Geometric Markov Chain Monte Carlo

Pennington, Thomas

Award date:
2022

Awarding institution:
University of Bath

Geometric Markov Chain Monte Carlo

submitted by

Thomas Lloyd Pennington

for the degree of Doctor of Philosophy

of the

University of Bath
Department of Mathematical Sciences

July 2022

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Signature of Author: Thomas Lloyd Pennington

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally.

Signature of Author: Thomas Lloyd Pennington

Contents

Nomenclature

Superscripts

A^{-}	Moore-Penrose pseudo-inverse of matrix A
A^{T}, q^{T}	Transpose of matrix A, vector q

Other symbols

$0_{p\ p}$	The $p \times p$ matrix of all zero entries
$1_{p\ p}$	The $p \times p$ identity matrix
$x \cdot y$	The Euclidean dot product; equivalent to $x^{T}y$
N	Number of Monte Carlo observations, i.e. Monte Carlo sample size
n	Number of observations, i.e. sample size
p	Number of parameters; dimension of Q

Spaces

Y	Set or space of possible values for observational data
Q	Statistical family
Q	Natural parameter space; open subset of R^{p}

List of Algorithms

List of Figures

Thesis Outline

Bayesian statistics treats the parameters of a data generation process as random variables to describe their uncertainty before and after data are observed. Prior information, which could include information from previous observations or expert opinion, is encoded in the form of a prior distribution on the process parameters. This distribution is updated to a posterior distributionby conditioning on new observations via Bayes' Theorem. One may be able to write down the posterior density up to a constant of proportionality, but often cannot directly compute expectations with respect to the posterior measure. In applications, posterior expectations are required to compute almost all informative quantities including the posterior mean, covariance and predictive distributions for future data.

The widely-used solution to this problem is Markov Chain Monte Carlo (MCMC), which computes approximations to expected values $E_{p(q)}(f(q))$ of a function f with respect to a target distribution $p(q)$. A Markov Chain is constructed whose stationary distribution is the target, and the quantity of interest is averaged over a simulation of the chain. The average over the simulated chain is known as the Monte Carlo estimator, which is an asymptotically unbiased approximation to the true expectation. The variance of the Monte Carlo estimator, and thus the accuracy in the approximation, depends on the autocorrelation of the underlying Markov Chain. Hence Markov Chains whose autocorrelation quickly decay are desirable.

This thesis studies MCMC algorithms which make use of Differential Geometry on the parameter space of the model in order to more ef ciently sample from the posterior distribution. Geometric structures on statistical models are independent of parameterization, and so sampling algorithms de ned in terms of these structures have certain invariance properties under transformations of parameter space.

To demonstrate this point, the left panel of Figure 0.0.1 shows several steps of a Gibbs sampler on the joint probability density of two correlated random variables q_1, q_2. The algorithm samples q_1 conditionally on the last value of q_2 and vice versa, and hence the sequence of samples appears to move in parallel to a co-ordinate axis at each step with right-angles between the lines joining successive samples. For distributions with high correlation between variables, the Gibbs sampler is known to be inef cient because the variance of the conditional distributions is small, and thus the algorithm is constrained to take small steps. The performance of the sampler is improved under a reparameterization to a co-ordinate system which is less correlated.

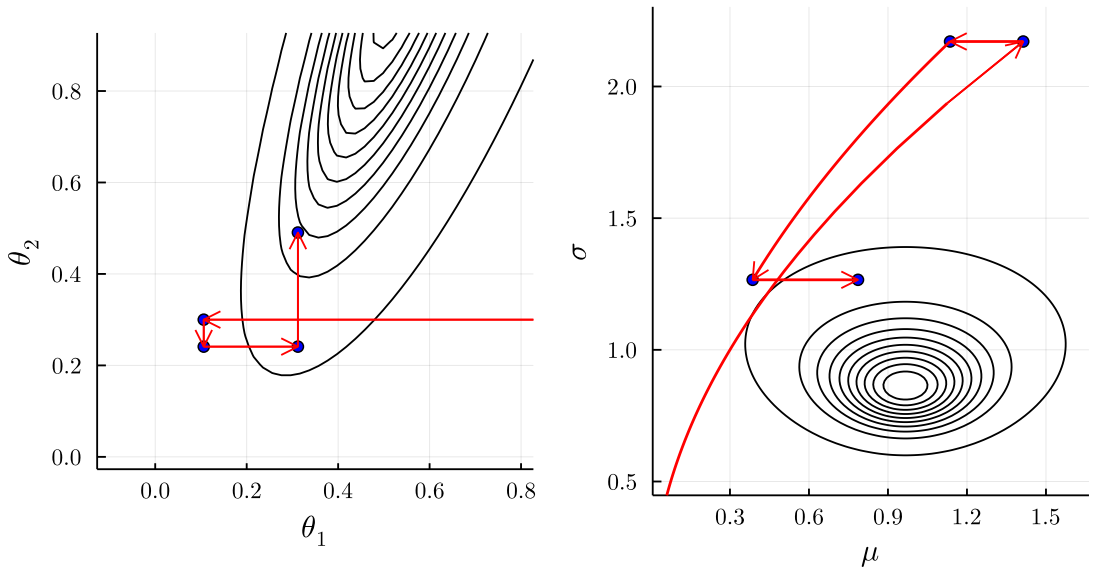

Figure 0.0.1: Left: Gibbs sampling from a bivariate target distribution. Right: The image of the samples under a nonlinear reparameterization with the transformed density. Contour lines roughly re ect boundaries between 10% quantile regions.

The right panel of Figure 0.0.1 shows the image of these samples and the lines joining them under a nonlinear reparameterization

$$m := \frac{q_2}{2q_1} \qquad s := \frac{\sqrt{1}}{2q_1}$$

Under this change of variable, the target density follows a transformation law involving the Jacobian determinant; see Corollary B.0.2 in Appendix B. As expected, the gure shows that in the new parameterization some lines joining successive samples are not straight and never meet each other perpendicularly. It seems therefore that straightness and orthogonality depend on co-ordinates.

In fact, Differential Geometry provides rigorous de nitions of orthogonality and straightness without reference to any particular co-ordinate system, and are consistent even under nonlinear changes of variable. These are respectively given by the Riemannian metric tensor and af ne connections. The thesis will develop sampling algorithms which move along geodesic curves, i.e. straight as determined by the af ne connections, and intersect orthogonally with respect to a particular Riemannian metric. Previous authors have used similar ideas to develop successful geometric sampling al-

gorithms; for example, the Riemannian variants of Hamiltonian Monte Carlo (HMC) and the Metropolis-Adjusted Langevin Algorithm (MALA) introduced by Girolami and Calderhead [39] often outperform their non-Riemannian versions, especially for hierarchical models with high curvature [11].

The algorithms developed in this thesis differ from previous work by making use of the dual geometry of statistical models studied in the field of Information Geometry. Whereas previous geometric sampling algorithms only use the Levi-Civita connection which is entirely determined by the Riemannian metric, the novel algorithms presented in this work use two coupled affine connections. There are therefore two ways of measuring flatness or curvature, arising from asymmetry in divergence functions such as Kullback-Leibler. The inherent dual geometry of a statistical model contains no information from any prior distribution, and perhaps for this reason has not been previously studied for use in Bayesian inference.

The thesis is divided into four parts. Part I discusses geometry for Bayesian inference; Chapter 1 shows how statistical families may be given a smooth manifold structure and recasts Bayesian inference in terms of measures on such manifolds. Chapter 2 reviews Differential and Information Geometry in detail, discussing how a metric tensor and connections are derived from a divergence function between probability distributions. Chapter 3 presents a novel dual geometry modified to include both prior and likelihood information for use in sampling, and discusses partitioning of parameter space into orthogonal submanifolds.

Part II concerns the use of geometry in sampling algorithms; Chapter 4 reviews MCMC, particularly those algorithms which resample the target distribution within a submanifold such as the Gibbs sampler and Hamiltonian Monte Carlo. Chapter 5 then presents two novel sampling algorithms called *Orthogonal Gibbs* and *Orthogonal Gradient*, analyzes their behaviour in certain cases and presents numerical results for several example models. Finally, Part III contains concluding remarks and Part IV is the collection of appendices.

List of Contributions

To my knowledge, the following are all novel contributions.

Chapter 1 gives a measure-theoretic co-ordinate free treatment of Bayes' theorem in Section 1.1.3. Although the discussion is fairly straightforward once all the terms are introduced in co-ordinate free form, I was not able to find a version of Bayes' theorem on a manifold in the literature.

Chapter 2 is mostly a review of differential and Information Geometry, but also includes some novel results.

- In Theorem 2.5.3, I show that the dual geometry arising from a divergence function restricted to a submanifold is exactly the dual geometry inherited from the ambient space in which the submanifold is embedded.
- If a dually-flat geometry is generated by a Bregman generator F in a global co-ordinate chart, Theorem 2.5.8 shows that a submanifold is e-flat, i.e. autoparallel with respect to the primal connection, if and only if the Bregman divergence in the submanifold generated by the restriction of F corresponds with the restricted Bregman divergence.

Chapter 3 suggests a dually-flat geometry for use in MCMC algorithms, called the Posterior Bregman Geometry, defined for any log-concave probability density. In some cases, the geometry has the same Riemannian metric as that suggested by Girolami and Calderhead [39].

- Proposition 3.1.2 shows that for exponential family models, the posterior Bregman divergence is given by the prior Bregman divergence plus the Kullback-Leibler divergence.
- Theorem 3.1.3 states that under weak conditions, log-concave densities can be written as $\exp(-D(q, q_0))$ up to a multiplicative constant where D is the p-Bregman divergence and q_0 is the mode when it exists.
- In Section 3.3.3, I generalize the mixed co-ordinates of Amari [3], [2] such that the corresponding foliations are general hyperplanes in primal and dual co-ordinates.

Theorem 3.3.11 shows that the primal component of generalized mixed coordinates is a primally-affine co-ordinate system for the dually-flat geometry within each fiber of the m-foliation. Furthermore, Theorem 3.3.12 gives an explicit form for the Bregman generator within the m-flat submanifold or fiber.

Chapter 5 introduces novel sampling algorithms which reduce to well-known methods in specific cases.

Algorithm 4 is a general form of novel Complementary Dual Submanifold algorithms that sample a target probability distribution conditioned on Riemannian-orthogonal e-flat and m-flat submanifolds. The algorithms recurse within the m-flat submanifold, making use of novel results from Chapter 3.

Theorem 5.1.1 and Theorem 5.1.2 give explicit disintegrations of the target density over the general e- and m-foliations. Furthermore, the general algorithm is shown to preserve the target distribution in Theorem 5.1.6.

Section 5.2 describes a specific case called Orthogonal Gibbs. Complementary primal and dual variables are recursively sampled conditionally on each other in a block Gibbs algorithm. Proposition 5.2.1 shows that Orthogonal Gibbs reduces to the regular deterministic-scan Gibbs sampler when the Euclidean divergence function is used.

In Section 5.2.2 and Appendix D, I show that for a multivariate Gaussian target the Orthogonal Gibbs algorithm is equivalent to the Gaussian Elimination solver for linear systems with a stochastic term in the update to the solution. Corollary 5.2.3 states that this sampling analogue to Gaussian Elimination produces independent samples from the Gaussian target distribution.

Section 5.3 describes another specific case of the general algorithm in which one samples the target distribution conditioned onto an e-geodesic in the direction of the negative log-density of the target, and then from the conditional on the orthogonal m-flat submanifold. I call this algorithm Orthogonal Gradient.

In Section 5.3 and Appendix D, I show that for a multivariate Gaussian target, the Orthogonal Gradient algorithm is equivalent to the Conjugate Gradient solver for linear systems, sampling along search directions instead of minimizing a cost function.

Part I

Geometry for Bayesian Inference

Chapter 1

Bayesian Inference and Statistical Families

This chapter recasts Bayesian inference in terms of measures on families of data-generating distributions, which will be treated as smooth manifolds. The special case of exponential families is examined, which are ubiquitous in statistical models, with several examples of Bayesian inference problems which will serve as test cases for MCMC later in the thesis.

1.1 Bayesian Inference in the Geometric Picture

We begin by briefly describing Bayesian inference and fixing notation to be used throughout the thesis.

In their book Bayesian Data Analysis [33], Gelman et al divide Bayesian analysis into three steps:

1. A joint probability model over observable and unobservable variables is constructed. Let $\theta = (\theta_i)_{i=1}^{p}$ be the p unknowns or unobservables, and $y = (y_j)_{j=1}^{n}$ be the n known or observable variables. The joint probability model is written

$$p(y_1, .., y_n, \theta_1, .., \theta_p) \quad p(y, \theta)$$

   We use $p(\,)$ to denote marginal probability densities or probability mass functions, and $p(\,|\,)$ for conditional densities or mass functions, with respect to some base measure. The probability model describes relationships between variables according to (e.g.) scientific theories, and also encodes one's uncertainty about the value of each of the variables. In particular, the conditional density $p(y|\theta)$ represents the distribution on observable data conditioned on unknown parameters θ, and the prior $p(\theta)$ is a probabilistic representation of uncertainty in the unknowns. By elementary probability theory, the joint distribution $p(y, \theta)$ is simply the product $p(y|\theta)p(\theta)$, so these two densities fully specify a joint probability model assuming they exist with respect to specified dominating measures.

2. The joint probability model is conditioned on the measured values of observables $y = y_{obs}$. Assuming the relevant densities exist, this is described by Bayes' theorem,

$$p(q|y = y_{obs}) = \frac{p(y_{obs}, q)}{p(y_{obs})} = \frac{p(y_{obs}|q)p(q)}{\int p(y_{obs}|q')p(q')dq'} \tag{1.1.1}$$

This conditional distribution of unknowns given knowns is called the posterior distribution. It represents one's updated uncertainty, or knowledge, about q given observed data y_{obs}. The likelihood is the density $p(y_{obs}|q)$ thought of purely as a function of q with the observed value of y_{obs} plugged in. The normalizing constant $p(y_{obs})$ is often a high-dimensional integral, and hence it is usually only possible to compute the posterior up to a constant of proportionality,

$$p(q|y_{obs}) \propto p(y_{obs}, q) = p(y_{obs}|q)p(q)$$

assuming the prior and likelihood can themselves be evaluated up to multiplicative constants.

3. The fit of the model to the data is assessed and inferences are drawn using the posterior distribution. To assess model fit, it is useful to predict new data based on the updated unknowns, and check how similar predicted and observed data are. This is achieved using the posterior predictive distribution,

$$p(y|y_{obs}) = \int p(y|q')p(q'|y_{obs})dq'$$

where y is a prediction for new data. One could compare a histogram of simulated data from the posterior predictive with actual observed data, for example - such diagnostics are known as posterior predictive checks [33].

If the model is deemed reasonable, one may wish to compute quantities of interest based on the posterior. These are expectations of the form

$$E_{p(q|y_{obs})}[f(q)] = \int f(q)p(q|y_{obs})dq \tag{1.1.2}$$

i.e. an average of the functional f with respect to the posterior distribution. For example, our interest may be in $E(q_i)$ for some $i \in \{1, .., p\}$, the mean of q_i with respect to the posterior, which represents a "best guess" or point estimate of the unknown variable q_i given observed data. Another example is given by the marginal variance of q_i,

$$Var_{p(q|y_{obs})}[q_i] := E_{p(q|y_{obs})}\left[q_i^2\right] - E_{p(q|y_{obs})}[q_i]^2$$

which can be interpreted as a measure of uncertainty in the value of q_i.

The primary computational challenge in Bayesian inference is in computing quantities of interest (1.1.2). This thesis studies Markov Chain Monte Carlo algorithms, which approximate quantities of interest by taking the sample mean of a set of simulated draws from the posterior distribution.

1.1.1 Reparameterization and Statistical Families

We have described Bayesian inference in terms of conditional densities on unknown variables given known variables. Here we introduce a different view of Bayesian inference that better suits the geometric picture; the prior and posterior are probability measures on *statistical families* of data-generating distributions, which can be viewed as smooth manifolds, i.e. generalizations of smooth surfaces in $\mathbb{R}^n$.

Firstly, we note that the probability model can be reparameterized by a transformation of the known or unknown variables. The following is a formal definition of the concept of a change of variable, which we adapt from Munkres [54].

Definition 1.1.1. Let $\Theta, X \subseteq \mathbb{R}^p$ be open subsets and $h : \Theta \to X$ be a bijection such that both h and its inverse h^{-1} are C^∞-smooth, i.e. its partial derivatives of all orders exist and are continuous on Θ. Then h is called a *diffeomorphism*.

Consider a transformation on the unknown variables θ of a probability model,

$$\theta \to x := h(\theta), \qquad h : \Theta \to X$$

where h is a diffeomorphism. In this context $\Theta, X \subseteq \mathbb{R}^p$ are open subsets called *parameter spaces*, sets of admissible values for θ and x respectively. The thesis will not consider countable or finite parameter spaces. Suppose further that data can take values in a set $Y \subseteq \mathbb{R}^n$. Under the transformation h, the conditional density of data given parameters transforms as a scalar function of θ, so that the conditional density given x evaluated at observed data $y = y_{obs}$ is

$$p(y_{obs}|x) := p(y_{obs}|\theta)|_{\theta = h^{-1}(x)} \qquad \text{for} \qquad x \in X \tag{1.1.3}$$

The marginal distribution $p(\theta)$, i.e. the prior, is a probability density in θ and therefore transforms according to the change of variable formula for probability densities, stated in Appendix B as Corollary B.0.2;

$$p(x) := \frac{p(\theta)}{|\det[Dh(\theta)]|}\Big|_{\theta = h^{-1}(x)} \qquad \text{for} \qquad x \in X \tag{1.1.4}$$

where $Dh(\theta)$ is the derivative matrix of h evaluated at θ, i.e. the matrix with (i,j)-entry

$$(Dh(\theta))_{ij} := \frac{\partial h_i}{\partial \theta_j}(\theta)$$

In measure-theoretic language, the transformed density (1.1.4) is the density of the pushed-forward prior distribution under the measurable function h, with respect to the Lebesgue measure on $\mathbb{R}^p$. The posterior distribution is represented by the probability density $p(\theta|y_{obs})$, which follows the same change of variable formula;

$$p(x|y_{obs}) = \frac{p(\theta|y_{obs})}{|\det[Dh(\theta)]|}\Big|_{\theta = h^{-1}(x)}$$

Substituting Bayes' rule (1.1.1) for $p(q|y_{obs})$ and the transformed prior (1.1.4) and conditional density (1.1.3), we find

$$p(x|y_{obs}) = \frac{p(y_{obs}|q)\,p(q)}{p(y_{obs})\,|\det[Dh(q)]|}\Big|_{q=h^{-1}(x)}$$
$$= \frac{p(y_{obs}|x)\,p(x)}{p(y_{obs})}$$

We see that Bayes' theorem holds in the transformed variable x. In other words, the process of Bayesian updating commutes with reparameterization of the model, so using notation such as $p(x|y_{obs})$ is justified. In conclusion, any diffeomorphism on the unknown variables is a reparameterization of a Bayesian model, and the model can therefore be represented in infinitely many equivalent parameterizations. We will therefore present a formal description of a Bayesian model which is independent of parameterization.

The observed variables $y = y_{obs}$ are a sample, i.e. a realization of a random variable with some unknown probability distribution q, called the population. One wishes to infer properties of q from the sample y. The space of all probability distributions over the data is usually too large to work with, and so statisticians work with restricted classes of distributions called "small worlds" [10] or families. The prior and posterior are then probability measures on these families, respectively representing information about the population before and after data are observed. This leads to the following formal definition:

Definition 1.1.2. Let $(Y, \mathcal{F})$ be a measurable space, where $Y \subseteq \mathbb{R}^n$ and $\mathcal{F}$ is a σ-algebra in Y. A statistical model or statistical family $\mathcal{Q}$ on $(Y, \mathcal{F})$ is a set of probability measures on $(Y, \mathcal{F})$.

In this picture, q indexes data-generating probability measures $q_q \in \mathcal{Q}$, where q ranges over some index set or parameter space. Reparameterization simply means a change of index, or a change of co-ordinates, to describe the family. For the prior and posterior to be represented in new co-ordinates, the co-ordinate change must be smooth and invertible with a smooth inverse, i.e. a diffeomorphism. We assert that inferences we draw about the population should not depend on parameterization; we are therefore interested in families which are invariant under diffeomorphisms. This is the study of differential geometry and smooth manifolds.

1.1.2 Statistical Families as Smooth Manifolds

We assume that all statistical families studied in the thesis are endowed with the structure of a smooth manifold, which will be defined in this subsection. The concept of a smooth manifold requires a topology. We will not be specific about what the topology is, and will simply assume that there is a given topology that makes a family $\mathcal{Q}$ into a topological p-manifold. We will comment later on how a statistical family may be given a topology and smooth manifold structure.

Before discussing geometry, we require the concept of a smooth manifold. This is the canvas on which geometry lives; it allows us to formally make sense of what it means for something to be independent of co-ordinates, and has the interpretation of being locally at so that small perturbations to a point are elements of a vector space.

Formally, a topological p-manifold Q is a topological space that is Hausdorff, second countable and locally Euclidean of dimension p [47]. Roughly, being Hausdorff and second countable means there are neither too many nor too few open sets. Locally Euclidean of dimension p means that every point in a manifold has a neighbourhood that is homeomorphic to an open set in R^p. An open subset $U \subset Q$ that is homeomorphic to an open subset $V \subset R^p$ is a co-ordinate domain, the homeomorphism $f : U \to V$ is a co-ordinate map, and the pair (U, f) is a co-ordinate chart. A collection of co-ordinate charts whose domains cover Q is called an atlas. Since manifolds are locally Euclidean of dimension p, every point in a p-manifold is contained in the domain of some co-ordinate chart, so an atlas must always exist.

In order to do calculus on a manifold, we need to make sense of when functions $f : Q \to R$ are differentiable or smooth. The following discussion follows Lee's Introduction to Smooth Manifolds [46], in which smoothness is de ned in terms of co-ordinate charts. Let (U, f) be a co-ordinate chart with $f : U \to V \subset R^p$ a co-ordinate map, and let $f : Q \to R$ be a scalar function on the manifold. Then f has a co-ordinate representation in the chart

$$f : V \to R, \qquad f(x) := f \circ f^{-1}(x) \text{ for } x \in V$$

Since f is a scalar function on R^p, its smoothness is well-de ned; f is C^∞-smooth, or simply smooth, if its partial derivatives of all orders exist and are continuous. So we could say that f is smooth if its co-ordinate representation in some chart is C^∞-smooth; however, this may depend on the choice of co-ordinate chart.

To overcome this ambiguity, suppose there exist two co-ordinate charts (U_1, f_1) and (U_2, f_2) such that their domain intersection $U_1 \cap U_2$ is non-empty. Then the composition

$$f_2 \circ f_1^{-1} : f_1(U_1 \cap U_2) \to f_2(U_1 \cap U_2)$$

maps co-ordinates represented in the rst chart to co-ordinates in the second and is called the transition map between the two co-ordinate charts. We de ne the charts to be smoothly compatible if $U_1 \cap U_2$ is empty or the transition map $f_2 \circ f_1^{-1}$ is a diffeomorphism. An atlas A is smooth if any two charts in A are smoothly compatible. A smooth atlas is said to be maximal if it is not properly contained in any larger smooth atlas. This just means that any chart that is smoothly compatible with every chart in A is already in A [46]. In fact, every smooth atlas is contained within a unique maximal smooth atlas, so it is enough to de ne a smooth atlas to know that a maximal one exists.

Finally we may de ne a smooth manifold. A smooth structure on a manifold is a maximal smooth atlas A, and a smooth manifold (Q, A) is simply a topological manifold Q equipped with a smooth structure A. This allows for a de nition of smooth functions $f : Q \to R$ on a smooth manifold (Q, A). A chart (U, f) is called smooth

when it belongs to the smooth structure. The function f is smooth if for every point $q \in Q$ a smooth chart $(U, f) \in \mathcal{A}$ exists whose co-ordinate domain U contains q, and the co-ordinate representation $f = f \circ f^{-1}$ is a smooth function on $\mathbb{R}^p$. The set of all smooth functions on Q is denoted by $C^\infty(Q)$.

Example 1.1.3. $\mathbb{R}^p$ is a smooth manifold when equipped with the Euclidean topology and a smooth structure containing charts of the form (U, I) where U is any open set and I is the identity map. Furthermore, any open subset U is a smooth manifold [46].

We may also define smoothness of maps between manifolds in a similar fashion; the following definition is from Chapter 2 of Lee [46]. Let Q, W be smooth manifolds and $F : Q \to W$. For any $q \in Q$, suppose there exist smooth charts (U, f_1) on Q with $q \in U$ and (V, f_2) on W such that $F(U) \subseteq V$. Note that $F := f_2 \circ F \circ f_1^{-1}$ is a representation of F in the given charts; $F(x)$ is the representation in the second chart of the image under F of the point whose representation in the first chart is x. We say F is smooth if for any $q \in Q$ there exist such a pair of smooth co-ordinate charts whose domains $U \subseteq Q$ and $V \subseteq W$ have $q \in U$ and $F(V) \subseteq W$, such that the co-ordinate representation $F : f_1(U) \to f_2(V)$ is smooth in the sense of real vector-valued functions. Finally, we can generalize Definition 1.1.1 by defining $F : Q \to W$ to be a diffeomorphism if it is bijective and smooth with a smooth inverse, in the sense of smooth functions between manifolds.

For brevity we will not explicitly refer to the formal structures defined above when they are implied from context. For example, we may treat a point $q \in Q$ and its co-ordinates $q = f(q)$ in some chart interchangeably, and keep in mind that the representation is local. In particular, we often talk of a point q where this can refer to a point in the manifold, a co-ordinate vector in $\mathbb{R}^p$, or both.

Parametric Families and The Smooth Manifold Chart Lemma

Our discussion of statistical families as smooth manifolds has thus far been abstract. In this section we will describe how statistical families can be given a smooth manifold structure from an existing parameterization.

Example 1.1.4. Let Q be the family of Gaussian probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where $\mathcal{B}(\mathbb{R})$ is the Borel s-algebra on $\mathbb{R}$,

$$Q = \left\{ \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}(x-m)^2 \right) d\lambda(x) : m \in \mathbb{R}, s^2 > 0 \right\},$$

where $d\lambda$ is the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

This example is typical of how statistical families are specified, i.e. a parameterized set of probability measures. Although a smooth manifold structure is not yet specified in this example, clearly the parameterization used to define the family is similar to a global smooth co-ordinate chart. We will show that there is a unique smooth manifold structure for which the parameterization provides a global smooth co-ordinate chart.

In general, we have the following definition from Shao [72] for families defined in this way.

Definition 1.1.5. A set of probability measures $\mathcal{Q} = \{\mathbb{Q}_q : q \in Q\}$ all on the same measurable space $(\mathcal{Y}, \mathcal{F})$ is a parametric family if the parameter space $Q \subseteq \mathbb{R}^p$ and each $\mathbb{Q}_q$ is a known probability measure whenever $q \in Q$ is known.

In Example 1.1.4, the parameter space is the open upper-half plane $Q = \{(m, s^2) : m \in \mathbb{R}, s^2 > 0\} \subseteq \mathbb{R}^2$, and clearly $\mathcal{Q}$ is a parametric family. For the parameterization to provide a valid co-ordinate chart which must be bijective, we clearly require the following property which is also adapted from Shao [72].

Definition 1.1.6. A parametric family $\mathcal{Q} = \{\mathbb{Q}_q : q \in Q\}$ is called identifiable if for any $q_1, q_2 \in Q$ such that $q_1 \neq q_2$, we have $\mathbb{Q}_{q_1} \neq \mathbb{Q}_{q_2}$.

Continuing the example, the parametric family $\mathcal{Q}$ of Gaussians is clearly identifiable. In order to show identifiable parametric families are smooth manifolds, we require the following result which is a simplification of the Smooth Manifold Chart Lemma given as Lemma 1.35 in Lee [46].

Lemma 1.1.7. Let Q be a set, and let $f : Q \to \mathbb{R}^p$ be a bijection between Q and an open subset $f(Q) \subseteq \mathbb{R}^p$. Then Q has a unique smooth p-manifold structure such that (Q, f) is a smooth chart.

Our main result in this section is the following, which essentially states that identifiable parametric families are naturally smooth manifolds with a global co-ordinate chart.

Theorem 1.1.8. Let $\mathcal{Q} = \{\mathbb{Q}_q : q \in Q\}$ be an identifiable parametric family such that $Q \subseteq \mathbb{R}^p$ is open. Define the mapping

$$f : Q \to \mathbb{R}^p, \qquad f(\mathbb{Q}_q) := q$$

then Q has a unique smooth p-manifold structure for which (Q, f) is a smooth chart.

Proof. We will show that f satisfies the conditions of Lemma 1.1.7. Firstly, $f(Q) = Q$ is an open subset of $\mathbb{R}^p$ by the assumptions of the theorem. It only remains to show that f is a bijection from Q to Q.

For surjectivity, for any $q \in Q$ there exists $\mathbb{Q}_q \in \mathcal{Q}$ such that $f(\mathbb{Q}_q) = q$. Hence $f : Q \to Q$ is surjective.

For injectivity, suppose $f(\mathbb{Q}_{q_1}) = f(\mathbb{Q}_{q_2})$ for some $\mathbb{Q}_{q_1}, \mathbb{Q}_{q_2} \in \mathcal{Q}$. By identifiability, it clearly follows that $q_1 = q_2$. Hence f is injective.

In conclusion, $f : Q \to Q$ is a bijection where $Q \subseteq \mathbb{R}^p$ is open. Thus by Lemma 1.1.7, Q has a unique smooth p-manifold structure in which (Q, f) is a smooth chart. □

Example 1.1.9. Continuing Example 1.1.4, the parametric family is

$$Q = \left\{ \frac{1}{\sqrt{2\pi s^2}} \exp\left[-\frac{1}{2s^2}(x-m)^2 \right] d\lambda(x) : m \in \mathbb{R}, s^2 > 0 \right\},$$

The parameter space of the family $Q = \{(m, s^2) : m \in \mathbb{R}, s^2 > 0\}$ is an open subset of $\mathbb{R}^2$, namely the open upper-half plane. Furthermore, the family is clearly identifiable. By Theorem 1.1.8, Q has a unique smooth manifold structure for which the mean and variance $\varphi_1 = (m, s^2)$ define a co-ordinate chart.

There is no canonical reason to parameterize in terms of mean and variance. We can also define a co-ordinate chart (Q, φ_2), with

$$\varphi_2(q) = \begin{bmatrix} m(q) \\ 1/s^2(q) \end{bmatrix} \qquad \text{for all } q \in Q$$

where $\varphi_1(q) := (m(q), s^2(q))$ are the components of the first co-ordinate map. This is a reparameterization in terms of mean and precision $t := 1/s^2$. The two charts are smoothly compatible since the transition map

$$(\varphi_2 \circ \varphi_1^{-1}) : Q \to Q$$
$$(\varphi_2 \circ \varphi_1^{-1})(m, s^2) = \left(m, \frac{1}{s^2}\right)$$

is a diffeomorphism. Thus φ_2 is in the same smooth structure as φ_1, and hence defines a co-ordinate chart on the smooth manifold structure on Q.

1.1.3 Co-ordinate Free Posterior Distribution

Equation (1.1.1) is Bayes' theorem for the real-valued unknown variables q, given observed variables y. It assumes the prior and posterior have densities in $\mathbb{R}^n$ and refers to a specific parameterization of the conditional density of the data $p(y|q)$. In this section, we reformulate Bayes' theorem in terms of probability measures on a statistical family. Although Bayes' theorem has been formulated before in measure-theoretic terms, e.g. by Stuart [74], to our knowledge this is the first time in the literature that the posterior has been stated with reference to a manifold of data-generating measures.

Prior Specification and Densities

Assuming a family Q is a smooth manifold, it has a topology which induces a Borel σ-algebra $\mathcal{B}(Q)$; see Rudin [66]. The pair $(Q, \mathcal{B}(Q))$ is a measurable space and so we can make sense of probability measures on Q. In particular, let $\pi : \mathcal{B}(Q) \to [0, 1]$ be a prior probability measure on $(Q, \mathcal{B}(Q))$. We can interpret this measure as weighting data-generating measures $q \in Q$ according to prior information available before data are observed.

We first recall the following definition from measure theory.

De nition 1.1.10. Let q, n be two positive measures on a measurable space (Y, F). If for any measurable set $U \in F$,

$$n(U) = 0 \quad \Rightarrow \quad q(U) = 0$$

then q is absolutely continuous with respect to n, and we write $q \ll n$. If Y is a countable union of sets E_i such that $n(E_i) < \infty$ for each i, then n is called σ-finite.

A famous result in Analysis is the Radon-Nikodym Theorem [66], which states that if $q \ll n$ then there exists a measurable function $h \in L^1(n)$ such that

$$q(U) = \int_U h dn$$

The function h is called the density or Radon-Nikodym derivative of q with respect to n, and is often written as $h = \frac{dq}{dn}$.

De nition 1.1.11. Let n be a positive σ-finite measure on the measurable space (Y, F). If the statistical family Q on (Y, F) is such that $q \ll n$ for all $q \in Q$, then Q is said to be dominated by n. In this case, the conditional data-generating density of $q \in Q$ with respect to n is the Radon-Nikodym derivative

$$\frac{dq}{dn} : Y \to \mathbb{R}$$

This de nition corresponds with our earlier notation of a conditional density $p(y|q)$ on a variable $y \in Y$ when the family Q is indexed by $q \in Q$.

To summarise, there are two measurable spaces of interest; rstly, the data space (Y, F) where observable data and events live, and secondly the statistical family and its Borel sets $(Q, B(Q))$. A statistician constructs a model, i.e. a statistical family Q of probability distributions on data space, each element of which is a candidate for the true distribution which generated the observed data. The prior probability distribution m is a measure on $(Q, B(Q))$ that represents prior uncertainty about how likely each distribution in Q is to generate the data. We can now describe the process of Bayesian analysis without reference to any particular parameterization.

Joint Distribution

The rst step of Bayesian analysis outlined in Section 1.1 is to construct a joint probability distribution over parameters and observable variables. In the co-ordinate free picture, the joint distribution is a measure on the product space of the statistical family and data space,

$$(Q \times Y, B(Q) \otimes F)$$

where $B(Q) \otimes F$ refers to the product σ-algebra of $B(Q)$ and F.

De nition 1.1.12. Let Q be a statistical family on the measurable space (Y, F) and let $m: B(Q) \to [0,1]$ be a probability distribution over Q. If $g: Q \to [0,1], g(q) = q(A)$ is a $B(Q)$-measurable function for all $A \in F$, de ne the joint probability measure as

$$(m \ Q) : B(Q) \ F \to [0,1]$$

$$(m \ Q)(U \ A) = \int_U q(A)\, dm(q) \qquad \text{for } U \in B(Q), A \in F$$

Co-ordinate Free Likelihood and Posterior

Let us de ne the likelihood in a co-ordinate free way.

De nition 1.1.13. Let Q be a statistical family on a measurable space (Y, F) dominated by a measure n. The likelihood function l_n with respect to n at a point $q \in Q$ is the data-generating density of q evaluated at the observed data y_{obs},

$$l_n : Q \to R$$

$$l_n(q) := \frac{dq}{dn}(y_{obs}) \qquad \text{for all } q \in Q$$

Note that this de nition of likelihood does not depend on any particular co-ordinate chart. The likelihood can be represented in a co-ordinate chart q in the usual way for functions on manifolds,

$$l_n : Q \to R \qquad l_n := l_n \ q^{-1}$$

where Q is the image of the q co-ordinate map. This co-ordinate representation agrees with the usual likelihood function $p(y_{obs}|q)$ assuming the same dominating measure n and co-ordinate chart $q = q(q)$. We assume the likelihood is a measurable function on $(Q, B(Q))$.

We can now write Bayes' theorem on a statistical family Q. Let l_n be the likelihood and mthe prior probability measure on Q.

De nition 1.1.14. The posterior probability measure given y_{obs}, written m, is de ned by the Radon-Nikodym derivative

$$\frac{dm}{dm}(q) = \frac{l_n(q)}{\int_Q l_n(q')\, dm(q')} \tag{1.1.5}$$

Thus for a Borel set $U \in B(Q)$,

$$m(U) = \int_U \frac{l_n(q)}{\int_Q l_n(q')\, dm(q')}\, dm(q)$$

When specifying the model and doing computations, it is usually easier to work with prior and posterior densities rather than measures. Prior and posterior densities on R^p are obtained by pushing-forward the probability measures under the co-ordinate charts and satisfy the usual statement of Bayes' theorem, as shown in the following proposition.

Proposition 1.1.15. Let Q be a p-dimensional smooth statistical family dominated by λ, and let μ be a prior probability measure on Q. Suppose (Q, f) is a global co-ordinate chart on Q with co-ordinates $x = f(q)$ for $q \in Q$. If the push-forward of μ under f admits a Lebesgue density $p(q)$, then the pushforward of the posterior probability measure under f has a Lebesgue density given by

$$\frac{l_n(q)\,p(q)}{\int l_n(q')\,p(q')\,dq'}$$

where $l_n := l_n \circ f^{-1}$ is the co-ordinate representation of the likelihood.

Proof. Firstly, note that since $f : Q \to \mathbb{R}^p$ is a homeomorphism, both f and f^{-1} are continuous and therefore measurable functions.

The pushed-forward prior and posterior measures under f are respectively defined by

$$f_*\mu(E) = \mu \circ f^{-1}(E)$$

$$f_*\nu(E) = \nu \circ f^{-1}(E)$$

for any Borel set E in $\mathbb{R}^p$. Substituting Definition 1.1.14 for ν, we find

$$\begin{aligned} f_*\nu(E) &= \nu \circ f^{-1}(E) \\ &= \int_{f^{-1}(E)} d\nu(q) \\ &= \frac{\int_{f^{-1}(E)} l_n(q)\,d\mu(q)}{\int_Q l_n(q')\,d\mu(q')} \end{aligned}$$

Let l^p be the p-dimensional Lebesgue density. By the assumptions of the proposition, the pushed-forward prior measure is $p(q)\,dl^p(q)$. Thus the expression in the previous equation can be rewritten as

$$f_*\nu(E) = = \frac{\int_E l_n \circ f^{-1}(q)\;p(q)\,dl^p(q)}{\int_{f(Q)} l_n(f^{-1}(q'))\,p(q')\,dl^p(q')}$$

since $l_n = l_n \circ f^{-1}$, this shows the required result. □

1.2 Exponential Families

We narrow our focus to exponential statistical families. A common assumption in statistical models is that data are generated by a probability distribution belonging to an exponential family; many elementary families are exponential, such as the normal, Bernoulli, binomial, beta and gamma distributions. In Chapter 2, we will show that geometric structures on exponential families have statistical interpretations.

The following definition and subsequent discussion is adapted from Brown [15] and Shao [72].

De nition 1.2.1. Let $Q = \{P_x | x \in X\}$ for $X \subseteq R^s$ be a parametric family on (Y, F) dominated by a s- nite measure n. The family is called a p-dimensional exponential family if the density of P_x with respect to n is

$$\frac{dP_x}{dn}(y) = h(y) \exp\left\{ c(x)^T T(y) - y(x) \right\} \quad \text{for all } x \in X \tag{1.2.1}$$

for some functions

$c : X \to R^p$,

$T : Y \to R^p$, measurable on (Y, F),

$h : Y \to R$, measurable on (Y, F),

$y : X \to R$, $y(x) = \log \int_Y h(y) \exp\{ c(x)^T t(y) \} \, dn(y)$

Note the de nition of y is necessary for each P_x to be a probability measure, i.e. $P_x(Y) = 1$. The parameter space X may not necessarily be a subset of R^p. We also introduce the following.

De nition 1.2.2. A parametric family $\{P_q | q \in Q\}$ of probability measures on the Borel sets of R^p is called a p-dimensional standard exponential family if the family is dominated by a s- nite measure n, $Q \subseteq R^p$, and

$$\frac{dP_q}{dn}(t) = \exp\left\{ q^T t - y(q) \right\} \quad \text{for all } q \in Q \tag{1.2.2}$$

where

$$y : Q \to R, \qquad y(q) = \log \int_{R^p} \exp\{ q^T t \} \, dn(t) \tag{1.2.3}$$

The parameter q of a standard exponential family is called its natural parameter , Q is the natural parameter space , and the function y is the cumulant generating function . Clearly any standard exponential family is an exponential family. Furthermore, the following proposition shows that one can always transform an exponential family into a standard one.

Proposition 1.2.3. Let $\{P_x | x \in X\}$ be a p-dimensional exponential family of probability measures dominated by a s- nite n on a measurable space (Y, F), where $X \subseteq R^s$. Let $T : Y \to R^p$, $h : Y \to R$, and $c : X \to R^p$ be as in (1.2.1) for this family. Let $Q = \{c(x) : x \in X\}$ and de ne the measure

$$n(A) := \int_{T^{-1}(A)} h(y) \, dn(y)$$

for Borel sets $A \subseteq R^p$. Then the family $Q = \{P_q | q \in Q\}$ is a p-dimensional standard exponential family dominated by n, where

$$P_q(A) := \int_A \exp\left\{ q^T t - y(q) \right\} \, dn(t) \tag{1.2.4}$$

is the push-forward measure of P_x under T where $q = c(x)$, and y is given by (1.2.3).

Proof. Firstly, note that the measure ν is well-defined since T and h are both measurable by Definition 1.2.1. For any Borel set $A \subseteq \mathbb{R}^p$ and measurable $f : \mathbb{R}^p \to \mathbb{R}$, the definition of ν implies that

$$\int_A f(t)\, d\nu(t) = \int_{T^{-1}(A)} f(T(y))\, h(y)\, d\mu(y) \tag{1.2.5}$$

In particular, for any $x \in \mathcal{X}$ let $f(t) := \exp\left(q^T t\right)$ where $q = c(x)$ and take $A = \mathbb{R}^p$. Equation (1.2.5) is then equivalent to

$$\exp(\psi(q)) = \exp(\psi(x)) \tag{1.2.6}$$

It follows that $\psi(x) = \psi(c(x))$ for any $x \in \mathcal{X}$. Dividing (1.2.5) by the non-zero quantity (1.2.6) for each Borel set $A \subseteq \mathbb{R}^p$ yields

$$\int_A \exp\left(q^T t - \psi(q)\right) d\nu(t) = \int_{T^{-1}(A)} \exp\left(c(x)^T T(y) - \psi(x)\right) h(y)\, d\mu(y)$$
$$, \quad P_q(A) = P_x(T^{-1}(A))$$

which shows that (1.2.4) is the pushforward of P_x under T for any $x \in \mathcal{X}$, where $q = c(x)$. The form (1.2.4) of P_q clearly shows that $\mathcal{Q}$ is a standard exponential family. □

Suppose a random variable y has a distribution in a parametric family $\mathcal{Q} = \{P_q | q \in \mathcal{Q}\}$. Recall that a function, or statistic, $T(y)$ is called *sufficient* for q if the conditional distribution of y given $T(y)$ does not depend on q. Sufficient statistics are interpreted as containing all of the information about the parameter q contained in a sample y. By the Neyman Factorization theorem for sufficient statistics [72], [71], the function T in Definition 1.2.1 is sufficient for the natural parameter $q = c(x)$. This function is therefore referred to as the *sufficient statistics* of the family. The standard form of an exponential family describes the distribution of its sufficient statistics $t := T(y)$.

The standard form of an exponential family is simplest for studying its properties and geometry. For example, the following proposition proven in Brown [15] shows that standard exponential families can be affinely transformed to an equivalent standard exponential family. This shows the natural parameterization is not unique.

Proposition 1.2.4. Let $\mathcal{Q} = \{P_q | q \in \mathcal{Q}\}$ be a p-dimensional standard exponential family dominated by a σ-finite measure ν. For $q \in \mathcal{Q}$, let t be a random variable with distribution P_q and define the transformed variable and parameter

$$t \to t' = Mt + t_0 \tag{1.2.7}$$
$$q \to q' = M^{-T} q + q_0 \tag{1.2.8}$$

where $M \in \mathbb{R}^{p \times p}$ is non-singular and $q_0 \in \mathbb{R}^p$. Then t' has the distribution

$$P_{q'}(A) := P_q\left(M^{-1}(A - t_0)\right) \tag{1.2.9}$$

for Borel sets $A \subset \mathbb{R}^p$, where $q = M^T(q^0 - q_0)$. Furthermore, the parametric family

$$Q := \{P_{q^0} | q^0 \in Q^0\}$$

where $Q^0 = \{M^T q + q_0\}$ is a p-dimensional standard exponential family.

The joint distribution of random variables distributed according to measures from an exponential family belongs to a product exponential family. Suppose two independent random variables $t_1, t_2 \in \mathbb{R}^p$ have distributions from the same p-dimensional standard exponential family Q with cumulant generating function ψ and dominating measure ν, and let t_1, t_2 respectively have natural parameters $q_1, q_2 \in \mathbb{R}^p$. Their joint distribution then has density

$$\exp\{q_1^T t_1 + q_2^T t_2 - \psi(q_1) - \psi(q_2)\}$$

with respect to the product measure $\nu \otimes \nu$. This is clearly a 2p-dimensional standard exponential family with sufficient statistics $t := (t_1, t_2) \in \mathbb{R}^{2p}$. Alternatively, if t_1, t_2 have the same parameter $q \in \mathbb{R}^p$, they are identically distributed and the joint density is

$$\exp\{q^T(t_1 + t_2) - 2\psi(q)\}$$

and we only need to redefine a sufficient statistic $t = t_1 + t_2$ and the dimension of the family is the same as the original one. This is made clear in the following example.

Example 1.2.5. The set of Gaussian distributions on the real line $\{\text{Normal}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$ is an exponential family. Rearranging the Gaussian density with respect to the Lebesgue measure gives

$$\begin{aligned} p(y|\mu, \sigma^2) &= \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log 2\pi\sigma^2\right\} \\ &= \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2\right\} \end{aligned} \tag{1.2.10}$$

Let us find the equivalent standard exponential family as per Proposition 1.2.3. Define natural parameters $q_1 = -\frac{1}{2\sigma^2}$ and $q_2 = \frac{\mu}{\sigma^2}$, a cumulant generating function $\psi(q_1, q_2) := -\frac{q_2^2}{4q_1} + \frac{1}{2}\log\left(-\frac{\pi}{q_1}\right)$, and sufficient statistics

$$T(y) = (t_1(y), t_2(y))^T = \begin{bmatrix} y^2, y \end{bmatrix}^T$$

This is a parabola in $\mathbb{R}^2$ parameterized by y. Let ν be the push-forward of the one-dimensional Lebesgue measure under T. Then we can rewrite the parametric family as a standard exponential family with dominating measure ν and natural parameter space $Q = (-\infty, 0) \times \mathbb{R}$.

Suppose there are n i.i.d. Gaussian random variables $y_1, .., y_n$ each with the same natural parameters q_1, q_2. The joint Lebesgue density is

$$\begin{aligned}
\prod_{i=1}^{n} \exp\left(q_1 y_i^2 + q_2 y_i - y(q) \right) &= \exp\left(\sum_{i=1}^{n} \left(q_1 y_i^2 + q_2 y_i \right) - y(q) \right) \\
&= \exp\left(q_1 \left(\sum_{i=1}^{n} y_i^2 \right) + q_2 \left(\sum_{i=1}^{n} y_i \right) - n y(q) \right) \\
&= \exp\left(n q_1 \left(\frac{1}{n} \sum_{i=1}^{n} y_i^2 \right) + n q_2 \left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) - n y(q) \right) \\
&= \exp\left(n q_1 \left(\frac{1}{n} \sum_{i=1}^{n} t_1(y_i) \right) + n q_2 \left(\frac{1}{n} \sum_{i=1}^{n} t_2(y_i) \right) - n y(q) \right)
\end{aligned}$$

In the last line, we have pulled a factor of n out of the sufficient statistics; this can be absorbed into the natural parameters with a simple re-scaling, i.e. by applying Proposition 1.2.4. In this new parameterization, the joint sufficient statistic is simply the sample mean of the individual sufficient statistics, corresponding to the negative second and positive first sample moments. In conclusion, the n data are distributed according to a two-parameter exponential family.

1.2.1 Minimality and Identifiability

In this subsection we will identify sufficient conditions for a standard exponential family to have a smooth manifold structure. Recall that Theorem 1.1.8 states that for a parametric family to have a smooth manifold structure in which its parameterization is a co-ordinate chart, it is sufficient for it to be identifiable with an open parameter space.

De nition 1.2.6. A p-dimensional standard exponential family is called *minimal* if the sufficient statistics $T(y)$ are linearly independent for all $y \in Y$, and there does not exist a $(p-1)$-dimensional subspace containing the natural parameter space Q.

The following proposition is shown by Geyer [35].

Proposition 1.2.7. A minimal standard exponential family is identifiable.

By Theorem 1.1.8, it follows that a minimal standard exponential family with an open parameter space has a unique smooth manifold structure in which the natural parameterization is a smooth co-ordinate chart.

An exponential family may not contain all possible probability distributions of exponential form with respect to a given measure. We define the *full natural parameter space* N to be the set of all permissible natural parameter values, i.e.

$$N = \left\{ q \in \mathbb{R}^p : y(q) = \int_{\mathbb{R}^p} \exp\left(q^T t \right) \, d\nu(t) < \infty \right\}$$

The associatedfull exponential family is $\exp(q^T t - y(q))\, dn(t) : q \in N$, and say a family is full if $Q = N$. Any s-finite measure n on Borel subsets of R^p has an associated full natural parameter space N_n, and therefore induces a standard exponential family of probability measures. If the natural parameter space Q of an exponential family has a non-empty interior, i.e. it contains a non-empty open set of R^p, then the family is said to be of full rank. Brown [15] defines an exponential family to be regular if its full natural parameter space N is an open subset of R^p.

Example 1.2.8. Consider the Lebesgue measure l with support restricted to $(0,\infty)$. The full natural parameter space is

$$N_l = N = \{q \in R : \int_0^\infty e^{qt}\, dl(t) < \infty\} = (-\infty, 0)$$

The family is therefore regular. The cumulant generating function is $y(q) = -\log(-q)$ for $q \in N$. The resulting exponential family is the family of exponential distributions, i.e.

$$\exp(qt + \log(-q))\, dl(t) \qquad q \in (-\infty, 0) \qquad t \in (0,\infty)$$

This is usually parameterized by a rate parameter $r = -q$; since r and q are related by an affine transformation, r is also a natural parameter if t is transformed according to $t \to -t$. A contrived family with parameter space $Q = \{-1, -2\} \subset N$ is standard exponential, but it does not have full rank.

The following example adapted from Amari [3] shows that the family of distributions whose support are all on the same discrete set is an exponential family.

Example 1.2.9. The set of probability measures on a finite set $Y = \{0,..,N\}$ is an exponential family. Let $P(y = i) = p_i$ for $i \in \{0,..,N\}$. Of course, there are $N+1$ outcomes with the constraint $\sum p_i = 1$, so the family must have N degrees of freedom (parameters). For $i = 1,..,N$, define natural parameters and sufficient statistics

$$q_i = \log \frac{p_i}{p_0} \qquad T_i(y) = \begin{cases} 1 & y = i \\ 0 & y \neq i \end{cases}$$

Take n to be the counting measure on $\{0,..,N\}$. The cumulant generating function is then

$$y(q) = \int \log \exp(q_i T_i(y))\, dn(y) = \log\left(1 + \sum_{i=1}^{N} \exp(q_i)\right)$$

With these definitions, it follows that for $i = 1,..,N$

$$\begin{aligned} \exp(q \cdot t(i) - y(q)) &= \exp(q_i - y(q)) \\ &= \frac{\exp(q_i)}{1 + \sum_{j=1}^{N} \exp q_j} = \frac{p_i / p_0}{1 + \sum_{j=1}^{N} p_j / p_0} = p_i \end{aligned}$$

and similarly $\exp(q \cdot T(0) - y(q)) = p_0$. So the set of discrete random variables taking a finite number of values can be identified with an exponential family.

1.2.2 Convexity and Legendre Transform

The geometry and properties of exponential families are based on convexity of the natural parameter space N and cumulant generating function ψ.

Proposition 1.2.10. For a minimal regular exponential family, the cumulant generating function $\psi : N \to \mathbb{R}$ is an infinitely differentiable strictly convex function and the natural parameter space N is a convex set.

The proof of proposition 1.2.10 is given by Kass and Vos [44], and is based on a simple application of Hölder's inequality. Convexity is not necessarily preserved under diffeomorphic changes of co-ordinates, so we consider ψ strictly as a function of natural parameters θ and not as a co-ordinate independent geometric object. However, the convexity of ψ gives it the following important property, also proven by Kass and Vos.

Proposition 1.2.11. Let $\psi : N \to \mathbb{R}$ be strictly convex and smooth where $N \subseteq \mathbb{R}^p$ is open and convex. Then the map

$$c : \mathbb{R}^p \to \mathbb{R}^p \qquad c(\theta) := \nabla \psi(\theta)$$

is a diffeomorphism.

Viewing the exponential family as a smooth manifold, this proposition provides another global smooth co-ordinate chart. Let (Q, f) be the co-ordinate chart for the natural parameterization, and define (Q, η), where

$$\eta : Q \to \mathbb{R}^p \qquad \eta(q) := (c \circ f)(q)$$

with $c(\theta) = \nabla \psi(\theta)$ as in the proposition. We will denote the co-ordinates in this chart as $\eta := c(\theta)$.

The statistical interpretation of this co-ordinate chart is given by the following theorem, which states that $\nabla \psi(\theta)$ is the expected value of the sufficient statistic vector T under the distribution with natural parameter θ.

Theorem 1.2.12. Let $Q = \{P_\theta | \theta \in N\}$ be a minimal regular exponential family of probability distributions on random vectors t with natural parameter θ. The expected value of the sufficient statistic t under P_θ is given by the gradient of the cumulant generating function ψ at θ,

$$\eta := E_{P_\theta}(t(y)) = \nabla \psi(\theta) \tag{1.2.11}$$

Proof. The member of Q with natural parameter θ is the distribution

$$P_\theta(A) = \int_A \exp\{\theta \cdot t - \psi(\theta)\} \, d\nu(t)$$

for Borel sets $A \subseteq \mathbb{R}^p$. Since the integrand is continuous in θ and t, we can differentiate

under the integral sign. The i^{th} component of the gradient of y is given by

$$\begin{aligned}\frac{\partial y}{\partial q^i} = \frac{\partial}{\partial q^i}\log\int_{R^n}\exp\left(q^T t\right)d\nu(t) &= \frac{1}{\int_{R^n}\exp(q^T t)\,d\nu(t)}\frac{\partial}{\partial q^i}\int_{R^n}\exp\left(q^T t\right)d\nu(t)\\ &= \frac{1}{y(q)}\int_{R^n}\frac{\partial}{\partial q^i}\exp\left(q^T t\right)d\nu(t)\\ &= \int_{R^n} t_i\exp\left(q^T t\right)y(q)\,d\nu(t)\\ &= \int_{R^n} t_i\,dP_q(t) = E_{P_q}(t_i)\end{aligned}$$

where t^i is the i^{th} component of the vector t. This holds for all $i = 1,..,p$ and (1.2.11) follows. □

Given this result, we call h the mean parameter or mean co-ordinates for the family, which range over the mean parameter space $M := \{r\,y(q) | q \in N\}$.

Example 1.2.13. Continuing from the family of discrete distributions in example 1.2.9, the mean sufficient statistic is simply the probability p_i;

$$h_i := E_{P_q}(t_i(y)) = E_{P_q}(\delta_i(y)) = p_i$$

By theorem 1.2.12, the gradient of $y = \log(1 + \sum_i \exp(q_i))$ gives the mapping from natural to mean parameters,

$$h_k = \frac{\partial y}{\partial q_k} = \frac{\exp(q_k)}{1 + \sum_{i=1}^{n}\exp(q_i)}$$

In the binary or Bernoulli case where $Y = \{0, 1\}$, this reduces to the sigmoid or expit function

$$h = \frac{\exp(q)}{1 + \exp(q)}$$

which is the inverse of the logit function $q = \log\frac{h}{1-h}$.

The Legendre transform associated with y refers to the variable transformation given by $h := r\,y(q)$, and also the following transformation of y to a convex function of h.

De nition 1.2.14. Let $y : N \to R$ be a strictly convex smooth function on an open convex set $N \subseteq R^p$. The Legendre dual or convex conjugate of y is the function

$$y^* : M \to R \tag{1.2.12}$$

$$y^*(h) := \sup_{q \in N}\left\{q^T h - y(q)\right\} \tag{1.2.13}$$

or equivalently,

$$y^*(h) = q^T h - y(q) \qquad \text{where } q \text{ is such that } h = r\,y(q) \tag{1.2.14}$$

where $M = \{r\,y(q) | q \in N\}$.

To see that the de nitions (1.2.14) and (1.2.13) are equivalent, note that since γ is smooth the supremum in (1.2.13) is attained by $q_0 \in N$ depending on h such that

$$0 = r_q \left(q^T h - \gamma(q) \right)_{q=q_0} = h - r\gamma(q_0)$$

and thus q_0 satis es $r\gamma(q_0) = h$.

The term dual refers to the fact that the Legendre transform is an involution; it is its own inverse, such that the Legendre transform applied to γ^* and h recovers γ and q, as shown in the following theorem.

Theorem 1.2.15. Let $\gamma : N \to \mathbb{R}$ be an in nitely differentiable, strictly convex function on a convex set $N \subseteq \mathbb{R}^n$. The Legendre dual (1.2.13) satis es

$$r_q \gamma^* \left(r_h \gamma^*(h) \right) = h$$

where r_h refers to the gradient with respect to h, i.e. the inverse of $h(q) = r_q\gamma(q)$ is $c^{-1}(h) = r_h\gamma^*(h)$. Furthermore, the Legendre dual of $\gamma^*(h)$ is $\gamma(q)$.

Proof. Using the alternative de nition (1.2.14) of the Legendre dual, The gradient $r_h\gamma^*(h)$ has i^{th} component

$$\frac{\partial \gamma^*}{\partial h_i} = \frac{\partial}{\partial h_i}\left(\sum_{j=1}^{d} q_j h_j - \gamma(q) \right) = \sum_{j=1}^{d} \frac{\partial q_j}{\partial h_i} h_j + \frac{\partial h_j}{\partial h_i} q_j - \sum_{k=1}^{d} \frac{\partial q_k}{\partial h_i} \frac{\partial \gamma}{\partial q_k}$$

$$= \delta_j^i q_j + \sum_{j=1}^{d} \frac{\partial q_j}{\partial h_i} h_j - \sum_{k=1}^{d} \frac{\partial q_k}{\partial h^i} h_k = q_i$$

where q is such that $h = r\gamma(q)$. Hence $q = r_h\gamma^*(h)$ if and only if $h = r\gamma(q)$. The Legendre dual of γ^* is therefore

$$\begin{aligned} \gamma^{**}(q) &= h^T q - \gamma^*(h) \qquad \text{where } h \text{ is such that } q = r_h\gamma^*(h) \\ &= h^T q - q^T h + \gamma(q) = \gamma(q) \end{aligned}$$

□

Whenever co-ordinates q and h are related via a Legendre transform in this way, we will refer to them as primal and dual co-ordinates respectively.

When γ is the cumulant generating function of a natural exponential family, the Legendre dual $\gamma^*(h)$ coincides with the negative entropy of the suf cient statistics t. Calculating directly,

$$\begin{aligned} H(q) &:= E(\log p(y|q)) \\ &= E_{P_q}\left(q^T t - \gamma(q) \right) \\ &= q^T E_{P_q}(t) - \gamma(q) \\ &= q^T r\gamma(q) - \gamma(q) = \gamma^*(c_h(q)) \end{aligned}$$

The Hessian of the cumulant generating function $\nabla^2\psi$ is the Jacobian matrix of the natural-to-mean diffeomorphism $c(\theta) := \nabla\psi(\theta)$, and also has geometrical and statistical interpretations. First, we recall the definition of Fisher information.

Definition 1.2.16. Let $\mathcal{Q} = \{P_\theta : \theta \in \Theta\}$ be a parametric family dominated by a measure ν, such that P_θ has density $p(y|\theta)$ with respect to ν. The Fisher information matrix, or simply Fisher matrix, for the family is the matrix with i-j element

$$I_{ij}(\theta) = E_{P_\theta}\left[\frac{\partial}{\partial\theta_i}\log p(y|\theta)\frac{\partial}{\partial\theta_j}\log p(y|\theta)\right] \tag{1.2.15}$$

The definition holds for any statistical family, not just exponential families. Note that the Fisher matrix can be viewed as a matrix-valued function on the parameter space $I : \Theta \to \mathbb{R}^{p\times p}$. The matrix plays an important role in statistics; for example, its inverse appears in the Cramer-Rao lower bound. Under mild conditions, the Fisher matrix can be written as

$$I(\theta) = E_{P_\theta}\left[\nabla^2_\theta \log p(y|\theta)\right] \tag{1.2.16}$$

A proof of this alternative form is given in Chapter 3 of Shao [72].

The following theorem shows that for exponential families, the Fisher matrix is simply the Hessian of the cumulant-generating function.

Theorem 1.2.17. For a standard exponential family with natural parameter θ, the Fisher Information matrix $I(\theta)$ and the covariance of $T(y)$ coincide, and are given by the Hessian matrix of ψ,

$$I(\theta) = \mathrm{Cov}_{P_\theta}(t(y)) = \nabla^2\psi(\theta)$$

A proof can be found on page 18 of Kass and Vos [44]. Finally, we note the following relationship between the Hessian of a convex function and its Legendre dual.

Theorem 1.2.18. Let $\psi : \mathcal{N} \to \mathbb{R}$ be a strictly convex function with Legendre dual ψ^*. The inverse of the Hessian of ψ at θ is the Hessian matrix of ψ^* evaluated at $\eta = \nabla\psi(\theta)$, i.e.

$$\nabla^2_\theta\psi(\theta)\nabla^2_\eta\psi^*(\eta) = 1_{p\times p} \tag{1.2.17}$$

Proof. Let $\eta = \nabla\psi(\theta)$ as in the theorem statement. Firstly, by Theorem 1.2.15 we have $\theta = \nabla\psi^*(\eta)$. The i^{th} component of this is

$$\theta^i = \frac{\partial\psi^*}{\partial\eta^i}(\eta)$$

To evaluate the Hessian of ψ^*, note that by the chain rule

$$\frac{\partial}{\partial\eta^j} = \sum_{s=1}^{p}\frac{\partial\theta^s}{\partial\eta^j}\frac{\partial}{\partial\theta^s}$$

Substituting this into the i-k component of the left-hand side of (1.2.17) and applying the chain rule gives

$$\begin{aligned}
\sum_{j=1}^{p} \frac{\partial^2 y}{\partial q^i q^j} \frac{\partial^2 y}{\partial h^j \partial h^k} &= \sum_{j=1}^{p} \frac{\partial}{\partial q^i} \frac{\partial y}{\partial q^j} \frac{\partial}{\partial h^j} \frac{\partial y}{\partial h^k} \\
&= \sum_{j=1}^{p} \frac{\partial h^j}{\partial q^i} \frac{\partial}{\partial h^j} q^k \\
&= \sum_{j=1}^{p} \sum_{s=1}^{p} \frac{\partial h^j}{\partial q^i} \frac{\partial q^s}{\partial h^j} \frac{\partial q^k}{\partial q^s} \\
&= \sum_{j=1}^{p} \frac{\partial h^j}{\partial q^i} \frac{\partial q^k}{\partial h^j} = \frac{\partial q^k}{\partial q^i} = g_i^k
\end{aligned}$$

which is the i-k component of the right-hand side of (1.2.17). □

1.3 Generalized Linear Models

Many statistical families in applications are subsets of other families which are smooth manifolds in their own right. In Chapter 2 we will discuss subfamilies in more generality; here we will describe subfamilies called generalized linear models, a common type of regression model in statistics.

Let Q be a one-dimensional exponential family with sufficient statistic y and cumulant generating function $y(q)$. Consider n independent observations $y := (y_i)_{i=1}^n$ such that each observation was generated from some distribution $P_{q_i} \in Q$. The joint density or mass function of all observations y for fixed natural parameters $q = (q_1, .., q_n)$ is

$$\begin{aligned}
p(y|q) = \prod_{i=1}^{n} p(y_i|q_i) &\propto \exp\left(\sum_{i=1}^{n} (y_i q_i - y(q_i)) \right) \\
&\propto \exp\left(y \cdot q - \sum_{i=1}^{n} y(q_i) \right)
\end{aligned} \tag{1.3.1}$$

where proportionality refers to multiplicative constants independent of q. We see that the n observations are generated by a distribution P_q in an n-dimensional exponential family, which we refer to as Q^n, with natural parameter $q \in R^n$ and cumulant generating function $y^n(q) := \sum_{i=1}^n y(q_i)$. Suppose further that for each observation y_i we have a fixed vector $x_i \in R^p$ of covariates, one often wishes to find a relationship between the covariates and observations. A generalized linear model (GLM) posits a relationship of the form

$$f^{-1}(E(y_i)) = \sum_{j=1}^{p} b_j X_{ij} \qquad i = 1, .., n \tag{1.3.2}$$

where $f : R \to R$ is some (usually nonlinear) function called the link, X_{ij} is the j^{th} component of the i^{th} covariate vector x_i, and b_j is a coefficient corresponding to the j^{th} covariate dimension. The generalized linear model is parameterized by the p coefficients $(b_j)_{j=1}^{p}$ - assuming $p < n$, this implies the GLM is a p-dimensional family G that is a subfamily of Q^n.

In general, a GLM G is not an exponential family [3]. However a special choice of link function f does make G an exponential family in its own right. Note that by Theorem 1.2.12, the left hand-side of equation (1.3.2) is $f^{-1}(h_i)$ where h_i is the i^{th} mean parameter. Suppose we make f the natural-to-mean mapping of Q, $f(q_i) := \frac{dy}{dq} = h_i$, so that

$$f^{-1}(h_i) = q_i = \sum_{j=1}^{p} b_j X_{ij} \qquad i = 1,..,n \tag{1.3.3}$$

This choice of link function gives a linear relation between the natural parameters and coefficients; in vector form the above equation is $q = Xb$, where $X = (X_{ij})$ is the design matrix of the GLM and $b = (b_j)$ is the coefficient vector. Such a link function f is called the canonical link function for the underlying exponential family Q. The vector form of the GLM relation shows that the set $G \subset Q^n$ is exactly the image of X in natural co-ordinates.

A GLM with a canonical link function is a p-dimensional exponential family. To see this, substitute $q = Xb$ into the joint mass/density (1.3.1) of $P_q \in Q$;

$$\begin{aligned} p(y|b) &\propto \exp\left(y \cdot q - \sum_{i=1}^{n} y(q_i) \right) \\ &\propto \exp\left(y \cdot (Xb) - \sum_{i=1}^{n} y \sum_{j=1}^{p} X_{ij} b_j \right) \\ &\propto \exp(t(y) \cdot b - y(b)) \end{aligned}$$

where $t(y) := X^T y$ is the sufficient statistic and $y(b) := \sum_{i=1}^{n} y \sum_{j=1}^{p} X_{ij} b_j$ is the cumulant generating function of G.

1.4 Conjugate Priors

The discussion thus far has related entirely to the likelihood or data-generating distribution $p(y|q)$. To fully specify a model for Bayesian inference, we require a prior distribution $p_0(q)$ which represents belief about the parameters before data are observed. After observations, we update our belief using Bayes' Theorem,

$$p_1(q) := p(q|y) \propto p(y|q) p_0(q) \tag{1.4.1}$$

Exponential family likelihoods admit conjugate families of priors; families which are closed under Bayesian updating according to (1.4.1). For example, choosing a prior

$p_0 \in C$ where C is a conjugate family of priors for the likelihood $p(y|q)$ implies the posterior is a member of the same family, $p_1 \in C$.

For the numerical experiments in this report, we use posterior distributions resulting from exponential families or GLM likelihoods with conjugate priors. This may seem too restrictive, since such models are often very easy to sample from, with dedicated algorithms for independent sampling - for example the classical conjugate pairs like Poisson-Gamma or Bernoulli-Beta models. However, the models used in our numerical experiments are not from any well-known family and are non-trivial to sample. We use conjugate priors because they provide an easy way to extend the dual geometry to include prior information. Furthermore, a Diaconis conjugate-exponential posterior (see below) has a known expectation which can be used to measure the performance of MCMC. Of course, an eventual goal of the research is to extend the geometric methods to the broadest class of posterior distributions; conjugate-exponential distributions provide a simple but non-trivial problem to test and understand our methods before we extend them. Finally, exponential families or GLMs with conjugate priors are extremely practical and widely-used statistical models - there are many problems with real data that can be used to benchmark our methods.

1.4.1 The Diaconis Conjugate Prior

For the standard exponential family $p(y|q) \propto \exp\left[q^T y - \psi(q)\right]$, the conjugate prior introduced by Diaconis and Ylvisaker [25] is

$$p_{n_0,y_0}(q) \propto \exp\left\{n_0\left[q^T y_0 - \psi(q)\right]\right\} \tag{1.4.2}$$

normalized so that p_{n_0,y_0} is a probability density over Q, where $n_0 > 0$ and $y_0 \in R^p$ are hyperparameters. This is conjugate to $p(y|q)$, since

$$\begin{aligned} p(q|y) &\propto p(y|q)p_{n_0,y_0}(q) \propto \exp\left\{q^T(y + n_0 y_0) - (1 + n_0)\psi(q)\right\} \\ &= \exp\left[(1 + n_0)\left[q^T \frac{y + n_0 y_0}{1 + n_0} - \psi(q)\right]\right] \end{aligned} \tag{1.4.3}$$

We see that the posterior density is $p_{n',y'}$, i.e. the same form as the prior (1.4.2) with updated hyperparameters

$$n_0 \to n' := 1 + n_0 \qquad y_0 \to y' := \frac{y + n_0 y_0}{1 + n_0} \tag{1.4.4}$$

We can find an interpretation for these hyperparameters by incorporating n data $\{y_i\}$ for $i = 1, .., n$. This is equivalent to iterating (1.4.4) n times to yield updated hyperparameters

$$n_0 \to n' := n + n_0 \qquad y_0 \to y' := \frac{n\bar{y} + n_0 y_0}{n + n_0} \tag{1.4.5}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the data. The update $n_0 \to n_0 + n$ motivates viewing n_0 as a prior sample size. Diaconis and Ylvisaker [25] show that

$$E_{q \sim p_{n_0,y_0}(q)} \left[E_{y \sim p(y|q)} [y|q] \right] = E_{q \sim p_{n_0,y_0}(q)} [\nabla \psi(q)] = y_0 \tag{1.4.6}$$

This gives an interpretation for the hyperparameter y_0; it is a prior prediction for the mean of the data. When the hyperparameters are updated with an observation, this is reweighted in (1.4.5) with the "true observations" according to the prior and observed sample sizes. Furthermore, we can use this relation to check the convergence of MCMC to the posterior distribution p_1; we only need to compare the Monte Carlo estimate $\frac{1}{N}\sum_{k=1}^{n} \nabla \psi(q^{(k)})$ of the samples $q^{(1)}, \ldots, q^{(N)}$ with the true quantity $y^0 = \frac{n\bar{y} + n_0 y_0}{n + n_0}$.

The Diaconis conjugate simplifies to the usual conjugate prior for specific exponential families. For example, the Diaconis conjugate prior for the Gaussian model from example 1.2.5 reduces to the Normal-Inverse Gamma prior; see Section 1.5.2.

1.5 Example Models

This section describes several specific models and their properties. The models will be used as test cases for MCMC algorithms.

1.5.1 A Linear Gaussian Inverse Problem

Suppose a vector-valued observation $y \in \mathbb{R}^p$ is distributed according to a multivariate Gaussian distribution with a known covariance matrix Γ^{-1} and unknown mean m,

$$y|m \sim \text{Normal}(y; m, \Gamma^{-1})$$

In other words, the distribution has a known precision matrix $\Gamma \in \mathbb{R}^{p \times p}$, which is symmetric positive-definite. The likelihood function can then be written

$$p(y|m) \propto \exp\left(m^T \Gamma y - \frac{1}{2} m^T \Gamma m \right)$$

up to multiplicative factors not involving m. Clearly this is an exponential family with natural parameter $q := \Gamma^T m = \Gamma m$ and cumulant generating function $\psi(q) = \frac{1}{2} q^T \Gamma^{-1} q$. As expected, m is a mean parameter satisfying $m = \nabla \psi(q) = \Gamma^{-1} q$.

We adopt a Gaussian prior in mean parameters for the model,

$$m \sim \text{Normal}(m; m_0, L_0^{-1})$$

where $m_0 \in \mathbb{R}^p$ is the prior mean and $L_0 \in \mathbb{R}^{p \times p}$ is the prior precision matrix. A straightforward calculation shows that the posterior distribution is then

$$m|y \sim \text{Normal}(m; L^{-1} w, L^{-1})$$

where $L := L_0 + \Gamma$ is the posterior precision and $w := \Gamma y + L_0 m_0$.

	Natural Parameters	m, s^2 Parameters
Parameters	$q_1 = \frac{1}{2s^2}, q_2 = \frac{m}{s^2}$	$s^2 = \frac{1}{2q_1}, m = \frac{q_2}{2q_1}$
Log Posterior Density (Up to additive constant)	$n_0 \quad q_1 a + q_2 b \quad \frac{q_2^2}{4q_1} + \frac{1}{2}\log(q_1)$	$\frac{n_0}{2s^2} \quad (m \quad b)^2 \quad b^2 \quad a \quad \frac{6+n_0}{2} \quad \log(s^2)$
Mode	$q_1 = \frac{1}{2(a+b^2)}, q_2 = \frac{b}{2(a+b^2)}$	$m = b, s^2 = \frac{n_0}{n_0+6}(a+b^2)$

Table 1.5.1: The normal observation model

1.5.2 Gaussian with Unknown Mean and Variance

Recall from example 1.2.5 that the two-parameter family of Gaussian distributions

$$y \quad \text{Normal} \quad y; m, s^2$$

is an exponential family with natural parameters $q_1 = \frac{1}{2s^2}, q_2 = \frac{m}{s^2}$, $y(q) = \frac{q_2^2}{4q_1} + \frac{1}{2}\log \frac{p}{q_1}$, and suf cient statistics $t_1(y) = \quad y^2$ and $t_2(y) = y$. Consider the Diaconis conjugate prior for this exponential family:

$$p_0(q) = \exp \quad n_0(q^T t_0 \quad y(q))$$

for some prior hyperparameters $n_0 > 0, t_0 = (a, b)^T$ where $a < 0$ and $b \in R$. Pushing forward this distribution to m, s^2 yields

$$\begin{aligned} p_0(m, s^2) &\propto \exp \quad n_0 \quad a\frac{1}{2s^2} + b\frac{m}{s^2} \quad \frac{m^2}{2s^2} \quad \frac{1}{2}\log 2ps^2 \quad \det \frac{\partial(q_1, q_2)}{\partial(m, s^2)} \\ &= \exp \quad \frac{n_0}{2s^2} \quad m^2 \quad 2bm \quad a \quad \frac{n_0}{2}\log 2ps^2 \quad \frac{1}{2(s^2)^3} \\ &\propto \frac{\exp \quad \frac{n_0}{2s^2} \quad (m \quad b)^2 \quad b^2 \quad a}{\sqrt{2p\frac{s^2}{n_0}}} \frac{1}{(s^2)^{\frac{5+n_0}{2}}} \\ &\propto \frac{\exp \quad \frac{n_0}{2s^2}(m \quad b)^2}{\sqrt{2p\frac{s^2}{n_0}}} \frac{\exp \quad \frac{n_0(b^2+a)}{2}\frac{1}{s^2}}{(s^2)^{\frac{5+n_0}{2}}} \end{aligned}$$

which is a normal-inverse-gamma distribution [33], the usual conjugate prior for normally-distributed data with unknown mean and variance. The properties of the model are shown in table 1.5.1. Note that the mode of the distribution in the m, s^2 parameterization depends on the n_0 hyperparameter, whereas there is no such dependence for the mode in the natural parameterization. In particular, the natural parameter mode (q_1, q_2) transformed to the m, s^2 parameterization is

$$m(q_1, q_2) = b \qquad s^2(q_1, q_2) = \quad a + b^2$$

The s^2 component here disagrees with s^2; in other words, the process of computing the mode does not necessarily commute with changes of parameterization. This implies that the mode or MAP estimator depends on co-ordinates, and thus cannot be geometric. In contrast, the maximum likelihood estimator (MLE) is known to commute with reparameterization [71].

1.5.3 Logistic Regression

Suppose we have a binary response $y_i \in Y = \{0, 1\}$, so that the underlying exponential family Q is the 1-simplex discussed in example 1.2.9. The cumulant generating function is then $\gamma(\theta_i) := \log(1 + \exp(\theta_i))$ and the natural-to-mean diffeomorphism is given by the expit or sigmoid function,

$$p_i = E_{p(y_i|\theta_i)}(y_i) = \gamma'(\theta_i) = \frac{1}{1 + \exp(-\theta_i)} =: \sigma(\theta_i)$$

Here $p_i := P(y_i = 1) = 1 - P(y_i = 0)$ is the mean parameter of the i^{th} observation. Thus the canonical GLM for Q uses the expit function σ as the link function, so that the link relation (1.3.3) is given by

$$\sigma^{-1}(p_i) = \theta_i = \sum_{j=1}^{p} X_{ij} \beta_j$$

This particular GLM is known as logistic regression. The log-likelihood function is given by

$$\log p(y|\beta) = \sum_{i=1}^{n} \sum_{j=1}^{p} y_i X_{ij} \beta_j - \log\left(1 + \exp\left(\sum_{j=1}^{p} X_{ij} \beta_j\right)\right)$$

Since it is a canonical GLM, this is a regular exponential family in β. The prior for the logistic regression coefficients will depend on the application of interest. A common choice is a Gaussian prior, in which case the log posterior density takes the form

$$\log p(\beta|y) = \sum_{i=1}^{n} \sum_{j=1}^{p} y_i X_{ij} \beta_j - \log\left(1 + \exp\left(\sum_{j=1}^{p} X_{ij} \beta_j\right)\right) - \frac{1}{2} \beta^T \Lambda^{-1} \beta$$

i.e. the prior has zero mean and covariance matrix Λ. This best suits data and covariates that have been standardized and centered.

Chapter 2

Information and Differential Geometry

The previous chapter recast Bayesian inference in a co-ordinate free form. In this chapter, we extend this picture by imposing additional co-ordinate free geometric structure on statistical families - this is the study of Information Geometry. Concepts such as curvature, arc-lengths of curves and even the Pythagorean Theorem can be made sense of for statistical families, and such concepts often have neat statistical interpretations.

We follow the approach of Amari [3] and derive all geometrical structures from a divergence function. A divergence is a kind of squared distance function which may be asymmetric in its arguments. In fact, asymmetric divergence functions give rise to an interesting dualistic structure which is not present in classical Riemannian geometry.

Formal details, including several proofs, are contained in appendix A.

2.1 Tangent Space, Cotangent Space and Tensors

2.1.1 Tangent Vectors

Smooth manifolds are often informally described as generalizations of smooth surfaces in that they are *locally* at . In this section we will make this precise; attached to each point of the manifold is a vector space called tangent space, whose elements are called tangent vectors. Differential Geometry is essentially described in terms of maps which act on tangent vectors. We will follow Chapter 3 of Lee [46].

Our aim is to formalize the notion of a directional vector whose base is at a point of the manifold, similar to how tangent vectors to curves in $\mathbb{R}^p$ are attached to a point of the curve to which they are tangential. This can be made sense of in terms of derivations of smooth functions on Q. For a vector $v \in \mathbb{R}^p$, the directional derivative of a smooth function $f : \mathbb{R}^p \to \mathbb{R}$ at a point $q_0 \in \mathbb{R}^p$ is

$$\nabla_v f(q_0) := v[f(q_0)] = \sum_{i=1}^{d} v_i \frac{\partial f}{\partial q_i}\Big|_{q=q_0} = \left(\sum_{i=1}^{d} v_i \frac{\partial}{\partial q_i}\right) f\Big|_{q=q_0}$$

We might therefore think of v as a real-valued operator acting on smooth functions. Let $V(f) := r_v f(q_0)$ be this directional derivative operator, which we view as a map $C^\infty(R^p) \to R$ where $C^\infty(R^p)$ is the space of in nitely-differentiable functions $R^p \to R$. Then V is a linear operator satisfying the Leibniz(product) rule at q_0,

$$V(fg) = f(q_0)V(g) + g(q_0)V(f) \tag{2.1.1}$$

Note that Leibniz rule associates V with the point q_0, and hence formalizes the notion of a tangent vector being attached to a point. Indeed, one can show that the set of derivations at $q_0 \in R^p$ forms a p-dimensional vector space with a basis given by the partial derivatives with respect to each co-ordinate evaluated at q_0,

$$\partial_i|_{q_0} : C^\infty(R^p) \to R \qquad \partial_i|_{q_0}(f) := \frac{\partial f}{\partial q^i}_{q=q_0}$$

for $i = 1, .., p$.

The concept of an operator acting on smooth functions can be generalized to smooth manifolds, since the notion of a smooth function is well-de ned (see Section 1.1.2). This is how tangent vectors on a manifold are de ned.

De nition 2.1.1. A tangent vector or derivation at $q \in Q$ is a linear map $X : C^\infty(Q) \to R$ that satis es the Leibniz rule at q,

$$X(fg) = f(q)X(g) + g(q)X(f) \qquad \text{for all } f, g \in C^\infty(Q)$$

The set of all tangent vectors at q is called the tangent space to Q at q and is denoted by T_qQ.

Analogously to the space of derivations at $a \in R^p$, Tangent space forms a vector space when equipped with the canonical addition and scalar multiplication; for all $X, Y \in T_qQ$, $c \in R$ and $f \in C^\infty(Q)$ we de ne

$$(X + Y)(f) := X(f) + Y(f), \qquad (cX)(f) := cX(f)$$

which are clearly both tangent vectors in T_qQ.

The Differential of a Function and Co-ordinate Basis

Performing explicit calculations with tangent vectors requires a basis. One such basis is provided by a generalized derivative, called the differential of a co-ordinate map. The differential of a map between two manifolds is a linear mapping between their respective tangent spaces; it is a linear approximation to the map.

De nition 2.1.2. Let Q, Q' be smooth manifolds and let $F : Q \to Q'$ be a smooth map between them. The differential dF at $q \in Q$ is the mapping $dF_q : T_qQ \to T_{F(q)}Q'$ such that for all $V \in T_qQ$, the tangent vector $dF_q(V) \in T_{F(p)}Q'$ acts on $f \in C^\infty(Q')$ via

$$dF_q(V)(f) := V(f \circ F)$$

Since $f : Q' \to \mathbb{R}$ and $F : Q \to Q'$ the composition $(f \circ F) : Q \to \mathbb{R}$ is a smooth function on Q and hence can be acted on by a tangent vector $V \in T_q Q$, and hence dF is well-defined. Lemma A.0.2 states that the differential dF_q is a linear isomorphism whenever F is a diffeomorphism.

An important example of differentials is given by co-ordinate maps. If (U, q) is a co-ordinate chart, then $q : U \to \mathbb{R}^p$ is a map between manifolds, since open subsets of $\mathbb{R}^p$ are smooth manifolds - see Example 1.1.3. The differential acts on a function $g : \mathbb{R}^p \to \mathbb{R}$ via

$$dq_q(V)(g) = V(g \circ q)$$

Suppose for example that g is the co-ordinate representation of a smooth function $g : Q \to \mathbb{R}$, i.e. $g = g \circ f^{-1}$. Then clearly

$$df_q(V)(g) = V(g \circ f) = V(g \circ f^{-1} \circ f) = V(g)$$

In words, the co-ordinate differential of a tangent vector applied to the co-ordinate representation of a smooth function agrees with the tangent vector applied to the *pure*, i.e. co-ordinate independent, function.

Since the co-ordinate map $q : U \to f(U)$ is a diffeomorphism onto its range, by Lemma A.0.2 the differential dq_q is a linear isomorphism. Since for any $q \in Q$, the tangent space $T_{q(q)}\mathbb{R}^p$ is a p-dimensional vector space, this implies that $T_q Q$ is also p-dimensional. Furthermore, the differential gives an explicit basis for tangent space as the following theorem shows.

Theorem 2.1.3. Let Q be a smooth manifold with a co-ordinate chart (U, q) for $U \subset Q$. Then for any $q \in Q$, a basis for tangent space $T_q Q$ is given by the p tangent vectors

$$\partial_{i|q,q} : C^\infty(Q) \to \mathbb{R}$$

$$\partial_{i|q,q}(f) := \frac{\partial}{\partial q^i}\bigg|_{q=f(q)} f \circ f^{-1} \qquad \text{for all } f \in C^\infty(Q), \quad i = 1, .., p$$

See Appendix A for a proof.

We refer to the basis $\{\partial_{i|q,q} | i = 1, .., p\}$ as the co-ordinate vectors. When the co-ordinate chart and/or point $q \in Q$ is clear from context, we will abuse notation slightly and write the i^{th} co-ordinate vector as $\partial_{i|q}$ or even ∂_i.

Tangent vectors have an important interpretation, especially in the context of dynamics, of being velocity vectors to curves. As usual, we follow Lee [46]. Let $c : J \to Q$ be a smooth map, where $J \subset \mathbb{R}$ is an open interval. Such a map is a smooth curve on Q; for example, the path of a particle around the manifold parameterized by time $t \in J$. Define the velocity of c at $t = t_0 \in J$ as the tangent vector $c'(t_0) \in T_{c(t_0)} Q$,

$$c'(t_0) = dc\left(\frac{d}{dt}\bigg|_{t=t_0}\right) \tag{2.1.2}$$

Note that $\frac{d}{dt}_{t=t_0} \quad \frac{\partial}{\partial t}_{t=t_0}$ is a tangent vector in $T_{t_0}\mathbb{R}$, and thus is mapped to a vector in $T_{c(t_0)}Q$ under the differential dc. We consider how the velocity vector acts on a smooth function f and unroll the definitions;

$$\begin{aligned} c'(t_0)(f) &:= dc\left(\frac{d}{dt}_{t=t_0}\right)(f) \\ &= \frac{d}{dt}_{t=t_0}(f \circ g) = (f \circ g)'(t_0) \end{aligned}$$

Comparing this to analogous derivatives of functions along curves in $\mathbb{R}^n$, this justifies calling $c'(t_0)$ a velocity. In fact, any tangent vector $V \in T_qQ$ is the velocity vector of some smooth curve [46].

The Tangent Bundle and Vector Fields

The disjoint union of tangent spaces indexed by all points $q \in Q$ is called the tangent bundle,

$$TQ := \bigsqcup_{q \in Q} T_qQ$$

Elements of the tangent bundle are therefore tuples (q, V) where $q \in Q$ and $V \in T_qQ$. If Q is p-dimensional, one can show [46] that the tangent bundle is a $2p$-dimensional smooth manifold. The canonical projection $\pi : TQ \to Q$ maps elements of the tangent bundle to their associated point in the manifold; $\pi((q, V)) = q$. Physically, we can interpret the tangent bundle as the space of possible positions and velocities for some particle in Q, that is, the tangent bundle is the state space for second-order dynamics such as Newton's laws. We will return to this in our discussion of Hamiltonian Monte Carlo in Chapter 4.

In $\mathbb{R}^p$, we have the notion of vector fields, i.e. maps $\mathbb{R}^p \to \mathbb{R}^p$ which are thought of as placing a vector at each point of the space $\mathbb{R}^p$. We generalise this to a manifold as follows:

Definition 2.1.4. Let Q be a smooth manifold. A *vector field* is a smooth map $X : Q \to TQ$ such that the image under X of any point $q \in Q$ is of the form $(q, \_)$, i.e.

$$X \circ \pi = \mathrm{Id}_Q$$

where Id_Q is the identity map on Q.

Since vector fields are maps between smooth manifolds, their smoothness is well-defined. Formally, the value $X(q)$ of a vector field is a tuple (q, V) in the tangent bundle; for brevity, we will follow the convention of only writing the tangent vector part of $X(q)$.

Definition 2.1.5. The set of all vector fields on a smooth manifold Q is written $\mathfrak{X}(Q)$.

In any co-ordinate chart (U, q), a vector field can be expanded in the co-ordinate vector basis,

$$X(q) = \sum_{i=1}^{p} X^i(q) \partial_{i|q} \tag{2.1.3}$$

In fact, each of the p co-ordinate vectors define vector fields, called co-ordinate vector fields,

$$\partial_i : Q \to TQ$$
$$\partial_i(q) = \partial_{i|q}$$

Furthermore, addition and scalar multiplication of vector fields are defined pointwise on Q so that equation (2.1.3) can be written more succinctly as $X = \sum_{i=1}^{p} X^i \partial_i$. Note that the notation ∂_i does not specify the co-ordinate chart q; expressions involving the co-ordinate vector fields should be assumed to hold in any chart over the relevant co-ordinate domain.

Since each smooth curve has a velocity vector, for any vector field X it is natural to consider curves whose velocity vector matches X at each point. In particular, an integral curve of a vector field X is a smooth map $c : J \to Q$, J an interval in $\mathbb{R}$, such that $c'(t) = X(c(t))$ for all $t \in J$. Clearly, such curves are solutions to ordinary differential equations determined by the vector field when written in co-ordinate form.

Vector fields can act on smooth functions to produce new smooth functions. Let X be a vector field on a smooth manifold Q, and let $f : Q \to \mathbb{R}$ be a smooth function. Then we may define a smooth function Xf whose value at each point $q \in Q$ is the action of the tangent vector $X(q) \in T_qQ$ on f,

$$Xf : Q \to \mathbb{R} \qquad (Xf)(q) = X(q)(f) \tag{2.1.4}$$

This construction can be applied recursively; if $Y \in \mathfrak{X}(Q)$ is another vector field, one can consider $YXf = Y(Xf)$. Finally, we can also scale a vector field by a smooth function; for $X \in \mathfrak{X}(Q)$ and $f \in C^\infty(Q)$, define the vector field $fX \in \mathfrak{X}(Q)$ by

$$(fX)(q) = f(q)X(q) \qquad \text{for all } q \in Q$$

2.1.2 Cotangent Vectors

Since tangent space T_qQ is a vector space, we can study the dual space T_q^*Q; the set of linear functionals $a : T_qQ \to \mathbb{R}$. This dual space is a vector space called *cotangent space*, and its elements are called *1-forms* or *covectors*. Elementary linear algebra informs us that the dual space is also a vector space of the same dimension as T_qQ, and hence the same dimension as Q.

An example of a 1-form is the differential of a smooth function $f : Q \to \mathbb{R}$, which maps tangent vectors X to their action on f;

$$d_qf : T_qQ \to \mathbb{R}$$
$$d_qf(X) := X(f)$$

This is clearly a linear functional for any smooth $f : Q \to \mathbb{R}$, so $df \in T_q^*Q$. Note that this coincides with Definition 2.1.2 of the differential of a map between manifolds, if we identify tangent spaces of $\mathbb{R}$ with $\mathbb{R}$ itself; see Lee [46].

Just as the co-ordinate vectors, i.e. partial derivatives, form a basis for tangent spaces, the following proposition shows that the differentials of the co-ordinate functions form a basis for cotangent space which is dual to the co-ordinate vector basis.

Proposition 2.1.6. Let Q be a smooth p-dimensional manifold with $q \in Q$, and let (U, q) be a co-ordinate chart whose domain contains q, with co-ordinates labelled q^j, $j = 1, .., p$. The differentials $\{d_q q^j\}_{j=1}^p$ of the local co-ordinate functions form a basis of cotangent space T_q^*Q that is dual to the co-ordinate tangent vector basis $\{\partial_i\}_{i=1}^p$, i.e. they satisfy

$$d_q q^j(\partial_i) = \partial_i(q^j) = \delta_i^j$$

for all $i, j = 1, .., p$, where $\delta_i^j = 1$ when $i = j$ and is 0 otherwise.

We refer to the basis in the previous proposition as the co-ordinate differential basis or simply the dual basis. For any $a \in T_q^*Q$, it is easy to verify that $a = \sum_i a(\partial_i)\, d_q q^i$, i.e. the components of a covector with respect to the dual basis are given by $a(\partial_i)$, $i = 1, .., p$.

Just as elements of T_qQ are interpreted as tangent vectors to curves, covectors can be seen as gradients of smooth functions $f : Q \to \mathbb{R}$. For example, the expansion of the differential of a smooth function in the dual basis is

$$df_q = \sum_{i=1}^{p} df_q(\partial_i)\, dq^i = \sum_{i=1}^{p} \partial_i(f_q)\, dq^i = \sum_{i=1}^{p} \frac{\partial f}{\partial q^i}\bigg|_{q=q(q)} dq^i$$

Put simply, the coefficients of the differential of f in the dual basis are the components of the gradient vector of the co-ordinate representation f.

As for tangent vectors, we can define the cotangent bundle T^*Q as the disjoint union of all cotangent spaces over all points of the manifold,

$$T^*Q := \bigsqcup_{q \in Q} T_q^*Q$$

covectors have a physical interpretation of being momenta, their action on a velocity vector gives twice the associated kinetic energy. The cotangent bundle is therefore the state space containing positions and momenta of a physical system or particle.

Einstein Summation Notation

We briefly describe a convenient notational convention that is regularly used in Differential Geometry and will be employed throughout the thesis. We will not write sums over indices (e.g. $\sum_{i=1}^{p}$) explicitly, and instead assume that any index that appears in

both upper and lower positions is summed over. For example, we can expand a vector field X at $q \in Q$ in some co-ordinate vector basis as

$$X = \sum_{j=1}^{p} X^j \partial_j \qquad X^j \partial_j$$

In particular, arrays with upper/lower indices respectively denote the coefficients of a tangent/covector with respect to a co-ordinate/differential basis. The basis elements themselves have indices in the opposite positions.

Statistical Scores

For the case when Q is a family of probability distributions, cotangent space has a special interpretation.

Definition 2.1.7. Let Q be a smooth statistical family with dominating measure ν. For a fixed $y \in Y$, Let $L_{\nu,y}(q)$ be the log conditional data-generating density, i.e.

$$L_{\nu,y} : Q \to \mathbb{R}$$
$$L_{\nu,y}(q) = \log \frac{dq}{d\nu}(y) \tag{2.1.5}$$

Define the score at q as the differential of $L_{\nu,y}$ at $q \in Q$,

$$S_{y,q} = d(L_{\nu,y})_q$$

Note that $L_{\nu,y}$ is a co-ordinate free definition of the log-likelihood. Since the action of the score on a tangent vector is a function of y, it is a random variable. If y is distributed according to the distribution q, the same point at which the score is evaluated, then the random variable has expectation zero.

Proposition 2.1.8. Let Q be a smooth statistical family of probability measures on a data space Y. Then for any $q \in Q$ and tangent vector $V \in T_q Q$,

$$E_{y \sim q}\, S_{q,y}(V) = 0$$

Proof. Note that $S_{q,y}(V) = V(L_{\nu,y})$. Choose a co-ordinate chart (U, q) containing q, and write

$$V = V^i \partial_i \qquad L_{\nu,y}(q) = \log p(y|q)$$

Then the action of the score on V follows as:

$$S_{q,y}(V) = V^i \frac{\partial}{\partial q^i}\Big|_{q=q(q)} L_{\nu,y} = V^i \frac{\partial}{\partial q^i}\Big|_{q=q(q)} \log p(y|q)$$

Noting that $\frac{\partial}{\partial q^i} \log p = \frac{1}{p}\frac{\partial p}{\partial q^i}$ and taking and expectation over y yields:

$$\begin{aligned} E\left[S_{q,y}(V)\right] &= V^i E\left[\frac{\partial}{\partial q^i}\Big|_{q=q(q)} \log p(y|q)\right] \\ &= V^i \int \frac{1}{p(y|q)} \left[\frac{\partial}{\partial q^i}\Big|_{q=q(q)} p(y|q)\right] p(y|q)\, dn(y) \\ &= V^i \int \frac{\partial}{\partial q^i}\Big|_{q=q(q)} p(y|q)\, dn(y) \end{aligned}$$

Exchanging the integral and derivatives, this yields the result;

$$\begin{aligned} V^i \int \frac{\partial}{\partial q^i}\Big|_{q=q(q)} p(y|q)\, dn(y) &= V^i \frac{\partial}{\partial q^i}\Big|_{q=q(q)} \int p(y|q)\, dn(y) \\ &= V^i \frac{\partial}{\partial q^i}\Big|_{q=q(q)} (1) = 0 \end{aligned}$$

□

2.1.3 Tensors and Tensor Fields

Just as covectors are linear maps $T_qQ \to \mathbb{R}$, we can consider multilinear maps, i.e. functions of multiple tangent and/or covectors, which turn out to be useful objects in geometry; for example, an inner-product on tangent spaces defines concepts such as length and angles on a manifold, giving rise to Riemannian geometry. The following definition is adapted from Calin and Udriste [17].

Definition 2.1.9. Let Q be a smooth manifold. A type (r,k)-tensor A_q at $q \in Q$ is a map

$$A_q : \underbrace{T_q^*Q \times \cdots \times T_q^*Q}_{r \text{ times}} \times \underbrace{T_qQ \times \cdots \times T_qQ}_{k \text{ times}} \to \mathbb{R}$$

which is multilinear, i.e. linear in each of its $r+k$ slots.

Tensors for which $r = 0$ are known as covariant, whereas tensors with $k = 0$ are contravariant

Example 2.1.10. Covectors are $(0,1)$-tensors, since they are linear functions of a single tangent vector, $T_qQ \to \mathbb{R}$.

Tangent vectors can be viewed as $(1,0)$-tensors, i.e. linear maps $T_q^*Q \to \mathbb{R}$. To see this, for $V \in T_qQ$, the value of V on a covector $w \in T_q^*Q$ is simply $w(V)$. In other words, the dual space of T_q^*Q, $(T_q^*Q)^*$, can be identified with T_qQ.

Tensors have representations with respect to the co-ordinate and differential bases; we may think of tensors as being multidimensional arrays whose elements are the value of the tensor applied to each combination of basis vectors/covectors. For example, suppose C is a $(1,2)$-tensor, i.e. a function of a single covector w and two tangent vectors V, W. Then a simple expansion into basis vectors/covectors and using multilinearity yields:

$$C(w,V,W) = C(w_i dq^i, V^j \partial_j, W^k \partial_k) = w_i V^j W^k \underbrace{C(dq^i, \partial_j, \partial_k)}_{:= C^i_{jk}}$$

where as usual Einstein summation notation is used. We define the components of C with respect to the co-ordinate / differential bases as $C^i_{jk} := C(dq^i, \partial_j, \partial_k)$; note that these are the components of a three-dimensional array, with upper indices representing covector (contravariant) slots and lower indices representing vector (covariant) slots. Dependence on the point in the manifold $q \in Q$ is implicit in the above expression.

There is a notion of tensor *fields* which generalises vector fields, and includes covector fields as a special case. We will not give the detailed definitions in this thesis, which can be found in Lee [46]. Roughly, we can consider *bundles* of tensors, i.e. disjoint unions of vector spaces indexed by points in the manifold, generalizing the concept of the tangent bundle. Tensor fields are then *smooth sections* of these bundles, i.e. maps from the manifold to the bundle whose value at any $q \in Q$ projects back to q. For our purposes, we may think of tensor fields as a position-dependent tensor whose components with respect to the co-ordinate or differential bases are all smooth functions. For example, the differential of a function defines a smooth covector field.

Definition 2.1.11. The set of all covector fields on a smooth manifold Q is written $\mathcal{X}(Q)$.

An important example of tensors in the context of statistical families is given by the Fisher information. Firstly, note that the score is the differential of a smooth function and is therefore a covector field. We write S_y to denote this covector field,

$$S_y := d\mathcal{L}_{n,y}$$

The score is therefore a covector field that depends on data $y \in \mathcal{Y}$; we could therefore define the score as a map $S : \mathcal{Y} \to \mathcal{X}(Q)$, where $\mathcal{X}(Q)$ is the set of covector fields on Q. Proposition 2.1.8 is then the statement that for all points $q \in Q$ and any vector $V \in T_q Q$, $E_{y \sim q} S_y(q)(V) = 0$. The score allows us to re-define the Fisher information in a co-ordinate free manner.

Definition 2.1.12. Let Q be a smooth statistical family and $q \in Q$. The Fisher information of $V, W \in T_q Q$ is the covariance of their respective scores, i.e.

$$I_q(V,W) = E_{q \sim Q} S_{q,y}(V) S_{q,y}(W)$$

Note that since the expected value of the score is 0 for any tangent vector, the covariance takes the form given in the de nition. The Fisher information is a tensor; it is linear in each of its two slots by the linearity of the score covector and the bilinearity of the covariance function on random variables. Note that De nition 2.1.12 agrees with our earlier De nition 1.2.16 of the Fisher information matrix; the components of the Fisher information at $q \in Q$ with respect to the co-ordinate vector basis are

$$\begin{aligned} I_q\left(\partial_i|_q, \partial_j|_q\right) &= E_{q\in Q}\left[S_{q,y}(\partial_i)\, S_{q,y}\left[\partial_j\right]\right] \\ &= E_{q\in Q}\left[\partial_i\left[L_{n,y}\right]\partial_j\left[L_{n,y}\right]\right] \\ &= E_{q\in Q}\left[\frac{\partial}{\partial q^i}\Big|_{q=f(q)} \log p(y|q)\, \frac{\partial}{\partial q^j}\Big|_{q=f(q)} \log p(y|q)\right] \end{aligned}$$

where $p(y|q)$ is the conditional data-generating density represented in the relevant co-ordinate chart.

Proposition 2.1.13. The Fisher information is a positive semi-de nite covariant 2-tensor eld, i.e. for all $q \in Q$,

$$I_q(V, V) \geq 0$$

with equality if $V = 0$.

Proof. We rst show that $I(0, 0) = 0$. Note that the action of the score on the 0 vector is 0;

$$S_{q,y}(0) = dL_{n,y}(0) = 0(L_{n,y}) = 0$$

where $L_{n,y}$ is the conditional density de ned in equation (2.1.5). By the de nition of the Fisher information, we have

$$I_q(0, 0) = E_{y\sim q}\left[S_{q,y}(0)\, S_{q,y}(0)\right] = E[0 \cdot 0] = 0$$

For a general tangent vector $V \in T_qQ$, we have

$$\begin{aligned} I_q(V, V) &= E_q\left[S_{q,y}(V) S_{q,y}(V)\right] \\ &= E_q\left[S_{q,y}(V)^2\right] \\ &\geq E_q\left[S_{q,y}(V)\right]^2 \geq 0 \end{aligned}$$

□

2.2 Divergence Functions

A divergence or contrast function is a measure of distance or proximity on a manifold, although in general they are not metrics. Divergences induce geometric structure - both classical Riemannian geometry and Information Geometry.

Firstly we formally de ne a divergence on a manifold. The de nition is given rst on $\mathbb{R}^p$ and then on manifolds in terms of co-ordinate charts. The following is from Calin and Udriste [17] and Amari [3];

De nition 2.2.1. Let $E \subseteq R^p$ be open. A divergence or contrast function on E is a smooth function $D : E \times E \to R$ satisfying

i) $D(q, q_0) \geq 0$ for all $q, q_0 \in E$

ii) For all $q, q_0 \in E$, $D(q, q_0) = 0$ if and only if $q_0 = q$

iii) The Hessian matrix along the diagonal $q_0 = q$, i.e. the matrix with $i \; j$ entry

$$(G(q_0))_{ij} := \frac{\partial}{\partial q^i}\frac{\partial}{\partial q^j} D(q, q_0)\Big|_{q=q_0} \tag{2.2.1}$$

is a strictly positive-de nite matrix-valued smooth function of $q \in E$.

Conditions i and ii are shared with the de nition of a metric on E, but divergences are not necessarily symmetric in their arguments nor do they satisfy the triangle inequality. The following lemma will be useful in later calculations and geometric de nitions; a proof is provided by Calin and Udriste [17].

Lemma 2.2.2. Let D be a divergence on an open set $E \subseteq R^p$. Then the first-order partial derivatives on the diagonal vanish:

$$\frac{\partial}{\partial q^i} D(q, q_0)\Big|_{q_0=q} = \frac{\partial}{\partial q_0^i} D(q, q_0)\Big|_{q_0=q} = 0 \qquad \text{for } i = 1, .., p \tag{2.2.2}$$

Furthermore, the following equivalent expressions for the Hessian on the diagonal (2.2.1) hold:

$$\begin{aligned}(G(q))_{ij} &= \frac{\partial}{\partial q^i}\frac{\partial}{\partial q^j} D(q, q_0)\Big|_{q=q_0} = \frac{\partial}{\partial q_0^i}\frac{\partial}{\partial q_0^j} D(q, q_0)\Big|_{q=q_0} \\ &= \frac{\partial}{\partial q_0^j}\frac{\partial}{\partial q^j} D(q, q_0)\Big|_{q=q_0} = \frac{\partial}{\partial q^i}\frac{\partial}{\partial q_0^j} D(q, q_0)\Big|_{q=q_0}\end{aligned}$$

The condition iii together with Lemma 2.2.2 imply that a divergence on R^p can be approximated with the quadratic form

$$D(q, q_0) = \frac{1}{2}(q - q_0)^T G(q_0)(q - q_0) + O\left(\|q - q_0\|^3\right) \tag{2.2.3}$$

as $q - q_0 \to 0$, where $G(q_0)$ is the matrix de ned in condition iii. This suggests we can interpret a divergence as half the squared distance from q_0 to q under the norm induced by $G(q_0)$. This, along with the fact that G is symmetric positive-de nite, will motivate an inner product on tangent spaces when we discuss Riemannian geometry in Section 2.3.1.

Example 2.2.3. Let $A \in \mathbb{R}^{p \times p}$ be a symmetric positive-definite matrix. Define a Euclidean divergence function $D_A : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ given by

$$D_A(q, q_0) := \frac{1}{2}(q - q_0)^T A (q - q_0)$$

The first two conditions i, ii follow immediately from the fact that A is positive-definite. Condition iii is satisfied since the Hessian on the diagonal is A.

We can now formally define a divergence function on a manifold. We define a divergence on a manifold by requiring that its co-ordinate representation is a divergence on an open set of $\mathbb{R}^p$.

Definition 2.2.4. Let Q be a smooth p-manifold. A divergence or contrast function on an open set $U \times Q$ is a smooth function $D : U \times U \to \mathbb{R}$, such that any co-ordinate chart (U, f) where $f : U \to \mathbb{R}^p$ makes

$$D_f : f(U) \times f(U) \to \mathbb{R}, \qquad D_f(q, q_0) := D(f^{-1}(q), f^{-1}(q_0))$$

a divergence function on the open subset $f(U) \times \mathbb{R}^p$ in the sense of definition 2.2.1.

This definition is from Calin and Udrişte [17], who also provide a proof (Theorem 11.2.1) of the following result, which shows that the choice of co-ordinate chart is arbitrary.

Theorem 2.2.5. Let Q be a smooth p-manifold, and let there be two co-ordinate charts (U, f) and (U', f'). Let $D : U \times U \to \mathbb{R}$ be a divergence with respect to the first co-ordinate chart in the sense of Definition 2.2.4, i.e. so that

$$D_f : f(U) \times f(U) \to \mathbb{R}, \qquad D_f(q, q_0) := D(f^{-1}(q), f^{-1}(q_0))$$

is a divergence on $f(U)$. Then the function

$$D_{f'} : f'(U \cap U') \times f'(U \cap U') \to \mathbb{R}, \qquad D_{f'}(q, q_0) := D(f^{-1}(q), f^{-1}(q_0))$$

is a divergence on $f'(U \cap U')$.

From the latter theorem and definition we deduce the following proposition, which allows us to define divergences on a manifold by extension of divergences in $\mathbb{R}^p$.

Proposition 2.2.6. Let Q be a smooth p-manifold with co-ordinate chart (U, q). Let $D_f : f(U) \times f(U) \to \mathbb{R}$ be a divergence on the co-ordinate range $f(U) \times \mathbb{R}^p$. Then the function

$$D : U \times U \to \mathbb{R}, \qquad D(q_1, q_2) := D_f(f(q_1), f(q_2))$$

is a divergence on $U \times Q$.

Proof. Let (U, f') be a co-ordinate chart with the same co-ordinate domain U, but an arbitrary co-ordinate map f' not necessarily the same as f. To show that D is a divergence on U, De nition 2.2.4 states that we must show that $D_{f'}(q', q_0') := D(f'^{-1}(q'), f'^{-1}(q_0'))$ is a divergence on $f'(U)$.

Note that by de nition

$$D(f^{-1}(q), f^{-1}(q_0)) = D_f(q, q_0)$$

for all $q, q_0 \in f(U)$. By Theorem 2.2.5, $D_{f'}$ is a divergence on $f'(U)$, and hence D is a divergence on U. □

Any divergence also has a dual divergence

De nition 2.2.7. Let D be a divergence on $E \subseteq \mathbb{R}^p$. The dual divergence D^* of D is

$$D^* : E \times E \to \mathbb{R} \qquad D^*(q, q_0) := D(q_0, q)$$

Note that any dual divergence is a divergence; conditions i and ii of De nition 2.2.1 are trivially satis ed, and condition iii holds by Lemma 2.2.2. A divergence is only equal to its dual when it is symmetric in its arguments, such as for the squared Euclidean distance.

2.2.1 The f-Divergences

We now study divergences on statistical families. There is a canonical class of divergences on a statistical family introduced by Csiszar [21].

De nition 2.2.8. Let Q be a statistical family on a measurable space $(Y, \mathcal{F})$. An f-divergence is a function of the form

$$D_f : Q \times Q \to \mathbb{R}, \qquad D_f(q_1, q_2) := \int f\left(\frac{dq_1}{dq_2}(y)\right) dq_2(y) \tag{2.2.4}$$

where $f : (0, \infty) \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$. If in addition $f''(1) = 1$, the f-divergence is called standard.

Since the elements q_1, q_2 of a statistical family are measures, i.e. probability distributions, the integral in (2.2.4) is well-de ned. Let us show that an f-divergence satis es the conditions of a divergence.

Theorem 2.2.9. An f-divergence is a divergence on Q. In a co-ordinate chart (U, q) with co-ordinates labelled q, if there exists a measure which dominates Q then the Hessian matrix on the diagonal is proportional to the Fisher Information,

$$\frac{\partial}{\partial q^i}\frac{\partial}{\partial q^j} D_f(q, q_0)\Big|_{q=q_0} = f''(1) E_{p(y|q_0)}\left[\frac{\partial}{\partial q^i}\log p(y|q)\frac{\partial}{\partial q^j}\log p(y|q)\right]_{q=q_0}$$

where $p(y|q_0)$ and $p(y|q)$ are respectively the densities with respect to of the measures $f^{-1}(q_0), f^{-1}(q) \in Q$. In particular, the Hessian is exactly the Fisher Information when the f-divergence is standard.

Proof. We check the three conditions of divergences.

i)

$$D_f(q_1, q_2) = \int_Y f\left(\frac{dq_1}{dq_2}(y)\right) dq_2(y) = E_{y\sim q}\left[f\left(\frac{dq_1}{dq_2}(y)\right)\right]$$

$$\geq f\left(E_{y\sim q}\left[\frac{dq_1}{dq_2}(y)\right]\right) \quad \text{by Jensen's inequality}$$

$$= f\left(\int_Y dp(y)\right) = f(1) = 0 \quad \text{since } p \text{ 2 } Q \text{ is a probability measure}$$

ii) Let $q_1 \neq q_2$. Then by the strict convexity of f, Jensen's inequality gives

$$D_f(q_1, q_2) = E_{y\sim q}\left[f\left(\frac{dq_1}{dq_2}(y)\right)\right] > f\left(E_{y\sim q}\left[\frac{dq_1}{dq_2}(y)\right]\right) = f(1) = 0$$

iii) Assume there exists a co-ordinate chart (U, q) whose co-ordinates we label $q \in f(U) \subseteq \mathbb{R}^p$. We write

$$D_f(q, q_0) = \int_Y f\left(\frac{p_0}{p}\right) p\, dy$$

for the induced divergence on $f(U)$, where $p_0 := p(y|q_0)$ and $p := p(y|q)$ are shorthand notation for the densities with respect to a measure n which dominates Q. Furthermore, let $\partial_i := \frac{\partial}{\partial q^i}$ denote the partial derivative with respect to the i^{th} component in the second slot of the divergence. We assume we can bring such differential operators under the integral sign with respect to y. Then

$$\partial_j D_f(q, q_0) = \int_Y f'\left(\frac{p_0}{p}\right)\left(-\frac{p_0}{p^2}(\partial_j p)\right) p\, dy + \int_Y f\left(\frac{p_0}{p}\right) \partial_j p\, dy$$

and by a direct calculation

$$\partial_i \partial_j D_f(q, q_0) = \int_Y f''\left(\frac{p_0}{p}\right)\left(-\frac{p_0}{p^2}(\partial_i p)\right)\left(-\frac{p_0}{p^2}(\partial_j p)\right) p\, dy \tag{2.2.5}$$

$$+ \int_Y f'\left(\frac{p_0}{p}\right)\left(\partial_i\left(-\frac{p_0}{p}(\partial_j p)\right) + \frac{p_0}{p^2}(\partial_i p)(\partial_j p)\right) dy \tag{2.2.6}$$

$$+ \int_Y f\left(\frac{p}{p_0}\right) \partial_i \partial_j p\, dy \tag{2.2.7}$$

We now evaluate this expression term-by-term at $q = q_0$, i.e. $p_0 = p$ in the shorthand notation. The third term (2.2.7) vanishes since $f(1) = 0$;

$$\int_Y f\left(\frac{p}{p_0}\right) \partial_i \partial_j p\, dy \Big|_{q=q_0} = 0$$

Similarly the second term (2.2.6) also vanishes;

$$\int_Y f'\left(\frac{p_0}{p}\right)\left[\partial_i\left(\frac{p_0}{p}\right)(\partial_j p) + \frac{p_0}{p^2}(\partial_i p)(\partial_j p)\right] dy \Big|_{q=q_0}$$

$$= \int_Y f'\left(\frac{p_0}{p}\right)\left[-\frac{p_0}{p^2}(\partial_i p)(\partial_j p) - \frac{p_0}{p}\partial_i\partial_j p + \frac{p_0}{p^2}(\partial_i p)(\partial_j p)\right] dy \Big|_{q=q_0}$$

$$= f'(1)\int_Y \partial_i\partial_j p\, dy \Big|_{q=q_0} = f'(1)\,\partial_i\partial_j \underbrace{\int_Y p\, dy}_{=1} = 0$$

Finally for the first term (2.2.5), note that $\frac{1}{p}\partial_i p = \partial_i \log p$. Thus we have

$$\int_Y f''\left(\frac{p_0}{p}\right)\left[-\frac{p_0}{p^2}(\partial_i p)\right]\left[\frac{p_0}{p^2}(\partial_j p)\right] p\, dy \Big|_{q=q_0} = \int_Y f''\left(\frac{p_0}{p}\right)\left[-\frac{p_0}{p}\right]^2 (\partial_i p)(\partial_j p)\, p\, dy \Big|_{q=q_0}$$

$$= f''(1)\int_Y (\partial_i \log p)|_{q=q_0}\,(\partial_j \log p)\Big|_{q=q_0} p_0\, dy$$

$$= f''(1)\,E_{p_0}\left[(\partial_i \log p)|_{q=q_0}\,(\partial_j \log p)\Big|_{q=q_0}\right]$$

Therefore the Hessian along the diagonal $q = q_0$ is proportional to the Fisher Information with proportionality constant $f''(1)$. When the divergence D_f is standard, $f''(1) = 1$ and the Hessian is exactly the Fisher information.

□

Therorem 2.2.9 shows that all standard f-divergences agree to second-order in any co-ordinate map in which they are defined. In fact, all f-divergences induce the same geometry - this will be clarified below.

Examples of f-Divergences

Let $f(u) = -\log(u)$, which clearly satisfies $f(1) = 0$ and is standard, i.e. $f''(1) = 1 \neq 0$. Then the corresponding f-divergence is the Kullback-Leibler divergence,

$$D_f(q_1, q_2) = \int_Y \log\left(\frac{dq_1}{dq_2}\right) dq_2$$

If $f(u) := u^2 - 1$, then the corresponding f-divergence is

$$D_f(q_1, q_2) = \int_Y \left(\frac{dq_1}{dq_2}\right)^2 - 1\, dq_2$$

Assuming both measures have respective densities p_1 and p_2 with respect to some dominating measure n, this can be written

$$D_f(q_1, q_2) = \int_Y \frac{(p_1 - p_2)^2}{p_2} dn$$

This is the Pearson c^2 distance, so-called because it is the test statistic in a c^2 test for categorical data.

2.2.2 The Bregman Divergence Induced by a Convex Function

The f-divergences introduced in the previous subsection are defined in terms of measures; they depend on a statistical family structure. In this subsection we introduce another class of divergences called Bregman divergences [3], [14] that do not make reference to points of the manifold being measure-valued. Any convex function on R^p induces a Bregman divergence on R^p; we can then therefore use convex functions in a co-ordinate chart to construct divergences on a manifold via Proposition 2.2.6. Finally, Bregman divergences can be made to include prior information by a suitable choice of convex function.

We begin with the definition of a Bregman divergence.

Definition 2.2.10. Let $F : R^p \to R$ be a smooth strictly convex function. The Bregman divergence corresponding to F is the function

$$D_F : R^p \times R^p \to R \qquad D_F(q, q_0) = F(q) - F(q_0) - \nabla F(q_0) \cdot (q - q_0)$$

This is also known as the Bregman divergence generated by F, and F in this context is a Bregman generator. Note that the Bregman divergence is the error at q in a linear Taylor approximation to F around q_0. There is inconsistency among authors around the ordering of the arguments q, q_0 in D_F; we have followed the convention used by Amari [3] of making the point being Taylor-expanded around the second argument. Other authors such as Calin and Udrişte [17] use the opposite ordering.

Proposition 2.2.11. The Bregman divergence D_F corresponding to any smooth convex function $F : R^p \to R$ is a divergence on R^p.

Proof. We check the conditions of divergence from Definition 2.2.1.

i) and ii). Non-negativity $D_F(q, q_0) \geq 0$ is equivalent to the first-order condition of a convex function F;

$$F(q) \geq F(q_0) + \nabla F(q_0) \cdot (q - q_0) \qquad \text{for all } q, q_0 \in R^p \tag{2.2.8}$$

with equality if and only if $q_0 = q$. This is a consequence of convexity for a differentiable function F; see Boyd and Vandenberghe [13].

iii) The Hessian on the diagonal of $D_F(q, q_0)$ is precisely the Hessian matrix of F,

$$\frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} D(q, q_0)\Big|_{q=q_0} = \frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} F(q)\Big|_{q=q_0}$$

Since F is strictly convex, the second-order condition [13] guarantees that the Hessian of F is strictly positive-definite.

□

Suppose the smooth manifold Q is equipped with a global co-ordinate chart (Q, f) where $f : Q \to \mathbb{R}^p$. A Bregman divergence D_F on the co-ordinates of the chart induces a divergence D on the manifold by Proposition 2.2.6, and this divergence has representations in other co-ordinate charts. However, the induced divergence may not necessarily be of Bregman type in other charts

Bregman Divergence on Exponential Families

Recall from Section 1.2 that a standard exponential family contains probability measures of the form

$$\exp\left(q^T t - y(q)\right) dr(t)$$

when parameterized in terms of natural parameters q.

Proposition 2.2.12. The Kullback-Leibler divergence between two elements of an exponential family written in the natural parameterization is the Bregman divergence of the cumulant generating function y.

Proof. The Kullback-Leibler divergence from P_{q_0} to P_q, both elements in the same standard exponential family, is given by

$$\begin{aligned} D(q_0, q) &= \int \log \frac{p(t|q_0)}{p(t|q)} dP_{q_0}(t) = \int (q_0 - q)\, t - y(q_0) - q\, t + y(q) dP_{q_0}(t) \\ &= \int (q_0 - q)\, t + y(q) - y(q_0) dP_{q_0}(t) \\ &= (q_0 - q) \underbrace{\int t dP_{q_0}(t)}_{= r\, y(q_0)} + (y(q) - y(q_0)) \underbrace{\int dP_{q_0}(t)}_{= 1} \\ &= y(q) - y(q_0) - r\, y(q_0)\, (q - q_0) \end{aligned}$$

which is clearly the Bregman divergence generated by y. In the third line, we made use of the fact that P_{q_0} is a probability distribution and Theorem 1.2.12, i.e. the expectation under P_{q_0} of the sufficient statistic is given by $r\, y(q_0)$. □

Dual Co-ordinates

In Section 1.2.2, we discussed the Legendre transform F^* of a convex function F, whose argument is the dual variable $h = \nabla F(q)$. If (U, q) is a co-ordinate chart on a smooth manifold Q, then a convex function $F : f(U) \to \mathbb{R}$ defines a dual co-ordinate chart (U, h), where

$$h : U \to \mathbb{R}^p, \qquad h(q) = \nabla F(q(q))$$

The Legendre transform $F^*(h)$ is a convex function of h, and hence defines its own Bregman divergence. The Bregman divergences induced by F and F^* are the dual divergences of each other, as the following proposition shows.

Proposition 2.2.13. Let $E \subseteq \mathbb{R}^p$ be open and $F : E \to \mathbb{R}$ be convex. For $q, q_0 \in E$, let $h = \nabla F(q)$ and $h_0 = \nabla F(q_0)$. The Bregman divergence D_F satisfies

$$D_F(h_0, h) = D_{F^*}(q, q_0) = D_F(q, q_0)$$

where F^* is the Legendre dual of F.

Proof. By the definition of the Bregman divergence for F^*, we have

$$D_{F^*}(h_0, h) = F^*(h) - F^*(h_0) - \nabla F^*(h_0) \cdot (h - h_0) \tag{2.2.9}$$

Recall the definition of F^* from equation (1.2.14) in Chapter 1,

$$F^*(h_0) := q_0 \cdot h_0 - F(q_0) \qquad \text{where } q_0 \text{ is such that } h_0 = \nabla y(q_0) \tag{2.2.10}$$

Substituting (2.2.10) and the equivalent statement for $F^*(h)$ into (2.2.9) yields

$$D_{F^*}(h_0, h) = q \cdot h - F(q) - q_0 \cdot h_0 + F(q_0) - \nabla F^*(h_0) \cdot h + \nabla F^*(h_0) \cdot h_0$$

By Theorem 1.2.15, $\nabla F^*(h_0) = q_0$, and hence

$$\begin{aligned} D_{F^*}(h_0, h) &= q \cdot h - F(q) - q_0 \cdot h_0 + F(q_0) - q_0 \cdot h + q_0 \cdot h_0 \\ &= F(q_0) - F(q) - h \cdot (q - q_0) \\ &= F(q_0) - F(q) - \nabla F(q) \cdot (q - q_0) \\ &= D_F(q, q_0) = D_F(q, q_0) \end{aligned}$$

where the last equality is from the definition of the dual divergence. □

The following proposition shows how we may write a Bregman divergence such that one point is represented in the q co-ordinate system, and the other in the h co-ordinate system.

Proposition 2.2.14. Let $E \subseteq \mathbb{R}^p$ be open and $F : E \to \mathbb{R}$ be convex. The Bregman divergence D_F with generator F can be written

$$D_F(q, q_0) = F^*(h_0) + F(q) - q \cdot h_0 \tag{2.2.11}$$

where $h_0 := \nabla F(q_0)$, F^* is the Legendre dual of F, and $\cdot$ denotes the Euclidean dot product on E.

Proof. Substituting (2.2.10) into the right-hand side of (2.2.11) yields the result,

$$\begin{aligned} F\ (h_0) + F(q)\quad q\ \ h_0 &= q_0\ \ h_0\quad F(q_0) + F(q)\quad q\ \ h_0 \\ &= F(q)\quad F(q_0) + r\ F(q_0)\ \ (q_0\quad q) \\ &= D_F(q, q_0) \end{aligned}$$

□

2.3 Geometry from Divergence Functions

In this section we derive various geometric objects from a divergence. These involve taking partial derivatives of the divergence, or equivalently applying vector fields to the smooth function defined by the divergence. To aid with this, we employ a helpful notation introduced by Amari et al [58].

Suppose $D : Q \quad Q \to R$ is a divergence on a smooth manifold Q. Fix $q_1 \in Q$ and think of $D(q_1, q_2)$ as a smooth function of only q_2;

$$D(q_1, \ ) : Q \to R \qquad D(q_1, \ )(q_2) := D(q_1, q_2)$$

We may apply a vector field $X \in \mathfrak{X}(Q)$ to this function to yield a new smooth function $X_{;2}D$ which can be evaluated at q_2. This is still a function of q_1 and q_2,

$$X_{;2}D : Q \quad Q \to R \qquad (X_{;2}D)(q_1, q_2) := X(D(q_1, \ ))(q_2)$$

Of course, we can also form the smooth function $X_{;1}D$ by fixing q_2 and allowing q_1 to vary; we use the subscript $;1$ or $;2$ to emphasize which slot of the divergence the vector field is acting on. Note that Lemma 2.2.2 implies

$$(X_{;1}D)(q, q) = (\ X_{;2}D)(q, q) = 0$$

evaluating the new functions produced by the vector field at the same point $q = q_1 = q_2$ is zero. This is trivial to show by expanding X in the co-ordinate vector basis.

Since $X_{;1}D$ and $X_{;2}D$ are both functions on the product space $Q \quad Q$, we can repeat this process; fix a point and apply a vector field with respect to one of the variables. For example, for vector fields $X, Y \in \mathfrak{X}(Q)$, write $X_{;1}Y_{;2}D$ for the function produced by first fixing q_1 and applying Y with respect to q_2, then fixing q_2 and applying X with respect to q_1.

$$X_{;1}Y_{;2}D : Q \quad Q \to R \qquad (X_{;1}Y_{;2}D)(q_1, q_2) := X\left((Y_{;2}D)(\ , q_2)\right)(q_1)$$

Amari et al [58] introduce the following simplifying notation. For the $r + s$ vector fields $X^1, ..., X^r, Y^1, ..., Y^s$, define the smooth function

$$D_{;\ X^1\quad X^r, Y^1\quad Y^s} : Q \to R$$

$$D_{;\ X^1\quad X^r, Y^1\quad Y^s}(q) = X^1_{;1}\quad X^r_{;1}Y^1_{;2}\quad Y^s_{;2}D\ \ (q, q)$$

In other words, we apply the vector fields in the indexed order via the process described above, and then evaluate the resulting function at $q_1 = q_2 = q$.

2.3.1 Riemannian Geometry

To do Euclidean geometry in a vector space such as $\mathbb{R}^p$, we only require an inner-product to measure lengths and angles. The manifold Q is locally approximated around each $q \in Q$ by the vector space T_qQ. This motivates equipping every tangent space with its own inner-product so that we can do geometry locally on every tangent space, and then define notions like length and area over global curves and surfaces by integrating over tangent spaces. This is the core idea of Riemannian geometry.

An inner-product on a vector space is a symmetric positive-definite bilinear function of two vectors; it is therefore a symmetric positive-definite $(0,2)$-tensor. On a manifold, this generalises to the following:

Definition 2.3.1. Let Q be a smooth manifold. A *Riemannian metric* is a smooth $(0,2)$-tensor field g on Q which satisfies for all $q \in Q$:

1. Symmetry: $g_q(V,W) = g_q(W,V)$ for all tangent vectors $V, W \in T_qQ$.
2. Positive-definiteness: $g_q(V,V) \geq 0$ for all $V \in T_qQ$, with equality if and only if V is the zero tangent vector.

Here g_q is the $(0,2)$-tensor-valued evaluation of g at $q \in Q$.

A *Riemannian manifold* is the pairing of a smooth manifold Q with a particular metric g, i.e. the tuple (Q,g). The components of the metric at $q \in Q$ in a particular co-ordinate chart are the p^2 numbers $g_{ij} := g_q(\partial_i, \partial_j), i,j = 1,\ldots,p$. Hence the co-ordinate representation of the inner-product is

$$g_q(V,W) = \sum_{i,j=1}^{d} V^i W^j g_q(\partial_i, \partial_j) = v^T G(q) w$$

where $G(q) = (g_q(\partial_i, \partial_j))_{i,j=1}^{p}$ for co-ordinates $q = q(q)$, and of course $v = V^i$ and $w = W^i$. As usual, we will surpress $q \in Q$ in our notation and simply write g when q is clear from context, or when expressions hold over the entire manifold. Clearly $G(q)$ is a symmetric positive-definite matrix.

Since g is an inner-product, it enforces a duality pairing, i.e. an isomorphism between T_qQ and its dual T_q^*Q. These spaces are vector spaces of the same dimension and were therefore always isomorphic, but there is no canonical isomorphism until we introduce the metric. For any tangent vector $V \in T_qQ$, define its dual as the linear functional $\hat{V} \in T_q^*Q$, such that $\hat{V}(W) := g(V,W)$ for any vector W. Expanding this expression in the dual bases with $\hat{V} = V_i dq^i$ yields

$$\begin{aligned} V_i dq^i \, W^j \partial_j &= g\left(V^l \partial_l, W^j \partial_j\right) \\ \Rightarrow \quad V_i W^j \delta^i_j &= V^l W^j g_{lj} \\ \Rightarrow \quad V_j W^j &= V^l W^j g_{lj} \end{aligned}$$

Since this expression holds for all vectors W, it follows that $V_j = g_{lj} V^l$, where Einstein summation applies as usual. In co-ordinates, this implies that the components of the dual covector are given by $\flat v = G(q)^T v = G(q) v$ since $G(q)$ is symmetric. Given an arbitrary covector $a \in T_q^* Q$, we deduce that the dual vector, i.e. given by the inverse of the isomorphism implicitly defined above, must be $V = g^{ij} a_j \partial_i$ where g^{ij} are the elements of the inverse matrix $G^{-1}(q)$. Indeed, for any vector W we recover

$$\begin{aligned} V^\flat(W) = g(V, W) &= g^{ij} a_j W^k g(\partial_i, \partial_k) \\ &= g^{ij} a_j W^k g_{ik} = a_j W^k \delta^j_k \\ &= a_j W^j = a(W) \end{aligned}$$

One can show that g^{ij} are the components of the inner-product on $T_q^* Q$ induced by g. In conclusion, we have the following relationships between components of dual vectors and covectors,

$$V^j = g^{ij} V_j \qquad V_i = g_{ij} V^j \tag{2.3.1}$$

The metric and its inverse respectively lower and raise indices, with the corresponding musical isomorphisms called *flat* and *sharp* from tangent space to cotangent space and vice versa.

Length and Angles

Clearly the Riemannian metric is an inner product on each tangent space $T_q Q$; this was our motivation for defining the metric. We can therefore make sense of concepts such as the norm of a tangent vector $V \in T_q Q$,

$$\|V\|_g^2 := g(V, V)$$

and the angle θ between two tangent vectors V, W,

$$\cos\theta := \frac{g(V, W)}{\|V\|_g \|W\|_g}$$

In particular, V and W are *orthogonal* when $g(V, W) = 0$. We can also define the length of a curve. Recall that if $x : (t_0, t_1) \to \mathbb{R}^p$ is a smooth curve in $\mathbb{R}^p$ equipped with the usual Euclidean product, then the arc-length is given by

$$S = \int_{t_0}^{t_1} \left\| \frac{dx}{dt} \right\|_2 dt$$

If $c : (t_0, t_1) \to Q$ is a smooth curve in a Riemannian manifold Q with metric g, this motivates the following definition of arc-length,

$$S = \int_{t_0}^{t_1} \|c'(t)\|_g dt = \int_{t_0}^{t_1} \sqrt{g(c'(t), c'(t))}\, dt$$

Metric from Divergence

Given that a Riemannian metric and a divergence both correspond to a form of squared distance measure, it is natural to ask whether a Riemannian metric can be induced by a divergence. Recall that a divergence D on $\mathbb{R}^p$ can be expanded to second-order as

$$D(q, q_0) \quad \frac{1}{2}(q \quad q_0)^T G_D(q)(q \quad q_0)$$

where

$$(G_D(q_0))_{ij} := \frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} D(q, q_0)\Big|_{q=q_0} \tag{2.3.2}$$

Clearly, this motivates a Riemannian metric whose components form the Hessian matrix on the diagonal of the divergence, $G_D(q)$. Ay et al [58] give the following co-ordinate independent de nition of this metric:

De nition 2.3.2. Let Q be a smooth manifold, and let $D : Q \quad Q \to \mathbb{R}$ be a divergence. The Riemannian metric induced by D is the $(0,2)$-tensor eld

$$g^{(D)}(V, W) = \quad D_{;}(V, W)$$

for all $V, W \in \mathcal{X}(Q)$.

Note that this de nition implies that for any $q \in Q$,

$$g^{(D)}(V, W)(q) = \quad (V_{;1}, W_{;2} D)(q)$$

As expected, the co-ordinate representation of the metric is simply the Hessian matrix G_D from equation (2.3.2) as shown in the following proposition.

Proposition 2.3.3. The tensor eld $g^{(D)}$ is a Riemannian metric whose components in a co-ordinate chart (U, f) are given by the Hessian on the diagonal of the divergence in the co-ordinate chart,

$$g^{(D)}(\partial_i, \partial_j) := \frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} D_f(q, q_0)\Big|_{q=q_0}$$

where $D_f(q, q_0) := D(f^{-1}(q), f^{-1}(q_0))$ is the co-ordinate representation of D.

Proof. Computing the i-j component in the co-ordinate chart gives

$$\begin{aligned} g^{(D)}(\partial_i, \partial_j)(q) &= \quad \partial_{i;1}\partial_{j;2} D \quad (q) \\ &= \quad \frac{\partial}{\partial q^i}\frac{\partial}{\partial q_0^j} D_f(q, q_0)\Big|_{q=q_0} \end{aligned}$$

where $q_0 = f(q)$. By Lemma 2.2.2, this is equal to the Hessian on the diagonal,

$$g^{(D)}(\partial_i, \partial_j)(q) = \frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} D_f(q, q_0)\Big|_{q=q_0}$$

To see that $g^{(D)}$ is a Riemannian metric, let $v = (V^i)$ and $w = (W^j)$ be the components of two vector fields V, W in the co-ordinate chart. Writing $g^{(D)}(V, W)$ in co-ordinates and applying the product rule, we find

$$\begin{aligned} g^{(D)}(V,W) &= g^{(D)}\left(V^i\partial_i, W^j\partial_j\right) \\ &= V^i(q)\frac{\partial}{\partial q^i}\left(W^j(q)\frac{\partial}{\partial q_0^j}D_f(q,q_0)\right)_{q=q_0} \\ &= V^i(q)\frac{\partial W^j}{\partial q^i}\left(\frac{\partial}{\partial q_0^j}D_f(q,q_0)\right)_{q=q_0} + V^i(q)W^j(q)\frac{\partial}{\partial q^i}\left(\frac{\partial}{\partial q_0^j}D_f(q,q_0)\right)_{q=q_0} \end{aligned}$$

By Lemma 2.2.2, the first term in the final equation vanishes, and hence

$$g^{(D)}(V,W) = V^i(q)W^j(q)g^{(D)}\left(\partial_i,\partial_j\right) = v(q)^T G_D(q)w(q)$$

In co-ordinates the inner-product of V, W therefore has the form $v^T G_D w$, where $G^{(D)}$ is the Hessian on the diagonal defined in equation (2.3.2). Clearly $G^{(D)}$ is a symmetric matrix, and is positive-definite by Definition 2.2.1 of a divergence on $\mathbb{R}^p$. It follows that $g^{(D)}$ is a Riemannian metric. □

Any divergence therefore gives Q the structure of a Riemannian manifold. We review the induced metric components for f- and Bregman divergences.

Let D_f be an f-divergence on Q. By Theorem 2.2.9 the components of the induced metric are proportional to the components of the Fisher information,

$$g_{ij}^{(D_f)}(q_0) = f''(1)I_{ij}(q_0)$$

Let D_F be a Bregman divergence on $\mathbb{R}^p$ with Bregman generator $F : \mathbb{R}^p \to \mathbb{R}$. A simple calculation shows that the Hessian on the diagonal, and hence the matrix of an induced Riemannian metric on $\mathbb{R}^p$, is given by the Hessian matrix of F,

$$g_{ij}^{(D_F)}(q_0) = \frac{\partial}{\partial q^i}\frac{\partial}{\partial q^j}F(q)_{q=q_0}$$

If Q is a smooth manifold with a global co-ordinate chart (U, q), by Proposition 2.2.6 the Bregman divergence induces a divergence, and hence a Riemannian metric, on Q.

2.3.2 Connections

Equipping a smooth manifold with a Riemannian metric allows us to define concepts of length and angle between tangent vectors. However, there are further geometric concepts that are not defined with a metric alone; for example, the notion of a smooth

curve being straight, analogous to a straight line in Euclidean geometry. A reasonable de nition of straightness might be that the velocity vector $c'(t)$ of the curve does not change over a suf ciently small step along the curve, i.e. from $q = c(t)$ to $q = c(t + dt)$. However, this requires being able to compare vectors from two distinct tangent spaces, namely $T_{c(t)}Q$ and $T_{c(t+dt)}Q$. We therefore require a connection which connects tangent spaces for in nitesimally close points in Q. Connections are extra information on a manifold in addition to the Riemannian metric, and enable us to differentiate vector elds with respect to other vector elds and de ne curvature. Furthermore, the choice of connection(s) on a smooth manifold is what separates Riemannian geometry from information geometry.

Consider how we might de ne the derivative of a vector eld Y with respect to a different vector eld X. This is closely related to the idea of connecting tangent spaces, because we need to somehow compare $Y(q) \in T_qQ$ with $Y(q) \in T_qQ$ in order to nd the rate of change of Y, where q is a point close to q in the direction of $X(q)$. We could de ne such a derivative in terms of its properties; e.g. linearity and an appropriate form of the product rule. This brings us to the following de nition of a connection, adapted from the de nitions given by Calin [17] and Lee [45]. Recall that $C^\infty(Q)$ and $\mathfrak{X}(Q)$ are respectively the sets of smooth functions and vector elds on Q.

De nition 2.3.4. Let Q be a smooth manifold. A connection ∇ on TQ is a map $\nabla : \mathfrak{X}(Q) \times \mathfrak{X}(Q) \to \mathfrak{X}(Q)$, written $\nabla_X Y$ for $X, Y \in \mathfrak{X}(Q)$, satisfying

i) Linearity over $C^\infty(Q)$ in X;

$$\nabla_{f_1X_1 + f_2X_2} Y = f_1 \nabla_{X_1} Y + f_2 \nabla_{X_2} Y$$

for all $f_1, f_2 \in C^\infty(Q)$ and all $X_1, X_2, Y \in \mathfrak{X}(Q)$.

ii) Linearity over $\mathbb{R}$ in Y;

$$\nabla_X (a_1Y_1 + a_2Y_2) = a_1 \nabla_X Y_1 + a_2 \nabla_X Y_2$$

for all $a_1, a_2 \in \mathbb{R}$ and all $X, Y_1, Y_2 \in \mathfrak{X}(Q)$.

iii) The Leibniz or product rule;

$$\nabla_X (fY) = (Xf)Y + f \nabla_X Y$$

for all $f \in C^\infty(Q)$ and all $X, Y \in \mathfrak{X}(Q)$.

The vector $\nabla_X Y$ is the covariant derivative of Y in the direction of X. We can compute its components in a co-ordinate vector basis:

$$\begin{aligned} \nabla_X Y &= \nabla_{X^i \partial_i} Y^j \partial_j \\ &= X^i \nabla_{\partial_i} Y^j \partial_j && \text{by linearity over } C^\infty(Q) \text{ in } X \\ &= X^i \partial_i Y^j \partial_j + Y^j \nabla_{\partial_i} \partial_j && \text{by the product rule} \end{aligned} \tag{2.3.3}$$

The covariant derivative is therefore determined by the regular partial derivative, plus a term involving the connection applied to the basis vectors, $\nabla_{\partial_i}\partial_j$. We follow the convention of writing this in terms of connection coefficients $\Gamma_{ij}^k \in C^\infty(Q)$, defined such that $\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \partial_k$. Changing the dummy index j to k in the first term in (2.3.3) yields the co-ordinate form of the covariant derivative,

$$\nabla_X Y = X^i \left(\frac{\partial Y^k}{\partial q^i} + \Gamma_{ij}^k Y^j \right) \partial_k$$

k^{th} component

Geodesics

We return to one of our motivating examples for connections from the beginning of this subsection; defining a notion of *straightness* for a smooth curve $c : J \to Q$, J an open interval in $\mathbb{R}$, such that the velocity vector $c'(t)$ is unchanged over small steps dt along the curve. Since the velocity vector at t represents the direction tangential to the curve, it seems reasonable to characterize such curves with the condition $\nabla_{c'(t)} c'(t) = 0$; the covariant derivative of the velocity vector along the curve, i.e. in the direction of the velocity vector, is zero. However, $c'(t)$ is not strictly a vector field defined over all of Q, but rather a vector field along a curve, and so the covariant derivative is not well-defined. Lee [45] shows that given a connection ∇, there is a unique total derivative operator $\nabla_{c'(t)}$ along any smooth curve $c(t)$ satisfying analogous properties to those in Definition 2.3.4 and matches the covariant derivative for ∇ when the vector fields along a curve can be extended to regular vector fields defined over all of Q. We shall omit the details here for brevity, and state that one of the properties of the total derivative along a curve $c(t)$ is the following product rule:

$$\nabla_{c'(t)} fV = f'V + f \nabla_{c'(t)} V$$

where $V : J \to TQ$ is a vector field along $c(t)$, i.e. satisfying $V(t) \in T_{c(t)}Q$ for all $t \in J$, and $f : J \to \mathbb{R}$ with derivative $\frac{df}{dt} \equiv f'$.

We are now able to make sense of curves $c(t)$ satisfying $\nabla_{c'(t)} c'(t) = 0$, which are called *geodesics*. Suppose there is some co-ordinate chart containing the entire curve, and write $c'(t) = q^i \partial_i$ for the velocity vector in co-ordinates. We compute the condition for geodesics in co-ordinates;

$$\begin{aligned}
\nabla_{c'(t)} c'(t) = 0 &\iff \nabla_{q^i \partial_i} q^j \partial_j = 0 \\
&\iff \dot{q}^i \partial_j + q^j \nabla_{q^i \partial_i} \partial_j = 0 \\
&\iff \dot{q}^i \partial_j + q^j q^i \nabla_{\partial_i} \partial_j = 0 \\
&\iff \dot{q}^k + q^j q^i \Gamma_{ij}^k \partial_k = 0
\end{aligned}$$

Writing time derivatives explicitly, the geodesics satisfy the system of p second-order ordinary differential equations,

$$\frac{d^2q^k}{dt^2}(t) + G^k_{ij}(q(t))\frac{dq^i}{dt}(t)\frac{dq^j}{dt}(t) = 0 \qquad i = 1,..,p \tag{2.3.4}$$

called the geodesic equations. A parameterized solution of these equations is called a geodesic flow, usually exact analytic solutions are not available. For any point $q \in Q$ and vector $V \in T_qQ$, one can show [45] that there exists a unique geodesic $c(t)$ such that $c(t_0) = q$, $c'(t_0) = V$. Clearly, different connections may produce different geodesics.

For a general vector field $V(t)$ along a fixed curve $c(t)$, we say V is parallel along $c(t)$ if $\nabla_{c'(t)}V(t) = 0$ for all $t \in J$. Visually, we can picture $V(t)$ as an arrow pointing in the same direction at each point on $c(t)$. In co-ordinates, the condition for parallelism along $c(t)$ takes the form

$$V^k + V^j q^i G^k_{ij} = 0 \qquad k = 1,..,p$$

where $q^i(t)$ are the parameterized co-ordinates for $q = c(t)$. These are p linear ordinary differential equations for the p components of $V(t)$. By taking an initial vector $V(0) = V_0 \in T_{c(0)}Q$, we can solve the system of ODEs to obtain a unique parallel vector field over the entire curve $c(t)$. This is known as parallel transport of V_0, which we write as

$$V(t) = P^{\nabla}_{c(0)\to c(t)}V(0)$$

Geodesics are commonly known as curves which have extremal, i.e. least, arc-length. In fact this holds for a particular connection that is uniquely defined by the Riemannian metric, known as the Levi-Civita connection, but may not hold for general connections. The Levi-Civita connection uniquely satisfies two important properties which we now state.

Definition 2.3.5. Let (Q, g) be a Riemannian manifold. A connection ∇ on TQ is called metrical, or simply a metric connection, if it satisfies

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z) \tag{2.3.5}$$

for all vector fields $X, Y, Z \in \mathfrak{X}(Q)$.

In co-ordinates, a metric connection satisfies $\partial_i g_{ij} = G^l_{ij}g_{lk} + G^l_{ik}g_{jl}$, or equivalently $\partial_i g_{ij} = G_{ijk} + G_{ikj}$ where $G_{ijk} := G^l_{ij}g_{lk}$. We state without proof that parallel transport is an isometry between tangent spaces for metrical connections [45], i.e. for a smooth curve $c(t)$ we have

$$g_{c(0)}(V,W) = g_{c(t)}\left(P^{\nabla}_{c(0)\to c(t)}V, P^{\nabla}_{c(0)\to c(t)}W\right) \qquad \text{for all } V, W \in T_{c(0)}Q$$

If $X, Y \in \mathfrak{X}(Q)$ are two vector fields, their Lie bracket is the vector field $[X, Y]$ which acts on a smooth function $f \in C^\infty(Q)$ via

$$[X,Y]f := XYf - YXf$$

The Lie bracket measures the commutativity of two vector fields; for example, by the equality of mixed partial derivatives we have $[\partial_i, \partial_j] = 0$.

De nition 2.3.6. Let Q be a smooth manifold. A connection r on TQ is called symmetric or torsion-free if

$$r_X Y - r_Y X = [X, Y]$$

for all $X, Y \in \mathfrak{X}(Q)$.

Substituting co-ordinate vector fields for X, Y yields the symmetry condition in co-ordinate form,

$$\Gamma^k_{ij} = \Gamma^k_{ji} \qquad \text{for all } i, j, k = 1, .., p \tag{2.3.6}$$

As stated, these two properties uniquely characterize the Levi-Civita connection. This is stated in the following theorem from Lee [45].

Theorem 2.3.7 (The Fundamental Theorem of Riemannian Geometry). Let (Q, g) be a Riemannian manifold. There exists a unique symmetric metric connection on TQ called the Levi-Civita connection .

A geodesic with respect to the Levi-Civita connection is exactly the curve with minimal arc-length among all curves with the same endpoints; see Chapter 6 of Lee [45].

Information Geometry: Connections from Divergence

Riemannian geometry typically utilizes only the Levi-Civita connection because of its unique properties of symmetry, compatibility with the metric and extremal geodesics. Information geometry on the other hand has an interesting structure that is unique among geometries; there are two dually coupled connections induced by a divergence, neither of which coincide with the Levi-Civita connection corresponding to $g^{(D)}$ in general. The induced connections have interesting statistical interpretations. The following de nition is from Ay et al [58].

De nition 2.3.8. Let Q be a smooth manifold with divergence $D : Q \times Q \to \mathbb{R}$. The connections $r^{(D)}, r^{(D^*)}$ respectively induced by D and its dual D^*, are given by

$$g^{(D)}\left(r^{(D)}_X Y, Z\right) = D_{;}(XY, Z)$$

$$g^{(D)}\left(r^{(D^*)}_X Y, Z\right) = D^*_{;}(XY, Z) = D_{;}(Z, XY)$$

for vector fields $X, Y, Z \in \mathfrak{X}(Q)$.

Lemma 2.3.9. The connection coef cients for $r^{(D)}$ and $r^{(D^*)}$ in a co-ordinate chart are given by

$$\Gamma^{(D)}_{ijk}(q) = \left.\frac{\partial^2}{\partial q^i \partial q'^j} \frac{\partial}{\partial q'^k_0} D(q, q_0)\right|_{q = q_0} \tag{2.3.7}$$

$$\Gamma^{(D^*)}_{ijk}(q) = \left.\frac{\partial^2}{\partial q^i_0 \partial q'^j_0} \frac{\partial}{\partial q'^k} D(q, q_0)\right|_{q = q_0} \tag{2.3.8}$$

Proof. See Ay et al [58]. □

One can verify that these are indeed connections by checking that the connection coefficients transform appropriately under changes of co-ordinates; see for example Amari [3]. In fact, the connections are symmetric.

Proposition 2.3.10. Let Q be a smooth manifold with divergence $D : Q \times Q \to \mathbb{R}$. The induced connections ∇, ∇^* are symmetric.

Proof. We will show the result in co-ordinate form, i.e. showing equation (2.3.6) holds for both ∇, ∇^*. By the interchangeability of mixed partial derivatives, it follows from Lemma 2.3.9 that $\Gamma^{(D)}_{ijk} = \Gamma^{(D)}_{jik}$ and $\Gamma^{(D^*)}_{ijk} = \Gamma^{(D^*)}_{jik}$. Multiplying by g^{lk} and contracting over k to raise the index yields the result in co-ordinate form, i.e. equation (2.3.6). □

Given that the induced connections are symmetric, one might wonder whether they are also metric connections, and hence coincide with Levi-Civita. This is not the case, but it turns out the connections are *dually coupled* with respect to the metric.

Definition 2.3.11. The connections ∇, ∇^* on a smooth manifold Q are called *dual* with respect to a Riemannian metric g if

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \tag{2.3.9}$$

for all vector fields $X, Y, Z \in \mathfrak{X}(Q)$.

We refer to ∇ and ∇^* respectively as the primal and dual connections, although we note that the choice of which is primal and which dual is arbitrary by the involutive nature of duality. If one parallel transports V, W with respect to the primal and dual connections respectively, the metric is preserved; see Nielsen [57].

$$g_{c(0)}(V,W) = g_{c(t)}\left(P^{\nabla}_{c(0)\to c(t)} V, P^{\nabla^*}_{c(0)\to c(t)} W\right) \qquad \text{for all } V, W \in T_{c(0)}Q$$

Clearly dual connections generalize metric connections; the Levi-Civita connection is self-dual, since we can replace both ∇ and ∇^* with the Levi-Civita connection in the above equation. The following theorem shows duality for the connections induced by a divergence, and is proven in Ay et al [58].

Theorem 2.3.12. Let Q be a smooth manifold with divergence $D : Q \times Q \to \mathbb{R}$. The induced connections $\nabla^{(D)}, \nabla^{(D^*)}$ are dual with respect to the induced metric $g^{(D)}$.

The structure (Q, g, ∇, ∇^*) such that ∇, ∇^* are dual with respect to g is called a *conjugate connections manifold* [57]. We have shown that a divergence on a manifold induces this structure.

2.4 Dually Flat Geometry

2.4.1 Curvature and Flatness

In general, the parallel transport of a vector depends on the curve $c(t)$. In particular, the parallel transport of a vector around a closed loop, i.e. from $q = c(0)$ back to $q = c(1)$, may not coincide with the original vector; in Euclidean space however, one observes that parallel transport is not dependent on the curve, and transporting a vector around a loop leaves it unchanged. A notion of how the manifold is curved could therefore be defined based on how much vectors are changed by parallel transport around a loop. Amari [3] and Ay et al [58] both show that the parallel transport of a vector around an infinitesimally small loop leaves the vector unchanged if and only if the curvature tensor of the manifold is zero.

Definition 2.4.1. Let Q be a smooth manifold with a connection ∇. The Riemann-Christoffel curvature tensor is the map

$$R : \mathcal{X}(Q) \times \mathcal{X}(Q) \times \mathcal{X}(Q) \to \mathcal{X}(Q)$$
$$R(X, Y, Z) = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

Note that R is usually interpreted as a $(1,3)$-tensor; it takes three vector fields X, Y, Z and a covector field w as input, and produces a smooth function $w(R(X, Y, Z))$. A manifold is called ∇-flat if and only if the curvature tensor defined by ∇ vanishes everywhere, i.e. $R_q = 0$ for all $q \in Q$. One can express the curvature tensor in local co-ordinates via the following formula - see for example Ay et al [58],

$$R^k_{lij} = \partial_i \Gamma^k_{jl} - \partial_j \Gamma^k_{il} + \Gamma^k_{im}\Gamma^m_{jl} - \Gamma^k_{jm}\Gamma^m_{il} \tag{2.4.1}$$

A co-ordinate chart (U, φ) is called affine if the connections coefficients vanish at all points $q \in U$, $\Gamma_{ijk}(\varphi(q)) = 0$ for all $i, j, k = 1, .., p$. Equivalently, the covariant derivative of the co-ordinate vector fields vanishes along co-ordinate curves,

$$\nabla_{\partial_i} \partial_j = 0 \qquad \text{for } i, j = 1, .., p$$

It follows that the co-ordinate curves of an affine co-ordinate system are geodesics, and co-ordinate vectors are parallel transported along co-ordinate curves. Substituting the vanishing connection coefficients into equation (2.4.1), it follows that the curvature tensor is zero wherever affine co-ordinates exist. Conversely, flatness implies the existence of locally affine co-ordinates.

Of course, the set of Cartesian co-ordinate charts is closed under affine transformations. This is extended to manifolds by the following proposition.

Proposition 2.4.2. Let Q be a smooth manifold, and let (U, φ) be an affine co-ordinate chart. Then the transformed co-ordinate chart (U, ψ) is an affine co-ordinate chart, where the map ψ is defined as

$$\psi : U \to \mathbb{R}^p \qquad \psi(q) = C\varphi(q) + d$$

where $C \in \mathbb{R}^{p \times p}$ is non-singular and $d \in \mathbb{R}^p$.

Proof. Calculating the connection coefficients via the covariant derivative along the coordinate directions,

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \nabla_{\frac{\partial q^m}{\partial x^i}\frac{\partial}{\partial q^m}} \frac{\partial q^l}{\partial x^j}\frac{\partial}{\partial q^l} = \frac{\partial q^m}{\partial x^i}\frac{\partial q^l}{\partial x^j}\nabla_{\frac{\partial}{\partial q^m}}\frac{\partial}{\partial q^l}$$
$$= C_i^m C_j^l \nabla_{\frac{\partial}{\partial q^m}}\frac{\partial}{\partial q^l} = 0$$

since q is an affine co-ordinate system, where $C = \{C_i^m\}$. □

2.4.2 Dual Flatness

A conjugate connections manifold which is flat with respect to both ∇ and ∇^* is called dually-flat. The following theorem shows that dual-flatness is implied by ∇- or ∇^*-flatness.

Theorem 2.4.3 (Dual Flatness Theorem). Let (Q, g, ∇, ∇^*) be a conjugate connection manifold. The manifold is ∇-flat if and only if it is ∇^*-flat.

Theorem 2.4.3 is a special case of The Fundamental Theorem of Information Geometry [57], which states that the scalar curvatures of ∇ and ∇^* are the same when one of them is constant; the stated theorem is for the case when the scalar curvatures are zero.

Consider the immediate properties of dually flat manifolds. There must exist two co-ordinate systems; one which is affine for ∇, and the other affine for ∇^*. We refer to these respectively as primal and dual co-ordinates. Note that the primal and dual co-ordinate systems may not necessarily be the same. By the definition of affine co-ordinates, the connection coefficients for the primal connection must vanish in the primal co-ordinate system, $\Gamma_{jk}^i = 0$, and likewise for the dual connection coefficients in the dual co-ordinate system, but primal connection coefficients may be non-vanishing in dual co-ordinates and vice-versa. It follows that geodesics with respect to ∇, called primal geodesics, are straight lines in the primal co-ordinate system, and dual geodesics which are geodesic with respect to ∇^* and are straight lines in the dual co-ordinate system.

Recall that dual co-ordinates were already defined in the context of Bregman divergences in Section 2.2.2; we shall show that Bregman divergences induce dually-flat geometries and the definitions coincide, i.e. Legendre dual co-ordinates $h = \nabla F(q)$ are ∇^*-affine co-ordinates in the geometry induced by D_F.

Theorem 2.4.4. Let Q be a smooth manifold with a global co-ordinate chart (Q, q). Let $F : q(Q) \subseteq \mathbb{R}^p \to \mathbb{R}$ be convex. Then the conjugate connections manifold induced by the Bregman divergence D_F is dually-flat.

Proof. Recall that the Bregman divergence D_F is given by

$$D_F(q, q_0) = F(q) - F(q_0) - \nabla F(q_0) \cdot (q - q_0)$$

A direct calculation for the induced primal connection coefficients in the co-ordinate system q from equation (2.3.7) yields

$$\Gamma_{ijk}^{(D_F)} = \frac{\partial^2}{\partial q^i \partial q^j} \frac{\partial}{\partial q_0^k} D_F(q, q_0) \Big|_{q=q_0} = 0 \qquad \text{for all } i, j, k \tag{2.4.2}$$

It follows that q is a r-affine co-ordinate system, and hence the manifold must be r-flat. By the Dual Flatness Theorem 2.4.3, it is also r-flat and hence is dually-flat. □

The proof of Theorem 2.4.4 shows that q is a primal co-ordinate system for the dually-flat manifold induced by D_F. The following proposition shows that the earlier definition of dual co-ordinates coincides with our new definition of r-affine co-ordinates.

Proposition 2.4.5. Let Q be a smooth manifold with global co-ordinate chart (Q, q), and let $F : q(Q) \to \mathbb{R}$ be convex. The dual Legendre co-ordinate chart (Q, h) defined via

$$h : Q \to \mathbb{R}^p, \qquad h(q) = \nabla F(q(q))$$

is a r^{D_F}-affine co-ordinate system, i.e. dual co-ordinates in the dually-flat manifold induced by D_F.

Proof. We need to show that the connections coefficients for r vanish in the h co-ordinate system, i.e. $\Gamma_{ijk}^{D_F}(h) = 0$ for all i, j, k. Definition 2.3.8 states that the dual connection r is equivalent to the primal connection induced by the dual divergence D_F. By Proposition 2.2.13, the dual Bregman divergence is another Bregman divergence generated by F,

$$D_F(q, q_0) = D_F(h_0, h)$$

Equation (2.4.2) must hold with h replacing q and D_F replacing D_F. □

Amari [3] shows that for any dually-flat manifold, there exists a convex function F on the primal co-ordinates such that the same geometry is induced by D_F, the so-called canonical divergence. Dually-flat manifolds are therefore uniquely characterized by their geometry being induced by Bregman divergences.

Cartesian co-ordinates are an affine system in which Euclidean geometry is simply described. Analogously, the primal and dual co-ordinates in a dually-flat manifold give the simplest description of the dually-flat geometry. Primal and dual geodesics have simple closed-form descriptions in their respective co-ordinate systems, and hence dually-flat geometry may provide a significant computational advantage over Riemannian geometry, in which Levi-Civita geodesics are usually not available in closed form and require expensive numerical integrators. This has motivated us to research MCMC algorithms that make use of dually-flat geometry, which is discussed in part II.

Example 2.4.6 (Regular exponential families are dually flat). By Proposition 2.2.12, the KL-divergence between two points in a regular exponential family coincides with the

Bregman divergence induced by the cumulant generating function ψ of the family in natural co-ordinates. By Theorem 2.4.4, the geometry induced by this divergence function is dually- at, where the natural parameters θ are primal co-ordinates and mean parameters $\eta = \nabla \psi(\theta)$ are dual co-ordinates.

By Proposition 2.4.2, any af ne transformation of the primal co-ordinates is another primal af ne co-ordinate system, and similarly for the dual co-ordinate system. Thus there are in nitely many primal and dual af ne co-ordinate systems. Primal and dual co-ordinate charts are coupled if they are the Legendre transform of each other, i.e. if $\eta = \nabla_\theta F(\theta)$ and $\theta = \nabla_\eta F^*(\eta)$ for the Bregman generator F corresponding to the canonical Bregman divergence on Ω. The following proposition shows how coupled co-ordinate charts change under af ne transformations.

Proposition 2.4.7. Let θ and η be dually coupled co-ordinate charts on a dually- at manifold Ω. Write $\tilde{\theta} = C\theta + d$ for an af ne-transformed co-ordinate system, where $C \in \mathbb{R}^{p \times p}$ is non-singular and $d \in \mathbb{R}^{p \times 1}$. Then $\tilde{\theta}$ is coupled to the dual co-ordinate system $\tilde{\eta}$ de ned by

$$\tilde{\eta} := C^{-T}\eta$$

where the transpose is the usual adjoint with respect to the Euclidean inner-product in $\mathbb{R}^p$.

Proof. Note that $\tilde{F}(\tilde{\theta}) := F(C^{-1}(\tilde{\theta} - d))$ is the canonical Bregman generator in transformed primal co-ordinates. For any θ let $\eta = \nabla_\theta F(\theta)$. Then by the chain rule,

$$\begin{aligned}
\nabla_{\tilde{\theta}} \tilde{F}(\tilde{\theta}) &= C^{-T} \nabla_\theta F(C^{-1}(\tilde{\theta} - d)) \\
&= C^{-T} \nabla_\theta F(C^{-1}(C\theta + d - d)) \\
&= C^{-T} \nabla_\theta F(\theta) = C^{-T}\eta = \tilde{\eta}
\end{aligned}$$

Therefore $\tilde{\theta}$ and $\tilde{\eta}$ are dually-coupled co-ordinate systems under the Bregman generator $\tilde{F}$. □

2.5 Geometry of Subfamilies

In Section 1.3, we brie y discussed statistical families which are subfamilies of larger families. We can now offer more precise de nitions for subfamilies and discuss how geometry can be inherited from the ambient larger family. This section contains two novel contributions which relate the inherited dually- at geometry to that of the ambient space; Theorem 2.5.3 and Theorem 2.5.8.

Suppose Ω is a p-dimensional smooth manifold, and N is an n-dimensional statistical family which we call the *ambient* family or manifold. For example, N could be a product of n 1-dimensional families modelling each of n i.i.d. observations in a statistical application, where co-ordinates are given by the mean parameters for each observation. Suppose further there is a smooth map $L : \Omega \to N$; continuing the example, L may represent a forward model relating p unknown parameters in Ω with the

means of the n observations. Such a structure is common in statistical modelling. We assume that $p < n$, i.e. there are fewer unknown parameters than observations.

Our goal is to view the image $L(Q) \subset N$ as a statistical family which is equivalent to Q, and to study how geometric structures on the ambient manifold are inherited by Q. This requires some further assumptions on L, which we adapt from Lee [46];

Definition 2.5.1. A smooth map $L : Q \to N$ between smooth manifolds Q, N is a *smooth immersion* if the differential dL_q has rank $p = \dim Q$ for all points $q \in Q$.

Definition 2.5.2. A smooth immersion $L : Q \to N$ between smooth manifolds Q, N is a *smooth embedding* that is a homeomorphism onto its image $L(Q) \subset N$, where the image is endowed with the subspace topology.

The image $L(Q)$ of a smooth embedding is an *embedded submanifold* of N. Lee [46] shows that embedded submanifolds are smooth manifolds; co-ordinate charts (U, b) for Q naturally give rise to smooth charts $(L(U), b \circ L^{-1})$ on the image. Note that since we assume $p < n$, any smooth embedding of Q into N must have full rank. We will adopt the convention of using $q \in \mathbb{R}^n$ as representing co-ordinates for N, and $b \in \mathbb{R}^p$ for co-ordinates in Q.

2.5.1 Inherited Geometry

Given the setup described above, suppose the ambient manifold N has geometric structures, i.e. connections and a Riemannian metric. These objects are essentially maps that take tangent vectors or vector fields as inputs; we can therefore use the differential dL of the embedding to map tangent vectors in T_qQ into $T_{L(q)}N$ where geometric objects are defined, and hence define an *inherited* geometry on Q.

In particular, suppose h is a $(0,k)$-tensor on $T_{L(q)}N$, i.e. a multilinear function of k tangent vectors. Define the *pullback* of h by L as the $(0,k)$-tensor $L^* h$ on T_qQ,

$$L^* h\left(V^1, \ldots, V^k\right) := h\left(dL_q\left(V^1\right), \ldots, dL_q\left(V^k\right)\right) \qquad \text{for all } V^1, \ldots, V^k \in T_qQ$$

The pullback of a tensor field is defined such that the tensor at each point $L(q) \in T_{L(q)}N$ is pulled back by L to T_qQ, for all $q \in Q$. Since the differential is linear, the pullback is clearly multilinear and is therefore a tensor.

The inherited Riemannian metric $\tilde{g}$ is the pullback of the Riemannian metric g on N by L,

$$\tilde{g}(V, W) := L^* g(V, W) \qquad \text{for all } V, W \in T_qQ$$

To see that $\tilde{g}$ is a Riemannian metric on Q, note that symmetry follows immediately from the symmetry of g. For positive-definiteness, let $V \in T_qQ$ and note

$$\tilde{g}(V, V) = g\left(dL_q(V), dL_q(V)\right) \geq 0$$

with equality if and only if $dL_q(V) = 0$, which holds if and only if $V = 0$ since dL_q has full rank. In co-ordinates, the inherited metric has components

$$g^{fl}_{ab} = g_{ij} \quad L(b) \quad \frac{\P L_i}{\P b_a} \frac{\P L_j}{\P b_b} \tag{2.5.1}$$

where $L = q \quad L \quad b^{\ 1}$ is the co-ordinate representation of L. Note that for any $q \ 2$ Q, the image $Im \quad dL_q \qquad T_{L(q)}N$ represents the space of vectors which are tangent to the embedded submanifold $L(Q)$. Any vector $V \ 2 \ T_{L(q)}N$ can be decomposed into tangential and normal components,

$$V = V^T + V^?$$

where $V^T \ 2 \ Im \quad dL_q$ and $V^?$ is in the g-orthogonal complement $Im \quad dL_q \ ^?$. Since dL_q is a linear isomorphism, we can make the identi cation $T_qQ \quad Im \quad dL_q$.

We can also derive a connection on Q inherited from an ambient connection r on N . Recall that the connection r on N is a map $X(N) \quad X(N) \ ! \quad X(N)$, so r applied to vector elds on Q mapped via the differential yields a vector eld on N ,

$$r_{\ dL(X)}dL(Y) \ 2 \ X(N) \qquad \text{for all } X, Y \ 2 \ X(Q) \tag{2.5.2}$$

Strictly speaking, $dL(X)$ and $dL(Y)$ do not de ne vector elds over all of N , and so we need to extend the vector elds; see Lee [45] for details.

Decomposing the covariant derivative into its tangential and normal components yields

$$r_{\ dL(X)}dL(Y) = \quad r_{\ dL(X)}dL(Y) \quad ^T + \quad r_{\ dL(X)}dL(Y) \quad ^? \tag{2.5.3}$$

The tangential term can be identi ed with a vector eld on Q, and hence de nes the inherited connection on Q,

$$r^{fl}_X Y := dL^{\ 1} \quad r_{\ dL(X)}dL(Y) \quad ^T$$

Lee [45] shows that this is a valid connection on Q, and is symmetric when r is symmetric. Amari [3] gives the following formula for the inherited connection coef cients:

$$G^{fl}_{abc} = B^j_c B^k_a \P_k \quad B^i_b \quad g_{ij} + B^i_a B^j_b B^k_c G_{jk} \tag{2.5.4}$$

where we use the shorthand $B^i_a = \frac{\P L_i}{\P b_a}$ for the differential in co-ordinates.

The normal component in (2.5.3) is a measure of how curved the embedded submanifold is within N , and is called the second fundamental form or embedding curvature,

$$P : X(Q) \quad X(Q) \ ! \quad X(N) \qquad P(X,Y) := \quad r_{\ dL(X)}dL(Y) \quad ^?$$

for any vector fields $X, Y \in \mathfrak{X}(Q)$. One can show [17] that the second fundamental form is symmetric in its arguments, although it is not a tensor.

Equation (2.5.3) is equivalent to the Gauss formula [45] $\nabla_{dL(X)} dL(Y) = dL\left(\nabla^{\mathcal{I}}_X Y\right) + P(X, Y)$.

If the ambient manifold is equipped with a divergence D, we can define an inherited divergence

$$Dj_Q : Q \times Q \to \mathbb{R}, \qquad Dj_Q(q_1, q_2) = D(L(q_1), L(q_2))$$

Clearly Dj_Q satisfies the positivity (i) and non-degeneracy (ii) conditions of Definition 2.2.1 in any co-ordinate chart, since D is a divergence and L is injective. The Hessian matrix on the diagonal is computed to be

$$\frac{\partial}{\partial b_0^a}\frac{\partial}{\partial b^b} Dj_Q(b_0, b)\Big|_{b=b_0} = \frac{\partial L_i}{\partial b_a}\frac{\partial L_j}{\partial b_b}\frac{\partial}{\partial q_i}\frac{\partial}{\partial q_j} D(q, q_0)\Big|_{q=q_0} \tag{2.5.5}$$

which is the (a, b)-component of a positive-definite matrix, because the co-ordinate differential $\frac{\partial L_i}{\partial b_a}$ has full rank. It follows that Dj_Q is indeed a divergence.

Suppose the geometry of N is a conjugate connections structure induced by a divergence D. It is natural to ask whether the geometry induced by Dj_Q is the same as the geometry inherited from N. The following theorem shows that this is indeed the case, and is a novel contribution to the literature.

Theorem 2.5.3. Let N be a smooth manifold with Riemannian metric and dual connections induced by a divergence $D : N \times N \to \mathbb{R}$. Let Q be an embedded submanifold of N. The Riemannian metric, primal and dual connections on Q inherited from N coincide with the metric and connections induced by the inherited divergence D_Q.

Proof. We will show that the induced and inherited metric and dual connections coincide by showing that they are equal in any co-ordinate chart. Firstly, the metric induced by Dj_Q is the Hessian on the diagonal; the right-hand sides of equations (2.5.5) and (2.5.1) agree when g_{ij} are the components of the metric induced by D.

To show the induced and inherited connections coincide, we explicitly compute the primal connection coefficients for Dj_Q. Let $q = L(b)$ and $b_0 = L(b_0)$. Write $B^i_a = \frac{\partial L_i}{\partial b_a}(b)$ and $B^k_{c_0} = \frac{\partial L_k}{\partial b_c}(b_0)$ for the co-ordinate differential at b and b_0 respectively. Then by equation (2.3.7), the primal connection coefficients induced by Dj_Q

$$\begin{aligned} G^{Dj_Q}_{abc} &= \frac{\partial}{\partial b^a}\frac{\partial}{\partial b^b}\frac{\partial}{\partial b_0^c} Dj_Q(b, b_0)\Big|_{b=b_0} \\ &= B^i_a \frac{\partial}{\partial q^i} B^j_b \frac{\partial}{\partial q^j} B^k_{c_0} \frac{\partial}{\partial q_0^k} D(q, q_0)\Big|_{q=q_0} \end{aligned}$$

Since $B^k_{c_0}$ is a function of b_0, it is a constant with respect to the partial derivatives $\frac{\partial}{\partial q^i}$ and $\frac{\partial}{\partial q^j}$, and hence

$$\begin{aligned} G^{D|_Q}_{abc} &= B^i_a B^k_{c_0} \frac{\partial}{\partial q^i} \left(B^j_b \frac{\partial}{\partial q^j} \frac{\partial}{\partial q^k_0} D(q, q_0) \right)_{q=q_0} \\ &= B^i_a B^k_{c_0} \frac{\partial B^j_b}{\partial q^i} \frac{\partial}{\partial q^j} \frac{\partial}{\partial q^k_0} D(q, q_0)_{q=q_0} \\ &\quad B^i_a B^k_{c_0} B^j_b \frac{\partial}{\partial q^i} \frac{\partial}{\partial q^j} \frac{\partial}{\partial q^k_0} D(q, q_0)_{q=q_0} \end{aligned} \tag{2.5.6}$$

Note that by Lemma 2.2.2, the metric induced by D on N may be written

$$g^{(D)}_{jk} = \frac{\partial}{\partial q^j} \frac{\partial}{\partial q^k_0} D(q, q_0)_{q=q_0}$$

Substituting this and the expression (2.3.7) for the induced primal connection , we have

$$G^{D|_Q}_{abc} = B^i_a B^k_c \frac{\partial B^j_b}{\partial q^i} g^{(D)}_{jk} + B^i_a B^j_b B^k_c G^{(D)}_{ijk}$$

This coincides with the coef cients (2.5.4) for the connection inherited by Q. The proof for the induced dual connection is identical. □

Since the primal and dual connections induced by $D|_Q$ are dual, inherited primal and dual connections are also dual with respect to the inherited metric by Theorem 2.5.3. Since inherited and induced geometries are the same, we will simply refer to the geometry on Q when Q is an embedded submanifold.

2.5.2 Autoparallel Submanifolds

De nition 2.5.4. Let ∇ be a connection on N and let Q be an embedded submanifold in N. Then Q is called ∇-autoparallel if the second fundamental form P corresponding to ∇ vanishes, $P(X, Y) = 0$, for all vector elds $X, Y \in \mathcal{X}(Q)$.

In the case when N is dually- at, ∇-autoparallel manifolds are called e- at and ∇^*-autoparallel manifolds are m- at . The terms come from the fact that e- at subfamilies of exponential families are themselves exponential families, and m- at subfamilies of mixtures are themselves mixture families; see Amari [3]. The following theorem from Calin and Udriste [17] shows that e- at and m- at submanifolds are both dually- at manifolds in their own right.

Theorem 2.5.5. Let N be a dually- at manifold and Q an embedded submanifold in N. If Q is either ∇-autoparallel or ∇^*-autoparallel, then Q is a dually- at manifold.

The following proposition helps build intuition for e- and m- at submanifolds; they are characterized by linear conditions in primal and dual co-ordinates respectively, so e- at submanifolds look like hyperplanes in primal co-ordinates and like curved hypersurfaces in dual co-ordinates and vice versa for m- at submanifolds.

Proposition 2.5.6. Let N be a dually- at n-manifold. An embedded p-submanifold Q is e- at if and only if there exist co-ordinates b for Q such that the co-ordinate embedding L has the af ne form in primal co-ordinates,

$$q = L(b) := Ab + d \tag{2.5.7}$$

for some matrix $A \in R^{n \times p}$ and vector $d \in R^n$.

Dually, an embedded p-submanifold Q is m- at if and only if there exist co-ordinates g for Q such that the co-ordinate embedding L has the af ne form in dual co-ordinates,

$$h = L(g) := Ag + d \tag{2.5.8}$$

for some matrix $A \in R^{n \times p}$ and vector $d \in R^n$.

A proof is given in appendix A. Clearly primal and dual geodesics, which are straight lines in their respective co-ordinate systems, are e- at and m- at submanifolds respectively. The following example considers when generalized linear models are e- or m- at.

Example 2.5.7. Consider generalized linear models discussed in Section 1.3. The mean parameters $h_i = E(y_i)$ for each of $i = 1, .., n$ data are related to coef cients $b_1, .., b_p$ by the relationship

$$h_i = f\left(\sum_{j=1}^{p} b_j X_{ij}\right) \qquad i = 1, .., n \tag{2.5.9}$$

where X_{ij} are constant covariates. We consider b as co-ordinates for a submanifold or subfamily embedded in an n-dimensional exponential family, In example 2.4.6, we showed that regular exponential families are dually at, and mean parameters $h = (h_i)$ are af ne dual co-ordinates. Clearly if f is the identity function, then the generalized linear model is m- at because (2.5.9) becomes $h_i = \sum_j X_{ij} b_j$, which is linear in dual co-ordinates. It follows that b are dual co-ordinates for the generalized linear model. Alternatively, suppose f is the component-wise primal to dual map, i.e. the canonical link function, so that (2.5.9) becomes linear in primal co-ordinates,

$$q_i = \sum_{j=1}^{p} b_j X_{ij} \qquad i = 1, .., n$$

In this case the generalized linear model is e- at, and b are primal co-ordinates.

When the ambient manifold N is dually- at, its geometry is derived from a canonical divergence which takes the form of a Bregman divergence D_F in primal co-ordinates,

generated by a convex function F. A submanifold Q, with primal co-ordinate embedding $q = L(b)$, has the inherited divergence in co-ordinate form

$$D_F|_Q(b_1, b_2) := F(L(b_1)) - F(L(b_2)) - \nabla F(L(b_2)) \cdot (L(b_1) - L(b_2))$$

There is a slight abuse of notation here; this is really D_F restricted to the co-ordinate representation of Q.

We can also consider the restriction $F|_Q$ of the Bregman generator F,

$$F|_Q(b) := F(L(b))$$

The restriction is not necessarily convex, and hence the Bregman divergence it generates is not always well-defined. The following theorem, which is a new contribution to the field, shows that the inherited divergence coincides with the Bregman divergence for the restricted generator exactly when Q is e-flat.

Theorem 2.5.8. Let Q be a submanifold of a dually-flat manifold N, and let F be the Bregman generator for the canonical divergence on N. Then Q is e-flat if and only if there exists global co-ordinates b in a chart (Q, ϕ) for which $F|_Q$ is convex and

$$D_F|_Q(b_1, b_2) = D_{F|_Q}(b_1, b_2)$$

for all points $b_1, b_2 \in \phi(Q)$, where $D_F|_Q$ is the restriction of D_F to Q and $F|_Q$ is the restriction of F to Q.

Dually, Q is m-flat if and only if there exist global co-ordinates g in a chart (Q, ϕ) for which $F^*|_Q$ is convex and

$$D_{F^*}|_Q(g_1, g_2) = D_{F^*|_Q}(g_1, g_2)$$

for all points $g_1, g_2 \in \phi(Q)$.

Proof. We will show the result characterizing e-flat submanifolds only; the proof for the m-flat case is identical.

Suppose Q is e-flat. By Proposition 2.5.6, there exists co-ordinates b such that the co-ordinate embedding for Q into N takes the affine form (2.5.8). Then writing the restriction of D_F to Q in b co-ordinates,

$$\begin{aligned} D_F(L(b_1), L(b_2)) &= F(L(b_1)) - F(L(b_2)) - \nabla F(L(b_2)) \cdot (L(b_1) - L(b_2)) \\ &= F|_Q(b_1) - F|_Q(b_2) - \nabla F(L(b_2)) \cdot (A(b_1 - b_2)) \\ &= F|_Q(b_1) - F|_Q(b_2) - A^T \nabla F(L(b_2)) \cdot (b_1 - b_2) \end{aligned}$$

Note that $A^T \nabla F$ is the gradient of the restriction of F,

$$A^T \nabla F(L(b_2))_i = A^k_i \frac{\partial F}{\partial q^k}(L(b_2)) = \frac{\partial L^k}{\partial b^i} \frac{\partial F}{\partial q^k}(L(b_2)) = \frac{\partial F|_Q}{\partial b^i}(b_2)$$

Hence the inherited divergence has Bregman generator $F|_Q$,

$$D_F|_Q(b_1, b_2) = F|_Q(b_1) - F|_Q(b_2) - \nabla_b F|_Q(b_2) \cdot (b_1 - b_2) = D_{F|_Q}(b_1, b_2)$$

Conversely, suppose $D_{F|_Q} \equiv D_F|_Q$ holds when Q is a submanifold; we aim to show Q is e- at. Expanding the de nitions, it follows that for all b_1, b_2,

$$\begin{aligned} r_F L(b_2) \cdot (L(b_1) - L(b_2)) &= r_b F|_Q(b_2) \cdot (b_1 - b_2) \\ \Leftrightarrow r_F L(b_2) \cdot (L(b_1) - L(b_2)) &= DL(b_2)^T r_F L(b_2) \cdot (b_1 - b_2) \\ &= r_F L(b_2) \cdot DL(b_2)(b_1 - b_2) \end{aligned}$$

where DL is the Jacobian matrix of L. Since this is true for all b_1, b_2, we must have

$$\begin{aligned} L(b_1) - L(b_2) &= DL(b_2)(b_1 - b_2) \\ \Rightarrow L(b_1) &= L(b_2) + DL(b_2)(b_1 - b_2) \end{aligned}$$

Differentiating this equation with respect to components of b_1 shows that the Jacobian DL is independent of b, and hence L is an af ne transformation. By Proposition 2.5.6, it follows that Q is e- at. □

2.5.3 The Pythagorean Theorem and Projections

A remarkable fact of dually- at manifolds is the following generalization of the Pythagorean Theorem from Amari [3]. Note that two curves $c_1(t), c_2(t)$ on Q are orthogonal at a point of intersection $q = c_1(t_0) = c_2(t_0)$ if their velocity vectors are orthogonal at q, i.e. if $g(c_1'(t_0), c_2'(t_0)) = 0$.

Theorem 2.5.9 (Amari's Generalized Pythagorean Theorem). Let Q be a dually- at manifold and let $q_1, q_2, q_3 \in Q$ be distinct points. If the dual geodesic connecting q_1 and q_2 is orthogonal at q_2 to the primal geodesic connecting q_2 and q_3, then

$$D(q_1, q_3) = D(q_1, q_2) + D(q_2, q_3)$$

where D is the canonical divergence inducing the dually- at structure on Q.

Note that the dual divergence D^* also satis es the Pythagorean Theorem, and hence under the conditions of Theorem 2.5.9,

$$D^*(q_3, q_1) = D^*(q_3, q_2) + D^*(q_2, q_1)$$

Clearly this generalizes the usual Pythagorean Theorem, since it includes the case when D is proportional to the squared Euclidean distance. Our interest in the Pythagorean Theorem is that it allows projections onto submanifolds, which we make use of in sampling algorithms in Chapter 5. Recall that in an inner-product space, the orthogonal projection of a point x onto an af ne subspace U is exactly the point in U which minimizes the distance to x. This result can also be generalized to dually- at spaces.

Let Q be a submanifold embedded in a dually- at ambient manifold N. For $s \in N$, a point $q \in Q$ is an e-projection of s onto Q if the primal geodesic connecting s and q is orthogonal to any tangent vector to Q at q. Similarly, q is the m-projection of s onto Q if the dual geodesic connecting the two points is orthogonal to any tangent vector to Q at q. The following theorem shows the e- and m-projections onto e- and m- at submanifolds respectively minimize the divergence among elements in Q.

Theorem 2.5.10. Let Q be a submanifold embedded in an ambient dually- at manifold N, and let $s \in N$. If Q is e- at, the m-projection q_m of s onto Q is exactly the point that minimizes the divergence from s to q in Q,

$$q_m = \operatorname{argmin}_{q \in Q} D(s, q)$$

Similarly, if Q is m- at the e-projection q_e of s onto Q is the point which minimizes the dual divergence,

$$q_e = \operatorname{argmin}_{q \in Q} D(q, s)$$

Proof. The following roughly reproduces the proof from Amari [3]. Let $q \in Q$ be arbitrary, and let q_m be the m-projection of s onto Q. By de nition of q_m, the velocity vector of the e-geodesic at q_m must be orthogonal to that of the m-geodesic connecting s to q_m. Hence the conditions of the Pythagorean Theorem 2.5.9 are satis ed, and

$$D(s, q) = D(s, q_m) + D(q_m, q)$$

which is clearly minimized by $q = q_m$. The proof for the e-projection onto an m- at submanifold is similar. □

Chapter 3

Independence and Orthogonal Submanifolds

In order to use Information Geometry to sample from a Bayesian posterior distribution, a divergence function should be chosen that includes information from both the prior and observed data. Traditionally, Information Geometry works with the expected geometry, i.e. induced by an f-divergence and hence includes neither prior nor observational information. This approach has been successful in geometric descriptions of classical, i.e. frequentist or non-Bayesian, statistics. There are also geometries which make use of observed data in the classical setting, such as Bradley and Efron's observed Fisher information [27] that replaces the expectation in the Fisher information (1.2.15) with evaluating y at the observed values $y = y_{obs}$. To my knowledge, no other authors have proposed a conjugate connections geometry that uses prior information.

In this chapter a Bregman divergence is suggested that includes all available information about the posterior, called the posterior Bregman divergence (PBD). The resulting geometry will be shown to have an important property for sampling algorithms, namely that orthogonal directions are close to independent sampling directions under the posterior. The parameter space may be decomposed into a fiber bundle, or foliation, of e- or m-flat submanifolds which are orthogonal to each other. The target distribution admits a disintegration, a generalization of conditional probability, over the orthogonal foliations. This idea will be central to the construction of new sampling algorithms in Chapter 5.

3.1 Posterior Bregman Geometry

In this section we introduce a dually-flat geometry induced by the posterior density, and present several results that justify its use. The geometry and two of the results are novel contributions to the literature; namely Proposition 3.1.2 and Theorem 3.1.3.

Let $p(q)$ be a smooth probability density supported on a convex set $Q \subseteq \mathbb{R}^p$. The density p is called log-concave if $F(q) := -\log p(q)$ is a strictly convex function. Any

log-concave probability density therefore de nes a Bregman divergence on Q, which we call the p-Bregman divergence,

$$D_F(q, q_0) \quad D_{\log p}(q, q_0) \qquad \text{for } q, q_0 \; 2 \; Q$$

The geometry induced by D_F, called the p-Bregman geometry, is immediate from the results in Chapter 2. In particular, D_F induces a dually- at structure on Q with primal co-ordinates q, dual co-ordinates de ned by $h := \; r \quad \log p(q)$ and Riemannian metric with components $G(q) = \; r \;^2 \log p(q)$ in primal co-ordinates.

Now suppose that a Bayesian posterior distribution $p(qjy) \; \mu \; p(yjq)p_0(q)$ is log-concave. In this context, we call $D_{\log p(\ jy)}$ the posterior Bregman divergence (PBD) and the induced geometry the posterior Bregman geometry. We similarly de ne the prior Bregman divergence/geometry when the prior density p_0 is log-concave. Calculating the p-Bregman divergence only requires p to be known up to a multiplicative factor. If two convex functions F_1, F_2 are related via $F_1(q) := F_2(q) + K$ for some constant K, then it follows immediately from the De nition 2.2.10 of a Bregman divergence that $D_{F_1} \quad D_{F_2}$. In particular, the Bregman divergence generated by $F(q) := \quad \log p(q) + K$ is the same as that generated by $\log p$.

Relationships between the posterior, prior, and likelihood-based divergences are implied by the following Lemma.

Lemma 3.1.1. Let $F_1, F_2 : Q \; ! \quad R$ be two smooth strictly convex functions on a convex set $Q \quad R$ with corresponding Bregman divergences D_{F_1}, D_{F_2}. De ne $F_1(q) := F_1(q) + c \quad q$. Then for all $q, q_0 \; 2 \; Q$, the following hold:

i) $D_{F_1+F_2}(q, q_0) = D_{F_1}(q, q_0) + D_{F_2}(q, q_0)$

ii) $D_{F_1}(q, q_0) = D_{F_1}(q, q_0)$.

Proof. Calculating directly,

i)

$$\begin{aligned} D_{F_1}(q, q_0) + D_{F_2}(q, q_0) &= F_1(q) \quad F_1(q_0) + F_2(q) \quad F_2(q_0) \\ &\quad r \quad F_1(q_0) \quad (q \quad q_0) \; r \quad F_2(q_0) \quad (q \quad q_0) \\ &= (F_1 + F_2)(q) \quad (F_1 + F_2)(q_0) \; r \quad (F_1 + F_2)(q_0) \quad (q \quad q_0) \\ &= D_{F_1+F_2}(q, q_0) \end{aligned}$$

ii)

$$\begin{aligned} D_{F_1}(q_0, q) &= F_1(q) \quad F_1(q_0) \quad (q \quad q_0) \; r \; F_1(q_0) \\ &= F_1(q) \quad F_1(q_0) + c \quad (q \quad q_0) \quad (r \; F_1(q) + c) \quad (q \quad q_0) \\ &= F_1(q) \quad F_1(q_0) \; r \quad F_1(q) \quad (q \quad q_0) = D_{F_1}(q, q_0) \end{aligned}$$

□

Since $p(q|y) \propto p_0(q)p(y|q)$, Lemma 3.1.1 implies that if the prior and likelihood are both convex functions of q then the posterior Bregman divergence decomposes into a sum of Bregman divergences respectively generated by the prior and likelihood. In particular, we have the following novel result relating the Kullback-Leibler, prior and posterior Bregman divergences for exponential family likelihoods.

Proposition 3.1.2. Let Q be a regular exponential family with log-concave prior $p_0(q)$ in natural parameters. Given an observation vector y, we have

$$D_{-\log p(\cdot|y)}(q, q_0) = D_{KL}(q, q_0) + D_{-\log p_0}(q, q_0) \tag{3.1.1}$$

where D_{KL} is the Kullback-Leibler divergence on Q.

Proof. By Proposition 2.2.12, the KL-divergence from q to q_0 is the Bregman divergence generated by the cumulant generating function ψ of the exponential family Q. Thus by the properties of the Bregman divergence from Lemma 3.1.1 we have

$$\begin{aligned} D_{KL}(q, q_0) + D_{-\log p_0}(q, q_0) &= D_{\psi}(q, q_0) + D_{-\log p_0}(q, q_0) \\ &= D_{\psi - \log p_0}(q, q_0) \\ &= D_F(q, q_0) \end{aligned}$$

where $F(q) := \langle q, y \rangle + \psi(q) - \log p_0(q)$. Since F and $-\log p(\cdot|y)$ differ only by a constant independent of q, the result follows. □

Note that the right-hand side of (3.1.1) does not depend on the observations y; this can also be viewed as a consequence of the invariance of Bregman divergences to terms of the form $\langle c, q \rangle$ in the Bregman generator from Lemma 3.1.1. This implies that for exponential family likelihoods, the posterior Bregman geometry does not depend on observed data. Although the divergence is invariant to adding terms of the form $\langle c, q \rangle$ in the Bregman generator, the operation induces a change of origin $h \to h' = h + c$ in dual co-ordinates. This is analogous to a change of origin of the Cartesian co-ordinates in Euclidean geometry, which leaves the Euclidean distance unchanged. In conclusion, two different observation vectors y and y' may shift the origin of the dual co-ordinate system under the posterior Bregman geometry for exponential family likelihoods, but intrinsic geometric objects are independent of observations.

The decomposition (3.1.1) together with Theorem 2.2.9 implies that the metric tensor for posterior Bregman geometry has components

$$G(q) = \nabla^2 \log p(q|y) = I(q) - \nabla^2 \log p_0(q)$$

in the primal co-ordinate system, where $I(q)$ is the Fisher information matrix. This coincides exactly with the Riemannian metric proposed by Girolami and Calderhead [39] for use in their Riemannian Hamiltonian Monte Carlo algorithm.

The following theorem is a new contribution. Although a fairly trivial observation, the result helps provide intuition for understanding the posterior Bregman geometry.

Theorem 3.1.3. Let $F : E \to \mathbb{R}$ be a convex function on $E = \mathbb{R}^p$, let p be a probability density supported on $Q \subseteq E$ such that $p(q) \propto \exp(-F(q))$ for all $q \in Q$, and let $q_0 \in E$ be such that $\nabla F(q_0) = 0$. Then

$$p(q) \propto \exp(-D_F(q, q_0)) \text{ for all } q \in Q$$

where D_F is the Bregman divergence generated by F, or equivalently p-th Bregman divergence.

Proof. Note that for any point $q \in Q$, $F(q) = -\log p(q) + K$ for some log-normalizing constant K. The Bregman divergence generated by F from q to q_0 is

$$\begin{aligned} D_F(q, q_0) &= F(q) - F(q_0) - \nabla F(q_0)(q - q_0) \\ &= F(q) - F(q_0) \\ &= -\log p(q) + K - F(q_0) \end{aligned}$$

and thus

$$p(q) = \exp(K - F(q_0)) \exp(-D_F(q, q_0)) \propto \exp(-D_F(q, q_0))$$

□

Suppose for simplicity that $Q = E = \mathbb{R}^p$. If the posterior has a unique mode $q_0 \in Q$, i.e. if the MAP estimator exists, then it follows that $\nabla \log p(q_0|y) = 0$, and hence Theorem 3.1.3 states that the negative log posterior density at q is the divergence from the mode q_0 to q up to additive constants. For example, a multivariate Gaussian random vector $q \sim \text{Normal}(q_0, L^{-1})$ has density

$$p(q) \propto \exp\left(-\frac{1}{2}(q - q_0)^T L (q - q_0)\right)$$

The p-Bregman divergence is then clearly half the squared Mahalonobis distance from example 2.2.3. The induced geometry has a constant Riemannian metric L in primal co-ordinates, and hence the primal and dual connections coincide and the geometry is Euclidean.

More generally, Theorem 3.1.3 shows that log-concave probability densities with a mode q_0 have level sets which are Bregman spheres centred at q_0, i.e. the locus of points with fixed Bregman divergence to q_0.

3.2 Orthogonality and Independence

The novel algorithms we introduce in the thesis will make use of the following heuristic: orthogonal directions in the posterior Bregman geometry are nearly independent sampling directions. In this section, we offer some justification for this heuristic.

Firstly, consider the case of a multivariate Gaussian distribution $\text{Normal}(0, L^{-1})$. As discussed in the previous section, the resulting posterior Bregman geometry is Euclidean with a fixed metric $G = L$ in primal co-ordinates. Since L is symmetric positive-definite, it defines a global weighted inner-product on $\mathbb{R}^p$,

$$(u, v)_L := u^T L v$$

The following proposition shows that a sample from the Gaussian has components with respect to a L-orthogonal basis which are stochastically independent. The formal statement of the result is novel, but is based on an observation made by Fox [30].

Proposition 3.2.1. Let $v^{(1)}, .., v^{(p)} \in \mathbb{R}^p$ be a basis of mutually L-orthogonal vectors, i.e.

$$v^{(i)T} L v^{(j)} = 0$$

for all $i \neq j$. Then for a random variable $q \sim \text{Normal}(0, L^{-1})$, distinct components of q in the $\{v^{(j)}\}$ basis are mutually independent.

Proof. The i^{th} component of q in the basis is given by

$$\phi_i := \frac{v^{(i)T} L q}{v^{(i)T} L v^{(i)}}$$

Let $d_i := v^{(i)T} L v^{(i)}$, (q_m) and $(v_s^{(i)})$ be the components of q $v^{(s)}$ respectively in the standard basis. Computing the covariance of ϕ_i and ϕ_j using Einstein summation notation, we find

$$\begin{aligned}
\text{Cov}(\phi_i, \phi_j) &= \frac{1}{d_i d_j} \text{Cov}\left(v_s^{(i)} L_{sm} q_m, v_l^{(j)} L_{lt} q_t\right) \\
&= \frac{1}{d_i d_j} v_s^{(i)} L_{sm} v_l^{(j)} L_{lt} \text{Cov}(q_m, q_t) \\
&= \frac{1}{d_i d_j} v_s^{(i)} L_{sm} v_l^{(j)} L_{lt} L^{-1}_{mt} \\
&= \frac{1}{d_i d_j} v_s^{(i)} L_{sm} v_l^{(j)} \delta_m^l \\
&= \frac{1}{d_i d_j} v_s^{(i)} L_{sm} v_m^{(j)}
\end{aligned}$$

Since $v^{(i)}$ and $v^{(j)}$ are L-orthogonal, it follows that $\text{Cov}(\phi_i, \phi_j) = 0$ whenever $i \neq j$. □

This fact can be exploited for sampling by building up a L-orthogonal or conjugate basis and independently sampling the components. This idea is the basis of conjugate directions sampling suggested by Fox and Parker [29], [59]. The algorithms developed in Chapter 5 will turn out to be conjugate directions samplers when applied to Gaussian target distributions.

Recall that co-ordinate curves of primal and dual co-ordinates intersect orthogonally, because for $i \neq j$

$$\begin{aligned}
g\left(\frac{\partial}{\partial q^i}, \frac{\partial}{\partial h_j}\right) &= g\left(\frac{\partial}{\partial q^i}, \frac{\partial q^m}{\partial h_j} \frac{\partial}{\partial q^m}\right) \\
&= \frac{\partial q^m}{\partial h_j} g\left(\frac{\partial}{\partial q^i}, \frac{\partial}{\partial q^m}\right) \\
&= \frac{\partial h_i}{\partial q^m} \frac{\partial q^m}{\partial h_j} = 0
\end{aligned}$$

One might therefore wonder if in the posterior Bregman geometry, primal and dual variables are independent. This is con rmed to be true in the Gaussian case, for which dual variables are given by a simple linear relation $h := L\,q$.

Proposition 3.2.2. Let $q \sim \text{Normal}(0, L^{-1})$, and de ne $h = L\,q$. Let $(q_i)_{i=1}^p$ and $(h_i)_{i=1}^p$ be the respective components of q and h in the standard basis. Then for $i \neq j$, the random variables q_i and h_j are independent.

Proof. Using Einstein summation notation,

$$\begin{aligned} \text{Cov}(q_i, h_j) &= \text{Cov}(q_i, L_{js} q_s) \\ &= L_{js}\,\text{Cov}(q_i, q_s) \\ &= L_{js} L^{-1}_{is} = 0 \end{aligned}$$

□

For non-Gaussian distributions, there are fewer concrete results linking orthogonality and independence. Girolami [52] notes that

$$\frac{\partial^2}{\partial q_i \partial q_j} \log p(q) = 0 \text{ for all } q \quad \Rightarrow \quad p(q) = F(q_i) H(q_j)$$

for some functions F, H. In our terms, if two variables with joint density p have orthogonal co-ordinate curves everywhere in parameter space under the p-Bregman divergence, then the variables are independent under p. However, since the Hessian of a log probability density is not a tensor, the result does not extend to arbitrary transformations of variables. Future work could extend the orthogonality results for Gaussians, perhaps by studying copulas and local correlations.

3.3 Mixed Co-ordinates and Foliations

To exploit orthogonal directions in a sampler, ideally we would use a global co-ordinate system which has mutually orthogonal co-ordinate vectors, i.e. with a diagonal metric tensor. However, DeTurck and Yang [23] show that such co-ordinates are only guaranteed to exist for manifolds of dimension 3, and even when they do exist it is not clear how to construct them in an ef cient manner. Instead, we will utilise the orthogonal relationship between primal and dual variables by making use of foliations

Euclidean spaces of dimension p are trivially partitioned into k-dimensional subspaces, or hyperplanes, which are all orthogonal to another subspace of dimension $p - k$ where $1 \leq k < p$. For example, in $\mathbb{R}^p$ we can consider an af ne subspace in which the rst k co-ordinates are allowed to vary, and the nal $p - k$ components are xed. Any such space is orthogonal to the af ne space with rst k co-ordinates xed and $p - k$ complementary components varying. In this section we will extend this to dually- at manifolds, where orthogonality is of course de ned in terms of the Riemannian metric.

The key generalization is the ber bundle [46], [12], which for our purposes is a smooth surjective map $F : Q \to M$, where Q and M are p- and $(p-k)$-dimensional smooth manifolds respectively and $1 \leq k < p$. In this context Q is called the total space, M is the base space and each k-dimensional smooth manifold $F^{-1}(z)$ is called a ber corresponding to $z \in \mathrm{Im} F$. The map F is also called the projection for the ber bundle. Note that there is an additional technical condition for F to be a ber bundle which we assume is always satis ed; see Chapter 10 of Lee [46] for details.

3.3.1 k-cut Mixed Co-ordinates

Suppose Q is a p-dimensional dually- at manifold. The primal $q = (q^1, .., q^p)$ and dually-coupled $h = (h^1, .., h^p)$ af ne co-ordinates can be mixed into a new co-ordinate system. For $1 \leq k \leq p$, construct the vector whose rst k components are dual and remaining $p-k$ components are primal,

$$x = (h^1, .., h^k, q^{k+1}, .., q^p) \tag{3.3.1}$$

which we refer to as the *k-cut mixed co-ordinate system*, or more broadly *mixed co-ordinates*. The k-cut system was introduced by Amari [2] to study hierarchies of probability distributions.

Suppose $Q \subseteq \mathbb{R}^n$ is the primal parameter space, and $F : Q \to \mathbb{R}$ is the canonical Bregman generator so that the i^{th} dual variable is $h^i = \frac{\P F}{\P q^i}(q)$. The transition map from primal to k-cut co-ordinates is the map $w : Q \to \mathbb{R}^p$, with i^{th} component

$$w^i(q) = \begin{cases} \frac{\P F}{\P q^i}(q) & \text{for } i \leq k \\ q^i & \text{for } i > k \end{cases} \tag{3.3.2}$$

We have omitted k in the notation w for clarity. The Jacobian matrix of (3.3.2), i.e. the co-ordinate differential, has components

$$\frac{\P w^i}{\P q^j}(q) = \begin{cases} \frac{\P^2 F}{\P q^i \P q^j}(q) & \text{for } i \leq k \quad j = 1, .., p \\ \delta^i_j & \text{for } i > k \quad j = 1, .., p \end{cases} \tag{3.3.3}$$

The Jacobian matrix therefore has rst k rows equal to the q representation of the metric, and last $p-k$ rows equal to the identity;

$$\frac{\P x^i}{\P q^j} = \frac{\P x}{\P q} = \begin{bmatrix} G_{[\,k,\,k]} & G_{[\,k,>k]} \\ 0 & 1_{(p-k)\,(p-k)} \end{bmatrix}$$

where $G_{[\,k,\,k]}$ and $G_{[\,k,>k]}$ respectively refer to the upper-left $k \times k$ and the upper-right $k \times (p-k)$ blocks of the metric Hessian $r^2 F$ in primal co-ordinates, and $1_{(p-k)\,(p-k)}$ is the $(p-k) \times (p-k)$ identity matrix. The Jacobian determinant is therefore the upper-left $k \times k$ block of the metric,

$$\begin{aligned} \det \frac{\P x}{\P q} &= \det \begin{bmatrix} G_{[\,k,\,k]} & G_{[\,k,>k]} \\ 0 & 1_{(p-k)\,(p-k)} \end{bmatrix} \\ &= \det G_{[\,k,\,k]} \end{aligned} \tag{3.3.4}$$

Since the $p \times p$ matrix $G(q)$ is positive-definite, it follows from Sylvester's criterion [37] that the $k \times k$ minor $\det G_{[k,k]} > 0$, and hence the Jacobian determinant is non-zero. Since this is true for all $q \in Q$, w is a diffeomorphism. We have shown that the transition map between primal and k-cut co-ordinates is a diffeomorphism, and hence the k-cut system is in the smooth structure of Q.

Although $x^i = q^i$ for $i > k$, the associated co-ordinate vectors $\frac{\partial}{\partial x^i}$ and $\frac{\partial}{\partial q^i}$ because each partial derivative involves holding different variables constant. Writing sums explicitly, we find that for $i > k$

$$\frac{\partial}{\partial q^i} = \sum_{j=1}^{p} \frac{\partial x^j}{\partial q^i} \frac{\partial}{\partial x^j} = \sum_{j=1}^{k} g_{ij} \frac{\partial}{\partial x^j} + \frac{\partial}{\partial x^i}$$

Recall from Theorem 1.2.18 that the inverse of the metric has components $g^{ij} = \frac{\partial^2 F}{\partial h^i \partial h^j}$. The inverse function theorem states that the Jacobian of the inverse transition map w^{-1} is the inverse of the Jacobian (3.3.3), and hence

$$\frac{\partial q^i}{\partial x^j} = \begin{cases} g^{ij} & \text{for } j \leq k, \quad i = 1, .., p \\ \delta^i_j & \text{for } j > k, \quad i = 1, .., p \end{cases} \tag{3.3.5}$$

where g^{ij} is evaluated at $h = r F(q)$ such that $w(q) = x$. Indeed, a simple calculation shows that $\frac{\partial x^i}{\partial q^j} \frac{\partial q^j}{\partial x^l} = \delta^i_l$, resuming the use of Einstein notation.

The key property of the k-cut system is that the i, j co-ordinate vectors are orthogonal whenever $i \leq k$ and $j > k$, or vice versa. This is stated in the following theorem, which gives the metric in mixed co-ordinates.

Theorem 3.3.1. The metric g_{ij} in k-cut co-ordinates takes the form

$$g_{ij} = \begin{cases} g^{ij}(h) & \text{for } i \leq k \quad j \leq k \\ g_{ij}(q) & \text{for } i > k \quad j > k \\ 0 & \text{for } i \leq k \quad j > k \\ 0 & \text{for } i > k \quad j \leq k \end{cases} \tag{3.3.6}$$

where q is such that $w(q) = x$, and $h = r F(q)$.

Proof. The components of the metric in the k-cut system are

$$g_{ij} = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \frac{\partial q^l}{\partial x^i} \frac{\partial q^m}{\partial x^j} g\left(\frac{\partial}{\partial q^l}, \frac{\partial}{\partial q^m}\right) = \frac{\partial q^l}{\partial x^i} \frac{\partial q^m}{\partial x^j} g_{lm}$$

where g_{lm} are the components of the metric in the primal q co-ordinate system. We consider each of the possible cases; firstly, when $i, j \leq k$ then the Jacobian equals the inverse metric g^{li} by (3.3.5),

$$g_{ij} = g^{li} g^{mj} g_{lm} = \delta^i_m g^{mj} = g^{ij}$$

Likewise, when $i, j > k$ the Jacobian is the identity δ^i_j,

$$g_{ij} = \delta^l_i \delta^m_j g_{lm} = g_{ij}$$

Finally when $i > k$ and $j \leq k$, or vice versa, the components of the metric are 0:

$$g_{ij} = \delta^l_i g^{jm} g_{lm} = \delta^l_i \delta^j_l = \delta^j_i = 0$$

since $i \neq j$ in this case. The proof for $i \leq k$ and $j > k$ is similar. □

Note that to actually evaluate the metric using (3.3.6), we must first convert the mixed co-ordinates to either fully-primal or fully-dual co-ordinates, and then compute the relevant blocks of the metric and its inverse. Written as a matrix, the metric in mixed co-ordinates takes the form

$$\begin{pmatrix} G_{[k \times k]}(w^{-1}(x)) & 0_{[k \times (p-k)]} \\ 0_{[(p-k) \times k]} & G_{[(p-k) \times (p-k)]}(w^{-1}(x)) \end{pmatrix}$$

3.3.2 Orthogonal Foliation

Mixed co-ordinates allow us to partition a dually-flat manifold in two ways; one with e-flat submanifolds, and the other with m-flat submanifolds. For $c \in R^k$, define $M_k(c)$ as the set of points in Q whose first k dual components are fixed,

$$h^i = c^i \qquad \text{for } i = 1, .., k \tag{3.3.7}$$

and remaining co-ordinates are allowed to vary freely. In dual co-ordinates, this is represented by the embedding

$$\begin{pmatrix} h^1 \\ \vdots \\ h^p \end{pmatrix} = \begin{pmatrix} 0_{[k \times (p-k)]} \\ 1_{[(p-k) \times (p-k)]} \end{pmatrix} \begin{pmatrix} h^{k+1} \\ \vdots \\ h^p \end{pmatrix} + \begin{pmatrix} c^1 \\ \vdots \\ c^k \\ 0_{[(p-k) \times 1]} \end{pmatrix} \tag{3.3.8}$$

By proposition 2.5.6, $M_k(c)$ is therefore a $(p-k)$-dimensional m-flat submanifold. Similarly for $d \in R^{p-k}$, define $E_k(d)$ as the set of points with fixed complementary primal co-ordinates,

$$q^i = d^{i-k} \qquad \text{for } i = k+1, .., p \tag{3.3.9}$$

and other co-ordinates varying freely. A similar argument to the above shows $E_k(d)$ is a k-dimensional e-flat submanifold.

Suppose a point $q \in Q$ has k-cut co-ordinates (c, d), i.e. first k dual co-ordinates given by (3.3.7) and remaining $p-k$ primal co-ordinates by (3.3.9). Then q belongs uniquely to the intersection $M_k(c) \cap E_k(d)$. For $c \neq c'$, the intersection $M_k(c) \cap M_k(c')$

is empty, and likewise $d \neq d'$ implies $E_k(d) \cap E(d')$ is empty. Clearly then, the E_k and M_k submanifolds both partition the ambient manifold Q,

$$Q = \bigcup_c M_k(c) = \bigcup_d E_k(d)$$

where unions are taken over the range of the $(h^1, .., h^k)$ and $(q^{k+1}, .., q^p)$ co-ordinate functions respectively. The partition into E_k submanifolds is known as the *e*-foliation, and the partition into M_k submanifolds is the *m*-foliation [3]. The submanifolds themselves are called the leaves of the foliation. The *e*-foliation is a fiber bundle with a projection map that picks out the $k+1, .., p$ components of the primal co-ordinate vector,

$$F_E : Q \to \mathbb{R}^{p-k} \qquad F_E(q) = \left(q^i\right)_{i=k+1}^{p}$$

such that each leaf $E_k(d)$ is a k-dimensional fiber for $d = F_E(q)$, and the base space is a convex subset $\mathrm{Im} F_E \subseteq \mathbb{R}^{p-k}$. Similarly, the *m*-foliation is also a fiber bundle with projection

$$F_M : Q \to \mathbb{R}^k \qquad F_M(q) = \left(\frac{\partial F}{\partial q^i}\right)_{i=1}^{k}$$

with base space $\mathrm{Im} F_M \subseteq \mathbb{R}^k$, and fibers $M_k(c)$ for $c = F_M(q)$. More generally, a foliation in differential geometry is a partition of the manifold into submanifolds such that there exists a co-ordinate chart in which each submanifold is described by holding constant particular co-ordinates [46].

The *e*-foliation and *m*-foliation are orthogonal in the following sense. Let q have mixed co-ordinates $x = (c, d)$ such that $q \in M_k(c) \cap E_k(d)$ as above. Then tangent vectors V to the embedded submanifold $E_k(d)$ must consist of only the first k mixed co-ordinate basis vectors; writing sums explicitly,

$$V = \sum_{i=1}^{k} V^i \frac{\partial}{\partial x^i}$$

Tangent vectors W to the embedded submanifold $M_k(d)$ only have non-zero components;

$$W = \sum_{j=k+1}^{p} W^j \frac{\partial}{\partial x^j}$$

Then by Theorem 3.3.1, tangent vectors to $E_k(c)$ and $M_k(d)$ are mutually orthogonal,

$$g(V, W) = \sum_{i=1}^{k} \sum_{j=k+1}^{p} V^i W^j g_{ij} = 0$$

3.3.3 General Foliations of e- and m- at Submanifolds

Since vector spaces can be partitioned with arbitrary colinear hyperplanes, foliations of dually- at manifolds can be constructed from arbitrary e- or m- at submanifolds. This generalizes the foliations arising from mixed co-ordinates. This more general concept is a novel contribution, and in particular Proposition 3.3.9, Theorem 3.3.11 and Theorem 3.3.12 are new results in Information Geometry.

For a matrix $A \in R^{p \times k}$ of full rank, let $B \in R^{p \times (p-k)}$ be a matrix whose columns form a basis for $\mathrm{Ker}\, A^T = (\mathrm{Im}\, A)^{\perp}$, where A^T is the usual transpose or adjoint of A with respect to the Euclidean inner-product; a canonical choice for B is any $(p-k)$ linearly independent columns of the projection matrix $1_{p \times p} - A\,(A^T A)^{-1} A^T$. Recall that the primal co-ordinate chart covers all of Q, and takes values in a convex subset $Q \subseteq R^p$. The primal co-ordinates of any $q \in Q$ can be decomposed as

$$q = Aa + Bb \tag{3.3.10}$$

where a and b belong to appropriate convex subsets $A \subseteq R^k$ and $B \subseteq R^{p-k}$ respectively. We fix $b := Bb$, so that a is a parameter for the k-dimensional e- at submanifold

$$E_{A,b} = \left\{ q = Aa + b : a \in A \subseteq R^k \right\} \tag{3.3.11}$$

in primal co-ordinates, a convex subset of the hyperplane through b spanned by the columns of A. Note that the convex subset A depends on b in general, although we suppress this in our notation. We adopt the notation $\{E_{A,b}\}$ for the collection $\{E_{A,b} \mid b \in \mathrm{Im}\, B\}$. Clearly elements of $\{E_{A,b}\}$ are disjoint and we have the partition

$$Q = \bigcup_{b \in \mathrm{Ker}(A^T)} E_{A,b}$$

Note that by proposition 2.5.6, any e- at submanifold can be written in the form (3.3.11) and is an element of such a partition. Similarly, dual co-ordinates h for any point $q \in Q$ can be decomposed as

$$h = Dd + Cg \tag{3.3.12}$$

where $D \in R^{p \times k}$ and $C \in R^{p \times (p-k)}$ are respectively defined analogously to A and B above. Let $M_{C,d}$ be the $(p-k)$-dimensional m- at submanifold defined by fixing $d := Dd$ and allowing g to vary, i.e.

$$M_{C,d} = \left\{ h = Cg + d : g \in R^{p-k} \right\} \tag{3.3.13}$$

Then $\{M_{C,d}\}$ is another partition of Q. Any m- at submanifold is an element of such a partition.

We use the general e- and m- at partitions to construct general mixed co-ordinates. Rewriting equation (3.3.10) in a more compact form,

$$q = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{3.3.14}$$

Suppose we let $D = A$ and $C = B$ in (3.3.12), so that h is decomposed into components $Ad \in \mathrm{Im}\, A$ and $Bg \in (\mathrm{Im}\, A)^{?}$. By Proposition 2.4.7, the dual co-ordinates corresponding to $q = (a, b)$ are given by

$$h = \begin{bmatrix} A^T \\ B^T \end{bmatrix} h = \begin{bmatrix} A^T Ad \\ B^T Bg \end{bmatrix}$$

For these dually coupled co-ordinate systems, the k-cut mixed co-ordinates are given by

$$x = \begin{bmatrix} A^T Ad \\ b \end{bmatrix} \tag{3.3.15}$$

In this mixed co-ordinate system, the submanifolds $E_{A,b}$ and $M_{B,d}$, with $b := Bb$ and $d := Ad$, are respectively leaves of the e- and m-foliation each corresponding to holding b and d constant, analogous to the submanifolds E_k and M_k in the previous subsection. By the results in the previous subsection, it follows that $E_{A,b}$ and $M_{B,d}$ intersect orthogonally with respect to the Riemannian metric at a unique point. The generalized e- and m-foliations are fiber bundles with projections $F_E(q) = B\, q$ and $F_M(q) = A^T r\, F(q)$ respectively, where $B := B^T B)^{-1} B^T$ is the generalized inverse of B. Of course, if A and B respectively have columns consisting of the first k and last $(p - k)$ standard basis vectors, then the matrix $A \quad B$ in (3.3.14) is the $p \times p$ identity matrix; $A^T Ad = h_1, .., h_k$ are the first k dual co-ordinates and the complementary $p - k$ primal co-ordinates are $b = q_{k+1}, .., q_p$. Hence generalized mixed co-ordinates (3.3.15) really are a generalization of regular mixed co-ordinates (3.3.1). We use a tuple (A, B, F) to denote a generalized mixed co-ordinate system, where A, B are matrices as defined above and F is a Bregman generator.

The following definition is inspired by the concept of an orthogonal complement in an inner-product space.

Definition 3.3.2. Two submanifolds E and M are called dual complements of each other if E is e- flat and M is m- flat, and the submanifolds are orthogonal to each other at any points of intersection.

Clearly $E_{A,b}$ and $M_{B,d}$ are dual complements whenever $A^T B = 0$.

Since $q = Aa + Bb$ and $r\, F(q) = Ad + Bg$, the conversion map from primal to generalized mixed co-ordinates is given by the respective projections for the fiber bundles of the m- and e-projections.

$$w(q) = \begin{bmatrix} A^T r\, F(q) \\ B\, q \end{bmatrix}$$

where $B := B^T B)^{-1} B^T$ is the generalized inverse of B.

G-Orthogonal Projections

De nition 3.3.3. Let $L \in R^{p \times p}$ be a symmetric positive-de nite matrix. A matrix $P \in R^{p \times p}$ is a L-orthogonal projection if

$$PP = P \tag{3.3.16}$$

$$P^T L = L P \tag{3.3.17}$$

Proposition 3.3.4. Let $P \in R^p$ be a L-orthogonal projection matrix. Then P is self-adjoint in the inner-product weighted by L, i.e.

$$(Pv)^T L w = v^T L (Pw)$$

for all $v, w \in R^p$.

Proof. Since P satis es the second condition (3.3.17) for L-orthogonal projections,

$$(Pv)^T L w = v^T P^T L w = v^T L Pw = v^T L (Pw)$$

□

De nition 3.3.5. Let $L \in R^{p \times p}$ be a symmetric positive-de nite matrix and $A \in R^{p \times k}$ have full rank, where $1 \leq k \leq p$. De ne the L-orthogonal projection onto the columns of A as

$$P_A^L = A \left(A^T L A\right)^{-1} A^T L$$

For example, the matrix $AA = A\left(AA^T\right)^{-1}A^T$ is the 1_p $_p$-orthogonal projection onto the columns of A, i.e. orthogonal with respect to the standard Euclidean inner-product on R^p. The de nition is justi ed by the following proposition.

Proposition 3.3.6. The matrix P_A^L is a L-orthogonal projection satisfying

$$\mathrm{Im} P_A^L = \mathrm{Im} A$$

In particular, the restriction of P_A^L to $\mathrm{Im} A$ is the identity map.

Proof. The rst condition (3.3.16) follows immediately;

$$\begin{aligned} P_A^L P_A^L &= A \left(A^T L A\right)\left(A^T L A\right)\left(A^T L A\right) A^T L \\ &= A \left(A^T L A\right)^{-1} A^T L = P_A^L \end{aligned}$$

For the second condition (3.3.17), we calculate:

$$\begin{aligned} \left(P_A^L\right)^T L &= \left(A \left(A^T L A\right)^{-1} A^T L\right)^T L \\ &= L A \left(A^T L A\right)^{-1} A^T L = L P_A^L \end{aligned}$$

To show $\mathrm{Im}\, P_A^L = \mathrm{Im}\, A$, let $v \in \mathrm{Im}\, P_A^L$. Then there exists $w \in R^p$ such that

$$v = P_A^L w = A \left(A^T L A\right)^{-1} A^T L w \in \mathrm{Im}\, A$$

and it follows that $\mathrm{Im}\, P_A^L \subseteq \mathrm{Im}\, A$. Conversely, let $w \in \mathrm{Im}\, A$; then $w = Aa$ for some $a \in R^k$, and

$$P_A^L w = A \left(A^T L A\right)^{-1} A^T L Aa = Aa = w \in \mathrm{Im}\, P_A^L$$

which shows $\mathrm{Im}\, A \subseteq \mathrm{Im}\, P_A^L$, and thus $\mathrm{Im}\, A = \mathrm{Im}\, P_A^L$. This calculation shows that $P_A^L w = w$ for any $w \in \mathrm{Im}\, A$, and hence the restriction of P_A^L to $\mathrm{Im}\, A$ is the identity map.

□

The matrix $P_A^{G(q)}$ is the $G(q)$-orthogonal projection onto the columns of A, a linear operator on the tangent space at the point represented by q.

Proposition 3.3.7. The matrix $1_{p\times p} - P_A^L$ is a L-orthogonal projection matrix satisfying

$$\mathrm{Ker}\left(1_{p\times p} - P_A^L\right) = \mathrm{Im} A \tag{3.3.18}$$

Proof. Condition (3.3.16) follows from the fact that P_A^L is a L-orthogonal projection matrix,

$$\left(1_{p\times p} - P_A^L\right)\left(1_{p\times p} - P_A^L\right) = 1_{p\times p} - 2P_A^L + P_A^L P_A^L = 1_{p\times p} - P_A^L$$

and similarly for (3.3.17),

$$\left(1_{p\times p} - P_A^L\right)^T L = L - P_A^{L\,T} L = L - L P_A^L = L\left(1_{p\times p} - P_A^L\right)$$

Finally, to show (3.3.18), we first let $w \in \mathrm{Im}\, A$. Then by Proposition 3.3.6, $P_A^L w = w$ and hence

$$\left(1_{p\times p} - P_A^L\right) w = w - w = 0$$

and so it follows that $w \in \mathrm{Ker}\left(1_{p\times p} - P_A^L\right)$.

Conversely, for any $w \in \mathrm{Ker}\left(1_{p\times p} - P_A^L\right)$, we have

$$\left(1_{p\times p} - P_A^L\right) w = 0$$

$$\Leftrightarrow \quad w = P_A^L w$$

This implies $w \in \mathrm{Im}\, A$ since $P = A\left(A^T L A\right)^{-1} A^T L$. Therefore $\mathrm{Ker}\left(1_{p\times p} - P_A^L\right) = \mathrm{Im}\, A$. □

Parameterization of m- at submanifolds

Our primary use of generalized mixed co-ordinates is in the construction of parameterizations of arbitrary m- at submanifolds, required for the algorithms in Chapter 5. For xed d := Ad, the (p k)-dimensional submanifold $M_{B,d}$ may be described in generalized mixed co-ordinates by holding d constant and allowing b to vary; it is the level set of the projection function

$$F_M : R^k \quad R^{p\ k} ! \quad R^k \qquad F_M(a,b) := A^T r\ F(Aa + Bb) \tag{3.3.19}$$

corresponding to the level $F_M = A^T Ad$. We consider using $b\ 2\ R^{p\ k}$ as a parameter for the submanifold, in a co-ordinate embedding as follows.

De nition 3.3.8. For a p-dimensional dually- at submanifold Q, with $M_{B,d}$ an m- at submanifold of Q, de ne a primal embedding of $M_{B,d}$ as a map

$$L_d : R^{p\ k} ! \quad R^p \tag{3.3.20}$$

satisfying $A\ r\ F(L_d(b)) = d = A\ d$ and $B\ L_d(b) = b$ for all $b\ 2\ B\ Q$.

Note that the subscript d indicates which m- at submanifold is being embedded, i.e. d selects a ber of the ber bundle F_M. The following is a novel contribution to the literature to our knowledge, and shows that there is one unique primal embedding which is stated explicitly.

Proposition 3.3.9. For any m- at submanifold $M_{B,d}$, there exists a unique primal embedding of $M_{B,d}$ given by

$$L_d(b) = Aa(b) + Bb \tag{3.3.21}$$

where $a : R^{p\ k} ! \quad R^k$ is given by

$$a(b) = \operatorname{argmin}_a \underbrace{F(Aa + Bb) \quad a^T A^T Ad}_{:= j\ (a)} \tag{3.3.22}$$

Proof. First we show that (3.3.21) and (3.3.22) together de ne a valid primal embedding. For the rst condition, the gradient of j at the minimizer a(b) must be zero, i.e.

$$0 = r_a j\ (a(b)) = A^T r\ F(Aa(b) + Bb) \quad A^T Ad$$

Noting that $L_d(b) = Aa(b) + Bb$, this is equivalent to $A\ r\ F(L_d(b)) = d$. The second condition $B\ L_d(b) = b$ of De nition 3.3.8 follows immediately from (3.3.21), since columns of A and B are mutually orthogonal.

To show uniqueness, suppose R_d is a primal embedding. The combined columns of A and B form a basis for R^p, and thus

$$R_d(b) = Aa(b) + Bb(b)$$

for some functions a and b. By the second condition in De nition 3.3.8, we have

$$b = B\ R_d(b) = B\ Bb(b) = b(b)$$

and thus $b(b) = b$ for all b. By the first condition in the de nition, we find

$$\begin{aligned} & A\ r\ F\ R_d(b) = d \\ , \quad & A^T r\ F\ R_d(b) \quad A^T A d = 0 \\ , \quad & A^T r\ F(Aa(b) + Bb) \quad A^T A d = 0 \\ & , r\ _a j\ (a(b)) = 0 \end{aligned}$$

Note that the function j is convex and has a global minimizer since F is strictly convex. Thus

$$a(b) = a(b)$$

and hence $R_d(b) \quad L_d(b)$ for all b. Thus the primal embedding is unique. □

The following Lemma gives the differential matrix of the primal embedding.

Lemma 3.3.10. Let $M_{B,d}$ be a m- at submanifold and let $= A\ d$. The co-ordinate differential matrix of the m- at embedding $= L_d(b)$ is given by

$$DL_d(b) = \quad 1_p \ _p \quad P_A^{G(b)} \quad B$$

where $G(b) := r\ ^2F(L_d(b))$ is the Riemannian metric tensor evaluated at $L_d(b)$.

Proof. Differentiating (3.3.21), the i-j component of the differential matrix is given by

$$\frac{\P L^i}{\P b^j} = A^i_l \frac{\P a^l}{\P b^j} + B^i_j$$

The differential $Da = \frac{\P a^l}{\P b^j}$ is given by applying the Implicit Function Theorem B.0.4 to $F_M(a, b) = A^T A d$ where F is de ned in (3.3.19).

$$Da = \quad \frac{\P a^i}{\P a^j} \quad = \quad \frac{\P F_M}{\P a} \quad ^1 \frac{\P F_M}{\P b} \tag{3.3.23}$$

Note that

$$\begin{aligned} \frac{\P F^i}{\P a^j} &= \frac{\P}{\P a^j} A^{li} \frac{\P F}{\P q^l}(Aa + Bb) \\ &= \frac{\P q^s}{\P a^j} A^{li} \frac{\P^2 F}{\P q^l \P q^s}(Aa + Bb) \\ &= A^{sj} A^{li} \frac{\P^2 F}{\P q^l \P q^s}(Aa + Bb) = \quad A^T G(b) A \ _{i,j=1,..k} \end{aligned}$$

And similarly

$$\begin{aligned}\frac{\partial F^i}{\partial b^j} &= \frac{\partial}{\partial b^j} A^{li} \frac{\partial F}{\partial q^l}(Aa + Bb) \\ &= \frac{\partial q^s}{\partial b^j} A^{li} \frac{\partial^2 F}{\partial q^l \partial q^s}(Aa + Bb) \\ &= B^{sj} A^{li} \frac{\partial^2 F}{\partial q^l \partial q^s}(Aa + Bb) = \left(A^T G(b) B\right)_{i=1,..,k, j=1,..,p\ \ k}\end{aligned}$$

Substituting these expressions into (3.3.23), the result follows.

$$\begin{aligned}DL_d(b) &= \left(A\ \ A^T G(b) A\right)^{-1} A^T G(b) B + B \\ &= \left(1_p\ \ _p\right)\left(A\ \ A^T G(b) A\right)^{-1} A^T G(b)\ \ B \\ &= \left(1_p\ \ _p\right)\ \ P_A^{G(b)}\ \ B\end{aligned}$$

□

We will discuss how to evaluate the implicitly de ned parameterization L_d in Section 5.1.3. Recall from subsection 2.5.2, and in particular Theorem 2.5.5, that $M_{B,d}$ inherits geometric structures from the ambient manifold Q and is a dually- at manifold in its right. Furthermore, the following novel result shows $(M_{B,d}, b)$ is a co-ordinate chart on $M_{B,d}$ which is af ne with respect to the inherited primal connection.

Theorem 3.3.11. The parameterization b given by the unique primal embedding de nes af ne co-ordinates with respect to the inherited primal connection on $M_{B,d}$

Proof. We will show that the coef cients of the inherited primal connection (2.5.4)

$$\Gamma_{abc} = \frac{\partial L_d^j}{\partial b^c}\frac{\partial}{\partial b^a}\left(\frac{\partial L_d^i}{\partial b^b}\right)^{!} g_{ij} + \frac{\partial L_d^i}{\partial b^a}\frac{\partial L_d^j}{\partial b^b}\frac{\partial L_d^k}{\partial b^c} G_{jk} \qquad (3.3.24)$$

vanish, where $\frac{\partial L_d^i}{\partial b^a}$ is the i-a component of the co-ordinate differential matrix DL_d given in Lemma 3.3.10. Since $q = L_d(b)$ are primal co-ordinates, the connection coef cients G_{jk} are zero and the second term vanishes. The remaining term is the G-inner product of the c^{th} column of DL_d with the b^{th} column of the component-wise partial derivative with respect to b^a, $\frac{\partial}{\partial b^a} DL_d$.

To nd $\frac{\partial}{\partial b^a} DL_d$, we rst differentiate $\left(A^T G A\right)^{-1}$. Differentiating both sides of the identity $\left(A^T G A\right)^{-1} A^T G A = I$ and rearranging yields

$$\frac{\partial}{\partial b^a}\left(A^T G A\right)^{-1} = -\left(A^T G A\right)^{-1} A^T \frac{\partial G}{\partial b^a} A \left(A^T G A\right)^{-1} \qquad (3.3.25)$$

where we have use the fact that A^TGA^{-1} is symmetric by symmetry of the Hessian matrix G. Differentiating the expression for DL_d from (3.3.10) and substituting (3.3.25) for $\frac{\partial}{\partial b^a}A^TGA^{-1}$ then yields

$$\begin{aligned}\frac{\partial}{\partial b^a}DL_d &= -A\frac{\partial}{\partial b^a}A^TGA^{-1}A^TGB - A(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}B \\ &= A(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}A(A^TGA)^{-1}A^TGB - A(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}B\end{aligned}$$

and hence

$$\frac{\partial}{\partial b^a}DL_d = A(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}(A(A^TGA)^{-1}A^TG - I)B$$

The G inner-product of this with DL_d is then

$$\begin{aligned}DL_d^TG\frac{\partial}{\partial b^a}DL_d &= B^TGA(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}(A(A^TGA)^{-1}A^TG - I)B \\ &\quad - B^TGA(A^TGA)^{-1}(A^TGA)(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}(A(A^TGA)^{-1}A^TG - I)B \\ &= B^TGA(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}(A(A^TGA)^{-1}A^TG - I)B \\ &\quad - B^TGA(A^TGA)^{-1}A^T\frac{\partial G}{\partial b^a}(A(A^TGA)^{-1}A^TG - I)B \\ &= 0\end{aligned}$$

Thus the coefficients of the inherited primal connection are zero. □

Since b are primal co-ordinates on the dually-flat manifold $M_{B,d}$, there must exist a Bregman generator ϕ such that the inherited divergence on $M_{B,d}$ agrees with a Bregman divergence in b co-ordinates generated by ϕ. This is given explicitly by the following novel result.

Theorem 3.3.12. The restricted divergence on $M_{B,d}$ agrees with the Bregman divergence generated by ϕ, where

$$\phi(b) = F(L_d(b)) - d^TA^TAa(b) = F(L_d(b)) - d^TL_d(b) \tag{3.3.26}$$

Furthermore, the Legendre dual co-ordinate system is given by

$$\nabla\phi(b) = B^TBg = B^T\nabla F(L_d(b)) \tag{3.3.27}$$

and the metric on $M_{B,d}$ in b co-ordinates is

$$\nabla^2\phi(b) = B^TG(L_d(b))\left(1_p \otimes_p A(A^TG(L_d(b))A)^{-1}A^TG(L_d(b))\right)B \tag{3.3.28}$$

Proof. As usual, let F be the Bregman generator in primal co-ordinates on Q. By Theorem 2.5.3, the geometry on $M_{B,d}$ is de ned by the restriction of D_F to $M_{B,d}$, which we write in b co-ordinates as

$$D_F j_{M_{B,d}}(b, b_0) := D_F\ L_d(b), L_d(b_0) \tag{3.3.29}$$

where b and b_0 are the primal parts of the generalized mixed co-ordinate system (3.3.15) respectively corresponding to points $q, q_0 \ 2\ M_{B,d}$. Let $h := r\ F(L_d(b))$ and $h_0 := r\ F\ L_d(b_0)$ be dual co-ordinates in Q corresponding to q and q_0 respectively. Then by proposition 2.2.13, the inherited divergence may be written

$$D_F j_{M_{B,d}}(b, b_0) := D_F\ L_d(b), L_d(b_0)\ = D_F\ (h_0, h)$$

Since q and q_0 are points in $M_{B,d}$, by de nition there exists $g, g_0 \ 2\ R^{p\ k}$ such that $h_0 = Bg_0 + d$ and $h = Bg + d$. We must have

$$D_F j_{M_{B,d}}(b, b_0) = \underbrace{D_F\ (Bg_0 + d, Bg + d)}_{:= D_F\ j_{M_{B,d}}(g_0, g)}$$

Analogously to equation (3.3.29), we de ne a restricted dual divergence on $M_{B,d}$,

$$D_F\ j_{M_{B,d}}(g_0, g) := D_F\ (Bg_0 + d, Bg + d)$$

By Theorem 2.5.8, this coincides with the Bregman divergence generated by the restriction of F to $M_{B,d}$,

$$D_F j_{M_{B,d}}(b, b_0) = D_F\ j_{M_{B,d}}(g_0, g) = D_{F\ j_{M_{B,d}}}(g_0, g)$$

This is a Bregman divergence generated by $F\ j_{M_{B,d}}$ in the g co-ordinate system. The Legendre dual co-ordinate corresponding to this convex function has i^{th} component

$$\begin{aligned} g_i &= \frac{\P}{\P g^i} F\ j_{M_{B,d}} = \frac{\P}{\P g^i} F\ (Bg + d) \\ &= \frac{\P h^l}{\P g^i}\frac{\P F}{\P h^l} = B^l_i q^l = \quad B^T q\ _i \end{aligned}$$

where we have used the involution property of the Legendre transform of F, i.e. $r\ _h F\ (h) = q$ such that $r\ _q F(q) = h$. Since $q = Aa + Bb$ with $A^T B = B^T A = 0$, we must have

$$g\ = B^T q = B^T B b \qquad , \qquad b = \quad B^T B \quad ^1 g$$

since B is assumed to have full rank. We will complete the proof by showing that the Legendre transform of $F\ j_{M_{B,d}}$ agrees with (3.3.26). Firstly, note that by de nition of the Legendre transform

$$F\ j_{M_{B,d}} \quad (g\ ) := g \quad g \quad F\ (h) \tag{3.3.30}$$

where g is such that $\nabla_g F|_{M_{B,d}}(g) = g$, and $h = Bg + d$ depends on g implicitly. We substitute $F(h)$ for its definition in terms of the Legendre transform, and expand q and h in components in $\mathrm{Im}\, A$ and the orthogonal complement $\mathrm{Im}\, B$,

$$\begin{aligned} F(h) &= q \cdot h - F(q) \\ &= (Aa(b) + Bb) \cdot (Bg + d) - F(q) \\ &= d^T Aa(b) + g^T B^T Bb - F(q) \\ &= d^T Aa(b) + g^T g - F(q) \end{aligned}$$

Substituting this into (3.3.30) yields the result (3.3.26),

$$F|_{M_{B,d}}(g) = F(q) - d^T Aa(b) = \underbrace{F(L_d(b)) - d^T L_d(b)}_{=:\Phi(b)}$$

To show that dually coupled co-ordinates are given by $b = B^T Bg$, we note that $b = (B^T B)^{-1} g$ and differentiate via the chain rule. Alternatively, differentiating Φ directly,

$$\begin{aligned} \nabla_b \Phi(b) &= \nabla_b \left[F(L_d(b)) - d^T L_d(b) \right] \\ &= DL_d(b)^T \nabla_q F(q) - d^T q \\ &= \left[B - A \left(A^T G(L_d(b)) A \right)^{-1} A^T G(L_d(b)) B \right]^T (\nabla F(q) - d) \end{aligned} \tag{3.3.31}$$

where we have substituted DL_d from Lemma 3.3.10. Note that we have the decomposition

$$\nabla F(q) = h = Ad + Bg = d + Bg$$

Substituting this into (3.3.31) and using $A^T B = 0_{k \times (p-k)}$ yields (3.3.27),

$$\begin{aligned} \nabla_b \Phi(b) &= \left[B^T - B^T GA \left(A^T GA \right)^{-1} A^T \right] (d + Bg - d) \\ &= B^T Bg = B^T \nabla F(q) \end{aligned}$$

Another differentiation yields the Hessian matrix of Φ, and hence the metric on the m-flat submanifold in b co-ordinates (3.3.28),

$$\begin{aligned} \nabla_b^2 \Phi(b) &= \nabla_q \left[B^T \nabla F(q) \right] \cdot DL_d(b) \\ &= B^T \nabla^2 F(q) \left[B - A \left(A^T G(L_d(b)) A \right)^{-1} A^T G(L_d(b)) B \right] \\ &= B^T G(L_d(b)) \left[I - A \left(A^T G(L_d(b)) A \right)^{-1} A^T G(L_d(b)) \right] B \end{aligned}$$

□

Note that Theorem 3.3.11 can be viewed as a corollary of Theorem 3.3.12; in the proof of Theorem 2.4.4 we showed that a co-ordinate system in which a Bregman divergence is de ned is af ne in the induced geometry.

The metric (3.3.28) in the b co-ordinate system is a Schur complement of the ambient metric in (a, b) co-ordinates. The transformation (3.3.14) from q to (a, b) transforms the metric according to

$$G(q) \; ! \quad A \quad B \;^{T} G(q) \quad A \quad B \quad = \quad \begin{matrix} A^T \\ B^T \end{matrix} \quad G(q) \quad A \quad B \quad = \quad \begin{matrix} A^T G A & A^T G B \\ B^T G A & B^T G B \end{matrix}$$

The Schur complement of the block $A^T G A$ in this matrix is given by

$$B^T G B \quad B^T G A \;\; A^T G A \quad {}^{1} A^T G B$$

which coincides with the metric (3.3.28) on $M_{B,d}$ in the b co-ordinate system.

Mapping of e- at into m- at Submanifolds

In the previous subsection, we studied the map $L_d : R^{p \;\; k} \; ! \quad R^p$ that represents an embedding of an m- at submanifold in primal co-ordinates. In this subsection, we will show how this can be used to construct a map which sends e- at submanifolds to m- at submanifolds.

Consider the e- at submanifold de ned by the image of B in primal co-ordinates. For any point $q \; 2 \; R^p$, the projection onto the image of B is given by $b = Bb = BB \; q$. The coef cients b are also the parameters of the primal embedding of a $(p \quad k)$-dimensional m- at submanifold $M_{B,d}$, such that

$$L_d(b) = Aa_d(b) + Bb$$

This motivates the following de nition.

De nition 3.3.13. Let Q be a p-dimensional dually- at manifold. De ne

$$L_d : R^p \; ! \quad R^p$$
$$L_d(q) = L_d(B \; q)$$

The mapping L_d maps points in R^p to $M_{B,d}$. Clearly, the map preserves the Euclidean-orthogonal projection of q onto Im B, i.e.

$$B \; L_d \, (q) = B \; q \qquad) \qquad BB \; L_d \, (q) = BB \; q$$

Furthermore, by de nition $L_d(q)$ belongs to the ber labelled by d;

$$A \; r \; F(L_d(q)) = A \; r \; F(L_d(B \; q)) = d$$

3.3.4 Decomposition of Divergence

Mixed co-ordinates admit simple e- and m-projections onto the leaves of the orthogonal foliations. Suppose a point $q \in Q$ has mixed co-ordinates $x = (h_{[k]}, q_{[>k]})$, where $h_{[k]} \in \mathbb{R}^k$ are the k dual components and $q_{[>k]} \in \mathbb{R}^{p-k}$ are the $p-k$ primal components. Consider the e-projection of q onto the m- at submanifold $M_k(c)$, the set of points whose dual co-ordinates satisfy (3.3.7). The following is a generalization of Theorem 6 from Amari [2].

Proposition 3.3.14. The e-projection of $q \in Q$ with mixed co-ordinates $x_q = (h_{[k]}, q_{[>k]})$ onto $M_k(c)$ has mixed co-ordinates

$$x_{P_e q} = (c, q_{[>k]})$$

Similarly, the m-projection of q onto $E_k(d)$ has mixed co-ordinates

$$x_{P_m q} = (h_{[k]}, d)$$

Proof. Note that by de nition, $x_q \in E_k(q_{[>k]})$. The leaf $E_k(q_{[>k]})$ must intersect $M_k(c)$ at the point $P_e q$ with mixed co-ordinates $(c, q_{[>k]})$. Suppose that the full primal co-ordinates of q are $q_q = (q_{[k]}, q_{[>k]})$, and let the full primal co-ordinates of $P_e q$ be $(q_c, q_{[>k]})$. Then the e-geodesic connecting q with $P_e q$ may be parameterized in primal co-ordinates as

$$q(t) = \begin{matrix} (1-t)q_{[k]} + tq_c \\ q_{[>k]} \end{matrix}$$

The e-geodesic clearly lies entirely in the e- at submanifold $E_k(q_{[>k]})$, and thus its velocity vector at $P_e q$ is orthogonal to any tangent vector to $M_k(d)$. It follows that $P_e q$ is the e-projection of q onto $M_k(d)$.

The proof for the m-projection of q onto $E_k(d)$ is similar; the m-geodesic connecting x_q and $(h_{[k]}, d)$ lies in the m- at leaf $M_k(h_{[k]})$, and hence is orthogonal to $E_k(c)$ at $(h_{[k]}, d)$. □

Part II

Algorithms

Chapter 4

Review of MCMC and Conditional Sampling on Submanifolds

This chapter presents a brief review of Markov Chain Monte Carlo (MCMC) and recent MCMC algorithms which aim to improve performance using geometrically motivated ideas. Particular attention is given to algorithms which resample the target distribution constrained to a submanifold, such as the Gibbs sampler and Hamiltonian Monte Carlo. Such general conditional sampling formalized in terms of a disintegrationof the target distribution over a ber bundle.

4.1 Markov Chain Monte Carlo

Suppose $p(q)$ is a probability density on R^p, and $q^{(0)}, q^{(1)}, .. 2 R^p$ are the iterates of some stochastic process that is meant to simulate$p$. We use superscripts with brackets to emphasise different iterates of the process rather than components of a vector. The associatedMonte Carlo Estimatorfor $m := E_p(f(q))$ is the Monte Carlo sample mean

$$m = \frac{1}{N} \sum_{i=0}^{N} {}^{1} f \quad q^{(i)} \tag{4.1.1}$$

where $f : R^p ! \quad R$ is some function of the random vector q. The expected value $m$of f is an example of a quantity of interestwhich may summarize the posterior, such as a variance, or is otherwise relevant to the area of application. Computing quantities of interest is the goal of MCMC. The number N is the Monte Carlo sample size, a completely distinct quantity to the sample size n which is the number of observations.

If the iterates $q^{(0)}, .., q^{(N \ 1)}$ are i.i.d., i.e. independently drawn directly from $p(q)$, the method is called Ordinary Monte Carlo In this case, one can easily show that the Monte Carlo estimator is unbiased with variance $Var_p(f)/N$, and is asymptotically normally distributed by the Central Limit Theorem. This means convergence is rather slow; to halve the Monte Carlo Standard Error (MCSE), $StdDev_p(f)/\sqrt{N}$, we must take four times as many Monte Carlo samples.

In practice, i.i.d. sampling from p is not possible for many target distributions. Instead, correlated samples are typically drawn by simulating a Markov chain $q^{(0)}, q^{(1)}, \ldots$ whose stationary distribution is $p(q)$; this is Markov Chain Monte Carlo (MCMC). We brie y recount the introduction by Geyer [36].

Recall that a stochastic process is Markov if the conditional distribution of $q^{(i+1)}$ given $q^{(0)}, .., q^{(i)}$ only depends on $q^{(i)}$, in which case the process is determined by the following two distributions:

1. An initial distribution $q^{(0)} \quad p^{(0)}$ for the rst iterate of the chain;
2. A transition kernel k, such that the conditional probability density of $q^{(i+1)}$ given $q^{(i)}$ is written $k\ q^{(i+1)} j q^{(i)}$. We assume the kernel is stationary, i.e. it does not explicitly depend on i.

Suppose the marginal distribution of $q^{(i)}$ has probability density $p^{(i)}(q)$ for all i. The transition kernel must satisfy

$$p^{(i+1)}(q) = \int_Q k(q j q') p^{(i)}(q') dq' \tag{4.1.2}$$

The i-step transition kernel $k^{(i)}(q^{(i)} j q^{(0)})$ of the chain is the conditional probability density of $q^{(i)}$ given $q^{(0)}$. Clearly, the i-step transition kernel satis es

$$p^{(i)}(q) = \int_Q k^{(i)}(q j q') p^{(0)}(q') dq' \qquad \text{for all } q \ 2 \ Q$$

If we enforce that the chain starts at a given point $q^{(0)}$, i.e. the initial distribution is $p^{(0)}(q) = d(q \quad q^{(0)})$, then the above equation shows that the marginal distribution of $q^{(i)}$ is $p^{(i)}(q) = k^{(i)}(q j q^{(0)})$.

A Markov chain is *stationary* if the marginal distribution $p^{(i)}$ does not depend on i. A distribution $p(q)$ is called a *stationary* or *invariant* distribution of the transition kernel k if it is preserved by k,

$$p(q) = \int_Q k(q j q') p(q') dq' \tag{4.1.3}$$

A suf cient condition for p to be invariant is if the transition kernel is *reversible* with respect to p,

$$k(q' j q) p(q) = k(q j q') p(q') \text{ for all } q', q \ 2 \ Q \tag{4.1.4}$$

This condition is also known as *detailed balance* with respect to p. To see that reversibility implies invariance, we substitute (4.1.4) into the right-hand side of (4.1.3),

$$\int_Q k(q j q') p(q') dq' = \int_Q k(q' j q) p(q) dq' = p(q)$$

since $k(\ j q)$ is a probability density.

If the initial distribution $p^{(0)}$ is an invariant distribution for k, the marginals $p^{(i)}$ p for all i 0 by induction on (4.1.2) and hence the Markov chain is stationary. In particular, if the transition kernel is reversible with respect to the initial distribution, then $p = p^{(0)}$ is an invariant distribution and hence the chain is stationary. In this case the joint distribution of $q^{(i+1)}$ and $q^{(i)}$ has density

$$k(q^{(i+1)}|q^{(i)})p(q^{(i)}) = k(q^{(i)}|q^{(i+1)})p(q^{(i+1)})$$

This shows that the joint distribution of $q^{(i)}$ and $q^{(i+1)}$ is exchangeable, so that the laws for moving forwards or backwards in time are the same; a Markov chain with this property is called *reversible*. We have shown that if k is reversible with respect to the initial distribution, then the chain itself is reversible and stationary.

The key question for MCMC is whether the Monte Carlo estimator (4.1.1) is asymptotically unbiased and what its asymptotic variance is, when the samples $q^{(0)}, .., q^{(N-1)}$ are the iterates of a Markov chain whose invariant distribution is p rather than i.i.d. samples from p. Under suitable conditions, the Markov chain Central Limit Theorem is satisfied for stationary chains, which states that the Monte Carlo estimator (4.1.1) for the Markov chain is asymptotically normally distributed with mean $m = E_p(f)$ and variance $\frac{s^2}{N}$, where

$$s^2 = \mathrm{Var}(f(q^{(i)})) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(f(q^{(i)}), f(q^{(i+k)})) \tag{4.1.5}$$

Since the chain is stationary, the index i is arbitrary. Roberts and Rosenthal [65] give various conditions for the Markov chain CLT to hold; it is sufficient that $E_p(f^2) < \infty$, the chain is reversible and geometrically ergodic, meaning that if p is the invariant distribution, then for p-almost all $q^{(0)} \in Q$,

$$\|k^{(i)}(\cdot|q^{(0)}) - p(\cdot)\|_{TV} \leq M(q^{(0)})r^i \tag{4.1.6}$$

for $i = 1, 2, ..$, some $r < 1$ and $M(q^{(0)}) < \infty$, where $\|\cdot\|_{TV}$ is the total variation norm. Geometric ergodicity simply means that after starting the chain from almost any point, the marginals converge to p geometrically in total variation. We emphasize that this is not "geometric" in the sense of using geometry. If M is a constant independent of q_0, then the chain is uniformly ergodic.

In practice, for a given invariant distribution p one cannot simulate the corresponding stationary Markov chain, because this requires being able to draw a sample from p for the initial distribution so that $p^{(0)} = p$, which is assumed to be impossible or infeasible or else we should be doing Ordinary Monte Carlo. However, the Markov chain CLT still holds when the chain is geometrically ergodic, or under similar conditions, if we start the chain at p-almost any $q^{(0)} \in Q$ [65]. The iterates in the asymptotic variance (4.1.5) still refer to those of the stationary chain.

4.1.1 The Metropolis-Hastings Framework

MCMC requires the construction of a Markov chain whose transition kernel is reversible with respect to the target $p(q)$, satisfies conditions for the Markov chain CLT to hold

and can be simulated on a computer. A general method for constructing such chains is the Metropolis-Hastings algorithm, which draws a sample from a proposal kernel $q(q' \mid q^{(i)})$ which can be simulated. The proposed sample q' is then accepted, i.e. the next iterate of the chain becomes $q^{(i+1)} := q'$, with probability

$$a = \min\left(\frac{q(q^{(i)} \mid q')\, p(q')}{q(q' \mid q^{(i)})\, p(q^{(i)})}, 1\right) \tag{4.1.7}$$

otherwise the proposal is rejected; we set $q^{(i+1)} := q^{(i)}$. Since the acceptance probability (4.1.7) involves a ratio of target densities, we need only evaluate p up to multiplicative constants, so it is not necessary to know normalization constants of p. One can show [36], [65] that the Markov chain resulting from the Metropolis-Hastings algorithm is reversible with respect to $p(q)$. All reversible sampling algorithms are in fact special cases of Metropolis-Hastings, including Gibbs updates and Hamiltonian Monte Carlo; this can be seen by substituting any kernel k which is reversible with respect to p for the proposal q in (4.1.7) - such proposals are always accepted.

An important case of the Metropolis-Hastings algorithm is when the proposal kernel is symmetric, i.e. $q(q' \mid q) = q(q \mid q')$ for all $q', q \in Q$, when the acceptance probability (4.1.7) becomes

$$a = \min\left(\frac{p(q')}{p(q^{(i)})}, 1\right) \tag{4.1.8}$$

This case is called the Metropolis algorithm. The proposal is always accepted if the target density is higher at q' than at the previous iterate $q^{(i)}$. For example, the (Gaussian) Random Walk Metropolis algorithm uses multivariate normal proposals of the form

$$q' \sim \text{Normal}(q, l^2 S)$$

where S is some covariance matrix, and $l > 0$ is some scaling factor. Clearly the proposal kernel $q(q' \mid q) = q(q \mid q')$ and the acceptance probability reduces to (4.1.8).

Note that Markov chains resulting from the Metropolis-Hastings procedure are not necessarily geometrically ergodic, and the Markov chain CLT is not guaranteed to hold in general. Counter-examples may be constructed with a result of Jarner and Hansen [42], who show that if a Random Walk Metropolis algorithm is geometrically ergodic, then the target distribution p must have at least exponentially light tails. For example, Livingstone and Girolami [50] offer a Cauchy target distribution as a counter-example, which has heavier than exponential tails, and hence the Random Walk Metropolis algorithm is not geometrically ergodic. Robert and Casella [62] give sufficient conditions for the marginal distributions in a Metropolis-Hastings algorithm to converge to p in the total variation norm, although not necessarily geometrically as in (4.1.6).

4.1.2 Measuring Performance

If a Markov chain is geometrically or uniformly ergodic, the convergence rate r in (4.1.6) gives one measure of performance; the rate at which the marginal distributions of the

MCMC iterates converge to the target distribution, with lower values indicating faster convergence. In practice, there may be analytic results for bounds on the convergence rate for some samplers under specific assumptions, but estimation of r from MCMC output for a general model is usually not possible.

The goal of MCMC is of course to estimate expectations with respect to the target distribution, and hence practitioners often use proxies for the accuracy of the Monte Carlo estimator as measures of performance. For example, the asymptotic variance of the Monte Carlo estimator has a term involving the covariance, or equivalently correlation, of f under successive iterates of the chain. Rewriting the s^2 (4.1.5) part of the asymptotic variance $\frac{s^2}{N}$,

$$s^2 = \mathrm{Var}_p(f(q^{(i)})) \underbrace{\left(1 + 2\sum_{k=1}^{\infty} \frac{\mathrm{Cov}(f(q^{(i)}), f(q^{(i+k)}))}{\mathrm{Var}(f(q^{(i)}))}\right)}_{=:t} = t\,\mathrm{Var}_p(f(q^{(i)})) \qquad (4.1.9)$$

where t is called the integrated autocorrelation time (IAT). Note that some authors such as Liu [49] take the IAT to be $\frac{1}{2}$ the value defined here; we follow the convention of Roberts and Rosenthal [63]. When the chain consists of independent samples from p, (4.1.9) shows that $t = 1$. The presence of autocorrelation in the Markov chain usually results in $t > 1$.

Another proxy for the asymptotic variance is the effective sample size (ESS) N_{eff}. This is defined as the equivalent number of independent samples from p such that the variance of the Monte Carlo estimators for independent and correlated samples are equal:

$$\frac{\mathrm{Var}\left(f\left(q^{(i)}\right)\right)}{N_{\mathrm{eff}}} = \frac{s^2}{N}, \quad N_{\mathrm{eff}} = N\frac{\mathrm{Var}(f(q^{(i)}))}{s^2} = \frac{N}{t}$$

This gives an interpretation for the IAT; t is asymptotically the number of stationary MCMC samples required to reduce the variance by an amount equivalent to a single independent sample. Low IATs, or equivalently high effective sample sizes, are therefore indicative of good MCMC performance.

Any metric used to compare MCMC algorithms should take into account the computational cost of producing the correlated samples. A popular measure of efficiency is the number of effective samples per second of computation, N_{eff}/T, where T is the runtime of the MCMC algorithm.

4.1.3 The Gibbs Sampler

The Gibbs sampler introduced by Geman and Geman [34] draws from conditional distributions of the target distribution. As usual, suppose there exists a target density $p(q)$ on $\mathbb{R}^p$, and decompose the current state q into d subvectors $q = (q_1, \ldots, q_d)$. For this subsection, we use subscripts to index the subvectors to distinguish them from the iterates of the Markov chain, for which we use superscripts. We allow the number of components in each subvector q_j to vary with $j = 1, \ldots, d$. We adopt the convention of

Gelman et al [33] of writing q_{-j} for the union of all subvectors not including subvector q_j,

$$q_{-j} = (q_1, .., q_{j-1}, q_{j+1}, .., q_d)$$

A Gibbs update draws q_j from the target conditioned on the current values of the complementary subvectors, i.e.

$$\text{Set } q_j = q_j^0 \text{ where } q_j^0 \sim p(q_j | q_{-j})$$

This Gibbs update for the j^{th} subvector is reversible with respect to the full target distribution. To see this, we adopt the small abuse of notation $q = (q_j, q_{-j})$, so that the kernel for the Gibbs update may be written

$$k^j(q_j^0, q^0_{-j} | q_j, q_{-j}) = \begin{cases} p(q_j^0 | q_{-j}) & \text{if } q_j^0 = q_j \\ 0 & \text{if } q_j^0 \neq q_j \end{cases}$$

If $q^0_{-j} = q_{-j}$, then reversibility (4.1.4) follows from the direct calculation

$$p(q_j^0 | q_{-j}) p(q_j, q_{-j}) = p(q_j^0, q^0_{-j}) p(q_j | q^0_{-j})$$

and otherwise both sides of (4.1.4) are 0. Hence the Gibbs update kernel is reversible with respect to the target. Gibbs updates are therefore Metropolis-Hastings samplers with a proposal $q = p(q_j^0 | q_{-j})$, which is always accepted.

Algorithm 1 Systematic-scan Gibbs Sampler

```
for i = 1, .., N do
    for j = 1, .., d do
        q_j^(i) ~ q_j^0 ~ p(q_j | q_1^(i), .., q_{j-1}^(i), q_{j+1}^(i-1), .., q_d^(i-1))
    end for
end for
```

In order to facilitate moves around the entire state space Q, i.e. to ensure an irreducible Markov Chain, Gibbs updates can be combined in various ways. Firstly, one could update each subvector in order of index, called systematic-scan Gibbs sampling [49], shown in algorithm 1. Since the Gibbs updates are applied consecutively, the kernel for an entire systematic-scan Gibbs sweep is the composition of the kernels for each Gibbs update. The composition kernel preserves p, but is not reversible with respect to the target distribution in general even though each Gibbs update is [36]. Alternatively, the random-scan Gibbs sampler chooses j randomly from $\{1, .., d\}$, and applies the corresponding Gibbs update for q_j. The transition kernel for a random-scan Gibbs iteration is a mixture $k = \sum_j \lambda_j k^j(q^0 | q)$, where k^j is the transition kernel for the j^{th} Gibbs update chosen with probability λ_j. Since each k^j is reversible with respect to p, one can easily show that the random-scan or mixture kernel is also reversible with respect to p.

4.2 Resampling using Disintegrations

In this section, we describe a framework for more general conditional sampling. Suppose we resample conditioned on some smooth function of the current state $F(q)$. The update is therefore constrained to a surface or submanifold of the state space corresponding to a level set of F. Liu and Sabatti [48], [49] describe such an update in terms of a randomly selected element of a group of transformations of state space. Geyer [36] notes that conditioning on arbitrary functions of the state in this way is equivalent to a regular Gibbs update under a change of variable, i.e. in a different co-ordinate chart. We will take the approach of Betancourt et al [12], in which the target distribution is disintegrated over the fiber bundle or foliation defined by F.

We first make clear what is meant by conditioning the target distribution on a submanifold. Since lower-dimensional submanifolds have Lebesgue measure zero, the textbook approach to conditioning using a ratio of probability densities is not well-defined. Instead, we adopt the concept of a *disintegration* [19], [12] to ensure regular conditional probabilities are well-defined. As usual, let Q be a p-dimensional smooth manifold equipped with its Borel sets so that it is a measure space with m a sigma-finite measure on Q, let $F : Q \to \mathbb{R}^k$ be measurable, and denote the push-forward of m to $\mathbb{R}^k$ under F by m^F, which we also require to be sigma-finite. We state the definition of a disintegration from Chang and Pollard [19], adapted to be slightly less general to suit our purposes here.

Definition 4.2.1. *A set of measures $\{m_z \mid z \in \mathbb{R}^k\}$ on the Borel sets of Q is a disintegration of m with respect to F if the following properties are satisfied:*

1. *Each m_z is a sigma-finite measure on the Borel sets of Q such that $m_z(F \neq z) = 0$, i.e. all mass is concentrated on the preimage or fiber $F^{-1}(z)$.*
2. *For a measurable function $h : Q \to \mathbb{R}$, the map*
$$z \mapsto \int_Q h(q)\, dm_z(q)$$
is a measurable function on $\mathbb{R}^k$.
3. *For all measurable $h : Q \to \mathbb{R}$,*
$$\int_Q h\, dm = \int_{\mathbb{R}^k} \left(\int_Q h(q)\, dm_z(q) \right) dm^F(z) \tag{4.2.1}$$

The integral in (4.2.1) is well-defined by the two previous conditions. Chang and Pollard [19] show that the measures m_z are probability measures on Q if the pushforward measure m^F is sigma-finite, and thus disintegrations define a *regular* conditional probability. The function F may be thought of as a projection map of a fiber bundle with total space Q and base space a subset of $\mathbb{R}^k$; the disintegration splits a measure into an integral over the base space of measures on the fibers. Betancourt et al [12] describe disintegrations in the differential-geometric terms of volume forms on a smooth

manifold and apply their theory to study Hamiltonian Monte Carlo, in which the relevant fibre bundle is defined by the level sets of the Hamiltonian function; we will explore this further in Section 4.2.1.

In order to understand the relevance of disintegrations for conditional sampling, we must first recast some of our earlier definitions into a measure-theoretic form. The following generalization of a Markov kernel is adapted from that given by Betancourt et al [12]. The Borel sets of Q are denoted by $\mathcal{B}(Q)$.

Definition 4.2.2. A Markov kernel $k : Q \times \mathcal{B}(Q) \to [0,1]$ is a map which is a measurable function in its first argument when any $B \in \mathcal{B}(Q)$ is plugged into its second argument,

$$k(\,\cdot\,, B) : Q \to [0,1]$$

and is a probability measure in its second argument for any $q \in Q$ as its first argument,

$$k(q, \,\cdot\,) : \mathcal{B}(Q) \to [0,1]$$

The measure-theoretic generalization of a stationary distribution is given by the following.

Definition 4.2.3. Let $k : Q \times \mathcal{B}(Q) \to [0,1]$ be a Markov kernel. A measure $\mu : \mathcal{B}(Q) \to [0,1]$ is an invariant distribution of k if

$$\mu(B) = \int_Q k(q, B)\, \mu(dq) \tag{4.2.2}$$

for any Borel set $B \in \mathcal{B}(Q)$.

The following result shows that if the elements of the disintegration of μ are stationary distributions of a Markov kernel, then so is μ

Theorem 4.2.4. Let $(Q, \mathcal{B}(Q), \mu)$ be a probability space where Q is a p-dimensional manifold and let $\{\mu_z \mid z \in \mathbb{R}^k\}$ be a disintegration of μ with respect to the fibre bundle $F : Q \to \mathbb{R}^k$ with $k < p$. If $k : Q \times \mathcal{B}(Q) \to [0,1]$ is a Markov kernel on Q that preserves each μ_z, i.e.

$$\int_Q k(q, U)\, \mu_z(dq) = \mu_z(U) \qquad \text{for all } z \in \mathbb{R}^k, U \in \mathcal{B}(Q) \tag{4.2.3}$$

then k preserves μ.

Proof. By the third condition (4.2.1) of the definition of a disintegration, for any Borel set U we have

$$\int_Q k(q, U)\, \mu(dq) = \int_{\mathbb{R}^k} \left(\int_Q k(q, U)\, \mu_z(dq) \right) \mu^F(dz)$$

Substituting (4.2.3) into the right-hand side of this, we find

$$\int_Q k(q, U)\, \mu(dq) = \int_{\mathbb{R}^k} \mu_z(U)\, \mu^F(dz)$$

Noting that $m_z(U) = \int_Q 1_U(q) m_z(q)$ where 1_U is the indicator function for U, and applying the disintegration (4.2.1) again with $h(q) = 1_U(q)$,

$$\begin{aligned}\int_Q k(q,U)m(dq) &= \int_{R^k} \int_Q 1_U(q) m_z(q) \; m^F(dz) \\ &= \int_Q 1_U(q) m(dq) = m(U)\end{aligned}$$

We see that the condition (4.2.2) for k to preserve m is satisfied. □

This theorem provides another generalization of a Gibbs update; we can sample from the element m_z of a disintegration, where $F(q^{(j)}) = z$. Such an update moves within the submanifold defined by the level set $\{q | F(q) = z\}$, analogously to a Gibbs update which resamples within the level sets of one or several co-ordinate functions.

4.2.1 Review of Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a sampling algorithm based on Hamiltonian dynamics which exploits geometric ideas, and in particular disintegrations of an extended target distribution. The method was first developed in 1987 by Duane et al [26], who called it Hybrid Monte Carlo. A variant of HMC, the No U-Turn Sampler (NUTS) [41], has been successful as the primary sampling algorithm for Bayesian inference in software packages such as Stan [18] and PyMC3 [69]. We briefly review HMC and its geometry here.

Given a target probability measure on a p-dimensional smooth manifold Q, HMC samples from an extended $2p$-dimensional distribution on the cotangent bundle T^*Q as defined in Section 2.1.2, which is called phase-space in this context. Although Hamiltonian dynamics and HMC can be described without co-ordinates [4], [12], we will assume the existence of a global co-ordinate chart (Q, q) to simplify our exposition, where co-ordinates take values in the set $q(Q) = Q \subseteq R^p$. Elements of cotangent space can then be identified with pairs (q, w) of two vectors in R^p, where q are co-ordinates for a point $q \in Q$, analogous to the degrees of freedom of a physical system, and w is the vector of components of a covector in T_q^*Q and is physically analogous to a momentum vector.

Suppose that the probability measure on Q is pushed forward under the global co-ordinate map to a p-dimensional Lebesgue density $p(q)$, whose normalization may not be known. One chooses some log-concave conditional momenta distribution, written $p(w|q)$, so that the joint distribution on position and momenta is

$$p(q,w) = p(q)p(w|q) = \frac{1}{Z}\exp(-H(q,w)) \tag{4.2.4}$$

where the negative log joint density is called the Hamiltonian or energy function H,

$$H(q,w) := -\log p(q,w) = \underbrace{-\log p(q)}_{:= V(q)} \underbrace{-\log p(w|q)}_{:= K(q,w)} \tag{4.2.5}$$

and Z is a normalization constant. The individual terms $V(q) := -\log p(q)$ and $K(q,w) := -\log p(w|q)$ are respectively called the *potential* and *kinetic* energy, in direct analogy to their physical counterparts. The joint distribution (4.2.4), also known as the Gibbs distribution, is supported on the product space $Q \times \mathbb{R}^p$.

Starting from an initial position $q^{(0)}$, the basic Hamiltonian Monte Carlo algorithm generates a proposal $z' = (q', w')$ via the following two steps:

1. Sample momenta from the conditional distribution $w_0 \sim p(w|q = q^{(0)})$. The conditional distribution is usually chosen to be multivariate Gaussian, so we may use the Cholesky decomposition to draw independent samples from this distribution.

2. With initial conditions $q(0) = q_0$ and $w(0) = w^{(0)}$, solve Hamilton's equations

$$\frac{dq^i}{dt} = \frac{\partial H}{\partial w_i} \qquad \frac{dw_i}{dt} = -\frac{\partial H}{\partial q^i}, \qquad i = 1, .., p \tag{4.2.6}$$

   until a fixed time T. Set the proposal to be the point at the end of this trajectory with a negated momentum vector, $q' = q(T)$ and $w' = -w(T)$. In most practical cases, Hamilton's equations cannot be solved exactly and a numerical timestepping method must be employed.

The proposal (q', w') is accepted or rejected using the canonical (joint) target density $p(q,w) \propto \exp(-H(q,w))$. One can show [8] that the Hamiltonian Monte Carlo proposal kernel is symmetric, and hence the acceptance probability (4.1.7) reduces to the ratio of densities

$$\begin{aligned} a &= \min\left(\frac{p(q', w')}{p(q^{(0)}, w^{(0)})}, 1\right) \\ &= \min\left(\frac{\frac{1}{Z}\exp(-H(q', w'))}{\frac{1}{Z}\exp(-H(q^{(0)}, w^{(0)}))}, 1\right) \\ &= \min\left\{\exp(H(q^{(0)}, w^{(0)}) - H(q', w')), 1\right\} \end{aligned}$$

The Hamiltonian H is conserved [55] along exact solutions of the system (4.2.6), that is

$$\frac{d}{dt} H(q(t), w(t))\Big|_{t=t_0} = 0$$

for all t_0, and so $H(q_0, w_0) = H(q', w')$. Exact solutions of Hamilton's equations therefore yield proposals which are always accepted. However, in practice the numerical integration of the dynamical system does not exactly conserve the Hamiltonian, and hence the acceptance probability can be less than 1. HMC samples the joint distribution (4.2.4) on position and momenta, with the position variables being marginal samples from the intended target distribution p.

Hamiltonian Monte Carlo only requires the target density $p(q,w)$ up to a multiplicative constant, because this is equivalent to an additive constant in the Hamiltonian

$H = \log p(q, w)$. Solutions of Hamilton's equations (4.2.6) are invariant under addition of a constant to the Hamiltonian since (4.2.6) only involves partial derivatives of H. Thus, the entire HMC algorithm can sample from a non-normalized density.

Since exact solutions of (4.2.6) preserve the Hamiltonian, the trajectories are confined to the level sets of H;

$$H^{-1}(E) = \{(q, w) \in Q \times \mathbb{R}^p | H(q, w) = E\}$$

The Hamiltonian can be viewed as the projection map of a bundle, with total space given by $Q \times \mathbb{R}^p$, base space $\mathbb{R}$ and the level sets of the Hamiltonian making up the fibers. The joint distribution (4.2.4) admits a disintegration over this bundle whose elements are called the *microcanonical distributions* [12], [8], which are conditional probability distributions on each level set or fiber. Hamiltonian dynamics (4.2.6) can be shown to preserve the microcanonical distributions [12].

To fully specify the Hamiltonian and hence the joint distribution on position and momenta, one must choose a conditional distribution $p(w|q)$ for the momenta given position. Suppose the conditional momenta given q has a zero-mean multivariate normal distribution, whose covariance matrix G depends on q in general;

$$p(p|q) = \text{Normal}(p|0, G(q))$$

which yields the Hamiltonian

$$H = \frac{1}{2} p^T G^{-1}(q) p + \frac{1}{2} \log \det G(q) - \log p(q) \propto \frac{1}{2} p^T G^{-1}(q) p + \frac{1}{2} \log \det G(q) + V(q)$$

The term $\frac{1}{2} p^T G^{-1}(q) p$ corresponds to the kinetic energy of a particle on a p-dimensional Riemannian manifold, whose metric has components $G(q)$ in the (Q, q) co-ordinate chart [4]. The resulting HMC algorithm is therefore called Riemannian Manifold Hamiltonian Monte Carlo (RM-HMC) introduced by Girolami and Calderhead [39]. They suggest a metric given by the Fisher infomation plus the Hessian of the negative log prior density,

$$G(q) = I(q) - \nabla^2 \log p_0(q) \tag{4.2.7}$$

As we noted in Chapter 3, Proposition 3.1.2 implies that this is the metric in the posterior Bregman geometry for an exponential family likelihood. Alternatively, a constant metric $G(q) \equiv M$ may be chosen, resulting in a flat curvature tensor and hence is referred to as Euclidean HMC.

A similar algorithm is based on Langevin diffusions, called the Metropolis Adjusted Langevin Algorithm (MALA) [64], [7]. Girolami and Calderhead [39] also extend MALA to the Riemannian manifold MALA (RM-MALA) algorithm. Both RM-HMC and RM-MALA admit simplified versions for which the connection coefficients are set to zero everywhere, reducing the required computational cost.

Riemannian manifold HMC/MALA algorithms usually outperform their Euclidean counterparts in terms of effective sample size; see Girolami and Calderhead [39], and the review by Livingstone and Girolami [50]. For models whose prior or posterior

density is not log-concave, Betancourt [9] suggests a general SoftAbs metric for RM-HMC which is a modified version of the Hessian matrix with eigenvalues forced to be positive. For hierarchical models, RM-HMC with the SoftAbs metric has been shown to have favourable performance and exploration of parameter space compared to Euclidean HMC [11]. RM-HMC has an interesting use case in sampling from densities on submanifolds isometrically embedded in $\mathbb{R}^p$, for which the algorithm simplifies and is called Geodesic Monte Carlo [16].

Chapter 5

Sampling using Dually-Flat Geometry

In this chapter I will describe the main contributions of the thesis; MCMC algorithms which alternately sample the target distribution conditioned on orthogonal e- and m-at submanifolds. For any given Bregman generator F the algorithms sample on orthogonal foliations in the dually-at geometry generated by F, i.e. orthogonal everywhere in the r^2F metric. The hope is that when $F = \log p$ for a log-concave density p, so that orthogonality is with respect to the Hessian of the negative log density, orthogonal directions are close to independent sampling directions and the resulting Markov chains have low autocorrelation.

I will suggest two speci c algorithms following this strategy. Firstly, the Orthogonal Gibbs algorithm samples primal co-ordinates conditioned on dual co-ordinates. Each step of the sampler is equivalent to a Gibbs update in the k-cut mixed co-ordinate system, with k incremented after each update. When F is the Euclidean Bregman generator $\frac{1}{2}q^Tq$, Orthogonal Gibbs reduces to the usual deterministic-scan Gibbs sampler. Secondly, the Orthogonal Gradient algorithm samples along an e-geodesic in the direction of the gradient $r F$ at the current sample, and then samples on the m-at submanifold orthogonal to the e-geodesic.

For a multivariate Gaussian target distribution with Bregman generator $F = \log p$, the induced metric r^2F is constant. In this case both of the above algorithms reduce to sampling analogues of conjugate directions solvers for symmetric positive-de nite linear systems. I will show in this chapter and Appendix D that Orthogonal Gibbs and Orthogonal Gradient respectively become identical to the Gaussian Elimination (GE) and Conjugate Gradient (CG) solvers with added stochastic terms. These results re-ect previous work by other researchers on the relationship between sampling and optimization algorithms, in particular by Fox and Parker [30], [31], who have developed similar, though not identical, conjugate directions sampling algorithms [59], [29]. I will also prove that the Orthogonal Gibbs sampler produces independent samples for a multivariate Gaussian target distribution with the p-Bregman generator.

I will derive both algorithms from a general procedure for sampling in generalized

mixed co-ordinates. Theorem 5.1.6 will show that this general algorithm preserves the target distribution, and hence so do the Orthogonal Gibbs and Orthogonal Gradient algorithms.

Finally, I will present numerical examples with Gaussian and non-Gaussian targets for which both algorithms appear to converge to the correct distribution, and for which the samples appear to be near-independent. The novel samplers are compared to deterministic-scan Gibbs and the No U-Turn Sampler (NUTS).

5.1 General Recursive Algorithm

We now introduce a class of sampling algorithms which we refer to as Dual Complement Submanifold Samplers. Suppose as usual that $p(q)$ is a probability density supported on a convex subset of $\mathbb{R}^p$, and D_F is a Bregman divergence on $\mathbb{R}^p$. We stress that D_F need not be the p-Bregman divergence, although we will see that this geometry is often the best choice when p is log-concave. The general algorithm and all results stated in this section are novel contributions to the literature.

As described in the introduction of this chapter, the main idea is to first sample the target distribution conditioned onto an e-flat submanifold, and then on a dual-complementary m-flat submanifold, as illustrated in figure 5.1.1. As we discussed in Section 4.2, the notion of conditioning on a submanifold is made precise by considering a disintegration of p over a fiber bundle for which the particular submanifold is a fiber. Since the generalized mixed co-ordinate systems of Section 3.3.3 induce a fiber bundles given by foliations, the overall scheme we suggest can be viewed as a Metropolis-within-Gibbs sampler in which primal co-ordinates are resampled conditionally on complementary dual co-ordinates and vice versa.

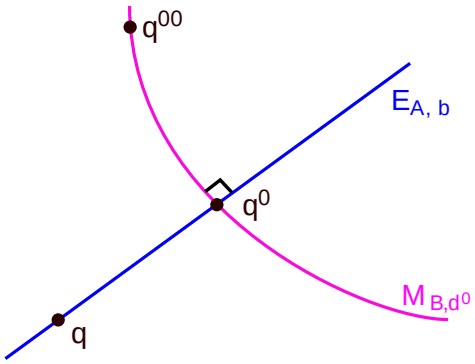

Figure 5.1.1: Sampling on complementary dual submanifolds. Starting from an initial point q, the component within the e-flat submanifold $E_{A,b}$ is resampled to obtain a new point q^0. The algorithm then resamples on the dual complement m-flat submanifold M_{B,d^0}, which intersects $E_{A,b}$ orthogonally with respect to the Riemannian metric. This yields a new sample q^{00}.

5.1.1 Overview of General Algorithm

Suppose q is the current iterate of the algorithm in primal co-ordinates. We generate $l < p$ linearly independent vectors in $\mathbb{R}^p$, which can depend on q in general, and arrange the vectors into the columns of a $p \times l$ matrix A. The choice of A at each step determines the specific algorithm within the class of dual complement submanifold samplers; two choices for A we will study are the standard basis vectors and the gradient of the generator at q, $\nabla F(q)$. In all cases we always assume l is much smaller than p. In primal co-ordinates, the affine subspace through q spanned by the columns of A represents an l-dimensional coordinate submanifold $E_{A,b}$ where b is the orthogonal projection, with respect to the Euclidean inner-product, of q onto the kernel or nullspace of A. The submanifold therefore has the affine parameterization $q' = Aa' + b$ in primal co-ordinates where a' ranges over a convex subset of $\mathbb{R}^l$.

The algorithm first samples p conditioned on $E_{A,b}$. In Theorem 5.1.1, we will give an explicit disintegration of p over the general coordinate-foliation defined by A, and show that we may sample from elements of the disintegration by sampling $a^{(1)}$ from the l-dimensional probability density

$$p^A(a') \propto p(Aa' + b) \tag{5.1.1}$$

with respect to the l-dimensional Lebesgue measure $\lambda^l(da')$, and then embedding the sample into $\mathbb{R}^p$ via $q' = Aa^{(1)} + b$. Then q' is a sample from the conditional distribution $p|_{E_{A,b}}$, a probability measure on $\mathbb{R}^p$ which is concentrated on the submanifold $E_{A,b}$. This step requires us to draw from the l-dimensional density (5.1.1) using some other sampling algorithm which we refer to as the subsampler. In principle any MCMC algorithm may be used as the subsampler; we offer some suggestions in Section 5.1.4. Since l is assumed to be very small, e.g. $l = 1$, (5.1.1) can be sampled by a black-box or generic subsampling algorithm that requires no hand-tuning and can produce independent or near-independent samples after some thinning. This is essentially a Metropolis-within-Gibbs update to resample the primal component a' of the generalized mixed co-ordinates.

The algorithm then samples the target conditioned on the dual complement submanifold to $E_{A,b}$ through q'. Recall that this is the unique $(p-l)$-dimensional maffine submanifold through q' orthogonal to $E_{A,b}$ with respect to the Riemannian metric induced by D_F. Given a matrix $B \in \mathbb{R}^{p \times (p-l)}$ whose columns form a basis for the Euclidean-orthogonal complement $(\operatorname{Im} A)^\perp$, i.e. such that $A^T B = 0$, the results in Section 3.3.3 show that the dual complement submanifold is $M_{B,d'}$. This is the set of points whose dual co-ordinates are the plane $Bg' + d'$ where g' ranges over a convex subset of $\mathbb{R}^{p-l}$, and $d' = Ad' = AA^\dagger h'$ is the Euclidean-orthogonal projection of the dual co-ordinates $h' = \nabla F(q')$ of q' onto $(\operatorname{Im} B)^\perp = \operatorname{Im} A$. Recall that there exists a parameterization (3.3.20) of $M_{B,d'}$ in primal co-ordinates, called its primal embedding $L_{d'}(b)$ where $b \in \mathbb{R}^{p-l}$. In Theorem 5.1.2, we will give an explicit disintegration of p over the general m-foliation, for which the element concentrated on $M_{B,d'}$ can be sampled

by sampling a density on b,

$$p_{d'}(b) \propto \frac{p(L_{d'}(b))}{\det(A\ G(L_{d'}(b))\ A)}$$

and then embedding into R^p by evaluating $q'' := L_d(b'')$ at the sample $b'' \sim p_{d'}$. This yields a new sample q'' from the element of the disintegration concentrated on $M_{B,d'}$. The new point q'' then forms the next iterate of the chain. In 5.1.3, we will discuss how to explicitly evaluate the primal embedding by solving an l-dimensional convex optimization problem.

In order to sample from $p_{d'}$, we might consider using a black-box subsampler again. However, since l is much smaller than p, the $(p - l)$-dimensional probability distribution $p_{d'}$ is likely to be just as difficult to sample as the full target distribution p. Instead we suggest using recursion; apply the same overall algorithm to sample $p_{d'}$. By Theorem 2.5.5, $M_{B,d'}$ has a dually-flat geometry, and Theorems 3.3.11 and 3.3.12 respectively state that b are primally affine co-ordinates in this geometry with Bregman generator

$$\phi_{d'}(b) := F(L_{d'}(b)) - d'^{T} L_{d'}(b)$$

and $p_{d'}$ is supported on a convex subset of R^{p-l}, namely $B - Q$. Hence all the assumptions of the algorithm listed at the beginning of this section are satisfied in its recursive application to sample $p_{d'}$ using the inherited Bregman divergence D_ϕ. In its first recursive call, the algorithm samples on an l-dimensional submanifold within $M_{B,d}$, which is e-flat with respect to the intrinsic dually-flat geometry of $M_{B,d'}$, but in general is not e-flat in the ambient geometry of R^p equipped with D_F. The algorithm then recurses again on the dual complement submanifold $M_{B_2,d_2'}$, which is m-flat when viewed either as a submanifold of $M_{B,d'}$ or Q. The algorithm therefore samples on a system of submanifolds embedded within each other, as illustrated as a tree structure in figure 5.1.2.

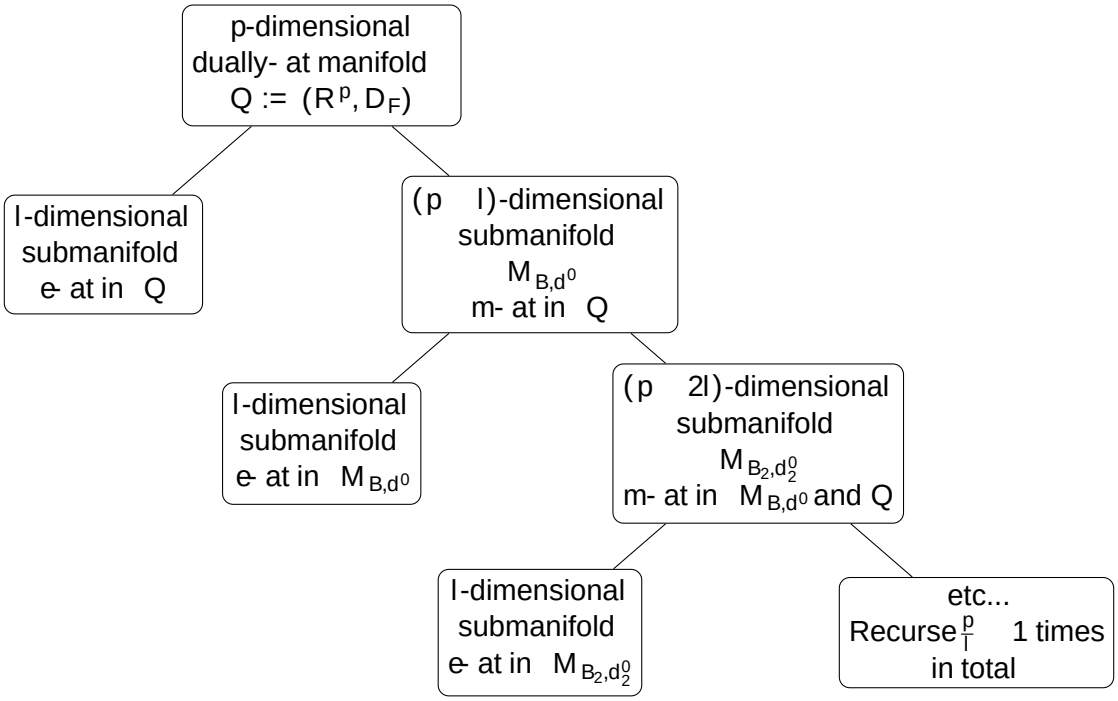

Figure 5.1.2: The system of embedded submanifolds used for conditional sampling in a dual complement submanifold sampler.

We assume l divides p so that after $p/l - 1$ recursive function calls, the target distribution is l-dimensional. This is a base case for the recursion, in which the subsampler is called to sample directly from this low-dimensional probability density. The result is then returned up the chain of recursive calls, and eventually embedded into the original space R^p to yield q''.

After stating the relevant disintegration theorems in Section 5.1.2, we will discuss how to evaluate the primal embedding $L_{d'}(b)$ in Section 5.1.3. Finally, in Section 5.1.4 we will state the recursive algorithm in a general form and prove that it preserves the target distribution p.

5.1.2 Disintegration over e- and m-foliations

The dual complement submanifold algorithm requires us to sample from p conditioned on e- and m- at submanifolds. This is made precise by the concept of a disintegration over a ber bundle discussed in Section 4.2. Here we state explicit disintegrations of a target density over e- and m- at submanifolds, which are novel results.

Consider the rst step of the algorithm, in which p is resampled on an e- at submanifold. For $l < p$, let $A \in R^{p \times l}$ have full rank and $\{E_{A,b}\}$ be an e-foliation of Q, where each $E_{A,b}$ is a l-dimensional e- at submanifold spanned by the columns of A in primal co-ordinates, where b ranges over $(\mathrm{Im}\, A)^{?}$. As discussed in 3.3.3, thee-foliation

is a ber bundle with projection map $F_E(q) := B\,q$ where $B \in R^{p\ (p\ l)}$ with $A^T B = 0$, and $B := B^T B\ ^1 B^T$ is the generalized inverse of B. Furthermore, we have the decomposition $q = Aa + Bb$ with $b := Bb$ so that each $E_{A,b}$ corresponds to xing b and allowing a to vary. Each b selects a different ber, i.e. level set of F_E. Then we have the following theorem, proved in Appendix B.

Theorem 5.1.1 (Disintegration over e-Foliations). Let $p(q)$ be a Lipschitz-continuous probability density on a convex open subset $Q \subset R^p$, and let $A \in R^{p\ l}$ and $B \in R^{p\ (p\ l)}$ be full-rank matrices with $A^T B = 0_{l\ (p\ l)}$. For $b \in B\,Q$, de ne the measure m_b on Q such that for a Borel set $U \subset R^p$,

$$m_b(U) := \int_{U \setminus F_E^{-1}(\{b\})} p(q)\ \frac{q}{\det(B^T B)}\, H^l(dq) \tag{5.1.2}$$

$$= \int_{A\,(U \setminus F_E^{-1}(\{b\}))} p(Aa + Bb)\ \det\ A\quad B\ \ ^{l}\, l^{l}(da) \tag{5.1.3}$$

where H^l is the l-dimensional Hausdorff measure on R^p and l^l is the l-dimensional Lebesgue measure on R^l. Then the measures $\{m_b\}$ form a disintegration of the target measure $p(q)\,l^p(dq)$ over the ber bundle $F_E : Q \to R^{p\ l}$, $F_E(q) = B\,q$.

The upshot of Theorem 5.1.1 is that for each b we can sample from m_b, which can be interpreted as a conditioning of p on the ber $F^{-1}\{b\}$, by sampling the a component from a density proportional to $p(Aa + Bb)$.

For the same matrices A and B, we also have the m-foliation $\{M_{B,d}\}$, a set of m- at submanifolds equivalent to the bers of the projection map $F_M(q) := A\,r\,F(q)$. For each $d \in R^l$, the ber $M_{B,d} = F^{-1}(\{d\})$ is an m- at submanifold, where $d = Ad$. The following theorem gives a disintegration over the m-foliation.

Theorem 5.1.2 (Disintegration over m-Foliations). Let $p(q)$ be a Lipschitz-continuous probability density on an open convex subset $Q \subset R^p$, A and B be matrices as in Theorem 5.1.1, and $X = r\,F(Q)$. For $d \in A\,X$, de ne the probability measure m_d on Q such that for a Borel $U \subset R^p$,

$$m_d(U) = \int_{U \setminus F_M^{-1}(\{d\})} p(q)\, r\frac{\det\ A^T A}{\det\ A^T G(q)^2 A}\, H^{p\ l}(dq) \tag{5.1.4}$$

where $H^{p\ l}$ is the $(p\ l)$-dimensional Hausdorff measure on R^p, $G(q) := r^2 F(q)$ and $F_M : R^p \to R^l$, $F_M(q) = A\,r\,F(q)$ for a smooth convex function $F : R^p \to R$. Then the measures $\{m_d\}$ form a disintegration of the target measure $p(q)\,l^p(dq)$ over the bers of F_M. Furthermore, we have

$$m_d(U) = \int_{B\,(U \setminus F_M^{-1}(\{d\}))} p(L_d(b))\ \frac{\det\ A\quad B}{\det(A\ G(L(b))\ A)}\ l^{p\ k}(db) \tag{5.1.5}$$

where $l^{p\ k}$ is the $(p\ k)$-dimensional Lebesgue measure on $R^{p\ k}$ and L_d is the primal embedding of $F_M^{-1}(\{d\})$.

The proof is also given in Appendix B. Put simply, we can sample from m_d by sampling b from the density

$$p_d(b) \mu \frac{p(L_d(b))}{\det(A^T G(L_d(b)) A)} \tag{5.1.6}$$

and then embedding into R^p with the primal embedding $L_d(b)$.

5.1.3 Numerical Evaluation of Primal Embedding

Before stating the recursive algorithm explicitly, we must discuss how to evaluate the primal embedding $L_d(b)$ of an m- at submanifold. Recall that for matrices $A 2 R^{p l}$ and $B 2 R^{p (p l)}$ both of full rank with $A^T B = 0$, we decompose the primal and dual co-ordinates of a point $q 2 Q$ as

$$q = Aa + Bb \qquad h = r F(q) = Ad + Bg$$

We consider holding d xed and allowing g to vary; g is then a co-ordinate chart for the $(p l)$-dimensional m- at submanifold $M_{B,Ad}$. In primal co-ordinates q, this submanifold is a hypersurface embedded in R^p that is curved in general. We seek a parameterization $q = L_d(b)$ of this hypersurface. The Implicit Function Theorem B.0.4 guarantees the local existence of a function $a_d(b)$ which gives the a component such that $q = Aa_d(b) + Bb =: L_d(b)$.

Let us x b and consider how to compute $a_d(b)$. Note that differentiating $F(Aa + Bb)$ with respect to a yields

$$\frac{\P}{\P a^i} F(Aa + Bb) = \sum_s \frac{\P q^s}{\P a^i} \frac{\P F}{\P q^s} = \sum_s A_i^s h^s$$
$$) r_a F(Aa + Bb) = A^T h = A^T Ad$$
$$) r_a \left[F(Aa + Bb) \quad a^T A^T Ad \right]_{a = a_d(b)} = 0$$

This is a system of l equations in l variables for $a_d(b)$. The equations are equivalent to the Legendre transform for the restricted convex function $a 7! F(Aa + Bb)$, where the corresponding restricted dual variable is $A^T Ad$. It follows that a_d is the solution of the unconstrained convex optimization problem

$$a_d = \operatorname{argmin}_a \underbrace{F(Aa + Bb) \quad a^T A^T Ad}_{:= j(a)} \tag{5.1.7}$$

This agrees with the expression for the primal embedding given in Proposition 3.3.9. The gradient and Hessian of the objective function are respectively given by

$$r j(a) = A^T r F(Aa + Bb) \quad A^T Ad \tag{5.1.8}$$
$$r^2 j(a) = A^T r^2 F(Aa + Bb) A \tag{5.1.9}$$

Note that (5.1.9) is proportional to the denominator in the integrand of (5.1.5) for the same a and b. Thus, since computing the conditional density on an m- at submanifold requires evaluating the Hessian term (5.1.9) anyway, one may as well use Newton's method to solve the optimization problem (5.1.7) because the asymptotic time complexity of the overall sampler is unchanged; the cost of evaluating a log-determinant and solving a linear system of the same size both scale cubically with the dimension of the matrix. Since l is assumed to be small, solving the $l \quad l$ linear system at each step of Newton's method has insigni cant cost. The entire method for evaluating $L_d(b)$ is listed as algorithm 2.

Algorithm 2 Evaluating the primal embedding $L_d(b)$ of $M_{B,Ad}$

```
function MFLATEMBED(b, F, d, A, B, a_0)
    b   Bb                          . a_0 here is the initial guess for Newton's method
    a_d   NEWTON(a 7! F(Aa + b)   a^T A^T Ad, a_0)
    return Aa_d + b
end function
```

In De nition 3.3.13, we de ned a map $L_d : R^p \; ! \; R^p, L_d(q) = L_d \; B \; q$. Algorithm 3 is a slight modi cation of 2 which allows evaluation of L_d.

Algorithm 3 Evaluating the parameterization $L_d(q)$ of $M_{B,Ad}$

```
function MFLATEMBED2(b, F, d, A, a_0)     . a_0 here is the initial guess for Newton's
method
    a_d   NEWTON(a 7! F(Aa + b)   a^T A^T Ad, a_0)
    return Aa_d + b
end function
```

5.1.4 General Dual Complement Submanifold Algorithm

Finally we can state the recursive dual complement submanifold algorithm, which is listed as Algorithm 4. The algorithm is implemented as a function RECDCSSAMPLER, which takes three arguments:

$q \; 2 \; R^p$, the initial iterate of the algorithm.

$\log p : R^p \; ! \; R$, a function returning the target log probability density up to additive constants.

$F : R^p \; ! \; R$, a smooth convex Bregman generator.

The function recursively calls itself on line 22, and so we must allow for the dimension of each of these inputs to vary. On line 2, the algorithm calls the following function.

Algorithm 4 General Recursive Dual Complement Submanifold Algorithm

```
function RecDCSSampler(q, F, log p)
    A ← BasisVectors(q, F)
    a ← A^T q
    b ← q − A a^(0)
    a' ← SubSampler(a ↦ log p(Aa + b), a)
    q' ← A a' + b
    if Dimension(q) == 1 then
        q'' ← q'
        return q''
    end if
    d' ← A^T r F(q')
    B ← ComplementBasis(A)
    function F#(b)
        q ← MFlatEmbed(b, F, d, A, B, a')
        return F(q) − d'^T A^T q
    end function
    function log p_{d'}(b)
        q ← MFlatEmbed(b, F, d, A, B, a')
        return log p(q) − log det A^T r ∇²F(q) A
    end function
    b' ← B^T q'
    b'' ← RecDCSSampler(b', F#, log p_{d'})
    q'' ← MFlatEmbed(b'', F, d, A, B, a')
    return q''
end function
```

De nition 5.1.3. For $q \in R^k$ and a convex Bregman generator $F : R^k \to R$ where $k \geq l$, the function BASISVECTORS(q, F) returns some $k \times l$ matrix A of rank l.

The varying size k of the inputs allows for the BASISVECTORS function to be called within recursive calls of the algorithm. For example, the first time BASISVECTORS is called with arguments q and F, and thus $k = p$. Within the first recursive call to sample $b \in R^{p-l}$, the BASISVECTORS function is called with $k = p - l$, then $k = p - 2l$ in the next recursive call and so on, each time returning a full-rank $k \times l$ matrix A. The choice of which matrix the BASISVECTORS returns determines the exact algorithm; choosing standard basis vectors leads to the orthogonal Gibbs algorithm which we will discuss in Section 5.2, and choosing ∇F gives the orthogonal gradient algorithm to be discussed in Section 5.3.

On lines 3 and 4, the initial iterate is decomposed into the form $Aa + b$. As stated in Section 5.1.1, the a component of this is to be resampled, which requires some other MCMC algorithm called the subsampler. A subsampler is any function satisfying the following.

De nition 5.1.4. Let $\log p : R^l \to R$ be a log probability density function up to an additive constant. Then a subsampler, written SUBSAMPLER$(\log p, \cdot)$, is a Markov kernel which preserves p, i.e. if $a \sim p$ then the result $a' \sim$ SUBSAMPLER$(\log p, a)$ is also distributed according to p.

Thus lines 5 and 6 use the subsampler to resample the a component of q from the relevant density (5.1.1), yielding an updated sample q'. Note that the optimal choice of subsampler will undoubtedly depend on the overall target distribution p, but we can offer some generic suggestions. After several nested recursive calls, the target p is defined by repeated applications of (5.1.6). Thus subsamplers that rely on derivative information of the log-density such as Hamiltonian Monte Carlo or MALA will require expensive derivatives of log-determinants of the Riemannian metric, and so we suggest using derivative-free subsamplers. Furthermore, in general (5.1.6) is not known to be log-concave in b, and there is little we can easily infer about its scale to tune any parameters of the subsampler such as a step-size. For $l = 1$, p is a univariate density, for which a reasonable choice might be the Adaptive Rejection Metropolis Sampler (ARMS) of Gilks [38], which can be derivative-free and does not require log-concavity of the target or hand tuning of parameters. The slice sampler [56] also satisfies these properties. For $k > 1$, the T-walk of Christen and Fox [20] is a "black-box" option that may offer reasonable performance.

If the input q has dimension l, then the base case of the recursion has been reached and caught by the if-statement on line 7. The new sample $q'' = q'$ is returned. If the dimension is larger than l, then the algorithm recurses on the dual complement to $E_{A,b}$. This requires a choice of basis for the dual complement, given by the following function.

De nition 5.1.5. For $A \in R^{k \times l}$, COMPLEMENTBASIS(A) returns a $k \times (k - l)$ matrix B which has Euclidean-orthogonal columns, has rank $k - l$, and satisfies $A^T B = 0_{l \times (k-l)}$.

For any set of vectors, we can always find a Euclidean-orthogonal set spanning the same space, and thus the condition of Euclidean-orthogonal columns is not restrictive. The complement basis B initialised on line 12 defines the primal co-ordinates b for the dual complement, in which the log target p_{d^0} and Bregman generator F_{d^0} of the dual complement are defined.

Since the aim is to resample from m_{d^0} from Theorem 5.1.2, the algorithm defines a new target log density function on line 17 which evaluates p_{d^0} given in (5.1.6). This requires evaluating the primal embedding L_{d^0} with a call to MF LAT EMBED. Similarly, the inherited Bregman generator for the dual complement is defined on line 13, using the result from Theorem 3.3.12 also evaluates the primal embedding. We assume that the gradient and Hessian, i.e. the inherited dual co-ordinates and metric, are evaluated using their expressions given in the Theorem.

Finally, the algorithm calls itself recursively, passing the b co-ordinate of q^0 as the initial iterate, the inherited Bregman generator and log p_{d^0} as the target log density. This returns a new sample b^{00}, which is embedded into $\mathbb{R}^p$ on line 23 as q^{00}, the final result of the algorithm.

The following novel theorem shows that Algorithm 4 preserves the target distribution.

Theorem 5.1.6. Let F and p respectively be a Bregman generator and a probability density on $\mathbb{R}^p$. Suppose l divides p and the BASIS VECTORS and COMPLEMENT BASIS functions are well-defined for inputs of dimension $k = l, 2l, .., p$. If $q \sim p$, algorithm 4 applied to p returns $q^{00} \sim p$.

Proof. We show the result using induction on the dimension p of the target distribution and Bregman generator. Since l divides p, we write $p = nl$ for some $1 \leq n \leq p$ and argue by induction on n.

Base case. Suppose the dimensions of the target distribution and initial iterate q are both $p = l$, i.e. $n = 1$. Then the matrix A on line 2 is a square $l \times l$ matrix of full rank. Thus $a = A^{-1}q$ and so $b = 0$ on line 4. The subsampler on line 5 therefore has a target density argument defined by $a \mapsto p(Aa)$, i.e. the pullback $p^{A^{-1}}$ of p by the linear map A or equivalently its pushforward by A^{-1}. Since $q \sim p$, then $a^{(0)} = A^{-1}q$ is distributed according to the same distribution $p^{A^{-1}}$. By definition, the subsampler preserves this distribution, implying $a^0 \sim p^{A^{-1}}$ and hence in line 6 $q^0 = Aa$ is distributed according to p. The if statement on line 7 is triggered and the algorithm returns $q^{00} = q^0 \sim p$.

Induction step. Suppose that the statement is true when the dimension of the target distribution, and hence the dimension of q, is $p = ln_0$ for some $n_0 \geq 1$. We refer to this assumption as the inductive hypothesis.

Consider the algorithm with input target distribution p and initial iterate q both of dimension $l(n_0 + 1)$. By Theorem 5.1.1, in particular (5.1.3), lines 5 to 6 must preserve m_b, the element of the disintegration over the e-foliation concentrated on the fiber containing q, i.e. such that $b = F_E(q)$. By Theorem 4.2.4, the same lines must therefore preserve the overall target distribution p, and so $q^0 \sim p$.

Note that A is a $l(n_0 + 1)$ l matrix, and thus the dimension of the density p_{d^0} defined in line 17 is $l(n_0 + 1)$ $l = ln_0$. Thus the recursive call of the function in line 22 must preserve p_{d^0} by the inductive hypothesis. Thus by Theorem 5.1.2, lines 22-23 define a random update $q^0 7! q^{00}$ which preserves m_{d^0}, the measure in the disintegration over the m-foliation concentrated on an m- flat submanifold containing q^0. Hence Theorem 4.2.4 implies the update also preserves the overall target distribution p, and so q^{00} p. □

Iterative Dual Complement Submanifold Algorithms

If implemented as-is, Algorithm 4 is severely computationally inefficient. The call to MFLATEMBED at each level of recursion requires the solution of progressively nested optimization problems, since $F^{\#}$ itself is evaluated by a call to MFLATEMBED with different arguments. Furthermore, the algorithm is linear recursive[1], meaning that its execution requires temporary storage of data proportional to p at each level of recursion; for example, the representation of inherited Bregman generators and conditional target distributions, and the variables d^0 and A must all be kept in memory at each recursive call. For large enough p, too many instructions will be pushed to the computer's call stack and cause a stack overflow. Clearly a different implementation is required. For each of the Orthogonal Gibbs and Orthogonal Gradient algorithms, we will state a more efficient iterative implementation derived in Appendix C which is equivalent to the recursive implementation, i.e. produces the same samples given the same input arguments and pseudorandom seed.

5.2 Orthogonal Gibbs Sampling

We now introduce a special case of the general algorithm called Orthogonal Gibbs. Suppose we choose the BASISVECTORS function to return the first l columns of the p p identity matrix, where p is dimension of the input q to the BASISVECTORS function;

$$A = \text{BASISVECTORS}(q, F) = 1_{p\ l} = \begin{pmatrix} 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \\ & & & \ddots & \\ 0 & 0 & 0 & & \\ \vdots & \vdots & \vdots & & \end{pmatrix} = \begin{matrix} 1_{l\ l} \\ 0_{(p\ l)\ l} \end{matrix} \qquad (5.2.1)$$

(l columns)

and choose the complement basis to be the remaining p l standard basis vectors of dimension p,

$$B = \text{COMPLEMENTBASIS}(A) = \begin{matrix} 0_{l\ (p\ l)} \\ 1_{(p\ l)\ (p\ l)} \end{matrix} \qquad (5.2.2)$$

Since $q = Aa + Bb$, as we remarked in Section 3.3.3, we recover the standard mixed co-ordinates of Section 3.3.1. In particular,

$$a = q_{\le l} := (q^1, .., q^l) \qquad b = q_{>l} := (q^{l+1}, ..., q^p)$$
$$d = h_{\le l} := (h^1, .., h^l) \qquad g = h_{>l} := (h^{l+1}, ..., h^p)$$

Let us reinterpret algorithm 4 with this choice of A and B, for which $A = A^T$ and $B = B^T$. First we sample $a = q_{\le l}$ proportionally from the target density, with $b^0 = q^{(0)}_{>l}$ fixed. This is equivalent to a Gibbs update for the block $q_{\le l}$, sampling conditioned on the current value of the complementary block $q_{>l}$. We then compute the first l dual co-ordinates for the updated state,

$$d^0 = h^{(1)}_{\le l} = A^T \nabla F\left(\begin{pmatrix} q^{(1)}_{\le l} \\ q^{(0)}_{>l} \end{pmatrix}\right) = \sum_{j=1}^{} \frac{\partial F}{\partial q^j}\left(\begin{pmatrix} q^{(1)}_{\le l} \\ q^{(0)}_{>l} \end{pmatrix}\right) A_{j\le l}$$

This dual block is distributed according to the conditional distribution of $h_{\le l}$ given the primal co-ordinates $q^{(0)}_{>l}$. Thus, the first step of the algorithm is equivalent to a block-Gibbs step in the mixed co-ordinate system.

The second step of the algorithm is the recursion; we call the algorithm again with $p \to p - l$, and inherited Bregman generator and target distribution respectively given by

$$F(q_{>l}) := F(q_{>l}) - h_{\le l}{}^T A^T L_{h_{\le l}}(q_{>l}) \tag{5.2.3}$$

$$p(q_{>l}) \propto \frac{p - L_{h_{\le l}}(q_{>l})}{\det G_{[\le l,\, \le l]}\, L_{h_{\le l}}(q_{>l})} \tag{5.2.4}$$

where $G_{[\le l\ \le l]}(q) = A^T \nabla^2 F(q) A$ is the upper-left $l \times l$ block of the metric, and $L_{h_{\le l}}(q_{>l})$ is the co-ordinate embedding of the m- flat manifold M_k described in Section 5.1.3. This recursive call of the algorithm returns a sample from the conditional distribution of $b = q_{>l}$ given $h^{(1)}_{\le l}$; it is the complementary Gibbs update, sampling the primal $q_{>l}$ block given the latest values of the dual components $h_{\le l}$. In conclusion, the algorithm with this choice of A and B is simply deterministic-scan block Gibbs in mixed co-ordinates, working with blocks $h_{\le l}$ and $q_{>l}$, using recursion to sample one of the conditionals. For this reason, we refer to this algorithm as "Orthogonal Gibbs", where "orthogonal" refers to the fact that the algorithm moves on Riemannian-orthogonal submanifolds.

The recursive form of the algorithm is inefficient for most target distributions. In Appendix C, we derive an equivalent iterative version of the Orthogonal Gibbs sampler, listed as Algorithm 5. The algorithm iterates over $n := p/l$ blocks of components of q, each of size l. At the j^{th} iteration, the j^{th} block of q is resampled conditioned on the $1, .., j-1$ blocks represented in dual co-ordinates, and the $j+1, .., n$ blocks in primal co-ordinates. This conditional distribution has a target log-density computed in the function SUBTARGETGIBBS.

Algorithm 5 Orthogonal Gibbs Sampler

```
function SUBTARGETGIBBS(z, q, h, j, F, log p)
    a^0 ← NEWTON(a ↦ F*(a; z, q_{[jl+1:p]}), a^T h, a_0 := q_{[1:(j-1)l]})
    q^0 ← ∇F*(a^0; z, q_{[jl+1:p]})
    return log p(q^0) − log det G_{[(j-1)l, (j-1)l]}(q^0)
end function

function ORTHOGONAL GIBBS(q^(0), F, log p)
    p ← DIMENSION(q^(0))
    n ← p/l                          . Calculate number of blocks. We assume k divides p.
    z^(1) ← SUBSAMPLER(z ↦ log p(z, q^(0)_{[(l+1):p]}), q^(0)_{[1:l]})
    q^(1) ← ∇F*(z^(1); q^(0)_{[(l+1):p]})
    h^(1) ← (∂F/∂q^i(q^(1)))_{i=1}^{l}

    for j = 2, .., n do
        z^(j) ← SUBSAMPLER(z ↦ SUBTARGETGIBBS(z, q^(j-1), h^(j-1), j, F, log p),
                           q^(j-1)_{[(j-1)l+1:jl]})
        a^(j) ← NEWTON(a ↦ F*(a; z, q^(j-1)_{[jl+1:p]}), a^T h^(j-1), a_0 := q^(j-1)_{[1:(j-1)l]})
        q^(j) ← ∇F*(a^(j); z^(j); q^(0)_{[jl+1:p]})
        h^(j) ← h^(j-1); (∂F/∂q^i(q^(j-1)))_{i=(j-1)l+1}^{jl}
    end for
    return q^(n)
end function
```

Figure 5.2.1 shows the iterates of two runs of the sampler for a bivariate target distribution p using the posterior Bregman geometry $F(q) := -\log p(q)$. The target distribution is the conjugate prior on the natural parameters of a univariate normal distribution. Blue lines show the e- at submanifold, or e-geodesic, sampled on when conditioning on primal variables. Red lines show m- at submanifolds or m-geodesics sampled when conditioning on dual variables. Note that the m-geodesics are nearly straight, suggesting that the Bregman generator is nearly quadratic, i.e. the distribution is close to a multivariate Gaussian.

The following novel proposition further justi es the name Orthogonal Gibbs; when the Euclidean geometry is used, the algorithm reduces to the usual Gibbs sampler.

Proposition 5.2.1. For the Euclidean Bregman generator $(q) := \frac{1}{2}q^T q$, the Orthogonal Gibbs sampler reduces to the block deterministic-scan Metropolis-within-Gibbs algorithm, where the subsampler is used to sample the full conditionals.

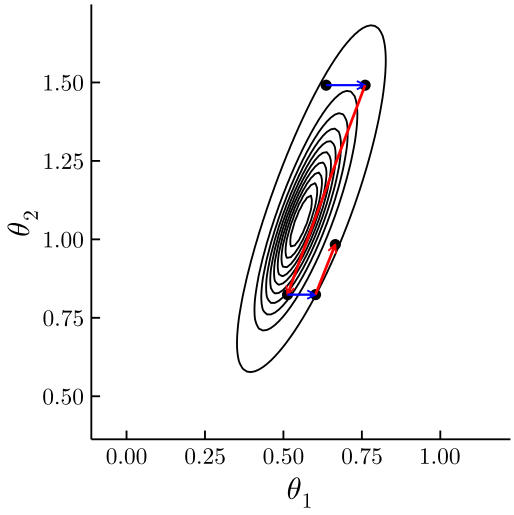

Figure 5.2.1: Orthogonal Gibbs sampling from a bivariate target distribution.

Proof. Note that for the Euclidean Bregman generator, r F(q) = q. Thus $q^{(j)}_{[1:jl]} = h^{(j)}$ for all j = 1, .., p, and the optimization problem solved on lines 2 and 16 simply reduces to

$$a^{(j)} = q^{(j)}_{[1:(j\ \ 1)l]}$$

Furthermore, since the metric in this case is simply the identity G $1_{p\ \ p}$, the log-determinant on line 4 is zero. Thus the SUBTARGETGIBBS function call on line 15 returns

$$\log p \quad q^{(j\ \ 1)}_{[1:(j\ \ 1)l]}, z, q^{(j\ \ 1)}_{[jl+1:p]}$$

This is the logarithm of the conditional density of the j^{th} block given all the complementary set of variables up to an additive constant independent of z, since

$$p \quad q_{j(l\ \ 1)+1:jl} = z \; j \; q_{[1:(j\ \ 1)l]} = q^{(j\ \ 1)}_{[1:(j\ \ 1)l]}, q_{[jl+1:p]} = q^{(j\ \ 1)}_{[jl+1:p]}$$
$$\mu p \quad q^{(j\ \ 1)}_{1:(j\ \ 1)l}, z, q^{(j\ \ 1)}_{[jl+1:p]}$$

Therefore the subsampler samples from this conditional density, and hence the algorithm is equivalent to the deterministic-scan Metropolis-within-Gibbs sampler. □

5.2.1 Computational Cost

Let us consider the computational cost of the Orthogonal Gibbs algorithm, i.e. Algorithm 5. At the j^{th} iteration for j 2, the algorithm requires the following computation:

Evaluation of the SUBTARGETGIBBS function. This is called a number s_j of times by the sub-sampler, and each call involves a Newton iteration for a l(j 1)-dimensional convex optimization problem, a log-determinant of a l(j 1) l(j

Operation	Size at iteration j	Number of evaluations at iteration j
F		$s_j d_j$
$\log p$		s_j
∇F	$(j-1)l$ components	$s_j d_j$
∇F	l components	1
$\nabla^2 F$	$(j-1)l \times (j-1)l$ sub matrix	$s_j d_j$
Newton inner-solve	$(j-1)l \times (j-1)l$ linear system	$s_j d_j$
$\log\det$	$(j-1)l \times (j-1)l$ matrix	s_j

Table 5.2.1: Breakdown of computational operations required for the Orthogonal Gibbs algorithm. Here s_j is the number of times SUBSAMPLER calls SUBTARGETGIBBS at iteration j, and d_j is the maximum number of Newton steps at iteration j.

1) sub-matrix of the Hessian of F, and at least one evaluation of the target log-density.

Another Newton solve on line 16. However, this Newton solve has already been performed within SUBTARGETGIBBS, and can be cached to improve performance.

On line 18, the j^{th} block of dual components is computed, which are l components of the gradient vector.

A breakdown of these computations is shown in Table 5.2.1. Assuming the Hessian matrix $\nabla^2 F$ is dense at each iteration, the most expensive operations for large $p = nl$ are the log-determinant and the linear solve at each Newton step, both of which scale cubically with the dimension of their inputs. Suppose then that the cost of the j^{th} iteration of Algorithm 5 is bounded above by $K(j-1)^3 l^3$ FLOPs for some constant K. Here K depends on s_j and d_j, which we assume are bounded as n increases. Note that for $j = 1$ there are no Newton solves or Hessian evaluations, and the cost of the iteration is dominated by s_1 evaluations of the target log-density, which we assume is negligible compared to later iterations. Then the total number of FLOPs is bounded above by

$$\sum_{j=1}^{n} K(j-1)^3 l^3 = Kl^3 \sum_{j=1}^{n-1} j^3 = \frac{Kl^3}{4} n^2 (n-1)^2$$

Since $n = p/l$, this implies that the asymptotic cost of Algorithm 5 is $O(p^4)$ as $p \to \infty$.

5.2.2 Relationship to Gaussian Elimination

There is a natural link between our suggested algorithms and conjugate directions solvers for linear systems. In particular, when the Bregman generator is a quadratic form so that its Hessian matrix is constant, the notion of orthogonality with respect to the Riemannian metric is clearly equivalent to conjugacy i.e. orthogonality with respect to an

inner-product weighted by a symmetric positive-definite matrix. Updating the sample at each iteration of the general Algorithm 4 is then equivalent to resampling the coefficients in a mutually conjugate basis, analogously to the iterations of a conjugate directions solver.

One can employ a modified Gram-Schmidt process to transform any basis of $\mathbb{R}^p$ into a mutually conjugate basis; see Theorem D.0.2 in Appendix D. In particular, one can transform the standard basis in $\mathbb{R}^p$ into a conjugate basis for which the corresponding conjugate directions solver is the Gaussian Elimination algorithm applied to the matrix used to weight the inner-product [40, 76]. Since the Orthogonal Gibbs sampling algorithm resamples in the direction of the first standard basis vector at each level of recursion, one might therefore suspect a link with Gaussian Elimination. As stated earlier, the relationship between sampling and optimization has been studied by previous authors such as Fox [30], but to our knowledge this is the first time a sampling analogue of Gaussian Elimination has been presented.

Suppose that in Algorithm 5, the target distribution is a multivariate Gaussian of the form

$$q \quad \text{Normal}\left(L^{-1}w, L^{-1}\right) \tag{5.2.5}$$

and suppose we choose the posterior Bregman geometry, so that the Bregman generator is given by

$$F(q) = \frac{1}{2}q^T L\, q \quad q^T w \tag{5.2.6}$$

Note that (5.2.5) is the form of the Bayesian posterior distribution from Section 1.5.1. In Appendix D, we show that under this choice of target distribution and geometry, Algorithm 5 reduces to a Gaussian Elimination Sampler, listed as Algorithm 6. The algorithm produces an independent sample $q^{(p)}$ from the target distribution and the solution m of $Lm = w$. Note that lines 8-15 apply elementary row operations to L, w and V^T; this transforms the linear system $Lm = w$ to its reduced row echelon form, i.e. making it upper-triangular.

The following theorem, proven in Appendix D, shows that the algorithm also produces the Cholesky factor of L as a biproduct.

Theorem 5.2.2. Let L^0 and V result from the application of Algorithm 6 to a symmetric positive-definite matrix $L \in \mathbb{R}^{p \times p}$ and $w \in \mathbb{R}^p$. Let D be the diagonal matrix whose diagonal is the diagonal of L^0. The lower-triangular Cholesky factor L of, i.e. such that $L = LL^T$, is given by

$$L := L^{0\,T} D^{-1/2}$$

In fact, the algorithm produces independent samples from the target distribution after p iterations, as the following corollary shows.

Corollary 5.2.3. The final iterate $q^{(p)}$ of the Gaussian Elimination sampler, Algorithm 6, is stochastically independent of the initial iterate $q^{(0)}$ and is a sample from the target distribution $\text{Normal}\left(L^{-1}w, L^{-1}\right)$.

Algorithm 6 Gaussian Elimination Sampler

$V \leftarrow 1_{p \times p}$. Set V to identity matrix
for $j = 1, .., p$ do
  $v^{(j)} \leftarrow V_j$. Column j of V
  $z_j \sim \text{Normal}(0, 1/L_{jj})$
  $e_j \leftarrow w_j - \sum_{s=1}^{p} L_{js} q_s^{(j-1)} / L_{jj}$
  $q^{(j)} \leftarrow q^{(j-1)} + z_j \, e_j \, v^{(j)}$
  $m^{(j)} \leftarrow m^{(j-1)} + w_j / L_{jj} \, v^{(j)}$
  for $i = j + 1, .., p$ do . If $j = p$, the for loop is skipped
    $C_{ij} \leftarrow L_{ij} / L_{jj}$
    for $s = 1, .., p$ do
      $L_{is} \leftarrow L_{is} - C_{ij} \, L_{js}$
      $V_{si} \leftarrow V_{si} - C_{ij} \, V_{sj}$
    end for
    $w_i \leftarrow w_i - C_{ij} \, w_j$
  end for
end for

5.3 Orthogonal Gradient Sampling

The second specific algorithm we suggest is inspired by the conjugate gradient method in optimizaton [73], [40]. One can imagine sampling conditionally on an e-geodesic whose tangent vector is proportional to the gradient of the Bregman generator at the current sample, and then sampling on the complementary dual submanifold to the geodesic as per Algorithm 4. We therefore suggest

$$A = \text{BASISVECTORS}(q, F) := \nabla F(q) \tag{5.3.1}$$

and B any orthonormal basis for the Euclidean-orthogonal complement to the span of this vector, e.g. the orthonormalized first $k - 1$ linearly-independent columns of

$$1_{k \times k} - \nabla F(q) \nabla F(q)^T \tag{5.3.2}$$

where k is the dimension of q.

Note that in (5.3.1), we take the gradient of the Bregman generator rather than the target log-density. This choice is made to avoid taking gradients of log-determinants of the metric, which arise in the target log-density at nested levels of recursion in Algorithm 4. If we use the posterior Bregman generator, i.e. take $F(q) = -\log p(q)$, then we can include derivative information about the posterior without differentiating log-determinants, since the gradient of the inherited generator F is given in Theorem 3.3.12 as $B^T \nabla F$, i.e. the projection of the gradient onto the Euclidean-orthogonal complement to the span of ∇F.

We call the resulting algorithm Orthogonal Gradient , using the term orthogonal to emphasize the local Riemannian inner-product, rather than conjugate which typically refers to a globally-constant inner-product. In Appendix C, we show that the sampler may be implemented in an iterative form, listed as Algorithm 7. The initial state $q^{(0)}$ must satisfy $r\ F\ q^{(0)} \neq 0$, so that $r^{(1)} \neq 0$.

Algorithm 7 Orthogonal Gradient Sampler

```
function SUBTARGET(a, r, d, A, c, F, log p, a_0)
    b^0 ← ar + c
    q^0 ← MFLATEMBED2(F, b^0, d, A, a_0)
    return log p(q^0) − log det A^T r ∇²F(q^0) A
end function

function ORTHOGONAL GRADIENT(q^(0), F, log p)
    p ← DIMENSION(q^(0))
    r^(1) ← r F q^(0)
    a^(0) ← r^(1)^T q^(0) / r^(1)^T r^(1)
    c^(1) ← q^(0) − a^(0) r^(1)
    a^(1) ← SUBSAMPLER(a ↦ log p ∘ ar^(1) + c^(1), a^(0))
    q^(1) ← a^(1) r^(1) + c^(1)
    b^(1) ← q^(1)
    d_1 ← r^(1)^T r F q^(1) / r^(1)^T r^(1)
    A^(1) ← r^(1)
    for j = 2, .., p do
        r^(j) ← A^(j) d r F q^(j−1)
        c^(j) ← 1_p − p (r^(j) r^(j)^T) / (r^(j)^T r^(j)) c^(j−1)
        a_0^(j) ← A^(j−1) q^(j−1)
        a_0^(j) ← r^(j)^T q^(j−1) / r^(j)^T r^(j)
        a^(j) ← SUBSAMPLER(a ↦ SUBTARGET(a, r^(j), d^(j−1), A^(j−1), c^(j), F, log p, a_0^(j)), a_0^(j))
        b^(j) ← a^(j) r^(j) + c^(j)                    . Update b in the direction of R^(j)
        q^(j) ← MFLATEMBED2(F, b^(j), d^(j−1), A^(j−1), a_0^(j))    . We already did this
    embedding in pfl - can be re-used
        d_j ← r^(j)^T r F q^(j) / r^(j)^T r^(j)
        A^(j) ← A^(j−1) r^(j)                          . Concatenate r^(j) onto right side of A^(j−1)
    end for
    return q^(n)
end function
```

5.3.1 Relationship to Conjugate Gradient Solver

As we remarked in Section 5.2.2, the dual complement submanifold samplers reduce to sampling along conjugate directions under the Euclidean Bregman generator (5.2.6). For the Orthogonal Gradient sampler with this geometry and the multivariate Gaussian target distribution (5.2.5), we show in Appendix D that the Orthogonal Gradient sampler reduces to a Conjugate Gradient (CG) sampler, listed as Algorithm 8.

Recall that each step of the conjugate gradient method minimizes the quadratic cost function (5.2.6) along a conjugate direction produced by a modified Gram-Schmidt process applied to the negative gradients, i.e. residuals, of the cost function at each iterate [40], [73]. Similarly, Algorithm 8 produces conjugate vectors by applying modified Gram-Schmidt to each iterate, and samples along the new conjugate direction at each step. The sampler and solver are illustrated for a two-dimensional example in figure 5.3.1. In particular, if we replace line 8 in Algorithm 8 with

$$\lambda^{(j)} \quad v^{(j)T} g^{(j-1)} / r^{(j)},$$

i.e. removing the randomly sampled variable $z^{(j)}$, then we would recover exactly the Conjugate Gradient solver for the linear system $Lm = w$.

Since the Conjugate Gradient solver minimizes along the conjugate directions, the gradients or residuals are mutually Euclidean-orthogonal, i.e.

$$g^{(i)T} g^{(j)} = 0 \quad \text{for} \quad i, j \in \{1, .., p\} \text{ such that } i \neq j \tag{5.3.3}$$

where $g^{(i)} := Lq^{(i)} - w$ for all $i = 1, .., p$; see Saad [68]. This fact allows efficient implementations of the Conjugate Gradient algorithm, since only one conjugate direction needs to be stored at each stage. For our sampler Algorithm 8, equation (5.3.3) no longer holds because the algorithm does not optimize the cost function along conjugate directions. We must therefore orthogonalize the gradients by a Gram-Schmidt procedure on line 12, and the sampler requires storage of three vectors $g^{(j)}, r^{(j)}, v^{(j)}$. While this is less efficient than the Conjugate Gradient solver, which requires storage of two vectors, we note that this is a large improvement over storing all the conjugate vectors $\{v^{(1)}, ..., v^{(p)}\}$.

Conjugate Gradient sampling algorithms have been studied previously by Fox and Parker [59] and also in separate works by Fox [29], [30]. There is a difference between the algorithms introduced in these works and Algorithm 8; the CG sampler suggested by Fox and Parker [59] samples from the subspace spanned by the conjugate vectors produced by the usual CG solver, i.e. using the gradient at the solution of the line-search at each step such that (5.3.3) is satisfied, and so the conjugate directions are entirely deterministic when conditioned on the initial state. This contrasts with Algorithm 8, which uses the gradient at the current sample $q^{(j-1)}$ to construct stochastic conjugate directions. Similarly, Fox [29] suggests yet another conjugate directions sampler in which the convex function F and right-hand side w are essentially modified at each iteration, such that in our notation

$$F^{(j)}(q) = \frac{1}{2} q^T L q - w^{(j)T} q$$

Future work could involve generalizing this idea to non-Gaussian targets, in which the Bregman generator F is updated in an analogous way.

Algorithm 8 Conjugate Gradient Sampler

```
function CGSAMPLER(q^(0), L, w)
    g^(0) ← L q^(0) − w
    v^(1) ← g^(0)
    r^(1) ← g^(0)
    for j = 1, .., p do
        r^(j) ← v^(j)T L v^(j)
        z^(j) ← Normal(0, 1/r^(j))
        l^(j) ← z^(j) v^(j)T g^(j−1) / r^(j)
        q^(j) ← q^(j−1) + l^(j) v^(j)
        g^(j) ← g^(j−1) + l^(j) L v^(j)
        d_j ← r^(j)T g^(j) / r^(j)T r^(j)
        r^(j+1) ← (1 + d_j) r^(j) − l^(j) L v^(j)
        v^(j+1) ← r^(j+1) − r^(j+1)T g^(j) / r^(j) l^(j) v^(j)
    end for
    return q^(p)
end function
```

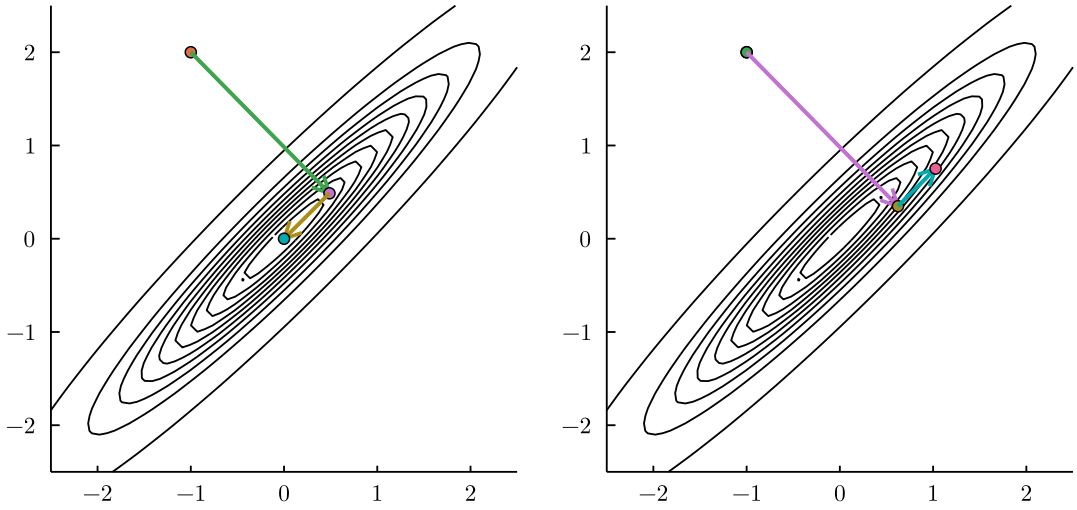

Figure 5.3.1: The Conjugate Gradient solver (left) and sampler (right) in two-dimensions for $w = 0$ and L a correlation matrix with variance $s_1^2 = L_{11} = 1.0$ in the x-direction, $s_2^2 = L_{22} = 1.0$ in the y-direction and correlation 0.95, and the same initial iterate $q^{(0)} = (-1, 2)^T$. The first conjugate direction $v^{(1)}$ is simply the residual $-Lq^{(0)}$ at the intial iterate in both cases, but the step length is chosen differently; the solver minimizes a one-dimensional cost function at each step and hence moves deterministically, whereas the sampler makes stochastic moves along the search directions.

Early Termination and Preconditioning

The Conjugate Gradient solver finds the solution in fewer than p steps when L has repeated eigenvalues [40]. For the Conjugate Gradient sampler, this property translates into pathological early termination which we will discuss here, following the discussion in Fox [29].

De nition 5.3.1. Let $g \in R^p$ be a vector and $L \in R^{p \times p}$ be a matrix. The j-order Krylov subspace generated by L and g is

$$K_j(L, g) := \text{Span}\{g, Lg, .., L^{k-1}g\}$$

Clearly $K_j(L, g) \subseteq K_{j+1}(L, g)$ for any $j \geq 1$. One can show that if L has $m \leq p$ distinct eigenvalues, then the dimension of the Krylov subspace of any order is at most m [29], [68]. The relevance of the Krylov subspaces is made clear by the following proposition.

Proposition 5.3.2. For $j = 1, .., p$, the j^{th} iterate $q^{(j)}$ in Algorithm 8 is in the j-order Krylov subspace generated by L and $g^{(0)}$, afinely shifted by $q^{(0)}$;

$$q^{(j)} \in q^{(0)} + K_j(L, g^{(0)})$$

Proof. Clearly, line 9 implies that

$$q^{(j)} \in q^{(0)} + \text{Span}\{v^{(1)}, .., v^{(j)}\}$$

and hence the result follows if we can show

$$\text{Span}\{v^{(1)}, .., v^{(j)}\} \subset K_j(L, g^{(0)}) \tag{5.3.4}$$

We will show this by induction. For $j = 1$, we have $r^{(1)} = v^{(1)} = -g^{(0)}$ and clearly (5.3.4) holds.

Assume an inductive hypothesis that the following both hold for some $j = j_0 \in \{1, .., p-1\}$,

$$\text{Span}\{v^{(1)}, .., v^{(j)}\} \subset K_j(L, g^{(0)}) \tag{5.3.5}$$

$$\text{Span}\{r^{(1)}, .., r^{(j)}\} \subset K_j(L, g^{(0)}) \tag{5.3.6}$$

Line 12 with $j = j_0$ states that

$$r^{(j_0+1)} := (1 + \alpha_{j_0} \beta^{(j_0)}) r^{(j_0)} - \beta^{(j_0)} L v^{(j_0)}$$

By the inductive hypothesis, the vectors $v^{(j_0)}$ and $r^{(j_0)}$ are elements of the Krylov subspace $K_{j_0}(L, g^{(0)})$. Thus, $v^{(j_0)} = \sum_{s=0}^{j_0-1} u_s L^s g^{(0)}$ for some coefficients $\{u_s\}_{s=1}^{j_0}$, and hence

$$L v^{(j_0)} = L \sum_{s=0}^{j_0-1} u_s L^s g^{(0)} = \sum_{s=0}^{j_0-1} u_s L^{s+1} g^{(0)} \in K_{j_0+1}(L, g^{(0)})$$

It follows that $r^{(j_0+1)} \in K_{j_0+1}(L, g^{(0)})$, which shows (5.3.6) holds for $j = j_0 + 1$. Finally, line 13 states that

$$v^{(j_0+1)} := r^{(j_0+1)} - (r^{(j_0+1)^T} g^{(j_0)} / r^{(j_0)^T} g^{(j)}) v^{(j_0)}$$

which implies $v^{(j_0+1)} \in K_{j_0+1}(L, g^{(0)})$. Thus by the principle of induction, (5.3.4) holds for all $j = 1, ..p$ and the result follows. □

Suppose L has $m < p$ distinct eigenvalues. Then Proposition 5.3.2 implies that the final iterate $q^{(p)}$ is constrained to a m-dimensional subspace, and hence the full state space is not properly explored. In particular, during the execution of the algorithm we will find $v^{(j)} = 0$ for all $j > m$.

The simplest example of this pathology, noted by Fox [29], is the case when L is the $p \times p$ identity matrix. In this case L has p repeated eigenvalues all equal to 1, and a simple calculation shows $v^{(2)} = r^{(2)} = 0$.

The problem can be addressed by a two-sided preconditioning For a non-singular matrix U, note that the linear system $Lm = w$ is equivalent to

$$\underbrace{U^T L U}_{:= L} U^{-1} m = \underbrace{U^T w}_{:= w}$$

We let $L := U^T L U$ and $w := U^T w$. One can then solve the system $Lz = w$ for z, and recover $m = Uz$. In the sampling context, for $q \sim \text{Normal}(L^{-1}w, L^{-1})$ we set $x = U^{-1}q$, so that

$$\begin{aligned} E(x) &= U^{-1}m = U^{-1}Uz = L^{-1}w \\ \text{Cov}(x) &= U^{-1}L^{-1}U^{-T} = U^{-1}UL\,U^T U^{-T} = L \end{aligned}$$

The Conjugate Gradient sampler can then be used to sample $x \sim \text{Normal}(L^{-1}w, L^{-1})$, since L is clearly symmetric positive-de nite. The idea is to choose U such that L has p distinct eigenvalues so that the algorithm successfully completes all iterations, and then set $q = Ux$ for each sample x.

5.4 Numerical Experiments

In order to con rm the theoretical results for the two novel algorithms and assess their performance, numerical experiments on various models are presented here. The novel algorithms presented in this chapter were implemented in a Julia package called GenericBayes.jl [60]. The package allows running the algorithms on arbitrary posterior distributions and Bregman generators. All numerical experiments were run on an Intel i5-3470 CPU clocked at 3.20 GHz with 8 GB of memory.

5.4.1 Multivariate Gaussian Target

Our rst experiment will demonstrate that Algorithms 6 and 8 both converge to the correct target distribution and the samples are independent. Recall that these algorithms sample from a multivariate Gaussian distribution with mean $L^{-1}w$ and covariance matrix L^{-1},

$$q \sim \text{Normal}(L^{-1}w, L^{-1})$$

We select L and w based on an example given by Fox [29]. Let U be a $p \times p$ upper-triangular matrix with ones on the diagonal, and $\text{Uniform}(0,1)$ random variables on the super-diagonal. We then let w be a p-dimensional vector of ones, and $L := U^T U$. Clearly this choice of L is symmetric positive-de nite.

The Gaussian Elimination and Conjugate Gradient samplers were both used to sample from this target distribution, with dimension $p = 10$. Both samplers were

started from a randomly chosen position in the unit hypercube, i.e. Uniform $(0, 1)$ for all $i = 1, .., 10$. For both algorithms, each sample is obtained by taking the last iterate of the algorithm, where the initial iterate is the previous sample. Figure 5.4.1 shows the convergence of the sample mean and covariance towards their respective true values $L^{1}w$ and L^{1}. There is a clear $\overline{N}$-convergence in the 2-norm for both quantities when computed with either algorithm, as expected for MCMC. This suggests both algorithms are converging to the correct target distribution.

Table 5.4.1 shows the minimum effective sample size and the equivalent maximum integrated autocorrelation time for both samplers with $N = 10^6$ samples. Clearly both samplers have a maximum IAT close to 1, implying near independent samples. This con rms our analytic result in Corollary 5.2.3 for the Gaussian Elimination sampler, and hints at the same independence result for the Conjugate Gradient sampler.

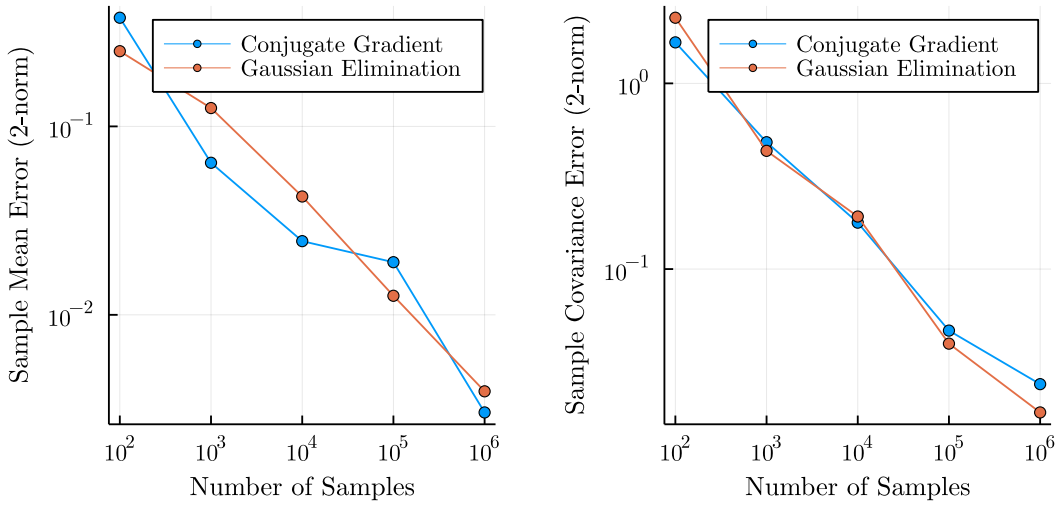

Figure 5.4.1: Error in the 2-norm for the sample mean and covariance for both samplers.

Sampler	Effective Sample Size	Integrated Autocorrelation Time
Gaussian Elimination	992764	1.00729
Conjugate Gradient	992818	1.00723

Table 5.4.1: Minimum effective sample size (ESS) and the equivalent maximum integrated autocorrelation times (IATs) for both sampling algorithms. The number of samples $N = 10^6$.

Circular Conditional Autoregression

The next example is another multivariate Gaussian distribution from Held and Rue writing in Chapter 13 of [32]. For f 2 [0, 1) and k > 0, de ne the p p precision matrix

$$L = \frac{k}{2}\begin{pmatrix} 2 & f & & & & & f \\ f & 2 & f & & & & \\ & f & 2 & f & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & f & 2 & f & \\ & & & & f & 2 & f \\ f & & & & & f & 2 \end{pmatrix}$$

The resulting multivariate Gaussian distribution q Normal (0, L ^{1}) is a conditional autoregression (CAR) model with a periodic boundary. The precision L is a circulant matrix [67], for which there is a simple formula for the eigenvalues yielding

$$l_j = k \quad 1 \quad f \cos\left(\frac{2p\,j}{p}\right), \qquad j = 1, .., p$$

Note that $l_k = l_{p\ k}$ for any k = 1, .., p 1, and so L has repeated eigenvalues. This is shown in Figure 5.4.2 for p = 25.

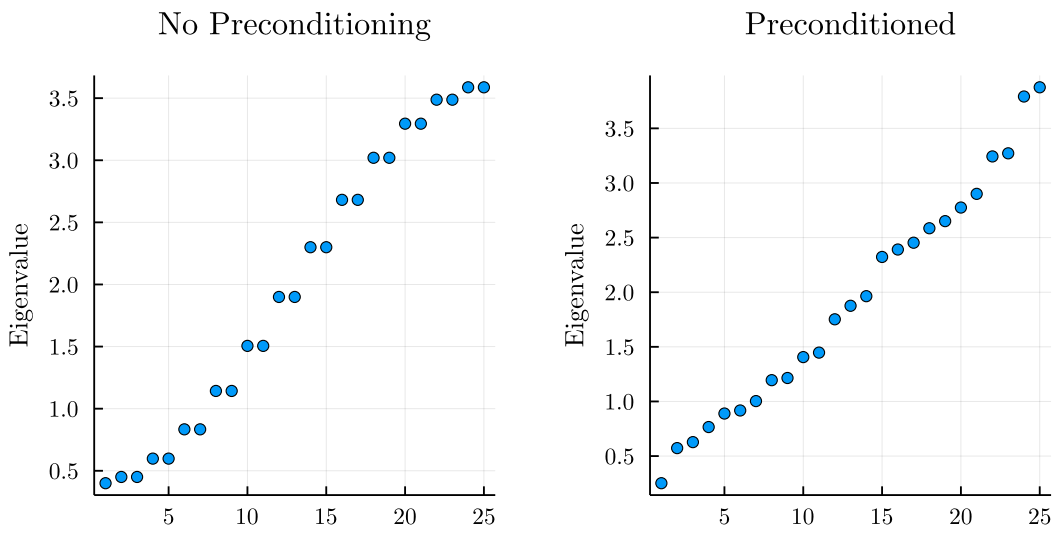

Figure 5.4.2: Left: Eigenvalues of the CAR precision matrix in increasing order for p = 25, f = 0.8 and k = 2. Note the staircase pattern indicating pairs of equal eigenvalues. Right: Eigenvalues of the preconditioned CAR precision matrix; all eigenvalues are now distinct.

As expected from our discussion in Section 5.3.1, Algorithm 8 terminates early after 14 iterations when sampling this target distribution with $p = 25$. The problem is alleviated by a preconditioning $L = U^T L U$, where U is the matrix de ned above, i.e. ones on the diagonal and Uniform(0,1) random variables on the super-diagonal. The Conjugate Gradient sampler successfully completes all $p = 25$ iterations with this preconditioner, and the samples can be transformed back with $q := Ux$.

5.4.2 Generalized Linear Models

Logistic Regression

The next numerical experiment uses a logistic regression model from Chapter 10 of McElreath [53]. The full probability model is

$$\begin{aligned} y_i &\quad \text{Bernoulli}(l_i) \\ \log \frac{l_i}{1 \quad l_i} &= b_1 X_{i1} + b_2 X_{i2} + g_{\text{ACTOR}[i]} \\ g_s &\quad \text{Normal}(0, 10) \qquad \text{for all} \qquad s = 1, .., 7 \\ b_1 &\quad \text{Normal}(0, 10) \\ b_2 &\quad \text{Normal}(0, 10) \end{aligned}$$

Here $g_{\text{ACTOR}[i]}$ represents one of seven different offset variables $g_1, .., g_7$, each corresponding to a particular individual who undergoes multiple trials. For trial i, the variable ACTOR[i] is the individual taking part in the trial. There are therefore 9 parameters in total; $g_1, .., g_7, b_1, b_2$. The logistic regression is used to model an experiment intended to assess the social characteristics of chimpanzees. After a suitable encoding of the experimental data, the design matrix X has 504 rows, each representing a trial in the experiment, with 9 columns each corresponding to a parameter. The data are such that every element of the design matrix is either 1 or 0, so no scaling or centering is necessary. The model ts the framework of a Generalized Linear Model with canonical link function discussed in Section 1.3.

As noted by McElreath, the posterior is skewed and therefore provides an example of a target distribution which is not close to a multivariate Gaussian. We ran the Orthogonal Gradient and Orthogonal Gibbs samplers on the model, and for comparison also ran the usual deterministic-scan Gibbs sampler and a Julia implementation of the No U-Turn Sampler (NUTS). All of the sampling algorithms sampled $N = 10,000$ samples, starting from the origin $q^{(0)} = 0\ 2\ R^9$. For Gibbs, Orthogonal Gibbs and Orthogonal Gradient, the univariate slice sampler outlined by Neal [56] was used as the subsampler for drawing from conditionals. A small amount of hand-tuning the parameters of the slice sampler was done to roughly optimize the effective sample sizes for these samplers.

Table 5.4.2 shows posterior means and standard deviations for each sampler, and the posterior mode or MAP estimator. These summary statistics are in close agreement

Parameter	Orthogonal Gradient	Orthogonal Gibbs	Gibbs	NUTS	Posterior mode
g_1	-0.729 (0.261)	-0.735 (0.268)	-0.729 (0.271)	-0.736 (0.267)	-0.721
g_2	5.367 (1.437)	5.408 (1.523)	5.408 (1.544)	5.38 (1.519)	4.714
g_3	-1.041 (0.261)	-1.041 (0.281)	-1.043 (0.281)	-1.045 (0.278)	-1.024
g_4	-1.038 (0.267)	-1.043 (0.277)	-1.042 (0.28)	-1.042 (0.28)	-1.024
g_5	-0.729 (0.262)	-0.733 (0.266)	-0.732 (0.268)	-0.733 (0.272)	-0.721
g_6	0.22 (0.257)	0.221 (0.267)	0.224 (0.268)	0.219 (0.269)	0.213
g_7	1.789 (0.382)	1.785 (0.386)	1.794 (0.389)	1.786 (0.389)	1.733
b_1	0.832 (0.258)	0.831 (0.261)	0.829 (0.261)	0.834 (0.265)	0.817
b_2	-0.132 (0.289)	-0.131 (0.299)	-0.13 (0.301)	-0.13 (0.3)	-0.127

Table 5.4.2: Mean (standard deviation) for each sampler and the posterior mode (MAP).

Sampler	Runtime (s)	Effective Sample Size (min, max)	ESS per sec (min, max)	Integrated Autocorrelation Time (min, max)
Orthogonal Gradient	631.6	9441.3, 10065.8	14.9, 15.9	0.993, 1.059
Orthogonal Gibbs	192.0	9003.3, 10191.3	46.9, 53.1	0.981, 1.111
Gibbs	33.0	2396.7, 9079.6	72.5, 274.8	1.101, 4.172
NUTS	5.1	5952.3, 11117.6	1175.2, 2195.0	0.899, 1.68

Table 5.4.3: Effective sample sizes and integrated autocorrelation times for each sampler. Each sample was run for $N = 10,000$ samples. Minima and maxima are taken over all 9 parameters.

between all of the samplers, and also with McElreath's results for the same model using the Stan [18] inference engine. The parameter means all appear to be close to the posterior mode values, except for g_2 whose marginal posterior distribution is skewed.

Table 5.4.3 shows performance diagnostics for each of the samplers. The Orthogonal Gibbs and Orthogonal Gradient algorithms perform well in terms of effective sample size, with minimum and maximum values both close to the number of samples $N = 10,000$. Both Gibbs and NUTS attain their minimum ESS for the b_2 parameter, whose autocorrelation is shown for each sampler in Figure 5.4.3. The autocorrelation clearly decays more quickly for Orthogonal Gibbs/Gradient than for Gibbs and NUTS.

However, table 5.4.3 shows that this statistical performance comes with the cost of much longer runtimes than Gibbs and NUTS, resulting in Orthogonal Gibbs/Gradient being least performant in terms of effective samples per second. The long runtimes of our algorithms are of course due to the j-dimensional Newton optimization and log-determinant evaluation over several times for each of $j = 1,..,p$. As we discussed previously, this leads to a $O(p^4)$ scaling in time complexity for both samplers.

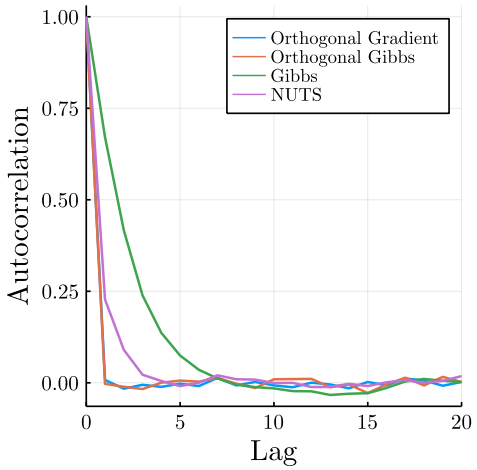

Figure 5.4.3: Empirical autocorrelation of b_1 for each of the four samplers.

Part III

Conclusion

Conclusion and Future Directions

The thesis has introduced new MCMC algorithms which exploit dually- at geometry in order to more ef ciently sample from a posterior distribution. The Orthogonal Gibbs algorithm samples primal variables given dual variables in the k-cut co-ordinate system, increasing k after each iteration. The Orthogonal Gradient algorithm samples the distribution constrained to the line spanned by the steepest descent direction of the negative log-density, and then recursively samples from the target constrained to a submanifold orthogonal to the line. Both algorithms successively sample on submanifolds in parameter space which are orthogonal with respect to the Riemannian metric, and are therefore closely related to conjugate directions solvers for linear systems when the metric is constant. For Gaussian target distributions, the Orthogonal Gibbs sampler has been proven to produce independent samples and numerical experiments suggest a similar result for the Orthogonal Gradient sampler.

The dually- at geometry on parameter space used by the algorithms is speci ed by a Bregman divergence, which may or may not depend on the target distribution. Full exponential families have a natural dually- at geometry for which the relevant Bregman divergence is generated by the cumulant generating function of the family. For log-concave posterior distributions, I have introduced the Posterior Bregman Divergence generated by the negative log-density of the posterior. The metric induced by this divergence has the property that orthogonal directions are independent sampling directions in the case of a Gaussian target, motivating the new sampling algorithms. Methods using Bregman divergences have been suggested to solve other numerical problems such as convex optimization [61] and clustering [6], but to my knowledge the new sampling algorithms are the rst to explicitly use dually- at geometry induced by a Bregman divergence.

The new algorithms can be compared to previous geometric MCMC algorithms; in particular the Riemannian variants of Hamiltonian Monte Carlo (HMC) and the Metropolis Adjusted Langevin Algorithm (MALA) introduced by Girolami and Calderhead [39]. These sampling algorithms use a Riemannian, rather than dually- at, structure to facilitate ef cient sampling from the posterior. They require numerical integration of ODEs or SDEs, whereas the Orthogonal Gibbs/Gradient algorithms make use of explicit geodesic ows in the dually- at geometry and require no numerical integration. Hamiltonian Monte Carlo disintegrates an extended target distribution over the level sets of the Hamiltonian function; similarly, the new algorithms disintegrate the

target over ber bundles de ned by the orthogonal foliations of Information Geometry. While HMC/MALA and their Riemannian variants require the gradient of the target log-density, Orthogonal Gibbs/Gradient only use derivatives of the Bregman generator and do not necessarily require any derivatives of the target log-density, unless the Bregman generator depends on the log-density. Future work could explore the possibility of Bregman generators which are cheap approximations the negative target log-density, perhaps by subsampling of data in a similar fashion to Stochastic Gradient HMC [51].

Hamiltonian Monte Carlo may require signi cant hand-tuning of algorithmic parameters; namely the step-size used in numerical integration of Hamilton's equations and the number of integration time steps. One must carefully choose these parameters to attain an optimal acceptance rate. In contrast, the algorithms introduced in this thesis are equivalent to a sequence of Gibbs updates in various co-ordinate systems, and hence their proposals are always accepted. The only hand-tuning required is the number of subsamples per iteration and any algorithmic parameters of the subsampler algorithm used to draw from conditional distributions. In the regression model given in Section 5.4, the slice sampler used for subsampling required only minimal hand-tuning.

For target distributions which are not log-concave and thus have a log-density whose Hessian is not positive-de nite, it is not clear how a Bregman divergence should be chosen for use in the new algorithms. To solve this problem for RM-HMC, the SoftAbs metric has been suggested by Betancourt [9] which forces all of the eigenvalues of the Hessian to be positive. Unfortunately, it does not seem possible to extend this idea to a Bregman divergence, which requires a globally de ned convex function rather than a local Riemannian metric. To deal with densities which are not log-concave, I suggest the following directions for future research.

One could use a Bregman generator de ned as the smooth convex function which is closest to the (non-convex) negative log-density in some metric, e.g. the in nity norm.

The Orthogonal Gradient/Gibbs algorithms may be generalized to an arbitrary conjugate connections geometry, i.e. more general than dually- at. Perhaps numerical integration could be used to move along a r -geodesic, and then similarly on an orthogonal r -geodesic, with a Metropolis-Hastings accept/reject step to correct for numerical errors similar to Hamiltonian Monte Carlo. This approach would allow the geometry to speci ed by any divergence function, rather than being restricted to Bregman divergences.

When the Fisher Information is singular, the parameter space Q can be understood as a ber bundle whose bers represent classes of equivalent data-generating densities. The quotient space with respect to this equivalence relation has a natural non-singular positive-de nite metric; see Xavier and Barroso [77]. I suggest investigating whether a natural Bregman generator is also induced on the quotient space, such that the Orthogonal Gibbs/Gradient algorithm could be used to sample in the quotient space, and some simple algorithm such as random walk

could be used to sample within the bers. This is very similar, and perhaps equivalent to, the idea of Likelihood Informed Subspace (LIS) of Cui et al [22].

A barrier to the practical application of the novel algorithms introduced in this thesis is their computational cost. When run for all p iterations, Orthogonal Gradient and Orthogonal Gibbs both have an asymptotic time complexity of $O(p^4)$ assuming a reasonable cost model, which may be prohibitively expensive for models with many parameters. To alleviate the computational cost, I offer the following suggestions.

In optimization, conjugate gradient algorithms are often restarted, i.e. terminated before reaching d iterations where d is the dimension of the space on which the cost function is de ned. If it can be shown that each iteration of the Orthogonal Gradient algorithm preserves the target distribution, it may be possible to restart the algorithm before reaching the full number of iterations, and therefore avoid much of the computational cost. The number of iterations would then be an algorithmic parameter which balances computational expense with the statistical performance of the algorithm, i.e. the autocorrelation of the resulting chain.

If applied to models for which the likelihood is expensive to evaluate such as PDE-constrained models in engineering and Uncertainty Quanti cation applications, the computational cost of the Orthogonal Gibbs/Gradient algorithms may be deemed acceptable in comparison to that of likelihood evaluations. In other words, the favourable statistical performance of the novel algorithms make them a good choice when the forward model is already expensive to evaluate.

Reformulating the algorithms in the co-ordinate free terms of differential forms could offer a solution; much of the computational cost arises from evaluating a determinant at each step, and perhaps this could be reduced by nding a relationship between the volume forms on orthogonal submanifolds.

On the way to deriving the new algorithms, the thesis has also provided several new results in Information Geometry. These include characterizations of the inherited geometry on submanifolds in terms of divergences given by Theorem 2.5.3 and a result on restrictions of Bregman divergences characterizing e- at submanifolds given in Theorem 2.5.8.

The new algorithms presented in this thesis will hopefully prove useful in statistical applications. This work and that of previous authors has shown that orthogonality, duality, geodesics and other ideas from Information Geometry can be leveraged in computational statistics to do inference more ef ciently. I hope this continues to be an active area of research so that the bene ts can be fully realized in statistical applications.

Part IV

Appendices

Appendix A

Details of Smooth Manifold Theory

Proposition A.0.1. Let T_aR^p be the set of linear derivations at $a \in R^p$, i.e. the set of linear maps $C^\infty(R^p) \to R$ satisfying the Leibniz rule (2.1.1). Then T_aR^p is a vector space of dimension p with a basis defined by the p derivations

$$\partial_{ia} : C^\infty(R^p) \to R$$
$$\partial_{ia}(f) := \frac{\partial f}{\partial x^i}_{x=a}$$

for $i = 1, .., p$.

Proof. The following is based on the proof of proposition 3.2 from Lee [46]. For any two derivations $V, W \in T_aR^p$ we define their scalar multiplication and linear combination respectively as

$$(\lambda V)(f) := \lambda V(f) \tag{A.0.1}$$
$$(\lambda_1 V + \lambda_2 W)(f) := \lambda_1 V(f) + \lambda_2 W(f) \tag{A.0.2}$$

for any $\lambda, \lambda_1, \lambda_2 \in R$ and $f \in C^\infty(R^p)$. It is easy to verify that linear combinations defined in this way are also derivations, and hence T_aR^p is a vector space. The zero element of the vector space 0_a is defined via $0_a(f) := 0$ for all $f \in C^\infty(R^p)$.

To show that $\{\partial_{ia} | i = 1, .., p\}$ is a basis for the space of derivations at a, we first need to show that the set is linearly independent. To this end, suppose that for some $v^i \in R$ for $i = 1, .., p$ we have

$$\sum_{i=1}^{p} v^i \partial_{ia} = 0_a$$

Then for any $f \in C^\infty(R^p)$,

$$\sum_{i=1}^{p} v^i \frac{\partial f}{\partial x^i}_{x=a} = 0$$

In particular, set f to be the j^{th} co-ordinate x^j for each of $j = 1, .., p$. Since $\frac{\partial x^j}{\partial x^i} = 1$ when $i = j$ and 0 otherwise, it follows that $v^j = 0$ for all $j = 1, .., p$ and hence $\{\partial_{ia}\}$ is a linearly independent set.

It remains to show that $\{\partial_{ija}\}$ spans T_aR^p and hence is a basis. To this end let $W \in T_aR^p$ be an arbitrary derivation. By Taylor's theorem, any $f \in C^\infty(R^p)$ can be written

$$f(x) = f(a) + \sum_{i=1}^{p} \frac{\partial f}{\partial x^i}\Big|_{x=a} (x^i - a^i) + \sum_{i,j=1}^{p} (x^i - a^i)(x^j - a^j) \int_0^1 (1-t) \frac{\partial^2 f}{\partial x^i \partial x^j}(a + t(x-a))\, dt$$

Consider the action of W on the right-hand side of this equation, which is the sum of its action on each term by linearity. The first term is constant, and hence $W(f(a)) = f(a)W(1)$ by linearity. But by the Leibniz rule (2.1.1)

$$W(1) = W(1 \cdot 1) = 1 \cdot W(1) + 1 \cdot W(1) \quad \Rightarrow \quad W(1) = 0$$

and hence $W(f(a)) = 0$. Similarly, the third term vanishes;

$$\begin{aligned} W\left(\sum_{i,j=0}^{p} C_{ij}(x^i - a^i)(x^j - a^j) \right) &= \sum_{i,j=0}^{p} C_{ij} W\left((x^i - a^i)(x^j - a^j) \right) \\ &= \sum_{i,j=0}^{p} C_{ij} \left[(x^j - a^j)\big|_{x=a} W(x^i - a^i) + (x^i - a^i)\big|_{x=a} W(x^j - a^j) \right] \\ &= \sum_{i,j=0}^{p} C_{ij} \left[0 \cdot W(x^i - a^i) + 0 \cdot W(x^j - a^j) \right] = 0 \end{aligned}$$

where we have set $C_{ij} := \int_0^1 (1-t) \frac{\partial^2 f}{\partial x^i \partial x^j}(a + t(x-a))\, dt$ for $i, j = 1, .., p$. Finally we are left with

$$W(f) = W\left(\sum_{i=1}^{p} \frac{\partial f}{\partial x^i}\Big|_{x=a} (x^i - a^i) \right) \tag{A.0.3}$$

$$= \sum_{i=1}^{p} \frac{\partial f}{\partial x^i}\Big|_{x=a} (W(x^i) - W(a^i)) \tag{A.0.4}$$

By the above argument, the action of W on a constant function vanishes and so $W(a^i) = 0$ for all $i = 1, .., p$. Hence we conclude that for any $W \in T_aR^p$, we can write

$$W(f) = \sum_{i=1}^{p} W(x^i) \partial_{ija}(f) \qquad \text{for all } f \in C^\infty(R^p)$$

$$\Rightarrow \quad W = \sum_{i=1}^{p} W(x^i) \partial_{ija}$$

and hence the set $\{\partial_{ija} | i = 1, .., p\}$ forms a basis for T_aR^p, which must therefore be a p-dimensional vector space. □

Lemma A.0.2. If $F : Q \to Q'$ is a diffeomorphism, its differential $dF : T_qQ \to T_{F(q)}Q'$ is a linear isomorphism.

Proof. Linearity follows trivially from the de nition of the differential; if $V_1, V_2 \in T_qQ$ then

$$\begin{aligned} dF_q(\lambda_1 V_1 + \lambda_2 V_2)(f) &= (\lambda_1 V_1 + \lambda_2 V_2)(f \circ F) \\ &= \lambda_1 V_1(f \circ F) + \lambda_2 V_2(f \circ F) \\ &= \lambda_1 dF_q(V_1)(f) + \lambda_2 dF_q(V_2)(f) \end{aligned}$$

To show injectivity, suppose that some tangent vector $V \in T_qQ$ is mapped to the zero element in $T_{F(q)}Q'$ under the differential;

$$\begin{aligned} dF_q(V) = 0_{F(q)}) \quad & dF_q(V)(f) = 0 \qquad \text{for all } f \in C^\infty(Q') \\ \Rightarrow \quad & V(f \circ F) = 0 \qquad \text{for all } f \in C^\infty(Q') \end{aligned}$$

Then for any $f \in C^\infty(Q)$, set $f = f \circ F^{-1}$; this is valid since F is bijective. We then have

$$V(f \circ F) = V(f \circ F^{-1} \circ F) = V(f) = 0 \qquad \text{for all } f \in C^\infty(Q)$$

which of course implies that V is the zero element in T_qQ, and hence dF_q is injective. For surjectivity, let $W \in T_{F(q)}Q'$. Then de ne a tangent vector in T_qQ as $V(f) := W(f \circ F^{-1})$. Then for any $f \in T_{F(q)}Q'$ we have

$$dF_q(V)(f) = V(f \circ F) = W(f \circ F \circ F^{-1}) = W(f)$$

and hence the differential at q is surjective; it is therefore a linear bijection or linear isomorphism. □

Theorem 2.1.3. Let Q be a smooth manifold with a co-ordinate chart (U, q) for $U \subset Q$. Then for any $q \in Q$, a basis for tangent space T_qQ is given by the p tangent vectors

$$\begin{aligned} &\partial_{i|q,q} : C^\infty(Q) \to \mathbb{R} \\ &\partial_{i|q,q}(f) := \frac{\partial}{\partial q^i}_{q=f(q)} f \circ f^{-1} \qquad \text{for all } f \in C^\infty(Q), \quad i = 1, .., p \end{aligned}$$

Proof. By lemma A.0.2 and the fact that f is a diffeomorphism, the differential of the co-ordinate map $df_q : T_qQ \to T_{f(q)}\mathbb{R}^p$ is a linear isomorphism. Therefore the preimage of any basis in $T_{f(q)}\mathbb{R}^p$ is a basis for T_qQ. In fact, the basis stated in the theorem is exactly the preimage of the basis $\{\partial_{i|f(q)}\}$ de ned in proposition A.0.1. Indeed,

$$\begin{aligned} df_q(\partial_{i|f,q})(f) &= \partial_{i|f,q}(f \circ f) \\ &= \frac{\partial}{\partial x^i}_{x=f(q)} (f \circ f \circ f^{-1}) \\ &= \frac{\partial}{\partial x^i}_{x=f(q)} (f) \\ &= \partial_{i|f(q)}(f) \end{aligned}$$

□

The following is a restatement of proposition 2.1.6.

Proposition A.0.3. Let Q be a smooth p-dimensional manifold with $q \in Q$, and let (U, f) be a co-ordinate chart whose domain contains q, with co-ordinates labelled q^j, $j = 1, .., p$. The differentials $\{ d_q q^j \}_{j=1}^p$ of the local co-ordinate functions form a basis of cotangent space $T_q^* Q$ that is dual to the co-ordinate tangent vector basis $\{ \partial_i |_q \}_{i=1}^p$, i.e. they satisfy

$$d_q q^j(\partial_i) = \partial_i(q^j) = \delta_i^j$$

for all $i, j = 1, .., p$, where $\delta_i^j = 1$ when $i = j$ and is 0 otherwise.

Proof. The duality property follows trivially from the de nition of the co-ordinate basis;

$$d_q q^j(\partial_i) = \partial_i(q^j) = \frac{\partial q^j}{\partial q^i}_{f(q)} = \delta_i^j$$

To show that the co-ordinate differentials form a basis for cotangent space, rst we check linear independence; for some λ_k, $k = 1, .., p$ we set

$$\begin{aligned} & \lambda_k dq^k = 0 \\ \Rightarrow & \sum_{k=1}^p \lambda_k dq^k(V) = 0 \qquad \text{for all tangent vectors } V \in T_q Q \\ \Rightarrow & \sum_{k=1}^p \lambda_k V(q^k) = 0 \qquad \text{for all tangent vectors } V \in T_q Q \end{aligned}$$

Setting $V = \partial_{|jq,f}$ for each $i = 1, .., p$, we obtain

$$\sum_{k=1}^p \lambda_k \partial_{|jq,f}(q^k) = \sum_{i=1}^p \lambda_k \delta_i^k = \lambda_i = 0 \qquad \text{for all } i = 1, .., p$$

And hence $\{ dq^k | k = 1, .., p \}$ is a linearly independent set. Finally, to show that the set spans $T_q^* Q$, for a given linear functional $a \in T_q^* Q$ let $a_i := a(\partial_i)$ for $i = 1, .., p$. Then, we see that the linear combination of co-ordinate differentials with coef cients given by $\{ a_i \}$ is exactly a; for any $V = V^k \partial_k \in T_q Q$,

$$\begin{aligned} \sum_{i=1}^p a_i dq^i(V) &= \sum_{i=1}^p a(\partial_i) dq^i(V^k \partial_k) \\ &= \sum_{i=1}^p V^k a(\partial_i) dq^i(\partial_k) \\ &= \sum_{i=1}^p a(V^k \partial_i) \delta_k^i = \sum_{i=1}^p a(V^i \partial_i) = a(V) \end{aligned}$$

□

Proposition 2.5.6. Let N be a dually- at n-manifold. An embedded p-submanifold Q is e- at if and only if there exist co-ordinates b for Q such that the co-ordinate embedding L has the af ne form in primal co-ordinates,

$$q = L(b) := Ab + d \tag{2.5.7}$$

for some matrix $A \in R^{n \times p}$ and vector $d \in R^n$.

Dually, an embedded p-submanifold Q is m- at if and only if there exist co-ordinates g for Q such that the co-ordinate embedding L has the af ne form in dual co-ordinates,

$$h = L(g) := Ag + d \tag{2.5.8}$$

for some matrix $A \in R^{n \times p}$ and vector $d \in R^n$.

Proof. We show the result for the e- at case; the proof of the m- at case is identical. Assume rst that Q is e- at so that the second fundamental form vanishes for all vector elds. Note that $dL\, \nabla^{fl}_X Y$ is the tangential component of $\nabla_{dL(X)} dL(Y)$ in general, where ∇^{fl} is the primal connection on Q. Thus by the Gauss formula (2.5.3), the second fundamental form vanishes if and only if

$$dL\, \nabla^{fl}_X Y = \nabla_{dL(X)} dL(Y) \tag{A.0.5}$$

for any vector elds $X, Y \in \mathcal{X}(Q)$, where ∇^{fl} and ∇ are the primal connections on Q and N respectively. By Theorem 2.5.5,Q is dually- at and hence there exists a ∇^{fl}-af ne co-ordinate system b. This af ne system must satisfy

$$\nabla^{fl}_{\frac{\partial}{\partial b^a}} \frac{\partial}{\partial b^b} = 0 \qquad a, b = 1, .., p$$

Setting $X = \frac{\partial}{\partial b^a}$ and $Y = \frac{\partial}{\partial b^b}$ in (A.0.5), we deduce that

$$\nabla_{dL\, \frac{\partial}{\partial b^a}} dL\, \frac{\partial}{\partial b^b} = 0$$

Writing $dL(X) = B^i_a \frac{\partial}{\partial q^i}$ where $B^i_a := \frac{\partial L^i}{\partial b^a}$ is the co-ordinate differential and similarly for $dL(Y)$, we compute

$$\nabla_{B^i_a \frac{\partial}{\partial q^i}} B^j_b \frac{\partial}{\partial q^j} = B^i_a \left(\frac{\partial B^j_b}{\partial q^i} \frac{\partial}{\partial q^j} + B^j_b \nabla_{\frac{\partial}{\partial q^i}} \frac{\partial}{\partial q^j} \right) = 0$$

Note that since q is an af ne co-ordinate system, we have $\nabla_{\frac{\partial}{\partial q^i}} \frac{\partial}{\partial q^j} = 0$. Therefore

$$B^i_a \frac{\partial B^j_b}{\partial q^i} \frac{\partial}{\partial q^j} = 0$$

$$\Leftrightarrow \qquad B^i_a \frac{\partial B^j_b}{\partial q^i} = 0 \qquad \text{for all } j = 1, .., n, \qquad a, b = 1, .., p$$

$$\Leftrightarrow \qquad \frac{\partial}{\partial b^a} B^j_b = 0 \qquad \text{for all } j = 1, .., n, \qquad a, b = 1, .., p$$

where we have used the chain rule $\frac{\partial}{\partial b^a} = B^i_a \frac{\partial}{\partial q^i}$ to go from the second to the final line. It follows that the co-ordinate differential, or Jacobian of L, B^j_b is a constant for all b. Consequently the co-ordinate embedding L must be an affine transformation.

Conversely, suppose there exist co-ordinates b for Q, in which the co-ordinate embedding takes the form

$$q = L(b) := Ab + d$$

Note that the differential of the embedding in co-ordinates is a constant $\frac{\partial L^i}{\partial b^a} = A^i_a$, the i-a component of the matrix A. The co-ordinates b must therefore be $\nabla^{(1)}$-affine co-ordinates, because the connection coefficients (2.5.4) vanish:

$$G_{abc} = A^j_c A^k_a \partial_k \left(A^i_b \right) g_{ij} + A^i_a A^j_b A^k_c G_{jk} = 0$$

where we have used the fact that A^i_b is constant, and q are affine co-ordinates for N so $G_{jk} = 0$. Consequently Q is $\nabla^{(1)}$-flat, and hence dually-flat by Theorem 2.4.3. We will show that Q is e-flat in N by showing the equivalent statement (A.0.5). Let $X = X^a \frac{\partial}{\partial b^a}$ and $Y = Y^b \frac{\partial}{\partial b^b}$ be two arbitrary vector fields on Q. Since A^i_a is constant and q are affine co-ordinates, the r-covariant derivative of the embedded vector fields in N is

$$r_{dL(X)} dL(Y) = r_{A^i_a X^a \frac{\partial}{\partial q^i}} A^j_b Y^b \frac{\partial}{\partial q^j} = A^i_a A^j_b X^a \frac{\partial Y^b}{\partial q^i} \frac{\partial}{\partial q^j} \tag{A.0.6}$$

Similarly, the embedding of the $\nabla^{(1)}$-covariant derivative is

$$\begin{aligned} dL\left(\nabla^{(1)}_X Y \right) &= dL\left(\nabla^{(1)}_{X^a \frac{\partial}{\partial b^a}} Y^b \frac{\partial}{\partial b^b} \right) = dL\left(X^a \frac{\partial Y^b}{\partial b^a} \frac{\partial}{\partial b^b} \right) \\ &= A^i_a A^j_b X^a \frac{\partial Y^b}{\partial q^i} \frac{\partial}{\partial q^j} \end{aligned} \tag{A.0.7}$$

where we have used the fact that b is $\nabla^{(1)}$-affine, and substituted the chain rule $\frac{\partial}{\partial b^a} = A^i_a \frac{\partial}{\partial q^i}$. Comparing equations (A.0.6) and (A.0.7), we see that $dL\left(\nabla^{(1)}_X Y \right) = r_{dL(X)} dL(Y)$ for all vector fields X, Y on Q, and hence the second fundamental form is zero and Q is e-flat. □

Appendix B

Elementary Geometric Measure Theory and The Implicit Function Theorem

Throughout the thesis, we often need to represent a target probability density $p : R^p \to R$ in different co-ordinate systems. Furthermore, the orthogonal submanifold samplers in Chapter 5 require a disintegration of the target over submanifolds in R^p which are level sets of dual or primal co-ordinate functions. In this appendix we describe tools for dealing with such tasks from the field of Geometric Measure Theory. The standard reference for the topic is Federer [28], although we mostly refer to the introduction in a sampling context given by Diaconis et al [24]. We briefly review the concepts and state the main results in notation consistent with the thesis, omitting technical and analytic details. We assume that $1 \leq k < p$.

A k-dimensional submanifold embedded in R^p inherits an area measure from the Lebesgue measure on R^p. Intuitively, we imagine thickening the submanifold so that it has a p-dimensional volume, and hence an area measure is approximately the volume of a subset of the thickened submanifold divided by the volume of a $(p-k)$-dimensional ball. This can be made rigorous [24], [28] to define the k-dimensional Hausdorff measure H^k on R^p.

There are two ways to specify a k-dimensional embedded submanifold Q of R^p;

- Explicitly, i.e. with a smooth embedding or parameterization $L : V \to R^p$ where $V \subseteq R^k$.
- Implicitly, i.e. the level set of a function $F : R^p \to R^{p-k}$ where DF has full rank.

The area formula allows us to compute the Hausdorff measure for an explicitly parameterized submanifold Q. Recall that dq and db respectively refer to the p- and k-dimensional Lebesgue measures on the q and b co-ordinate charts.

Theorem B.0.1 (Area Formula). Let $L : \mathbb{R}^k \to \mathbb{R}^p$ be a smooth embedding with $k \le p$. Then for any Borel $U \subset \mathbb{R}^k$ and Borel-measurable $h : \mathbb{R}^p \to \mathbb{R}$,

$$\int_U h(L(b)) J_k L(b)\, db = \int_{L(U)} h(q)\, \mathcal{H}^k(dq)$$

where

$$J_k L(b) := \sqrt{\det DL(b)^T DL(b)}$$

is the k-dimensional Jacobian of L.

The area formula as stated by Federer and Diaconis et al is more general than this; it is only necessary that L be Lipschitz, so that differentiability is guaranteed almost everywhere by Rademacher's theorem. For a Riemannian manifold, the Riemannian volume element in a co-ordinate chart q is the measure $\sqrt{\det G(q)}\, dq$. Suppose $\mathbb{R}^p$ is endowed with the Euclidean metric $G(q) = 1_{p \times p}$, so that the inherited metric on the submanifold Q is $DL^T DL$. Then Theorem B.0.1 shows that the k-dimensional Hausdorff measure coincides with the inherited Riemannian volume element.

Diaconis et al [24] describe how to use the area formula to sample from the Hausdorff measure on an embedded submanifold. In particular, the area formula generalizes the change of variable formula for probability densities, as the proof of the following shows.

Corollary B.0.2 (Change of Variable Formula). Let $p : Q \to \mathbb{R}$ be a probability density on $Q \subset \mathbb{R}^p$, and let $f : Q \to X$ be a diffeomorphism where $X = f(Q) \subset \mathbb{R}^p$. Then

$$\int_{f(V)} \frac{p(f^{-1}(x))}{|\det (Df(f^{-1}(x)))|}\, dx = \int_V p(q)\, dq$$

for a Borel set $V \subset Q$.

Proof. Set $L(x) := f^{-1}(x)$ for all $x \in X$, so that $k = p$. Since f is a diffeomorphism, then $L = f^{-1}$ is also a diffeomorphism and hence is a smooth embedding. By the inverse function theorem [54], the differential matrices of L and f, which are both square and of full rank, are related via

$$DL(x) = \left(Df(f^{-1}(x)) \right)^{-1} \tag{B.0.1}$$

Thus the p-dimensional Jacobian of f is given by

$$\begin{aligned} J_p L(x) &= \sqrt{\det DL(x)^T DL(x)} = \sqrt{\det DL(x)^T \det DL(x)} \\ &= \sqrt{\det DL(x)^2} = \det DL(x) \\ &= \det \left(Df(f^{-1}(x)) \right)^{-1} \end{aligned}$$

The result follows from Theorem B.0.1 with $g(q) := p(q)$, $U = f^{-1}(V)$, and noting that the Hausdorff measure agrees with the Lebesgue measure when $k = p$. □

The co-area formula gives a disintegration over the k-dimensional level sets of a function $F : \mathbb{R}^p \to \mathbb{R}^{p-k}$.

Theorem B.0.3 (Co-Area Formula). Let $F : \mathbb{R}^p \to \mathbb{R}^{p-k}$ be Lipschitz. Then

$$\int_{\mathbb{R}^p} h(q) J_{p-k} F(q) \, dq = \int_{\mathbb{R}^{p-k}} \int_{F^{-1}(z)} h(q) \, \mathcal{H}^k(dq) \, dz$$

where

$$J_{p-k} F(q) = \sqrt{\det DF(q) DF(q)^T}$$

Diaconis et al [24] use the co-area formula to derive a disintegration of a probability measure over the level sets of F.

The orthogonal submanifold sampling algorithms in Chapter 5 sample from conditional distributions on implicitly defined submanifolds, where the conditional densities are defined with respect to the Hausdorff measure. We require an explicit parameterization and its differential matrix in order to use the area formula to compute the density with respect to the Lebesgue measure for sampling. The existence of such a parameterization, and an explicit form for the differential matrix, is provided by the Implicit Function Theorem. We adapt the stated theorem from Lee [46] and Kantorovitz [43].

Theorem B.0.4 (Implicit Function Theorem). Let $F : \mathbb{R}^k \times \mathbb{R}^{p-k} \to \mathbb{R}^k$ be smooth on a neighbourhood of $(a_0, b_0) \in \mathbb{R}^k \times \mathbb{R}^{p-k}$, and let $c = F(a_0, b_0)$. If the matrix

$$\frac{\partial F}{\partial a}(a_0, b_0) := \left(\frac{\partial F^i}{\partial a^j}(a_0, b_0) \right)_{i,j=1,\ldots,k}$$

is non-singular, then there exist neighbourhoods $W \subset \mathbb{R}^k$ of a_0 and $V_0 \subset \mathbb{R}^{p-k}$ of b_0, and a smooth function $a : V_0 \to W_0$ such that $F(a, b) = c$ for $(a, b) \in V_0 \times W_0$ if and only if $a = a(b)$. Furthermore, the differential matrix of a satisfies

$$Da = \frac{\partial a^i}{\partial a^j} = \frac{\partial F}{\partial a}^{-1} \frac{\partial F}{\partial b}$$

Lemma B.0.5. For $1 \le k < p$, let $A \in \mathbb{R}^{p \times k}$ and $B \in \mathbb{R}^{p \times (p-k)}$ be two matrices of full rank with $A^T B = 0_{k \times (p-k)}$. Then

$$\sqrt{\det(A^T A) \det(B^T B)} = |\det [A \; B]|$$

where $[A \; B]$ is the $p \times p$ matrix whose first k columns are those of A and final $p-k$ columns are those of B.

Proof. The statement is equivalent to

$$\det [A \; B]^T [A \; B] = \det [A \; B]^2$$

We rewrite the right-hand side using elementary properties of the determinant:

$$\begin{aligned}\det\begin{bmatrix}A & B\end{bmatrix}^2 &= \det\begin{bmatrix}A & B\end{bmatrix}\det\begin{bmatrix}A & B\end{bmatrix} = \det\begin{bmatrix}A & B\end{bmatrix}^T\det\begin{bmatrix}A & B\end{bmatrix}\\ &= \det\begin{bmatrix}A^T\\ B^T\end{bmatrix}\det\begin{bmatrix}A & B\end{bmatrix} = \det\begin{bmatrix}A^T\\ B^T\end{bmatrix}\begin{bmatrix}A & B\end{bmatrix}\\ &= \det\begin{bmatrix}A^TA & A^TB\\ B^TA & B^TB\end{bmatrix}\end{aligned}$$

Note that by assumption, $A^TB = 0_{k\times(p-k)}$ and $B^TA = \left(A^TB\right)^T = 0_{(p-k)\times k}$. The result then follows from the formulas for block determinants; see e.g. [75].

$$\det\begin{bmatrix}A & B\end{bmatrix}^2 = \det\begin{bmatrix}A^TA & 0_{k\times(p-k)}\\ 0_{(p-k)\times k} & B^TB\end{bmatrix} = \det A^TA\,\det B^TB$$

□

Theorem 5.1.1 (Disintegration over co-Foliations). Let $\rho(q)$ be a Lipschitz-continuous probability density on a convex open subset $\Omega \subset \mathbb{R}^p$, and let $A \in \mathbb{R}^{p\times l}$ and $B \in \mathbb{R}^{p\times(p-l)}$ be full-rank matrices with $A^TB = 0_{l\times(p-l)}$. For $b \in B^T\Omega$, define the measure m_b on Ω such that for a Borel set $U \subset \mathbb{R}^p$,

$$m_b(U) := \int_{U\cap F_E^{-1}(\{b\})} \rho(q)\,\frac{1}{\sqrt{\det(B^TB)}}\,\mathcal{H}^l(dq) \tag{5.1.2}$$

$$= \int_{A^T(U\cap F_E^{-1}(\{b\}))} \rho(Aa+Bb)\,\det\begin{bmatrix}A & B\end{bmatrix}\,\mathcal{I}^l(da) \tag{5.1.3}$$

where $\mathcal{H}^l$ is the l-dimensional Hausdorff measure on $\mathbb{R}^p$ and $\mathcal{I}^l$ is the l-dimensional Lebesgue measure on $\mathbb{R}^l$. Then the measures $\{m_b\}$ form a disintegration of the target measure $\rho(q)\mathcal{I}^p(dq)$ over the fiber bundle $F_E : \Omega \to B^T\Omega$, $F_E(q) = B^Tq$.

Proof. First we check the conditions for (5.1.2) to be a disintegration. Since the integrand in (5.1.2) is measurable and the Hausdorff measure is σ-finite, it follows that m_b is σ-finite and $\int_{\mathbb{R}^p} h\,dm_b$ is a measurable function of b for any measurable h. Furthermore, since the domain of integration in (5.1.2) is $U \cap F_E^{-1}(\{b\})$, then m_b must concentrate on the fiber $F_E^{-1}(\{b\})$. Thus the first two conditions in Definition 4.2.1 are satisfied.

The third condition for $\{m_b\}$ to be a disintegration is guaranteed by the co-area formula, stated in appendix B as Theorem B.0.3.

Note that $DF_E = B^T = B^TB\left(B^TB\right)^{-1}B^T$, and hence the $(p-k)$-dimensional Jacobian of F_E is given by

$$\sqrt{\det(DF\,DF^T)} = \sqrt{\det\left(B^TB\right)^{-1}B^TB\left(B^TB\right)^{-T}}$$

$$= \sqrt{\det\left(B^TB\right)^{-T}} = \frac{1}{\sqrt{\det(B^TB)}}$$

Thus applying the co-area formula given in Theorem B.0.3 to $h(q) := p(q)\sqrt{\det(B^T B)}$,

$$\begin{aligned}
&\int_{R^{p-k}} \int_{F_E^{-1}} p(q)\sqrt{\det(B^T B)}\, H^k(dq)\, l^{p-k}(db) \\
&= \int_{R^p} p(q)\sqrt{\det(B^T B)}\frac{1}{\sqrt{\det(B^T B)}} l^p(dq) \\
&= \int_{R^p} p(q) l^p(dq)
\end{aligned}$$

Equation (5.1.3) follows from the area formula, stated in Theorem B.0.1. For each $b \in R^{p-k}$, define a co-ordinate embedding $L_b(b) = Aa + Bb$, so that the differential matrix of L_b is given by $DL_b = A$ and the k-dimensional Jacobian is $J_k L_b(a) = \sqrt{\det(A^T A)}$. The preimage of $U \cap F_E^{-1}(\{b\})$ under L is given by $A \quad U \cap F_E^{-1}(\{b\}) \quad .$ The area formula with $h(q) = p(q)\sqrt{\det(B^T B)}$ then yields

$$\begin{aligned}
&\int_{U \cap F_E^{-1}(\{b\})} p(q)\sqrt{\det(B^T B)}\, H^k(dq) \\
&= \int_{A\,(U \cap F_E^{-1}(\{b\}))} p(Aa + Bb)\sqrt{\det(B^T B)}\, J_k L_b(a)\, l^k(da) \\
&= \int_{A\,(U \cap F_E^{-1}(\{b\}))} p(Aa + Bb)\sqrt{\det(A^T A)\det(B^T B)}\, l^k(da) \\
&= \int_{A\,(U \cap F_E^{-1}(\{b\}))} p(Aa + Bb)\,\det\ A \quad B\ \, l^k(da)
\end{aligned}$$

where in the last equality we have used Lemma B.0.5. □

Theorem 5.1.2 (Disintegration over m-Foliations). Let $p(q)$ be a Lipschitz-continuous probability density on an open convex subset $\Omega \subseteq R^p$, A and B be matrices as in Theorem 5.1.1, and $X = r\,F(\Omega)$. For $d \in A\,X$, define the probability measure m_d on Ω such that for a Borel $U \subseteq R^p$,

$$m_d(U) = \int_{U \cap F_M^{-1}(\{d\})} p(q) \frac{r}{\sqrt{\det\ A^T A}} H^{p-l}(dq) \qquad (5.1.4)$$

where H^{p-l} is the $(p-l)$-dimensional Hausdorff measure on R^p, $G(q) := r^2 F(q)$ and $F_M : R^p \to R^l$, $F_M(q) = A\,r\,F(q)$ for a smooth convex function $F : R^p \to R$. Then the measures $\{m_d\}$ form a disintegration of the target measure $p(q)l^p(dq)$ over the fibers of F_M. Furthermore, we have

$$m_d(U) = \int_{B\,(U \cap F_M^{-1}(\{d\}))} p(L_d(b))\ \frac{\det\ A \quad B}{\det(A\ G(L(b))\ A)}\ l^{p-k}(db) \qquad (5.1.5)$$

where l^{p-k} is the $(p-k)$-dimensional Lebesgue measure on R^{p-k} and L_d is the primal embedding of $F_M^{-1}(\{d\})$.

Proof. We first show $\{m_d\}$ is a disintegration. The first two conditions in Definition 4.2.1 are satisfied by the same arguments in the proof of Theorem 5.1.1

For the third condition, the differential matrix of F_M is

$$DF_M(q) = A\,\nabla^2 F(q) = A\,G(q)$$

and hence the $(p-k)$-dimensional Jacobian of F_M is

$$\begin{aligned} J_{p-k}F_M(q) &= \sqrt{\det\, DF_M(q)DF_M(q)^T} = \sqrt{\det\, A\,G(q)G(q)(A)^T} \\ &= \sqrt{\det\, (A^TA)^{-1}A^TG(q)G(q)A(A^TA)^{-T}} \\ &= \sqrt{\det\, A^TG(q)^2A}\ \det\, (A^TA)^{-\frac{1}{2}} \\ &= \frac{\sqrt{\det\, A^TG(q)^2A}}{|\det(A^TA)|} \end{aligned}$$

where, to go from the second to the third line, we have used the fact that A^TA and A^TG^2A are both $k \times k$ square matrices. By the co-area formula with $h(q) := \frac{\rho(q)}{J_{p-k}(q)}$, we then have

$$\begin{aligned} \int_{\mathbb{R}^k}\int_{F_M^{-1}(d)} \frac{\rho(q)}{J_{p-k}F_M(q)}\mathcal{H}^{p-k}(dq)\ \lambda^k(dd) &= \int_{\mathbb{R}^p} \frac{\rho(q)}{J_{p-k}F_M(q)}J_{p-k}F_M(q)\lambda(dq) \\ &= \int_{\mathbb{R}^p}\rho(q)\lambda(dq) \end{aligned}$$

To show (5.1.5), we use the area formula. The parameterization of the level sets of F_M is the co-ordinate embedding function $L_d(b)$ given in (3.3.20) with differential matrix from Lemma 3.3.10 as

$$DL_d(b) = \begin{bmatrix} A & A^TGA \end{bmatrix}^{-1}\begin{bmatrix} A^TGB \\ +B \end{bmatrix}$$

where we use the shorthand $G := G(L_d(b))$. The k-dimensional Jacobian of L_d is calculated as

$$\begin{aligned} \sqrt{\det\, DL_d^TDL_d} &= \sqrt{\det \begin{bmatrix} B^T & B^TGA(A^TGA)^{-1}A^T \end{bmatrix}\begin{bmatrix} B \\ A(A^TGA)^{-1}A^TGB \end{bmatrix}} \\ &= \sqrt{\det\, B^TB + B^TGA(A^TGA)^{-T}A^TA(A^TGA)^{-1}A^TGB} \end{aligned}$$

where dependence on b has been omitted for brevity, and we have used the mutual orthogonality of the columns of A and B, i.e. A^TB and B^TA are zero matrices. Define a matrix $V := \left(A^TGA\right)^{-T}A^TA\left(A^TGA\right)^{-1}$, so that

$$\sqrt{\det\, DL_d^TDL_d} = \sqrt{\det\,(B^TB + B^TGAVA^TGB)}$$

Then the area formula with $h(L_d(b)) := p(L_d(b)) \det A^TA \det A^TG(L_d(b))^2A^{1/2}$ yields

$$\begin{aligned} m_d(U) &= \int_{U \setminus F_M^{-1}(f\,dg)} p(L_d(b)) \det A^TA \det A^TG^2A^{1/2} H^{p-k}(dL_d(b)) \\ &= \int_{U \setminus F_M^{-1}(f\,dg)} p(L_d(b)) \det A^TA \sqrt{\frac{\det(B^TB + B^TGAVA^TGB)}{\det(A^TG^2A)}} I^{p-k}(db) \end{aligned} \tag{B.0.2}$$

Then (5.1.5) follows if we can show the integrand in the previous equation is equal to the integrand in (5.1.5) for all b, i.e.

$$\begin{aligned} \det A^TA \sqrt{\frac{\det(B^TB + B^TGAVA^TGB)}{\det(A^TG^2A)}} &= \frac{\det A\ B}{\det(A\ GA)} \\ , \quad \det A^TA \frac{\det B^TB + B^TGAVA^TGB}{\det(A^TG^2A)} &= \left(\frac{\det A\ B}{\det(A\ GA)}\right)^2 \\ , \quad \det A^TA \frac{\det B^TB + B^TGAVA^TGB}{\det(A^TG^2A)} &= \frac{\det(B^TB)}{\det(A\ GA)^2} \\ , \quad \det A^TA \det(A\ GA)^2 \det B^TB + B^TGAVA^TGB &= \det A^TG^2A \det(B^TB) \end{aligned} \tag{B.0.3}$$

where, to go from the second to the third equivalent statement, we have applied Lemma B.0.5. Our objective then is to show that (B.0.3) is true for all matrices A, B satisfying the conditions of Lemma B.0.5, and any symmetric positive-de nite $G \in \mathbb{R}^{p\ p}$, and where $V := A^TGA^{\ T}A^TA\ A^TGA^{\ 1}$.

Firstly, consider the left-hand side of (B.0.3). Note that $A\ GA$, A^TA and V are all $k\ k$ non-singular matrices. Thus by the elementary properties of the determinant,

$$\begin{aligned} \det A^TA \det(A\ GA)^2 &= \det A^TA \det A^TA^{\ 1}A^TGA^{\ 2} \\ &= \det A^TA\ A^TA^{\ 1}A^TGA\ A^TA^{\ 1}A^TGA \\ &= \det A^TGA\ A^TA^{\ 1}A^TGA = \det V^{\ 1} \end{aligned}$$

Finally, writing the left-hand side of (B.0.3) as the determinant of a block matrix,

$$
\begin{aligned}
&\det\ V^{-1}\ \det\ B^T B + B^T G A V A^T G B \\
&= (-1)^k \det\ V^{-1}\ \det\ B^T B + B^T G A V A^T G B \\
&= (-1)^k \det \begin{matrix} V^{-1} & A^T G B \\ B^T G A & B^T B \end{matrix} \\
&= (-1)^k \det\ B^T B\ \det\ V^{-1} - A^T G B\ B^T B\ ^{-1} B^T G A \\
&= (-1)^k\ (-1)^k \det\ B^T B\ \det\ V^{-1} + A^T G B\ B^T B\ ^{-1} B^T G A \\
&= \det\ B^T B\ \det\ A^T G A\ A^T A\ ^{-1} A^T G A + A^T G B\ B^T B\ ^{-1} B^T G A \\
&= \det\ B^T B\ \det\ A^T G\ A\ A^T A\ ^{-1} A^T + B\ B^T B\ ^{-1} B^T\ G A \\
&= \det\ B^T B\ \det\ A^T G\ AA^{\dagger} + BB^{\dagger}\ G A \\
&= \det\ B^T B\ \det\ A^T G^2 A
\end{aligned}
$$

which is the right-hand side of (B.0.3). Note that the last equality follows because $AA^{\dagger} + BB^{\dagger}$ is the $p \times p$ identity matrix; decomposing any $q \in \mathbb{R}^p$ as $q = Aa + Bb$, we see that $AA^{\dagger} + BB^{\dagger}$ is the identity map,

$$
\begin{aligned}
AA^{\dagger} + BB^{\dagger}\ q &= AA^{\dagger} + BB^{\dagger}\ Aa + Bb \\
&= A\ A^T A\ ^{-1} A^T + B\ B^T B\ ^{-1} B^T\ Aa + Bb \\
&= A\ A^T A\ ^{-1} A^T Aa + B\ B^T B\ ^{-1} B^T Bb \\
&= Aa + Bb = q
\end{aligned}
$$

In conclusion, equation (B.0.3) is true. It follows that the integrand in the area formula (B.0.2) equals

$$p\ L_d(b)\ \frac{\det\ A\ \ B}{\det(A\ GA)}$$

and hence the result (5.1.5) holds. □

Appendix C

Translating Recursion into Iteration

The purpose of this appendix is to show that Algorithm 4, the general recursive dual complement sampler, can be written into a general iterative form, and from this derive the iterative Orthogonal Gibbs and Orthogonal Gradient samplers.

We restate algorithm 4 below.

C.1 General Iterative Algorithm

Intuitively, one can imagine iterating over the submanifolds on the left-hand branches of gure 5.1.2 rather than recursing on the right-hand branches. Each of these submanifolds is either e- at within Q, or e- at within an m- at submanifold of Q. The idea behind the iterative algorithm is to use the map L_d given in de nition 3.3.13 to work with e- at submanifolds instead.

To develop intuition for this iterative process, consider stepping through Algorithm 4. Lines 5-6 resample the target distribution on an e- at submanifold $E_{A,b}$, yielding an updated sample

$$q^0 = Aa^0 + b \tag{C.1.1}$$

We use tildes to represent variables in the rst level of recursion in the algorithm; within the rst recursive call, the BASISVECTORS function is evaluated on line 2 with a p l dimensional input $q \quad b^0$ and inherited generator $F \quad \tilde{H}$, yielding a matrix $A \in R^{(p\ l)\ l}$. On lines 5-6 within the same recursive call, the target density $p = p_{d^0}$ is resampled along a submanifold

$$b = Aa + b \tag{C.1.2}$$

parameterized by $a \in R^l$, and $b = b^0 \quad AA\ b^0$. Here b are co-ordinates for the dual complement submanifold, but we may also view b as coef cients of $B \in R^{p\ p\ l}$, the complement basis to A, such that b are co-ordinates for the e- at submanifold spanned by the columns of B in primal co-ordinates. With this identi cation, the resampling step (C.1.2) within the rst recursive call is

$$Bb = BAa + Bb$$

Algorithm 4 General Recursive Dual Complement Submanifold Algorithm

```
function RecDCSSampler(q, F, log p)
    A ← BasisVectors(q, F)
    a ← A^T q
    b ← q − A a^(0)
    a' ← SubSampler(a ↦ log p(Aa + b), a)
    q' ← A a' + b
    if Dimension(q) == 1 then
        q'' ← q'
        return q''
    end if
    d' ← A^T r F(q')
    B ← ComplementBasis(A)
    function F#(b)
        q ← MFlatEmbed(b, F, d, A, B, a')
        return F(q) − d'^T A^T q
    end function
    function log p_{d'}(b)
        q ← MFlatEmbed(b, F, d, A, B, a')
        return log p(q) − log det A^T r ∇²F(q) A
    end function
    b' ← B^T q'
    b'' ← RecDCSSampler(b', F#, log p_{d'})
    q'' ← MFlatEmbed(b'', F, d, A, B, a')
    return q''
end function
```

By definition of $L_{d'}$, we have

$$L_{d'}(BAa + Bb) = L_{d'}(Aa + b)$$

Continuing the algorithm further, we must have $b = Bb$ where B is the result of COMPLEMENTBASIS(A). Thus, we can keep track of a variable b in an appropriate e- at submanifold of Q, which is mapped onto the corresponding m- at submanifold. This motivates the following recursive definition.

Definition C.1.1. The BASISVECTORS2 function is defined as follows for $q \in \mathbb{R}^p$, a Bregman generator $F : \mathbb{R}^p \to \mathbb{R}$ and $A^{(l)} \in \mathbb{R}^{p \times k}$ for $k \geq l$;

```
1: function BASISVECTORS2(q, F, A^(l))
2:     A ← A^(l)[:,1:l]                          . First 1,..,l columns of A^(l)
3:     B ← COMPLEMENTBASIS(A)
4:     d ← A^T ∇F(q)
5:     F(b) := F ∘ L_d(b) − d^T A^T L_d(b)
6:     if NCOLS(A^(l)) == l then
7:         R ← BASISVECTORS(B^T q, F)
8:     else
9:         R ← BASISVECTORS2(B^T q, F, B^T A^(l)[:, 1:l])
10:    end if
11:    return [B R]
12: end function
```

On line 5 in this definition, we have used the shorthand L_d for the evaluation of MFLATEMBED for this A, B, d, i.e. the primal embedding into $M_{B,Ad}$.

The general iterative algorithm is listed as Algorithm 9. The algorithm defines

$$R^{(1)} \leftarrow \text{BASISVECTORS}(q^{(0)}, F)$$
$$R^{(j)} \leftarrow \text{BASISVECTORS2}(q^{(j-1)}, F, A^{(j-1)}) \quad \text{for} \quad j = 2, .., p$$

C.2 Equivalence to Recursive Algorithm

In this section we will prove that the general iterative and recursive algorithms return the same final sample given the same input arguments and pseudorandom seed, which is stated as Theorem C.2.5. The proof rests on an inductive argument in which we replace the recursive call in algorithm 4 with a call to the iterative algorithm 9 to sample from the density $p_{d'}$, and by the inductive hypothesis, recover the same b'' as returned from the recursive call.

Throughout the section, we assume that the Newton optimization in MFLATEMBED and MFLATEMBED2 is solved exactly.

Algorithm 9 Iterative Dual Complementary Submanifold Algorithm

```
function SUBTARGET(a, R, d, A, c, F, log p, a_0)
    b^0 ← Ra + c
    q^0 ← MFLATEMBED2(F, b^0, d, A, a_0)
    return log p(q^0) − log det A^T ∇²F(q^0) A
end function

function ITERDCSSAMPLER(q^(0), F, log p)
    p ← DIMENSION(q^(0))
    n ← p / l                          . Calculate number of blocks. We assume k divides p.
    R^(1) ← BASISVECTORS(q^(0), F)
    a^(0) ← R^(1) q^(0)
    c^(1) ← q^(0) − R^(1) a^(0)
    a^(1) ← SUBSAMPLER(a ↦ log p(R^(1) a + c^(1)), a^(0))
    q^(1) ← R^(1) a^(1) + c^(1)
    b^(1) ← q^(1)
    A^(1) ← R^(1)
    for j = 2, .., n do
        R^(j) ← BASISVECTORS2(q^(j−1), F, A^(j−1))
        d^(j−1) ← A^(j−1) ∇F(q^(j−1))

        c^(j) ← (1_p − P) R^(j) R^(j)ᵀ c^(j−1)

        a_0^(j) ← A^(j−1)ᵀ q^(j−1)

        a_0^(j) ← R^(j)ᵀ q^(j−1)

        a^(j) ← SUBSAMPLER(a ↦ SUBTARGET(a, R^(j), d^(j−1), A^(j−1), c^(j), F, log p, a_0^(j)), a_0^(j))

        b^(j) ← R^(j) a^(j) + c^(j)                . Update b in the direction of R^(j)
        q^(j) ← MFLATEMBED2(F, b^(j), d^(j−1), A^(j−1), a_0^(j))

        A^(j) ← A^(j−1) R^(j)                . Concatenate R^(j) onto right side of A^(j−1)
    end for
    return q^(n)
end function
```

Lemma C.2.1. Let $d, \hat{d} \in \mathbb{R}^l$ and define

$$D = \begin{matrix} d \\ \hat{d} \end{matrix} \in \mathbb{R}^{2l}$$

Let $A \in \mathbb{R}^{p \times l}$ have full rank with orthonormal complementary basis $B \in \mathbb{R}^{p \times (p-l)}$. Let $A \in \mathbb{R}^{(p-l) \times k}$, and define

$$A' = \begin{matrix} A & BA \end{matrix} \in \mathbb{R}^{p \times (l+k)}$$

Let $F : \mathbb{R}^p \to \mathbb{R}$ be a convex Bregman generator, and define $\hat{F}$ to be the inherited Bregman generator $\hat{F}(b) := B^T \nabla F \circ L_d(b)$ for all b where L_d is the primal embedding into $M_{B,Ad}$. Let $b \in \mathbb{R}^{p-l}$ and $\hat{b} := Bb$. If

$$q_{A'} \quad \text{MFLATEMBED2}(F, b, D, A', a_0) \tag{C.2.1}$$

and

$$b_A \quad \text{MFLATEMBED2}(F, b, d, A, a_1) \tag{C.2.2}$$

$$q_A \quad \text{MFLATEMBED}(b_A, F, d, A, B, a_2) \tag{C.2.3}$$

where a_0, a_1, a_2 are starting values such that the Newton iteration in each function call converges, then

$$q_{A'} = q_A$$

Proof. Let $B \in \mathbb{R}^{(p-l) \times (p-l-k)}$ be a complement basis to A. Then $B' := BB \in \mathbb{R}^{p \times (p-l-k)}$ is a complement basis to $A' \in \mathbb{R}^{p \times (l+k)}$, since it clearly has full rank and

$$A'^T B' = \begin{matrix} A^T B' \\ A^T \hat{B}^T B' \end{matrix} = \begin{matrix} A^T BB \\ A^T \hat{B}^T BB \end{matrix} = \begin{matrix} 0_{l \times (p-2l)} \\ A^T B \end{matrix} = \begin{matrix} 0_{l \times (p-2l)} \\ 0_{l \times (p-2l)} \end{matrix} = 0_{(2l) \times (p-2l)}$$

We will show that the (A', B', F) generalized mixed co-ordinates of $q_{A'}$ and q_A are the same, and hence $q_{A'} = q_A$. By Definition 3.3.13 The generalized mixed co-ordinates for $q_{A'}$ are

$$\begin{pmatrix} (A')^T \nabla F(q_{A'}) \\ (B')^T q_{A'} \end{pmatrix} = \begin{matrix} D \\ \hat{B}^T \hat{B}^T b \end{matrix} = \begin{matrix} D \\ \hat{B}^T \hat{B}^T \hat{B} b \end{matrix} = \begin{matrix} D \\ \hat{B}^T b \end{matrix} \tag{C.2.4}$$

Consider (A', B', F)-mixed co-ordinates of q_A. For the dual component,

$$A'^T \nabla F(q_A) = \begin{matrix} A^T \nabla F(q_A) \\ A^T \hat{B}^T \nabla F(q_A) \end{matrix} \tag{C.2.5}$$

Clearly (C.2.3) is a primal embedding into the m-flat submanifold $M_{B,Ad}$, and hence the first block of the above is $A^T \nabla F(q_A) = d$. For the second component, recall that $\hat{B}^T \nabla F(L_d(b)) = \nabla_b \hat{F}(b)$, and thus

$$A^T \hat{B}^T \nabla F(q_A) = A^T \nabla_b \hat{F} \, b_A$$

Since (C.2.2) is a primal embedding into $M_{B,Ad}$, we therefore have

$$A^T B^T \nabla F(q_A) = d$$

Thus the dual component (C.2.5) of q_A agrees with the dual component of $q_{A'}$ from (C.2.4),

$$A'^T \nabla F(q_A) = \begin{matrix} d \\ d \end{matrix} = D$$

Finally, the primal component of q_A in the mixed co-ordinate system is given by

$$B'^T q_A = B^T B^T q_A = B^T b_A = B^T b$$

where the second equality follows from the fact that $q_A = L_d(b)$, and the final equality follows from (C.2.2). Thus q_A and $q_{A'}$ have the same (A', B', F) generalized mixed co-ordinates, and hence we must have $q_A = q_{A'}$. □

Lemma C.2.2. Let $A, \bar{A}, A', B, d, \bar{d}, D, F, \bar{F}$ be defined as in Lemma C.2.1. Let $c \in \mathbb{R}^{p-l}$ and let $R \in \mathbb{R}^{(p-l)\times l}$ have full rank and columns which are mutually Euclidean-orthogonal to the columns of A'. For a log-density $\log p$, let $\log p_d$ be defined as (5.1.6), i.e. for all $q \in \mathbb{R}^{p-l}$,

$$\log p_d(q) = \log p\big(L_d(q)\big) - \log\det\big(A^T \nabla^2 F\big(L_d(q)\big) A\big)$$

Then for $a \in \mathbb{R}^l$,

$$\textsc{SubTarget}(a, BR, D, A', Bc, F, \log p) = \textsc{SubTarget}(a, R, d, A, c, \bar{F}, \log p_d) \quad \text{(C.2.6)}$$

where SUBTARGET is the function defined on line of Algorithm 9.

Proof. Consider stepping through SUBTARGET with each set of arguments on either side of (C.2.6). We use tildes to refer to the execution of the function with the arguments on the right-hand side, and absence of tildes to indicate the execution of the function with arguments on the left-hand side of (C.2.6).

Firstly, from line 2 we have $b' = Ra + c$. Multiplying this by B, we find

$$Bb' = BR + Bc$$

The matrix BR and vector Bc are exactly the R and c argument passed to the function on the left-hand side of (C.2.6). Thus we have $b' = \tilde{b}'$.

Consider line 3 for each set of arguments. Since $b' = \tilde{b}'$, the conditions of Lemma C.2.1 are satisfied, and we identify $q_{A'}$ from the lemma with q' from the execution of SUBTARGET on the left-hand side of (C.2.6), and b_A from the lemma with $\tilde{q}'$ from the function called with the arguments on the right-hand side of the same equation.

To show the result (C.2.6), it remains to show that line 4 gives the same result for both sets of arguments, i.e.

$$\log p(q') - \log\det\big(A^T \nabla^2 F(q') A\big) = \log p_d(\tilde{q}') - \log\det\big(\bar{A}^T \nabla^2 \bar{F}(\tilde{q}') \bar{A}\big) \quad \text{(C.2.7)}$$

Since we made the identifications $q^0 \$ b_A$ and $q^0 \$ q_A$, it follows from Lemma C.2.1 that

$$L_d\ q^0 = q^0$$

where L_d is the primal embedding into $M_{B,Ad}$. For brevity, from here on we write G for $r^2F\ L_d(q^0)$. Expanding the definition of $\log p_d$ in the right-hand side of (C.2.7),

$$\begin{aligned}
&\log p_d(q^0) \quad \log\det A^T r^2 F A \\
&= \log p\ L_d(q^0) \quad \log\det A^T G A \quad \log\det A^T r^2 F(q^0) A \\
&= \log p\ L_d(q^0) \quad \log\left[\det A^T G A \det A^T r^2 F(q^0) A\right] \qquad \text{(C.2.8)}
\end{aligned}$$

Note by Theorem 3.3.12, we have the expression for r^2F,

$$r^2F(q^0) = B^T G\ 1_p\ _p \quad A\ A^T G A \quad {}^1 A^T G\ B$$

and thus the determinant in (C.2.8) can be written as

$$\begin{aligned}
&\det A^T G A \det A^T r^2 F(q^0) A \\
&= \det A^T G A \det A^T B^T G\ 1_p\ _p \quad A\ A^T G A \quad {}^1 A^T G\ B A \\
&= \det A^T G A \det A^T B^T G B A \quad A^T B^T G A\ A^T G A \quad {}^1 A^T G B A \\
&= \begin{vmatrix} A^T G A & A^T G B A \\ A^T B^T G A & A^T B^T G B A \end{vmatrix} \\
&= A\ B A\ {}^T G\ A\ B A \\
&= A^{0\ T} G A^0
\end{aligned}$$

where, to go from the third to the fourth line, we have written a product of determinants as a determinant of a block matrix using the Schur complement property [75]. Thus it follows from (C.2.8) that

$$\log p_d(q^0) \quad \log\det A^T r^2 F A = \log p\left(L_d(q^0)\right) \quad A^{0\ T} r^2 F\ L_d(q^0)\ A^0$$

and hence (C.2.7) holds. Thus the returned results for both sets of arguments to the SUBTARGET function are equal. □

Lemma C.2.3. Suppose $B 2 R^{p\ (p\ l)}$ and $R 2 R^{(p\ l)\ l}$ are matrices of full rank where the columns of B are orthonormal, and let $R = BR$. Then

$$R = R\ B^T$$

Proof. By the orthonormality of B, the pseudo-inverse of R is

$$\begin{aligned} R^{\dagger} &= BR^{\dagger} = R^T B^T BR^{\dagger} R^T B^T \\ &= R^T R^{\dagger} R^T B^T = R^{\dagger} B^T \end{aligned} \tag{C.2.9}$$

□

Lemma C.2.4. Let $q^{(0)}, .., q^{(n)}$ be the iterates of algorithm 9 with arguments $q^{(0)}$, $\log p$ and F where $p > 1$ so that $n > 1$, and let $z^{(1)}$ represent the state of the pseudorandom generator after the first call to the subsampler in line 13. Let A be $A^{(1)}$ defined in line 10 for this $q^{(0)}$, and let B be the result of COMPLEMENT BASIS(A). Set $d = A^{\dagger} \nabla F\left(q^{(1)}\right)$. Define $\tilde{q}^{(0)} = B^T q^{(1)}$, and let $\tilde{q}^{(0)}, \tilde{q}^{(1)}, \ldots, \tilde{q}^{(n-1)}$ be the resulting iterates of algorithm 9 with initial iterate $\tilde{q}^{(0)}$ and target log-density and Bregman generator respectively given by

$$\begin{aligned} \log \tilde{p}_d(q) &:= \log p\left(L_d(q)\right) - \log\det\left(A^T \nabla^2 F(L_d(q))\, A\right) \\ \tilde{F}(q) &:= F(L_d(q)) - d^T A^T L_d(q) \end{aligned}$$

Assume that immediately before executing the algorithm with these arguments, the state of the pseudorandom generator is $\tilde{z}^{(0)} = z^{(1)}$. Then for $j = 0, .., n-1$, we have

$$L_d\left(\tilde{q}^{(j)}\right) = q^{(j+1)} \tag{C.2.10}$$

where L_d is the primal embedding into $M_{B,Ad}$. Moreover,

$$R^{(j+1)} = BR^{(j)} \tag{C.2.11}$$

for $j = 1, .., n-1$, where $R^{(j)}$ is the $R^{(j)}$ matrix from Algorithm 9 corresponding to the iterates $q^{(0)}, .., q^{(n-1)}$.

Proof. Throughout the proof, we will use tildes to refer to variables in the execution of algorithm 9 with iterates $\tilde{q}^{(0)}, .., \tilde{q}^{(n-1)}$, and the abscence of tildes to indicate the execution of the algorithm with iterates $q^{(0)}, q^{(1)}, .. q^{(n)}$. We will show by induction that each of the following holds for $j = 1, .., n-1$,

$$R^{(j+1)} = BR^{(j)} \tag{C.2.12}$$

$$A^{(j+1)} = A - BA^{(j)} \tag{C.2.13}$$

$$c^{(j+1)} = Bc^{(j)} \tag{C.2.14}$$

$$b^{(j+1)} = Bb^{(j)} \tag{C.2.15}$$

$$a^{(j+1)} = a^{(j)} \tag{C.2.16}$$

where $R^{(1)} := A^{(1)}$. In addition, we will show the following hold for $j = 0, 1, .., n - 1$,

$$z^{(j+1)} = z^{(j)} \tag{C.2.17}$$

$$A \; r \; F \; q^{(j+1)} = d \tag{C.2.18}$$

$$L_d \; q^{(j)} = q^{(j+1)} \tag{C.2.19}$$

where $z^{(j)}$ is the state of the pseudorandom generator after j calls to the subsampler. Note that the line (C.2.12) corresponds to the result (C.2.11), and the final line (C.2.19) is the result (C.2.10) of the Lemma.

Base case. Since $d = A \; r \; F(q^{(1)})$, then (C.2.18) holds trivially for $j = 0$ and by definition $q^{(1)} \in M_{B,Ad}$. Thus the (A, B) generalized mixed co-ordinates of $q^{(1)}$ are

$$\begin{matrix} d \\ B^T q^{(1)} \end{matrix} = \begin{matrix} d \\ q^{(0)} \end{matrix}$$

by definition of $q^{(0)}$. Since L_d is the primal embedding into $M_{B,Ad}$, the generalized mixed co-ordinates of $L_d \; q^{(0)}$ agree with the above; thus $L_d \; q^{(0)} = q^{(1)}$ and (C.2.19) holds for $j = 0$. Finally, by the assumptions of the lemma we have $z^{(1)} = z^{(0)}$, and hence (C.2.17) is trivially true for $j = 0$.

We show directly that equations (C.2.12)-(C.2.19) hold for $j = 1$. For (C.2.12), note that by line 10 we have

$$R^{(1)} = \textsc{BasisVectors}(q^{(0)}, F) \tag{C.2.20}$$

Furthermore, by line 18

$$R^{(2)} = \textsc{BasisVectors2}(q^{(1)}, F, A^{(1)}) = \textsc{BasisVectors2}(q^{(1)}, F, R^{(1)})$$

Since $A^{(1)} = R^{(1)}$ has l columns, the if-statement on line 6 in Definition C.1.1 is triggered, and hence

$$\begin{aligned} R^{(2)} &= B \; \textsc{BasisVectors}(B^T q^{(1)}, F) \\ &= B \; \textsc{BasisVectors}(q^{(0)}, F) \\ &= B R^{(1)} \end{aligned}$$

by (C.2.20), where $B = \textsc{Complement Basis}(A)$. Thus (C.2.12) holds for $j = 1$. This result, with the fact that $A^{(1)} \quad A$ and line 28, implies

$$A^{(2)} = A \; R^{(2)} = A \; B R^{(1)} = A \; B A^{(1)}$$

which shows (C.2.13) for $j = 1$.

To show (C.2.14) for $j = 1$, we manipulate the expression on line 21,

$$c^{(2)} = 1_p \; _p \; R^{(2)} \; R^{(2)} \quad c^{(1)}$$

Substituting the identity $1_{p\ p} = AA + BB^T$ and $R^{(2)} = BR^{(1)}$ from (C.2.12),

$$\begin{aligned} c^{(2)} &= AA + BB^T \quad BR^{(1)} \quad BR^{(1)} \quad c^{(1)} \\ &= BB^T \quad BR^{(1)} \quad R^{(1)} \quad B^T \quad c^{(1)} \\ &= B \quad 1_{(p\ l)\ (p\ l)} \quad R^{(1)} \quad R^{(1)} \quad B^T c^{(1)} \end{aligned} \tag{C.2.21}$$

where we have used the fact that $c^{(1)} = q^{(0)} \quad AA \quad q^{(0)}$, and thus $c^{(1)} \ 2 \ (\mathrm{Im}\, A)^?$. By line 14, we have

$$B^T c^{(1)} = B^T \quad q^{(1)} \quad R^{(1)} a^{(1)} = B^T q^{(1)} = q^{(0)}$$

since B is a complement basis to $R^{(1)} = A$. Substituting this into (C.2.21);

$$c^{(2)} = B \quad 1_{(p\ l)\ (p\ l)} \quad R^{(1)} \quad R^{(1)} \quad q^{(0)} = Bc^{(1)}$$

Thus (C.2.14) holds for $j = 1$.

Consider the call to SUBTARGET on line 24 returning $a^{(2)}$, i.e. with the arguments

$$a^{(2)} \quad \text{SUBSAMPLER}(a\ 7!\ \text{SUBTARGET} \quad a, R^{(2)}, d^{(1)}, A^{(1)}, c^{(2)}, F, \log p, a_0^{(2)} \quad , a_0^{(2)})$$

Then b^0 on line 2 within the function call with these arguments is

$$b^0 = R^{(2)} a + c^{(2)} = B \quad R^{(1)} a + c^{(1)}$$

Therefore q^0 on line 3 satis es

$$q^0 = L_d \quad b^0 = L_d \quad B^T b^0 = L_d \quad R^{(1)} a + c^{(1)}$$

since $d = d^{(1)}$. The returned value of the log-density of the subtarget is then

$$\begin{aligned} &\log p \quad L_d \quad B^T b^0 \quad \log \det \quad A^{(1)} \quad {}^T r \ {}^2 F \quad B^T b^0 \quad A^{(1)} \\ &= \log p \quad L_d \quad B^T b^0 \quad \log \det \quad A^T r \ {}^2 F \quad B^T b^0 \quad A \\ &= \log p_d \quad B^T b^0 \\ &= \log p_d \quad R^{(1)} a + c^{(1)} \end{aligned}$$

Note that this is exactly the same log-density passed to the subsampler in line 13 when the algorithm I TERDCSSAMPLER is passed a target log-density $\log p_d$. By assumption,

the state of the pseudorandom generator is the same before each call to the subsampler, i.e. $z^{(1)} = z^{(0)}$, and the subsampler is passed the same initial iterate for a since

$$a^{(0)} = \quad R^{(1)} \quad q^{(0)} = \quad R^{(1)} \quad B^T q^{(1)} = \quad R^{(2)} \quad q^{(1)} = a_0^{(2)}$$

Thus the subsampler is passed the same target log-density, initial iterate and pseudo-random seed, and hence we must have $a^{(2)} = a^{(1)}$, i.e. (C.2.16) holds for $j = 1$. The pseudorandom state must also be updated in the same way, such that $z^{(2)} = z^{(1)}$ and thus (C.2.17) is true for $j = 1$. Furthermore, (C.2.15) holds for $j = 1$, since lines 25, 14 and 15 imply

$$b^{(2)} = R^{(2)} a^{(2)} + c^{(2)} = B \quad R^{(1)} a^{(1)} + c^{(1)} \quad = Bq^{(1)} = Bb^{(1)}$$

By the properites of the primal embedding, (C.2.19) holds for $j = 1$;

$$\begin{aligned} q^{(2)} &= L_d \quad b^{(2)} \quad = L_d \quad B^T b^{(2)} \\ &= L_d \quad R^{(1)} a^{(1)} + c^{(1)} \quad = L_d \quad q^{(1)} \end{aligned}$$

Since $q^{(2)}$ is in the image of the primal embedding into $M_{B,Ad}$, then (C.2.18) follows by de nition for $j = 1$;

$$A \; r \; F \quad q^{(2)} \quad = d$$

Thus (C.2.12)-(C.2.19) all hold for $j = 1$.

Induction step. Assume the inductive hypothesis that (C.2.12)-(C.2.19) all hold for some $j = j_0 \quad 1$, and consider each equation in the case $j_0 + 1$.

Firstly consider (C.2.12). From line 18 with $j = j_0 + 2$,

$$R^{(j_0+2)} = \textsc{BasisVectors2}(q^{(j_0+1)}, F, A^{(j_0+1)}) \tag{C.2.22}$$

By the inductive hypothesis (C.2.13) holds for $j = j_0$, and thus the submatrix of $A^{(j_0+1)}$ with the rst l columns removed is $BA^{(j_0)}$;

$$\begin{aligned} A^{(j_0+1)} &= \quad A \quad BA^{(j_0)} \\ ) \quad A^{(j_0+1)}_{[:,\ 1:l]} &= BA^{(j_0)} \end{aligned}$$

The columns of A are rst l columns of $A^{(j_0+1)}$. Thus by De nition C.1.1, $R^{(j_0+2)} = BR$, where B is the result of $\textsc{Complement Basis}(A)$, and

$$\begin{aligned} R &:= \textsc{BasisVectors2}(B^T q^{(j_0+1)}, F, A^{(j_0+1)}_{[:,\ 1:l]}) \\ &= \textsc{BasisVectors2}(B^T q^{(j_0+1)}, F, B^T BA^{(j_0)}) \\ &= \textsc{BasisVectors2}(B^T q^{(j_0+1)}, F, A^{(j_0)}) \end{aligned}$$

By the inductive hypothesis, $L_d\ q^{(j_0)} = q^{(j_0+1)}$ holds for $j = j_0$, and thus $q^{(j_0)} = B^T q^{(j_0+1)}$ by De nition 3.3.8. It follows that

$$R = \text{BASISVECTORS2}(q^{(j_0)}, F, A^{(j_0)}) = R^{(j_0+1)}$$

Substituting this into equation (C.2.22) yields $R^{(j_0+2)} = BR^{(j_0+1)}$, and thus (C.2.12) holds for $j = j_0 + 1$.

Next consider (C.2.13). By line 28 for $j = j_0 + 1$ and the inductive hypothesis, we have

$$A^{(j_0+2)} = A^{(j_0+1)}\ R^{(j_0+2)} = A\ BA^{(j_0)}\ R^{(j_0+2)}$$

Since (C.2.12) is proven for $j = j_0 + 1$, we have $R^{(j_0+2)} = BR^{(j_0+1)}$. Substituting this and applying line 28 again for $A^{(j_0+1)}$, we have

$$A^{(j_0+2)} = A\ BA^{(j_0)}\ BR^{(j_0+1)} = A\ BA^{(j_0+1)}$$

and thus (C.2.13) holds for $j = j_0 + 1$.

Next we show (C.2.14) holds for $j = j_0 + 1$. From line 21, we have

$$c^{(j_0+2)} = 1_{p\ p}\ R^{(j_0+2)}\ R^{(j_0+2)}\ c^{(j_0+1)} \tag{C.2.23}$$

Note that (C.2.12) holds for $j = 1, .., j_0, j_0 + 1$, so we substitute $R^{(j_0+1)} = BR^{(j_0)}$ and $R^{(j_0+2)} = BR^{(j_0+1)}$ into (C.2.23), applying Lemma C.2.3, and with the identity $1_{p\ p} = AA + BB$ we recover

$$c^{(j_0+2)} = AA + BB^T\ BR^{(j_0+1)}\ BR^{(j_0+1)}\ c^{(j_0+1)}$$

$$c^{(j_0+2)} = BB^T\ BR^{(j_0+1)}\ R^{(j_0+1)}\ B^T\ c^{(j_0+1)}$$

$$= B\ 1_{(p\ l)\ (p\ l)}\ R^{(j_0+1)}\ R^{(j_0+1)}\ B^T c^{(j_0+1)}$$

By the inductive hypothesis, $B^T c^{(j_0+1)} = B^T B c^{(j_0)} = c^{(j_0)}$ and hence

$$c^{(j_0+2)} = B\ 1_{(p\ l)\ (p\ l)}\ R^{(j_0+1)}\ R^{(j_0+1)}\ c^{(j_0)}$$

This is exactly the expression for $c^{(j_0+1)}$ from line 21 multiplied by B, so it follows $c^{(j_0+2)} = Bc^{(j_0+1)}$ and hence (C.2.14) holds for $j = j_0 + 1$.

To show (C.2.16) is true for $j = j_0 + 1$, rst substitute (C.2.13) for $j = j_0$ into the

expression on line 20 for $d^{(j_0+1)}$,

$$
\begin{aligned}
d^{(j_0+1)} &= A^{(j_0+1)} \; r_q F\left(q^{(j_0+1)}\right) \\
&= A\left(BA^{(j_0)}\right) \; r_q F\left(q^{(j_0+1)}\right) \\
&= \left(A\left(BA^{(j_0)}\right)^T A\left(BA^{(j_0)}\right)\right)^{-1} A\left(BA^{(j_0)}\right)^T r_q F\left(q^{(j_0+1)}\right) \\
&= \begin{pmatrix} A^TA & 0_{l\,(p-l)} \\ 0_{(p-l)\,l} & A^{(j_0)T}A^{(j_0)} \end{pmatrix}^{-1} \begin{pmatrix} A^T r_q F\left(q^{(j_0+1)}\right) \\ A^{(j_0)T}B^T r\, F\left(q^{(j_0+1)}\right) \end{pmatrix} = \begin{pmatrix} A\, r_q F\left(q^{(j_0+1)}\right) \\ A^{(j_0)}\, B^T r\, F\left(q^{(j_0+1)}\right) \end{pmatrix}
\end{aligned}
$$

By the inductive hypothesis, $q^{(j_0+1)} = L_d\, q^{(j_0)}$. Hence by the formula for the inherited generator in Theorem 3.3.12, we have $B^T r\, F\left(L_d\, q^{(j_0)}\right) = r\, F(q^{(j_0)})$. Thus

$$
d^{(j_0+1)} = \begin{pmatrix} A\, r\, F\left(q^{(j_0+1)}\right) \\ A^{(j_0)}\, r\, F\left(q^{(j_0)}\right) \end{pmatrix} \tag{C.2.24}
$$

By line 20, $A^{(j_0)}\, r\, F\left(q^{(j_0)}\right) = d^{(j_0)}$, and by (C.2.18) in the inductive hypothesis

$$
A\, r\, F\left(q^{(j_0+1)}\right) = d
$$

Substituting these into (C.2.24), we find

$$
d^{(j_0+1)} = \begin{pmatrix} d \\ d^{(j_0)} \end{pmatrix}
$$

Now, the conditions of Lemma C.2.2 are satisfied when we make the following identifications:

$$
\begin{array}{llll}
A^0 \$ A^{(j_0+1)} & R \$ R^{(j_0+1)} & R \$ R^{(j_0+2)} & A \$ A^{(j_0)} \\
D \$ d^{(j_0+1)} & d \$ d^{(j_0)} & c \$ c^{(j_0+1)} & c \$ c^{(j_0+2)}
\end{array}
$$

Then by Lemma C.2.2, for any a we have

$$
\begin{aligned}
&\textsc{SubTarget}\left(a, R^{(j_0+2)}, d^{(j_0+1)}, A^{(j_0+1)}, c^{(j_0+2)}, F, \log p\right) \\
&= \textsc{SubTarget}\left(a, R^{(j_0+1)}, d^{(j_0)}, A^{(j_0)}, c^{(j_0+1)}, F, \log p_d\right)
\end{aligned}
$$

Therefore the log-density passed to the subsampler on line 24 is the same function for iterations $j = j_0 + 1$ and $j = j_0 + 2$ of the for-loop for the variables with and without tildes respectively. Furthermore, the initial iterate passed to the subsampler $a_0^{j_0+1} = a_0^{j_0+2}$, since from line 23,

$$
\begin{aligned}
a_0^{(j_0+2)} &:= R^{(j_0+2)}\, q^{(j_0+1)} \\
&= R^{(j_0+1)}\, B^T L_d\, q^{(j_0)} = R^{(j_0+1)}\, q^{(j_0)} = a_0^{(j_0+1)}
\end{aligned}
$$

where we have used the property of the primal embedding into $M_{B,Ad}$, and Lemma C.2.3. Finally, by the inductive hypothesis (C.2.17) holds for $j = j_0$ and so $z^{(j_0+1)} = z^{(j_0)}$, and so the subsampler is called with the same pseudorandom seed or state. In conclusion, the subsampler is called in each case with the same pseudorandom state, target log-density, and initial iterate, and therefore the returned result is the same for each case;

$$a^{(j_0+2)} = a^{(j_0+1)}$$

which shows (C.2.16) holds for $j = j_0 + 1$. Additionally, the state of the pseudorandom generator must be updated in the same way, and thus (C.2.17) holds for $j = j_0 + 1$; $z^{(j_0+2)} = z^{(j_0+1)}$.

To show (C.2.15) for $j = j_0 + 1$, we substitute equations (C.2.14) and (C.2.12), both of which have been proven for $j = j_0 + 1$, into the expression on line 25;

$$\begin{aligned} b^{(j_0+2)} = R^{(j_0+2)}a^{(j_0+2)} + c^{(j_0+2)} &= BR^{(j_0+1)}a^{(j_0+1)} + Bc^{(j_0+1)} \\ &= B\ R^{(j_0+1)}a^{(j_0+1)} + c^{(j_0+1)} \\ &= Bb^{(j_0+1)} \end{aligned}$$

and hence (C.2.15) holds for $j = j_0 + 1$. Finally, the conditions of Lemma C.2.1 are satisfied with the above identifications and $b \$ b^{(j_0+1)}$. Line 26 takes the form in each case:

$$\begin{aligned} q^{(j_0+2)} &\quad \text{MFLATEMBED2}(F, b^{(j_0+2)}, d^{(j\ \ 1)}, A^{(j_0+1)}, a_0^{(j_0+2)}) \\ q^{(j_0+1)} &\quad \text{MFLATEMBED2}(F, b^{(j_0+1)}, d^{(j_0+1)}, A^{(j_0)}, a_0^{(j_0+1)}) \end{aligned}$$

It follows from Lemma C.2.1, specifically (C.2.3), that

$$q^{(j_0+2)} = L_d\ \ q^{(j_0+1)}$$

and hence (C.2.19) is proven for $j = j_0 + 1$. Thus $q^{(j_0+2)}$ is an element of the submanifold $M_{B,Ad}$, and hence (C.2.18) holds for $j = j_0 + 2$,

$$A\ r\ F\ \ q^{(j_0+2)}\ \ = d$$

We have shown that (C.2.12)-(C.2.19) hold for $j = j_0 + 1$, assuming they hold for $j = j_0$. By the principle of induction, (C.2.12)-(C.2.19) hold for all $j = 1, .., n\ \ 1$, and in particular (C.2.19) holds for $j = 0, .., n\ \ 1$. □

Theorem C.2.5. Suppose algorithms 4 and 9 are initialized with the same pseudorandom seed, and the same target distribution p, Bregman generator F and initial iterate $q^{(0)}$ are passed as arguments to both algorithms. Suppose also that the Newton optimization in calls to algorithm 2 is replaced with the exact minimization. Then the algorithms produce the same output, i.e.

$$\text{RECDCSSAMPLER}(q^{(0)}, F, \log p) = \text{ITERDCSSAMPLER}(q^{(0)}, F, \log p) \qquad \text{(C.2.25)}$$

Proof. For clarity, we label the iterates of algorithm 9 as $q_{\mathrm{IT}}^{(1)}, .., q_{\mathrm{IT}}^{(n)}$, and let q^0 and q^{00} refer to these variables in algorithm 4. We fix the block-size l and use an inductive argument on the number of blocks $n = p/l$.

Base case. Suppose $n = 1$, i.e. $p = l$. Note that A defined on line 2 of algorithm 4 is exactly $A^{(1)}$ defined on line 10 of algorithm 9. Furthermore, lines 3-6 of algorithm 4 are identical to lines 11-14 of algorithm 9 with $a \$ a^{(0)}$, $a^0 \$ a^{(1)}$, $b \$ b^{(1)}$ and $q^0 \$ q^{(1)}$. Since the subsampler on these lines is called with the same pseudorandom seed, we have $q_{\mathrm{IT}}^{(1)} = q^0$. In the recursive algorithm 4, the if-statement on line 7 is triggered, and the algorithm terminates and returns $q^{00} = q^0$. Similarly in algorithm 9, the for-loop on line 7 ranges from 2 to n, and thus is skipped since $n = 1$, and the algorithm returns $q_{\mathrm{IT}}^{(n)} = q_{\mathrm{IT}}^{(1)} = q^0$ on line 30. Thus (C.2.25) holds for inputs of size $p = l$.

Induction step. Assume an inductive hypothesis that (C.2.25) holds for inputs of size $p = n_0 l$ for some $n_0 \geq 1$ when the pseudorandom generator is initialized with the same state or seed for both algorithms.

Consider the execution of each algorithm with inputs of size $p = (n_0 + 1)l$. The same argument we used in the base case applies, and we have $q_{\mathrm{IT}}^{(1)} = q^0$. Furthermore, since the subsampler is called for the first time in each algorithm with identical inputs and the same pseudorandom seed, the underlying states of the pseudorandom generators used for each algorithm are identical after this first call to the subsampler.

In line 12 of algorithm 4, the matrix B is defined, which has dimension $p \times n_0 l$ since $p - l = n_0 l$. Thus the vector $b^0 = B^T q^0$ initialized in line 21 has dimension $n_0 l$, and the dynamically defined functions ∇f_{d^0} and $\log p_{d^0}$ both accept inputs of this dimension. Thus by the inductive hypothesis, the recursive call on line 22 is equivalent to calling the iterative algorithm, i.e. the line can be replaced with

$$b^{00} \leftarrow \textsc{IterDCSSampler}(B^T q_{\mathrm{IT}}^{(1)}, \nabla f_{d^0}, \log p_{d^0})$$

Within this inner call to the iterative algorithm, the iterates $b^{(0)} = B^T q_{\mathrm{IT}}^{(1)}, b^{(1)}, \ldots, b^{(n_0)}$ are produced, with the final iterate $b^{(n_0)}$ being returned and stored in b^{00}. Thus by Lemma C.2.4 with $n = n_0 + 1$, we have $q^{00} = L_{d^0}(b^{00}) = q_{\mathrm{IT}}^{(n_0+1)}$, and so both algorithms return the same value. Hence (C.2.25) holds for $n_0 + 1$. By the principle of induction, (C.2.25) holds for all $n \geq 1$. □

C.3 Explicit Algorithms

The following Lemma allows simplification of each specific algorithm.

Lemma C.3.1. If the matrices $R^{(j)}$ in Algorithm 9 are mutually orthogonal so that $(R^{(j_2)})^T R^{(j_1)} = 0_{l \times l}$ whenever $j_2 \neq j_1$, then the vector $d^{(j)}$ satisfies

$$d^{(j)} = \begin{pmatrix} d^{(j-1)} \\ R^{(j)} r F q^{(j)} \end{pmatrix} \quad \text{for } j = 2, .., n$$

Proof. By line 20, we have for each $j = 2, .., n$

$$d^{(j)} = A^{(j)} \; r \; F \; q^{(j)}$$

By line 28, $A^{(j)} = A^{(j-1)} \; R^{(j)}$, and thus

$$\begin{aligned} d^{(j)} &= \begin{pmatrix} A^{(j-1)^T} A^{(j-1)} & 0_{(j-1)l \times l} \\ 0_{l \times (j-1)l} & R^{(j)^T} R^{(j)} \end{pmatrix}^{-1} \begin{pmatrix} A^{(j-1)^T} r \; F \; q^{(j)} \\ R^{(j)^T} r \; F \; q^{(j)} \end{pmatrix} \\ &= \begin{pmatrix} A^{(j-1)} \; r \; F \; q^{(j)} \\ R^{(j)} \; r \; F \; q^{(j)} \end{pmatrix} \end{aligned} \tag{C.3.1}$$

By line 26, $q^{(j)}$ is an element of the submanifold

$$M_{B^{(j-1)}, A^{(j-1)} d^{(j-1)}} = \{ q \mid A^{(j-1)} \; r \; F(q) = d^{(j-1)} \}$$

and thus the result follows from (C.3.1);

$$d^{(j)} = \begin{pmatrix} A^{(j-1)} \; r \; F \; q^{(j)} \\ R^{(j)} \; r \; F \; q^{(j)} \end{pmatrix} = \begin{pmatrix} d^{(j-1)} \\ R^{(j)} \; r \; F \; q^{(j)} \end{pmatrix}$$

□

C.3.1 Orthogonal Gibbs

Note that BASISVECTORS2 simplifies in the special cases of the Orthogonal Gibbs and Orthogonal Gradient algorithm as the following propositions show; the general form in Definition C.1.1 allows us to reason about both cases.

Proposition C.3.2. Let BASISVECTORS and COMPLEMENTBASIS be as defined as in the Orthogonal Gibbs algorithm, i.e. by equations (5.2.1) and (5.2.2) respectively. Then for $j = 1, .., n = p/l$, the matrix $R^{(j)}$ in Algorithm 9 has columns given by the $(j-1)l + 1, ..., jl$ standard basis vectors, i.e. for any Bregman generator F,

$$R^{(j)} = \begin{pmatrix} 0_{(j-1)l \times l} \\ 1_{l \times l} \\ 0_{(n-j)l \times l} \end{pmatrix} \quad \text{for} \quad j = 1, .., n \tag{C.3.2}$$

Proof. Note that the definitions (5.2.1) and (5.2.2) only depend on the dimension of q, not on the value of q or the Bregman generator F. Fix the block size l so that the dimension of q is $p = nl$. We will prove (C.3.2) by induction on the number of blocks n; in particular, we will show the following for each $j = 1, .., n$;

$$R_n^{(j)} = \begin{pmatrix} 0_{(j-1)l \times l} \\ 1_{l \times l} \\ 0_{(n-j)l \times l} \end{pmatrix} \tag{C.3.3}$$

The subscript n indicates the number of blocks, i.e. that the dimension of the inputs to the algorithm is $p = nl$.

Base case. Suppose $n = 1$, so that the input dimension is $p = l$. Then the only case to check is $R^{(1)}$, which from line 10 of Algorithm 4 is given by

$$R_1^{(1)} := \text{BASISVECTORS}(q^{(0)}, F)$$

Note BASISVECTORS is defined for the Orthogonal Gibbs in (5.2.1) as the matrix whose columns are the first l standard basis vectors of dimension $p = l$. Thus

$$R_1^{(1)} = \text{BASISVECTORS}(q^{(0)}, F) = 1_{l \times l}$$

which is (C.3.3) for $j = n = 1$. Thus the inductive hypothesis (C.3.3) holds for $n = 1$.

Induction step. Assume the inductive hypothesis that (C.3.3) holds for all $j = 1, .., n_0$ for some $n_0 \geq 1$, i.e.

$$R_{n_0}^{(j)} = \begin{pmatrix} 0 & 1 \\ 0_{(j-1) \times l} & \\ 1_{l \times l} & \\ 0_{(n_0 - j) \times l} & \end{pmatrix} \tag{C.3.4}$$

Consider the case when the size of the inputs is $p = (n_0 + 1)l$, i.e. $n = n_0 + 1$. For $j = 1$, the statement on line 10 of the algorithm and the choice of BASISVECTORS for Orthogonal Gibbs (5.2.1) imply

$$R_{n_0+1}^{(1)} = \text{BASISVECTORS}(q, F) = \begin{matrix} 1_{l \times l} \\ 0_{n_0 l \times l} \end{matrix}$$

which agrees with (C.3.3) for $j = 1$ and $n = n_0 + 1$. Furthermore, since line 28 concatenates each $R_{n_0+1}^{(j)}$ to the right-side of $A_{n_0+1}^{(j-1)}$, then for all $j = 1, .., n_0 + 1$ the first l columns of $A_{n_0+1}^{(j)}$ are

$$A := A_{n_0+1}^{(1)} = R_{n_0+1}^{(1)} = \begin{matrix} 1_{l \times l} \\ 0_{n_0 l \times l} \end{matrix}$$

Thus when we step through BASISVECTORS2 from Definition C.1.1, the matrix B defined on line 3 of BASISVECTORS2 is

$$B := \text{COMPLEMENTBASIS}(A) = \begin{matrix} 0_{l \times n_0 l} \\ 1_{n_0 l \times n_0 l} \end{matrix} \tag{C.3.5}$$

by the COMPLEMENTBASIS function (5.2.2) for Orthogonal Gibbs.

Now, let $q^{(0)}, .., q^{(n_0)}$ be the iterates of ITERDCSSAMPLER as defined in Lemma C.2.4, i.e. corresponding to initial iterate $B^T q^{(1)}$, Bregman generator F and target log-density $\log p_d$, and with the state of the pseudorandom generator identical to the state after the first call to SUBSAMPLER producing $a^{(1)}$. Then, by the result (C.2.11) of the lemma, for $j \geq 2$ we have

$$R^{(j)} = BR^{(j-1)}, \quad R_{n_0+1}^{(j)} = BR_{n_0}^{(j-1)}$$

Then, substituting (C.3.5) and the inductive hypothesis (C.3.3) for $n = n_0$ and $j \geq 1$, we find

$$R^{(j)}_{n_0+1} = \begin{bmatrix} & & 0 & & 1 \\ & 0_{n_0 \times n_0} & & 0_{(j-2)\times(n_0+1-j)} & \\ & & 1 & & \\ 1_{n_0 \times n_0} & & & 0_{(n_0+1-j)\times(n_0+1-j)} & \end{bmatrix}$$

$$= \begin{bmatrix} 0 & & 1 & & 0 & & 1 \\ & 0_{(j-2)\times(j-2)} & & & & 0_{(j-1)\times(n_0+1-j)} & \\ & & 1 & & & & \\ 0 & & & & 1 & & \end{bmatrix} = \begin{bmatrix} 0 & & 1 & & 0 & & 1 \\ & 0_{(j-2)\times(j-2)} & & & & 0_{(j-1)\times(n_0+1-j)} & \\ & & 1 & & & & \\ 0 & & & & 0 & & \end{bmatrix}$$

which agrees with (C.3.3) for $n = n_0 + 1$ and $j \geq 2$. Hence by the principle of induction, (C.3.2) is true for all $n \geq 1$. □

C.3.2 Orthogonal Gradient

Lemma C.3.3. Let BASISVECTORS be defined as in the Orthogonal Gradient algorithm, i.e. defined by (5.3.1). Then the $p \times 1$ vectors $R^{(j)}$ $r^{(j)}$ in Algorithm 9 satisfy the following for $j = 2, .., p$:

$$r^{(j)} = \left(\sum_{i=1}^{j-1} \frac{r^{(i)} r^{(i)T}}{r^{(i)T} r^{(i)}} \right) 1_{p \times p} \; r \; F \; q^{(j-1)} \qquad j = 2, .., p \tag{C.3.6}$$

and for $j = 1$,

$$r^{(1)} = r \; F \; q^{(0)} \tag{C.3.7}$$

Proof. Note that for the Orthogonal Gradient algorithm the block size $l = 1$, and hence the number of blocks is $n = p$. We show the result using induction on the dimension p.

Base case. Suppose the dimension is $p = 1$. Then the only case to check is $j = 1$, which is given on line 10 of Algorithm 4 as

$$r^{(1)} \; R^{(1)} := \text{BASISVECTORS}(q^{(0)}, F) = r \; F \; q^{(0)}$$

which satisfies (C.3.7).

Induction step. Suppose (C.3.6) and (C.3.7) hold for some input dimension $p = p_0 \geq 1$, and consider the case when the input dimension is $p := p_0 + 1$.

For $j = 1$, the same argument as in the base case applies, and we have

$$r^{(1)} \; R^{(1)} := \text{BASISVECTORS}(q^{(0)}, F) = r \; F \; q^{(0)} \tag{C.3.8}$$

which satisfies (C.3.7).

Note that for $j \geq 2$, we have the definition

$$R^{(j)} \quad \text{BASISVECTORS2}(q^{(j-1)}, F, A^{(j-1)})$$

The first column of $A^{(j-1)}$ is $r^{(1)}$, which is the matrix A defined on line 2 of BASISVECTORS2 in Definition C.1.1 since $l = 1$. The matrix B is the complement basis of A, defined on line 3 of the same function.

Let $q^{(0)}, .., q^{(p_0)}$ be the iterates of ITERDCSSAMPLER as defined in Lemma C.2.4, i.e. corresponding to initial iterate $B^T q^{(1)}$, Bregman generator F and target log-density $\log p_d$, and with the state of the pseudorandom generator identical to the state after the first call to SUBSAMPLER producing $a^{(1)}$. Then, by the result (C.2.11) of the lemma, for $j = 2, .., p_0 + 1$ we have $r^{(j)} = Br^{(j-1)}$ where $r^{(j)}$ is the equivalent of the $r^{(j)}$ vector for the iterates $q^{(0)}, .., q^{(p_0)}$.

For $j = 2$, we apply the inductive hypothesis for $j = 1$ and $p = p_0$, and (C.2.11);

$$r^{(2)} = Br^{(1)} = \quad Br \; F \; q^{(0)} \quad = \quad Br \; F \; q^{(0)}$$

By Theorem 3.3.12, we have $r\,F(q^{(0)}) = B^T r \; F \; L_d \; q^{(0)}$. By Lemma C.2.4, $L_d \; q^{(0)} = q^{(1)}$, and hence

$$r^{(2)} = \quad Br \; F \; q^{(0)} \quad = \quad BB^T r \; F \; q^{(1)}$$

Finally, we substitute the identity $BB^T = 1_{(p_0+1)\;(p_0+1)} \quad AA^{\prime} = 1_{(p_0+1)\;(p_0+1)} \quad \frac{r^{(1)} r^{(1)\,T}}{r^{(1)\,T} r^{(1)}}$.

$$r^{(2)} = \quad 1_{(p_0+1)\;(p_0+1)} \quad \left(\frac{r^{(1)} r^{(1)\,T}}{r^{(1)\,T} r^{(1)}} \right) \; r \; F \; q^{(1)} \qquad \text{(C.3.9)}$$

which shows (C.3.6) for $j = 2$ and $p = p_0 + 1$.

For $j \quad 3$, we apply the formula $r^{(j)} = Br^{(j-1)}$ from Lemma C.2.4. By the inductive hypothesis,

$$r^{(j)} = Br^{(j-1)} = B \left(\sum_{i=1}^{j-2} \frac{r^{(i)} r^{(i)\,T}}{r^{(i)\,T} r^{(i)}} \quad 1_{p_0 \;\; p_0} \right) r \; F(q^{(j-2)})$$

By Theorem 3.3.12, we have $r\,F(q^{(j-2)}) = B^T r \; F \; L_d \; q^{(j-2)}$. By Lemma C.2.4, $L_d \; q^{(j-2)} = q^{(j-1)}$, and hence

$$\begin{aligned} r^{(j)} &= \left(\sum_{i=1}^{j-2} B \frac{r^{(i)} r^{(i)\,T}}{r^{(i)\,T} r^{(i)}} \quad B \right) B^T r \; F \; q^{(j-1)} \\ &= \left(\sum_{i=1}^{j-2} \frac{B r^{(i)} r^{(i)\,T} B^T}{r^{(i)\,T} r^{(i)}} \quad BB^T \right) r \; F \; q^{(j-1)} \qquad \text{(C.3.10)} \end{aligned}$$

Applying Lemma C.2.4 once more for $i = 1, .., j-2$, we have $r^{(i+1)} = Br^{(i)}$ and, since B has orthonormal columns,

$$r^{(i)\,T} r^{(i)} = r^{(i+1)\,T} B^T B r^{(i+1)} = r^{(i+1)\,T} r^{(i+1)}$$

Substituting this and $1 = AA + BB^T$ into (C.3.10) yields

$$
\begin{aligned}
r^{(j)} &= \sum_{i=1}^{j-2} \frac{r^{(i+1)} r^{(i+1)T}}{r^{(i+1)T} r^{(i+1)}} + AA \Big) 1_{(p_0+1)\ (p_0+1)} - r\ F\ q^{(j-1)} \\
&= \sum_{i=1}^{j-2} \frac{r^{(i+1)} r^{(i+1)T}}{r^{(i+1)T} r^{(i+1)}} + \frac{r^{(1)} r^{(1)T}}{r^{(1)T} r^{(1)}} \Big) 1_{(p_0+1)\ (p_0+1)} - r\ F\ q^{(j-1)} \\
&= \sum_{i=1}^{j-1} \frac{r^{(i)} r^{(i)T}}{r^{(i)T} r^{(i)}} \Big) 1_{(p_0+1)\ (p_0+1)} - r\ F\ q^{(j-1)}
\end{aligned}
$$

which is the result (C.3.6) for $p = p_0 + 1$ and $j \geq 3$.

Thus by the principle of induction, (C.3.6) holds for all $p \geq 1$. □

Appendix D

Details of Conjugate Directions Samplers

In this appendix, we consider how the Dual Complement Submanifold algorithms are simplified when the Bregman generator is of the Euclidean form

$$F(q) := \frac{1}{2} q^T L q - w^T q \tag{D.0.1}$$

where L is a $p \times p$ symmetric positive-definite matrix. Geometry of this kind may arise from the p-Bregman divergence of a Gaussian target distribution, but may also be used as an approximate geometry for more complex distributions. We immediately see that (D.0.1) implies that the metric $\nabla^2 F(q) = L$, a constant matrix. Thus the Riemannian manifold simply takes the structure of an inner-product space, and the notion of orthogonality is equivalent to the following.

Definition D.0.1. Vectors $v, w \in \mathbb{R}^p$ are called *conjugate* if they are orthogonal in the L-weighted inner-product, i.e. $v^T L w = 0$. A set $\{v^{(1)}, .., v^{(n_0)}\}$ of $1 \leq n_0 \leq p$ vectors in $\mathbb{R}^p$ is a *mutually conjugate set* if every distinct pair of its elements are conjugate, i.e. if $v^{(i_2)T} L v^{(i_1)} = 0$ for all distinct $i_1, i_2 \in \{1, .., n_0\}$. A matrix $V \in \mathbb{R}^{p \times n_0}$ is a *conjugate matrix* if its columns form a mutually conjugate set.

Intuitively, algorithm 9 updates iterates $q^{(j)}$ in mutually conjugate directions at each iteration; this will be shown rigorously in this section. There is an obvious analogy to Conjugate Directions methods in optimization [40], the best known of which is the Conjugate Gradient method [73], [70]. Indeed, in Section D.2 we will show that the Orthogonal Gradient sampler reduces to a sampling variant of the Conjugate Gradient algorithm when F is given by (D.0.1). Note that for the Bregman generator (D.0.1), the dual co-ordinate vector corresponding to q is the residual for the equation $Lq = w$,

$$h := \nabla F(q) = Lq - w$$

For simplicity in this section we assume $k = 1$, i.e. the BASISVECTORS function outputs a single column vector on line 18 of algorithm 9,

$$R^{(j)} \quad r^{(j)} = \text{BASISVECTORS}(q, j)$$

We suspect that results in this section can be generalized to $k > 1$ however, so further work could explore block conjugate directions samplers. Since each $r^{(j)}$ is not modified, line 28 implies that for all $j = 1, .., p$,

$$A^{(j)} = \quad r^{(1)} \qquad r^{(j)}$$

The only assumption on $\{r^{(j)}\}_{j=1}^{p}$ is linear independence, so that each $A^{(j)}$ is of full rank; there is no requirement for the set to be mutually conjugate. The following well-known algorithm allows any such set of linearly independent vectors to be transformed into a mutually conjugate basis for the same space. A proof can be found in Axler [5].

Theorem D.0.2 (Modified Gram-Schmidt Process). Let $1 \quad n_0 \quad p$, and let $\{r^{(j)}\}_{j=1}^{n_0}$ be a set of n_0 linearly independent vectors in $\mathbb{R}^p$. Set $v^{(1)} = r^{(1)}$, and for $j = 2, .., n_0$ define

$$v^{(j)} = r^{(j)} \quad \sum_{i=1}^{j-1} \frac{r^{(j)T} L v^{(i)}}{v^{(i)T} L v^{(i)}} v^{(i)} = \quad 1_p \quad p \quad \left(\sum_{i=1}^{j-1} \frac{v^{(i)} v^{(i)T} L}{v^{(i)T} L v^{(i)}} \right) \quad r^{(j)} \qquad \text{(D.0.2)}$$

Then for all $j = 1, .., n_0$, the set $\{v^{(1)}, .., v^{(j)}\}$ is mutually conjugate and spans the same space as $\{r^{(1)}, .., r^{(j)}\}$.

Note that the rank-1 matrix $\frac{v^{(i)} v^{(i)T} L}{v^{(i)T} L v^{(i)}}$ is the L-orthogonal projection matrix onto the span of $v^{(i)}$. In particular, the following lemma shows that each $v^{(j)}$ is the projection of $r^{(j)}$ onto the L-orthogonal complement to the image of $A^{(j-1)}$.

Lemma D.0.3. Let $\{r^{(j)}\}_{j=1}^{n}$ be $1 \quad n \quad p$ linearly independent vectors in $\mathbb{R}^p$, and let $\{v^{(j)}\}_{j=1}^{n}$ be a conjugate set resulting from applying the Gram-Schmidt process of Theorem D.0.2 to $\{r^{(j)}\}_{j=1}^{n}$. If $A^{(j)} = \quad r^{(1)} \qquad r^{(j)}$ and $V^{(j)} = \quad v^{(1)} \qquad v^{(j)}$ then for $j = 1, .., n$,

$$P^L_{A^{(j)}} = P^L_{V^{(j)}} = \sum_{i=1}^{j} \frac{v^{(i)} v^{(i)T} L}{v^{(i)T} L v^{(i)}} \qquad \text{(D.0.3)}$$

and in particular, we have

$$v^{(j)} = \quad 1_p \quad p \quad P^L_{A^{(j-1)}} \quad r^{(j)} \qquad \text{(D.0.4)}$$

for all $j = 2, .., n$.

Proof. Firstly, by Definition 3.3.3,

$$P^L_{V^{(j)}} = V^{(j)} \quad V^{(j)T} L V^{(j)} \quad {}^{-1} V^{(j)T} L \qquad \text{(D.0.5)}$$

for any $j = 2, .., n$. By Theorem D.0.2, $V^{(j)}$ is a conjugate matrix and hence $V^{(j)T} L V^{(j)}$ is diagonal with i^{th} diagonal entry $d_i := v^{(i)T} L v^{(i)}$. Clearly then, the matrix

$$V^{(j)T} L V^{(j)} \quad {}^{1} V^{(j)T} L$$

has i^{th} row $\frac{1}{d_i} v^{(i)T} L$, and hence by (D.0.5) the matrix $P^L_{V^{(j)}}$ is

$$P^L_{V^{(j)}} = \sum_{i=1}^{j} \frac{v^{(i)} v^{(i)T} L}{d_i} = \sum_{i=1}^{j} \frac{v^{(i)} v^{(i)T} L}{v^{(i)T} L v^{(i)}}$$

which gives the second equality in (D.0.3). For the first equality in the same equation, we will show that $P^L_{A^{(j)}} q = P^L_{V^{(j)}} q$ for any $q \in R^p$. Using the decomposition $q = z + u$, where $z \in \text{Im}\, A^{(j)}$ and u is in the L-orthogonal complement to $\text{Im}\, A^{(j)}$, we have

$$P^L_{A^{(j)}} q = P^L_{A^{(j)}} z = z$$

where we have applied Proposition 3.3.6. Similarly, by Theorem D.0.2 we have $\text{Im}\, A^{(j)} = \text{Im}\, V^{(j)}$, and hence $P^L_{V^{(j)}} q = z$. Hence $P^L_{A^{(j)}} = P^L_{V^{(j)}}$. Substituting (D.0.3) with $j \$ j \quad 1$ into (D.0.2) yields result (D.0.4). □

The following proposition gives the decomposition of a vector in R^p in a conjugate basis.

Proposition D.0.4. Let $\{v^{(j)}\}_{j=1}^{p}$ be a basis of mutually conjugate vectors in R^p. Then for any $q \in R^p$,

$$q = \sum_{i=1}^{p} \frac{v^{(i)T} L q}{v^{(i)T} L v^{(i)}} v^{(i)}$$

Proof. Since $\{v^{(j)}\}_{j=1}^{p}$ is a basis for R^p, there exist coefficients $\{q^j\}$ such that

$$q = \sum_{i=1}^{p} q^j v^{(i)}$$

By conjugacy, for each $j = 1, .., p$ we have

$$v^{(j)T} L q = q^j v^{(j)T} L v^{(j)}$$

$$) \quad q^j = \frac{v^{(j)T} L q}{v^{(j)T} L v^{(j)}}$$

Note that since L is positive-definite, the denominator of this expression is non-zero. The result follows. □

The simple form of the Bregman generator (D.0.1) implies that the induced primal and dual connections given in (2.3.7) and (2.3.8) respectively are both zero, and hence the geometry is self-dual. It follows that a submanifold in this geometry is e-flat if and only if it is m-flat, with both cases corresponding to hyperplanes in q co-ordinates. In particular, the embedding L_d of M_{B,d^0} into $\mathbb{R}^p$ has a simple linear form.

Proposition D.0.5. Let $A \in \mathbb{R}^{p \times k}$ have full rank. Under the Bregman generator (D.0.1), the primal embedding L_d of $M_{B,Ad}$ takes the form

$$L_d(b) = \mathbf{1}_p \circ p \quad P\,Bb + PL^{-1}(Ad + w) \tag{D.0.6}$$

where

$$P := P_A^L = A\left(A^T L A\right)^{-1} A^T L$$

and B is a orthonormal basis for the Euclidean-orthogonal complement to $\mathrm{Im}A$.

Proof. Note that under (D.0.1), dual co-ordinates are given by the simple affine relation $h = \nabla F(q) = Lq - w$. Therefore the submanifold M_{B,d^0} is simply the hyperplane defined by the condition $A^T h = d$, i.e. the set of q satisfying

$$A^T(Lq - w) = d \quad , \quad A^T L q = A^T(Ad + w)$$

Substituting the usual decomposition $q = Aa + Bb$, we find an expression for a in terms of b,

$$\begin{aligned} A^T L(Aa + Bb) &= A^T(Ad + w) \\ , \quad A^T L A a &= -A^T L Bb + A^T(Ad + w) \\ , \quad a &= -\left(A^T L A\right)^{-1} A^T L Bb + \left(A^T L A\right)^{-1} A^T(Ad + w) \end{aligned}$$

Substituting this back into $q = Aa + Bb$, we recover (D.0.6).

$$\begin{aligned} q &= \underbrace{A\left(A^T L A\right)^{-1} A^T L}_{P} Bb + \underbrace{A\left(A^T L A\right)^{-1} A^T}_{PL^{-1}}(Ad + w) + Bb \\ &= \mathbf{1}_p \circ p \quad P\,Bb + PL^{-1}(Ad + w) \end{aligned}$$

□

Note that the matrix in front of b in (D.0.6) is exactly the differential matrix of L_d in the general case from Lemma 3.3.10 with $G = L$, as expected. We can use this linear embedding to find an explicit simplified form for the resampling distribution sampled at each iteration of algorithm 9.

Proposition D.0.6. For the Euclidean Bregman generator given in (D.0.1), the resampling step from lines 24 to 26 in algorithm 9 is equivalent to sampling $a^{(j)}$ from a probability density

$$p_{\mathrm{fl}}(a) \propto p\left(q^{(j-1)} + \left(a - \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}}\right) v^{(j)}\right) \tag{D.0.7}$$

where p is the overall target distribution, and then embedding into $\mathbb{R}^p$ via

$$q^{(j-1)} + a^{(j)} \left(\frac{r^{(j)\top} q^{(j-1)}}{r^{(j)\top} r^{(j)}} \right)^{!} v^{(j)} \tag{D.0.8}$$

where $\{v^{(j)}\}_{j=1}^{p}$ is the result of applying the Gram-Schmidt process to $\{r^{(j)}\}_{j=1}^{p}$.

Proof. From line 4, the distribution sampled at each iteration is proportional to

$$\frac{p(q^0)}{\det\left[A^{(j-1)\top} \nabla^2 F(q^0) A^{(j-1)} \right]}$$

for $q^0 = L_d(b^0)$. Since $\nabla^2 F = G = L$ is constant, we can disregard the denominator. Substituting the linear embedding $q^0 = L_d(b^0)$ from proposition D.0.5, we find

$$p(\cdot \mid \mu, p) \propto 1_{\{p = p\}} \mathcal{N}\left(P^{(j-1)} B b^0 + P^{(j-1)} L^{-1} A^{(j-1)} d^{(j-1)} + w \right)$$

where $P^{(j-1)} := P^{L}_{A^{(j-1)}}$ is the L-orthogonal projection matrix onto the columns of $A^{(j-1)}$. We aim to rewrite the argument of this expression,

$$q^0 = 1_{\{p = p\}} \mathcal{N}\left(P^{(j-1)} B b^0 + P^{(j-1)} L^{-1} A^{(j-1)} d^{(j-1)} + w \right) \tag{D.0.9}$$

Firstly, note that $B b^0 = b^0$. The expression for b^0 from line 2 of algorithm 9 is given by

$$\begin{aligned} b^0 &= a r^{(j)} + c^{(j)} = a r^{(j)} + \left(1_{\{p = p\}} \frac{r^{(j)} r^{(j)\top}}{r^{(j)\top} r^{(j)}} \right)^{!} c^{(j-1)} \\ &= a r^{(j)} + c^{(j-1)} \left(\frac{r^{(j)\top} c^{(j-1)}}{r^{(j)\top} r^{(j)}} \right)^{!} r^{(j)} = c^{(j-1)} + \left(a \frac{r^{(j)\top} c^{(j-1)}}{r^{(j)\top} r^{(j)}} \right)^{!} r^{(j)} \end{aligned}$$

where we have applied the update for $c^{(j)}$ on line 21. Substituting this as $b^0 = B b^0$ in (D.0.9) yields

$$\begin{aligned} q^0 &= 1_{\{p = p\}} \mathcal{N}\left(P^{(j-1)} c^{(j-1)} + \left(a \frac{r^{(j)\top} c^{(j-1)}}{r^{(j)\top} r^{(j)}} \right)^{!} r^{(j)} + P^{(j-1)} L^{-1} A^{(j-1)} d^{(j-1)} + w \right) \\ &= 1_{\{p = p\}} \mathcal{N}\left(P^{(j-1)} c^{(j-1)} + \left(a \frac{r^{(j)\top} c^{(j-1)}}{r^{(j)\top} r^{(j)}} \right)^{!} v^{(j)} + P^{(j-1)} L^{-1} A^{(j-1)} d^{(j-1)} + w \right) \end{aligned} \tag{D.0.10}$$

where in the second line we have used Lemma D.0.3. Note that from line 26 of algorithm 9, $q^{(j-1)}$ is the result of MF LAT EMBED2 called with arguments $A = A^{(j-2)}$ and $b = b^{(j-1)} = a^{(j-1)} r^{(j-1)} + c^{(j-1)}$. It follows from the definition of MF LAT EMBED2 in Section 5.1.3 that

$$q^{(j-1)} = A^{(j-2)} a + b^{(j-1)} = A^{(j-2)} a + a^{(j-1)} r^{(j-1)} + c^{(j-1)}$$

for some $a \in \mathbb{R}^{(j-1)}$. Thus, since $P^{(j-1)}$ is a projection onto the columns of $A^{(j-1)} = A^{(j-2)}\ r^{(j-1)}$, it follows from Proposition 3.3.7 that

$$(1_p - P^{(j-1)})q^{(j-1)} = (1_p - P^{(j-1)})c^{(j-1)} \tag{D.0.11}$$

Furthermore, since the $\{r^{(j)}\}$ vectors are mutually Euclidean-orthogonal, we have

$$r^{(j)T}q^{(j-1)} = r^{(j)T}c^{(j-1)} \tag{D.0.12}$$

Similarly, for the third term in (D.0.10),

$$\begin{aligned} A^{(j-1)}d^{(j-1)} + w &= A^{(j-1)}\ A^{(j-1)}\ r\ F\ q^{(j-1)} + w \\ &= A^{(j-1)}A^{(j-1)}\ L\,q^{(j-1)}\ w + w \\ &= A^{(j-1)}A^{(j-1)}\,L\,q^{(j-1)}\ A^{(j-1)}A^{(j-1)}\,w + w \end{aligned}$$

and hence

$$\begin{aligned} &P^{(j-1)}L^{-1}\,Ad^{(j-1)} + w = P^{(j-1)}L^{-1}\,AA\,L\,q^{(j-1)}\ AA\,w + w \\ &= A\ A^TL\,A^{-1}A^TAA\,L\,q^{(j-1)}\ A\ A^TL\,A^{-1}A^TAA\,w + A\ A^TL\,A^{-1}A^Tw \\ &= A\ A^TL\,A^{-1}A^TL\,q^{(j-1)} \\ &= P^{(j-1)}q^{(j-1)} \end{aligned} \tag{D.0.13}$$

Finally, the result follows from substituion of equations (D.0.11), (D.0.12), and (D.0.13) into (D.0.10),

$$\begin{aligned} q^0 &= 1_p - P^{(j-1)}\ q^{(j-1)} + a\ \left(\frac{r^{(j)T}q^{(j-1)}}{r^{(j)T}r^{(j)}}\right)\ v^{(j)} + P^{(j-1)}q^{(j-1)} \\ &= q^{(j-1)} + a\ \left(\frac{r^{(j)T}q^{(j-1)}}{r^{(j)T}r^{(j)}}\right)\ v^{(j)} \end{aligned}$$

□

Algorithm 10 General Conjugate Directions Sampler

```
function GCDSAMPLER(q^(0), L, log p)
    r^(1) ← BASISVECTORS1(q^(0), F)
    v^(1) ← r^(1)
    a^(1) ← SUBSAMPLER(a ↦ p(q^(0) + a (r^(1)T q^(0)) / (r^(1)T r^(1)) v^(1)))
    q^(1) ← q^(0) + a^(1) (r^(1)T q^(0)) / (r^(1)T r^(1)) v^(1)
    A^(1) ← r^(1)
    for j = 2, .., p do
        r^(j) ← BASISVECTORS2(q^(j-1), F, A^(j-1))
        v^(j) ← r^(j) − Σ_{i=1}^{j-1} (r^(j)T L v^(i)) / (v^(i)T L v^(i)) v^(i)
        a^(j) ← SUBSAMPLER(a ↦ p(q^(j-1) + a (r^(j)T q^(j-1)) / (r^(j)T r^(j)) v^(j)))
        q^(j) ← q^(j-1) + a^(j) (r^(j)T q^(j-1)) / (r^(j)T r^(j)) v^(j)
        A^(j) ← A^(j-1) r^(j)
    end for
    return q
end function
```

D.0.1 Gaussian Target Distributions

Lemma D.0.7. Let p be a multivariate Gaussian density on $\mathbb{R}^p$ with mean m and precision L, i.e. covariance matrix L^{-1}. Then for $q, v \in \mathbb{R}^p$, the function $z \mapsto p(q + zv)$ for $z \in \mathbb{R}$ is proportional to a one-dimensional Gaussian density with mean $m = v^T L (m - q) / v^T L v$ and variance $s^2 = 1 / v^T L v$.

Proof.

$$p(q + zv) \propto \exp\left[-\frac{1}{2}(q + zv - m)^T L (q + zv - m)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(z^2 v^T L v + 2z v^T L (q - m)\right)\right]$$

where we have dropped terms not involving z. Substituting the definitions of m and s^2 yields

$$p(q + zv) \propto \exp\left[-\frac{1}{2}\left(z^2 s^{-2} - 2zms^{-2}\right)\right]$$

$$\propto \exp\left[\frac{1}{2s^2}\left(z^2 - 2zm\right)\right] \propto \exp\left[\frac{1}{2s^2}(z - m)^2\right]$$

which is a Gaussian density with mean m and variance s^2. □

Theorem D.0.8. Let p be a multivariate Gaussian distribution on $\mathbb{R}^p$ with mean m and covariance L^{-1}. Then in algorithm 10 applied to target distribution p, lines 10 and 11 are equivalent to the update

$$q^{(j)} \quad q^{(j-1)} + \left(\frac{z^{(j)}}{\sqrt{r^{(j)}}} + \frac{v^{(j)T} L \left(m - q^{(j-1)}\right)}{r^{(j)}}\right) v^{(j)}$$

where $r^{(j)} = v^{(j)T} L v^{(j)}$ and $z^{(j)} \sim \text{Normal}(0, 1)$.

Proof. Line 10 samples $z^{(j)}$ from a distribution

$$p(z) \propto p\left(q^{(j-1)} + z \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}} v^{(j)}\right)$$

By Lemma D.0.7 with $q = q^{(j-1)} - \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}} v^{(j)}$ and $v = v^{(j)}$, this is proportional to a Gaussian distribution on z with mean and variance respectively given by

$$\begin{aligned} m &= \frac{v^{(j)T} L m}{v^{(j)T} L v^{(j)}} - \frac{v^{(j)T} L q^{(j-1)}}{v^{(j)T} L v^{(j)}} + \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}} - \frac{v^{(j)T} L v^{(j)}}{v^{(j)T} L v^{(j)}} \\ &= \frac{v^{(j)T} L \left(m - q^{(j-1)}\right)}{r^{(j)}} + \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}} \\ s^2 &= \frac{1}{v^{(j)T} L v^{(j)}} = \frac{1}{r^{(j)}} \end{aligned}$$

If z is a random variable with this distribution, then

$$z - \frac{r^{(j)T} q^{(j-1)}}{r^{(j)T} r^{(j)}} \tag{D.0.14}$$

is Gaussian with mean $v^{(j)T} L \left(m - q^{(j-1)}\right) / r^{(j)}$ and variance $1/r^{(j)}$. Clearly then (D.0.14) has the same Gaussian distribution as

$$\left(\frac{z}{\sqrt{r^{(j)}}} + \frac{v^{(j)T} L \left(m - q^{(j-1)}\right)}{r^{(j)}}\right) \tag{D.0.15}$$

where $z \sim \text{Normal}(0, 1)$. The result follows by replacing the random variable (D.0.14) with (D.0.15) in line 11 of algorithm 10. □

Lemma D.0.9. Let $\left\{v^{(j)}\right\}$ be the conjugate basis resulting from applying the modified Gram-Schmidt process to the basis $\left\{r^{(j)}\right\} \subset \mathbb{R}^p$. Then for all $j = 1, .., p$,

$$v^{(j)T} L v^{(j)} = v^{(j)T} L r^{(j)}$$

Proof. By the modified Gram-Schmidt process in Theorem D.0.2, we have

$$v^{(j)} = r^{(j)} - \sum_{i=1}^{j-1} \frac{r^{(j)T} L v^{(i)}}{v^{(i)T} L v^{(i)}} v^{(i)}$$

The result follows from the fact that $v^{(j)}$ is conjugate to $v^{(i)}$ for all $i = 1, .., j-1$. □

The general Conjugate Directions sampler for a multivariate Gaussian target distribution is listed as Algorithm 11.

Algorithm 11 General Conjugate Directions Sampler for a Gaussian Target

```
1: function GCDSAMPLER(q^(0), L, w)
2:     r^(1) ← BASISVECTORS1(q^(0), F)
3:     v^(1) ← r^(1)
4:     A^(1) ← r^(1)
5:     for j = 1, .., p do
6:         r^(j) ← v^(j)T L v^(j)
7:         z^(j) ← Normal(0, 1/r^(j))
8:         q^(j) ← q^(j-1) + z^(j) + (v^(j)T (w - L q^(j-1))) / r^(j) v^(j)
9:         r^(j+1) ← BASISVECTORS2(q^(j), F, A^(j))
10:        v^(j+1) ← r^(j+1) - Σ_{i=1}^{j} (r^(j+1)T L v^(i)) / r^(i) v^(i)
11:        A^(j+1) ← A^(j) r^(j+1)
12:    end for
13:    return q^(p)
14: end function
```

Theorem D.0.10. If Algorithm 11 terminates successfully after p iterations, then the returned vector is

$$q^{(p)} = m + \sum_{i=1}^{p} z^{(i)} v^{(i)}$$

where $\{z^{(i)}\}$ and $\{v^{(i)}\}$ are defined as in the algorithm, and $m = w$.

Proof. Note that by line 8 of algorithm 11,

$$q^{(j)} = q^{(j-1)} + \left(z^{(j)} + \frac{v^{(j)T} (w - L q^{(j-1)})}{r^{(j)}} \right) v^{(j)}$$

for $j = 1, .., p$. Thus by induction, for all $j = 1, .., p$ we must have

$$q^{(j)} = q^{(0)} + \sum_{i=1}^{j} \left(z^{(i)} + \frac{v^{(i)T} (w - L q^{(i-1)})}{r^{(i)}} \right) v^{(i)} \qquad \text{(D.0.16)}$$

In particular, for $j = p$,

$$q^{(p)} = q^{(0)} + \sum_{j=1}^{p} \left(z^{(j)} + \frac{v^{(j)T} w \quad L\, q^{(j\ 1)}}{r^{(j)}} \right) v^{(j)}$$

$$= q^{(0)} + \sum_{j=1}^{p} z^{(j)} v^{(j)} + \sum_{j=1}^{p} \left(\frac{v^{(j)T} L \quad m \quad q^{(j\ 1)}}{r^{(j)}} \right) v^{(j)} \qquad \text{(D.0.17)}$$

where we have substituted $w = L\, m$ to rewrite the final term. Equation (D.0.16) implies that for $j = 2, .., p$,

$$q^{(j\ 1)} \in q^{(0)} + \text{Span}\{ v^{(1)}, .., v^{(j\ 1)} \}$$

and therefore, since $v^{(j)}$ is conjugate to each of $v^{(1)}, .., v^{(j\ 1)}$,

$$v^{(j)T} L\, q^{(j\ 1)} = v^{(j)T} L\, q^{(0)}$$

for $j = 2, .., p$, which is also trivially true in the case $j = 1$. Substituting this result and $r^{(j)} = v^{(j)T} L\, v^{(j)}$ into (D.0.17),

$$q^{(p)} = q^{(0)} + \sum_{j=1}^{p} z^{(j)} v^{(j)} + \sum_{j=1}^{p} \left(\frac{v^{(j)T} L \quad m \quad q^{(0)}}{v^{(j)T} L\, v^{(j)}} \right) v^{(j)}$$

By Proposition D.0.4, the final term of this expression is $m \quad q^{(0)}$. Thus the result follows. □

D.1 Gibbs Sampling and Matrix Factorizations

In this section we show that the Orthogonal Gibbs algorithm is equivalent to the Gaussian Elimination sampler, Algorithm 6, when the Bregman generator is given by (D.0.1). Recall that the Gaussian elimination algorithm applies elementary row operations to a matrix $L \in R^{p \quad p}$ in order to transform it into reduced row echelon form as follows:

Algorithm 12 Gaussian Elimination

```
for j = 1, .., p  1 do
    for i = j + 1, .., p do
        C_ij  L_ij / L_jj
        for s = 1, .., p do
            L_is  L_is  C_ij  L_js
        end for
    end for
end for
```

The result L' of applying Gaussian Elimination to L can be expressed as a product of elementary matrices with L, where each elementary matrix acts on its multiplicand by applying an elementary row operation.

Recall that for Orthogonal Gibbs, the BASISVECTORS2 function returns the standard basis vectors, i.e. $r^{(j)} = e^{(j)}$ for all $j = 1, .., p$. We have shown in this appendix that the Euclidean Bregman generator reduces the general Dual Complement Submanifold algorithm to the conjugate directions sampler, Algorithm 10, which applies a modified Gram-Schmidt process to the $\{r^{(j)}\}$ basis vectors. Thus, let $V = \begin{bmatrix} v^{(1)} & \cdots & v^{(p)} \end{bmatrix}$ be the matrix whose columns are the conjugate vectors resulting from the modified Gram-Schmidt process applied to the standard basis vectors. The following theorem shows that V^T is the product of elementary matrices that transform L into its reduced row echelon form, i.e. $L' = V^T L$. The result is also shown in Chapter 4 of Hestenes [40].

Theorem D.1.1. Let V be the matrix whose columns are formed by the modified conjugate Gram-Schmidt process applied to the standard basis vectors. Then the matrix $V^T L$ is equivalent to L', the application of the Gaussian Elimination algorithm to L. In particular, for each $j = 1, .., p - 1$ the row vector $v^{(j)T} L$ is the j^{th} row of the matrix resulting from applying j iterations of the Gaussian Elimination algorithm to L.

Proof. We show the result by induction on the rows of L. Let L' be the reduced row echelon form of L, i.e. the result of applying Gaussian Elimination to L.

Base case. The Gaussian Elimination algorithm leaves the first row of L unchanged, so the first rows of L and L' are identical. From the Gram-Schmidt process, the first conjugate vector is equal to the first standard basis vector; $v^{(1)} = e^{(1)}$. Therefore $v^{(1)T} L$ is simply the first row of L, which agrees with the first row of L'.

Induction step. Let $j \in \{1, .., p - 1\}$. Assume the inductive hypothesis that for all $i = 1, .., j$, the row vector $v^{(i)T} L$ is the i^{th} row of L'. We seek to show the equivalent statement holds for $j + 1$. By the modified Gram-Schmidt process given in Theorem D.0.2, we have

$$v^{(j+1)} = e^{(j+1)} - \sum_{i=1}^{j} \frac{v^{(i)T} L\, e^{(j+1)}}{v^{(i)T} L\, v^{(i)}} v^{(i)}$$

$$\Rightarrow v^{(j+1)T} L = e^{(j+1)T} L - \sum_{i=1}^{j} \frac{v^{(i)T} L\, e^{(j+1)}}{v^{(i)T} L\, v^{(i)}} v^{(i)T} L$$

Note that $e^{(j+1)T} L$ is the row vector corresponding to j^{th} row of L, and by the inductive hypothesis $v^{(i)T} L$ is the i^{th} row of L'. In particular, the coefficients in the summation become

$$\frac{v^{(i)T} L\, e^{(j+1)}}{v^{(i)T} L\, v^{(i)}} = \frac{L'_{i,j+1}}{L'_{ii}}$$

where in the denominator we have applied Lemma D.0.9. It follows that the s^{th} component of the row vector $v^{(j+1)^T} L$ is given by

$$v^{(j+1)^T} L\, e^{(s)} = L_{j+1,s} \sum_{i=1}^{j} \frac{L^0_{i,j+1}}{L^0_{ii}} L^0_{is} \qquad \text{(D.1.1)}$$

The right-hand side of this is exactly the s^{th} component of the $(j+1)^{th}$ row of L^0. Thus the inductive hypothesis holds for $j+1$.

Thus by the principle of induction, $L^0 = V^T L$. □

Since $V^T = V^T 1_{p\ p}$, Theorem D.1.1 implies that the conjugate vectors $\{v^{(j)}\}$ can be computed by applying elementary row operations to the identity matrix, where the operations are the same as those used to reduce L to its echelon form. Similarly, the same elementary row operations can be used to transform the linear system into an upper-triangular form,

$$\begin{aligned} L\,m &= w \\ \Rightarrow \quad V^T L\,m &= V^T w \end{aligned}$$

In Algorithm 6, we apply these elementary row operations to the identity, L and w.

Theorem 5.2.2. Let L^0 and V result from the application of Algorithm 6 to a symmetric positive-definite matrix $L \in R^{p\ p}$ and $w \in R^p$. Let D be the diagonal matrix whose diagonal is the diagonal of L^0. The lower-triangular Cholesky factor L of, i.e. such that $L = LL^T$, is given by

$$L := L^{0\ T} D^{-1/2}$$

Proof. By Theorem D.1.1, $L^0 = V^T L$. Lemma D.0.9 then implies that the j^{th} entry on the diagonal of D is given by

$$D_{jj} = v^{(j)^T} L\, e^{(j)} = v^{(j)^T} L\, v^{(j)}$$

In matrix form, $D = V^T L\, V$. Therefore

$$\begin{aligned} LL^T &= L^{0\ T} D^{-1/2} D^{-1/2} L^0 \\ &= L\, V D^{-1/2} D^{-1/2} V^T L \\ &= L\, V D^{-1} V^T L \\ &= L\, V (V^T L\, V)^{-1} V^T L \\ &= LL^{-1} L = L \end{aligned}$$

where we have used the fact that V is non-singular because its columns form a basis. Note that since L^0 is upper-triangular, then so is L, and thus L is the Cholesky factor of L. □

Corollary 5.2.3. The final iterate $q^{(p)}$ of the Gaussian Elimination sampler, Algorithm 6, is stochastically independent of the initial iterate $q^{(0)}$ and is a sample from the target distribution Normal $(L^{-1}w, L^{-1})$.

Proof. By Theorem D.0.10, the final iterate is given by

$$q^{(p)} = m + \sum_{j=1}^{p} z^{(j)} v^{(j)}$$

where $\{v^{(j)}\}$ are the conjugate vectors defined by the algorithm, and $z^{(j)} \sim \text{Normal}(0, 1/r^{(j)})$. Note that the $\{v^{(j)}\}$ are entirely deterministic, resulting from a Gram-Schmidt process applied to the standard basis vectors. It follows that the $z^{(j)}$ are independent random variables. We therefore write

$$q^{(p)} = m + VD^{-1/2}x$$

where D is a diagonal matrix whose diagonal elements are $\{r^{(j)}\}_{j=1}^{p}$ and $x \sim \text{Normal}(0, 1_p)$ is a vector of i.i.d. standard normal random variables. Clearly then, $q^{(p)}$ has a multivariate Gaussian distribution with mean m and covariance

$$\begin{aligned} VD^{-1/2}D^{-1/2}V^T &= VD^{-1}V^T \\ &= V(V^T L V)^{-1}V^T \\ &= L^{-1} \end{aligned}$$

Since neither m nor L^{-1} depend on $q^{(0)}$, it follows that $q^{(p)}$ is independent of $q^{(0)}$. □

D.2 The Conjugate Gradient Sampler

Consider Algorithm 10 when the BASISVECTORS function has the definition given in Section 5.3. Lemma C.3.3 states that the resulting basis vectors $\{r^{(j)}\}$ built up over the iterations of the general iterative algorithm are the result of a Gram-Schmidt process, with respect to the Euclidean inner-product on $\mathbb{R}^p$, applied to the negated gradients $\nabla_r F(q^{(j-1)})$ of the Bregman generator at each iterate. In Algorithm 10, the conjugate directions $\{v^{(j)}\}$ are the result of a modified Gram-Schmidt process, i.e. with respect to the L-weighted inner-product, applied to the $\{r^{(j)}\}$.

One might therefore expect that two Gram-Schmidt updates are required, in order to orthonormalize against the Euclidean inner-product and then the L-weighted inner-product. In fact, the following lemma shows that the $\{v^{(j)}\}$ can be computed directly via the modified Gram-Schmidt process applied to the negated gradients. This simplifies the conjugate gradient algorithm.

Lemma D.2.1. Let $\{r^{(j)}\}$ and $\{v^{(j)}\}$ be defined as in Algorithm 10 with the BASISVECTORS function defined as in Section 5.3. Then for $j \in \{2, \ldots, p\}$, we have

$$v^{(j)} = \sum_{i=1}^{j-1} \frac{v^{(i)T} L r - q^{(j-1)}}{v^{(i)T} L v^{(i)}} v^{(i)} - r + q^{(j-1)}$$

Proof. Let $j \in \{2, \ldots, p\}$. Line 9 of Algorithm 10 is the j^{th} step of the modified Gram-Schmidt process applied to $\{r^{(j)}\}$,

$$\begin{aligned} v^{(j)} &= r^{(j)} - \sum_{i=1}^{j-1} \frac{r^{(j)T} L v^{(i)}}{v^{(i)T} L v^{(i)}} v^{(i)} \\ &= \left(1_p - \sum_{i=1}^{j-1} \frac{v^{(i)} v^{(i)T} L}{v^{(i)T} L v^{(i)}}\right) r^{(j)} = \left(1_p - P^L_{V^{(j-1)}}\right) r^{(j)} \end{aligned} \tag{D.2.1}$$

where the last equality follows from Lemma D.0.3. From Lemma C.3.3, we have

$$r^{(j)} = \left(\sum_{i=1}^{j-1} \frac{r^{(i)} r^{(i)T}}{r^{(i)T} r^{(i)}} - 1_p\right) r + q^{(j-1)} = \left(P_{A^{(j-1)}} - 1_p\right) r + q^{(j-1)} \tag{D.2.2}$$

where $P_{A^{(j-1)}} - P^{1_p}_{A^{(j-1)}}$ is the Euclidean-orthogonal projection matrix onto the span of $\{r^{(1)}, \ldots, r^{(j-1)}\}$. Substitution of (D.2.2) into (D.2.1) then gives

$$\begin{aligned} v^{(j)} &= \left(1_p - P^L_{V^{(j-1)}}\right)\left(P_{A^{(j-1)}} - 1_p\right) r + q^{(j-1)} \\ &= \left(P_{A^{(j-1)}} - 1_p - P^L_{V^{(j-1)}} P_{A^{(j-1)}} + P^L_{V^{(j-1)}}\right) r + q^{(j-1)} \end{aligned} \tag{D.2.3}$$

Note that the span of $\{r^{(1)}, \ldots, r^{(j-1)}\}$ is identical to the span of $\{v^{(1)}, \ldots, v^{(j-1)}\}$ by Theorem D.0.2. Thus for any $x \in R^p$, the projection $P_{A^{(j-1)}} x$ is an element of $\mathrm{Im}\, A^{(j-1)} = \mathrm{Im}\, V^{(j-1)}$. Therefore by Proposition 3.3.7,

$$P^L_{V^{(j-1)}} P_{A^{(j-1)}} x = P_{A^{(j-1)}} x$$

Since this holds for any $x \in R^p$, we conclude that $P^L_{V^{(j-1)}} P_{A^{(j-1)}} = P_{A^{(j-1)}}$. Substituting this result into equation (D.2.3)

$$\begin{aligned} v^{(j)} &= \left(P_{A^{(j-1)}} - 1_p - P^L_{V^{(j-1)}} P_{A^{(j-1)}} + P^L_{V^{(j-1)}}\right) r + q^{(j-1)} \\ &= \left(P_{A^{(j-1)}} - 1_p - P_{A^{(j-1)}} + P^L_{V^{(j-1)}}\right) r + q^{(j-1)} \\ &= \left(P^L_{V^{(j-1)}} - 1_p\right) r + q^{(j-1)} \end{aligned}$$

Substituting the result of Lemma D.0.3 yields the result.

$$v^{(j)} = \sum_{i=1}^{j-1} \frac{v^{(i)T} L r - q^{(j-1)}}{v^{(i)T} L v^{(i)}} v^{(i)} - r + q^{(j-1)}$$

□

As in Algorithm 8, we de ne

$$g^{(j)} = L\,q^{(j)} \quad w$$

for $j = 1, .., p$.

Corollary D.2.2. For $j = 1, .., p$,

$$g^{(j)} = g^{(j\ 1)} + l^{(j)} L\,v^{(j)}$$

Proof. By the de niton of $g^{(j)}$ and line 9 of Algorithm 8, we have

$$\begin{aligned} g^{(j)} &= L\,q^{(j)} \quad w \\ &= L \quad q^{(j\ 1)} + l^{(j)} v^{(j)} \quad w \\ &= g^{(j\ 1)} + l^{(j)} L\,v^{(j)} \end{aligned}$$

which holds for $j = 1, .., p$. □

The following proposition justi es line 13 of Algorithm 8. The proof is adapted from Proposition 2.4.1 in Fox [29].

Proposition D.2.3. Let $f\,v^{(j)}g$ and $f\,r^{(j)}g$ be de ned as in Algorithm 8 for $ǂ 1, .., p$. Then for $j = 2, .., p$,

$$v^{(j)} = r^{(j)} \quad (r^{(j)T} g^{(j\ 1)} / r^{(j\ 1)} l^{(j\ 1)}) v^{(j\ 1)}$$

where $r^{(j)} = v^{(j)T} L\,v^{(j)}$, and

$$l^{(j)} \quad z^{(j)} \quad v^{(j)T} g^{(j\ 1)} / r^{(j)}$$

as de ned on line 8.

Proof. By Corollary D.2.2, the product $L\,v^{(j)}$ is proportional to the difference of successive gradient vectors,

$$g^{(j)} = \frac{1}{l^{(j)}} \quad g^{(j)} \quad g^{(j\ 1)}$$

Substituting this into the Gram-Schmidt process for $v^{(j)}$ from line 9 of the general conjugate directions algorithm, Algorithm 10, gives

$$\begin{aligned} v^{(j)} &= r^{(j)} \quad \sum_{i=1}^{j\ 1} \frac{r^{(j)T} L\,v^{(i)}}{v^{(i)T} L\,v^{(i)}} v^{(i)} \\ &= r^{(j)} \quad \sum_{i=1}^{j\ 1} \frac{r^{(j)T} \quad g^{(i)} \quad g^{(i\ 1)}}{r^{(i)} l^{(i)}} v^{(i)} \end{aligned} \tag{D.2.4}$$

for $j = 2, .., p$, where $v^{(1)} = r^{(1)}$. By Lemma C.3.3, $\{r^{(j)}\}$ is the Gram-Schmidt process applied to $\{g^{(j-1)}\}$, i.e.

$$r^{(j)} = g^{(j-1)} + \sum_{i=1}^{j-1} \frac{r^{(i)T} g^{(j-1)}}{r^{(i)T} r^{(i)}} r^{(i)} \qquad j = 2, .., p \tag{D.2.5}$$

Thus by Theorem D.0.2, $r^{(j)}$ is Euclidean-orthogonal to $\text{Span}\{g^{(0)}, .., g^{(j-2)}\}$ for $j = 2, .., p$. Thus all terms except one in (D.2.4) are zero, and hence the result follows;

$$v^{(j)} = r^{(j)} - \frac{r^{(j)T} g^{(j-1)}}{r^{(j-1)} \| r^{(j-1)} \|} v^{(j-1)}$$

□

Corollary D.2.4. For all $i = 1, .. p-1$ and $j > i$,

$$r^{(i)T} g^{(j)} = r^{(i)T} g^{(i)}$$

Proof. By Corollary D.2.2,

$$g^{(j)} = g^{(j-1)} + \| r^{(j)} \| v^{(j)}$$

Taking the inner product of this with $r^{(i)}$,

$$r^{(i)T} g^{(j)} = r^{(i)T} g^{(j-1)} + \| r^{(j-1)} \| r^{(i)T} v^{(j)}$$

Since $v^{(j)}$ is the j^{th} conjugate vector resulting from a modified Gram-Schmidt process applied to $\{r^{(k)}\}_{k=1}^{j}$, it follows that $v^{(j)}$ is conjugate to $r^{(i)}$. Thus $r^{(i)T} g^{(j)} = r^{(i)T} g^{(j-1)}$. Since this holds for any $j > i$, the result follows. □

The following proposition justifies lines 11 and 12 in Algorithm 8.

Proposition D.2.5.

$$r^{(j+1)} = (1 + \delta_j) r^{(j)} - \| r^{(j)} \| v^{(j)}$$

where

$$\delta_j := \frac{r^{(j)T} g^{(j)}}{r^{(j)T} r^{(j)}}$$

for $j = 1, .., p$.

Proof. Line 9 of Algorithm 10 is the j^{th} step of the modified Gram-Schmidt process applied to $\{r^{(j)}\}$, and hence

$$\begin{aligned} r^{(j+1)} &= g^{(j)} + \sum_{i=1}^{j} \frac{r^{(i)T} g^{(j)}}{r^{(i)T} r^{(i)}} r^{(i)} \\ &= g^{(j)} + \delta_j r^{(j)} + \sum_{i=1}^{j-1} \frac{r^{(i)T} g^{(j)}}{r^{(i)T} r^{(i)}} r^{(i)} \end{aligned}$$

Substituting the expression in Corollary D.2.2 for $g^{(j)}$,

$$r^{(j+1)} = d_j r^{(j)} \quad l^{(j)} L v^{(j)} \quad g^{(j-1)} + \sum_{i=1}^{j-1} \frac{r^{(i)T} g^{(j)}}{r^{(i)T} r^{(i)}} r^{(i)} \qquad \text{(D.2.6)}$$

By Corollary D.2.4, $r^{(i)T} g^{(j)} = r^{(i)T} g^{(i)} = r^{(i)T} g^{(j-1)}$ for all $i = 1, .., j-1$. Thus the last two terms in (D.2.6) are

$$g^{(j-1)} + \sum_{i=1}^{j-1} \frac{r^{(i)T} g^{(j)}}{r^{(i)T} r^{(i)}} r^{(i)} = g^{(j-1)} + \sum_{i=1}^{j-1} \frac{r^{(i)T} g^{(j-1)}}{r^{(i)T} r^{(i)}} r^{(i)} = r^{(j)}$$

Substituting this into (D.2.6) yields the result,

$$r^{(j+1)} = r^{(j)} + d_j r^{(j)} \quad l^{(j)} L v^{(j)}$$

□

Bibliography

[1] Harold Abelson and Gerald Jay Sussman. Structure and interpretation of computer programs. The MIT Press, 1996.

[2] S.-I. Amari. Information geometry on hierarchy of probability distributions. In: IEEE Transactions on Information Theory 47.5 (2001), pp. 1701–1711. ISSN: 0018-9448. DOI: 10.1109/18.930911. URL: http://dx.doi.org/10.1109/18.930911.

[3] Shun-ichi Amari. Information geometry and its applications. Springer, 2016.

[4] V. I Arnold, K Vogtmann, and A Weinstein. Mathematical Methods of Classical Mechanics. eng. Second Edition. Vol. 60. Graduate Texts in Mathematics. New York: Springer New York, 1989. ISBN: 1441930876.

[5] Sheldon Axler. Linear algebra done right. Vol. 2. Springer, 2015.

[6] Arindam Banerjee et al. Clustering with Bregman divergences. In: Journal of machine learning research 6.10 (2005).

[7] Julian Besag. Comments on Representations of knowledge in complex systems by U. Grenander and MI Miller. In: J. Roy. Statist. Soc. Ser. B 56.591-592 (1994), p. 4.

[8] Michael Betancourt. A Conceptual Introduction To Hamiltonian Monte Carlo. In: CoRR (2017). arXiv: 1701.02434 [stat.ME]. URL: http://arxiv.org/abs/1701.02434v1.

[9] Michael Betancourt. A General Metric for Riemannian Manifold Hamiltonian Monte Carlo. In: Geometric Science of Information (2013), pp. 327–334. ISSN: 1611-3349. DOI: 10.1007/978-3-642-40020-9_35. URL: http://dx.doi.org/10.1007/978-3-642-40020-9_35.

[10] Michael Betancourt. A Uni ed Treatment of Predictive Model Comparison. 2015. arXiv: 1506.02273 [stat.ME].

[11] Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. In: Current trends in Bayesian methodology with applications 79 (2015), p. 30.

[12] Michael Betancourt et al. The geometric foundations of hamiltonian monte carlo. In: Bernoulli 23.4A (2017), pp. 2257–2298.

[13] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. Convex optimization Cambridge university press, 2004.

[14] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming . In: USSR computational mathematics and mathematical physics 7.3 (1967), pp. 200 217.

[15] Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory . In: Ims. 1986.

[16] Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds . In: Scandinavian Journal of Statistics 40.4 (2013), pp. 825 845.

[17] Ovidiu Calin and Constantin Udrişte. Geometric modeling in probability and statistics. Vol. 121. Springer, 2014.

[18] Bob Carpenter et al. Stan: A probabilistic programming language . In: Journal of statistical software 76.1 (2017).

[19] Joseph T Chang and David Pollard. Conditioning as disintegration . In: Statistica Neerlandica 51.3 (1997), pp. 287 317.

[20] J. Andres Christen and Colin Fox. A general purpose sampling algorithm for continuous distributions (the t -walk) . In: Bayesian Anal. 5.2 (June 2010), pp. 263 281. DOI: 10.1214/10-BA603. URL: https://doi.org/10.1214/10-BA603 .

[21] Imre Csiszar. Information measures: A critical survey . In: Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes. 1974, pp. 73 86.

[22] T Cui et al. Likelihood-informed dimension reduction for nonlinear inverse problems . In: Inverse Problems 30.11 (Oct. 2014), p. 114015. ISSN: 1361-6420. DOI: 10.1088/0266-5611/30/11/114015 . URL: http://dx.doi.org/10.1088/0266-5611/30/11/114015 .

[23] Dennis M DeTurck and Deane Yang. Existence of elastic deformations with prescribed principal strains and triply orthogonal systems . eng. In: Duke mathematical journal 51.2 (1984), pp. 243 260. ISSN: 0012-7094.

[24] Persi Diaconis, Susan Holmes, and Mehrdad Shahshahani. Sampling from a manifold . In: Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton. Institute of Mathematical Statistics, 2013, pp. 102 125.

[25] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families . In: The Annals of statistics (1979), pp. 269 281.

[26] Simon Duane et al. Hybrid Monte Carlo . In: Physics Letters B 195.2 (Sept. 1987), pp. 216 222. ISSN: 0370-2693. DOI: 10.1016/0370-2693(87)91197-x . URL: http://dx.doi.org/10.1016/0370-2693(87)91197-x .

[27] Bradley Efron and David V Hinkley. Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information . eng. In: Biometrika 65.3 (1978), p. 457. ISSN: 0006-3444.

[28] Herbert Federer. Geometric measure theory. Springer, 2014.

[29] Colin Fox. A Conjugate Direction Sampler for Normal Distributions, with a Few Computed Examples. Tech. rep. 2008-1. University of Otago, 2008.

[30] Colin Fox. Polynomial accelerated mcmc and other sampling algorithms inspired by computational optimization . In: Monte Carlo and Quasi-Monte Carlo Methods 2012. Springer, 2013, pp. 349 366.

[31] Colin Fox, Albert Parker, et al. Accelerated Gibbs sampling of normal distributions using matrix splittings and polynomials . In: Bernoulli 23.4B (2017), pp. 3711 3743.

[32] Alan E Gelfand et al. Handbook of spatial statistics. CRC press, 2010.

[33] Andrew Gelman et al. Bayesian data analysis. CRC press, 2013.

[34] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images . In: IEEE Transactions on pattern analysis and machine intelligence 6 (1984), pp. 721 741.

[35] Charles Geyer. Asymptotics of Exponential Families. Jan. 2013.

[36] Charles Geyer. Introduction to markov chain monte carlo . In: Handbook of markov chain monte carlo (2011).

[37] George T Gilbert. Positive de nite matrices and Sylvester's criterion . In: The American Mathematical Monthly 98.1 (1991), pp. 44 46.

[38] Wally R Gilks, Nicky G Best, and Keith KC Tan. Adaptive rejection Metropolis sampling within Gibbs sampling . In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 44.4 (1995), pp. 455 472.

[39] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods . In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.2 (2011), pp. 123 214.

[40] Magnus Rudolph Hestenes. Conjugate direction methods in optimization. Vol. 12. Springer Science & Business Media, 2012.

[41] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. In: Journal of Machine Learning Research 15.1 (2014), pp. 1593 1623.

[42] Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms . In: Stochastic processes and their applications 85.2 (2000), pp. 341 361.

[43] Shmuel Kantorovitz. Several Real Variables. Vol. 3. Springer, 2016.

[44] Robert E Kass and Paul W Vos. Geometrical foundations of asymptotic inference. Vol. 908. John Wiley & Sons, 2011.

[45] John Lee. Introduction to Riemannian manifolds. Cham, Switzerland: Springer, 2018. ISBN: 9783319917542.

[46] John Lee. *Introduction to smooth manifolds*. New York London: Springer, 2012. ISBN: 9781441999818.

[47] John Lee. *Introduction to topological manifolds*. New York: Springer, 2000. ISBN: 0387987592.

[48] Jun S Liu and Chiara Sabatti. "Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation". In: *Biometrika* 87.2 (2000), pp. 353–369.

[49] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer New York, 2004. ISBN: 9780387763712. DOI: 10.1007/978-0-387-76371-2. URL: http://dx.doi.org/10.1007/978-0-387-76371-2.

[50] Samuel Livingstone and Mark Girolami. "Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions". eng. In: *Entropy (Basel, Switzerland)* 16.6 (2014), pp. 3074–3102. ISSN: 1099-4300.

[51] Yi-An Ma, Tianqi Chen, and Emily B. Fox. *A Complete Recipe for Stochastic Gradient MCMC*. 2015. arXiv: 1506.04696v2 [math.ST].

[52] Mark Girolami. *MCMC methods derived from Diffusions, Geodesics and Foliations on Riemannian manifolds*. Apr. 2012. URL: http://www.csml.ucl.ac.uk/userdata/lunch_talks/2012_04_27_mg.pdf (visited on 04/03/2019).

[53] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2020.

[54] James R Munkres. *Analysis on manifolds*. CRC Press, 2018.

[55] Radford M Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[56] Radford M Neal. "Slice sampling". In: *The annals of statistics* 31.3 (2003), pp. 705–767.

[57] Frank Nielsen. "An Elementary Introduction to Information Geometry". In: *Entropy* 22.10 (2020). ISSN: 1099-4300. DOI: 10.3390/e22101100. URL: https://www.mdpi.com/1099-4300/22/10/1100.

[58] Ay Nihat. *Information geometry*. Cham, Switzerland: Springer Nature, 2017. ISBN: 9783319564777.

[59] Albert Parker and Colin Fox. "Sampling Gaussian distributions in Krylov spaces with conjugate gradients". In: *SIAM Journal on Scientific Computing* 34.3 (2012), B312–B334.

[60] Thomas Pennington. *GenericBayes.jl - Sampling with dually-flat geometry*. Apr. 2022. URL: https://github.com/tpght/GenericBayes.jl.

[61] Garvesh Raskutti and Sayan Mukherjee. "The Information Geometry of Mirror Descent". In: *Geometric Science of Information*. Ed. by Frank Nielsen and Frédéric Barbaresco. Cham: Springer International Publishing, 2015, pp. 359–368. ISBN: 978-3-319-25040-3.