *not* symmetric $\text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) \neq \text{KL}(p(\boldsymbol{\theta} \mid \mathbf{x}) \parallel q(\boldsymbol{\theta}))$.
- It is non-negative $\text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) \geq 0$.
- It is minimized when the two distributions are equal, i.e. $\text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) = 0$ if and only if $q(\boldsymbol{\theta}) \stackrel{\text{a.e.}}{=} p(\boldsymbol{\theta} \mid \mathbf{x})$

With this in mind, we can concisely rewrite the variational approach as

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{D}}{\text{argmin}} \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x}))$$

Unfortunately, using the KL-divergence is often impractical because it would require knowing the normalization constant for $p(\boldsymbol{\theta} \mid \mathbf{x})$, which is likely to be unknown. Indeed one can rewrite (3.6) as:

$$\begin{aligned} \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) &:= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ln \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathbf{x})} \right) \right] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\ln(q(\boldsymbol{\theta})) - \ln \left(\frac{p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{x})} \right) \right] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(q(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}))] + \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x}))] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(q(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}))] + \ln(p(\mathbf{x})) \end{aligned}$$

A simple remedy to this, is to "shift" the KL-divergence to not depend on the model evidence $p(\mathbf{x})$ anymore. Thus, we define a new objective function called **Evidence Lower Bound** (ELBO), which takes the following form [4]:

$$(3.7) \quad \begin{aligned} \text{elbo}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) &= -\text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) + \ln(p(\mathbf{x})) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(q(\boldsymbol{\theta}))] \end{aligned}$$

Its name comes from the fact that it provides a lower bound for the model evidence. This can be seen by rewriting the formula above as

$$\ln(p(\mathbf{x})) = \text{elbo}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) + \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x}))$$

and using the fact that the KL-divergence is non negative leads to

$$\ln(p(\mathbf{x})) \geq \text{elbo}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x}))$$

Having defined this new objective function, one can observe that since $\ln(p(\mathbf{x}))$ is constant in \mathcal{D} , maximizing the ELBO is equivalent to minimizing the KL-divergence.

For a better insight into what it practically means to maximize the evidence lower bound, we can rewrite (3.7) following the work of Blei et al. [4]:

$$\begin{aligned}
 \text{elbo}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(q(\boldsymbol{\theta}))] \\
 &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}))] + \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(q(\boldsymbol{\theta}))] \\
 (3.8) \qquad &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ln(p(\mathbf{x} \mid \boldsymbol{\theta}))] - \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}))
 \end{aligned}$$

From (3.8) we can deduce that maximizing the evidence lower bound means

- Maximizing the expected log-likelihood. This can be interpreted as encouraging distributions of parameters $\boldsymbol{\theta}$ that explain the data \mathbf{x} the best.
- Minimizing the KL divergence between $q(\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$. This means that we require the approximate distribution $q^*(\boldsymbol{\theta})$ to be close to the prior.

3.2.2 Local Variational Methods

In the previous section we have seen that variational methods aim to maximizing a lower bound on the model evidence. The lower bound is found by bounding the whole posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$, so such methods are called *global* variational methods.

A simpler, but often equally powerful, strategy is to bound individual variables or terms within the posterior distribution and then combine those bounds to obtain one for the model evidence. Naturally, such methods are referred to as *local* variational methods [3].

In this thesis, we are interested in Bayesian logistic regression, so the model evidence is given by

$$p(\mathbf{y}) = \int_{\mathbb{R}^p} p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

where $p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and the likelihood is given by (1.12) as:

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = \prod_{i=1}^n p(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}) = \prod_{i=1}^n \pi_i(\boldsymbol{\beta})^{y_i} (1 - \pi_i(\boldsymbol{\beta}))^{1-y_i}$$

where $\pi_i(\boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})$. The strategy adopted by Jaakkola and Jordan [10] is to restrict the distributions in \mathcal{D} to be only Gaussian, and thus to obtain a Gaussian lower bound on the success probability $\pi_i(\boldsymbol{\beta})$, which is essentially a sigmoid transformation of the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, see Figure 3.1. One can show (Bishop [3], Jaakkola and Jordan [10]) that

$$(3.9) \qquad \pi_i(\boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}) \geq \sigma(\xi_i) \exp \left\{ \frac{\mathbf{x}_i^\top \boldsymbol{\beta} - \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right\}$$

where each ξ_i is a **variational parameter** which has to be optimized for the bound to be maximal, and

$$(3.10) \qquad \lambda(\xi_i) := \frac{1}{2\xi_i} \left[\sigma(\xi_i) - \frac{1}{2} \right]$$

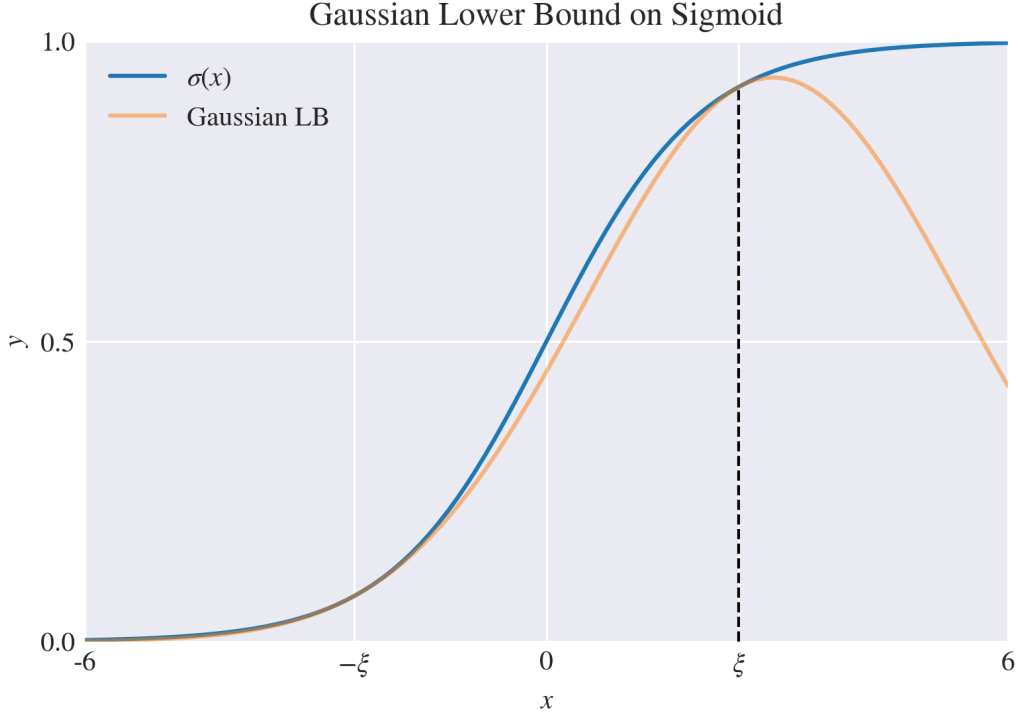


Figure 3.1: Example of Jaakkola and Jordan's lower bound

Our aim is to use (3.9) to bound each factor $p(y_i; \mathbf{x}_i, \boldsymbol{\beta})$ in (1.12). In order to do this, we can rewrite each term as follows:

$$\begin{aligned}
 p(y_i; \mathbf{x}_i, \boldsymbol{\beta}) &= \pi_i(\boldsymbol{\beta})^{y_i} (1 - \pi_i(\boldsymbol{\beta}))^{1-y_i} \\
 &= \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} (1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}))^{1-y_i} \\
 &= \left(\frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{y_i} \left(1 - \frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \right)^{1-y_i} \\
 &= \frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i\}}{(1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\})^{y_i}} \frac{1}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} (1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\})^{y_i} \\
 &= \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i\} \frac{1}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \\
 &= \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i\} \frac{\exp\{-\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1 + \exp\{-\mathbf{x}_i^\top \boldsymbol{\beta}\}} \\
 &= \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i\} \sigma(-\mathbf{x}_i^\top \boldsymbol{\beta})
 \end{aligned}
 \tag{3.11}$$

Now it is straightforward to apply the bound in (3.9) to (3.11)

$$\begin{aligned} p(y_i; \mathbf{x}_i^\top, \boldsymbol{\beta}) &\geq \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i\} \sigma(\xi_i) \exp\left\{-\frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i)((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2)\right\} \\ &= \sigma(\xi_i) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i - \frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i)((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2)\right\} \end{aligned}$$

and consequently we can bound the likelihood

$$(3.12) \quad p(\mathbf{y} | \boldsymbol{\beta}) \geq \tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \boldsymbol{\xi}) := \prod_{i=1}^n \sigma(\xi_i) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta} y_i - \frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i)((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2)\right\}$$

where $\boldsymbol{\xi} := (\xi_1, \dots, \xi_n)^\top$ is a vector of variational parameters².

Recall that in Section 3.1.1 we have found a bound on $p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$ that was an exponential of a quadratic function of $\boldsymbol{\theta}$, and then we have obtained a Gaussian approximation to the full posterior by normalizing it to become a proper normal distribution. Similarly, we can find the following bound

$$p(\mathbf{y} | \boldsymbol{\beta})p(\boldsymbol{\beta}) \geq \tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \boldsymbol{\xi})p(\boldsymbol{\beta})$$

on the unnormalized posterior distribution. Since the logarithm is a strictly increasing function, we can equally write

$$\begin{aligned} \ln(p(\mathbf{y}, \boldsymbol{\beta})) &= \ln(p(\mathbf{y} | \boldsymbol{\beta})p(\boldsymbol{\beta})) \\ &\geq \ln(\tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \boldsymbol{\xi})p(\boldsymbol{\beta})) \\ &= \ln(\tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \boldsymbol{\xi})) + \ln(p(\boldsymbol{\beta})) \\ &= \sum_{i=1}^n \left[\ln(\sigma(\xi_i)) + \mathbf{x}_i^\top \boldsymbol{\beta} y_i - \frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i)((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right] - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \text{const} \\ &= \sum_{i=1}^n \left[\mathbf{x}_i^\top \boldsymbol{\beta} \left(y_i - \frac{1}{2} \right) - \lambda(\xi_i)(\mathbf{x}_i^\top \boldsymbol{\beta})^2 \right] - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \text{const} \\ &= \sum_{i=1}^n \left[\boldsymbol{\beta}^\top \mathbf{x}_i \left(y_i - \frac{1}{2} \right) - \lambda(\xi_i)(\boldsymbol{\beta}^\top \mathbf{x}_i)^2 \right] - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \text{const} \\ &= \boldsymbol{\beta}^\top \sum_{i=1}^n \mathbf{x}_i \left(y_i - \frac{1}{2} \right) - \boldsymbol{\beta}^\top \left[\sum_{i=1}^n \lambda(\xi_i) \mathbf{x}_i \mathbf{x}_i^\top \right] \boldsymbol{\beta} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) + \text{const} \\ &= \boldsymbol{\beta}^\top \sum_{i=1}^n \mathbf{x}_i \left(y_i - \frac{1}{2} \right) - \boldsymbol{\beta}^\top \left[\sum_{i=1}^n \lambda(\xi_i) \mathbf{x}_i \mathbf{x}_i^\top \right] \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \text{const} \\ (3.13) \quad &= \boldsymbol{\beta}^\top \left[\sum_{i=1}^n \mathbf{x}_i \left(y_i - \frac{1}{2} \right) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right] - \frac{1}{2} \boldsymbol{\beta}^\top \left[2 \sum_{i=1}^n \lambda(\xi_i) \mathbf{x}_i \mathbf{x}_i^\top + \boldsymbol{\Sigma}_0^{-1} \right] \boldsymbol{\beta} + \text{const} \end{aligned}$$

where we have absorbed all terms not depending on $\boldsymbol{\beta}$ into a constant and rewritten the entire expression in terms of powers of $\boldsymbol{\beta}$. We can then complete the square using equations (A.6)-(A.8) to

²In equation (3.12) we have defined $\tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \boldsymbol{\xi}) = \tilde{p}(\mathbf{y} | \mathbf{X}_D^{\text{RV}} = \mathbf{X}_D^{\text{obs}}, \boldsymbol{\beta}; \boldsymbol{\xi})$ to be equal to the right-hand side of the equation. This is not a density because it is not normalized. If we were to normalize it, it would cease to be a bound [3].

obtain the form of our variational Gaussian approximation

$$(3.14) \quad q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$$

where the mean and variance-covariance matrix are given by

$$(3.15) \quad \boldsymbol{\Sigma}_v^{-1} = 2 \sum_{i=1}^n \lambda(\xi_i) \mathbf{x}_i \mathbf{x}_i^\top + \boldsymbol{\Sigma}_0^{-1}$$

$$(3.16) \quad \boldsymbol{\mu}_v = \boldsymbol{\Sigma}_v \left(\sum_{i=1}^n \mathbf{x}_i \left(y_i - \frac{1}{2} \right) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right)$$

So far, we have bounded each success probability $\pi_i(\boldsymbol{\beta})$ with a Gaussian lower bound depending on the variational parameter ξ_i , then we have used the independence assumption to factorize our likelihood and we have used the bound on the success probability to find a lower bound for the whole likelihood. Next, we managed to find a normal approximation to the posterior distribution by first bounding $\ln(p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta}))$ and then completing the square. Looking at equations (3.15)-(3.16) we quickly spot that they depend on the vector of variational parameters $\boldsymbol{\xi}$. Recall that in Variational Inference our aim is to maximize the evidence lower bound and therefore the parameters $\boldsymbol{\xi}$ have to be determined via maximization of the following bound

$$p(\mathbf{y}) = \int_{\mathbb{R}^p} p(\mathbf{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} \geq \int_{\mathbb{R}^p} \tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

Once again, since the logarithm is a strictly increasing function, maximizing the bound above is just equivalent to maximizing

$$(3.17) \quad \mathcal{B}(\boldsymbol{\xi}) := \ln \left(\int_{\mathbb{R}^p} \tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} \right).$$

There are two ways of maximizing (3.17). One method uses the EM algorithm, the second uses straight forward integration, luckily both lead to the same result, summarized in Algorithm 3 below.

Algorithm 3: Variational Approximation

1 Initialize $\boldsymbol{\xi} = (\xi_1^{(1)}, \dots, \xi_n^{(1)})^\top \in \mathbb{R}^n$ randomly.

2 **for** $j = 1, 2, \dots, \Delta$:

3 **for** $i = 1, 2, \dots, n$ **do**:

4 Find mean and variance-covariance matrix and update variational parameters.

$$\boldsymbol{\Sigma}_v = \left(2 \sum_{i=1}^n \lambda(\xi_i^{(j)}) \mathbf{x}_i \mathbf{x}_i^\top + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}$$

$$\boldsymbol{\mu}_v = \boldsymbol{\Sigma}_v \left(\sum_{i=1}^n \mathbf{x}_i \left(y_i - \frac{1}{2} \right) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right)$$

$$\xi_i^{(j+1)} = \sqrt{\mathbf{x}_i^\top (\boldsymbol{\Sigma}_v + \boldsymbol{\mu}_v \boldsymbol{\mu}_v^\top) \mathbf{x}_i}$$

end

5 **end**

Notice that one could use several stopping criteria however, in practice, this algorithm converges after just a very small number of iterations Δ . In the algorithm we write $\xi_i^{(j)}$ to indicate the approximation at iteration $j \in \{1, \dots, \Delta\}$ of the variational parameter associated with the i -th observation.

In the following section we will compare stochastic and deterministic approximations in the context of Bayesian logistic regression. The examples have been coded in Python.

4.1 No Explanatory variables and No prior, only likelihood

4.1.1 Laplace Approximation

Let $y_1, \dots, y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi)$. Then we define

$$\theta := \ln\left(\frac{\pi}{1-\pi}\right)$$

The success probability becomes

$$p(y_i = 1) = \frac{e^\theta}{1 + e^\theta}$$

The likelihood function becomes

$$\begin{aligned} \mathcal{L}(\mathbf{y} | \theta) &= \prod_{i=1}^n p(y_i = 1)^{y_i} (1 - p(y_i = 1))^{1-y_i} \\ &= \prod_{i=1}^n \left(\frac{e^\theta}{1 + e^\theta} \right)^{y_i} \left(\frac{1}{1 + e^\theta} \right)^{1-y_i} \\ &= \prod_{i=1}^n \frac{e^{\theta y_i}}{1 + e^\theta} \end{aligned} \tag{4.1}$$

the log-likelihood function becomes

$$\ell(\mathbf{y} | \theta) = \theta \sum_{i=1}^n y_i - n \ln(1 + e^\theta) \tag{4.2}$$

To find the mode we find the derivative with respect to θ and set it equal to zero

$$(4.3) \quad \frac{\partial \ell(\mathbf{y} | \theta)}{\partial \theta} = \sum_{i=1}^n y_i - n \frac{e^\theta}{1 + e^\theta}$$

setting this equal to zero we obtain

$$(4.4) \quad \frac{e^{\hat{\theta}}}{1 + e^{\hat{\theta}}} = \bar{y} \quad \implies \quad \hat{\theta} := \ln\left(\frac{\bar{y}}{1 - \bar{y}}\right)$$

next we take the second derivative with respect to θ and evaluate it at $\theta = \hat{\theta}$ to obtain the Hessian matrix at the mode.

$$(4.5) \quad \frac{\partial^2 \ell(\mathbf{y} | \theta)}{\partial \theta^2} \Big|_{\theta = \hat{\theta}} = -n\bar{y}(1 - \bar{y})$$

Then, Laplace approximation has the observed information matrix as its variance-covariance matrix

$$(4.6) \quad \mathcal{N}\left(\ln\left(\frac{\bar{y}}{1 - \bar{y}}\right), \frac{1}{n\bar{y}(1 - \bar{y})}\right)$$

4.1.2 Variational Approximation

Recall

$$(4.7) \quad \sigma(-\theta) = \frac{e^{-\theta}}{1 + e^{-\theta}} = \frac{1}{1 + e^\theta}$$

then in the likelihood equation we have

$$(4.8) \quad L(\mathbf{y} | \theta) = \prod_{i=1}^n L(y_i | \theta) = \prod_{i=1}^n e^{\theta y_i} \sigma(-\theta)$$

Each likelihood term can be bounded using the Jaakkola and Jordan's bound as

$$\begin{aligned} L(y_i | \theta) &\geq e^{\theta y_i} \sigma(\xi_i) \exp\left(-\frac{\theta + \xi_i}{2} - \lambda(\xi_i)(\theta^2 - \xi_i^2)\right) \\ &= \sigma(\xi_i) \exp\left(\theta y_i - \frac{\theta + \xi_i}{2} - \lambda(\xi_i)(\theta^2 - \xi_i^2)\right) \end{aligned}$$

The complete likelihood is bounded by

$$(4.9) \quad \begin{aligned} L(\mathbf{y} | \theta) &\geq \prod_{i=1}^n \sigma(\xi_i) \exp\left(\theta y_i - \frac{\theta + \xi_i}{2} - \lambda(\xi_i)(\theta^2 - \xi_i^2)\right) \\ &= \left(\prod_{i=1}^n \sigma(\xi_i)\right) \exp\left(\theta \sum_{i=1}^n y_i - \frac{n\theta}{2} + \sum_{i=1}^n \frac{\xi_i}{2} - \sum_{i=1}^n \lambda(\xi_i)(\theta^2 - \xi_i^2)\right) \end{aligned}$$

the log-likelihood is bounded by the following, where we have collected the powers of θ

$$(4.10) \quad \begin{aligned} \ell(\mathbf{y} | \theta) &\geq \sum_{i=1}^n \ln \sigma(\xi_i) + \theta \sum_{i=1}^n y_i - \frac{n\theta}{2} + \sum_{i=1}^n \frac{\xi_i}{2} - \sum_{i=1}^n \lambda(\xi_i)(\theta^2 - \xi_i^2) \\ &= \sum_{i=1}^n \ln \sigma(\xi_i) + \theta n\bar{y} - \frac{n\theta}{2} + \sum_{i=1}^n \frac{\xi_i}{2} - \sum_{i=1}^n \lambda(\xi_i)(\theta^2 - \xi_i^2) \\ &= \left[\sum_{i=1}^n \ln \sigma(\xi_i) + \sum_{i=1}^n \frac{\xi_i}{2} + \sum_{i=1}^n \lambda(\xi_i) \xi_i^2 \right] + \theta \left[n\left(\bar{y} - \frac{1}{2}\right) \right] + \theta^2 \left[-\sum_{i=1}^n \lambda(\xi_i) \right] \end{aligned}$$

Now notice that we can write a univariate Gaussian as follows

$$\begin{aligned} \mathcal{N}(\theta \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ (4.11) \quad &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\theta^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\theta\mu}{\sigma^2}\right) \end{aligned}$$

now taking the logarithm we obtain

$$\begin{aligned} \ln \mathcal{N}(\theta \mid \mu, \sigma^2) &= -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{\theta^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\theta\mu}{\sigma^2} \\ (4.12) \quad &= \left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right] + \theta \left[\frac{\mu}{\sigma^2}\right] + \theta^2 \left[-\frac{1}{2\sigma^2}\right] \end{aligned}$$

Comparing this with the bound on the log-likelihood we find

$$(4.13) \quad \frac{\hat{\mu}}{\hat{\sigma}^2} = n \left(\bar{y} - \frac{1}{2} \right)$$

and

$$(4.14) \quad -\frac{1}{2\hat{\sigma}^2} = -\sum_{i=1}^n \lambda(\xi_i)$$

from the latter we obtain

$$(4.15) \quad (\hat{\sigma}^2)^{-1} = 2 \sum_{i=1}^n \lambda(\xi_i)$$

and from the former we have

$$(4.16) \quad \hat{\mu} = \hat{\sigma}^2 n \left(\bar{y} - \frac{1}{2} \right)$$

4.2 No explanatory variables, prior and posterior

we put a prior on the parameter θ

$$(4.17) \quad p(\theta) = \mathcal{N}(\mu_0, \sigma_0^2)$$

for which we have

$$(4.18) \quad \ln(p(\theta)) = \left[-\frac{1}{2} \ln(2\pi\sigma_0^2) - \frac{\mu_0^2}{2\sigma_0^2}\right] + \theta \left[\frac{\mu_0}{\sigma_0^2}\right] + \theta^2 \left[-\frac{1}{2\sigma_0^2}\right]$$

then we have, considering only terms depending on θ

$$\begin{aligned} \ln(p(\mathbf{y} \mid \theta)p(\theta)) &= \ln(p(\mathbf{y} \mid \theta)) + \ln(p(\theta)) \\ &\geq \theta \left[n \left(\bar{y} - \frac{1}{2} \right) + \frac{\mu_0}{\sigma_0^2} \right] + \theta^2 \left[-\sum_{i=1}^n \lambda(\xi_i) - \frac{1}{2\sigma_0^2} \right] + \text{const} \end{aligned}$$

comparing with the logarithm of a normal distribution we then obtain

$$(4.19) \quad \frac{\hat{\mu}_V}{\hat{\sigma}_V^2} = n \left(\bar{y} - \frac{1}{2} \right) + \frac{\mu_0}{\sigma_0^2}$$

and

$$(4.20) \quad -\frac{1}{2\hat{\sigma}_V^2} = -\sum_{i=1}^n \lambda(\xi_i) - \frac{1}{2\sigma_0^2}$$

from the latter we obtain

$$(4.21) \quad (\hat{\sigma}_V^2)^{-1} = (\sigma_0^2)^{-1} + 2 \sum_{i=1}^n \lambda(\xi_i)$$

and from the former we get

$$(4.22) \quad \hat{\mu}_V = \hat{\sigma}_V^2 \left(\frac{\mu_0}{\sigma_0^2} + n \left(\bar{y} - \frac{1}{2} \right) \right)$$

4.2.1 Laplace with prior

4.2.2 Optimization of variational parameters



APPENDIX A - STATISTICS BACKGROUND

We review some fundamental concepts in probability and statistics: the Exponential family, Generalized Linear Models, the multivariate normal distribution and the KL-divergence. We also provide additional details to the derivation of the Gaussian bound used in Section 3.2.2.

A.1 Exponential Family of Distributions

Many of the most common probability distributions, such as Normal and Bernoulli, can be written, after an appropriate re-parametrization, in a common format which will allow us to construct a general theoretical framework for regression. In particular, we say that the probability distribution of a random variable Y belongs to the exponential family if we can rewrite it as

$$(A.1) \quad f_Y(y | \boldsymbol{\theta}) = h(y) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{T}(y) - A(\boldsymbol{\theta}))$$

where $\boldsymbol{\theta}$ is the vector of *natural parameters*, $\mathbf{T}(y)$ is a vector of *sufficient statistics*, and $A(\boldsymbol{\theta})$ is often called *log-partition function* and it ensures that the probability distribution function integrates to 1.

A.1.1 Bernoulli Distribution

A *binary* random variable Y that takes value 1 with probability π and value 0 with probability $1 - \pi$ is said to be Bernoulli. It has probability mass function defined below

$$(A.2) \quad p(y; \pi) = \pi^y (1 - \pi)^{1-y} \quad \text{for } y \in \{0, 1\}$$

and we denote this by $Y \sim \text{Bernoulli}(\pi)$. We can rearrange its probability mass function to show that the Bernoulli distribution belongs to the Exponential Family by noting that in this case we have

a scalar parameter $\boldsymbol{\theta} = (\pi)^\top \in \mathbb{R}$ rather than a vector of parameters, and by using the fact that the logarithm is the inverse of exponentiation:

$$\begin{aligned} p(y; \pi) &= \pi^y (1 - \pi)^{1-y} \\ &= \exp(y \ln(\pi) + (1 - y) \ln(1 - \pi)) \\ &= \exp(y \ln(\pi) + \ln(1 - \pi) - y \ln(1 - \pi)) \\ &= \exp\left(y \ln\left(\frac{\pi}{1 - \pi}\right) + \ln(1 - \pi)\right) \end{aligned}$$

In expression (A.1) we can set

$$\begin{aligned} h(y) &= 1 \\ \boldsymbol{\eta}(\boldsymbol{\theta}) &= \eta(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) \\ \mathbf{T}(y) &= T(y) = y \\ A(\boldsymbol{\theta}) &= A(\pi) = \ln(1 - \pi) \end{aligned}$$

which shows that the Bernoulli distribution belongs to the Exponential Family.

A.2 Generalized Linear Models

Generalized Linear Models are powerful for modelling the relationship between a set of *independent* or *explanatory* variables, and a set of *dependent* or *response* variables.

A.2.1 Explanatory Variables

Suppose that we have $n \times p$ random variables $X_{11}, \dots, X_{1p}, \dots, X_{n1}, \dots, X_{np}$ which we arrange into a so-called design matrix for convenience

$$\mathbf{X}_D^{\text{RV}} := \begin{pmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

where the superscript of \mathbf{X}_D^{RV} indicates that its elements are random variables, and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$. In addition, suppose that we have observed the realization of all these random variables, which we arrange in matrix form again

$$(A.3) \quad \mathbf{X}_D^{\text{obs}} := \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

where each row of $\mathbf{X}_D^{\text{obs}}$ is an observation of p **explanatory variables**.

Throughout this thesis we will only consider the observed design matrix $\mathbf{X}_D^{\text{obs}}$, meaning that whenever explanatory variables will be involved, we will be considering realizations $X_{ij} = x_{ij}$.

A.2.2 Response Variables

In addition to recording n observations \mathbf{x}_i , we also consider n corresponding random variables $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and their realizations $\mathbf{y} = (y_1, \dots, y_n)^\top$, where each Y_i follows the same member of the exponential family.

$$f_{Y_i}(y_i | \boldsymbol{\theta}) = h(y_i) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{T}(y_i) - A(\boldsymbol{\theta}))$$

To model the relationship between each pair of Y_i and \mathbf{x}_i , we choose a suitable function $g: \mathbb{R} \rightarrow \mathbb{R}$ that is *invertible* and *differentiable* and we assume

$$(A.4) \quad g(\mathbb{E}[Y_i | \mathbf{X}_i = \mathbf{x}_i]) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

where $\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^\top$ is a vector of parameters in \mathbb{R}^p .

A.3 Multivariate Normal Distribution

Recall that a multivariate normal distribution with mean $\boldsymbol{\mu}$ and with $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ can be written as

$$(A.5) \quad \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

Very often one works with quadratic functions of \mathbf{x} and wants to "complete the square" to rewrite such expressions as multivariate Gaussian distributions. By taking the natural logarithm of expression (A.5) we can then rearrange the terms in powers of \mathbf{x}

$$\begin{aligned} \ln(\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})) &= -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \det(\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\ &= -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \det(\boldsymbol{\Sigma}) - \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} + \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &= -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \det(\boldsymbol{\Sigma}) - \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &= -\frac{1}{2} (D \ln(2\pi) + \det(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) - \frac{1}{2} (-2 \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) - \frac{1}{2} (\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}) \\ &= -\frac{1}{2} (\mathcal{A}_0 + \mathbf{x}^\top \mathcal{A}_1 + \mathbf{x}^\top \mathcal{A}_3 \mathbf{x}) \end{aligned}$$

where

$$(A.6) \quad \mathcal{A}_0 := D \ln(2\pi) + \det(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

$$(A.7) \quad \mathcal{A}_1 := -2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

$$(A.8) \quad \mathcal{A}_2 := \boldsymbol{\Sigma}^{-1}$$

One can then use equations (A.6)-(A.8) to complete the square.

A.4 Information Theory

The Kullback-Leibler divergence of a continuous distribution $p(\boldsymbol{\theta} | \mathbf{x})$ with respect to a continuous distribution $q(\boldsymbol{\theta})$, denoted by $\text{KL}(p(\boldsymbol{\theta} | \mathbf{x}) || q(\boldsymbol{\theta}))$, can be interpreted as the amount of information lost when $q(\boldsymbol{\theta})$ is used to approximate $p(\boldsymbol{\theta} | \mathbf{x})$ [5]. It is defined as

$$(A.9) \quad \text{KL}(p(\boldsymbol{\theta} | \mathbf{x}) || q(\boldsymbol{\theta})) := \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{x})} \left[\ln \left(\frac{p(\boldsymbol{\theta} | \mathbf{x})}{q(\boldsymbol{\theta})} \right) \right] = \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{x}) \ln \left(\frac{p(\boldsymbol{\theta} | \mathbf{x})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}$$

Notice that if $p(\boldsymbol{\theta} | \mathbf{x})$ is our true posterior distribution, it will generally be unknown, which means that the integration above is intractable. Deterministic methods for approximating $p(\boldsymbol{\theta} | \mathbf{x})$ by minimizing (A.9) are generally called Expectation Propagation methods, and will not be covered in this thesis. A much simpler problem to solve, is to minimize the reversed KL-divergence $\text{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} | \mathbf{x}))$. In what follows we show that the KL-divergence is non-negative.

A.4.1 Non-negativity Property

One can use the power series definition of e^x to show

$$\begin{aligned} e^{x-1} &:= \sum_{k=1}^{\infty} \frac{(x-1)^k}{k!} \\ &= 1 + (x-1) + \frac{(x-1)^2}{2} + \dots \\ &= x + \frac{(x-1)^2}{2} + \dots \\ &\geq x \end{aligned} \quad \text{for } x > 0$$

where equality $e^{x-1} = x$ holds for $x = 1$. Since also the logarithm is a strictly increasing function, one has

$$(A.10) \quad x - 1 \geq \ln(x) \quad \forall x > 0$$

Now consider the negative KL-divergence

$$\begin{aligned} -\text{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} | \mathbf{x})) &= - \int_{\Theta} q(\boldsymbol{\theta}) \ln \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} | \mathbf{x})} \right) d\boldsymbol{\theta} \\ &= \int_{\Theta} q(\boldsymbol{\theta}) \ln \left(\frac{p(\boldsymbol{\theta} | \mathbf{x})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} \\ &\leq \int_{\Theta} q(\boldsymbol{\theta}) \left(\frac{p(\boldsymbol{\theta} | \mathbf{x})}{q(\boldsymbol{\theta})} - 1 \right) d\boldsymbol{\theta} && \text{using (A.10)} \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{x}) - q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta} - \int_{\Theta} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

so that $\text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{x})) \geq 0$ with equality achieved only when $\frac{p(\boldsymbol{\theta} \mid \mathbf{x})}{q(\boldsymbol{\theta})} = 1$ or equivalently $p(\boldsymbol{\theta} \mid \mathbf{x}) = q(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \Theta$.

A.5 Details of Local Variational Methods

A.5.1 Optimization of Parameters: EM Algorithm

The EM algorithm is a procedure that can be used to find maximum likelihood estimates of a parameter $\boldsymbol{\xi}$ governing a model of observed¹ data $\mathbf{Y} = \mathbf{y}$ and unobserved or latent variables $\boldsymbol{\beta}$. Namely, suppose we aim to find the maximum likelihood estimate $\hat{\boldsymbol{\xi}}_{\text{MLE}}$ maximizing the marginal likelihood for observed data given below

$$(A.11) \quad \mathcal{L}(\boldsymbol{\xi}; \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\xi}) = \int_{\mathbb{R}^p} p(\mathbf{y}, \boldsymbol{\beta} \mid \boldsymbol{\xi}) d\boldsymbol{\beta}$$

which is found by marginalizing the complete-data log-likelihood $p(\mathbf{y}, \boldsymbol{\beta} \mid \boldsymbol{\xi})$ over the latent variables $\boldsymbol{\beta}$. The EM-algorithm then proceeds as follows

Algorithm 4: EM Algorithm

- 1 Initialize maximum likelihood estimate $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^p$ with a guess.
- 2 Until convergence:
- 3 **E-step:** Define the function $Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)})$ as the expected value of complete-data log-likelihood $\ln(p(\mathbf{y}, \boldsymbol{\beta} \mid \boldsymbol{\xi}))$ with respect to $p(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\xi}^{(t)})$, the conditional distribution over the latent variables, given the observed data \mathbf{y} and the current estimate of the maximum likelihood estimator $\boldsymbol{\xi}^{(t)}$.

$$Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)}) := \mathbb{E}_{p(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\ln(p(\mathbf{y}, \boldsymbol{\beta} \mid \boldsymbol{\xi}))]$$

- 4 **M-step:** Maximize the function $Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)})$ with respect to $\boldsymbol{\xi}$ to find the new estimate of the maximum likelihood estimator $\boldsymbol{\xi}^{(t+1)}$

$$\boldsymbol{\xi}^{(t+1)} := \arg \max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi} \mid \boldsymbol{\xi}^{(t)})$$

end

Notice that the marginal likelihood defined in (A.11) is nothing more than $\tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi}) p(\boldsymbol{\beta})$ encountered in Section 3.2.2 where $\tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi})$ is defined in equation (3.12). Also notice that in our discussion of local variational methods, we state the our goal is to maximize the bound on the marginal likelihood given by

$$p(\mathbf{y}) \geq p(\mathbf{y} \mid \boldsymbol{\xi}) := \int_{\mathbb{R}^p} \tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} \int_{\mathbb{R}^p} \tilde{p}(\mathbf{y} \mid \boldsymbol{\beta}; \boldsymbol{\xi}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

¹Recall from Appendix A.2 that we consider our explanatory variables $\mathbf{X}_D^{\text{obs}}$ as being fixed constants. By observed data we mean the observations of the random variable \mathbf{Y} of response variables.

with respect to the variational parameters ξ . Since logarithms are strictly increasing we have that $p(\mathbf{y}) \geq p(\mathbf{y} | \xi)$ implies

$$\ln(p(\mathbf{y})) \geq \ln(p(\mathbf{y} | \xi)) = \ln\left(\int_{\mathbb{R}^p} \tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \xi) p(\boldsymbol{\beta}) d\boldsymbol{\beta}\right) =: \mathcal{B}(\xi)$$

where $\mathcal{B}(\xi)$ was defined in the same way in equation (3.17).

It follows, that we can perform local variational inference and find the values of ξ maximizing the evidence lower bound by performing the EM-algorithm described in Algorithm 4 as this will deliver the value of ξ maximizing the marginal likelihood $p(\mathbf{y} | \xi)$, and since the marginal likelihood $p(\mathbf{y} | \xi)$ is exactly our bound on the model evidence $p(\mathbf{y})$, the two methods will be equivalent. Notice we will use the variational approximate posterior $q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} | \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$ as the conditional distribution for the Expectation, where its mean and variance-covariance matrix are given by (3.16) and (3.15) respectively.

Algorithm 4 can be simplified because, in this case, it is easy to find an expression for $Q(\xi | \xi^{(t)})$ and maximizing it with respect to ξ_i . Let us rewrite $Q(\xi | \xi^{(t)})$ as follows

$$\begin{aligned} Q(\xi | \xi^{(t)}) &= \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\ln(p(\mathbf{y}, \boldsymbol{\beta} | \xi))] \\ &= \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\ln(\tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \xi) p(\boldsymbol{\beta}))] \\ &= \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\ln(\tilde{p}(\mathbf{y} | \boldsymbol{\beta}; \xi))] + C && \text{no } \xi \text{ dependence} \\ &= \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} \left[\sum_{i=1}^n \left(\ln(\sigma(\xi_i)) + \mathbf{x}_i^\top \boldsymbol{\beta} y_i - \frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \right] + C && \text{by (3.12)} \\ &= \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} \left[\sum_{i=1}^n \left(\ln(\sigma(\xi_i)) + \boldsymbol{\beta}^\top \mathbf{x}_i y_i - \frac{\boldsymbol{\beta}^\top \mathbf{x}_i + \xi_i}{2} - \lambda(\xi_i) (\mathbf{x}_i^\top \boldsymbol{\beta} \boldsymbol{\beta}^\top \mathbf{x}_i - \xi_i^2) \right) \right] + C \\ &= \sum_{i=1}^n \left(\ln(\sigma(\xi_i)) + \boldsymbol{\mu}_v^\top \mathbf{x}_i \left(y_i - \frac{1}{2} \right) - \frac{\xi_i}{2} - \lambda(\xi_i) \mathbf{x}_i^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_i + \lambda(\xi_i) \xi_i^2 \right) + C && \text{by (3.14)} \\ &= \sum_{i=1}^n \left(\ln(\sigma(\xi_i)) - \frac{\xi_i}{2} - \lambda(\xi_i) \mathbf{x}_i^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_i + \lambda(\xi_i) \xi_i^2 \right) + C \end{aligned}$$

where C at each step we have defined C to be a constant absorbing all terms not depending on ξ .

Now we can take the derivative of this expression with respect to ξ_k for $k \in \{1, \dots, n\}$

$$\begin{aligned} \frac{\partial Q(\xi | \xi^{(t)})}{\partial \xi_k} &= \frac{\partial}{\partial \xi_k} \left(\sum_{i=1}^n \left(\ln(\sigma(\xi_i)) - \frac{\xi_i}{2} - \lambda(\xi_i) \mathbf{x}_i^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_i + \lambda(\xi_i) \xi_i^2 \right) + C \right) \\ &= \frac{\partial}{\partial \xi_k} \left(\ln(\sigma(\xi_k)) - \frac{\xi_k}{2} - \lambda(\xi_k) \mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k + \lambda(\xi_k) \xi_k^2 \right) \\ &= \frac{1}{\sigma(\xi_k)} \sigma'(\xi_k) (1 - \sigma(\xi_k)) - \frac{1}{2} - \lambda'(\xi_k) \mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k + \lambda'(\xi_k) \xi_k^2 + 2\lambda(\xi_k) \xi_k && \text{by (3.3)} \\ &= \frac{1}{2} - \sigma(\xi_k) - \lambda'(\xi_k) \mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k + \lambda'(\xi_k) \xi_k^2 + 2\xi_k \frac{1}{2\xi_k} \left(\sigma(\xi_k) - \frac{1}{2} \right) && \text{by (3.10)} \\ \text{(A.12)} \quad &= -\lambda'(\xi_k) \left(\mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \xi^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k - \xi_k^2 \right) \end{aligned}$$

We can now set this equal to zero to find an update formula for ξ_k . Before doing so, we can show the following facts:

- The bound found by Jaakkola and Jordan and given in equation (3.9) is symmetric around 0.
- The function $\lambda(\xi_i)$ is strictly decreasing for $\xi_i > 0$.

First, we can notice that the Sigmoid function has the following useful property

$$(A.13) \quad \sigma(-x) = 1 - \sigma(x)$$

We can then use it to show that $\lambda(\xi_i)$ is symmetric around zero:

$$(A.14) \quad \begin{aligned} \lambda(-\xi_i) &= \frac{1}{2(-\xi_i)} \left(\sigma(-\xi_i) - \frac{1}{2} \right) \\ &= -\frac{1}{2\xi_i} \left(1 - \sigma(\xi_i) - \frac{1}{2} \right) \\ &= \frac{1}{2\xi_i} \left(\sigma(\xi_i) - \frac{1}{2} \right) \\ &= \lambda(\xi_i) \end{aligned}$$

Finally, we can show that bound (3.9) is indeed symmetric around $\xi_i = 0$

$$\begin{aligned} \sigma(-\xi_i) \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} - (-\xi_i)}{2} - \lambda(-\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) &= \sigma(-\xi_i) \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \quad \text{by (A.14)} \\ &= \frac{e^{-\xi_i}}{1 + e^{-\xi_i}} \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} + \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \quad \text{def. of } \sigma(\xi_i) \\ &= \frac{1}{1 + e^{-\xi_i}} \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} - \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \\ &= \frac{e^{\xi_i}}{1 + e^{\xi_i}} \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} - \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \\ &= \sigma(\xi_i) \exp \left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta} - \xi_i}{2} - \lambda(\xi_i) ((\mathbf{x}_i^\top \boldsymbol{\beta})^2 - \xi_i^2) \right) \end{aligned}$$

which finishes the proof. One can also inspect the following bound graphically as both ξ_i and $\mathbf{x}_i^\top \boldsymbol{\beta}$ change, as illustrated by Figure A.1. Since the bound is symmetric, we can restrict ourselves to considering values of \mathbf{x}_i that are non-negative.

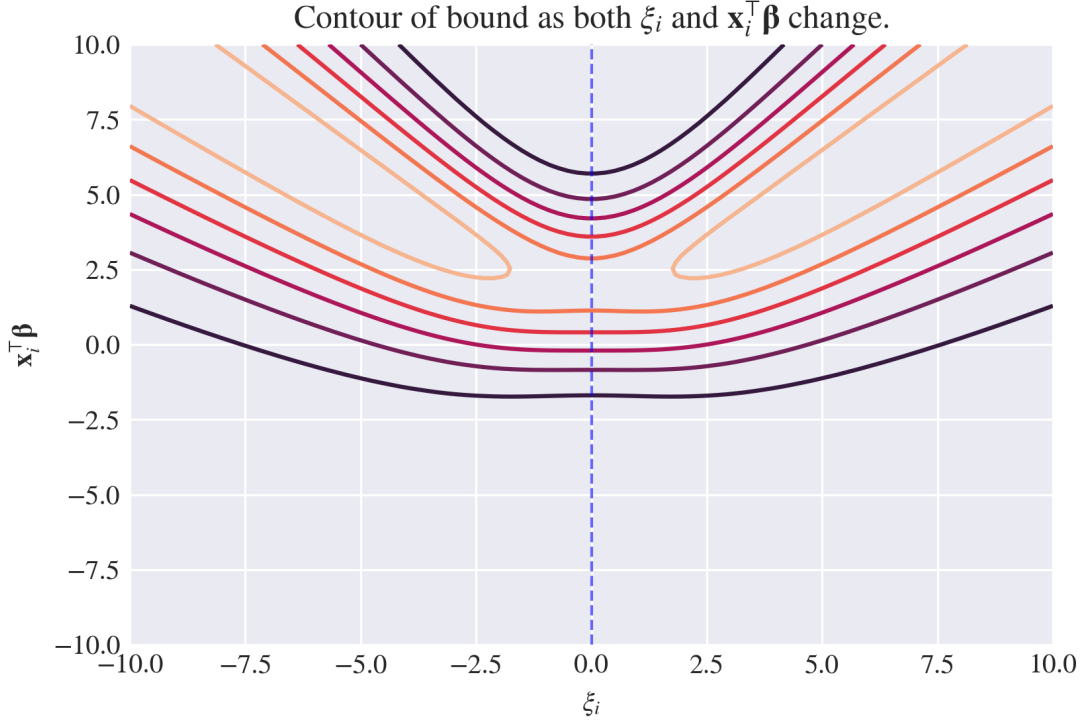


Figure A.1: Illustration of Symmetry for variational bound

Now consider $\xi_i > 0$. We can take the derivative of $\lambda(\xi_i)$ and show that it is strictly negative for all $\xi_i > 0$, and thus it is a strictly decreasing function. Indeed we have

$$\begin{aligned}
 \frac{d\lambda(\xi_i)}{d\xi_i} &= \frac{d}{d\xi_i} \left(\frac{1}{2\xi_i} \left(\sigma(\xi_i) - \frac{1}{2} \right) \right) \\
 &= -\frac{1}{2\xi_i^2} \left(\frac{e^{\xi_i}}{1+e^{\xi_i}} - \frac{1}{2} \right) + \frac{1}{2\xi_i} \left(\frac{e^{\xi_i}(1+e^{\xi_i}) - e^{2\xi_i}}{(1+e^{\xi_i})^2} \right) \\
 &= -\frac{1}{4\xi_i^2} \left(\frac{e^{\xi_i} - 1}{1+e^{\xi_i}} \right) + \frac{e^{\xi_i}}{2\xi_i(1+e^{\xi_i})^2} \\
 &= \frac{-e^{2\xi_i} + 1 + 2\xi_i e^{\xi_i}}{4\xi_i^2(1+e^{\xi_i})^2}
 \end{aligned}$$

Now consider the numerator and the series expansion of the exponential function:

$$\begin{aligned}
 1 + 2\xi_i \sum_{k=0}^{\infty} \frac{\xi_i^k}{k!} - \sum_{k=0}^{\infty} \frac{(2\xi_i)^k}{k!} &= 1 + \sum_{k=0}^{\infty} \frac{2\xi_i^{k+1}}{k!} - 1 - \sum_{k=1}^{\infty} \frac{2^k \xi_i^k}{k!} \\
 &= \sum_{k=0}^{\infty} \frac{2\xi_i^{k+1}}{k!} - \sum_{k=1}^{\infty} \frac{2^k \xi_i^k}{k!} \\
 &= \sum_{k=0}^{\infty} \frac{2\xi_i^{k+1}}{k!} - \sum_{j=0}^{\infty} \frac{2^{j+1} \xi_i^{j+1}}{(j+1)!} \quad \text{define } j := k-1 \\
 &= \sum_{k=0}^{\infty} \frac{2\xi_i^{k+1}}{k!} \left(1 - \frac{2^k}{k+1}\right) \\
 &< 0 \quad \text{By Bernoulli inequality, since } \xi_i > 0
 \end{aligned}$$

Therefore $\lambda'(\xi_i) < 0$ for $\xi_i > 0$, and in particular $\lambda'(\xi_i) \neq 0$ for $\xi_i > 0$. It follows that

$$\frac{\partial Q(\boldsymbol{\xi} | \boldsymbol{\xi}^{(t)})}{\partial \xi_k} = 0 \quad \implies \quad \xi_k^2 = \mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k$$

Since the bound is symmetric around 0 it follows that we can consider the positive square root so that the updated parameter is given by

$$\xi_k^{(t+1)} = \sqrt{\mathbf{x}_k^\top \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] \mathbf{x}_k}$$

Now we can find the expectation above by recalling our variational distribution in (3.14) and by using

$$\text{Var}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}] = \mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] - \left(\mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}]\right)^2$$

to obtain

$$\mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}^\top \boldsymbol{\beta}] = \text{Var}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}] + \left(\mathbb{E}_{p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\xi}^{(t)})} [\boldsymbol{\beta}]\right)^2 = \boldsymbol{\Sigma}_v + \boldsymbol{\mu}_v \boldsymbol{\mu}_v^\top$$

which immediately leads to a computable update formula for the variational parameters

$$\xi_k^{(t+1)} = \sqrt{\mathbf{x}_k^\top (\boldsymbol{\Sigma}_v + \boldsymbol{\mu}_v \boldsymbol{\mu}_v^\top) \mathbf{x}_k}$$

Putting everything together we obtain Algorithm 3.



APPENDIX B - COMPUTING

Begins an appendix some some some some some some some some some some some
some some some some some some some some some some some some some some
some some some some some some some some some some some some some some
some

B.1 some

BIBLIOGRAPHY

- [1] Christophe Andrieu, Nando De Freitas, and et al.
An introduction to mcmc for machine learning, 2003.
- [2] John E. Angus.
The probability integral transform and related results.
SIAM Review, 36(4):652–654, 1994.
ISSN 00361445.
URL <http://www.jstor.org/stable/2132726>.
- [3] Christopher M. Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.
- [4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe.
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859–877, 2017.
doi: 10.1080/01621459.2017.1285773.
URL <https://doi.org/10.1080/01621459.2017.1285773>.
- [5] Anderson D. R Burnham, K. P.
Model Selection and Multimodel Inference, page 51.
Springer, 2nd edition, 2002.
ISBN 9780387953649.
- [6] Bernard D. Flury.
Acceptance–rejection sampling made easy.
Siam Review - SIAM REV, 32, 09 1990.
doi: 10.1137/1032082.
- [7] Luc Devroye.
Non-uniform random variate generation.
1986.
- [8] D. Gamerman and H.F. Lopes.

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition.

Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2006.

ISBN 9781584885870.

URL https://books.google.co.uk/books?id=yPvECi_L3bwC.

- [9] W.R. Gilks, S. Richardson, and D. Spiegelhalter.

Markov Chain Monte Carlo in Practice.

Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

ISBN 9780412055515.

URL http://books.google.com/books?id=TRXrMWY_i2IC.

- [10] Tommi S. Jaakkola and Michael I. Jordan.

A variational approach to bayesian logistic regression models and their extensions, 1996.

- [11] David J. C. MacKay.

Information Theory, Inference & Learning Algorithms.

Cambridge University Press, New York, NY, USA, 2002.

ISBN 0521642981.

- [12] Radford M Neal.

Probabilistic inference using markov chain monte carlo methods.

1993.

- [13] Christian P. Robert and George Casella.

Monte Carlo Statistical Methods (Springer Texts in Statistics).

Springer-Verlag, Berlin, Heidelberg, 2005.

ISBN 0387212396.