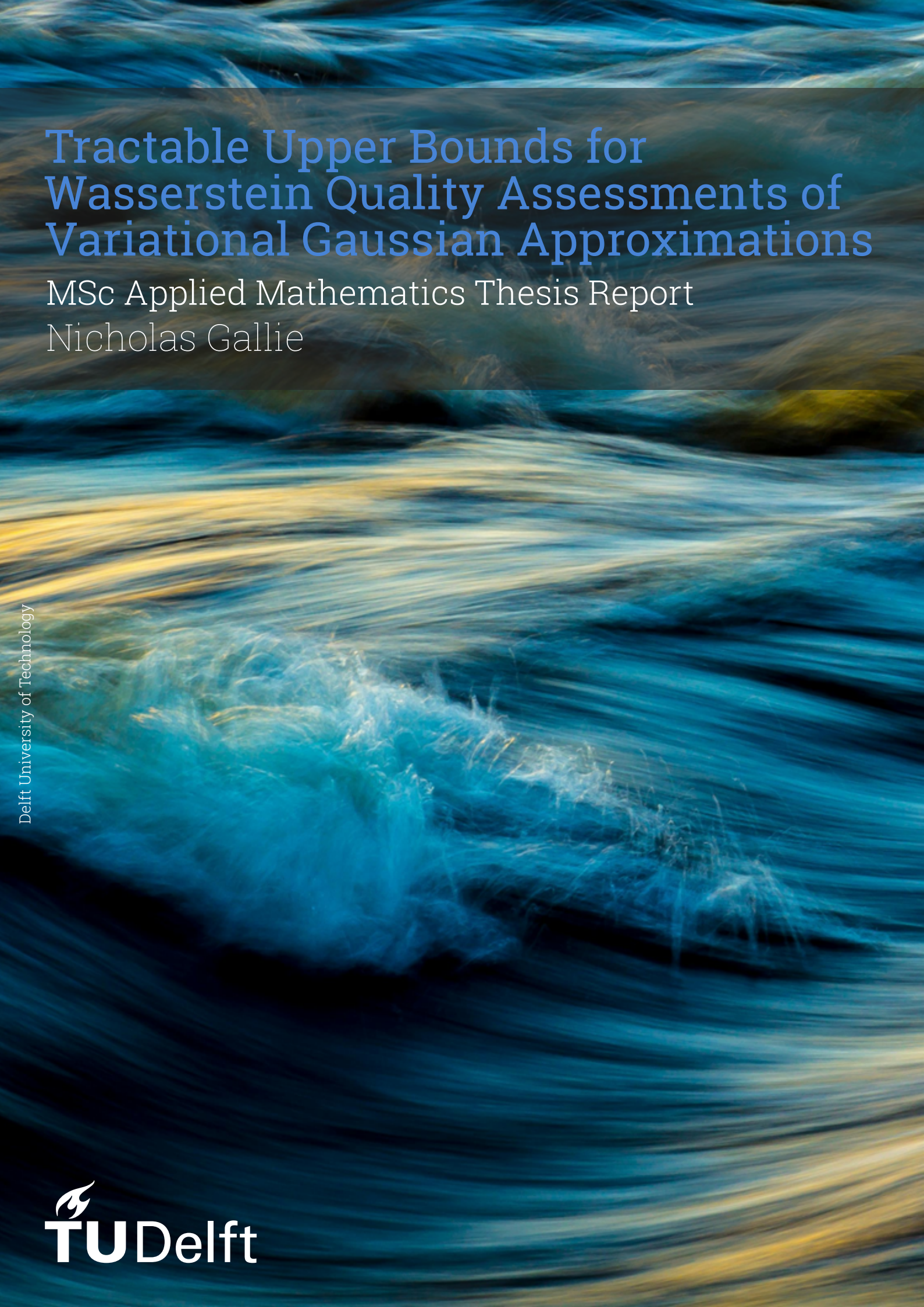

Tractable Upper Bounds for Wasserstein Quality Assessments of Variational Gaussian Approximations

MSc Applied Mathematics Thesis Report

Nicholas Gallie

Delft University of Technology

TU Delft

Tractable Upper Bounds for Wasserstein Quality Assessments of Variational Gaussian Approximations

by

Nicholas Gallie

Supervisor:	Dr. Richard Kraaij
Supervisor:	Dr. Joris Bierkens
External supervisor:	Dr. Havva Yoldaş
Project duration:	January 2025 — October 2025
Faculty:	Electrical Engineering, Mathematics and Computer Science

Cover:	Macro Photography of Water Waves, by PixaBay under CC0 (available on Pexels)
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Abstract

Variational inference comprises a family of statistical methods to obtain the optimal approximation of a target probability distribution using some reference class of distributions and a cost function, commonly the Kullback-Leibler (KL) divergence. Recent work on variational inference has yielded a fast, stable set of mean and covariance evolutions which dynamically yield variational Gaussian approximations $\{p_t^L\}_{t\geq 0}$ via a restriction to Gaussian measures of the well-known JKO scheme. The sequence of Gaussian measures thus generated converges towards the KL-optimal Gaussian approximation of the VI target e^{-V}: it may also be used to approximate the entire sequence of distributions $\{p_t^J\}_{t\geq 0}$ generated by a JKO gradient flow directed at this same target, which supports practical usage of Gaussian VI as well as fast, approximate modelling of the Fokker-Planck PDE. However, it is not immediately clear whether this Gaussian sequence offers valid, helpful approximations of the original JKO gradient flow. In this work, three upper bounds for the sequence of distances $\{W_2(p_t^L, p_t^J)\}_{t\geq 0}$ are obtained by exploiting the Riemannian structure of the W_2 manifold and the shared properties of the Gaussian and JKO evolutions. Numerical simulations support the validity of these bounds and test their performance in both ordinary and exceptional scenarios. One of the bounds may be computed solely using the Gaussian p_t^L and the potential V, thus offering a tractable estimator for the suitability of variational Gaussian approximations which retains the attractive properties of Wasserstein distances whilst avoiding their computational demands.

Preface

This text encapsulates approximately nine months of work I performed towards the end of my master's degree at TU Delft. My thesis project has taken me on an incredible journey through mathematical topics I thought were well beyond my grasp and towards the contemporary frontier of research into theoretical machine learning. I am immensely pleased to have had the opportunity to reach one small corner of this frontier during my master's degree, and I am immensely grateful to all those who supported me in doing so.

I would first like to thank my thesis supervisors, whose guidance and encouragement throughout this project were essential for its completion, and whose exceptional time and energy investments into my work did not go unnoticed. I would like express my sincere gratitude to Dr. Richard Kraaij, who initially proposed the objectives pursued in this work and sparked my interest in optimal transport theory and gradient flows: Richard was an unwavering source of support throughout my thesis project, assisting me with both theoretical and practical challenges as well as offering essential insight into how my results could be constructed. I would also like to extend my wholehearted thanks to Dr. Joris Bierkens, who introduced me to Richard and his project after I expressed an interest in working together: Joris never hesitated to offer me extensive and invaluable insight into statistics, machine learning and the many mathematical topics which I encountered during this project. Additionally, I would like to thank Dr. Havva Yoldaş for agreeing to join my thesis committee and taking the time to consider my work in detail.

I would also like to thank all of my friends in the TU Delft community for keeping me sane and happy throughout my time here. Of these, Max deserves special praise for his mathematical humour, intelligent critiques and eternal curiosity in my work. Special thanks also go to Valdimar and Erio for joining me on this journey through the Applied Mathematics degree. I am sincerely grateful to my colleagues in the IP organising team for the time and memories we shared together over the past two years, and for reminding me of the skills and challenges I can also pursue outside mathematics.

Last but certainly not least, I must thank my close friends and family for their relentless support, without which I would not have been able to complete this thesis project at all. Thank you, Dad, for always reminding me of what I can achieve when I put in the effort; and thank you most of all, Sarah, for always reminding me why it's worth the effort in the first place.

Nicholas Gallie
Rotterdam, September 2025

Declaration on the use of generative AI: I acknowledge the use of the generative AI model ChatGPT for seeking relevant literature, for debugging certain elements of my simulation code and for providing accessible explanations of unfamiliar mathematical topics. All mathematical results and writing produced for this project were created without the use of generative AI models.

Notation

The following table lists the principal notation used in this work. Please note that this table is only intended to serve as a general support reference, so it does not cover every instance of mathematical notation used in this text and unfortunately may not eliminate all ambiguities. Many notational choices made here are intended to reflect notation used in key references, in order to support further reading.

Symbol	Definition
ρ, ϱ, p	Probability density (or occasionally a measure: this should hopefully be clear from context).
μ, ν	Probability measure (or occasionally a density).
m	The mean of a (typically Gaussian) probability distribution.
Σ	The covariance matrix of a (typically Gaussian) probability distribution.
v_t	The velocity field associated with a particular JKO/Bures-JKO evolution p_t.
p_t^L	The (Gaussian) Bures-JKO probability density at time t, evolving to the Lambert et al. ODEs.
p_t^J	The JKO probability evolution at time t, evolving according to the JKO FPE.
u_t^i	One of the three upper bounds for the quantity $W_2(p_t^L, p_t^J)$ obtained in this work.
c_t, c_t^i	The bound coefficients for u_t^1, u_t^2 obtained in Chapter 4.
d	Dimension of the ambient space.
S_d	The space of symmetric matrices of size $d \times d$.
S_d^{++}	The space of symmetric positive-definite matrices of size $d \times d$.
π	Target probability density, to be approximated via the JKO or Bures-JKO schemes.
V	Potential function for the target π.
α	The modulus of convexity for the potential V.
β	The (inverse) temperature of the ambient space, as specified for the Langevin diffusion and the FPE.
Ω	The ambient vector space over which probability distributions are defined: this is typically $\mathbb{R}^d$, except in Chapter 5 where Ω refers to the specific discrete grid used to perform a particular experiment.
$\Pi(\mu, \nu)$	The set of couplings for the two input probability measures μ, ν.
T	Depending on the context, either: an optimal transport map $T(x)$, the stopping time of a numerical simulation or part of the notation for a tangent space (see following entry). The notation v^T is also used throughout this text to denote the (horizontal) transpose of the (column) vector v.
T_pM	Tangent space of the member p of a Riemannian manifold M.
$\mathcal{P}_2(\mathbb{R}^d)$	The Wasserstein W_2 manifold of probability measures over $\mathbb{R}^d$ with finite second moments.
$\mathrm{BW}(\mathbb{R}^d)$	The Bures-Wasserstein manifold over $\mathbb{R}^d$, i.e. the set of Gaussian probability densities over $\mathbb{R}^d$ equipped with the Wasserstein 2-distance.

Symbol	Definition
$\langle \cdot, \cdot \rangle_p$	p-inner product of the two input arguments.
$\lvert \cdot \rvert, \lvert\lvert \cdot \rvert\rvert$	The Euclidean norm of the input argument.
$\nabla_x f(x)$	The gradient w.r.t. x of the function f. If x is not specified and only ∇ is visible, the argument with which to take the gradient should hopefully be clear from context.
$\nabla \cdot f(x)$	The divergence of the function f w.r.t. its argument x.
$\Delta f(x)$	The Laplacian of the function f w.r.t. its argument x.
$\nabla_{W_2} F(p)$	The Wasserstein W_2 gradient (i.e. direction of fastest ascent within $\mathcal{P}_2(\mathbb{R}^d)$) for the functional F, when evaluated at the point p.
$\nabla_{\mathrm{BW}} F(p)$	The Bures-Wasserstein gradient (i.e direction of fastest ascent within $\mathrm{BW}(\mathbb{R}^d)$) for the functional F, when evaluated at the point p.
∂_t	Shorthand for $\frac{d}{dt}$, used in some sections,
δ	Used to denote first variations; the exact manner in which this is done is context-dependent - please see Section 2.5 for more details.
$\mathbb{E}_p(\cdot)$	The expectation of the input argument w.r.t. the probability distribution p. If p is not specified, then the choice of distribution should hopefully be clear from context.
W_t	A standard Brownian motion.
$W_2(\cdot, \cdot)$	The Wasserstein-2 distance between the two input arguments.
W_2	Either the Wasserstein-2 distance in general, or specifically the quantity $W_2(p_t^L, p_t^J)$ (particularly in Chapter 5): the correct interpretation should hopefully be clear from context.
$\mathrm{KL}(\cdot \lvert\lvert \cdot)$	The Kullback-Leibler divergence between the two input arguments.
$I(\cdot \lvert \cdot)$	The relative Fisher information between the two input arguments.
$F(p)$	A functional defined over probability measures p, most commonly the context-specific Kullback-Leibler divergence $\mathrm{KL}(p \lvert\lvert \pi)$.

Contents

1

Introduction

1.1. Context

The statistical approach to machine learning seeks to use statistical approaches to uncover relationships from large, complex data sets. The true functions which connect variables of interest are often unknown and may, in whole or in part, be fundamentally unknowable due to mathematical or practical restrictions. We must therefore seek approximations of these functions using statistical learning techniques. A fundamental example of the need to develop such approximations lies at the heart of Bayesian statistics: consider the goal of developing a stronger understanding about some variable of interest x which depends on another variable y, using some prior knowledge $p(x)$ together with an empirical data set $p(y|x)$. In Bayesian statistics, we obtain this *posterior* understanding $p(x|y)$ from the *prior* understanding $p(x)$ and the labelled data $p(y|x)$ through the following relationship:

$$p(x|y) \propto p(y|x)p(x) \tag{1.1}$$

This relationship is simple enough to describe in an abstract sense, but readers familiar with Bayesian statistics will know that complications arise when we attempt to convert our abstract concept of the posterior into an actual probability distribution, using the probability distributions $p(x), p(y|x)$. The normalisation coefficient required for $p(x|y)$ is often intractable, so practitioners must resort to either numerical approximation (e.g. by numerical integration or Monte Carlo methods) or statistical approximation. In this latter case, a suitable surrogate for the original function is sought using the available data and knowledge about its origins: for Bayesian statistics, the objective then becomes to find an approximation for the posterior. Unfortunately, the quality of these approximations cannot in general be known before they are tested against new data — which requires new data to be used for testing, along with time and other resources. If we were able to know the quality of our approximations while they are being computed, we could potentially exploit this insight to obtain these approximations more efficiently. The pursuit of this insight is necessary in modern machine learning, where loss functions are used to guide the parameters of neural networks towards appropriate values. As we shall see, however, there are certain learning frameworks whose mathematical properties yield further details on approximation quality without requiring additional data.

Let us now turn our attention towards variational inference (VI), which comprises a broad family of Bayesian statistical techniques seeking the ”best” approximation of a posterior distribution (see Section 2.3 for more details). Within the VI paradigm, the ”best” approximation of the posterior is the member of a chosen family of probability distributions which minimises the Kullback-Leibler (KL) divergence relative to the posterior: this minimisation is often achieved through some iterative algorithm which yields a sequence of probability measures converging towards the optimal approximation. On paper, minimisation of the KL divergence may be achieved through the well-known JKO scheme [57], which asymptotically achieves the VI objective by propagating the Fokker-Planck Equation (FPE) (2.9) towards the target posterior, thus generating a converging sequence of probability distributions $\{p_t^J\}_{t\geq 0}$.

One condition which permits the JKO scheme to be applied is for the VI target π to be *log-concave*, i.e. it must have the form $\pi \propto e^{-V(x)}$ for a convex *potential* function V: targets of this form arise naturally in many Bayesian scenarios. Although the JKO scheme offers an interesting and well-studied framework for researchers modelling the FPE, it is unfortunately not a very practical method to work with if the goal is to perform VI. Each of the discrete optimisation steps presents a challenging (or even intractable) optimisation problem for real-world usage, while numerical propagation of the FPE is computationally expensive and may introduce significant approximation errors.

To address these issues, the JKO scheme was recently employed by Lambert et al. in [65] to perform VI using Gaussian approximations. More precisely, the authors of [65] were able to demonstrate that a certain pair of mean-covariance ODEs (3.8) corresponds to a restriction of the JKO FPE to the space of Gaussian measures (i.e. the "Bures-JKO scheme"): by propagating these ODEs forward in time, we obtain a sequence of Gaussian distributions $\{p_t^L\}_{t\geq 0}$ converging towards the optimal VI approximation of a given target. We thus obtain a computationally cheaper alternative to the "full" JKO scheme based on Gaussian approximations: this alternative applies not just to the asymptotic approximation of the target but also to the entire sequence of intermediate propagation steps. The intermediate Bures-JKO step p_t^L is not intended to serve as an ideal [1] Gaussian approximation of a full JKO step p_t^J, but understanding the suitability of such approximations may nonetheless assist practitioners using the Bures-JKO ODEs by Lambert et al. (and, through further research, potentially related forms of Gaussian VI) by offering new stopping criteria and "early warning systems" for poor initialisation choices. Of course, these tools are already available via comparison functionals such as $\mathrm{KL}(\cdot||p_t^J)$: however, the bounds obtained in this project leverage both the advantages of optimal transport comparisons and the geometric behaviour of the JKO and Bures-JKO evolutions. Furthermore, the substitution of a JKO FPE with its Gaussian counterpart may have downstream applications wherever the FPE is used: having dedicated tools to monitor the quality of this substitution will undoubtedly benefit researchers and practitioners seeking to do so. The JKO/Bures-JKO dyad thus offers an intriguing case for studying Gaussian approximations of a dynamical process.

A different dynamical setting where Gaussian approximations are also used is Kalman filtering. Gaussian distributions are used by Kalman filters to model both the uncertainty inherent to their state processes and the noise in their measurement processes. This is true both of the original Kalman filter and of many of its variants, including the Unscented Kalman Filter (UKF), which is a popular algorithm for tracking and controlling non-linear systems in fields ranging from robotics to economics. As noted by Lambert et al. in [65] and as we shall see in Section 3.4, the mean-covariance ODEs of the Bures-JKO scheme are mathematically equivalent to a special case of the continuous-time UKF where the observation update is disregarded (i.e. where the Kalman gain is set to zero). We may then consider whether insight into the suitability of Gaussian approximations in the JKO/Bures-JKO case may be translated into the broader framework provided by the UKF. This would not be a straightforward adaptation due to the more general stochastic process described by the UKF (consider a potential function which varies over time) and so such a task lies beyond the scope of this thesis project: however, the work presented here takes a small step in this direction and hopefully encourages others to continue doing so.

The general question motivating this research project is: what if it were possible to know how appropriate these dynamic Gaussian approximations are — in real time? Of course, we cannot *know* this with absolute certainty in real-world settings, due to the unavoidable imperfections in our data collection and the unpredictability of most real-world dynamic processes. Moreover, the question of what constitutes a "good" approximation has yet to be settled: in this text, a "good" approximation is one which minimises the Wasserstein-2 distance, denoted here with the shorthand W_2. Wasserstein distances offer symmetric, flexible comparison metrics which interpret their arguments as distributions of "mass" which cost effort to move, which naturally resembles the particle-based interpretation of the JKO-scheme FPE as a Wasserstein gradient flow (see Subsection 2.1.3 and Section 2.2 for details). The interpretation of Wasserstein spaces as Riemannian manifolds also offers valuable geometric techniques which were

[1] In this context, the "ideal" Gaussian approximation of p_t^J would be that which minimises some comparison functional of interest, e.g. the forward KL divergence $\mathrm{KL}(p_t^J||\cdot)$, the reverse KL divergence $\mathrm{KL}(\cdot||p_t^J)$ or the Wasserstein W_2 distance $W_2(\cdot, p_t^J)$. In the first case, we know that p_t^L is not the optimal approximation of p_t^J due to mismatched moments: see Subsection 4.1.2 for an explanation. For the other two cases listed here, it has not been investigated whether p_t^L is the optimal approximation of p_t^J: however, these possibilities are likely disprovable through similar recourse to the moment-matching arguments used for the forward KL case above.

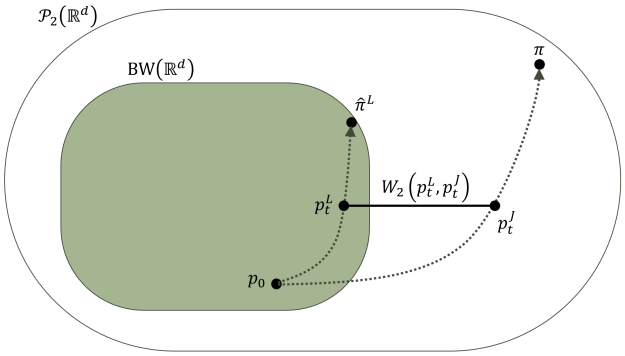

Figure 1.1: the basic scenario considered for this research project — p_t^J represents a JKO gradient flow converging towards a target π, and p_t^L represents a Bures-JKO gradient flow converging towards the optimal approximation of π. The central quantity of interest is the distance $W_2(p_t^L, p_t^J)$ and its evolution over time: the W_2 distance function offers a powerful, intuitive geometric comparison between the two gradient flows but is difficult to compute. In this work, three upper bounds u_t^1, u_t^2, u_t^3 for $W_2(p_t^L, p_t^J)$ were obtained and tested. For simplicity, this diagram assumes that $p_0^L = p_0^J$. The terms used here are explained more thoroughly in Chapter 4.

exploited to obtain the results in Chapter 4. However, Wasserstein distances are usually mathematically intractable to work with on paper and computationally expensive to work with empirically. A strong incentive thus arises to obtain reasonable bounds for W_2 which permit retaining the benefits of this distance metric whilst reducing the computational costs otherwise required. The search for such bounds comprises the objective of this thesis project, which seeks to answer three research questions:

- Is it possible to obtain theoretical upper bounds on $W_2(p_t^L, p_t^J)$, i.e. the W_2 distance between a Bures-JKO and a JKO propagation?
- Can such a bound depend only on some initial value $W_2(p_0^L, p_0^J)$, the Gaussian/Bures-JKO propagation $\{p_t^L\}_{t\geq 0}$ and the potential V?
- Are these bounds close enough to the true value of $W_2(p_t^L, p_t^J)$ to be useful?

1.2. Related work

The following section presents a brief summary of the academic literature most relevant to the research questions posed above. A broader scope of literature evaluations related to this project is distributed throughout Chapter 2.

During the literature search performed for this project, no publications were found which explicitly attempted to compare the JKO/Bures-JKO pair in the manner proposed above. The reason for this literature gap is not immediately clear, although a prior lack of applications may have been a factor, since recent developments such as the ODE-based Bures-JKO scheme in [65] do offer new incentives for comparisons such as $W_2(p_t^L, p_t^J)$ to be studied in more detail. For two general JKO gradient flows evolving towards the same target, an explicit formula [2] for $\frac{d}{dt}\frac{1}{2}W_2^2(\cdot,\cdot)$ is known ([90], Thm. 5.24): furthermore, we may control $W_2(p_t^L, p_t^J)$ using a Grönwall-style bound (see the proof of Corollary 3 in [65] for an explanation). A similar bound for Fokker-Planck evolutions with non-gradient drifts was obtained in [11]. The convergence rate of two JKO gradient flows with the same target was studied in more detail in [112], with a focus on relative entropy and information-based comparisons. In the case where both gradient flows are Gaussian, readers may be aware that the W_2 distance between two Gaussian measures has a closed-form expression (see (3.1)). Furthermore, it is not immediately clear what the downstream applications of this Bures-JKO/Bures-JKO pair would be: hence, we will not consider this case any further.

By contrast, there has been more research into the convergence behaviour of a single gradient flow. Some work has been performed directly on bounding the behaviour of the JKO FPE, not necessarily via

[2] To evaluate this formula, the Kantorovich potentials for the W_2 distance in $\frac{d}{dt}\frac{1}{2}W_2^2(\cdot,\cdot)$ must be known, which is a highly non-trivial requirement. As demonstrated in Chapter 4, however, it is still possible to make productive use of this formula to bound the derivative $\frac{d}{dt}\frac{1}{2}W_2^2(\cdot,\cdot)$ even when the Kantorovich potentials are not themselves known.

the W_2 distance ([30], Proposition 3.8). Bounds have also been obtained for alternative JKO-style gradient flows, often involving substitutions of the Kullback-Leibler evaluation function and/or the Fokker-Planck dynamics ([16], Section 4.2; [10], Thm. 3.4; [13], Proposition 2.9; [71], Thm. 3.7). Research into the Bures-JKO scheme has also yielded convergence bounds for this particular gradient flow ([31], Section 5; [65], Corollary 3 and Thm. 4). More general upper bounds on the W_2 distance are also an object of study: well-known examples include the Evolution Variational Inequality ([75], Section 3.1) and the Talagrand's transportation inequality [99]; a convenient summary of transport inequalities may be found in [42], and other controls for Wasserstein distances are listed in [81]. Besides the interpretation of VI as a Wasserstein gradient flow, Wasserstein distances have also been studied as a tool for assessing the quality of VI approximations [9, 53] — an application which may also be extracted from the bounds provided in this report, and which is implicitly employed in Chapter 5.

1.3. This work

The key knowledge contributions generated by this thesis project are:

- Three distinct upper bounds u_t^1, u_t^2, u_t^3 were obtained for the W_2 distance between the JKO and Bures-JKO propagations.
- One of these bounds (u_t^3) requires only an initial W_2 value, the Bures-JKO propagation and the potential in order to be computed.
- Numerical experiments suggest that these bounds provide reasonable approximations of W_2 in ordinary settings; however, this may not be the case in extreme cases or when the assumption of log-concavity is broken.

The remainder of this text is structured as follows:

- Chapter 2 provides the essential mathematical background required for this project. Familiarity is assumed with Bayesian statistics, (measure-theoretic) probability theory, real analysis, vector/-matrix calculus and some basic functional analysis. Chapter 2 also provides a broader literary context for this text: the mathematical topics covered here interact both with each other and with other relevant topics in mathematics and machine learning which readers may be interested in.
- Chapter 3 provides a detailed account of the central result published in [65], which establishes the Bures-JKO ODEs as being a consequence of a Wasserstein gradient flow. Two distinct proofs of this result are reproduced here, supported by detailed computations and expanded commentary intended to deepen readers' understanding of the Bures-JKO scheme and support the search for W_2 bounds.
- Chapter 4 provides three new upper bounds for the W_2 distance between the JKO and Bures-JKO propagations, along with discussions on their relative theoretical merits and limitations. Two of these bounds are proposed as conjectures, whereas for the remaining bound a complete proof is provided.
- Chapter 5 provides the descriptions and outcomes of numerical experiments performed to test the bounds obtained in Chapter 4. A detailed explanation of the approach used is available in Chapter 5, along with descriptions and plots of the experimental results, which are then evaluated and compared with each other.
- Chapter 6 provides reflections on the theoretical and practical results obtained, along with suggested directions for future research.

2

Mathematical Background

2.1. Optimal Transport

This section contains an overview of the background in optimal transport theory and gradient flows needed for the analysis presented later in this report. The primary reference for the definitions and notation used in this section is Filippo Santambrogio's book *Optimal Transport for Applied Mathematicians* [90]: for a more comprehensive treatment of the concepts presented below, we refer readers to Santambrogio's book or to [103].

2.1.1. Preliminaries

The original problem which defines optimal transport theory, first posed by Monge, concerns the most efficient way to transport mass from one distribution to another. Let X, Y be two sets with sets of probability measures $\mathcal{P}(X), \mathcal{P}(Y)$: furthermore, let $c(x, y)$ be the cost of transporting a particle at location $x \in X$ to $y \in Y$. The Monge Problem (MP) is:

$$\inf_T \left\{ \int_X c\left(x, T(x)\right) d\mu(x) : T_{\#}\mu = \nu \right\}. \tag{MP}$$

In the above we have: an initial mass distribution $\mu \in \mathcal{P}(X)$, a target mass distribution $\nu \in \mathcal{P}(Y)$ and a possible mapping $T : X \to Y$ [1]. The objective is to choose a mapping T which minimises the average transport cost expressed by the integral (i.e. the expectation of $c(x, T(x))$ under μ). The condition $T_{\#}\mu = \nu$, where $(T_{\#}\mu)(A) := \mu(T^{-1}(A))$ is the pushforward measure of μ through T, ensures that T actually maps the mass distributed by μ to the mass distributed by ν.

The constraint $T_{\#}\mu = \nu$ makes the Monge Problem difficult to find solutions for. The Kantorovich Problem allows the question of optimal transport to be framed in a different way:

$$\inf_\gamma \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}. \tag{KP}$$

Now, instead of considering an optimal transport map T, we consider an optimal transport *plan* $\gamma : X \times Y \to \mathbb{R}$, whereby $\gamma(x, y)$ conceptually represents the amount of mass moving from the point $x \in X$ to the point $y \in Y$. If, for a point x, multiple destinations y receive non-zero mass, then there cannot be a map T associated with γ: hence, the constraint $\gamma \in \Pi(\mu, \nu)$ is required, where $\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) : (\pi_X)_{\#}\gamma = \mu, (\pi_Y)_{\#}\gamma = \nu\}$. Here, π_X and π_Y represent the projections of $X \times Y$ onto X and Y, respectively.

[1] In general, μ, ν needn't be probability measures or have the same total mass, and may instead be general measures. However, the development of Wasserstein gradient flows and the core research behind this project is defined exclusively for probability measures, so μ, ν will be assumed as such for the remainder of this report.

The Kantorovich constraint is easier to work with than the Monge constraint, due to its linearity; as such, it becomes possible to prove that a solution exists for (KP) under mild conditions. To obtain a form for this solution, a dual to (KP) is defined:

$$\sup_{\varphi,\psi}\left\{\int_X \varphi(x)d\mu(x)+\int_Y \psi(y)d\nu(y) : \varphi\in C_b(X), \psi\in C_b(Y), \varphi(x)+\psi(y)\leq c(x,y)\right\}. \tag{DP}$$

It can be shown that the supremum of this problem is equal to the infimum of (KP). Note that $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ must be continuous and bounded functions such that $\varphi + \psi \leq c$. If we assume $c(x, y) = h(x - y)$ for some strictly convex function h, then the solution to (DP) must be obtained via a pair of functions φ, φ^c (known as Kantorovich potentials), where $\varphi^c(y) = \inf_x\{c(x, y) - \varphi(x)\}$. Furthermore, a formula for the optimal map T can be obtained in certain cases, such as when μ is absolutely continuous w.r.t Lebesgue measure:

$$T(x) = x - (\nabla h)^{-1}(\nabla\varphi(x)) \tag{2.1}$$

2.1.2. Wasserstein distances

In (KP), one frequent choice for the cost function c (or, equivalently, for the strictly convex $h(x - y)$) is $c(x, y) = |x - y|^p$, for $p \in [1, \infty) \cup \{\infty\}$. By inserting this cost into (KP), we obtain the Wasserstein p-distance between two measures μ and ν (both defined on a space Ω) [2]:

$$W_p(\mu,\nu) = \min_{\gamma}\left\{\int_{\Omega\times\Omega}|x-y|^p d\gamma(x,y) : \gamma\in\Pi(\mu,\nu)\right\}^{1/p}. \tag{2.2}$$

The Wasserstein distance is a true distance and can be used to define a metric space, i.e. the Wasserstein p-space $\mathbb{W}_p := (\mathcal{P}_p(\Omega), W_p)$ (where $\mathcal{P}_p(\Omega)$ is the set of Ω probability measures which are L^p-integrable).

Within the space $\mathbb{W}_p$, we can define curves of measures between μ and ν. To start, let us define $\pi_t(x, y) = (1 - t)x + ty$ for $t \in [0, 1]$: π_t is, of course, a convex combination of two points in Ω. By considering such combinations of all points in Ω as the pre-images and images of an optimal transport plan γ, we can define the sequence of measures $\mu_t := (\pi_t)_{\#}\gamma, t \in (0, 1)$, which is a curve (more precisely: a constant-speed geodesic) that moves through $\mathcal{P}_p(\Omega)$ from μ to ν.

One of the core behaviours we must see along such geodesics is the preservation of mass: if, at some point $x \in \Omega$, the mass assigned by μ_t is decreasing over time, then that mass must be redirected somewhere else (and vice versa). This principle underlies the continuity equation:

$$\partial_t\mu_t + \nabla\cdot(v_t\mu_t) = 0 \tag{2.3}$$

The vector field $v_t(x)$ represents the velocities (of mass) at each point $x \in \Omega$: the quantity $v_t(x)\mu_t(x)$ is therefore analogous to momentum. The divergence $\nabla \cdot (v_t(x)\mu_t(x))$ represents the net flow of mass inwards or outwards from the point x, and must balance exactly with the change in mass at x as represented by $\partial_t\mu_t(x)$.

2.1.3. Gradient flows in W2

Let us now restrict our attention to $\Omega = \mathbb{R}^d$. By definition, a gradient flow is a system of equations describing the motion of a particle $x(t)$ starting at some point $x(0) \in \mathbb{R}^d$ and always moving in the direction where some function f decreases most rapidly. We can write this definition as follows:

[2] In general, μ and ν do not need to be defined on a common space for the Wasserstein distance definition to be valid: indeed, this is one advantage of Wasserstein distances over other comparisons such as the KL divergence (KL). Nonetheless, this simplifying assumption is frequently made when defining Wasserstein distances, for instance in [65, 90], as many applications of Wasserstein distances do require comparing measures defined on a common space, e.g. on $\mathbb{R}^d$.

$$\begin{aligned} x(0) &= x_0 \\ x'(t) &= -\nabla f(x(t)) \end{aligned} \tag{2.4}$$

In the $\mathbb{W}_2$ space, we can use the continuity equation as described in (2.3) to obtain the following equation for a curve of probability measures $\varrho_t \subset \mathcal{P}_2(\Omega)$ and a functional $F : \mathcal{P}_2(\Omega) \to \mathbb{R}$:

$$\partial_t \varrho_t - \nabla \cdot \left(\varrho_t \nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right) \right) = 0 \tag{2.5}$$

It is clear that (2.5) is a special case of (2.3), implying that the term $\varrho_t \nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right)$ represents the "momentum" of particles distributed according to the evolving measure ϱ_t. Equation (2.5) employs the term $\frac{\delta F}{\delta \varrho}(\varrho_t(x))$, which is called the "first variation"[3] in [90]. This is the "functional derivative" definition of the first variation, as defined in Subsection 2.5.1, which has the following interpretation in this context: $\frac{\delta F}{\delta \varrho}$ resembles a "partial derivative" of F w.r.t. ϱ, and measures the sensitivity of the value of F to a change in the value of ϱ at a point $x \in \Omega$. The velocities of particles distributed by ϱ_t therefore depend on the functional F.

The functional F is used to define a "contour map" of its values over $\mathcal{P}_2(\Omega)$: a particle with an initial location in this landscape will "roll downhill" towards a local minimum. This is the fundamental motion being described by the gradient flow in (2.5). When used by authors within the mathematical, scientific and machine learning communities, the term "Wasserstein gradient flow" refers to the particle ϱ_t and its evolution in the $\mathbb{W}_2$ space over time as dictated by (2.5) or the discretisation (2.6), which explicitly requires the use of the W_2 distance.

For a time step of size τ, it is possible to create a time discretisation of the gradient flow of ϱ_t (as written in (2.5)) as an optimisation problem, regularised by the incorporation of the W_2 distance:

$$\varrho^\tau_{(k+1)} \in \underset{\varrho \in \mathcal{P}(\Omega)}{\operatorname{argmin}} \left(F(\varrho) + \frac{W_2^2(\varrho, \varrho^\tau_{(k)})}{2\tau} \right) \tag{2.6}$$

This provides an implicit Euler scheme which can be used to propagate ϱ_t through time and, in principle, perform numerical approximation of (2.5)[4]. Indeed, by choosing F appropriately, we can obtain various PDEs through (2.5); alternatively, we can propagate and/or study these PDEs through the discretisation (2.6). Prominent examples are the heat equation $\delta_t \varrho - \Delta \varrho = 0$ and the Fokker-Planck Equation, which is studied in more detail in Section 2.2. Besides the new mechanism for numerical simulation offered by (2.6), the characterisation of these equations as Wasserstein gradient flows can facilitate their study by offering new ways to show the existence of solutions and find properties of these solutions (e.g. convergence rate, stability).

2.1.4. Applications in machine learning

There are extensive connections between optimal transport theory and machine learning, covering both theoretical and practical aspects of both disciplines. An extensive survey of these connections lies beyond the scope of this project: the purpose behind this subsection is merely to offer an outline of key appearances of Wasserstein distances and Wasserstein gradient flows in machine learning research.

Wasserstein distances are commonly used by machine learning researchers and practitioners alike. The W_1 distance for discrete and empirical distributions can be estimated via linear programming meth-

[3] Readers who are unsure how this term is being used here are directed to Section 2.5, where an attempt has been made to define and disentangle the various uses of the "first variation" in relevant literature.

[4] In practice, the Wasserstein distance is often intractable and poses a challenging optimisation problem itself, thus diminishing the viability of this approach. Nonetheless, there have been attempts to use (modified) JKO discretisations to model certain systems (e.g. [15, 17, 18]), as well as theoretical efforts to study JKO-based simulations using tractable approximations to the Wasserstein distance (e.g. [17, 86]). More recently, there has also been research into neural network-based approximations which make the JKO step tractable (e.g. [82, 109]).

ods (a fact that is exploited in, e.g. [69, 88]): additionally, the W_2 distance for one-dimensional distributions can be computed analytically using the distribution CDFs. Implementations for computing the Wasserstein distance can be found in libraries for R and Python [38, 101, 105]. Academically, the history of applying Wasserstein distances/gradient flows to machine learning problems stretches at least as far back as 2000 [88]: early applications were found for computer vision tasks (e.g. image retrieval [88], contour matching [43], histogram comparison [69]), with another strand of research seeking improved computational approximations of Wasserstein distances [84, 94], including the well-known Sinkhorn approximation [27]. Wasserstein distances have also found numerous applications in neural network-based models: examples include the Wasserstein GAN [5], Wasserstein-based object detection [47] and open-set classification in signal processing [113]. More fundamental research into the benefits and limitations of the usage of Wasserstein distances in ML has also been performed (e.g. [39, 62, 95]). Conversely, Wasserstein distances have themselves been the subject of approximation using machine learning models [22, 70], as have Wasserstein gradient flows [1, 74, 106, 109], with this research broadly seeking improvements in speed, accuracy and versatility over existing approaches and approximations.

2.2. The JKO scheme and the Fokker-Planck Equation

The "JKO scheme" refers to the discretisation (2.6), applied to the functional F:

$$\begin{aligned} F(\varrho) &:= \mathcal{F}(\varrho) + \mathcal{V}(\varrho) \\ \mathcal{F}(\varrho) &:= \int_\Omega (\log \varrho(x))\, \varrho(x) dx \\ \mathcal{V}(\varrho) &:= \int_\Omega V(x) \varrho(x) dx \end{aligned} \tag{2.7}$$

Applying (2.7) to (2.6) yields the original problem studied by Jordan, Kinderlehrer and Otto in their publication [57], after which the "JKO scheme" is named. The term $V(x)$ is the *potential* of the stochastic particle $X_t \in \Omega$ governed by the following SDE, known as the *Langevin diffusion*:

$$dX_t = -\nabla V(X_t) + \sqrt{2\beta^{-1}} dW_t \tag{2.8}$$

Loosely speaking, the particle X_t will move preferentially towards regions of Ω with lower values of the potential $V(X_t)$: this principle drives modelling applications of the Langevin diffusion in the natural sciences. The parameter β acts as an "ambient temperature" which controls the overall speed at which X_t moves [5]. Without loss of generality, we shall in later chapters follow [65] and other relevant publications by assuming $\beta = 1$: for completeness, however, β is displayed in (2.9) and (2.10) below. At any time $t \geq 0$, the marginal probability distribution (density) for the location of X_t in Ω is given by the Fokker-Planck Equation (FPE):

$$\frac{d}{dt}\varrho_t = \beta^{-1} \Delta \varrho_t + \nabla \cdot (\varrho_t \nabla V) \tag{2.9}$$

The principal contribution made by [57] is that the FPE (2.9) arises as the limiting case as $\tau \to 0$ of the discretisation (2.6) applied to the functional F from (2.7). Under suitable conditions for V (in particular: convexity), the FPE as described in (2.9) will converge towards the stationary distribution

$$\begin{aligned} \pi(x) &:= \frac{1}{Z} \exp\left(-\beta V(x)\right) \\ Z &:= \int_\Omega \exp\left(-\beta V(x)\right) dx. \end{aligned} \tag{2.10}$$

[5] This effect is perhaps most visible through the FPE (2.9), where β^{-1} has a linear influence on the diffusion term $\Delta \varrho_t$. The influence of β^{-1} is ubiquitous throughout the ambient space Ω, which is why it is referred to as a "heat bath" in some publications, e.g. [51].

By choosing V to be formed through the combination of a prior probability density and a likelihood based on data (this is possible, for instance, when both the prior and likelihood are exponential-family distributions), we can have the FPE (2.9) converge towards a Bayesian posterior. Furthermore, by inserting $V(x) = \beta^{-1} \log \pi(x)$ into (2.7), it follows that this motion towards the posterior is governed by the KL divergence between ϱ_t and π. We can see this clearly by rewriting the discretisation (2.6):

$$\varrho^{\tau}_{(k+1)} \in \underset{\varrho}{\operatorname{argmin}} \left(\mathrm{KL}\left(\varrho||\pi\right) + \frac{W_2^2(\varrho, \varrho^{\tau}_{(k)})}{2\tau} \right) \tag{2.11}$$

At each time step, the procedure (2.11) will reduce $\mathrm{KL}\left(\varrho_{(k+1)}||\pi\right)$: this process will be repeated until no further reduction is possible, i.e. until $\mathrm{KL}\left(\varrho_{(k+1)}||\pi\right)$ has reached a minimum[6]. The JKO scheme therefore provides a dynamical mechanism for performing variational inference.

2.3. Variational Inference

2.3.1. Overview

In this section, a brief overview of variational inference is provided. The primary reference used is Christopher Bishop's *Pattern Recognition and Machine Learning* [8].

Variational inference (VI) derives its name from the *variational principle* whereby an optimal function is selected by minimising or maximising the value of a functional dependent on that function: a partial description of the *calculus of variations* used to perform this optimisation is provided in Section 2.5. More precisely, variational inference is a Bayesian statistical approach which seeks to minimise some functional measuring the similarity between a freely-chosen probability density p (freely chosen from some set $\mathcal{P}(\Omega)$) and the Bayesian posterior π. VI is most commonly performed using the Kullback-Leibler (KL) divergence:

$$\underset{p \in \mathcal{P}(\Omega)}{\operatorname{argmin}} \left\{ \mathrm{KL}(p||\pi) \right\} \tag{VI}$$

The Kullback-Leibler divergence between two probability densities is defined as:

$$\begin{aligned} \mathrm{KL}(p||\pi) &:= \mathbb{E}_p \left(\log \left(\frac{p}{\pi} \right) \right) \\ &= \int_{\Omega} \log \frac{p(x)}{\pi(x)} p(x) dx \end{aligned} \tag{KL}$$

As a functional for comparing an arbitrary density p with a target posterior π, (KL) has the favourable property that $\mathrm{KL}(p||\pi) \geq 0$, with $\mathrm{KL}(p||\pi) = 0$ if and only if $p = \pi$ almost everywhere (w.r.t. some reference measure). However, it should be noted that, unlike the Wasserstein distance, the KL divergence is not a true distance metric, as it is not symmetric and does not satisfy the Triangle Inequality.

The objective of variational Bayesian inference is to use the KL divergence to find the best approximation p^* to a (typically intractable) posterior π. Note that if π is intractable, then so is (KL). However, it is possible to reformulate (VI) into a tractable problem as follows, using $q(x, y) = q(y|x)q(x)$ for the prior $q(x)$ and the likelihood $q(y|x)$ used in the construction of π.

[6] Depending on the choices of V and $\mathcal{P}(\Omega)$, this minimum may be local or global. A common simplifying assumption is to require that V be convex, such that π is log-concave: for suitable choices of $\mathcal{P}(\Omega)$ (e.g. the set of Gaussian densities), we obtain a convex landscape from $\mathrm{KL}\left(\cdot||\pi\right)$ with a unique minimiser.

$$\begin{aligned}
\mathrm{KL}(p||\pi) &= \mathbb{E}_{p(x)}\left(\log\frac{p(x)}{\pi(x|y)}\right) \\
&= \mathbb{E}_{p(x)}\left(\log\frac{p(x)}{\frac{q(x,y)}{q(y)}}\right) \\
&= \mathbb{E}_{p(x)}\left(\log\frac{p(x)}{q(x,y)}\right) + \mathbb{E}_{p(x)}\left(\log q(y)\right) \\
&= -\mathcal{L}(p) + \log q(y)
\end{aligned} \tag{2.12}$$

In (2.12), the term $\log q(y)$ does not depend on our choice of p, so the question of minimising (KL) is equivalent to maximising the so-called Evidence Lower Bound (ELBO) $\mathcal{L}(p) := \mathbb{E}_{p(x)}\left(\log\frac{q(x,y)}{p(x)}\right)$. Since the likelihood and the prior may be assumed to have tractable forms, we have a viable expression for $q(x,y)$ and are able to compute $\mathcal{L}$ in practice to perform variational inference.

Different possibilities for obtaining variational approximations of π by making various restrictions to the possible distributions $\mathcal{P}(\Omega)$ which we can draw p from. A simple and common choice is to restrict $\mathcal{P}(\Omega)$ to be the set of product distributions across the dimensions of Ω, i.e. to assume independence between the constituent variables of Ω: $\mathcal{P}(\Omega) = \left\{p : p(x) = \prod_{i=1}^{d} q(x_i)\right\}$. This is known as *mean-field approximation*: optimising the ELBO in this setting ican be performed by the expectation maximisation (EM) algorithm. However, the assumption of independence between the components of p necessarily implies that any dependencies between the components of x will not be taken into consideration, which reduces the informativeness of p^* — perhaps substantially so. This provides a strong incentive to develop VI using other choices of $\mathcal{P}(\Omega)$.

2.3.2. Gaussian VI

Gaussian VI, A.K.A. "variational Gaussian approximation", is (VI) using Gaussian probability densities to approximate the posterior π. The restriction to Gaussian densities offers a partial solution to the dependency problem described at the end of Subsection 2.3.1. If we model a random variable $X \in \mathbb{R}^d$ using a Gaussian density, i.e. if $X \sim N(m, \Sigma)$, then linear dependencies between the constituent dimensions of X can be captured by the covariance terms in Σ. Non-linear dependencies cannot be modelled in this way, but the ability to capture linear dependencies still offers a substantial improvement over mean-field approximation (from Subsection 2.3.1) and is often sufficient for real-world applications. Indeed, this and many other favourable properties of Gaussian densities collectively motivate a strong interest in finding methods to perform Gaussian VI.

The early history of Gaussian VI is enmeshed inside that of machine learning research. During the literature search performed for this project, the earliest publication found which specifically explores Gaussian VI is the 1993 paper [50], where Gaussian VI (with diagonal covariances and a Gaussian posterior) is applied to fit neural network parameters. This method is called "ensemble learning"[7] in the follow-up paper [6], which extends the work of [50] to incorporate non-diagonal covariances. In [93], the usage scope of Gaussian VI is extended to the optimisation of hyperparameters for SVM and Gaussian Process classification. Some years after [50], the publication [52] returns to the idea of using Gaussian VI to fit neural networks, this time for multi-layer perceptron models. In 2009, the authors of [78] suggest that prior literature on Gaussian VI up to that point was scant due to the need to estimate $O(d^2)$ parameters (in the covariance), and propose an $O(2d)$ workaround using Gaussian processes. [19] studied Gaussian VI directly, obtaining various properties including sufficient conditions to establish convexity and differentiability of the Gaussian VI objective — as required for the publication [65] studied in detail for this project.

More recent work has continued to uncover new algorithms for performing Gaussian VI, for instance in [60, 64]. Both of these publications specifically define *online* algorithms for variational Gaussian

[7] This name may prove somewhat confusing for readers from the machine learning community, where "ensemble learning" now refers to a technique whereby multiple models are trained on part or all of a data set and their outputs combined (q.v. "Mixture of Experts").

approximation of a posterior — that is, algorithms which do not need to process the entire data set at once, which can impose significant memory requirements. The follow-up paper [63] offers an alternative to the "R-VGA" algorithm proposed in [64] with reduced memory requirements: extensions and generalisations of this method are proposed in [33, 56].

2.4. Kalman Filtering

Kalman filtering consists of a family of algorithms which generate sequential estimates of some unobserved process using observations generated by or otherwise related to that process. The ODEs provided by [65] for Gaussian VI are claimed by Lambert et al to be equivalent to a special case of the continuous limit of the Unscented Kalman Filter: this claim is studied more thoroughly in Section 3.4. However, the connection between 3.8 and 2.19, as explained in 3.4, may not be immediately clear to readers unfamiliar with the Unscented Kalman Filter. Therefore, a brief introduction to Kalman filtering is provided here: for more comprehensive explanations, we refer readers to [92].

2.4.1. Basic Kalman Filter

The Kalman Filter (KF) is an algorithm that generates estimates over discrete time steps $k = 0, 1, 2, ...$ for a *state process* x_k, which is a vector of unobserved variables that evolve over time. Estimates are obtained by combining initial assumptions about the unobserved vector with observations from a *measurement process* y_k, where at each time step k the observation y_k is assumed to depend in a known (or at least estimable), linear manner on x_k. The outcome of this process is a sequence of updates for the estimated mean vector $\hat{x}_k$ and covariance matrix P_k associated with the state process. The KF contains two steps —a prediction step and an update step— which, while intended to be alternated, can be performed in any order depending on the availability of observations from y_k: this flexibility is one of the key practical advantages of Kalman filtering.

The prediction step uses the prior mean and covariance estimates of the state process to produce new estimates of these quantities. Mathematically, this step can be written as follows:

$$\begin{aligned} \hat{x}_k &= F_k \hat{x}_{k-1} + B_k u_k \\ P_k &= F_k P_{k-1} F_k^T + Q_k. \end{aligned} \tag{2.13}$$

In the above, the following terms are used:

- $\hat{x}_{k-1}$: the previous estimate of the mean vector for the state process.
- F_k: the prediction matrix (A.K.A state transition model), which establishes a linear relationship between $\hat{x}_{k-1}$ and $\hat{x}_k$.
- u_k: known external influences on x_k, i.e. a control vector.
- B_k: the control matrix, which converts the control vector into the state process' coordinate space.
- Q_k: the covariance matrix of the noise in the state process, which is assumed to be Gaussian in nature.

The Kalman Filter's update step combines the estimates $\hat{x}_k, P_k$ with an observation y_k and its associated noise covariance R_k to produce new estimates $\tilde{x}_k, \tilde{P}_k$. This is performed as follows:

$$\begin{aligned} \tilde{x}_k &= \hat{x}_k + K(y_k - H_k \hat{x}_k) \\ \tilde{P}_k &= P_k - K H_k P_k \\ K &= P_k H_k^T \left(H_k P_k H_k^T + R_k \right)^{-1}. \end{aligned} \tag{2.14}$$

The additional terms introduced in (2.14) are:

- H_k: the observation model, which specifies the (presumed linear, known) relationship between the states x_k and the observations y_k. Equivalently, H_k transforms vectors from the state space to the observation space.

- R_k: the covariance matrix of the observation noise, which is assumed to be Gaussian.
- K: the Kalman gain. This term describes the relative influence of y_k over $\hat{x}_k$ on $\tilde{x}_k$. This can be seen by rewriting $\tilde{x}_k$ as an interpolation between $\hat{x}_k$ and y_k: $\tilde{x}_k = (I - KH_k)\hat{x}_k + Ky_k$. A "larger" K (as measured by its determinant) will place a greater weight on the measurement y_k, whereas a "smaller" K will favour the prediction $\hat{x}_k$. Importantly for Section 3.4: if K is the zero matrix, then $\tilde{x}_k$ depends entirely on $\hat{x}_k$, which is equivalent to using only the Kalman Filter's prediction steps without any observations.

Even if no observations are available, the KF must be initialised with some user-supplied $\hat{x}_0, P_0$. Furthermore: in the basic implementation of the filter, all the other terms used above (namely: $F_k, u_k, B_k, Q_k, H_k, R_k$) must be supplied by the user.

The KF can be described as a sequential Bayesian model, in the manner as Hidden Markov Models (with which Kalman Filters share a common graphical structure), MCMC techniques and many practical implementations of VI. To briefly summarise this interpretation: at each update step, the KF combines prior knowledge $\hat{x}_k$ with an observation y_k to produce a posterior estimate $\tilde{x}_k$. This procedure can be repeated whenever new observations become available, as the Markov property is assumed to hold along the sequence of hidden states, i.e. $p(x_k|x_0, ..., x_{k-1}) = p(x_k|x_{k1})$. Under suitable conditions (linearity, white noise from β_t, η_t), the mean and covariance updates prescribed in equations (2.14) are in fact the optimal estimators (according to the mean-squared error) for these moments of x_k. For a more comprehensive explanation of the Bayesian nature of Kalman filtering, we refer readers to [92], Section 6.

As presented above, the KF is constructed for a discrete-time Markov process

$$x_k = F_k x_{k-1} + B_k u_k + w_k, \tag{2.15}$$

where $w_k \sim N(0, Q_k)$. If each step $k \to k+1$ corresponds to a fixed-size time step Δt, then as we take the limit of (2.15) as $\Delta t \to 0$ (after finding appropriate infinitesimal versions for F_k, H_k, etc.) we obtain the continuous-time Markov process:

$$\begin{aligned} dx_t &= F(t)x_t dt + B(t)u(t)dt + d\beta_t \\ dy_t &= H(t)x_t dt + d\eta_t \end{aligned} \tag{2.16}$$

where β_t and η_t are independent Brownian motions with diffusion matrices $Q(t), R(t)$ respectively. Since, from (2.16), we can write $dx_t = f(x_t, t)dt + g(x_t, t)dW_t$ (i.e. a generalised Itô process), we see that the marginal probability density p_t of x_t obeys the Fokker-Planck equation and is generally intractable as a result. However, we are still able to obtain information about the location of x_t by tracking its moments. Equations (2.16) represent the state- and measurement-space models of the Kalman-Bucy Filter (KBF), which prescribes the following ODEs to track the mean $\hat{x}_t$ and covariance P_t of x_t ([55], Thm. 7.3):

$$\begin{aligned} \frac{d\hat{x}_t}{dt} &= F(t)\hat{x}_t + B(t)u(t) + K(t)\left(y_t - H(t)\hat{x}_t\right) \\ \frac{dP_t}{dt} &= F(t)P_t + P_t F^T(t) + Q(t) - P_t H(t)R^{-1}(t)H^T(t)P_t \\ K(t) &= P_t H^T(t)R^{-1}(t). \end{aligned} \tag{2.17}$$

2.4.2. Unscented Kalman Filter

One limitation of the basic KF is the assumption of linearity within the state-space and measurement models, as represented by the matrices F_k, H_k in (2.13) and (2.14), respectively. This limitation is also inherent to the Kalman-Bucy filter for continuous-time state space models. Let us drop the assumption of linearity to consider more general continuous-time states taking the form:

$$\begin{aligned} dx_t &= f(x_t, t)dt + L(t)d\beta_t \\ dy_t &= h(x_t, t)dt + V(t)d\eta_t. \end{aligned} \tag{2.18}$$

In this system of (Itô) SDEs, we have replaced the matrices F_t, H_t with more general non-linear transformations $f(\cdot, t), h(\cdot, t)$. The matrices $L(t), V(t)$ are the dispersion matrices of x_t, y_t, serving to transform the diffusions β_t, η_t into the state- and measurement-space coordinates, respectively. As with the Kalman-Bucy model in (2.16), the marginal distribution of x_t satisfies the FPE and is generally intractable. Furthermore, the generalisation introduced by f and h makes finding optimal solutions for the mean m_t and covariance P_t of x_t also intractable [91]. Hence, approximations must be used: one popular option is the Extended Kalman Filter (EKF), which uses a linearisation (i.e. first-order Taylor approximation) to generate mean and covariance updates. As with any linearisation, the EKF does not perform well in highly non-linear settings: this issue prompted the invention of the Unscented Kalman Filter (UKF), first proposed in [58]. First created for discrete-time systems, the UKF identifies changes to m_k and P_k (i.e. the mean and covariance of the discretised x_k, respectively) by tracking a set of *sigma points* in the state space over time (i.e. through repeated applications of f). The sigma points are chosen such that their mean and covariance equal m_k, P_k for all $k \geq 0$. For the sake of brevity, the details of this process (A.K.A. the *unscented transform*) are omitted here: we refer readers to [91] or [92], Section 8.8 for full descriptions of the unscented transform and the UKF. Note that the "Gaussian cubature" technique referred to by Lambert et al. in [65] (and utilised in Chapter 5) is a variant/special case of the unscented transform used in the UKF[8]. The Gaussian cubature method may also be considered as the multivariate extension of the "Gaussian quadrature" method for approximating univariate integrals using sigma points [67]: in academic literature, the terms are sometimes used interchangeably (e.g. in [65]).

As with the basic KF, it is possible to adapt the UKF to work directly on continuous-time systems such as (2.18): this is the principal contribution provided in [91], where the Unscented Kalman-Bucy Filter (UKBF) is defined. This algorithm makes use of the unscented transform to provide the following trajectories for m_t, P_t ([91], eq. (29)):

$$\begin{aligned} \frac{dm_t}{dt} &= f(X(t), t)w_m + K(t)\left(z(t) - h(X(t), t)w_m\right) \\ \frac{dP_t}{dt} &= X(t)Wf^T(X(t), t) + f(X(t), t)WX^T(t) + L(t)Q_c(t)L^T(t) - K(t)V(t)R_c(t)V^T(t)K^T(t) \\ K(t) &= X(t)Wh^T(X(t), t)\left(V(t)R_c(t)V^T(t)\right)^{-1}. \end{aligned} \tag{2.19}$$

Note that (2.19) has been written using the same notation found in [91] and introduces several new terms:

- $X(t)$: a matrix containing sigma points.
- w_m: a vector containing weights for the sigma points.
- W: a matrix containing weights for the sigma points.
- $Q_c(t)$: the diffusion matrix for β_t. Note that in [91], the ODEs in (2.19) are defined for general diffusion processes β_t and η_t. In the case that β_t is a standard Brownian motion, we have $Q_c(t) = I$.
- $R_c(t)$: the diffusion matrix for η_t. In the case that η_t is a standard Brownian motion, we have $R_C(t) = I$.

Without a thorough understanding of the unscented transform and the UKBF, it may not be immediately clear what the effects of $X(t), w_m$ and W are. Thankfully, a set of translations is available in [91], eq.

[8] This is explained in Section II.C of [44]: in particular, see Table I for a direct comparison of the parameters used by each model.

(58) and has been partially reproduced below in (2.20). Through these translations and their accompanying explanation in [91], we see that the usage of $X(t), w_m, W$ serves to generate approximate expectations under the marginal law p_t of x_t.

$$\begin{aligned}
\mathbb{E}\left(f(x_t,t)\right) &\approx f(X(t),t)w_m \\
\mathbb{E}\left(h(x_t,t)\right) &\approx h(X(t),t)w_m \\
\mathrm{Cov}\left(x_t, f(x_t,t)\right) &\approx X(t)Wf^T(X(t),t) \\
\mathrm{Cov}\left(x_t, h(x_t,t)\right) &\approx X(t)Wh^T(X(t),t)
\end{aligned} \tag{2.20}$$

We can thus rewrite (2.19) in terms of the approximate expectations given by (2.20):

$$\begin{aligned}
\frac{dm_t}{dt} &\approx \mathbb{E}\left(f(x_t,t)\right) + K(t)\left(z(t) - \mathbb{E}\left(h(x_t,t)\right)\right) \\
\frac{dP_t}{dt} &\approx \mathrm{Cov}\left(x_t, f(x_t,t)\right) + \mathrm{Cov}\left(f(x_t,t), x_t\right) + L(t)Q_c(t)L^T(t) - K(t)V(t)R_c(t)V^T(t)K^T(t) \\
K(t) &\approx \mathrm{Cov}\left(x_t, h(x_t,t)\right)\left(V(t)R_c(t)V^T(t)\right)^{-1}.
\end{aligned} \tag{2.21}$$

2.4.3. Connections to VI

As noted above, Kalman filters can be interpreted as sequential Bayesian algorithms: hence, there is extensive literature describing, investigating or exploiting this connection. However, there is relatively little literature specifically comparing VI and Kalman filters. The principal reason for this is that the fundamental VI optimisation problem (VI) is very broad, so it does not automatically correlate with a Kalman filter setting in the most general case. When VI is adapted to be performed sequentially, however, the connections become clearer. Indeed, for *online* Gaussian VI (as defined in Subsection 2.3.2), the similarities to Kalman filtering are obvious: both methods generate sequential Bayesian updates to a probability distribution (which we shall assume is stationary) based on a stream of incoming data. This resemblance was observed, for instance, in [60], where the authors propose (but do not explicitly show) that their algorithm for online Gaussian VI resembles the UKF. In [64], the main group behind [65] (i.e. Lambert, Bonnabel, Bach) show that their "R-VGA" algorithm (which is an "online" version of Gaussian VI) is equivalent to the EKF when used on regression problems, and equivalent to the original KFs when applied to (Bayesian) linear regression with Gaussian noise. Other recent publications into online Gaussian VI [33, 56] have also considered the interpretation of their models as Kalman filters, even using Kalman filters as the primary modelling paradigm in some cases [20, 21, 108]. VI has also been employed as a complexity-reduction technique to simplify the storage and computation of KF parameters in high dimensions [28], most notably the dispersion matrices.

One useful consequence of the connections between Kalman filtering and VI is the usage of the unscented transform/Gaussian cubature techniques to compute approximate expectations for VI and other sequential Bayesian models, as seen in e.g. [6, 52, 93]. This application is also exploited for the numerical computations performed in [65].

2.5. Calculus of Variations

This section has been included in order to guide readers unfamiliar with the calculus of variations. The principle tool from this branch of mathematics that is required here is the first variation of a functional; unfortunately, there are multiple distinct definitions of the terms "variation" and "first variation" used in the sources referenced in this work. To ease comprehension of the synthesis performed through this project, the two most relevant definitions of the "first variation" are presented together in this section, along with indications as to where each of them is used by authors whose work is used later on. Therefore, this section is by no means a detailed introduction to the calculus of variations, as descriptions of several key concepts (e.g. the Euler-Lagrange equation, higher-order variations, symmetries, usage in physics and machine learning) have been omitted for brevity: rather, Section 2.5 is intended to serve merely as a clarifying appendix.

2.5.1. Santambrogio definition (Functional derivative)

In [90] (Definition 7.12), Santambrogio offers the following implicit definition for the first variation of a functional $F : \mathcal{P}(\Omega) \mapsto \mathbb{R} \cup \{+\infty\}$. The first variation $\frac{\delta F}{\delta \varrho}(\varrho)$ is the measurable function which satisfies, for $\epsilon \in [0, 1]$ and an arbitrary test function $h \in \mathcal{P}(\Omega)$:

$$\frac{d}{d\epsilon} F(\varrho + \epsilon h)\Big|_{\epsilon=0} = \int_\Omega \frac{\delta F}{\delta \varrho}(\varrho)\, dh. \tag{2.22}$$

Alternatively, if we take F to be a function of probability densities instead of probability measures (which is more common in practice and is used for, e.g. descriptions of gradient flows), we can rewrite the integral above as:

$$\frac{d}{d\epsilon} F(\varrho + \epsilon h)\Big|_{\epsilon=0} = \int_\Omega \frac{\delta F}{\delta \varrho}(\varrho)\, h(x) dx. \tag{2.23}$$

This definition of the first variation can be found in literature related to optimal transport theory and its applications, e.g. [23, 24, 36, 110, 111] - with most of these sources citing [90] or the Villani book [104] as their source for this definition. This may imply that the "functional derivative" definition of the first variation has its roots in an influential text in optimal transport theory (e.g. Villani's book).

The object $\frac{\delta F}{\delta \varrho}$ is also known as the *functional derivative* in other texts. Informally, $\frac{\delta F}{\delta \varrho}(x)$ is analogous to a partial derivative, in that it measures the change in the value of F when the value of ϱ changes at the point $x \in \Omega$, specifically towards the value $h(x)$. Naturally, the definition above is only valid if a function $\frac{\delta F}{\delta \varrho}(\varrho)$ satisfying (2.22) actually exists. There are further (rather technical) requirements for the first variation to exist given in [90] (Def. 7.12), but they can be assumed to hold for the application of this "first variation" in Subsection 2.1.3 and so are omitted here for simplicity.

To provide examples of how definition (2.23) of the first variation can be applied in practice, let us consider the functionals $\mathcal{F}(\varrho) = \int_\Omega (\log \varrho(x))\, \varrho(x) dx$ and $\mathcal{V}(\varrho) = \int_\Omega V(x)\varrho(x) dx$ from (2.7). For $\mathcal{F}$, we have that

$$\begin{aligned}
\int_\Omega \frac{\delta \mathcal{F}}{\delta \varrho}(\varrho)\, dh &= \frac{d}{d\epsilon} \mathcal{F}(\varrho + \epsilon h)\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon} \int_\Omega \log\left(\varrho(x) + \epsilon h(x)\right) \left(\varrho(x) + \epsilon h(x)\right) dx \Big|_{\epsilon=0} \\
&= \int_\Omega \frac{d}{d\epsilon} \left(\log\left(\varrho(x) + \epsilon h(x)\right) \left(\varrho(x) + \epsilon h(x)\right)\right) dx \Big|_{\epsilon=0} \\
&= \int_\Omega \left(\frac{h(x)}{\varrho(x) + \epsilon h(x)} (\varrho(x) + \epsilon h(x)) + \log(\varrho(x) + \epsilon h(x)) h(x) \right) dx \Big|_{\epsilon=0} \\
&= \int_\Omega h(x)(1 + \log(\varrho(x) + \epsilon h(x)))dx \Big|_{\epsilon=0} \\
&= \int_\Omega h(x)(1 + \log \varrho(x))dx.
\end{aligned} \tag{2.24}$$

Following the implicit definition of $\frac{\delta \mathcal{F}}{\delta \varrho}$ expressed in (2.23), the last line above implies that $\frac{\delta \mathcal{F}}{\delta \varrho}(\varrho) = 1 + \log \varrho$. Applying the same method to $\mathcal{V}$ yields

$$\begin{aligned}
\frac{\delta \mathcal{V}}{\delta \varrho}(\varrho) &= \frac{d}{d\epsilon} \int_\Omega V(x)(\varrho(x) + \epsilon h(x))dx \Big|_{\epsilon=0} \\
&= \int_\Omega \frac{d}{d\epsilon} V(x)(\varrho(x) + \epsilon h(x))dx \Big|_{\epsilon=0} \\
&= \int_\Omega V(x)h(x)dx.
\end{aligned} \tag{2.25}$$

As before, the last line above implies that $\frac{\delta \mathcal{V}}{\delta \varrho}(\varrho) = V$.

2.5.2. Lambert et al. definition (Gateaux/Fréchet derivatives)

An alternative definition of the term "first variation", used in the key reference [65], is to use the LHS of (2.22) directly as the first variation $\delta_G F$ w.r.t. ϱ in the direction h:

$$\delta_G F(\varrho, h) := \frac{d}{d\epsilon} F(\varrho + \epsilon h)\Big|_{\epsilon=0} \tag{2.26}$$

The expression in (2.26) is also known as the *Gateaux derivative* of F. This derivative can also be expressed as:

$$\frac{d}{d\epsilon} F(\varrho + \epsilon h)\Big|_{\epsilon=0} = \lim_{\epsilon \searrow 0} \frac{F(\varrho + \epsilon h) - F(\varrho)}{\epsilon}. \tag{2.27}$$

It is also possible to define another derivative of F called the *Fréchet derivative*. Adapting the definition provided in [72]: let V, W be two normed vector spaces and let $F : U \to W$ for an open subset $U \subset V$. The Fréchet derivative, if it exists, is the bounded linear operator $\delta F(\varrho, \cdot) : V \to W$ satisfying, for all $h \in V$,

$$\lim_{||h||_V \to 0} \frac{||F(\varrho + h) - F(\varrho) - \delta F(\varrho, h)||_W}{||h||_V} = 0. \tag{2.28}$$

Alternatively, the condition in (2.28) can be converted into the following form, which shows that the Fréchet derivative explicitly provides a linear approximation of F at $\varrho + h$:

$$F(\varrho + h) - F(\varrho) - \delta F(\varrho, h) = o(||h||_V) \tag{2.29}$$

The condition that (2.28) must hold for *all* $h \in V$ is quite strong, so the existence of the Fréchet derivative is in many cases a non-trivial matter. One sufficient condition to establish the existence of the Fréchet derivative at $\varrho \in U$ is if F is analytic in a region containing a diagonal matrix containing the spectrum of ϱ ([72], Section 2.2). Equivalently, the Fréchet derivative exists if the functional F can be expressed as a power series (adapted from [59], Section 4.1). If the Fréchet derivative exists, it is equal to the Gateaux derivative: $\delta F(\varrho, h) = \delta_G F(\varrho, h)$. Furthermore, the existence of the Fréchet derivative implies that the Gateaux derivative can provide a linear approximation in the same manner as (2.29). The Fréchet and Gateaux derivatives possess many of the properties of the "conventional" derivative, such as the product rule (required for Section 3.2): $\delta(FG)(\varrho, h) = \delta F(\varrho, h)G(\varrho) + F(\varrho)\delta G(\varrho, h)$.

The Fréchet/Gateaux definition of the first variation[9] may be more common in mathematical literature, being found (for example) in [34, 54, 87]. A common source of confusion when encountering these derivatives in writing is that the Gateaux derivative is typically written with both arguments specified (e.g. $\delta F(\varrho, h)$), whereas this is occasionally omitted for the Fréchet derivative (e.g. $\delta F(\varrho)$ or even just δF). This discrepancy is likely rooted in the fact that the Fréchet derivative should exist for *any* test function h, even though the value of this derivative is, in general, not the same for different h.

2.5.3. Comparison and other definitions

The definitions of the term "first variation" provided in Subsection 2.5.1 and Subsection 2.5.2 are not equivalent, and will lead to different results when applied to the same functional F. Fundamentally, this is because the functional derivative (2.22) and the directional derivative (2.26) are different objects which are not interchangeable. To see this, consider the functional $\mathcal{F}(\varrho)$ from (2.7). In Subsection 2.5.1, we found that the first variation $\frac{\delta \mathcal{F}}{\delta \varrho}(\varrho) = 1 + \log \varrho$. When applying the definition (2.26), it is clear from (2.24) that we obtain $\delta F(\varrho, h) = \int_\Omega h(x)(1 + \log \varrho(x))dx$. In this case, definitions (2.22) and (2.26) are

[9] Unfortunately, even the term "Fréchet derivative" is not entirely free from ambiguity: in the textbook [96], the functional derivative $\frac{\delta F}{\delta \varrho}$ is referred to as the "*functional* (or *Fréchet*) derivative".

clearly not equal nor interchangeable: the former yields a function of a variable $x \in \Omega$, whereas the latter yields a functional of a functional h (in some set containing $\mathcal{P}(\Omega)$) which has been integrated over the space Ω, thus eliminating any dependency on the value of any point $x \in \Omega$.

Unfortunately, there are even more alternative definitions of the term ”first variation” to be found in other texts. For instance, some authors [41] define the ”first variation” as simply being a small increment in a certain direction, i.e. $\Delta F(\varrho, h) := F(\varrho + h) - F(\varrho)$, which is the linear component of the Fréchet approximation seen in (2.29).

In his work [80], which established the theoretical framework and justifications for Wasserstein gradient flows known collectively as ”Otto calculus”, Felix Otto also describes a ”first variation” of a curve $\rho^{(k)}(t) : [0,1] \to \mathcal{P}(\Omega)$ by crafting variations of that curve $\rho_\epsilon^{(k)}(t)$ involving a parameter ϵ and taking the derivative $\frac{d}{d\epsilon}\rho_\epsilon^{(k)}(t)\Big|_{\epsilon=0}$, which will find the ”optimal” curve between $\rho^{(k-1)}$ and $\rho^{(k)}$. This definition is also used by Otto et al. in the earlier paper establishing Wasserstein gradient flows themselves [57]. The ”first variation” seen here resembles (and is actually a special case of) the Gateaux definition (2.26), with the set of possible ”directions” for this derivative being restricted to a single curve indexed by time. This usage of the Gateaux derivative evokes the application of the calculus of variations most commonly employed in physics, which is to adhere to the *principle of stationary action* in Lagrangian mechanics by finding a trajectory which minimises the action integral (A.K.A. the *variational principle*). Unfortunately, this can be achieved by setting either the functional derivative (2.22) or the Gateaux derivative (2.26) to zero, which further contributes to the confusion in literature. Physicists may also define the ”first variation” through an integral resembling [14]:

$$\delta F(\varrho, h) = \int h \left(\frac{\partial F}{\partial \varrho} + \frac{\partial F}{\partial \varrho'} \right) dx \tag{2.30}$$

Although this may not be obvious at first glance, definition (2.30) is simply a special case of (2.26) when F is defined to be the integral of a Lagrangian over the state space. Textbook treatments of these forms of the ”first variation” can be found in [61, 14]: for accessible explanations of the physical interpretation of this variation, we refer readers to [45, 97].

An alternative approach to these misapprehensions is seen in the book [89] on calculus of variations, where Santambrogio appears to be avoiding the ”first variation” nomenclature issue entirely by avoiding use of this term altogether, preferring instead to focus directly on the optimisation problems typically handled through variational methods. Other authors appear to have adopted a similar approach in their textbooks, e.g. [100].

A full evaluation of the compatibility and strengths of this set of definitions would likely require a master’s thesis of its own, and so must be omitted here for brevity.

2.6. Riemannian Geometry

The space $\mathcal{P}_2(\mathbb{R}^d)$ may be interpreted as a Riemannian manifold, which permits concepts from differential calculus to be applied to Wasserstein gradient flows. In recognition of the seminal publication [80], which laid out the essential framework for the Riemannian nature of the 2-Wasserstein space, the study of such manifolds is denoted *Otto calculus*. Applications are seen, for instance, in Section 3.3, where we work with gradients over $\mathcal{P}_2(\mathbb{R}^d)$ and $\mathrm{BW}(\mathbb{R}^d)$ directly (as opposed to working in the product space $\mathbb{R}^d \times S_d^{++}$, which is isomorphic to $\mathrm{BW}(\mathbb{R}^d)$). For clarity and brevity, many ideas and results from Otto calculus (and more broadly Riemannian geometry) have been omitted, with only definitions essential to understanding Section 3.3 being included here. The material below has been adapted primarily from Chapter 1.3 of [37] and Section 2 of [77]; an accessible overview of the core results in Otto calculus may be found in Section 2 of [3].

Loosely speaking, a *manifold* M is a d-dimensional subspace of an ambient vector space (e.g. $\mathbb{R}^D$, where $d < D$) which is topologically similar to Euclidean space on a local level. More precisely, M is a manifold if each point $p \in M$ has a neighbourhood U_p homeomorphic to an open subset of the Euclidean space $\mathbb{R}^d$. These homeomorphisms φ_i and their domains U_i are called *charts* (U_i, φ_i): an

atlas $\{(U_i, \varphi i)\}$ is a set of charts which collectively cover the entirety of M. We may also define invertible *transition maps* $\tau_{a,b} : U_a \to U_b; \tau_{a,b}(r) := \varphi_b(\varphi_a^{-1}(r))$ between two charts (U_a, φ_a) and (U_b, φ_b), allowing them to be compared. If, for a given atlas, all transition maps are smooth (i.e. infinitely differentiable - denoted C^∞), then the atlas is considered smooth and M is a *smooth manifold*.

For a given point $p \in M$, let us consider all the curves in a smooth manifold M which pass through p. Without loss of generality, we can parametrise each of these curves as $\gamma : (-1, 1) \to M$ such that $\gamma(0) = p$. By considering the gradient (within the ambient space) of each curve γ as it passes through p, we can form the *tangent space*[10] T_pM of M at p, which is a d-dimensional vector space containing all possible directional derivatives which can be taken from the point p. More precisely:

$$T_pM := \{\dot{\gamma}(0) : \gamma : (-1, 1) \to M, \gamma(0) = p\} \tag{2.31}$$

It is possible to imbue manifolds with additional structure that ultimately permits derivatives to be taken. This is done by equipping M with a *metric* (A.K.A. *metric tensor*) g. In essence, g is a mapping $g : p \mapsto g_p$ which assigns to each point p an inner product $g_p : T_pM \times T_pM \to \mathbb{R}$, which allows us to define angles, lengths (via the induced norm $||\cdot||_{g_p}$) and distances (via the induced distance metric $d_{||\cdot||_{g_p}}(\cdot, \cdot)$) between elements of T_pM. We write $\langle \cdot, \cdot \rangle_p$ to denote the use of g_p, which in addition to the usual properties for inner products must also be positive-definite, i.e. $\langle r, r \rangle_p > 0, \forall r \neq 0$. For a given chart (U, φ) and a point $p \in U$, it is possible to obtain a coordinate basis $\{\partial_i\}_{i=1}^d$ for T_pM by mapping the canonical basis $\{e_i\}_{i=1}^d$ from $\mathbb{R}^d$ via φ^{-1}; for any g_p, we may define a (SPD) matrix G^p using $G^p_{i,j} = \langle \partial_i, \partial_j \rangle_p$ which captures the inner product g_p: $\langle r, q \rangle_p = r^T G^p q$. If the mappings $\mathcal{R}(p) = \langle \partial_i, \partial_j \rangle_p$ vary smoothly across M (for an appropriate choice of atlas and using transition maps where necessary), then g is a *Riemannian metric* and (M, g) is a *Riemannian manifold*.

Within Riemannian manifolds, many geometric concepts such as angles, volumes and distances may be defined, thus permitting ideas from calculus to be brought to the manifold setting. For this project, we are particularly interested in the definition of the *gradient* ∇F for a function $F : M \to \mathbb{R}$. For all $p \in M$, $\nabla F(p)$ is the unique element of T_pM such that, for an arbitrary curve $\gamma : (-1, 1) \to M, \gamma(0) = p$, the following holds:

$$\langle \nabla F(p), \dot{\gamma}(0) \rangle_p = \frac{d}{dt}\bigg|_{t=0} F(\gamma(t)) \tag{2.32}$$

Alternatively, we may consider a specific curve $\{p_t\}_{t \in \mathbb{R}} \subset M$ with a corresponding sequence of tangent vectors $\{v_t\}_{t \in \mathbb{R}}$, where $v_t \in T_{p_t}M$ for all t. Adhering to the notation used in [65], we may thus rewrite the definition (2.32) for ∇F as follows:

$$\langle \nabla F(p_t), v_t \rangle_p = \frac{d}{dt} F(p_t) \tag{2.33}$$

Informally speaking, the definition (2.33) of ∇F suggests the following interpretation. At the point $p_t \in M$, the object $\nabla F(p_t)$ is a vector indicating which direction the function F would grow the fastest, if p_t were to move in this direction at time t. However, p_t is actually moving in the direction v_t: the similarity between the directions $\nabla F(p_t)$ and v_t, captured by the inner product $\langle \nabla F(p_t), v_t \rangle_p$, determines the rate at which the value of F will change as we follow the true trajectory p_t over time.

The construction of T_pM raises the question of how to assign elements $q \in M$ near p to relevant vectors in T_pM and vice versa. These assignments are performed using the *logarithmic* and *exponential* maps, respectively. For a vector $v \in T_pM$, let us construct a geodesic $\gamma_v : [0, 1] \to M$ such that $\gamma_v(0) = p$, $\dot{\gamma}(0) = v$ and $|\dot{\gamma}_v(0)| = |v|$ (i.e. γ_v has constant speed). This geodesic is unique for any given v, and it has a unique endpoint $\gamma_v(1) = q$: we thus define the exponential map $\exp_p(v) = q$. Conversely: if we start from a point $q \in M$, there is a unique vector v such that a geodesic γ_v joins p and q in

[10] Note that $\dot{\gamma}(0)$ will indeed be tangential to M if M is smooth. If M contains a discontinuity, inflection point or singularity, then it is no longer smooth. If M includes linear segments (where tangents are not possible), then there will be discontinuities in the second derivatives of the transition maps and so M cannot be smooth. Therefore, the smoothness of M is sufficient to establish that $\dot{\gamma}(0)$ is tangent to M.

Property	Value chosen for $(\mathcal{P}_2(\mathbb{R}^d), W_2)$	Explanation
$g(p)$	$\int \langle \cdot, \cdot \rangle dp$	This choice of metric tensor is smooth for $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ and thus endows this space with Riemannian structure.
$T_p\mathcal{P}_2(\mathbb{R}^d)$	$\{\nabla\varphi \vert \varphi : \mathbb{R}^d \to \mathbb{R}\}$	Starting from p, moving in a given direction within $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is equivalent to applying some vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ to a set of particles distributed according to p. It may be shown that, in order to satisfy the continuity equation (2.5) for a curve $\{p_t\}_{t\geq 0} \in (\mathcal{P}_2(\mathbb{R}^d), W_2)$, we must have velocities (i.e. tangents) of the form described here.
$\exp_p(\nabla\varphi)$	$(\mathrm{id} + \nabla\varphi)_{\#}p$	The exponential map is defined so that the zero vector maps p to itself, i.e. $\exp_p(0) = (\mathrm{id})_{\#}p = p$.
$\log_p(q)$	$\nabla\varphi - \mathrm{id}$	The logarithmic map provides the optimal W_2 transport displacement between p and q, after subtracting for the identity displacement. Using (2.1), we may also write $\log_p(q)(x) = -T(x)$.

Table 2.1: a dictionary containing key features of the $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ manifold.

the manner described above. We may thus define the logarithmic map $\log_p(q) = v$. Note that the mappings provided by $\exp_p(\cdot)$ and $\log_p(\cdot)$ depend on the choice of metric tensor g, which determines "lengths" and "distances" in M and, consequently, the geodesics they are constructed with (and which in turn are used to construct $\exp_p(\cdot)$ and $\log_p(\cdot)$).

This text is concerned specifically with the $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ space and its interpretation as a Riemannian manifold. The basic characteristics of this manifold have been provided in Table 2.1, along with brief explanations. Further discussion of this manifold may be found in Subsection 3.3.1. However, a full justification of the Riemannian structure of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is omitted: we refer readers to Appendix B of [65] for suitable explanations.

3

Gaussian Variational Inference through Wasserstein Gradient Flows

In this chapter, we shall explore in detail the method for performing Gaussian variational inference established by Lambert et al. in their 2022 publication "Variational inference via Wasserstein gradient flows" [65]. This text thus comprises the primary source for the following chapter, with results from other sources being cited where necessary. In particular, we shall consider two different proofs of [**VI_WGF Thm 1**], both provided in [65]: for each of these, the proof has been reproduced in this chapter, with additional explanations and commentary added where appropriate. In Section 3.4, the connection between the Lambert et al. method and the Unscented Kalman Filter, proposed in [65], is investigated more thoroughly.

3.1. Summary

Let us first consider how the JKO scheme, as described in Section 2.2, might be used to describe Gaussian VI as described in Subsection 2.3.2. Let $\mathrm{BW}(\mathbb{R}^d)$ be the set of Gaussian densities on $\mathbb{R}^d$, equipped with the W_2 distance which, between Gaussians, has the closed form expression

$$W_2^2(p_1, p_2) = ||m_1 - m_2||^2 + \mathcal{B}^2(\Sigma_1, \Sigma_2) \tag{3.1}$$

where $\mathcal{B}^2(\Sigma_1, \Sigma_2)$ is the squared Bures distance between the covariance matrices

$$\mathcal{B}^2(\Sigma_1, \Sigma_2) = \mathrm{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right). \tag{3.2}$$

The restriction of the discretisation (2.11) to $\mathrm{BW}(\mathbb{R}^d)$ is referred to as the "Bures-JKO scheme" by Lambert et al., and specifies the following problem:

$$p_{t+h} \in \underset{p \in \mathrm{BW}(\mathbb{R}^d)}{\mathrm{argmin}} \left\{ \mathrm{KL}(p||\pi) + \frac{W_2^2(p, p_t)}{2h} \right\}. \tag{3.3}$$

Let us now consider the parametrisation of $p = N(m, \Sigma)$. By applying (3.1) and (3.2), we are able to obtain the following expression for what Lambert et al. call the "Bures-JKO scheme":

$$\underset{m \in \mathbb{R}^d, \Sigma \in S_d^{++}}{\mathrm{argmin}} \left\{ \mathrm{KL}(p||\pi) + \frac{1}{2h}||m - m_{k,h}||^2 + \mathcal{B}^2(\Sigma, \Sigma_{k,h}) \right\} \tag{3.4}$$

Let the posterior $\pi(x) \propto e^{-V(x)}$ for the potential V. Then taking the time-step limit of the scheme (3.3) leads to the following FPE describing the evolution of the solutions p_t:

$$\frac{d}{dt}p_t = \Delta p_t + \nabla \cdot (p_t \nabla V). \tag{3.5}$$

As with the general case in Section 2.2, (3.5) describes the marginal distribution of a particle x_t following a Langevin diffusion with stationary distribution p_π, the closest Gaussian approximation of π according to the left KL divergence:

$$dx_t = -\nabla V(x_t)dt + \sqrt{2}IdW_t, \tag{3.6}$$

The connections between π and (3.5), (3.6) provide opportunities to learn about the posterior. The diffusion (3.6) may be repurposed into a sampling mechanism which should empirically reconstruct p_π (q.v. "Langevin Monte Carlo"). Unfortunately, this method does not scale well to high dimensions (q.v. the "curse of dimensionality) and does not provide clear indications of convergence. A more direct approach would be to discretise (3.5) using Euler schemes, Runge-Kutta methods, etc. and thus track a sequence of Gaussian approximations to π. However, these techniques will become very expensive and inaccurate for the PDE (3.5), which must be propagated over an "infinite" (in practice: very large — the curse of dimensionality is also an issue here) number of dimensions to reasonably simulate p_t.

An alternative approach to learn about π using Gaussian approximations can be found in Särkkä's characterisation of the Unscented Kalman Filter [91]. For a Langevin particle $x_t \sim p_t$ (where p_t evolves according to the general JKO-FPE (2.9) in $\mathcal{P}_2(\mathbb{R}^d)$ and needn't be Gaussian), it may be shown ([65], Appendix B.4) that the mean and covariance of p_t behave as follows:

$$\begin{aligned} \dot{m}_t &= -\mathbb{E}\left(\nabla V(X_t)\right) \\ \dot{\Sigma}_t &= 2I - \mathbb{E}\left(\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)\right) \end{aligned} \tag{3.7}$$

By replacing $\{X_t\}_{t\geq 0}$ with a sequence of Gaussian random variables $\{Y_t : Y_t \sim p_t = N(m_t, \Sigma_t)\}_{t\geq 0}$, we can use the same formulas above to track m_t and Σ_t over time and thus obtain a Gaussian-restricted form of variational inference. Note that in general, the ODEs (3.7) and (3.8) are no longer describing the *same* mean and covariance evolutions - see Subsection 4.1.2 for more details.

$$\begin{aligned} \dot{m}_t &= -\mathbb{E}\left(\nabla V(Y_t)\right) \\ \dot{\Sigma}_t &= 2I - \mathbb{E}\left(\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)\right) \end{aligned} \tag{3.8}$$

Alongside Gaussian VI, Equations (3.8) also describe a special case of the UKBF: see Subsection 2.4.2 for a description of the UKBF, and Section 3.4 for a more thorough explanation of the interpretation of (3.8) as an UKBF. Although the specific characterisation of Gaussian VI as the UKBF ODEs (3.8) has not, to the author's knowledge, been postulated prior to [65], the connections between KF and VI have, in general, been thoroughly explored in previous research (see Subsection 2.4.3). In fact, the central contribution of Lambert et al. in [65] is to show that Gaussian VI may be interpreted as a Kalman filter because the latter emerges as a Wasserstein gradient flow applied to the former. More precisely, they show that up to first-order, taking the limit of the Bures-JKO discretisation (3.4) yields precisely the ODEs (3.8). This result is encapsulated in the following theorem, provided in [65]:

Theorem 3.1 ("VI-WGF" Theorem 1)**.** *Let $\{p_t\}_{t\geq 0} : p_t = N(m_t, \Sigma_t)$ be the limiting curve obtained via the Bures-JKO scheme* (3.3)*. Then, for a target posterior $\pi \propto \exp(-V)$, the Gaussian parameters m_t, Σ_t satisfy Särkkä's system of ODEs as given in* (3.8)*.*

Theorem 3.1 provides further justification for the connections between VI and KF seen in Subsection 2.4.3 by providing a new channel through which to express the former in terms of the latter. Furthermore, this channel allows for specific results about Wasserstein gradient flows and Kalman filters

to be applied to Gaussian VI, e.g. Corollary 3 in [65]. In [65], Lambert et al. provide three distinct proofs for Theorem 3.1: their presentation of these proofs omits various computations and technical details which have been verified in this chapter. Furthermore, the proofs themselves offer further insight into the system (3.8), the sequence of Gaussian approximations $\{p_t\}_{t\geq 0}$ and its relationship to the more general JKO scheme defined on $\mathcal{P}(\Omega)$.

3.2. Proof of Theorem 3.1 via Bures-JKO

Here, the proof of Theorem 3.1 provided by Lambert et al. using the Bures-JKO scheme is reproduced. Additional details of the proof not explicitly stated in [65] are provided. To assist the reader's comprehension of Lambert et al's original work, we shall for the most part adhere to their notation and follow the general structure of their proof.

3.2.1. Overview

The objective of Theorem 3.1 is to show that the curve $\{p_t\}_{t\geq 0}$ of Gaussian distributions which satisfies the JKO-FPE (2.9) also satisfies Särkkä's system of ODEs as given in (3.8). One way of doing so is to start from a special case of the JKO scheme discretisation (2.11) and take the step-size limit: this is precisely what Lambert et al. achieve in this proof. We can restrict the search space to that of Gaussian probability measures by introducing the closed-form Gaussian W_2 distance found in (3.1), which reduces the problem (2.11) to finding the mean and covariance m, Σ which solve the minimisation problem (3.4).

A solution $\hat{m}, \hat{\Sigma}$ for (3.4) is known to exist, as this problem is a special case of the optimisation problem (13) in [57] (which was shown to have a solution under conditions encompassing this present work). Let $L(m,\Sigma) := \mathrm{KL}(p||\pi) + \frac{1}{2h}||m - m_{k,h}||^2 + \mathcal{B}^2(\Sigma, \Sigma_{k,h})$. We can find m, Σ which minimise L by finding the stationary points, i.e. where $\nabla_m L = 0$ and $\nabla_\Sigma L = 0$ respectively. By considering solutions to $\operatorname{argmin} L(m_t, \Sigma_t)$ for unknown sequences $\{m_t\}_{t\geq 0}, \{\Sigma_t\}_{t\geq 0}$, which arise as the step size $h \searrow 0$, we can obtain expressions for $\dot{m}_t$ and $\dot{\Sigma}_t$.

3.2.2. Mean evolution

Let us consider a Gaussian density $p = N(m,\Sigma)$ and its negative entropy $H(p) = \mathbb{E}_p(\log p(x))$. From Identity B.1, we have $\nabla_m H(p) = 0$; from Identity B.2, we have $\nabla_m p(x) = -\nabla_x p(x)$. We can now take the gradient:

$$\begin{aligned}
\nabla_m \mathrm{KL}(p||\pi) &= \nabla_m \mathbb{E}_p\left(\log p(x) - \log \pi(x)\right) \\
&= 0 - \nabla_m \int_{\mathbb{R}^d} \log \pi(x) p(x) dx \\
&= \int_{\mathbb{R}^d} \log \pi(x) \left(-\nabla_m p(x)\right) dx \\
&= \int_{\mathbb{R}^d} \log \pi(x) \left(\nabla_x p(x)\right) dx.
\end{aligned} \tag{3.9}$$

At this point, Lambert et al. refer to their use of "integration by parts", which in this case is applied component-wise across the elements of the vector $\nabla p(x)$. By applying integration by parts for univariate functions, we have $\int_{\mathbb{R}} \log \pi(x) \frac{\partial}{\partial x_i} p(x) dx_i = \lim_{r\to\infty} [\log \pi(x) p(x)]_{x_i=-r}^{x_i=r} - \int_{\mathbb{R}} p(x) \frac{\partial}{\partial x_i} (\log \pi(x)) dx_i$. Since the Gaussian density $p(x) \to 0$ as $x_i \to \infty$ for any of the components x_i, and since $\log \pi(x) \propto -V(x)$ is assumed to be polynomial (or, at least, have bounded growth which is slower than that for the inverse exponential $p(x)$), we have that $\log \pi(x) p(x) \to 0$ as $x_i \to \infty$. Therefore, the boundary term vanishes and, by combining the component-wise results into a single vector, we obtain the desired result: $\int_{\mathbb{R}^d} \log \pi(x) \nabla_x p(x) dx = -\int_{\mathbb{R}^d} \nabla_x \log \pi(x) p(x) dx$. We therefore have that

$$\begin{aligned}
\nabla_m \mathrm{KL}(p||\pi) &= -\int_{\mathbb{R}^d} \nabla_x \log \pi(x) p(x) dx \\
&= -\mathbb{E}_p\left(\nabla_x \log \pi(x)\right)
\end{aligned} \tag{3.10}$$

and that

$$\begin{aligned}\nabla_m L(m,\Sigma) &= \nabla \mathrm{KL}(p||\pi) + \frac{1}{h}(m - m_t) \\ &= -\mathbb{E}_p\left(\nabla_x \log \pi(x)\right) + \frac{1}{h}(m - m_t).\end{aligned} \tag{3.11}$$

Since m represents a potential evolution step for m_t, we can rewrite $m = m_{t+h}$ and consider the limit $\lim_{h \searrow 0} \frac{m_{t+h} - m_t}{h} := \dot{m}_t$, which yields:

$$\begin{aligned}\dot{m}_t &= \mathbb{E}_p\left(\nabla_x \log \pi(x)\right) \\ &= -\mathbb{E}_p\left(\nabla_x V(x)\right)\end{aligned} \tag{3.12}$$

3.2.3. Covariance evolution

From Identity B.4, we have $\nabla_\Sigma H(p(x)) = -\frac{1}{2}\Sigma^{-1}$; from Identity B.5, we have $\nabla_\Sigma \mathbb{E}_p\left(\log \pi(x)\right) = \frac{1}{2}\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)$. We are thus able to write:

$$\begin{aligned}\nabla_\Sigma \mathrm{KL}(p||\pi) &= \nabla_\Sigma (H(p) - \mathbb{E}_p\left(\log \pi(x)\right) \\ &= -\frac{1}{2}\Sigma^{-1} - \frac{1}{2}\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\end{aligned} \tag{3.13}$$

Let us consider T^{Σ,Σ_t}, which is the optimal transport map between two Gaussian densities $N(0,\Sigma)$ and $N(0,\Sigma_t)$. It is known [7] that T^{Σ,Σ_t} has the form

$$T^{\Sigma,\Sigma_t} = \Sigma^{-\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}\Sigma_t\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} \tag{3.14}$$

and, by extension, that $T^{\Sigma,\Sigma_t} = (T^{\Sigma_t,\Sigma})^{-1}$. From [7], we know that the covariance gradient of the squared Bures distance is:

$$\nabla_\Sigma \mathcal{B}^2(\Sigma_t,\Sigma) = I - T^{\Sigma,\Sigma_t} \tag{3.15}$$

Recalling the definition $L(m,\Sigma) = \mathrm{KL}(p||\pi) + \frac{1}{2h}||m - m_{k,h}||^2 + \mathcal{B}^2(\Sigma,\Sigma_{k,h})$, we obtain:

$$\nabla_\Sigma L(m,\Sigma) = \frac{1}{2h}\left(I - T^{\Sigma,\Sigma_t}\right) - \frac{1}{2}\left(\Sigma^{-1} + \mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\right) \tag{3.16}$$

By setting $\nabla_\Sigma L(m,\Sigma) = 0$, we aim to find a solution for (3.4). This yields:

$$I = T^{\Sigma,\Sigma_t} + h\Sigma^{-1} + h\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right) \tag{3.17}$$

If we multiply the above expression by Σ separately on the left and the right, we obtain the following expressions for Σ:

$$\begin{aligned}\Sigma &= \Sigma T^{\Sigma,\Sigma_t} + hI + h\Sigma\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right) \\ \Sigma &= T^{\Sigma,\Sigma_t}\Sigma + hI + h\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\Sigma\end{aligned} \tag{3.18}$$

By summing these together, we obtain a symmetric[1] expression for Σ:

$$\Sigma = \frac{1}{2}T^{\Sigma,\Sigma_t}\Sigma + \frac{1}{2}\Sigma T^{\Sigma,\Sigma_t} + hI + \frac{1}{2}h\Sigma\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right) + \frac{1}{2}h\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\Sigma \tag{3.19}$$

For a fixed Σ_t, we can define the shorthand notation $T(\Sigma) := T^{\Sigma,\Sigma_t}$. As noted in Appendix A of [2]: for a vector $X \sim N(0,\Sigma)$, the Gaussian property $AX \sim N(0, A\Sigma A^T)$ implies that $T(\Sigma)X \sim N(0, T(\Sigma)\Sigma T(\Sigma)$, which must be equal to $N(0,\Sigma_t)$ by the definition of T as a transport map. In fact, from the expression for $T(\Sigma)$ provided in (3.14), we can directly verify that $T(\Sigma)\Sigma T(\Sigma) = \Sigma_t$:

$$\begin{aligned} T(\Sigma)\Sigma T(\Sigma) &= \Sigma^{-\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}\Sigma_t\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}\Sigma_t\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\Sigma_t\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} \\ &= \Sigma_t \end{aligned} \tag{3.20}$$

At this point, the remaining objective is to obtain a linear approximation (about t, in terms of h) of Σ_{t+h} using (3.19), which can then be passed into a limit to obtain an expression for $\dot{\Sigma}_t$. Doing so requires generic first-order approximations for Σ_{t+h} and $T(\Sigma_{t+h})$. It is tempting to simply define functions of time $f(s) := \Sigma_s, g(s) := T(\Sigma_s)$ and use ordinary calculus to obtain such approximations: indeed, inserting these approximations into (3.19) will yield the desired ODE for $\dot{\Sigma}_t$. However, this approach assumes that $f(s), g(s)$ are differentiable at $s = t$, which is not known *a priori*. An alternative approach would be to work with a broader mechanism for linear approximation which, when valid, works for all directions and thus encompasses the desired case $\Sigma_t \to \Sigma_{t+h}$. This notion motivates the use of variational methods[2], more precisely the Fréchet derivative (2.28), which, when it exists, exists in all directions accessible from the chosen reference point. The corresponding Fréchet linear approximations also exist in all possible directions, including the desired direction $\dot{\Sigma}_t$ (which, by extension, proves the differentiability of $f(s), g(s)$ and the existence of $\dot{\Sigma}_t$).

Ideally, we would simply find an expression for $\delta T(\Sigma_t, \dot{\Sigma}_t)$ and proceed with the linearisation of (3.19): unfortunately, this task is not tractable when working with the definitions of Fréchet or Gateaux derivatives. To handle any terms involving $\delta T(\Sigma_t, \dot{\Sigma}_t)$ later on, we will need at least an expression relating this Fréchet derivative to "known" terms. We can achieve this through the identity $T(\Sigma)\Sigma T(\Sigma) = \Sigma_t$ demonstrated above: since the Fréchet derivative of the right-hand-side is zero, we may be able to simplify expressions involving $\delta T(\Sigma_t, \dot{\Sigma}_t)$ in the linear approximation of (3.19) later on. The task then becomes to determine Fréchet differentiabiity of $F(\Sigma) := T(\Sigma)\Sigma T(\Sigma)$. For the functional[3] $G(\Sigma) := \Sigma, \Sigma \in S_d^{++}$, the Fréchet derivative exists for all $A \in S_d^{++}$ as G clearly has a power series representation (itself). Furthermore, the definition (2.28) immediately yields $\delta G(\Sigma, H) = H$. For $T(\Sigma)$, Fréchet differentiability is not immediately clear, and working with definition (2.28) is unfortunately not tractable. However, we do know the following:

- The Fréchet derivative obeys the product rule. (As noted in [48], Section 2)
- The Fréchet derivative obeys the chain rule, s.t. the composition of Fréchet-differentiable functionals is itself Fréchet-differentiable. (As noted in [48], Section 2)
- The matrix inverse function is Fréchet differentiable. (Follows from [48], Thm. 3.2 and the fact that the real scalar function $\}(a) = a^{-1}$ is analytic on the set $a > 0$)

[1] Enforcing symmetry in this way ensures that the ensuring linearisations yield symmetric (and thus appropriate) expressions for Σ. This turns out to be a necessary step, as evidenced by the usage of both results from Identity B.6 to obtain the symmetric ODE for Σ in (3.8). Alternatively, if we proceeded solely with, e.g. the first expression in (3.18), we would obtain an expression of the form $\dot{\Sigma}_t = ... + c\mathbb{E}_p\left(\nabla_x \log \pi(x) \otimes (x - m)\right)$, which is clearly asymmetric in general.

[2] Lambert et al. appear to acknowledge this reasoning in their work through their mentions of a "first variation", which, given the context, should refer to the Fréchet (or possibly the Gateaux) derivative. Unfortunately, this is not clear from what is presented in [65]; to make matters worse, the *other* definition of "first variation" (i.e. the definition (2.22) from Subsection 2.5.1) is used in Appendix F of [65].

[3] The use of the term "functional" here is somewhat arbitrary and may be confusing to readers expecting a mapping of an infinite-dimensional operator instead of a finite-dimensional matrix. However, this ambiguity has no impact on the existence or applicability of the Fréchet derivative to the functions (or functionals) G, T, F, as shown in this subsection. It may be helpful to interpret a matrix $\Sigma \in \mathbb{R}^{d\times d}$ as being an operator (or a function) acting on vectors $x \in \mathbb{R}^d$.

- The matrix square-root function is Fréchet differentiable ([29], Thm. 1.1).

By composing the facts listed above, we obtain that both $T(\Sigma)$ and $F(\Sigma)$ are Fréchet differentiable for all $\Sigma \in S_d^{++}$. From above, we know that $\delta F(\Sigma, H) = 0$ always, since F is a constant functional for which, from (2.28), it is clear that the Fréchet derivative is zero for all Σ and directions H. Meanwhile, we can apply the chain rule to $F(\Sigma) := T(\Sigma)\Sigma T(\Sigma)$ to obtain:

$$\delta F(\Sigma, H) = \delta T(\Sigma, H)\Sigma T(\Sigma) + T(\Sigma)\delta G(\Sigma, H)T(\Sigma + T(\Sigma\Sigma\delta T(\Sigma, H). \tag{3.21}$$

By setting $\Sigma \leftarrow \Sigma_t$ and $H \leftarrow \dot{\Sigma}_t$ we arrive at

$$\delta T(\Sigma_t, \dot{\Sigma}_t)\Sigma_t + \dot{\Sigma}_t + \Sigma_t \delta T(\Sigma_t, \dot{\Sigma}_t) = 0 \tag{3.22}$$

Equation (3.22) is a Sylvester equation which offers an implicit definition of the value of $\delta T(\Sigma, H)$. However, an explicit solution for $\delta T(\Sigma, H)$ is not needed to proceed with the linearisation of (3.19). With the existence of suitable linearisations now firmly established, we can now set $\Sigma \leftarrow \Sigma_{t+h}$ in (3.19) and use $\Sigma_{t+h} \approx \Sigma_t + h\dot{\Sigma}_t$ and $T(\Sigma_{t+h}) \approx T(\Sigma_t) + \delta T(\Sigma_t, \dot{\Sigma}_t)$ to obtain the first-order (in h) approximation

$$\begin{aligned}
\Sigma_{t+h} &\approx \Sigma_t + h\dot{\Sigma}_t \\
&\approx hI + \frac{1}{2}\left(T(\Sigma_{t+h})\Sigma_{t+h} + \Sigma_{t+h}T(\Sigma_{t+h}) + h\Sigma_{t+h}E_{p_t} + hE_{p_t}\Sigma_{t+h}\right) \\
&= hI + \frac{1}{2}\Big((I + \delta T(\Sigma_t, \dot{\Sigma}_t))(\Sigma_t + h\dot{\Sigma}_t) + (\Sigma_t + h\dot{\Sigma}_t)(I + \delta T(\Sigma_t, \dot{\Sigma}_t)) \\
&\qquad + h(\Sigma_t + h\dot{\Sigma}_t)E_{p_t} + hE_{p_t}(\Sigma_t + h\dot{\Sigma}_t)\Big) \\
&= hI + \frac{1}{2}\Big(2\Sigma_t + h\dot{\Sigma}_t + h\dot{\Sigma}_t + \delta T(\Sigma_t, \dot{\Sigma}_t)\Sigma_t + \Sigma_t\delta T(\Sigma_t, \dot{\Sigma}_t) + hdT\dot{\Sigma}_t + h\dot{\Sigma}_t\delta T(\Sigma_t, \dot{\Sigma}_t) \\
&\qquad + h\left(\Sigma_t E_{p_t} + E_{p_t}\Sigma_t\right) + h^2\left(\dot{\Sigma}_t E_{p_t} + E_{p_t}\dot{\Sigma}_t\right)\Big) \\
&= hI + \Sigma_t + \frac{1}{2}h\dot{\Sigma}_t + \frac{1}{2}\Big(d\Sigma + \delta T(\Sigma_t, \dot{\Sigma}_t)\Sigma_t + \Sigma_t\delta T(\Sigma_t, \dot{\Sigma}_t) + h\delta T(\Sigma_t, \dot{\Sigma}_t)\dot{\Sigma}_t \\
&\qquad + h\dot{\Sigma}_t\delta T(\Sigma_t, \dot{\Sigma}_t) + h\left(\Sigma_t E_{p_t} + E_{p_t}\Sigma_t\right) + h^2\left(\dot{\Sigma}_t E_{p_t} + E_{p_t}\dot{\Sigma}_t\right)\Big)
\end{aligned} \tag{3.23}$$

In (3.23), we see the emergence of the identity (3.22). We may rewrite the term $h\delta T(\Sigma_t, h\dot{\Sigma}_t)\dot{\Sigma}_t \approx hT(\Sigma_{t+h})\dot{\Sigma}_t - h\dot{\Sigma}_t$: if we divide by h then take the limit as $h \searrow 0$, we obtain $\dot{\Sigma}_t - \dot{\Sigma}_t = 0$. A symmetric argument leads to the same outcome for the term $h\dot{\Sigma}_t\delta T(\Sigma_t, h\dot{\Sigma}_t)$. By applying these same operations (dividing by h and setting $h \searrow 0$) to the entire last line of (3.23) and rearranging terms, we obtain an expression for $\dot{\Sigma}_t$ using the results from Identity B.6:

$$\begin{aligned}
\dot{\Sigma}_t &= 2I + \Sigma_t E_{p_t} + E_{p_t}\Sigma_t \\
&= 2I + \mathbb{E}_{p_t}\left(\nabla_x \log \pi(x) \otimes (x - m)\right) + \mathbb{E}_{p_t}\left((x - m) \otimes \nabla_x \log \pi(x)\right).
\end{aligned} \tag{3.24}$$

We thus obtain the specified ODE for the Gaussian covariance matrix Σ_t.

3.3. Proof of Theorem 3.1 via orthogonal projection

An alternative proof of Theorem 3.1 to that provided above can be constructed using Otto calculus. More precisely: by treating $\mathrm{BW}(\mathbb{R}^d)$ as a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$, we can find the orthogonal projection of the gradient of the functional $\mathrm{KL}(\cdot||\pi)$ from $\mathcal{P}_2(\mathbb{R}^d)$ onto $\mathrm{BW}(\mathbb{R}^d)$. This is the approach used in Appendix C.1 of [65], which has been reproduced here with additional clarifying details. Furthermore, the differential machinery enlisted for this proof offers further insight into the relationship between $\mathrm{BW}(\mathbb{R}^d)$ and $\mathcal{P}_2(\mathbb{R}^d)$, which is discussed below.

3.3.1. Preamble

Before proceeding, we must be sure that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ does indeed admit a Riemannian structure such that differential concepts such as tangents and gradients may be defined. This claim was famously first made by Otto in [80], with subsequent work such as [4] providing further justification and more recent publications offering concise summaries to readers seeking to cover this topic in more detail than is admissible here [32]. In order for orthogonal projections to be possible, we must also ensure that $T_p\mathrm{BW}(\mathbb{R}^d) \subset T_p\mathcal{P}_2(\mathbb{R}^d)$ for some point $p \in \mathrm{BW}(\mathbb{R}^d)$. As observed in Appendix B.3 of [65], this will be the case provided we choose to equip $\mathrm{BW}(\mathbb{R}^d)$ with the same Riemannian structure as $\mathcal{P}_2(\mathbb{R}^d)$. The choice of metric tensor is not a trivial consideration: alternative metric tensors have also been explored for $\mathrm{BW}(\mathbb{R}^d)$ (or, at least, for the space S_d^{++}), e.g. in [46, 102]. The manner in which points $q \in \mathrm{BW}(\mathbb{R}^d)$ near p are associated with vectors $v \in T_p\mathrm{BW}(\mathbb{R}^d)$ (i.e. the choices of exponential and logarithmic maps) also impacts the structure of the space being considered. This issue was noted in [25] (Appendix A.2), where the proposed solution is to conceptually adhere to optimal transport principles by defining $\exp_p(q) = \nabla\varphi_{p\to q} - \mathrm{id}$, for the optimal transport map $\nabla\varphi_{p\to q}$ between p and q. This convention ensures that geodesics $\gamma_{p\to q} : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ in $\mathcal{P}_2(\mathbb{R}^d)$ (and, by extension, $\mathrm{BW}(\mathbb{R}^d)$) conform to optimal transport mappings[4].

It is also important to be familiar with how scholars often work with $\mathrm{BW}(\mathbb{R}^d)$ in practice. Since any Gaussian density on $\mathbb{R}^d$ may be uniquely identified by its mean and covariance, we may construct a homeomorphism between $\mathrm{BW}(\mathbb{R}^d)$ and the finite-dimensional vector space $\mathbb{R}^d \times S_d^{++}$. Furthermore, the tangent space $T_p\mathrm{BW}(\mathbb{R}^d)$ is itself homeomorphic to $\mathbb{R}^d \times S^d$ [5]. Since many results are obtained by working inside these finite-dimensional vector spaces rather than the infinite-dimensional function spaces, many authors (including Lambert et al. in [65]) use the expressions $\mathrm{BW}(\mathbb{R}^d)$ and $T_p\mathrm{BW}(\mathbb{R}^d)$ both for the function spaces and for their associated vector spaces. Where reasonable, we continue to use this convention here.

The proof provided in this section is valid because of the following observation for the functional $F(p) := \mathrm{KL}(p||\pi)$, made by Lambert et al. in [65]: when the gradient $\nabla_{W_2}F(p)$ in $T_p\mathcal{P}_2(\mathbb{R}^d)$ is projected onto the tangent space $T_p\mathrm{BW}(\mathbb{R}^d)$, we obtain the same result as if we compute $\nabla_{\mathrm{BW}}F(p)$ within $T_p\mathrm{BW}(\mathbb{R}^d)$ directly. We can see this by considering the implicit definition of $\nabla_{\mathrm{BW}}F$: for a curve $\{p_t\}_{t\in\mathbb{R}} \subset \mathrm{BW}(\mathbb{R}^d)$ and its velocity vectors $\{v_t\}_{t\in\mathbb{R}} : v_t \in T_{p_t}\mathrm{BW}(\mathbb{R}^d)$, we have

$$\langle \nabla_{\mathrm{BW}}F(p_t), v_t\rangle_{p_t} = \frac{d}{dt}F(p_t) \tag{3.26}$$

Meanwhile, for $\nabla_{W_2}F(p)$, we can use the same curve $\{p_t\}_{t\in\mathbb{R}}$ (which also lies in $\mathcal{P}_2(\mathbb{R}^d)$) to obtain

$$\langle \nabla_{W_2}F(p_t), v_t\rangle_{p_t} = \frac{d}{dt}F(p_t) \tag{3.27}$$

We may now define the orthogonal projection $\mathrm{proj}_{\mathrm{BW}}\nabla_{W_2}F(p_t)$:

$$\mathrm{proj}_{\mathrm{BW}}\nabla_{W_2}F(p_t) := \underset{w\in T_{p_t}\mathrm{BW}(\mathbb{R}^d)}{\mathrm{argmin}} ||w - \nabla_{W_2}F(p_t)||^2 \tag{3.28}$$

[4] The same convention is also assumed by Altschuler et al. in [2], where in Appendix A.1 an illuminating example of the importance of vector mapping choice is made. When the Riemannian structure induced by the W_2 distance over $\mathcal{P}_2(\mathbb{R}^d)$ is applied to the subspace of zero-mean Gaussian distributions, the constant-speed W_2 geodesic $\{\Sigma_t\}_{t\in[0,1]}$ between two densities $N(0,\Sigma_0)$ and $N(0,\Sigma_1)$ has the form

$$\Sigma_t = ((1-t)I + tT(\Sigma))\,\Sigma_0\,((1-t)I + tT(\Sigma)) \tag{3.25}$$

From Section 2.6, we know that $\log_{\Sigma_0}(\Sigma_1) = \dot{\Sigma}_0$ by definition; in [2], it is shown that $\log_{\Sigma_0}(\Sigma_1) = T(\Sigma) - I$. However, if we directly take the time derivative of (3.25), we obtain $\dot{\Sigma}_0 = (T(\Sigma) - I)\Sigma + \Sigma(T(\Sigma) - I)$. Hence, following the usual approach to computing geodesic velocities clearly yields different outcomes to the optimal-transport convention described above.

[5] We know from Subsection 3.2.3 that $T(\Sigma)$ optimally maps $x \sim N(0,\Sigma)$ to $x' \sim N(0,\Sigma_t)$ under the W_2 distance. We may extend this procedure to incorporate Gaussians with non-zero mean, such that we have the optimal transport mapping $x \mapsto m_t + T(\Sigma)(x - m)$ between $N(m,\Sigma)$ and $N(m_t,\Sigma_t)$ This mapping is an affine transformation: since, by the convention adopted in the previous paragraph, $T_p\mathrm{BW}(\mathbb{R}^d)$ contains the optimal transport mappings between p and other Gaussians in $\mathrm{BW}(\mathbb{R}^d)$, we therefore have that $T_p\mathrm{BW}(\mathbb{R}^d)$ may be parametrised by $\mathbb{R}^d \times S^d$.

Since $T_{p_t}\mathcal{P}_2(\mathbb{R}^d)$ is a Hilbert space, inside which $T_{p_t}\mathrm{BW}(\mathbb{R}^d)$ is a closed subspace[6], we may apply the following basic property of orthogonal projection in Hilbert spaces, which holds true for any $v \in T_{p_t}\mathrm{BW}(\mathbb{R}^d)$:

$$\langle \mathrm{proj}_{\mathrm{BW}} \nabla_{W_2} F(p_t), v \rangle_{p_t} = \langle \nabla_{W_2} F(p_t), v \rangle_{p_t} \tag{3.29}$$

From (3.27), we then know that for the vector $v_t = \dot{p}_t$:

$$\langle \mathrm{proj}_{\mathrm{BW}} \nabla_{W_2} F(p_t), v_t \rangle_{p_t} = \frac{d}{dt} F(p_t) \tag{3.30}$$

Since $\mathrm{proj}_{\mathrm{BW}} \nabla_{W_2} F(p_t) \in T_{p_t}\mathrm{BW}(\mathbb{R}^d)$, this implies that $\mathrm{proj}_{\mathrm{BW}} \nabla_{W_2} F(p_t)$ satisfies (3.26), i.e.

$$\nabla_{\mathrm{BW}} F(p_t) = \mathrm{proj}_{\mathrm{BW}} \nabla_{W_2} F(p_t) \tag{3.31}$$

3.3.2. Core proof

In Appendix B.1 of [65], the inner product for $T_p\mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$\langle v, w \rangle_p := \int \langle v, w \rangle dp \tag{3.32}$$

The expression $\langle v, w \rangle$ inside the integral in (3.32) has varying meanings depending on the nature of the arguments v, w. If v, w are functions from $T_p\mathcal{P}_2(\mathbb{R}^d)$ proper, then $\langle v, w \rangle$ refers to the standard L^2 inner product. However, since we are able to work with the space $\mathbb{R}^d \times S^d$, Lambert et al. may also write $\langle v, w \rangle$ when v, w are finite-dimensional vectors in $\mathbb{R}^d$: in this case, $\langle v, w \rangle$ should be interpreted as the ordinary dot product between vectors. Additionally, if $\langle v, w \rangle$ is used when v, w are matrices in $\mathbb{R}^{d\times d}$, then $\langle v, w \rangle$ is the Frobenius inner product between these matrices. In this text, we aim to improve clarity by writing $\langle v, w \rangle_{\mathbf{V}}$ and $\langle v, w \rangle_{\mathbf{F}}$ for the vector and Frobenius inner products, respectively.

If we work with the parametric space $\mathbb{R}^d \times S^d$ instead of $T_p\mathrm{BW}(\mathbb{R}^d)$, we may consider the inner product (which becomes the dot product) between two velocities $(\bar{a}, \bar{S}), (a, S) \in \mathbb{R}^d \times S^d$:

$$\begin{aligned}
\langle (\bar{a}, \bar{S}), (a, S) \rangle_p &= \int \langle \bar{a} + \bar{S}(x - m_p), a + S(x - m_p) \rangle_{\mathbf{V}} dp(x) \\
&= \mathbb{E}_p \left(\sum_{i=1}^{d} \left(\bar{a}_i a_i + \bar{a}_i S(x - m_p)_i + \bar{S}(x - m_p)_i a_i + \bar{S}(x - m_p)_i S(x - m_p)_i \right) \right) \\
&= \sum_{i=1}^{d} \mathbb{E}_p \left(\bar{a}_i a_i + \bar{a}_i S(x - m_p)_i + \bar{S}(x - m_p)_i a_i + \bar{S}(x - m_p)_i S(x - m_p)_i \right) \\
&= \sum_{i=1}^{d} (\bar{a}_i a_i + 0 + 0) + \mathbb{E}_p \left((\bar{S}(x - m_p))^T S(x - m_p) \right) \\
&= \langle \bar{a}, a \rangle_{\mathbf{V}} + \mathbb{E}_p \left((\bar{S}(x - m_p))^T S(x - m_p) \right)
\end{aligned} \tag{3.33}$$

The term $\mathbb{E}_p \left((\bar{S}(x - m_p))^T S(x - m_p) \right)$ may be further simplified:

[6] We have a homeomorphism between $T_{p_t}\mathrm{BW}(\mathbb{R}^d)$ and the finite-dimensional vector space $\mathbb{R}^d \times S_d$. A well-known result from functional analysis states that any finite-dimensional subspace of a Banach space is closed; the homeomorphism permits us to extend this property to $T_{p_t}\mathrm{BW}(\mathbb{R}^d)$.

$$\begin{aligned}\mathbb{E}_p\left((\bar{S}(x-m_p))^T S(x-m_p)\right) &= \mathbb{E}_p\left((x-m_p)^T\bar{S}^T S(x-m_p)\right)\\ &= \mathbb{E}_p\mathrm{tr}\left((x-m_p)^T\bar{S}S(x-m_p)\right)\\ &= \mathbb{E}_p\mathrm{tr}\left(S(x-m_p)\otimes(x-m_p)^T\bar{S}\right)\\ &= \mathrm{tr}\left(S\mathbb{E}_p\left((x-m_p)\otimes(x-m_p)^T\right)\bar{S}\right)\\ &= \mathrm{tr}\left(\bar{S}\Sigma_p S\right)\\ &= \langle\bar{S},\Sigma_p S\rangle_{\mathbf{F}}\end{aligned} \tag{3.34}$$

We thus have:

$$\langle(\bar{a},\bar{S}),(a,S)\rangle_p = \langle\bar{a},a\rangle_{\mathbf{V}} + \langle\bar{S},\Sigma_p S\rangle_{\mathbf{F}} \tag{3.35}$$

Following the result (3.31) obtained in the prefious subsection, the remaining task for proving Theorem 3.1 via orthogonal projection is to determine the value of $\mathrm{proj}_{\mathrm{BW}}\nabla_{W_2}F(p_t)$. To do so, we make use of (3.35) as well as the identity (asserted in [65] to comprise a special case of Thm. 10.4.13 in [4]):

$$\nabla_{W_2}\mathrm{KL}(p||\pi) = \nabla\log\frac{p}{\pi} \tag{3.36}$$

Naturally, (3.36) also applies to the special case $\nabla_{\mathrm{BW}(\mathbb{R}^d)}\mathrm{KL}(p||\pi)$, since $\mathrm{BW}(\mathbb{R}^d)\subset\mathcal{P}_2(\mathbb{R}^d)$. We must be able to express $\nabla\log\frac{p}{\pi}$ with a pair $(\bar{a},\bar{S})\in\mathbb{R}^d\times S^d$, as $\nabla\log\frac{p}{\pi}$ is a vector in $T_p\mathrm{BW}(\mathbb{R}^d)$.

We are now ready to begin computing

$$\int\langle\nabla\log\frac{p}{\pi}(x),a+S(x-m_p)\rangle_{\mathbf{V}}dp(x) = \mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),a+S(x-m_p)\rangle_{\mathbf{V}} \tag{3.37}$$

By the bilinearity of the vector dot product, we are able to write:

$$\begin{aligned}\mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),a+S(x-m_p)\rangle_{\mathbf{V}} &= \mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),a\rangle_{\mathbf{V}} + \mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),S(x-m_p)\rangle_{\mathbf{V}}\\ &= \langle\mathbb{E}_p\nabla\log\frac{p}{\pi}(x),a\rangle_{\mathbf{V}} + \mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),S(x-m_p)\rangle_{\mathbf{V}}\end{aligned} \tag{3.38}$$

Once more by the bilinearity of the dot product[7], the second term above can be rewritten as

$$\begin{aligned}\mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),S(x-m_p)\rangle_{\mathbf{V}} &= \mathbb{E}_p\langle S\nabla\log\frac{p}{\pi}(x),x-m_p\rangle_{\mathbf{V}}\\ &= \mathbb{E}_p\langle\Sigma_p S\nabla\log\frac{p}{\pi}(x),\Sigma_p^{-1}(x-m_p)\rangle_{\mathbf{V}}\end{aligned} \tag{3.39}$$

From the computations in (B.16), we know that $\nabla_x p(x) = -\Sigma_p^{-1}(x-m)p(x)$. We may therefore write:

$$\begin{aligned}\mathbb{E}_p\langle\nabla\log\frac{p}{\pi}(x),S(x-m_p)\rangle_{\mathbf{V}} &= \mathbb{E}_p\langle\Sigma_p S\nabla\log\frac{p}{\pi}(x),-\frac{\nabla_x p(x)}{p(x)}\rangle_{\mathbf{V}}\\ &= -\int\left(\sum_{i=1}^{d}\left(\Sigma_p S\nabla\log\frac{p}{\pi}(x)\right)_i\left(\nabla_x p(x)\right)_i\right)\frac{1}{p(x)}p(x)dx\end{aligned} \tag{3.40}$$

[7] In this case, we extend this bilinearity to generate the identity $\langle v,Aw\rangle_{\mathbf{V}} = \langle Av,w\rangle_{\mathbf{V}}$ for a symmetric matrix A, which can be proven by rewriting the equivalent expression $\langle v,Aw\rangle_{\mathbf{V}} = v^TAw$

Here, Lambert et al. refer once more to their usage of "integration by parts", which involves a very similar component-wise operation to that used in Subsection 3.2.2. Applying univariate integration by parts to each of the components $i = 1, ..., d$ yields terms of the form $\lim_{r\to\infty}[\left(\Sigma_p S\nabla \log \frac{p}{\pi}(x)\right)_i p(x)]_{-r}^{r} - \int \frac{d}{dx_i}\left(\Sigma_p S\nabla \log \frac{p}{\pi}(x)\right)_i p(x)dx_i$, where ∇^{*i} is used to denote the gradient operation with the *second* derivative taken at the i^{th} position instead of the first. Since, as established in Subsection 3.2.2, the growth of p dominates both $\log p$ and $\log \pi$, the first term $\lim_{r\to\infty}[\left(\Sigma_p S\nabla \log \frac{p}{\pi}(x)\right)_i p(x)]_{-r}^{r} = 0$. We may recombine the second term across $i = 1, ...d$ to obtain:

$$\mathbb{E}_p\langle \nabla \log \frac{p}{\pi}(x), S(x - m_p)\rangle_{\mathbf{V}} = \int \sum_{i=1}^{d} \frac{d}{dx_i}\left(\Sigma_p S\nabla \log \frac{p}{\pi}(x)\right) p(x)dx \tag{3.41}$$

By the definition of the divergence operator $\nabla \cdot (\cdot)$, we have

$$\begin{aligned}
\mathbb{E}_p\langle \nabla \log \frac{p}{\pi}(x), S(x - m_p)\rangle_{\mathbf{V}} &= \mathbb{E}_p \nabla \cdot \left(\Sigma_p S\nabla \log \frac{p}{\pi}(x)\right) \\
&= \mathbb{E}_p\left(\sum_{i,j=1}^{d} (\Sigma_p S)_{i,j} \frac{d}{dx_i dx_j} \log \frac{p}{\pi}(x)\right) \\
&= \mathbb{E}_p\langle \Sigma_p S, \nabla_x^2 \log \frac{p}{\pi}(x)\rangle_{\mathbf{F}} \\
&= \langle \mathbb{E}_p\Sigma_p S, \mathbb{E}_p\nabla_x^2 \log \frac{p}{\pi}(x)\rangle_{\mathbf{F}} \\
&= \langle \mathbb{E}_p\nabla_x^2 \log \frac{p}{\pi}(x), \Sigma_p S\rangle_{\mathbf{F}}
\end{aligned} \tag{3.42}$$

We may now return to (3.38) and write:

$$\mathbb{E}_p\langle \nabla \log \frac{p}{\pi}(x), a + S(x - m_p)\rangle_{\mathbf{V}} = \langle \mathbb{E}_p\nabla \log \frac{p}{\pi}(x), a\rangle_{\mathbf{V}} + \langle \mathbb{E}_p\nabla_x^2 \log \frac{p}{\pi}(x), \Sigma_p S\rangle_{\mathbf{F}} \tag{3.43}$$

The RHS of (3.43) has the form $\langle \bar{a}, a\rangle_{\mathbf{V}} + \langle \bar{S}, \Sigma_p S\rangle_{\mathbf{F}}$ from (3.35), implying that the parameters corresponding to $\nabla_{\text{BW}} F(p_t)$ are:

$$(\bar{a}, \bar{S}) = \left(\mathbb{E}_p\nabla \log \frac{p}{\pi}(x), \mathbb{E}_p\nabla_x^2 \log \frac{p}{\pi}(x)\right) \tag{3.44}$$

Using the basic Gaussian identities $\mathbb{E}_p \log p = 0$ and $\mathbb{E}_p(\nabla_x^2 \log p(x)) = -\Sigma^{-1}$, we may rewrite (3.44) as:

$$\begin{aligned}
(\bar{a}, \bar{S}) &= \left(\mathbb{E}_p\nabla \log p(x) - \mathbb{E}_p\nabla \log \pi(x), \mathbb{E}_p\nabla_x^2 \log p(x) - \mathbb{E}_p\nabla_x^2 \log \pi(x)\right) \\
&= \left(-\mathbb{E}_p\nabla V, \mathbb{E}_p\nabla_x^2 V - \Sigma^{-1}\right)
\end{aligned} \tag{3.45}$$

We thus obtain the projection of $\nabla_{W_2}\text{KL}(p||\pi)$ onto $T_p\text{BW}(\mathbb{R}^d)$:

$$\text{proj}_{\text{BW}(\mathbb{R}^d)}\nabla_{W_2}\text{KL}(p||\pi) = \mathbb{E}_p\nabla V + (\mathbb{E}_p\nabla_x^2 V - \Sigma^{-1})(\cdot - m_p) \tag{3.46}$$

Using the results $\dot{m}_t = -\bar{a}$ and $\dot{\Sigma}_t = -\left(\bar{S}\Sigma_t + \Sigma_t\bar{S}\right)$ obtained in Appendix B.3 of [65] [8], we thus obtain $\dot{m}_t = -\mathbb{E}_p\nabla V$ and

[8] Note that, as presented in [65], these equations do not include the negative signs added here - this may be a typographical error, as the negative signs are necessary to ensure that p_t evolves in the right direction. For brevity, a full derivation of these expressions for $\dot{m}_t$ and $\dot{\Sigma}_t$ is omitted here.

$$\begin{aligned}\dot{\Sigma}_t &= -(\mathbb{E}_p \nabla_x^2 V - \Sigma^{-1})\Sigma - \Sigma(\mathbb{E}_p \nabla_x^2 V - \Sigma^{-1}) \\ &= 2I - \mathbb{E}_p(\nabla_x^2 V)\Sigma - \Sigma\mathbb{E}_p(\nabla_x^2 V)\end{aligned} \tag{3.47}$$

We may thus apply the results from Identity B.6 to obtain:

$$\dot{\Sigma}_t = 2I - \mathbb{E}_p\left(\nabla_x V(x) \otimes (x-m)^T + (x-m) \otimes \nabla_x V(x)^T\right) \tag{3.48}$$

The pair of ODEs (3.8) has thus been obtained by orthogonally projecting $\nabla_{W_2}\mathrm{KL}(p||\pi)$ onto $T_p\mathrm{BW}(\mathbb{R}^d)$.

3.3.3. Discussion

From Appendix B.4, we see that $\nabla_{W_2} KL(p||\pi)$ prescribes the same directions for mean and covariance as $\nabla_{BW} KL(p||\pi)$. If we are at a Gaussian p_0 and π is not Gaussian, then p_t^J will be heading out of the Gaussian subspace. At time $t = 0$, the means and covariances of p_t^J and p_t^L are pointing in the same direction,. However, since p_t^L, p_t^J are not geodesics in general, we cannot track them by tracing out their tangent vectors ($\nabla_{W_2} KL$ and $\nabla_{BW} KL$, respectively) and applying the exponential map(s). We also cannot assume that their means and covariances continue matching after some infinitesimal time has passed. To see this, consider the cross entropy $E_{p_t} \log \pi = \dot{m}_t$ when π is skew. After a short time Δt has passed, the Bures-JKO flow mean will be taking a step of $E_{p_{t+\delta t}^L} \log \pi$: $p_{t+\delta t}^L$ is still a Gaussian, so no skewness can be captured. Meanwhile, the JKO flow mean will be taking a step

More importantly: for the form of VI proposed in this paper, we are working with the "forward"/"left" KL divergence, which is known for its "mode-seeking" behaviour. That is, it will fit a Gaussian around one of the modes of π, ignoring points further out and underestimating their density. This is in contrast to the backward/right KL divergence, which attempts to match the moments of p_t^L and π (as I was previously looking for). If π is log-concave, then the mean of p_t^L will converge towards the (only) mode of π, which may or may not be the mean.

Consequently from the above: the JKO and Bures-JKO flows generally converge to distributions with different moments. This means that, even if $m_t^L = m_t^J$ and $\Sigma_t^L = \Sigma_t^J$ at a specific time t, they should not be equal for any time after t as they are heading to different values. In the event that the JKO and Bures-JKO flows do converge to the same moments (e.g. π is log-concave and symmetric), then we should be able to prove that that the moments follow the same trajectory to this point.

3.4. Interpretation as an Unscented Kalman Filter

In the introduction to [65], Lambert et al. claim that the ODEs (3.8) comprise "Särkkä's heuristic" for computing the moments of a Langevin diffusion x_t, with said heuristic being the UKF as described in Särkkä's publication [91]. Supported by the introduction to Kalman filtering given above in Section 2.4, the aim of this section is to illustrate why this claim is correct and how we gain a new perspective on Gaussian VI as a result.

Remark: the central insight needed to make this connection is that the "Unscented Kalman Filter" quoted in [65] is an "unobserved" case of the UKBF, whereby no observations from the measurement process are being incorporated into the flows for m_t, Σ_t. This fact is a necessary consequence of the characterisation of (3.8) as a form of VI, which does not sample directly from the true posterior π as other Bayesian methods do. More specifically, we can observe that although the ODEs in (3.8) partially describe the marginal law p_t of a particle x_t following a Langevin diffusion (as in (3.6)), no samples of x_t are used to compute (3.8). In fact, this will never be the case for any form of VI, which by definition consists of taking an expectation over the state space Ω (see Section 2.3 for details) and thus cannot incorporate information about a single particle $x_t \in \Omega$. In the context of Kalman filtering, this implies that the Kalman gain must be zero, since during an "update" step, the updated estimates will depend entirely on the previous prediction — and not at all on any hypothetical observation. To summarise this argument: if the Gaussian VI presented in [65] (or, indeed, any form of variational inference) is equivalent to a Kalman filter, it must be equivalent to an unobserved version of that filter.

Särkkä	Lambert et al.	Explanation
$f(x,t)$	$-\nabla V(x)$	This is the drift function from the Langevin diffusion (3.6) used for GVI.
$L(t)$	$\sqrt{2}I$	This is the diffusion term from the Langevin diffusion (3.6).
$Q_c(t)$	I	As specified in the introduction of [65], the diffusion process used in (3.6) is a standard Brownian motion, which has diffusion matrix I.
m_t	m_t	The mean vector of x_t.
P_t	Σ_t	The covariance matrix of x_t.
$K(t)$	0	The Kalman gain for the system.

Table 3.1: a dictionary mapping key terms in the Langevin diffusion for GVI (3.6) to the UKBF ODEs (2.19).

To proceed, we will also need to determine the specific values of the remaining non-zero terms in the UKBF ODEs (2.19): the results of doing so are provided in Table 3.1. With this information, we are now ready to formally present and prove Lambert et al.'s claim:

Lemma 3.2. *The Gaussian parameter ODEs in* (3.8) *constitute a special case of the UKBF ODEs in* (2.19)*. Specifically, equations* (3.8) *describe an unobserved instance of the UKBF with no information from the measurement process.*

Proof. First: following the remark above, we must have that the Kalman gain $K(t) = 0$. We can thus reduce (2.19) to:

$$\begin{aligned} \frac{dm_t}{dt} &= f(X(t),t)w_m \\ \frac{dP_t}{dt} &= X(t)Wf^T(X(t),t) + f(X(t),t)WX^T(t) + L(t)Q_c(t)L^T(t). \end{aligned} \tag{3.49}$$

By using (2.20), we can rewrite (3.49) using expectations:

$$\begin{aligned} \frac{dm_t}{dt} &= \mathbb{E}\left(f(x_t,t)\right) \\ \frac{dP_t}{dt} &= \mathrm{Cov}\left(x_t, f(x_t,t)\right) + \mathrm{Cov}\left(f(x_t,t), x_t\right) + L(t)Q_c(t)L^T(t). \end{aligned} \tag{3.50}$$

By applying the mappings from Table 3.1, we have:

$$\begin{aligned} \frac{dm_t}{dt} &= \mathbb{E}\left(-\nabla V(x_t)\right) \\ \frac{d\Sigma_t}{dt} &= \mathrm{Cov}\left(x_t, -\nabla V(x_t)\right) + \mathrm{Cov}\left(-\nabla V(x_t), x_t\right) + 2I. \end{aligned} \tag{3.51}$$

We now have the desired expression for m_t. To further simplify the expression for Σ_t, let us rewrite the cross-covariance terms in (3.51). Beginning with $\mathrm{Cov}\left(x_t, -\nabla V(x_t)\right)$: by using the linearity of expectation and the bilinearity of the outer product, we can write

$$\begin{aligned}
\operatorname{Cov}\left(x_t, -\nabla V(x_t)\right) &= \mathbb{E}\left(x_t \otimes -\nabla V(x_t)\right) - \mathbb{E}\left(x_t\right) \otimes \mathbb{E}\left(-\nabla V(x_t)\right) \\
&= \mathbb{E}\left(x_t \otimes -\nabla V(x_t)\right) - m_t \otimes \mathbb{E}\left(-\nabla V(x_t)\right) \\
&= \mathbb{E}\left(x_t \otimes -\nabla V(x_t)\right) - \mathbb{E}\left(m_t \otimes -\nabla V(x_t)\right) \\
&= \mathbb{E}\left(\left(x_t \otimes -\nabla V(x_t)\right) - \left(m_t \otimes -\nabla V(x_t)\right)\right) \\
&= \mathbb{E}\left(\left(x_t - m_t\right) \otimes -\nabla V(x_t)\right) \\
&= -\mathbb{E}\left(\left(x_t - m_t\right) \otimes \nabla V(x_t)\right).
\end{aligned} \tag{3.52}$$

A symmetric argument yields $\operatorname{Cov}\left(-\nabla V(x_t), x_t\right) = -\mathbb{E}\left(\nabla V(x_t) \otimes \left(x_t - m_t\right)\right)$. By applying these results to (3.51), we thus arrive at the expression for Σ_t in (3.8) and conclude the proof.

□

The interpretation of (3.8) as a special case of the UKBF is convenient for the numerical application of GVI as proposed in [65]. This is due to the use of "Gaussian quadrature rules", i.e. the unscented transform to compute the expectations required by (3.8) in practice. It should also be noted that Theorem 3.1 uses first-order approximations to arrive at (3.8): this ensures consistency between the two sets of equations, since the UKBF ODEs (2.19) are themselves merely first-order approximations of more complex dynamical systems (see [91], Appendix C).

4

Extended Analysis

In this chapter, three upper bounds for the distance $W_2(p_t^L, p_t^J)$ are proposed and their theoretical properties discussed. This work builds upon the general background in Chapter 2 and the detailed analysis performed in Chapter 3 of the Bures-JKO ODEs by Lambert et al. The bounds in Section 4.1 are presented as conjectures due to an unresolved logical gap in their construction; the bound in Section 4.2 is presented via a theorem with a complete proof.

Remark: for both Section 4.1 and Section 4.2, the following set of assumptions and notation will be made:

- Let $\{p_t^J\}_{t\geq 0} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a probability density evolving according to the Wasserstein gradient flow for the JKO scheme, i.e. the FPE (2.9). To enforce this evolution, we equip $\{p_t^J\}$ with the accompanying set of velocity fields $\{v_t^J\}_{t\geq 0} : v_t^J \in T_{P_t^J}\mathcal{P}_2(\mathbb{R}^d), v_t^J := \nabla_{W_2}\mathrm{KL}(p_t^J||\pi)$.
- Let $\{p_t^L\}_{t\geq 0} \subset \mathrm{BW}(\mathbb{R}^d)$ be a probability density evolving according to the Bures-Wasserstein gradient flow, encapsulated by the ODEs (3.8) for its mean m_t^L and covariance Σ_t^L. $\{p_t^L\}$ has the accompanying set of vector fields $\{v_t^L\}_{t\geq 0} : v_t^L \in T_{P_t^L}\mathrm{BW}(\mathbb{R}^d), v_t^L := \nabla_{\mathrm{BW}}\mathrm{KL}(p_t^L||\pi)$.
- Let $\pi \propto e^{-V} : \pi \in \mathcal{P}_2(\mathbb{R}^d)$ be the stationary target for both $\{p_t^J\}$ and $\{p_t^L\}$. We assume $V(x)$ to be a convex function (i.e. that $\nabla^2 V \succeq \alpha I$ for some $\alpha \geq 0$), making π log-concave.
- Let φ_t, ψ_t be the c-conjugate Kantorovich potentials corresponding to the optimal transport mapping used by $W_2(p_t^L, p_t^J)$.
- Under the specifications of the manifold $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ provided in Table 2.1, the term $\nabla\varphi_t$ describes the velocity of the W_2 geodesic travelling from p_t^L to p_t^J, while $\nabla\psi_t$ describes the velocity of the geodesic travelling from p_t^J to p_t^L (see Section 5.4 of [90] for more details). We describe these vectors with the shorthand notation $\nabla_1 := \nabla\varphi_t$ and $\nabla_2 := \nabla\psi_t$.

The following results rely on Grönwall's Lemma, which for convenience is provided below (without proof):

Lemma 4.1 (Grönwall's Lemma)**.** *If a differentiable function $f : [0,\infty) \to \mathbb{R}$ satisfies the following inequality for $t \in (0,\infty)$:*

$$\frac{d}{dt}f(t) \leq g(t)f(t) + b(t), \tag{4.1}$$

where g, b are L^1-integrable on $(0,\infty)$. Let $\mathcal{G}(t) := \int_0^t g(s)ds$. Then the following holds true for $t \in [0,\infty)$:

$$f(t) \leq e^{\mathcal{G}(t)}f(0) + \int_0^t e^{\mathcal{G}(t)-\mathcal{G}(s)}b(s)ds \tag{4.2}$$

Bound	Explanation	$g(t)$	$\mathcal{G}(t)$	$b(t)$
u_t^1	Conjecture 4.2	$-\frac{\sqrt{I(p_t^J\|\pi)I(p_t^J\|p_t^L)}}{\mathrm{KL}(p_t^J\|\|p_t^L)}$	$\int_0^t g(s)ds$	$\|\|v_t^L\|\|_{L^2(p_t^L)}$ $= \left(\|a_t\|^2 + \mathrm{tr}\left(S_t\Sigma_t^L S_t\right)\right)^{1/2}$
u_t^2	Conjecture 4.3	$-\frac{\sqrt{I(p_t^J\|p_t^L)}}{W_2(p_t^L,\tilde{p}_t^J)}$	$\int_0^t g(s)ds$	$\|\|v_t^L\|\|_{L^2(p_t^L)}$ $= \left(\|a_t\|^2 + \mathrm{tr}\left(S_t\Sigma_t^L S_t\right)\right)^{1/2}$
u_t^3	Theorem 4.4	-2α	$-2\alpha t$	$\|\|v_t^L - v_t^*\|\|_{L^2(p_t^L)}$ $= \left(-\|a_t\|^2 + \mathrm{tr}\left(\tilde{S}_t\Sigma_t^L\tilde{S}_t\right) + \mathbb{E}_{p_t^L}\|\nabla V\|^2\right)^{1/2}$

Table 4.1: the terms used in the three bounds obtained for $W_2(p_t^L, p_t^J)$ in Chapter 4. The alternative expressions for the velocity norms are obtained from Identities B.7 and B.8.

The three bounds u_t^1, u_t^2, u_t^3 obtained in this chapter all have the form seen in the RHS of (4.2), where we take $f(t) = W_2(p_t^L, p_t^J)$. The values of the terms in (4.2) for each bound are displayed in Table 4.1. Building upon the notation used in Appendices B and C of [65], the following shorthands shall be used throughout this chapter:

$$\begin{aligned} a_t &:= \mathbb{E}_{p_t^L}\left(\nabla_x V(x)\right) \\ S_t &:= \mathbb{E}_{p_t^L}\left(\nabla_x^2 V(x)\right) - (\Sigma_t^L)^{-1} \\ \tilde{S}_t &:= \mathbb{E}_{p_t^L}\left(\nabla_x^2 V(x)\right) - 3(\Sigma_t^L)^{-1} \end{aligned} \tag{4.3}$$

4.1. Information-based bounds

In essence, the method presented in this subsection relies on the creation of a third "particle" moving within $\mathcal{P}_2(\mathbb{R}^d)$, besides p_t^L and p_t^J. This particle, which we may call p_t^*, is initialised at p_t^J at time t and, for time $s \geq t$, follows the geodesic $\gamma(t) : \gamma(0) = p_t^L, \gamma(1) = p_t^J$ outwards from p_t^J, i.e. away from p_t^L. The speed at which p_t^* travels is determined by a coefficient c_t which is derived below. We are thus able to construct an analytically convenient sequence of distances $\{W_2(p_s^L, p_s^*)\}_{s\geq t}$ such that $\frac{d}{dt}W_2(p_s^L, p_s^*) \geq \frac{d}{dt}W_2(p_s^L, p_s^J)$ for all $s \geq t$, allowing us to apply Grönwall's Lemma and obtain the following bounds for $W_2(p_t^L, p_t^J)$:

- $\{u_t^1\}_{t\geq 0} : u_t^1 = e^{\mathcal{C}_t}W_2(p_0^L, p_0^J) + \int_0^t \|\|v_s^L\|\|_{L^2(p_s^L)}e^{\mathcal{C}_t - \mathcal{C}_s}ds$, where $c_t := \frac{\sqrt{I(p_T^J|\pi)I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)}$ and $\mathcal{C}_t := \int_0^t -c_s ds$.
- $\{u_t^2\}_{t\geq 0} : u_t^2 = e^{\mathcal{C}_t}W_2(p_0^L, p_0^J) + \int_0^t \|\|v_s^L\|\|_{L^2(p_s^L)}e^{\mathcal{C}_t - \mathcal{C}_s}ds$, where $c_t := \frac{\sqrt{I(p_t^J|p_t^L)}}{W_2(p_t^L,\tilde{p}_t^J)}$, $\tilde{p}_t^J := N(m_t^J, \Sigma_t^J)$ and $\mathcal{C}_t := \int_0^t -c_s ds$.

The goal is to obtain an upper bound for $\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J)$ and apply Grönwall's Lemma. To improve the tractability of such a bound, we should seek to make use of terms which are known — or which are already being computed, such as the constituent parts of v_t^L (see Identity B.7 for details). Hence, we consider possible bounds on $\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J)$ whilst keeping the first term $\langle\nabla_1, v_t^L\rangle_t^L$ in (4.6) fixed.

4.1.1. A first bound

Conjecture 4.2. *In addition to the assumptions listed above, let us assume that* $W_2(p_t^L, p_t^J) \neq 0$ *for all* $t \geq 0$. *Furthermore, let us define*

$$c_t := \frac{\sqrt{I(p_t^J|\pi)I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)},$$

and $\mathcal{C}_t := \int_0^t -c_s ds$. *Then we have:*

$$W_2(p_t^L, p_t^J) \leq e^{\mathcal{C}_t} W_2(p_0^L, p_0^J) + \int_0^t ||v_s^L||_{L^2(p_s^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds \tag{4.4}$$

Proof. Let us begin by stating Theorem 5.24 from [90], adapted to the present context. Using φ_t, ψ_t to denote the Kantorovich potentials associated with the distance $W_2(p_t^L, p_t^J)$:

$$\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) = \int_{\mathbb{R}^d} \nabla\varphi_t \cdot v_t^L p_t^L dx + \int_{\mathbb{R}^d} \nabla\psi_t \cdot v_t^J p_t^J dx. \tag{4.5}$$

Using the inner product $g : p \mapsto \langle \cdot, \cdot \rangle_t^L := \int \langle \cdot, \cdot \rangle dp$ provided in Table 2.1, we may rewrite (4.5) as a sum of inner products using the tangent spaces of p_t^L, p_t^J. Let us further simplify notation using the shorthand $\nabla_1 := \nabla\varphi_t$ and $\nabla_2 := \nabla\psi_t$ defined at the start of this chapter:

$$\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) = \langle \nabla_1, v_t^L \rangle_t^L + \langle \nabla_2, v_t^J \rangle_t^J \tag{4.6}$$

Note that each inner product term in (4.6) describes the motion of its respective particle against a stationary density p located at the other particle, e.g. $\langle \nabla_2, v_t^J \rangle_t^J = \frac{d}{dt}\frac{1}{2}W_2^2(p_t^J, p)$ [1]. We might therefore consider replacing $\langle \nabla_2, v_t^L \rangle_t^J$ with the corresponding inner product for a particle $\{p_t^*\}_{t\geq 0}$ that is at p_t^J at time t but is "faster" than p_t^J in the W_2^2 sense, i.e.

$$\frac{d}{dt}\frac{1}{2}W_2^2(p_t^J, p) \leq \frac{d}{dt}\frac{1}{2}W_2^2(p_t^*, p) \tag{4.7}$$

It would also be beneficial for the W_2^2 inner product expression for this particle to have a tractable form: a convenient choice for p_t^* is for p_t^* to follow the geodesic from p_t^J to p_t^L, i.e. for $v_t^* = -\nabla_2 = -\nabla\psi_t$. However, to ensure that the inequality (4.7) holds, we must introduce a coefficient $c_t \in \mathbb{R}$ to modulate the velocity of p_t^*. A visualisation of the relationship between this new particle and $\{p_t^L\}_{t\geq 0}, \{p_t^J\}_{t\geq 0}$ is provided in Figure 4.1. We therefore seek to solve the following for c_t:

$$\langle \nabla_2, v_t^J \rangle_t^J \leq \langle \nabla_2, -c_t \nabla_2 \rangle_t^J \tag{4.8}$$

Using the bilinearity of inner products, we may rearrange (4.8) to obtain:

$$c_t \leq \frac{\langle \nabla_2, v_t^J \rangle_t^J}{\langle \nabla_2, \nabla_2 \rangle_t^J} \tag{4.9}$$

Using the upper bound for the numerator $\langle \nabla_2, v_t^J \rangle_t^J \leq W_2(p_t^L, p_t^J)\sqrt{I(p_t^J|\pi)}$ from Identity B.9, as well as the fact that $\langle \nabla\psi, \nabla\psi \rangle_t^J = W_2^2(p_t^L, p_t^J)$, we obtain

$$c_t \leq \frac{\sqrt{I(p_t^J|\pi)}}{W_2(p_t^L, p_t^J)} \tag{4.10}$$

The potential of the Gaussian p_t^L is quadratic and convex, with a modulus of convexity $\alpha_t^L \geq 0$. The HWI inequality (Theorem 3 of [79]) therefore applies, which in this setting may be written as:

$$\mathrm{KL}(p_t^J||p_t^L) \leq W_2(p_t^J, p_t^L)\sqrt{I(p_t^J|p_t^L)} - \frac{\alpha_t^L}{2}W_2^2(p_t^J, p_t^L) \tag{4.11}$$

This inequality may be arranged as

[1] To see this, simply apply the same process illustrated in this proof, starting with Theorem 5.24 from [90] and using $v_p = 0$.

$$W_2(p_t^J, p_t^L) \geq \frac{\mathrm{KL}(p_t^J||p_t^L)}{\sqrt{I(p_t^J|p_t^L)}} + \frac{\alpha_t^L}{2}\frac{W_2^2(pt^J, p_t^L)}{\sqrt{I(p_t^J|p_t^L)}} \tag{4.12}$$

In particular, since $\alpha_t^L \geq 0$, we must have:

$$W_2(p_t^J, p_t^L) \geq \frac{\mathrm{KL}(p_t^J||p_t^L)}{\sqrt{I(p_t^J|p_t^L)}} \tag{4.13}$$

Consequently, we have $\frac{1}{W_2(p_t^L, p_t^J)} \leq \frac{\sqrt{I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)}$ and

$$c_t \leq \frac{\sqrt{I(p_t^J|\pi)I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)} \tag{4.14}$$

We thus obtain an upper bound for c_t which does not depend on $W_2(p_t^L, p_t^J)$. If we set c_t to be equal to this bound, we may be violating the condition (4.9): with further assumptions, however, it might be possible to ensure that this value for c_t is theoretically valid. The following subsection provides a brief discussion on why Conjecture 4.2 is likely true in at least some scenarios, but likely not true in others. These concerns notwithstanding, we thus choose to set c_t to be equal to this bound, where the numerator vanishes as $t \to \infty$ due to the term $I(p_T^J|\pi)$ and, if π is not Gaussian, the denominator tends towards a finite non-zero value. If the bound is valid, we have that

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) &\leq \langle \nabla_1, v_t^L\rangle_t^L + \langle \nabla_2, -c_t\nabla_2\rangle_t^J \\ &= \langle \nabla_1, v_t^L\rangle_t^L - c_t\langle \nabla_2, \nabla_2\rangle_t^J \\ &= \langle \nabla_1, v_t^L\rangle_t^L - c_t W_2^2(p_t^L, p_t^J)\end{aligned} \tag{4.15}$$

Applying once more the Cauchy-Schwarz expansion used in Identity B.9:

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) &\leq W_2(p_t^L, p_t^J)||v_t^L||_{L^2(P_t^L)} - c_t W_2^2(p_t^L, p_t^J) \\ &= W_2(p_t^L, p_t^J)\left(||v_t^L||_{L^2(P_t^L)} - c_t W_2(p_t^L, p_t^J)\right)\end{aligned} \tag{4.16}$$

At this point, note that by the chain rule for univariate calculus we may write

$$\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) = W_2(p_t^L, p_t^J)\frac{d}{dt}W_2(p_t^L, p_t^J) \tag{4.17}$$

Multiplying by $\frac{2}{W_2(p_t^L, p_t^J)}$ on both sides of (4.16) thus yields

$$\frac{d}{dt}W_2(p_t^L, p_t^J) \leq ||v_t^L||_{L^2(P_t^L)} - c_t W_2(p_t^L, p_t^J) \tag{4.18}$$

Assuming the definition of c_t is valid, we thus have an inequality fitting the requirements for Grönwall's Lemma. Specifically: we may identify the function $g(t) = -c_t$, which, by its smoothness and the fact that $\lim_{t\to\infty} c_t = 0$ [2], is L^1 integrable over $[0, \infty)$, as well as the function $b(t) = ||v_t^L||_{L^2(P_t^L)}$. Setting $\mathcal{C}_t = \int_0^t -c_s ds$, we may then write:

[2] This follows from the fact that $I(p_t^J|\pi) \to 0$ as $t \to \infty$, whilst neither of $I(p_t^J|p_t^L), \mathrm{KL}(p_t^J||p_t^L)$ will tend to zero provided $\pi \notin \mathrm{BW}(\mathbb{R}^d)$.

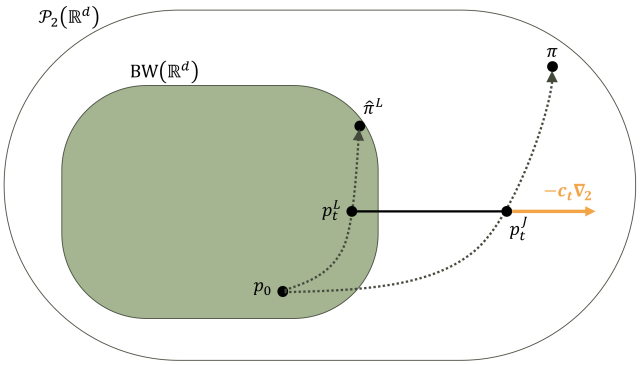

Figure 4.1: the basic scenario considered during the search for u_t^1 and u_t^2. The essential ingredient used in the attempted proofs of Conjecture 4.2 and Conjecture 4.3 is the orange outward-facing vector $-c_t\nabla_2$ which was formally introduced in (4.8). This vector points in the opposite direction to the W_2 geodesic connecting p_t^L, p_t^J: its magnitude is controlled by c_t, for which values have been proposed in Conjecture 4.2 and Conjecture 4.3.

$$W_2(p_t^L, p_t^J) \leq e^{\mathcal{C}_t} W_2(p_0^L, p_0^J) + \int_0^t ||v_s^L||_{L^2(p_s^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds. \tag{4.19}$$

□

4.1.2. Discussion and corollary bound

A full proof of Conjecture 4.2 was unfortunately not attained during this project. The remaining gap in reasoning stems from the chosen value of c_t potentially lying above the upper bound specified by (4.9) Considering the form of u_t^1, an excessively large value of c_t might cause this bound to "implode" by shrinking too rapidly, potentially becoming smaller than $W_2(p_t^L, p_t^J)$ and thus failing to provide a valid upper bound. This behaviour was not observed during the numerical simulations performed in Chapter 5, hinting that a complete solution to the conjecture exists. Moreover, it is suspected that Conjecture 4.2 is true at least when $p_0^L = p_0^J$ due to the following rationale. The Cauchy-Schwarz expansions applied in (4.10) and to the term $\langle \nabla_1, v_t^L \rangle_t^L$ in the first line of (4.16) should approximately cancel out, with the latter expansion being larger in scale due to the relatively larger value of $I(p_t^L|\pi)$ versus $I(p_t^J|\pi)$[3]. The additional inequality $\frac{1}{W_2(p_t^L, p_t^J)} \leq \frac{\sqrt{I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)}$ used to obtain (4.14) is unlikely to introduce a large error when π is log-concave [4] and so may also be dominated by the Cauchy-Schwarz expansion of $\langle \nabla_1, v_t^L \rangle_t^L$. The sign of this expansion is positive, which would justify the inequality (4.7) and therefore the validity of u_t^1.

As shown in Identity B.7, we may rewrite

$$||v_t^L||_{L^2(p_t^L)} = \left(|a_t|^2 + \mathrm{tr}(S_t \Sigma_t^L S_t)\right)^{1/2} \tag{4.20}$$

The terms a_t, S_t arise naturally through the geometric proof of Theorem 3.1 (see Section 3.3) and may be computed solely using p_t^L and V (e.g. as intermediate steps in Algorithm 1 of [65]). Unfortunately, the same cannot be said for $\mathcal{C}_t$. From the use of $\mathrm{KL}(p_t^J||p_t^L)$ and $\sqrt{I(p_t^J|p_t^L)}$ in c_t, it is clear that u_t^1 depends on the entirety of the information expressed by p_t^J. The advantage of using all this information is that we might expect u_t^1 to be a "well-informed" (and thus close) bound. The disadvantage is that

[3] This claim, in particular, is certainly not true if we drop the requirement that $p_0^L = p_0^J$, as we may arbitrarily choose p_0^J such that $I(p_t^J|\pi) \geq I(p_t^L|\pi)$. Further research may hopefully make the criteria for this "difference of Cauchy-Schwarz bounds" to apply more precise.

[4] There are only two ways a large error could be obtained here. First, we would need $\mathrm{KL}(p_t^J||p_t^L)$ to shrink significantly relative to $W_2(p_t^L, p_t^J)$: there is a limit to this shrinkage imposed by Talagrand's transportation inequality (originally proposed in [99]: see Definition 2 in [79] for a concise explanation). Second, we would need $I(p_t^J|p_t^L)$ to grow significantly relative to $\frac{1}{W_2(p_t^L, p_t^J)}$: this could occur if p_t^J assumes highly entropic, non-log-concave forms (e.g. multimodal forms), but such behaviour is restricted by the assumption of log-concavity in π.

in settings where p_t^J is unavailable (which encompasses many, if not most practical scenarios where Gaussian VI might be applied), we are unable to compute u_t^1 at all.

A variant of Conjecture 4.2 which does not invoke the HWI inequality[5] may be obtained using the Gelbrich bound for $W_2(\cdot,\cdot)$ (originally proposed in [40]: see Prop. 2.4 in [76] for a compact presentation). For two measures $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$ we have

$$W_2(\mu,\nu) \geq W_2\left(N(m_\mu,\Sigma_\mu), N(m_\nu,\Sigma_\nu)\right). \tag{4.21}$$

The W_2 distance between two Gaussian measures has the known closed-expression (3.1), making the Gelbrich bound tractable when $N(m_\mu,\Sigma_\mu), N(m_\nu,\Sigma_\nu)$ are known. We may now exploit this tractability by introducing a second bound u_t^2, based on the substitution in (4.10) made possible by (4.21). Note that the proof of this bound suffers from the same incompleteness as Conjecture 4.2: the relevant commentary provided above for u_t^1 also applies to u_t^2, which displays similar empirical performance to u_t^1 in Chapter 5.

Conjecture 4.3. *Let $W_2(p_t^L,p_t^J) \neq 0$ for all $t \geq 0$. Let $\tilde{p}_t^J := N(m_t^J,\Sigma_t^J)$ and $W_2(p_t^L,\tilde{p}_t^J) \neq 0$ for all $t \geq 0$. Let*

$$c_t := \frac{\sqrt{I(p_t^J|\pi)}}{W_2(p_t^L,\tilde{p}_t^J)}.$$

and $\mathcal{C}_t := \int_0^t -c_s ds$. Then we have

$$W_2(p_t^L,p_t^J) \leq e^{\mathcal{C}_t} W_2(p_0^L,p_0^J) + \int_0^t ||v_s^L||_{L^2(p_s^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds \tag{4.22}$$

Proof. The proof follows the same process as that for Conjecture 4.2: for brevity, we shall not reproduce it in full here. The key difference is that we redefine c_t using the Gelbrich bound $W_2(p_t^L,\tilde{p}_t^J)$ of $W_2(p_t^L,p_t^J)$. Starting with (4.9):

$$c_t \leq \frac{\langle \nabla_2, v_t^J \rangle_t^J}{\langle \nabla_2, \nabla_2 \rangle_t^J} \tag{4.23}$$

As before, we use the upper Cauchy-Schwarz bound for $\langle \nabla_2, v_t^J \rangle_t^J$ as presented in Identity B.9:

$$\langle \nabla_2, v_t^J \rangle_t^J \leq W_2(p_t^L,p_t^J)\sqrt{I(p_t^J|\pi)} \tag{4.24}$$

Note that the Gelbrich bound $W_2(p_t^L,p_t^J) \geq W_2(p_t^L,\tilde{p}_t^J)$ implies that

$$\frac{1}{W_2(p_t^L,p_t^J)} \leq \frac{1}{W_2(p_t^L,\tilde{p}_t^J)} \tag{4.25}$$

We may insert (4.24) and (4.25) into $\frac{\langle \nabla_2, v_t^J \rangle_t^J}{\langle \nabla_2, \nabla_2 \rangle_t^J}$ as follows:

$$\frac{\langle \nabla_2, v_t^J \rangle_t^J}{\langle \nabla_2, \nabla_2 \rangle_t^J} \leq \frac{\sqrt{I(p_T^J|\pi)}}{W_2(p_t^L,\tilde{p}_t^J)} \tag{4.26}$$

A new upper bound for c_t may thus be prescribed:

[5] Consequently, it may appear that u_t^2 does not require α-strong convexity for V at all. However, as we shall see in Subsection 5.2.4, u_t^2 may still fail to work correctly in non-convex settings: the reason for this remains unclear.

$$c_t \leq \frac{\sqrt{I(p_T^J|\pi)}}{W_2(p_t^L, \tilde{p}_t^J)} \tag{4.27}$$

Similarly to Conjecture 4.2, the bound (4.27) is suspected to satisfy the requirement (4.23). Furthermore, note that $W_2(p_t^L, \tilde{p}_t^J)$ cannot be obtained as a function of $W_2(p_t^L, p_t^J)$ [6]. We may therefore set $c_t = \frac{\sqrt{I(p_T^J|\pi)}}{W_2(p_t^L, \tilde{p}_t^J)}$ and proceed as in the remainder of the proof of Conjecture 4.2. □

Remark: although in most cases there are clearly multiple choices[7] of μ for which $W_2(p_t^L, \tilde{\mu}) = 0$, the condition $W_2(p_t^L, \tilde{p}_t^J) \neq 0$ is less restrictive than it may initially appear. In general, the moments of p_t^L and p_t^J will not converge towards the same final values. Assuming $\pi \in \mathcal{P}_2(\mathbb{R}^d)$, we will have $m_t^J \to m_\pi$ and $\Sigma_t^J \to \Sigma_\pi$. Meanwhile, p_t^J will centre itself about a mode of π, displaying the "mode-seeking behaviour" that is well-known within the machine learning community when minimising the *reverse* KL divergence $\mathrm{KL}(\cdot||\pi)$ (see Section 10.1.2 of [8] for a detailed explanation). Since the mode of π is not required to coincide with m_π in this setting, it follows that $m_t^L \neq m_t^J$ in general. The covariance of p_t^L will also shrink to capture a region about the mode of π, and is thus usually "smaller" than (i.e. has eigenvalues smaller than) Σ_π.

One curious implication of the mode-seeking behaviour of p_t^L described above is that in general, the Gaussian VI mean and covariance evolutions (as described by (3.8)) do *not* actually describe the same trajectories as the equivalent equations for a JKO flow p_t^J (as in (3.7)). This outcome may be somewhat unexpected given the presentation of these evolutions in [65], but is a necessary consequence of the divergence in moment trajectories described above. A further consequence resulting from this observation is that $\mathrm{KL}(p_t^J||p)$ is not minimised by $p = p_t^L$, i.e. p_t^L is not the optimal Gaussian projection of p_t^J under $\mathrm{KL}(p_t^J||\cdot)$: see Identity B.10 for more details.

Both u_t^1 and u_t^2 require the entirety of p_t^J to be known to the practitioner: as noted above, this assumption is unrealistic in many modelling/filtering scenarios. A somewhat more plausible scenario might be one in which *some* information about p_t^J is accessible, e.g. its mean and covariance [8], as needed for the denominator of Conjecture 4.3. It is tempting to construct a bound which requires only the mean and covariance of p_t^J using some simplifying assumption for $I(p_t^J|\pi)$, such as the following claim:

$$I(p_t^J|\pi) \overset{?}{\leq} I(p_t^L|\pi) \tag{4.28}$$

If $p_0^L = p_0^J$, then intuitively this assertion seems reasonable, as relative Fisher information tends to decrease as its two arguments become more similar. However, it is actually possible to disprove this claim via contradiction, meaning we cannot construct a bound this way without imposing further restrictions on p_t^L, p_t^J: see Appendix A for details.

4.2. Gradient-based bound

An alternative route for obtaining a Grönwall-style bound on $W_2(p_t^L, p_t^J)$ may be found by starting a new JKO flow[9] $\{p_s^*\}_{s \geq t} \subset \mathcal{P}_2(\mathbb{R}^d)$ at p_t^L and considering the evolution of $W_2(p_t^L, p_t^*)$. Specifically, the following bound is obtained:

[6] To see this, let us set $W_2(p_t^L, \mu) = \gamma, \gamma > 0$ and $W_2(p_t^L, \tilde{\mu}) = \xi, \xi \geq 0$ for some measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. In general, , there is not a unique μ which satisfies $W_2(p_t^L, \mu) = \gamma$: therefore, we cannot find the inverse mapping of the function $f : f(\gamma) = \xi$, which means that we cannot specify $\xi = W_2(p_t^L, \tilde{\mu})$ as a function of γ .

[7] All that is required of μ is for its mean and covariance to match those of p_t^L. For a fixed Gaussian measure p_t^L which, like all Gaussians, has skewness zero, we may consider a set of mean- and covariance-matching probability measures $\{\mu_\epsilon\}_{\epsilon \in \mathbb{R} - \{0\}}$ for which the skewness tensor has operator norm ϵ. We thus have a set of distinct, non-Gaussian measures μ_ϵ such that $W_2(p_t^L, \tilde{\mu}_\epsilon) = 0$ for all $\epsilon \in \mathbb{R} - \{0\}$.

[8] These moments could be estimated, for instance, if practitioners simultaneously run some algorithm to iteratively minimise the *forward* KL divergence, which induces "moment-matching behaviour" when used for Gaussian approximation (cf. the remark above; see [8], Section 10.1.2 for details).

[9] i.e. *not* a Bures-JKO flow: starting at p_t^L at time t, p_s^* moves across $\mathcal{P}_2(\mathbb{R}^d)$ towards π and $W_2(p_s^*, \pi) \to 0$ as $s \to \infty$.

$$\{u_t^3\}_{t\geq 0} : u_t^3 := e^{-2\alpha t} W_2(p_0^L, p_0^J) + \int_0^t e^{-2\alpha(t-s)} \left(\mathrm{tr}\left(\tilde{S}_s \Sigma_s^L \tilde{S}_s \right) + \mathbb{E}_{p_s^L} |\nabla V|^2 - |a_s|^2 \right)^{1/2} ds$$

We now demonstrate how such a bound may be obtained and comment on its features.

4.2.1. Result

Theorem 4.4. *Let the assumptions listed at the start of this chapter hold. Then we have*

$$W_2(p_t^L, p_t^J) \leq e^{-2\alpha t} W_2(p_0^L, p_0^J) + \int_0^t e^{-2\alpha(t-s)} \left(\mathrm{tr}\left(\tilde{S}_s \Sigma_s^L \tilde{S}_s \right) + \mathbb{E}_{p_s^L} |\nabla V|^2 - |a_s|^2 \right)^{1/2} ds \tag{4.29}$$

Proof. As with Conjecture 4.2, the objective is to bound $\frac{d}{dt} W_2(p_t^L, p_t^J)$ in a manner suitable for Grönwall's Lemma. We begin once more with (4.6) :

$$\frac{d}{dt} \frac{1}{2} W_2^2(p_t^L, p_t^J) = \langle \nabla_1, v_t^L \rangle_t^L + \langle \nabla_2, v_t^J \rangle_t^J \tag{4.30}$$

Let us now consider a new particle $\{p^*_{t+\delta t}\}_{\delta t \geq 0} \subset \mathcal{P}_2(\mathbb{R}^d)$ with $p_t^* = p_t^L$ velocity fields $\{v^*_{t+\delta t}\}_{\delta t \geq 0}$: $v^*_{t+\delta t} := \nabla_{W_2} \mathrm{KL}(p^*_{t+\delta t} || \pi)$. Starting at time t, p_t^* represents a new JKO-FPE flow from p_t^L to π. At time t (i.e. when $\delta t = 0$), we may thus write $\frac{d}{dt} \frac{1}{2} W_2^2(p_t^*, p_t^J) = \langle \nabla_1, v_t^* \rangle_t^L$ and introduce it as follows

$$\frac{d}{dt} \frac{1}{2} W_2^2(p_t^L, p_t^J) = \langle \nabla_1, v_t^L \rangle_t^L - \langle \nabla_1, v_t^* \rangle_t^L + \langle \nabla_1, v_t^* \rangle_t^L + \langle \nabla_2, v_t^J \rangle_t^J \tag{4.31}$$

The first two terms above may be combined as $\langle \nabla_1, v_t^L - v_t^* \rangle_t^L$; by a Cauchy-Schwarz expansion of the type employed in Identity B.9, we obtain

$$\langle \nabla_1, v_t^L - v_t^* \rangle_t^L \leq W_2(p_t^L, p_t^J) || v_t^L - v_t^* ||_{L^2(p_t^L)} \tag{4.32}$$

For the second two terms, note that by the α-strong convexity of V we may write (see Lemma 9 of [23] or Appendix D of [65]):

$$\begin{aligned} \mathrm{KL}(p_t^* || \pi) &\geq \mathrm{KL}(p_t^J || \pi) + \langle \nabla_2, v_t^J \rangle_t^J + \frac{\alpha}{2} W_2^2(p_t^*, p_t^J) \\ \mathrm{KL}(p_t^J || \pi) &\geq \mathrm{KL}(p_t^* || \pi) + \langle \nabla_1, v_t^* \rangle_t^L + \frac{\alpha}{2} W_2^2(p_t^*, p_t^J) \end{aligned} \tag{4.33}$$

We may add the two inequalities in (4.33) together and rearrange them to obtain

$$\begin{aligned} \langle \nabla_1, v_t^* \rangle_t^L + \langle \nabla_2, v_t^J \rangle_t^J &\leq -2\alpha W_2^2(p_t^*, p_t^J) \\ &= -2\alpha W_2^2(p_t^L, p_t^J) \end{aligned} \tag{4.34}$$

Inserting (4.32) and (4.34) into (4.31) yields

$$\begin{aligned} \frac{d}{dt} \frac{1}{2} W_2^2(p_t^L, p_t^J) &\leq W_2(p_t^L, p_t^J) || v_t^L - v_t^* ||_{L^2(p_t^L)} - 2\alpha W_2^2(p_t^L, p_t^J) \\ &= W_2(p_t^L, p_t^J) \left(|| v_t^L - v_t^* ||_{L^2(p_t^L)} - 2\alpha W_2(p_t^L, p_t^J) \right) \end{aligned} \tag{4.35}$$

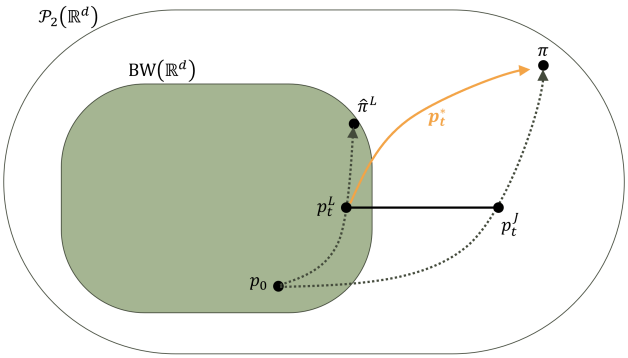

Figure 4.2: the basic scenario considered during the search for u_t^3. The essential ingredient used in the proof of Theorem 4.4 is the new JKO gradient flow p_t^*, depicted here with an orange arrow. Note that this flow is initialised at p_t^L and will converge towards the target π: when decomposed into their components, we should in principle be able to benefit from cancellations between the velocity vectors v_t^L, v_t^* at time t, which should hopefully lead to a more reliable bound. Moreover, as seen in the previous subsection, it is possible to construct v_t^L solely using p_t^L and V. of p_t^* will have Note the orange outward-facing vector $-c_t\nabla_2$ which was introduced in (4.8).

As with the proof of Conjecture 4.2, let us recall that from the univariate chain rule we have $\frac{d}{dt}\frac{1}{2}W_2^2(p_t^L, p_t^J) = W_2(p_t^L, p_t^J)\frac{d}{dt}W_2(p_t^L, p_t^J)$. Hence, we may cancel out $W_2(p_t^L, p_t^J)$ on both sides of (4.35) to obtain

$$\frac{d}{dt}W_2(p_t^L, p_t^J) \leq ||v_t^L - v_t^*||_{L^2(p_t^L)} - 2\alpha W_2(p_t^L, p_t^J) \tag{4.36}$$

This inequality satisfies the form required for Grönwall's Lemma: specifically, we may identify $g(t) = -2\alpha$ (such that $\mathcal{G}(t) = -2\alpha t$) and $b(t) = ||v_t^L - v_t^*||_{L^2(p_t^L)}$. Consequently:

$$W_2(p_t^L, p_t^J) \leq e^{-2\alpha t}W_2(p_0^L, p_0^J) + \int_0^t e^{-2\alpha(t-s)}||v_s^L - v_s^*||_{L^2(p_s^L)}ds \tag{4.37}$$

Introducing the expression for $||v_s^L - v_s^J||_{L^2(p_s^L)}$ found in Identity B.8 yields the final expression

$$W_2(p_t^L, p_t^J) \leq e^{-2\alpha t}W_2(p_0^L, p_0^J) + \int_0^t e^{-2\alpha(t-s)}\left(\mathrm{tr}\left(\tilde{S}_s\Sigma_s^L\tilde{S}_s\right) + \mathbb{E}_{p_s^L}|\nabla V|^2 - |a_s|^2\right)^{1/2}ds \tag{4.38}$$

□

4.2.2. Discussion

Needless to say, the first advantage of the bound u_t^3 over u_t^1 and u_t^2 is that its proof is actually complete. The second immediately visible benefit we obtain from Theorem 4.4 is that u_t^3 requires no further information about p_t^J other than the potential that p_t^J is heading towards. This means that u_t^3 can be computed solely using p_t^L and V — which is a more realistic scenario of what practitioners will actually have available when they attempt to estimate how accurate their Gaussian approximations $\{p_t^L\}$ are. The terms a_t and $\tilde{S}_t$ may be approximated numerically (e.g. via the Gaussian quadrature method from the UKF), and α may be estimated by computing Hessian matrices for each point on a grid of interest and obtaining the lowest eigenvalue out of all these matrices. Keen readers may observe that to compute u_t^3 we also need $W_2(p_0^L, p_0^J)$: in practice, we may choose p_0^J such that this term takes a convenient value (e.g. zero).

The loss of information about p_t^J in u_t^3 might lead to the notion that this bound will be less precise than u_t^1 or u_t^2. This is indeed a valid hypothesis, which will be tested empirically in Chapter 5. However, there is also a theoretical argument to be made in favour of u_t^3 actually being a more precise bound than u_t^1, u_t^2. The core of this argument lies in the cancellation of velocity components performed in Identity B.8, whereby considering the norm of the velocity difference $v_t^L - v_t^*$ allows us to eliminate various terms which may represent "common components" of both v_t^L and v_t^*. These cancellations

may yield a velocity field with a smaller $||\cdot||_{L^2(p_t^L)}$ norm than v_t^L itself, as used in u_t^1, u_t^2. Furthermore, the relative performance of the linear exponent $-2\alpha t$ versus the more complex $\int_0^t \frac{\sqrt{I(p_s^J|\pi)I(p_s^J|p_s^L)}}{\mathrm{KL}(p_s^J||p_s^L)}ds$ and $\int_0^t \frac{\sqrt{I(p_s^J|p_s^L)}}{W_2(p_s^L,\tilde{p}_s^J)}ds$ is not intuitively clear, and likely depends substantially on the choices of p_0^L, p_0^J and V.

5

Numerical Experiments

On paper, it is not immediately clear how the bounds for $W_2(p_t^L, p_t^J)$ obtained in Chapter 4 will perform against this distance and each other when applied to real examples. Therefore, a series of numerical experiments were performed as part of this project, with the principal findings presented in Section 5.2.

As a reminder, the three bounds for $W_2(p_t^L, p_t^J)$ obtained in Chapter 4 are:

- $u_t^1 := e^{\mathcal{C}_t^1} W_2(p_0^L, p_0^J) + \int_0^t ||v_t^L||_{L^2(p_t^L)} e^{\mathcal{C}_t^1 - \mathcal{C}_s^1} ds$, where $c_t^1 := \frac{\sqrt{I(p_t^J|\pi) I(p_t^J|p_t^L)}}{\mathrm{KL}(p_t^J||p_t^L)}$ and $\mathcal{C}_t^1 := \int_0^t -c_s^1 ds$.
- u_t^2 : the same form as u_t^1, but replacing $c_t^1, \mathcal{C}_t^1$ with $c_t^2 = \frac{\sqrt{I(p_T^J|\pi)}}{W_2(p_t^L, \tilde{p}_t^J)}$ (where $\tilde{p}_t^J = N(m_t^J, \Sigma_t^J)$) and $\mathcal{C}_t^2 := \int_0^t -c_s^2 ds$.
- $u_t^3 := e^{-2\alpha t} W_2(p_0^L, p_0^J) + \int_0^t e^{-2\alpha(t-s)} ||v_t^L - v_t^*||_{L^2(p_t^L)} ds$, where

$$||v_t^L - v_t^*||_{L^2(p_t^L)} = \left(\mathrm{tr}\left(\tilde{S}_s \Sigma_s^L \tilde{S}_s \right) + \mathbb{E}_{p_s^L} |\nabla V|^2 - |a_s|^2 \right)^{1/2}$$

  and $\tilde{S}_t := \mathbb{E}_{p_t^L} \nabla^2 V - 3(\Sigma_t^L)^{-1}$.

5.1. Setup

Each of the bounds u_t^1, u_t^2, u_t^3 relies on several intermediate coefficients which must be numerically approximated. Approximate values for a_t, S_t are obtained as intermediate steps during the numerical propagation of p_t^L (see [65], Section 4), and $\tilde{S}_t$ can be readily obtained from S_t, so these terms and their compositions $||v_t^L||_{L^2(p_t^L)}$, $||v_t^L - v_t^*||_{L^2(p_t^L)}$ do not pose substantial challenges to the simulations below. The Gelbrich bound $W_2(p_t^L, \tilde{p}_t^J)$ is also straightforward to compute once the moments of p_t^J are known: these are estimated numerically over the grid p_t^J is defined on. Similarly, the integrals $I(p_t^J|\pi), I(p_t^J|p_t^L)$ and $\mathrm{KL}(p_t^J||p_t^L)$ must also be estimated numerically over a finite grid $\Omega(N, dx, d)$. This isotropic grid, centred at zero for all examples in this report, has a side length N and resolution dx. By obtaining Hessian matrices for the values of $V(x)$ at each point $x \in \Omega(N, dx, d)$ (e.g. using finite differences), we can estimate the modulus of convexity α with the lowest out of all eigenvalues obtained from these Hessians. The term $\mathbb{E}_{p_t^L} |\nabla V|^2$ in u_t^3 is estimated using the same unscented-transform approximation used by Lambert et al. in [65] to compute a_t, S_t.

To prevent division-by-zero errors in c_t^1, c_t^2, a small constant $\epsilon = 10^{-15}$ has been added to the denominators of these terms in the code. We may thus encounter scenarios where c_t^1, c_t^2 actually fall above their theoretical upper bounds (4.14), (4.27): however, the anticipated margins of error are smaller than those resulting from approximation inaccuracies and other, more salient numerical issues.

Alongside the coefficients described above, we must propagate numerical approximations for $\{p_t^L\}, \{p_t^J\}$: in order to adequately assess the fit of each bound, we must also obtain values for $W_2(p_t^L, p_t^J)$ itself. The Gaussian particle $\{p_t^L\}$ is propagated using the code accompanying Lambert et al's publication [65],

published in the following GitHub repository [66]. This code applies the procedure briefly described in Section 4.1 of [65]: we estimate the update steps for m_t^L, Σ_t^L by performing fourth-order Runge-Kutta approximation of (3.8). The expectations $\mathbb{E}_{p_t^L} \nabla V$ and $\mathbb{E}_{p_t^L}(\nabla V \otimes (x - m_t^L))$ are estimated using the Gaussian cubature method (for more details, see Subsection 2.4.2). To preserve the positive-definite property of Σ_t^L, the code base [66] propagates the principal square root of Σ_t^L, a separate ODE for which is provided in Appendix I.2 of [65].

The general JKO flow $\{p_t^J\}$ is approximated using the FPlanck Python library [83]. This library performs grid-based numerical propagation of a probability density $p_t^J \in \mathcal{P}_2(\mathbb{R}^d)$ obeying the FPE using the method described in the accompanying publication [51]. The principle behind this method is to treat the FPE as a master equation involving an operator $\mathcal{L} : \mathcal{L}p := \Delta p + \beta \nabla \cdot (p \nabla V)$, i.e.:

$$\frac{d}{dt} p_t^J = \mathcal{L} p_t^J \tag{5.1}$$

Assuming the temperature β^{-1} and potential V are constant over time, then $\mathcal{L}$ is also constant over time and we may obtain p_t^J from p_0^J as follows ([51], Section IV):

$$p_t^J = e^{t\mathcal{L}} p_0^J \tag{5.2}$$

Let us discretise p_0^J over a finite grid $\Omega(N, dx, d)$ with $n = \frac{N}{dx}$ points along each axis, implying that the overall grid has size n^d. We may treat $\mathcal{L}$ as a large (and, in practice, sparse) matrix of dimensions (n^d, n^d): in fact, the Markovian nature of the Langevin diffusion $x_t \sim p_t^J$ makes this discrete $\mathcal{L}$ the transition matrix of a discrete-time, discrete-space Markov chain. This Markov chain is shown in [51] to satisfy the *local* detailed balance condition[1], which can be exploited to obtain tractable formulas for the non-zero entries in $\mathcal{L}$ (see Eqs. (5), (13), (14) from [51]). Propagation of p_t^J may then be formed in a numerically robust manner with the *expm_multiply()* function from SciPy [105].

As noted previously, proper evaluation of the bounds u_t^1, u_t^2, u_t^3 requires the true W_2 distances they are supposed to control. Between discrete distributions such as the discretisations of p_t^L, p_t^J, the W_2 distance becomes tractable and may be found via linear programming methods. For the experiments performed here, W_2 distances were computed using the *emd2()* function from the Python Optimal Transport (POT) library [38]. This function uses the network simplex algorithm, which was found in [12] to be the best performing out of several contenders: this algorithm iteratively moves towards the true value of $W_2^2(p_t^L, p_t^J)$ and will converge within $\mathcal{O}((n^d)^3)$ time (worst-case). More details on the network simplex algorithm are omitted here for brevity. Note that, as part of this research, attempts were also made to use the Sinkhorn approximation to the W_2 distance, since this is also possible via the *sinkhorn()* function in POT and should theoretically be faster: however, this approach did not perform as intended, yielding frequent numerical errors (infinite/NaN values) and requiring additional hyperparameter tuning to use. Hence, the solution utilised was to compute the "true" W_2 distance using *emd2()* for only a subset of the discrete time steps used during propagation, which was necessary due to the substantial cost of the W_2 computation — this being the most time- and memory-intensive component of the experiments seen below.

[1] Loosely speaking, the local detailed balance condition specifies that the change in entropy (i.e. "surprise") caused by a particular movement of mass is directly correlated with the probability of that transition occurring and inversely correlated with the probability of the *reverse* transition occurring. In the setting considered here, we may be more specific: the log-ratio between a transition probability $\mathcal{L}(x_i \to x_j)$ and its reverse motion $\mathcal{L}(x_j \to x_i)$ should be proportional to the change in entropy $\Delta S(i,j) := \frac{1}{n^d}(\nabla V(x_i) + \nabla V(x_j))$ for the transition $x_i \to x_j$, i.e.

$$\beta^{-1} \log \frac{\mathcal{L}(x_i \to x_j)}{\mathcal{L}(x_j \to x_i)} = \Delta S(i,j)$$

Keen readers may note that we are able to recover a relationship of this format from the *global* detailed balance condition, which requires that the Markov process have an equilibrium distribution π:

$$\pi(x_i)\mathcal{L}(x_i \to x_j) = \pi(x_j)\mathcal{L}(x_j \to x_i)$$

The crucial difference is that local detailed balance may still apply in situations where global detailed balance does not, e.g. because no equilibrium distribution exists. A detailed explanation of local detailed balance lies beyond the scope of this work: we refer readers to [51] or to [73] for a more thorough treatment of this condition.

Symbol	Parameter	Value	Explanation
β	Inverse temperature	$\beta = 1$	The effect of a varying β was not studied during this research, as doing so was not believed to offer substantial insight into the behaviour of u_t^1, u_t^2, u_t^3.
T	Experiment end time	In most cases: $T = 10$	Experiments were run for as long as needed to display relevant behaviour.
dt	Time step size	In most cases: $dt = 0.01$	For a fixed end time T, smaller experiment time steps naturally incur a higher real-world time cost: however, larger time steps were also observed to incur a higher real-world time cost, due to technical details behind the *expm_multiply()* function used by FPlanck (which, for brevity, must be omitted here). A large time step will also introduce inaccuracies in propagation. $dt = 0.01$ was found to strike a well-performing compromise in most cases.
N	Grid side length	Varies: generally, $8 \leq N \leq 20$	This parameter was chosen separately for each experiment, with the overall objective being to strike a balance between computational efficiency, excluding regions where the potential V became too large (which may lead to numerical errors) and not excluding regions of non-trivial mass from the target π.
dx	Grid resolution along each axis	In most cases: $dx = 0.4$	Wide grid spacing leads to approximation errors, whereas reducing the spacing very rapidly increases the computational cost of various steps. $dx = 0.4$ was empirically found to be a suitable compromise that does not leave remainders when applied to integer-valued grid lengths.
d	Dimensionality of π	$d = 2$	Higher values of d would increase the computational requirements for these experiments even further, as well as require non-trivial code adaptations to preserve the interoperability of the Lambert et al. code base and FPlanck.
skip	Interval between W_2 computations	Varies: generally, $50 \leq \mathrm{skip} \leq 200$	For a given timestep, value of $W_2(p_t^L, p_t^J)$ requires substantially more time to compute than any of its bounds or $\mathrm{KL}(p_t^J \| \| p_t^L)$; moreover, we needn't compute the W_2 values for each of the $\frac{T}{dt}$ time steps in order to understand its overall behaviour. The skip parameter determines how many time steps are skipped between successive W_2 computations: its value was tuned separately for each experiment to strike a balance between speed and expressivity.

Table 5.1: general values for key hyperparameters used in the following numerical experiments.

The inverse temperature β is a free parameter which may be specified for both the Bures-JKO and FPlanck code libraries. Furthermore, the time and space discretisations required for these experiments generate several hyperparameters which must be selected manually. The choices for these values used in this work is provided in Table 5.1. The numerical experiments presented below were performed using Python 3.11.5 on a Windows 11 computer (Intel Core i7-10510U CPU, 16GB RAM). A summary of the key software packages and the versions of these packages needed to prevent dependency conflicts is provided in Table 5.2. The specific parameter values used for each experiment are provided in Table 5.3.

Software (library)	Version
Python	3.11.5
NumPy	1.24.3
SciPy	1.11.4
FPlanck	0.2.2
POT	0.9.5

Table 5.2: required software versions for executing the code created to support this research project.

Experiment	p_0^L	p_0^J	T	dt	N	dx	skip
Generic	$N((5,5),2I)$	$N((-5,5),2I)$	60	0.01	20	0.4	1000
Gaussian, separate start	$N\left((3,3),\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$	$N((-3,3),2I)$	20	0.01	20	0.4	50
Gaussian, same start	$N\left((3,3),\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$	$N\left((3,3),\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$	10	0.01	20	0.4	20
Gaussian, off-grid	$N\left((3,3),\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$	$N\left((3,3),\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$	10	0.01	12	0.4	50
Banana	$N\left((-5,-2),\begin{pmatrix}25&0\\0&0.1\end{pmatrix}\right)$	$B\left((0,-1),\begin{pmatrix}2&0\\0&0.2\end{pmatrix},0.5\right)$	30	0.01	8	0.2	100
Symm. Bimodal	$N((0,5),2I)$	$N((0,5),2I)$	10	0.01	20	0.4	50
Asymm. Bimodal	$N((-5,0),2I)$	$N(((0,0),2I)$	60	0.01	20	0.4	100

Table 5.3: the full specifications of the parameters used for each experiment. Note that β, d are omitted, as for all experiments $\beta = 1$ and $d = 2$.

5.2. Experiments

Four categories of targets were considered for this work. Each of the seven experiments/case studies described below was chosen to help illustrate certain aspects of the behaviour of u_t^1, u_t^2, u_t^3. This section contains a detailed description of each experiment's setup and results as well as additional investigations into salient features found therein. A broader comparison between the outcomes of the experiments is provided in Section 5.3.

5.2.1. Generic target

The target explored in this subsection is intended to serve as an arbitrary, "typical" member of the set of log-concave target distributions, and has the form:

$$\begin{aligned} \pi &\propto e^{-V} \\ V(x) &:= 0.1|x+5|^4 + 10^{-5}e^{-((x_0+5)(x_1+5))^2} \end{aligned} \tag{5.3}$$

The potential $V(x)$ is indeed a convex function, as is visible in Figure 5.1; the modulus of convexity is approximately 0.1. The first simulation to approximate this target was initialised with $p_0^L = N((5,5),2I), p_0^J =$

$N((-5,5),2I)$ and run until $T = 60$. The evolutions of $W_2(p_t^L, p_t^J), \mathrm{KL}(p_t^L||p_t^J)$ as well as the bounds u_t^1, u_t^2, u_t^3 are displayed in Figure 5.2.

From Figure 5.2, we observe that each of the bounds u_t^1, u_t^2, u_t^3 does indeed provide an upper bound for $W_2(p_t^L, p_t^J)$: furthermore, these bounds converge towards stationary values on a similar timescale as the W_2 distance itself. At $t = 0$, all three bounds begin at $W_2(p_0^L, p_0^J) = 9.97860$. The information-based bounds u_t^1, u_t^2 follow a similar trajectory through the experiment, although u_t^1 is slightly lower throughout. For $t \lessapprox 10$, we see that u_t^1 and u_t^2 actually rise in value, even though W_2 is falling: given the construction of these bounds, this implies that the non-decreasing integral term $\int_0^t ||v_t^L||_{L^2(p_t^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds$ is growing more rapidly than the exponential term $e^{\mathcal{C}_t} W_2(p_0^L, p_0^J)$ is shrinking. For $t \gtrapprox 10$ we see u_t^1, u_t^2 contracting, eventually surpassing u_t^3 down to their equilibrium values $u_T^1 = 4.92108, u_T^2 = 6.45727$. u_t^3, on the other hand, undergoes only a very small increase in the first few timesteps before decreasing to $u_T^3 = 7.07805$.

5.2.2. Gaussian target

A Gaussian target offers a useful test case for the theoretical results obtained in Chapter 4, as we know more about the theoretical behaviour of p_t^L, p_t^J in this setting and may test our bounds (and our code) against this knowledge. Let $V(x) = (x - m_\pi)^T \Sigma_\pi^{-1} (x - m_\pi)$ be a Gaussian potential for

$$\begin{aligned} m_\pi &:= \begin{pmatrix} 0 \\ -2 \end{pmatrix} \\ \Sigma_\pi &:= \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix} \end{aligned} \tag{5.4}$$

The Gaussian $\pi = N(m_\pi, \Sigma_\pi)$ is log-concave, with α being equal to the smallest eigenvalue of Σ_π^{-1} [2]. Since π is Gaussian, we expect that both p_t^L and p_t^J should be able to perfectly approximate it, which would imply that $W_2(p_t^L, p_t^J) \to 0$. Let $p_0^L = N\left((3,3), \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right)$ and $p_0^J = N((-3,3), 2I)$. The results of running this experiment until $T = 20$ are displayed in Figure 5.3.

As in Subsection 5.2.1, u_t^1, u_t^2, u_t^3 all correctly provide upper bounds for the value of W_2. Starting at $W_2(p_0^J, p_0^J) = 6.03470$, the value of u_t^3 increases slightly until equilibrium is reached at $u_T^3 = 6.87083$. For $t \lessapprox 4$ we see u_t^1, u_t^2 increasing more sharply than u_t^3, but around $t \approx 4$ the information bound u_t^1 begins decreasing and thus diverging away from its Gelbrich variant u_t^2, which continues increasing until its equilibrium value $u_T^2 = 10.66889$. u_t^1, on the other hand, decreases to a stationary value $u_T^1 = 6.62913$ just below u_T^3. The proximal cause of the delayed divergence between u_t^1, u_t^2 is a spike in the value of c_t^1 relative to c_t^2, peaking around $5 \leq t \leq 7.5$: the spiking behaviour is displayed in Figure 5.4. The underlying source of this spike is unclear and remains an open question.

Another unexpected outcome from the Gaussian experiment, visible in Figure 5.3, is the convergence of $W_2(p_t^L, p_t^J)$ to the non-zero value $W_2(p_T^L, p_T^J) = 0.39957$. This issue stems from the restriction of p_t^J to the grid $\Omega(20, 0.4, 2)$, which means p_t^J may only move towards the best[3] approximation of π possible entirely on that grid, i.e. such that the sum of mass assigned by p_t^J over $\Omega(20, 0.4, 2)$ is 1. This means that, on $\Omega(20, 0.4, 2)$, p_t^J can only be *almost* Gaussian at best, and may not resemble a Gaussian at all in cases where significant amounts of the target's mass/information lie outside the grid boundaries — particularly if the missing mass is distributed unevenly relative to the grid boundaries. Meanwhile, p_t^L is propagated entirely independently from our choice of $\Omega(N, dx, d)$, with this grid only becoming relevant for estimating $W_2(p_t^L, p_t^J)$ and other terms used by the three bounds. The values of p_t^L obtained on $\Omega(N, dx, d)$ for these purposes are sampled from this grid-independent distribution, and so the *local* (or restricted) view of p_t^L we have on $\Omega(20, 0.4, 2)$ needn't resemble (i.e. approximate) the *global* form of π.

[2] Due to the method used for computing α in the code for this project, the initial value of α for this experiment was -0.0951. This was manually corrected before proceeding with propagation and bound computation.

[3] As a reminder: according to the JKO-FPE (2.9), p_t^J moves towards the "best" approximation of π in the VI sense, i.e. by minimising $\mathrm{KL}(p||\pi)$. Strictly speaking, $\mathrm{KL}(\cdot||\cdot)$ requires p and π to be defined on the same domain: although in practice $p = p_t^J$ is restricted in where it may assign non-zero mass, we may nonetheless assume that the domain of p_t^L is still $\mathbb{R}^d$ and that $p_t^J(x) = 0$ for all x outside the boundaries of $\Omega(N, dx, d)$. The condition that $p_t^J \ll \pi$ (a technical requirement for $\mathrm{KL}(p_t^J||\pi)$ to exist) is still met in this case.

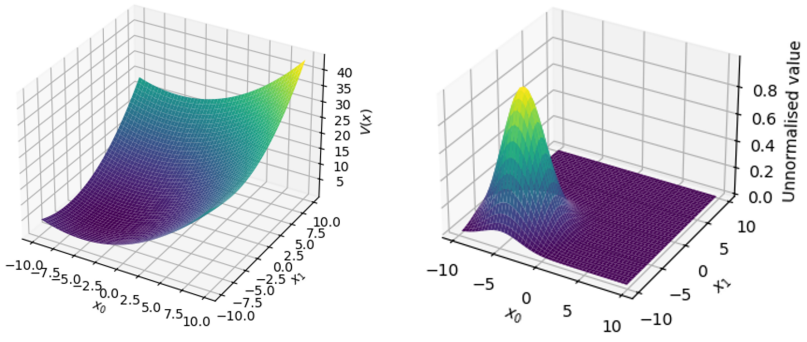

Figure 5.1: the potential $V(x)$ and the unnormalised density $e^{-V(x)}$ for the generic potential (5.3).

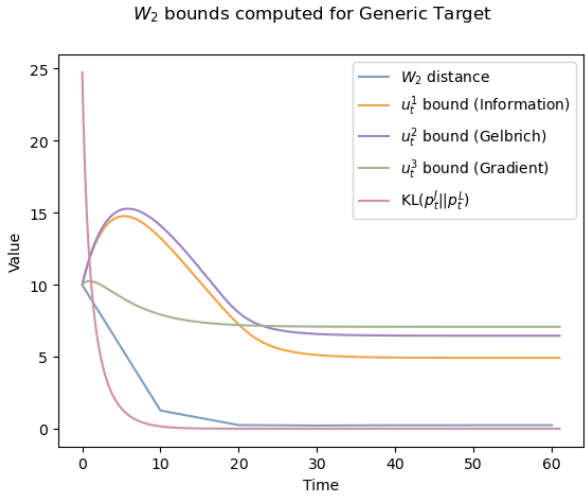

Figure 5.2: the W_2 values for the generic target experiment ($\text{skip} = 1000$), plotted against u_t^1, u_t^2, u_t^3. The Kullback-Leibler divergence $\text{KL}(p_t^J || p_t^L)$ is also plotted here, as it is computationally cheap relative to W_2 and helps establish the suitability of the linear interpolations between W_2 values displayed here.

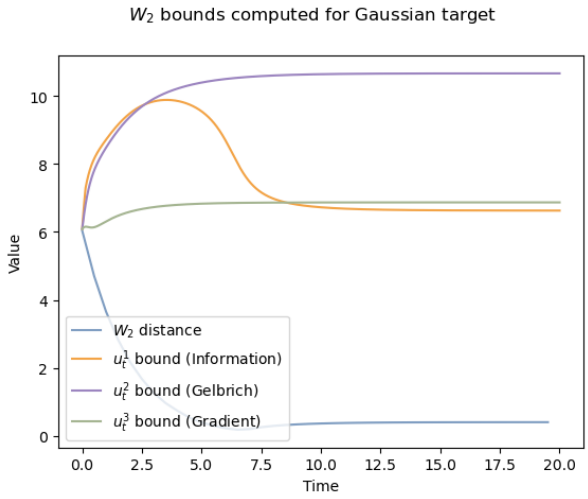

Figure 5.3: the W_2 values and bounds for the Gaussian experiment. Note the divergence between u_t^1 and u_t^2 starting at $t \approx 4$, as well as the non-zero equilibrium value of W_2.

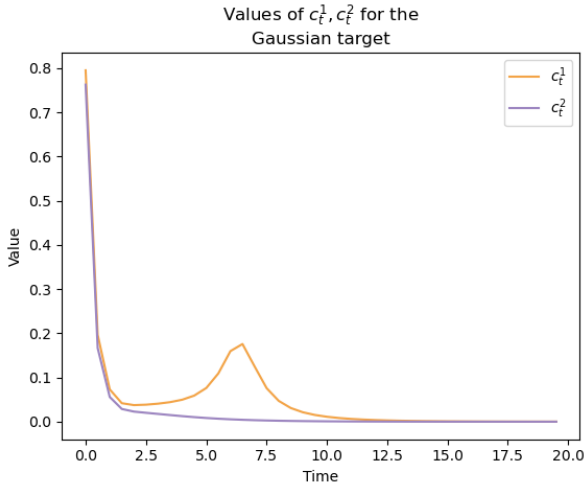

Figure 5.4: the evolution for the coefficients c_t^1, c_t^2 for the Gaussian experiment. Note the spike in the value of c_t^1 for $5 \lessapprox t \lessapprox 7.5$.

In fact, the restriction of p_t^L to $\Omega(N, dx, d)$ converges towards the best Gaussian approximation of the portion of π which can be approximated within $\Omega(N, dx, d)$: the sum of p_t^L over this grid needn't (and actually won't) be equal to 1 [4].

The crucial consequence of the previous paragraph is that p_t^L and p_t^J are generally behaving differently even as they seek to approximate the same (Gaussian) measure. Therefore, the computed value of $W_2(p_t^L, p_t^J)$ cannot and should not be precisely zero. We can see this conclusion more clearly by performing an altered version of the Gaussian experiment described above, setting both p_0^J and p_0^L to be the same Gaussian $N\left((3,3), \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right)$ and keeping the previous target $\pi = N(m_\pi, \Sigma_\pi)$. Since p_t^J is now following a Gaussian-to-Gaussian trajectory with the same start and end points as p_t^L, the geodesic convexity of $\mathrm{BW}(\mathbb{R}^d)$ in $\mathcal{P}_2(\mathbb{R}^d)$ [5] implies that $p_t^J = p_t^L$ and $W_2(p_t^L, p_t^J) = 0$ for all $t \geq 0$. The results of this "same-start Gaussian" experiment can be seen in Figure 5.5. As expected, the W_2 distance does not remain zero after $t = 0$ but rather fluctuates between 10^{-1} and 1 before converging towards its equilibrium value of 0.39925.

The difference in the behaviour of our numerical approximations of p_t^L, p_t^J becomes even more apparent when the Gaussian target's mean is moved outside the coverage of the grid. Specifically, let us reuse the starting points from the previous (same-start) Gaussian experiment and set $\pi = N((-8,-8), 2I)$: to ensure m_π falls outside the grid area, we set the grid to be $\Omega(12, 0.4, 2)$. If we run this "off-grid Gaussian" experiment up to $T = 10$ (by which time convergence has clearly occurred - see Figure 5.6), we find that $m_T^L = (-8.07989 - 8.07989)$, implying p_t^L was able to properly locate the mass of π: meanwhile, $m_T^J = (-5.63566 - 5.63566)$, which lies within the bounds of $\Omega(12, 0.4, 2)$ and illustrates how p_t^J is unable to properly account for off-grid mass. The Euclidean norm of $m_t^L - m_t^J$, displayed in Figure 5.6, further demonstrates how p_t^L, p_t^J display different numerical behaviour when propagated using this work's code base, even when their theoretical behaviour is identical.

The trajectories of $W_2(p_t^L, p_t^J)$ and $\mathrm{KL}(p_t^J || p_t^L)$ as well as the performance of the three bounds u_t^1, u_t^2, u_t^3 is displayed in Figure 5.7. The spike in the value of W_2 around $t \approx 1$ is likely caused by approximation errors of the sort described above, with $W_2(p_t^L, p_t^J)$ later decreasing to an equilibrium value near zero. because the grid approximations of both p_t^L and p_t^J will place most of their on-grid mass near the target mean $m_\pi = (-8,-8)$. The value of $\mathrm{KL}(p_t^J || p_t^L)$ fluctuates near the start of the experiment before

[4] This is true only in theory. For the experiments themselves, the values of p_t^L over $\Omega(20, 0.4, 2)$ must be normalised so that they do indeed sum to 1, as this is required for the *emd2()* function.

[5] The geodesic convexity of $\mathrm{BW}(\mathbb{R}^d)$ within $\mathcal{P}_2(\mathbb{R}^d)$ is a consequence of the the fact, noted in Section 3.3 (see Footnote 5), that the optimal transport map between two Gaussian measures has an affine form, which implies the geodesic connecting these measures also consists of Gaussian measures. This consequence is briefly explained in Appendix B.3 of [65]: a more complete treatment of the geometric properties of $\mathrm{BW}(\mathbb{R}^d)$ can be found in [98].

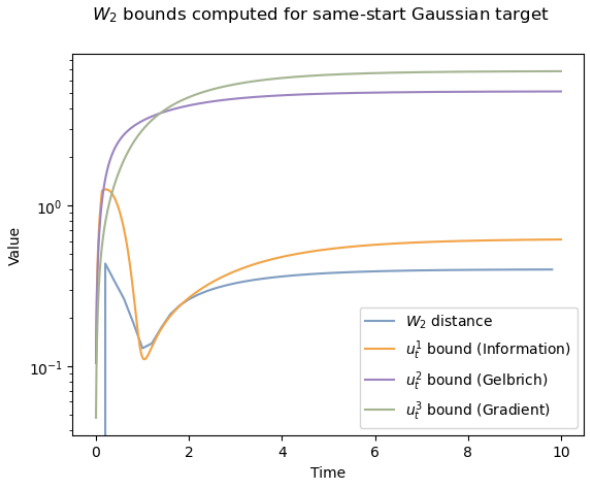

Figure 5.5: the evolution of W_2, u_t^1, u_t^2, u_t^3 for the same-start Gaussian experiment. The plot's y-axis is set to a logarithmic scale to improve visibility of the behaviour of each value after $t = 0$, when all values were automatically set to zero for this experiment. Note that u_t^1 dips slightly below W_2 around $t = 1$: this inconsistency is due to numerical approximation errors, which are amplified when working in this special case.

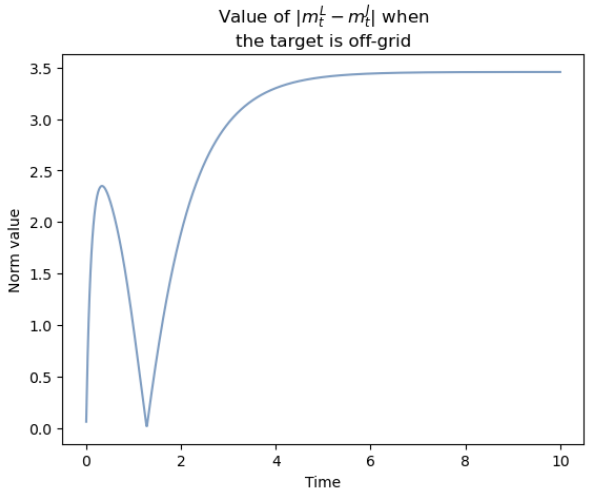

Figure 5.6: the evolution of $|m_t^L - m_t^J|$ over the duration of the off-grid Gaussian experiment. If both p_t^L, p_t^J were able to incorporate off-grid mass into their moment estimates, then this plot would be zero for all $t \geq 0$.

settling near the unusually high value of 7.81085. This unexpected behaviour is likely the result of approximation/grid errors and the particular behaviour of the *kl_div* object from SciPy which was used to compute these values: as such, it was not investigated further. Some fluctuation is also seen in the values of u_t^1, u_t^2, which may ultimately also be caused by approximation errors. u_t^3 assumes a lower, more steady trajectory than u_t^1, u_t^2, with all three bounds reaching equilibrium from $t \gtrapprox 5$ onwards.

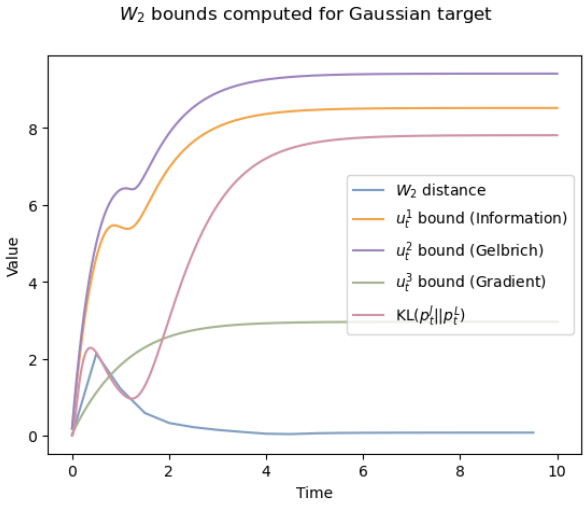

Figure 5.7: the evolution of $W_2, \mathrm{KL}, u_t^1, u_t^2, u_t^3$ over the duration of the off-grid Gaussian experiment.

5.2.3. Banana target

The two following subsections consider targets which do not satisfy the log-concave property specified in Chapter 4. The first of these involves a banana-shaped target, the potential of which is only "slightly" non-convex as it still has a unique minimiser (see Figure 5.8, in particular the density plot). This potential function V is defined via transformation of a vector $x \in \mathbb{R}^d$ using Gaussian parameters m_π, Σ_π as well as a distortion parameter $b_\pi \in \mathbb{R}$, as seen in the first line of (5.5) below. The transformed vector y is then passed through the same quadratic transformation used in a Gaussian potential. To prevent numerical errors arising from large potential values, a final transformation is then applied in the last line of (5.5).

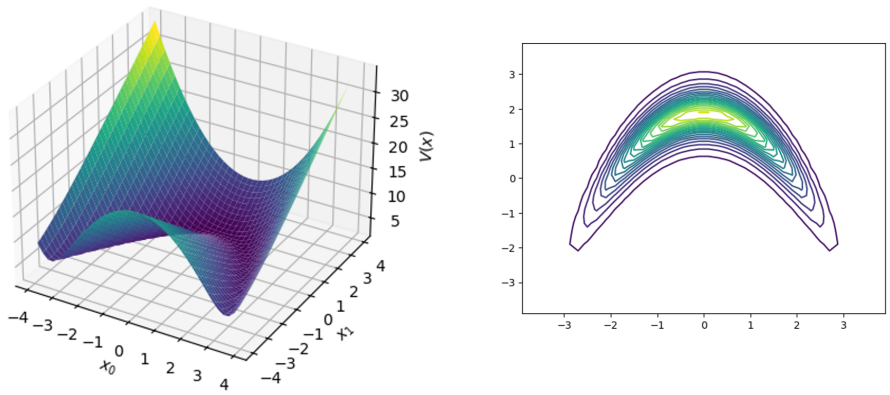

Figure 5.8: the potential $V = B(m_\pi, \Sigma_\pi, b_\pi)$ (left) and the density (right; represented here as a heat map) for the banana experiment, defined using the parameter values in (5.6). Note the non-convexity of the potential visible towards the bottom left of the potential plot (i.e. where $x_1 < 0$: the estimated modulus of convexity over the grid $\Omega(8, 0.2, 2)$ is $\alpha = -3.3663$.

$$
\begin{aligned}
y &= (x_0 - (m_\pi)_0, x_1 - (m_\pi)_1 - b\left((x_0 - (m_\pi)_0)^2 - (\Sigma_\pi)_{0,0}\right) \\
\tilde{V}(x) &:= y^T \Sigma_\pi^{-1} y \\
V(x) &:= \sqrt{|x - m_\pi|^2 + \tilde{V}(x)}
\end{aligned}
\tag{5.5}
$$

We may thus define a banana potential function $B(m_\pi, \Sigma_\pi, b_\pi) = V$. For the experiment displayed below, the following parameter values were used:

$$
\begin{aligned}
m_\pi &= (0, 2) \\
\Sigma_\pi &= \begin{pmatrix} 0.004 & 0 \\ 0 & 0.0004 \end{pmatrix} \\
b_\pi &= -0.5 \\
p_0^L &= N\left(\begin{pmatrix} -5 \\ -2 \end{pmatrix}, \begin{pmatrix} 25 & 0 \\ 0 & 0.1 \end{pmatrix}\right) \\
p_0^J &= B\left(\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 0.2 \end{pmatrix}, 0.5\right) \\
T &= 30 \\
N &= 8
\end{aligned}
\tag{5.6}
$$

The choices of p_0^L, p_0^J and V presented in (5.6) are deliberately specific and were not selected randomly or unintentionally. Rather, this combination of initial and final distributions was found to offer a strong example of the prominent multi-phase behaviour visible in Figure 5.9, which was also observed for other parameter choices. Beginning at $w_2(p_0^L, p_0^J) = 1.68938$, both u_t^1 and u_t^2 display two successive periods of rapid growth followed by slow growth. In the first of these periods (lasting from $t = 0$ to $t \approx 0.8$), u_t^1, u_t^2 are almost flat during the slow growth period. However, both u_t^1 and u_t^2 diverge rapidly away from W_2 early on during the experiment and continue growing indefinitely instead of converging to final upper bounds for W_2, which at time T has converged towards the value 0.86577. From $t \approx 4$ onwards, we see linear growth in u_t^1, u_t^2. The final values of the two bounds displayed in Figure 5.10 are $u_T^1 = 81.17605$ and $u_T^2 = 80.56184$.

A partial explanation for the behaviour of u_t^1, u_t^2 can be seen in Figure 5.11. The two peaks in the plot of the velocity field norm $||v_t^L||_{L^2(p_t^L)}$, occurring at $t \approx 1$ and $t \approx 5$, would (*ceteris paribus*) lead to increases in the value of the integral term $\int_0^t ||v_t^L||_{L^2(P_t^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds$ for u_t^1, u_t^2, which in term would explain the two rapid growth periods seen in u_t^1, u_t^2. However, we are presently unable to explain the double-peak behaviour of $||v_t^L||_{L^2(p_t^L)}$ itself, which will not be related to the bimodal phase undergone by p_t^J early on in the experiment (see Figure 5.13).

Meanwhile, the coefficients c_t^1, c_t^2 decrease logarithmically over the duration of the experiment, until settling into equilibrium values of $c_T^1, c_T^2 \approx 10^{-5}$ after $T \approx 27$. This rapid decay means that eventually, many values of $\mathcal{C}_t - \mathcal{C}_s$ will have values close to zero for a given value of t, and $e^{\mathcal{C}_t - \mathcal{C}_s}$ will have many values close to 1: see Figure 5.12 for a visualisation involving u_t^1 (u_t^2 exhibits similar behaviour). Given the convergence of $||v_t^L||_{L^2(p_t^L)}$ to the equilibrium value $||v_T^L||_{L^2(p_T^L)} = 2.47893$, the integral term $\int_0^t ||v_t^L||_{L^2(P_t^L)} e^{\mathcal{C}_t - \mathcal{C}_s} ds$ will eventually grow approximately linearly in terms of t and dominate the behaviour of u_t^1, u_t^2 - as seen in Figure 5.10. The steady logarithmic decay of c_t^1, c_t^2 cannot itself be explained at this time.

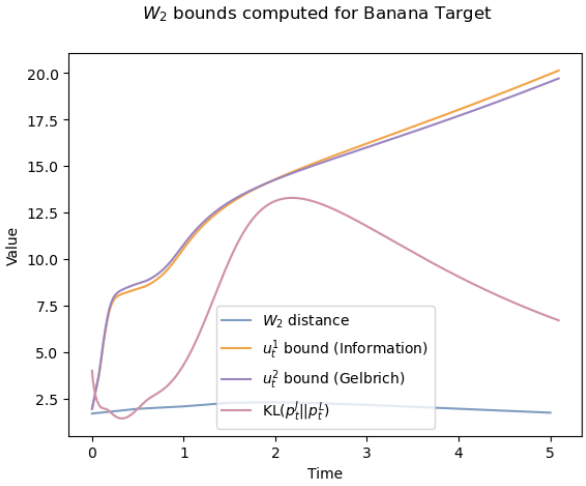

Figure 5.9: the evolution of $W_2, u_t^1, u_t^2, \mathrm{KL}$ for the banana experiment, displayed for $0 \leq t \leq 5$. u_t^3 is excluded as this bound grows exponentially when $\alpha < 0$. Note the promiment oscillation of $\mathrm{KL}(p_t^J || p_t^L)$, in particular relative to the stability of $W_2(p_t^L, p_t^J)$. This discrepancy most likely arises because p_t^J briefly becomes bimodal during its evolution from p_0^J to π (see Figure 5.13). The bimodality will affect the Kullback-Leibler divergence more strongly than the Wasserstein distance, as the latter only considers the work required to move mass whereas the former considers the ”relative surprise” of seeing a bimodal distribution instead of a unimodal Gaussian.

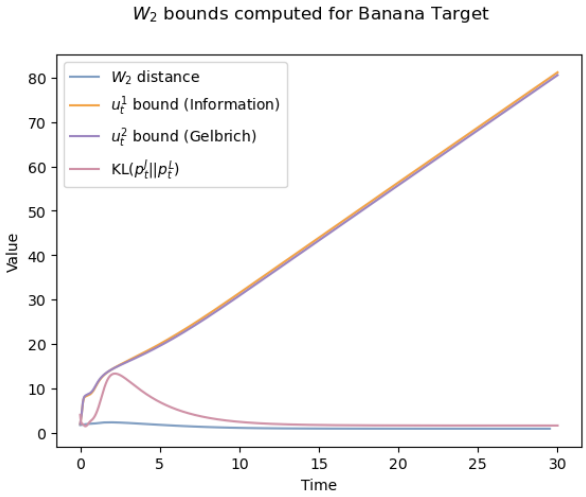

Figure 5.10: an extended version of Figure 5.9 which displays the values of $W_2, u_t^1, u_t^2, \mathrm{KL}$ until the stopping time $T = 30$ of the banana experiment. In this plot, the long-term linear behaviour of u_t^1, u_t^2 is more clearly visible.

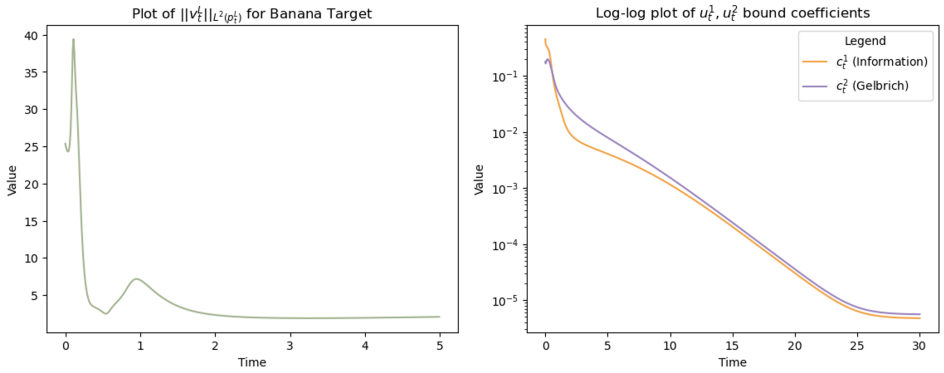

Figure 5.11: the left-hand plot displays the evolution of the velocity field norm $||v_t^L||_{L^2(p_t^L)}$ for $0 \leq t \leq 5$. Note the two peaks in the value of this series which approximately correspond to the periods of rapid growth for u_t^1, u_t^2 visible in Figure 5.9. The right-hand plot displays the log-scale evolution of the coefficients c_t^1, c_t^2 for $0 \leq t \leq 30$: according to the plot, these coefficients shrink approximately logarithmically before converging to equilibrium values $c_T^1, c_T^2 \approx 10^{-5}$ from $t \approx 27$ onwards.

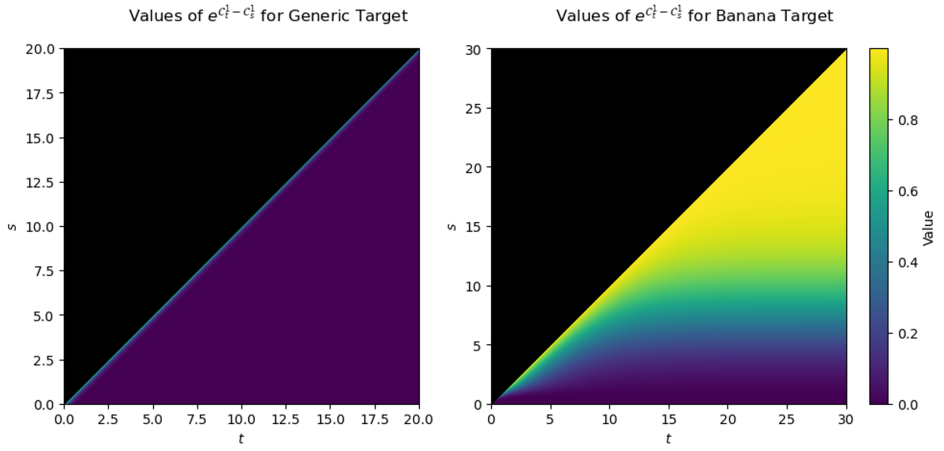

Figure 5.12: two heatmaps showing the values of $e^{\mathcal{C}_t^1 - \mathcal{C}_s^1}$ for the generic target and the banana target, respectively. Note how nearly all terms in the generic target heatmap are zero, with only a thin sliver of non-zero terms visible along the diagonal edge where $s \approx t$. Meanwhile, the banana target heatmap shows steady (and eventually linear) growth in the quantity of non-zero $e^{\mathcal{C}_t^1 - \mathcal{C}_s^1}$ values as t increases. Together with the non-zero equilibrium value of $||v_t^L||_{L^2(p_t^L)}$, the consequence of this pattern is that the integral term in u_t^1 will fail to converge.

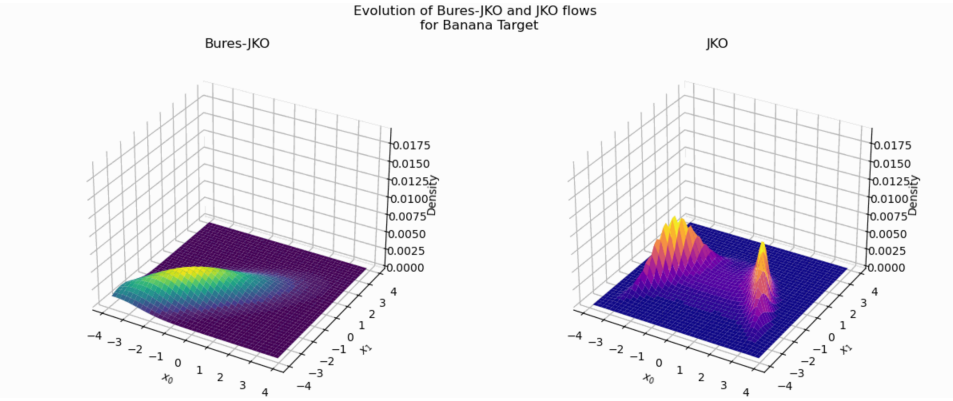

Figure 5.13: the evolution of p_t^L, p_t^J for the banana experiment, captured here at $t \approx 0.5$. Note the bimodal nature of p_t^J visible in the right-hand plot.

5.2.4. Bimodal target

Let us now turn our attention towards a ”strongly” non-convex potential which generates a bimodal distribution. The target for the following experiment is an evenly-weighted mixture of two Gaussian densities $\rho_1 = N(m_1, \Sigma_1), \rho_2 = N(m_2, \Sigma_2)$, which produces a potential of the form

$$V(x) = \log\left(\frac{\rho_1(x) + \rho_2(x)}{2}\right) \tag{5.7}$$

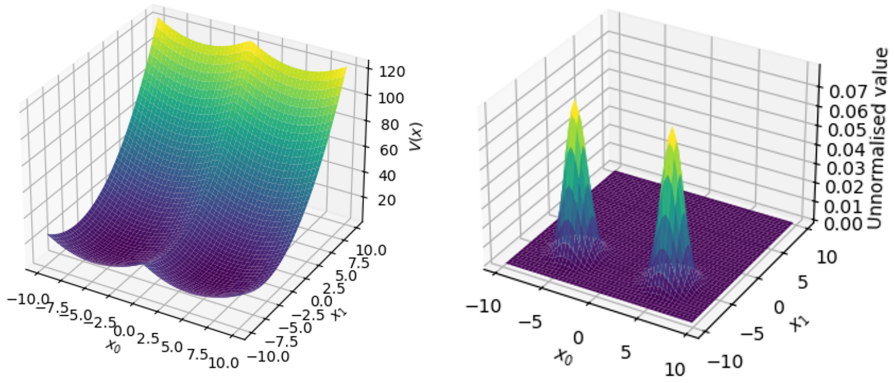

Figure 5.14: the potential (left) and the target (right) for the bimodal target. The modulus of convexity computed for this potential is $\alpha = -7.9823$.

A bimodal target $\pi \propto e^{-V}$ is thus defined for the symmetric bimodal experiment presented in this subsection: this target and its potential are displayed in Figure 5.14. The two Gaussian components are:

$$\begin{aligned} \rho_1 &= N\left((-5,-5), I\right) \\ \rho_2 &= N(((5,-5), I) \end{aligned} \tag{5.8}$$

For this experiment, we set $p_0^L = p_0^J = N((0,5), 2I)$, such that the two modes of π are equally distant from p_0^L, p_0^J: we thus define a symmetric bimodal experiment. The discretised grid used is $\Omega(20, 0.4, 2)$ and the experiment is run until $T = 10$. The evolution of $W_2(p_t^L, p_t^J)$ for this experiment as well as the bounds u_t^1, u_t^2 is displayed in Figure 5.15

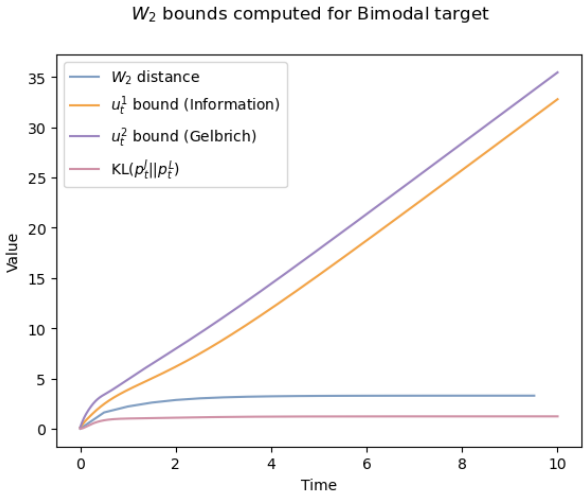

Figure 5.15: the evolution of $W_2, u_t^1, u_t^2, \mathrm{KL}$ for the bimodal target. The gradient-based bound u_t^3 is excluded as the negative value $\alpha = -7.9823$ induces exponential growth in this bound. The remaining bounds u_t^1, u_t^2 do not appear to be converging according to this plot: instead, we see long-term linear growth resembling that seen in Figure 5.10.

Starting from zero, we see that $W_2(p_t^L, p_t^J)$ increases gradually for $t \lessapprox 3$, eventually tapering off to a stationary value of $W_2(p_T^L, p_T^J) = 3.28609$. Since $p_0^L = p_0^J$, the two information-based bounds u_t^1, u_t^2 display linear growth (with parallel slopes) for $t \gtrapprox 1$ and do not converge to equilibrium values. This behaviour was also seen in Subsection 5.2.3, which suggests that the coefficients c_t^1, c_t^2 for the bimodal bounds u_t^1, u_t^2 are decreasing logarithmically: through Figure 5.16 we confirm that this is the case.

We may also consider how p_t^L, p_t^J and the bounds u_t^1, u_t^2 behave when the bimodal target is not symmetrically placed between p_0^L, p_0^J. Let us define:

$$\begin{aligned} \rho_1 &= N\left((0,-5), I\right) \\ \rho_2 &= N\left((0,0), I\right) \\ V(x) &= \log\left(\frac{\rho_1(x) + \rho_2(x)}{2}\right) \end{aligned} \tag{5.9}$$

Furthermore, let us set $p_0^L = N\left((-5,0), 2I\right)$ and $p_0^J = N\left((0,5), 2I\right)$. An asymmetric bimodal experiment is thus created, whereby both p_0^L and p_0^J are closer to ρ_2 than ρ_1: additionally, the mode ρ_2 lies directly between p_0^J and ρ_1. To properly observe the long-term behaviour of $W_2(p_t^L, p_t^J)$ as well as the information-based bounds u_t^1, u_t^2, this experiment must be run until $T = 60$.

The main results from the asymmetric bimodal experiment are displayed in Figure 5.17. The value of W_2 shows V-shaped behaviour, dropping down from $W_2(p_0^L, p_0^J) = 7.04173$ all the way down to approximately 0.48903 at $t \approx 3$ before slowly climbing again towards $W_2(p_T^L, p_T^J) = 2.67910$: this trajectory is a consequence of p_t^J assuming a unimodal, approximately Gaussian form centred around $N(m_2, \Sigma_2)$ as an intermediate step in its evolution (i.e. before p_t^J becomes bimodal). We see a similar V-shaped trajectory for $\mathrm{KL}(p_t^J || p_t^L)$. In contrast, the values of u_t^1, u_t^2 rise until $t \approx 2$ then decrease before gradually tapering off from $t \approx 20$ onwards [6]. The inverse correlation between u_t^1, u_t^2 and $W_2(p_t^L, p_t^J)$ may also be an outcome of the temporary similarity between p_t^L and p_t^J captured in Figure 5.18: however, the true cause was not uncovered during this project. Another striking feature of Figure 5.17 is the fact that u_t^1, u_t^2 both drop below the value of $W_2(p_t^L, p_t^J)$ before the end of the experiment. This is most likely a result of the non-convexity of the asymmetric bimodal potential function, making these bounds formally inapplicable to the setting of this experiment.

5.3. Evaluation

The final values of $W_2(p_t^L, p_t^J)$ as well as the three bounds u_t^1, u_t^2, u_t^3 are available in Table 5.4. The following section compares the results provided in Section 5.2 with each other and with the discussions provided in Chapter 4.

Experiment	W_2	u_T^1	u_T^2	u_T^3
Generic	0.24290	**4.92108**	6.45727	7.07805
Gaussian, separate start	0.39957	**6.62913**	10.66889	6.87083
Gaussian, same start	0.39925	**0.61413**	5.12652	6.84834
Gaussian, off-grid	0.07558	8.51917	9.41293	**2.95951**
Banana	0.86577	81.17605	**80.56184**	2.67×10^{88}
Symm. Bimodal	3.28609	**32.78878**	35.45650	1.25×10^{69}
Asymm. Bimodal	2.85395	2.67910	**2.10554**	7.09×10^{159}

Table 5.4: the final values $W_2(p_T^L, p_T^J), u_T^1, u_T^2, u_T^3$ at the end of each experiment provided in Chapter 5. The best performing (i.e. lowest) bound for each experiment is highlighted in bold. Note that according to this criterion, for the asymmetric bimodal experiment, the "best performing" bound actually lies below W_2 and is thus invalid.

The two cases which best represent appropriate scenarios for the bounds to be used in are the generic and separate-start Gaussian experiments. In these experiments, both u_t^1, u_t^2 provided valid approximations of the distance $W_2(p_t^L, p_t^J)$ and did not implode as was speculated in Subsection 4.1.2, thus

[6] Readers may have noticed that u_t^1, u_t^2 do not appear to reach true equilibrium in Figure 5.17, as they adopt a slightly positive linear growth pattern for $t \gtrapprox 40$. This may represent another instance of the linear growth trend seen in the other non-convex experiments (banana, symmetric bimodal): see Subsection 5.2.3 for a more detailed investigation into this phenomenon.

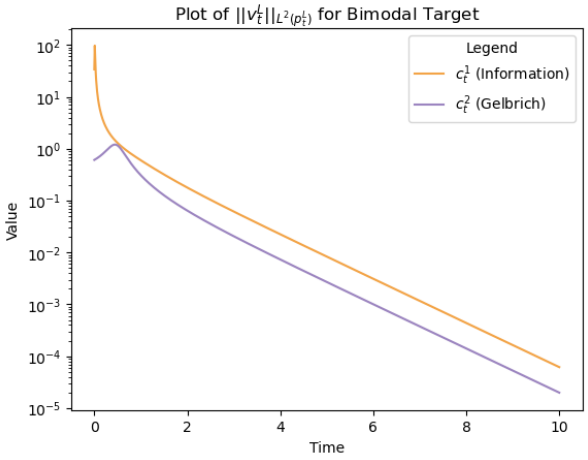

Figure 5.16: the evolution of c_t^1, c_t^2 for the bimodal target. The same logarithmic decay observed in Figure 5.11 is also visible here form $t \approx 1$ onwards.

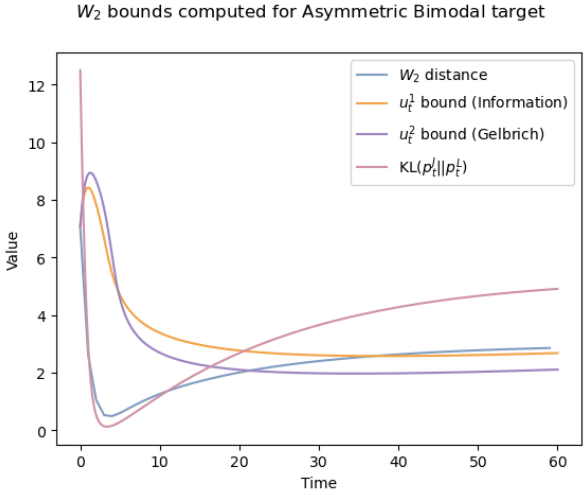

Figure 5.17: the evolution of $W_2, \mathrm{KL}, u_t^1, u_t^2$ for the asymmetric bimodal experiment. The most prominent features in this plot, visible over $0 \leq t \leq 10$, are the V-shaped trajectories of $W_2(p_t^L, p_t^J), \mathrm{KL}(p_t^J||p_t^L)$ and their inverse correlation with the upwards peaks visible in the trajectories of u_t^1, u_t^2. As with other non-log-concave targets ($\alpha = -3.0510$), the bound u_t^3 was excluded from plotting due to its exponential growth.

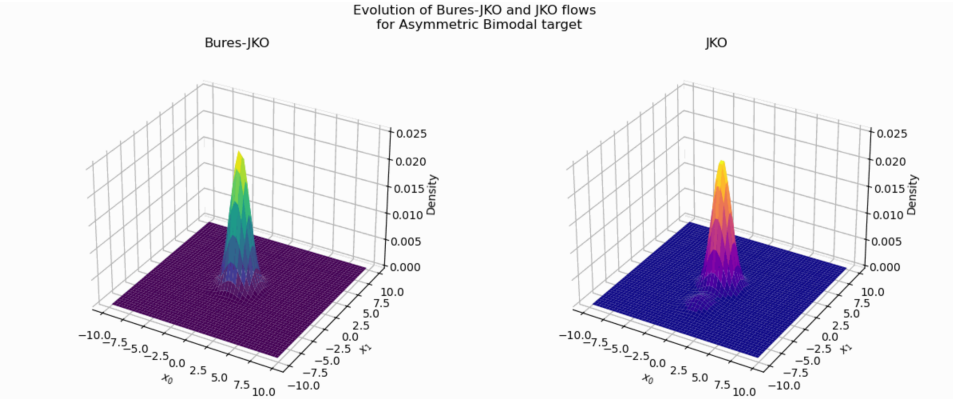

Figure 5.18: a snapshot of the propagation of p_t^L, p_t^J over the grid $\Omega(20, 0.4, 2)$ at time $t = 3$. Note that p_t^J is almost entirely unimodal and closely resembles p_t^L, which has already located the mass of the $N(m_2, \Sigma_2)$ peak in π. A second mode in p_t^J has just started growing when this snapshot was taken, and indicates where later increases in the values of $W_2(p_t^L, p_t^J)$ and $\mathrm{KL}(p_t^J||p_t^L)$ originate from.

providing empirical support for Conjecture 4.2 and Conjecture 4.3. The strong performance of u_t^1 also supports the notion (explored in Chapter 4) that incorporating the entirety of the information contained in p_t^J into the bound improves its accuracy. However, this notion is challenged by the comparatively weak performance of u_t^2 in the generic and separate-start Gaussian experiments, since this bound also requires knowing the entirety of p_t^J to compute the relative Fisher information $I(p_t^J|\pi)$. The superior performance of u_t^1 for these two experiments may be a result of the additional "knowledge" incorporated by the Kullback-Leibler comparison $\mathrm{KL}(p_t^J||p_t^L)$, or of explicitly exploiting the log-concave property required of π via invoking the HWI inequality to obtain $\mathrm{KL}(p_t^J||p_t^L)$ in the denominator in the first place. The value of u_t^3 for the generic and separate-start Gaussian experiments is higher than the best performer u_t^1, although in the separate-start Gaussian experiment the difference is small. These outcomes provide evidence against the hypothesis (q.v. Subsection 4.2.2) that the velocity cancellations performed in the construction of u_t^3 would outperform the incorporation of p_t^J into u_t^1. However, we cannot definitively conclude that u_t^1 will outperform u_t^3 based solely on two experiments, so this question remains open. Furthermore: the values of the gradient-based bound u_t^3 are of comparable magnitude to those of u_t^1, u_t^2 in the four experiments with convex potentials (generic and Gaussian). Since p_t^J may not be available in most applications of Bures-JKO VI, it is therefore important to note that the results for the convex experiments support the use of the tractable bound u_t^3 as a substitute for the stronger but less rigorous and accessible bounds u_t^1, u_t^2.

The large values of u_t^3 for the banana and bimodal experiments clearly indicate that this bound is unsuitable for use with non-convex potentials. It may be tempting to suggest that the same is not true for u_t^1, u_t^2, given that these bounds did not explode during the experiments provided: however, as noted in Section 5.2, they did not converge, either, even if convergence occurred for $W_2(p_t^L, p_t^J)$. Moreover, the application of u_t^1, u_t^2 to non-convex settings may produce other unintended outcomes such as upper bounds which lie below the quantity they are supposed to control — which was the case for the asymmetric bimodal experiment.

The remaining experiments (same-start and off-grid Gaussian, asymmetric bimodal) all comprise test cases selected to elicit special behaviour from the bounds and the experimental setup created to evaluate them. It is therefore ill-advised to draw strong conclusions about the performance of u_t^1, u_t^2, u_t^3 in "standard" settings based on their behaviour in these exceptional scenarios. For instance, the bound performance in the same-start and off-grid Gaussian experiments will almost certainly be affected by the grid-based approximation and associated numerical errors these experiments were seeking to highlight. Nonetheless, we may still gain insight into the performance of the bounds from these experiments precisely by considering their interactions with these numerical artifacts and experimental design goals. Let us consider the value of u_t^1 in the same-start Gaussian experiment, which is an order of magnitude smaller (and more accurate) than u_t^2, u_t^3: the numerical discrepancies between p_t^L and p_t^J may have led to a smaller denominator for c_t^1 (which relies on the entirety of the differences between p_t^J and p_t^L across the whole grid via $\mathrm{KL}(p_t^J||p_t^L)$) versus c_t^2, which in turn would have led to a faster contraction

in the $e^{\mathcal{C}_t} W_2(p_0^L, p_0^J)$ term for u_t^1. The denominator $\mathrm{KL}(p_t^J || p_t^L)$ of c_t^1 relies on the the discrepancies between p_t^J and p_t^L at each point on $\Omega(20, 0.4, 2)$ and so considers regions where both (approximate) distributions have similarly low mass, whereas the denominator of c_t^2 only considers the first two moments of p_t^L, p_r^J — which, as discussed in Subsection 5.2.2, exhibit anomalous differences during the propagation of the same-start Gaussian experiment due to limitations of grid-based approximation. The conclusion supported by this outcome, which aligns with the theoretical differences between $\mathrm{KL}(p_t^J || p_t^L)$ and $W_2(p_T^L, \tilde{p}_t^J)$, is that u_t^1 is more robust than u_t^2 against errors in approximating the location and spread of the underlying distributions p_t^L, p_t^J. The tradeoff, of course, is that u_t^1 is more sensitive to errors in shape approximation than u_t^2: future researchers might consider crafting experiments where the moments of p_t^L, p_t^J match even when their shapes do not.

6

Conclusion

The core motivation underlying this thesis project was to find tools to assess the suitability of the Bures-JKO scheme p_t^L (from [65]) as a method of VI, in particular when compared to the original JKO scheme p_t^J. The Wasserstein W_2 distance was chosen to measure suitability, due to the physical and geometric interpretations of mass, velocity and distance that this metric permits when applied to Wasserstein gradient flows. The principal mathematical objective for this project was to obtain upper bounds for the distance $W_2(p_t^L, p_t^J)$: three such bounds u_t^1, u_t^2, u_t^3 were obtained in Chapter 4. These bounds were then empirically tested in Chapter 5 through a series of numerical experiments intended to highlight both their strengths and weaknesses. The results of these experiments indicate that the bounds offer useful approximations of $W_2(p_t^L, p_t^J)$ under suitable conditions.

The "gradient-based" bound u_t^3 offers a particularly important tool for practitioners seeking to apply Lambert et al's Bures-JKO technique—or Gaussian VI in general. Since u_t^3 requires only the Gaussian approximation p_t^L, the potential function V for the Bayesian target and an initial value for $W_2(p_t^L, p_t^J)$ (which, for practical purposes, may be set to zero), this bound can be reliably computed even in situations when p_t^J is unavailable, as is the case for most real-world statistical modelling scenarios. Moreover, u_t^3 was found to have comparable empirical performance to the alternative bounds u_t^1, u_t^2 (which, by requiring the non-Gaussian p_t^J to be known, are far more demanding on the practitioner) when the potential is convex. However, the numerical experiments in Chapter 5 also revealed that u_t^3 explodes in value when the potential is not convex, making this property a hard requirement for u_t^3 to be applicable. A valuable direction for future research would be to try and relax the assumption of convexity/log-concavity for u_t^3 and possibly also for u_t^1, u_t^2.

Meanwhile, the "information-based" bounds u_t^1 and u_t^2 generally displayed improved performance and robustness to non-convexity when compared to u_t^3. The empirical performance of these two bounds was very similar in all but two of the experiments they were tested on (cf. the separate-start and same-start Gaussian experiments in Subsection 5.2.2). Unfortunately, as noted above, the requirement that p_t^J be known (or at least estimated) limits the applicability of u_t^1, u_t^2 to practical modelling scenarios. The status of these bounds as conjectural also warrants a more thorough investigation into their validity and limitations. Further research could additionally seek to refine these bounds by reducing their dependency on p_t^J: indeed, u_t^2 itself is the result of attempting to reduce the "knowledge" of p_t^J needed by u_t^1. An attempt to continue refining u_t^2 was also performed during this project (see Appendix A), but this was unfortunately unsuccessful.

The experiments performed in Chapter 5 form a crucial component of this project, as they offer essential insight into the bounds u_t^1, u_t^2, u_t^3 by mapping out their performance under various conditions. The practical challenges which had to be overcome in order to perform these experiments limited their quantity and scope. Therefore, there are more test cases which would further illuminate the performance of the W_2 bounds, including: targets with significant skewness, targets with a potential that is *exactly* convex (i.e. where the modulus of convexity $\alpha = 0$), and—of course—empirical data sets. Additionally, the restriction to two-dimensional experiments precludes the possibility of seeing if/how the W_2 bounds

behave differently in high-dimensional settings, which are ubiquitous in both the research and practice of modern machine learning.

During the introduction to this text, a long-term vision for estimating the suitability of Kalman Filter approximations was presented. Needless to say, this objective was not completed here: indeed, the bounds u_t^1, u_t^2, u_t^3 are not applicable as-is to a general filtering model, where the state transition function $f(x,t)$ needn't be static as it is in the Langevin diffusions associated with the JKO and Bures-JKO schemes. Nonetheless, it may eventually be possible to construct a comparable W_2 bound for the UKF using only the Gaussian parameters produced by this model as well as its model parameters (the most important of which is the state transition function). The outputs obtained from this thesis project, most notably Section 3.4 as well as the three W_2 bounds themselves, may hopefully offer future researchers clues that help solve this challenge.

In summary, this work has made a small contribution to the body of VI literature by presenting new methods to estimate the reliability of a dynamic VI technique. These bounds may be of use to practitioners seeking quality assessments or early stopping criteria for their VI models, and may even assist theoretical researchers in optimal transport, PDEs and Bayesian statistics by offering new stepping stones for proof construction. Consequently, this text has hopefully made the power and attractiveness of variational inference as a modelling solution.

References

[1] Fabian Altekrüger, Johannes Hertrich, and Gabriele Steidl. “Neural Wasserstein Gradient Flows for Discrepancies with Riesz Kernels”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 664–690. URL: https://proceedings.mlr.press/v202/altekruger23a.html.

[2] Jason M. Altschuler et al. “Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent”. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.

[3] Shun-ichi Amari and Takeru Matsuda. “Information geometry of Wasserstein statistics on shapes and affine deformations”. In: *Information Geometry* 7 (July 2024). DOI: https://doi.org/10.1007/s41884-024-00139-y.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. 2. ed. Lectures in Mathematics ETH Zürich. OCLC: 254181287. Basel: Birkhäuser, 2008. 334 pp. ISBN: 978-3-7643-8722-8 978-3-7643-8721-1.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 214–223. URL: https://proceedings.mlr.press/v70/arjovsky17a.html.

[6] David Barber and Christopher M Bishop. “Ensemble Learning for Multi-Layer Networks”. In: *Neural Information Processing Systems* 10 (Dec. 1997), pp. 395–401.

[7] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. “On the Bures–Wasserstein distance between positive definite matrices”. In: *Expositiones Mathematicae* 37.2 (2019), pp. 165–191. ISSN: 0723-0869. DOI: https://doi.org/10.1016/j.exmath.2018.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0723086918300021.

[8] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 9781493938438.

[9] Niloy Biswas and Lester Mackey. “Bounding Wasserstein Distance with Couplings”. In: *Journal of the American Statistical Association* 119.548 (2024), pp. 2947–2958. DOI: 10.1080/01621459.2023.2287773. eprint: https://doi.org/10.1080/01621459.2023.2287773. URL: https://doi.org/10.1080/01621459.2023.2287773.

[10] Adrien Blanchet, Vincent Calvez, and José A. Carrillo. “Convergence of the Mass-Transport Steepest Descent Scheme for the Subcritical Patlak–Keller–Segel Model”. In: *SIAM Journal on Numerical Analysis* 46.2 (2008), pp. 691–721. DOI: 10.1137/070683337. eprint: https://doi.org/10.1137/070683337. URL: https://doi.org/10.1137/070683337.

[11] François Bolley, Ivan Gentil, and Arnaud Guillin. “Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations”. In: *Journal of Functional Analysis* 263.8 (2012), pp. 2430–2457. ISSN: 0022-1236. DOI: https://doi.org/10.1016/j.jfa.2012.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0022123612002777.

[12] Nicolas Bonneel et al. “Displacement interpolation using Lagrangian mass transport”. In: *ACM Trans. Graph.* 30.6 (Dec. 2011), pp. 1–12. ISSN: 0730-0301. DOI: 10.1145/2070781.2024192. URL: https://doi.org/10.1145/2070781.2024192.

[13] Siwan Boufadene and François-Xavier Vialard. “On the Global Convergence of Wasserstein Gradient Flow of the Coulomb Discrepancy”. In: *SIAM Journal on Mathematical Analysis* 57.4 (2025), pp. 4556–4587. DOI: 10.1137/24M1647631. eprint: https://doi.org/10.1137/24M1647631. URL: https://doi.org/10.1137/24M1647631.

[14] Bruce van Brunt. *The calculus of variations*. New York: Springer, 2004. ISBN: 9781441923165.

[15] Clément Cancès et al. “Simulation of multiphase porous media flows with minimising movement and finite volume schemes”. In: *European Journal of Applied Mathematics* 30.6 (Oct. 2018), pp. 1123–1152. DOI: https://doi.org/10.1017/s0956792518000633.

[16] Guillaume Carlier and Clarice Poon. “On the total variation Wasserstein gradient flow and the TV-JKO scheme”. In: *ESAIM: COCV* 25 (2019), p. 42. DOI: 10.1051/cocv/2018042. URL: https://doi.org/10.1051/cocv/2018042.

[17] Guillaume Carlier et al. “Convergence of Entropic Schemes for Optimal Transport and Gradient Flows”. In: *Siam Journal on Mathematical Analysis* 49.2 (Jan. 2017), pp. 1385–1418. DOI: https://doi.org/10.1137/15m1050264.

[18] José A. Carrillo et al. “Primal Dual Methods for Wasserstein Gradient Flows”. In: *Foundations of Computational Mathematics* 22.2 (Mar. 2021), pp. 389–443. DOI: https://doi.org/10.1007/s10208-021-09503-1.

[19] Edward Challis and David Barber. “Gaussian Kullback-Leibler Approximate Inference”. In: *Journal of Machine Learning Research* 14.68 (2013), pp. 2239–2286. URL: http://jmlr.org/papers/v14/challis13a.html.

[20] Peter G Chang, Kevin Patrick Murphy, and Matt Jones. “On diagonal approximations to the extended Kalman filter for online training of Bayesian neural networks”. In: *Continual Lifelong Learning Workshop at ACML 2022*. 2022.

[21] Peter G. Chang et al. “Low-rank extended Kalman filtering for online learning of neural networks from streaming data”. In: *Proceedings of The 2nd Conference on Lifelong Learning Agents*. Ed. by Sarath Chandar et al. Vol. 232. Proceedings of Machine Learning Research. PMLR, 22–25 Aug 2023, pp. 1025–1071. URL: https://proceedings.mlr.press/v232/chang23a.html.

[22] Samantha Chen and Yusu Wang. “Neural approximation of Wasserstein distance via a universal architecture for symmetric and factorwise group invariant functions”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 9506–9517. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/1e5f58d98523298cba093f658cfdf2d6-Paper-Conference.pdf.

[23] Xiang Cheng and Peter Bartlett. “Convergence of Langevin MCMC in KL-divergence”. In: *Proceedings of Algorithmic Learning Theory*. Ed. by Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 186–211.

[24] Sinho Chewi. “An optimization perspective on log-concave sampling and beyond”. PhD Thesis. Massachusetts Institute of Technology: Massachusetts Institute of Technology, May 2023.

[25] Sinho Chewi et al. “Gradient descent algorithms for Bures-Wasserstein barycenters”. In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1276–1304. URL: https://proceedings.mlr.press/v125/chewi20a.html.

[26] John M. Cioffi. *Appendix C: Linear Algebra and Matrix Calculus*. Part of the online book “Data Transmission Theory”, available at: https://cioffi-group.stanford.edu/ . URL: https://cioffi-group.stanford.edu/doc/book/AppendixC.pdf.

[27] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Neural Information Processing Systems* 26 (Dec. 2013), pp. 2292–2300.

[28] Niladri Das, Jed A. Duersch, and Tommie A. Catanach. “Variational Kalman Filtering with H∞-Based Correction for Robust Bayesian Learning in High Dimensions”. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. 2022, pp. 7522–7528. DOI: 10.1109/CDC51059.2022.9992586.

[29] P. Del Moral and A. Niclas. “A Taylor expansion of the square root matrix function”. In: *Journal of Mathematical Analysis and Applications* 465.1 (2018), pp. 259–266. ISSN: 0022-247X. DOI: https://doi.org/10.1016/j.jmaa.2018.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0022247X18303901.

[30] Simone Di Marino and Filippo Santambrogio. “JKO estimates in linear and non-linear Fokker–Planck equations, and Keller–Segel: L^p and Sobolev bounds”. en. In: *Annales de l’I.H.P. Analyse non linéaire* 39.6 (2022), pp. 1485–1517. DOI: 10.4171/aihpc/36. URL: https://www.numdam.org/articles/10.4171/aihpc/36/.

[31] Michael Ziyang Diao et al. “Forward-Backward Gaussian Variational Inference via JKO in the Bures-Wasserstein Space”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 7960–7991. URL: https://proceedings.mlr.press/v202/diao23a.html.

[32] Hao Ding and Shizan Fang. “Geometry on the Wasserstein Space Over a Compact Riemannian Manifold”. In: *Acta Mathematica Scientia* 41.6 (Nov. 2021), pp. 1959–1984. DOI: https://doi.org/10.1007/s10473-021-0612-4.

[33] Gerardo Duran-Martin et al. *A unifying framework for generalised Bayesian online learning in non-stationary environments*. 2025. arXiv: 2411.10153 [stat.ML]. URL: https://arxiv.org/abs/2411.10153.

[34] Lev D. Elsgolc. *Calculus of variations*. Mineola, N.Y.: Dover, Jan. 2007. ISBN: 9780486457994.

[35] Hassan Emamirad and Arnaud Rougirel. “De Bruijn identities in different Markovian channels”. In: *Electronic Journal of Differential Equations* 2023.01-87 (Feb. 2023), pp. 12–12. DOI: https://doi.org/10.58997/ejde.2023.12.

[36] Jiaojiao Fan et al. “Variational Wasserstein gradient flow”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 6185–6215. URL: https://proceedings.mlr.press/v162/fan22d.html.

[37] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. European Mathematical Society, May 2023. ISBN: 9783985470501.

[38] Rémi Flamary et al. “POT: Python Optimal Transport”. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8. URL: http://jmlr.org/papers/v22/20-451.html.

[39] Charlie Frogner et al. “Learning with a Wasserstein Loss”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/a9eb812238f753132652ae09963a05e9-Paper.pdf.

[40] Matthias Gelbrich. “On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces”. In: *Mathematische Nachrichten* 147.1 (Jan. 1990), pp. 185–203. DOI: https://doi.org/10.1002/mana.19901470121.

[41] I M Gelfand and S V Fomin. *Calculus of Variations*. Courier Corporation, Apr. 2012. ISBN: 9780486135014.

[42] Nathael Gozlan and Christian Léonard. “Transport Inequalities. A Survey”. In: *arXiv e-prints*, arXiv:1003.3852 (Mar. 2010), arXiv:1003.3852. DOI: 10.48550/arXiv.1003.3852. arXiv: 1003.3852 [math.PR].

[43] K. Grauman and T. Darrell. “Fast contour matching using approximate earth mover’s distance”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 1. 2004, pp. I–I. DOI: 10.1109/CVPR.2004.1315035.

[44] Fredrik Gustafsson and Gustaf Hendeby. “Some Relations Between Extended and Unscented Kalman Filters”. In: *IEEE Transactions on Signal Processing* 60.2 (2012), pp. 545–555. DOI: 10.1109/TSP.2011.2172431.

[45] Patrick Hamill. *A student’s guide to Lagrangians and Hamiltonians*. Cambridge: Cambridge University Press, 2018. ISBN: 9781107042889.

[46] Andi Han et al. “On Riemannian Optimization over Positive Definite Matrices with the Bures-Wasserstein Geometry”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8940–8953. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/4b04b0dcd2ade339a3d7ce13252a29d4-Paper.pdf.

[47] Yuzhuo Han et al. “Wasserstein Loss based Deep Object Detection”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 4299–4305. DOI: 10.1109/CVPRW50498.2020.00507.

[48] Nicholas J. Higham and Samuel D. Relton. “Higher Order Fréchet Derivatives of Matrix Functions and the Level-2 Condition Number”. In: *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), pp. 1019–1037. DOI: 10.1137/130945259. eprint: https://doi.org/10.1137/130945259. URL: https://doi.org/10.1137/130945259.

[49] Bastian Hilder et al. “An inequality connecting entropy distance, Fisher Information and large deviations”. In: *Stochastic Processes and their Applications* 130.5 (2020), pp. 2596–2638. ISSN: 0304-4149. DOI: https://doi.org/10.1016/j.spa.2019.07.012. URL: https://www.sciencedirect.com/science/article/pii/S0304414919300687.

[50] Geoffrey E. Hinton and Drew van Camp. “Keeping the neural networks simple by minimizing the description length of the weights”. In: *Proceedings of the Sixth Annual Conference on Computational Learning Theory*. COLT ’93. Santa Cruz, California, USA: Association for Computing Machinery, 1993, pp. 5–13. ISBN: 0897916115. DOI: 10.1145/168304.168306. URL: https://doi.org/10.1145/168304.168306.

[51] Viktor Holubec, Klaus Kroy, and Stefano Steffenoni. “Physically consistent numerical solver for time-dependent Fokker-Planck equations”. In: *Phys. Rev. E* 99 (3 Mar. 2019), p. 032117. DOI: 10.1103/PhysRevE.99.032117. URL: https://link.aps.org/doi/10.1103/PhysRevE.99.032117.

[52] Antti Honkela and Harri Valpola. “Unsupervised Variational Bayesian Learning of Nonlinear Models”. In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press, 2004. URL: https://proceedings.neurips.cc/paper_files/paper/2004/file/8169e05e2a0debcb15458f2cc1eff0ea-Paper.pdf.

[53] Jonathan Huggins et al. “Validated Variational Inference via Practical Posterior Error Bounds”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 26–28 Aug 2020, pp. 1792–1802. URL: https://proceedings.mlr.press/v108/huggins20a.html.

[54] Matt Jacobs, Wonjun Lee, and Flavien Léger. “The back-and-forth method for Wasserstein gradient flows”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 27 (Mar. 2021). DOI: https://doi.org/10.1051/cocv/2021029. URL: https://www.esaim-cocv.org/articles/cocv/abs/2021/02/cocv200249/cocv200249.html.

[55] Andrew H Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Publications, Nov. 2007. ISBN: 9780486318196.

[56] Matt Jones, Peter Chang, and Kevin Murphy. “Bayesian Online Natural Gradient (BONG)”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 131104–131153. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/ecffd829f90b0a4b6aa017b6df15904f-Paper-Conference.pdf.

[57] Richard F Jordan, David Kinderlehrer, and Felix Otto. “The Variational Formulation of the Fokker–Planck Equation”. In: *Siam Journal on Mathematical Analysis* 29.1 (Jan. 1998), pp. 1–17. DOI: https://doi.org/10.1137/s0036141096303359.

[58] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. *A new approach for filtering nonlinear systems*. June 1995. DOI: https://doi.org/10.1109/ACC.1995.529783. URL: https://ieeexplore.ieee.org/document/529783.

[59] Peter Kandolf and Samuel D. Relton. “A Block Krylov Method to Compute the Action of the Fréchet Derivative of a Matrix Function on a Vector with Applications to Condition Number Estimation”. In: *SIAM Journal on Scientific Computing* 39.4 (Jan. 2017), A1416–A1434. DOI: https://doi.org/10.1137/16m1077969.

[60] Mohammad Emtiyaz Khan et al. “Vprop: Variational Inference using RMSprop”. In: *arXiv* (Dec. 2017). DOI: https://doi.org/10.48550/arxiv.1712.01038.

[61] Mark Kot. *A first course in the calculus of variations*. Providence, Rhode Island: American Mathematical Society, 2014. ISBN: 9781470414955.

[62] Daniel Kuhn et al. “Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning”. In: *Operations Research & Management Science in the Age of Analytics* (Aug. 2019), pp. 130–166. DOI: https://doi.org/10.1287/educ.2019.0198.

[63] Marc Lambert, Silvère Bonnabel, and Francis Bach. “The continuous-discrete variational Kalman filter (CD-VKF)”. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. 2022, pp. 6632–6639. DOI: 10.1109/CDC51059.2022.9992993.

[64] Marc Lambert, Silvère Bonnabel, and Francis Bach. “The recursive variational Gaussian approximation (R-VGA)”. In: *Statistics and Computing* 32.1 (Dec. 2021). DOI: https://doi.org/10.1007/s11222-021-10068-w.

[65] Marc Lambert et al. “Variational inference via Wasserstein gradient flows”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 14434–14447. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/5d087955ee13fe9a7402eedec879b9c3-Paper-Conference.pdf.

[66] Marc H. Lambert. *GitHub - marc-h-lambert/W-VI: The companion code for the paper “Variational inference via Wasserstein gradient flows (W-VI)”*. 2022. URL: https://github.com/marc-h-lambert/W-VI/tree/main.

[67] Jean B. Lasserre. “The existence of Gaussian cubature formulas”. In: *Journal of Approximation Theory* 164.5 (2012), pp. 572–585. ISSN: 0021-9045. DOI: https://doi.org/10.1016/j.jat.2012.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0021904512000135.

[68] Wuchen Li and Lexing Ying. “Hessian transport gradient flows”. In: *Research in the Mathematical Sciences* 6.4 (Oct. 2019). DOI: https://doi.org/10.1007/s40687-019-0198-9.

[69] Haibin Ling and Kazunori Okada. “An Efficient Earth Mover’s Distance Algorithm for Robust Histogram Comparison”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.5 (2007), pp. 840–853. DOI: 10.1109/TPAMI.2007.1058.

[70] Huidong Liu, Xianfeng GU, and Dimitris Samaras. “A Two-Step Computation of the Exact GAN Wasserstein Distance”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3159–3168. URL: https://proceedings.mlr.press/v80/liu18d.html.

[71] Tianle Liu et al. “Towards Understanding the Dynamics of Gaussian-Stein Variational Gradient Descent”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 61234–61291. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/c0ae487420ebc8d0ed7c541b4e3f09d4-Paper-Conference.pdf.

[72] Kathryn Lund and Marcel Schweitzer. “The Fréchet derivative of the tensor t-function”. In: *Calcolo* 60.3 (June 2023). DOI: https://doi.org/10.1007/s10092-023-00527-3.

[73] Christian Maes. “Local detailed balance”. In: *SciPost Phys. Lect. Notes* (2021), p. 32. DOI: 10.21468/SciPostPhysLectNotes.32. URL: https://scipost.org/10.21468/SciPostPhysLectNotes.32.

[74] Petr Mokrov et al. “Large-Scale Wasserstein Gradient Flows”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 15243–15256. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/810dfbbebb17302018ae903e9cb7a483-Paper.pdf.

[75] Matteo Muratori and Giuseppe Savaré. “Gradient flows and Evolution Variational Inequalities in metric spaces. I: Structural properties”. In: *Journal of Functional Analysis* 278.4 (2020), p. 108347. ISSN: 0022-1236. DOI: https://doi.org/10.1016/j.jfa.2019.108347. URL: https://www.sciencedirect.com/science/article/pii/S0022123619303416.

[76] Viet Anh Nguyen et al. “Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization”. In: *Mathematics of Operations Research* 48.1 (Dec. 2021). DOI: https://doi.org/10.1287/moor.2021.1176.

[77] 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.

[78] Manfred Opper and Cédric Archambeau. “The Variational Gaussian Approximation Revisited”. In: *Neural Computation* 21.3 (Mar. 2009), pp. 786–792. DOI: https://doi.org/10.1162/neco.2008.08-07-592.

[79] F. Otto and C. Villani. “Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality”. In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400. ISSN: 0022-1236. DOI: https://doi.org/10.1006/jfan.1999.3557. URL: https://www.sciencedirect.com/science/article/pii/S0022123699935577.

[80] Félix Otto. “The geometry of dissipative evolution equations: the porous medium equation”. In: *Communications in Partial Differential Equations* 26.1-2 (Jan. 2001), pp. 101–174. DOI: https://doi.org/10.1081/pde-100002243.

[81] Victor M. Panaretos and Yoav Zemel. “Statistical Aspects of Wasserstein Distances”. In: *Annual Review of Statistics and Its Application* 6.1 (Mar. 2019), pp. 405–431. DOI: https://doi.org/10.1146/annurev-statistics-030718-104938.

[82] Min Sue Park et al. “The deep minimizing movement scheme”. In: *Journal of Computational Physics* 494 (2023), p. 112518. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2023.112518. URL: https://www.sciencedirect.com/science/article/pii/S0021999123006137.

[83] John A. Parker. *GitHub - johnaparker/fplanck: Numerically solve the Fokker-Planck equation in N dimensions*. 2019. URL: https://github.com/johnaparker/fplanck.

[84] Ofir Pele and Michael Werman. “Fast and robust Earth Mover’s Distances”. In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 460–467. DOI: 10.1109/ICCV.2009.5459199.

[85] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Version 20121115. Nov. 2012. URL: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html.

[86] Gabriel Peyré. “Entropic Approximation of Wasserstein Gradient Flows”. In: *SIAM Journal on Imaging Sciences* 8.4 (2015), pp. 2323–2351. DOI: 10.1137/15M1010087. eprint: https://doi.org/10.1137/15M1010087. URL: https://doi.org/10.1137/15M1010087.

[87] Filip Rindler. *Calculus of Variations*. Springer, June 2018. ISBN: 9783319776378.

[88] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. “The Earth Mover’s Distance as a Metric for Image Retrieval”. In: *International Journal of Computer Vision* 40.2 (2000), pp. 99–121. DOI: https://doi.org/10.1023/a:1026543900054.

[89] Filippo Santambrogio. *A Course in the Calculus of Variations*. Springer Nature, Jan. 2024. ISBN: 9783031450365.

[90] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, Oct. 2015. ISBN: 9783319208282.

[91] Simo Sarkka. “On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems”. In: *IEEE Transactions on Automatic Control* 52.9 (Sept. 2007), pp. 1631–1641. DOI: https://doi.org/10.1109/tac.2007.904453.

[92] Simo Särkkä and Lennart Svensson. *Bayesian Filtering and Smoothing*. 2nd ed. Cambridge: Cambridge University Press, 2023. ISBN: 9781108926645.

[93] Matthias Seeger. “Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. MIT Press, 1999. URL: https://proceedings.neurips.cc/paper_files/paper/1999/file/404dcc91b2aeaa7caa47487d1483e48a-Paper.pdf.

[94] Sameer Shirdhonkar and David W. Jacobs. “Approximate earth mover’s distance in linear time”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587662.

[95] Nian Si et al. “Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 21260–21270. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/f3507289cfdc8c9ae93f4098111a13f9-Paper.pdf.

[96] Michael Stone and Paul M Goldbart. *Mathematics for physics : a guided tour for graduate students*. Cambridge ; New York: Cambridge University Press, 2010. ISBN: 9780521854030.

[97] Leonard Susskind and George Hrabovsky. *Classical mechanics : the theoretical minimum*. London: Penguin Books, 2014. ISBN: 9780141976228.

[98] Asuka Takatsu. “Wasserstein geometry of Gaussian measures”. In: *Osaka Journal of Mathematics* 48.4 (2011), pp. 1005–1026.

[99] M. Talagrand. “Transportation Cost for Gaussian and Other Product Measures.” In: *Geometric and functional analysis* 6.3 (1996), pp. 587–600. URL: http://eudml.org/doc/58238.

[100] John Robert Taylor. *Classical mechanics*. Sausalito, Calif.: University Science Books, Cop, 2005. ISBN: 9781891389221.

[101] Simon Urbanek and Yossi Rubner. *emdist: Earth Mover’s Distance*. 2023. DOI: https://cran.r-project.org/package=emdist. URL: https://doi.org/10.32614/CRAN.package.emdist.

[102] Raviteja Vemulapalli and David W. Jacobs. “Riemannian Metric Learning for Symmetric Positive Definite Matrices”. In: *CoRR* abs/1501.02393 (2015). arXiv: 1501.02393. URL: http://arxiv.org/abs/1501.02393.

[103] Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, Oct. 2008. ISBN: 9783540710509.

[104] Cedric Villani. *Topics in Optimal Transportation*. Chapter 8. S.L.: Amer Mathematical Society, 2003. ISBN: 9781470467265.

[105] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[106] Yifei Wang et al. “Optimal Neural Network Approximation of Wasserstein Gradient Direction via Convex Optimization”. In: *SIAM Journal on Mathematics of Data Science* 6.4 (2024), pp. 978–999. DOI: 10.1137/23M1573173. eprint: https://doi.org/10.1137/23M1573173. URL: https://doi.org/10.1137/23M1573173.

[107] Andre Wibisono, Varun Jog, and Po-Ling Loh. “Information and estimation in Fokker-Planck channels”. In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017, pp. 2673–2677. DOI: 10.1109/ISIT.2017.8007014.

[108] Manuel Wüthrich et al. “A New Perspective and Extension of the Gaussian Filter”. In: *The International Journal of Robotics Research* 35.14 (Dec. 2016), pp. 1731–1749.

[109] Chen Xu, Xiuyuan Cheng, and Yao Xie. “Normalizing flow neural networks by JKO scheme”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 47379–47405. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/93fce71def4e3cf418918805455d436f-Paper-Conference.pdf.

[110] Rentian Yao, Xiaohui Chen, and Yun Yang. “Wasserstein Proximal Coordinate Gradient Algorithms”. In: *Journal of Machine Learning Research* 25 (May 2024), pp. 1–66. DOI: https://doi.org/10.48550/arxiv.2405.04628.

[111] Mingxuan Yi and Song Liu. “Bridging the Gap between Variational Inference and Wasserstein Gradient Flows”. In: *International Conference on Learning Representations* 13 (2025). DOI: https://doi.org/10.48550/arxiv.2310.20090.

[112] Hiroaki Yoshida. “A Dissipation of Relative Entropy by Diffusion Flows”. In: *Entropy* 19.1 (2017). ISSN: 1099-4300. DOI: 10.3390/e19010009. URL: https://www.mdpi.com/1099-4300/19/1/9.

[113] Wei Zhang et al. “Open-Set Signal Recognition Based on Transformer and Wasserstein Distance”. In: *Applied Sciences* 13.4 (2023). ISSN: 2076-3417. DOI: 10.3390/app13042151. URL: https://www.mdpi.com/2076-3417/13/4/2151.

A

Further Bound Attempt

In this appendix, a counterexample for the alternative upper bound for $W_2(p_t^L, p_t^J)$ hypothesised at the end of Subsection 4.1.2 is provided. Let us begin with the constant c_t as used in the bound u_t^2, obtained in Conjecture 4.3. For u_t^2, the coefficient c_t is defined as being equal to the following upper bound ((4.27) in Subsection 4.1.2):

$$c_t \leq \frac{\sqrt{I(p_t^J|\pi)}}{W_2(p_t^L, \tilde{p}_t^J)} \tag{A.1}$$

The objective is to find an alternative expression for c_t where the denominator does not depend on p_t^J. One conjectured solution which would realise the objective above is that proposed in Subsection 4.1.2. We now provide a more formal definition of this conjecture and show why it cannot be true.

Conjecture A.1. *Let $\{p_t^L\}_{t\geq 0} \subset \mathrm{BW}(\mathbb{R}^d)$ be a Bures-JKO evolution and $\{p_t^J\}_{t\geq 0} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a full JKO evolution, both defined using the same potential function $V(x)$ and target $\pi : \pi \propto e^{-V}, \pi \in \mathcal{P}_2(\mathbb{R}^d)$. Furthermore, let $p_0^L = p_0^J$. Then for all $t \geq 0$ we have*

$$I(p_t^J|\pi) \leq I(p_t^L|\pi) \tag{A.2}$$

Informally, the relative Fisher information measures the average magnitude of the linear approximation of the log-ratio between its two arguments (with. If the two arguments of $I(\cdot|\cdot)$ become less similar, then the differences in their log-values should become more pronounced in at least some parts of their common domain. Since $p_t^J \to \pi$ as $t \to \infty$ but p_t^L does not, we would therefore expect $I(p_t^J|\pi) \leq I(p_t^L|\pi)$. This reasoning provides some intuition as to why Conjecture Conjecture A.1 might be true. Furthermore, we may rewrite $I(p_t|\pi) = ||\nabla_{W_2}\mathrm{KL}||_{L^2(p_t)}$; for p_t^J, we expect the "size" of its steepest-descent vector to tend to zero, whereas this will not be the case for a new full-JKO flow p_t^* starting at p_t^L since this new flow will always have a non-trivial distance to travel. As we shall see, however, it is possible to find a contradiction within Conjecture A.1 which means it cannot be true without further qualifying assumptions.

Let us now consider de Bruijn's Identity, which appears in many slightly different forms across research literature [1] but which for the purposes of this text may be stated as follows:

Identity A.2 (de Bruijn's Identity)**.** *Let $\{p_t\}_{t\geq 0} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a JKO evolution following the FPE (2.9) for a potential function V and a target $\pi \propto e^{-V}$. Then:*

$$\frac{d}{dt}\mathrm{KL}(p_t||\pi) = -I(p_t|\pi) \tag{A.3}$$

[1] For examples, see: [107], Theorem 6; [49], Eq. (4); [35], Thm. 7.1; [68], Corollary 5; and[24], Section 8.3.2; — see this last citation for a derivation of the specific form of de Bruijn's Identity presented here.

Note that Identity A.2 applies to both $I(p_t^J|\pi)$ and $I(p_t^L|\pi)$ [2], so Conjecture A.1 is true if and only if

$$\frac{d}{dt}\mathrm{KL}(p_t^J||\pi) \geq \frac{d}{dt}\mathrm{KL}(p_t^L||\pi) \tag{A.4}$$

However: if $\mathrm{KL}(p_0^J||\pi) = \mathrm{KL}(p_0^L||\pi)$ and $\mathrm{KL}(p_t^J||\pi) \to 0$ as $t \to \infty$ while $\mathrm{KL}(p_t^L||\pi)$ tends to some strictly positive value, then we would actually expect to see $\frac{d}{dt}\mathrm{KL}(p_t^J||\pi) \leq \frac{d}{dt}\mathrm{KL}(p_t^L||\pi)$ for at least one time period due to the Mean Value Theorem from univariate calculus. We have therefore arrived at a contradiction, meaning Conjecture A.1 cannot be true for all $t \geq 0$. Hence, we are unable to use $I(p_t^L|\pi)$ to bound/define c_t and thus generate a new bound for $W_2(p_t^L, p_t^J)$. Future research may seek to refine this conjecture by specifying more stringent criteria which ensure that $I(p_t^J|\pi) \leq I(p_t^L|\pi)$ (which would in turn restrict the valid use cases for the resulting W_2 bound), or by finding another replacement for $\sqrt{I(p_t^J|\pi)}$ in c_t altogether.

[2] This is possible by assuming that a new "full" JKO flow heading towards π may be started at each $t \geq 0$.

B

Supplementary Identities

Throughout this work, many minor results are used whose proofs are straightforward but can distract from a larger argument. To improve the clarity of the main text, these proofs have been collected here.

Most of the identities here are only used in a specific context: their presentation here mimics that of their origin, so descriptions of terminology, etc. are kept to a minimum. Instead, readers will be pointed to where the identity has been used, from where they can equip themselves with the necessary explanations for notation, conditions, et cetera. Identities are provided here in the approximate order they appear in the main body.

B.1. Identities for Chapter 3

Identity B.1. *The entropy of a Gaussian is invariant to translation. What this means in the context of Theorem 3.1 (specifically on p.22) is that, for a Gaussian probability density $\varrho \sim N(m, \Sigma)$:*

$$\nabla_m H(\varrho(x)) = 0, \tag{B.1}$$

where $H(\varrho) := \mathbb{E}_\varrho(\log \varrho)$ is the (negative) entropy.

Proof. We may expand $\nabla_m H(\varrho(x))$ as follows:

$$\begin{aligned} \nabla_m H(\varrho(x)) &= \nabla_m \mathbb{E}\left(\log\left((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\right)\right) \\ &= \nabla_m \mathbb{E}\left(\log\left((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\right)\right) - \frac{1}{2}\nabla_m \mathbb{E}\left((x-m)^T\Sigma^{-1}(x-m)\right) \end{aligned} \tag{B.2}$$

Using the cyclic property of the trace, we can rewrite:

$$\begin{aligned} (x-m)^T\Sigma^{-1}(x-m) &= \operatorname{tr}\left((x-m)^T\Sigma^{-1}(x-m)\right) \\ &= \operatorname{tr}\left((x-m)\otimes(x-m)^T\Sigma^{-1}\right) \end{aligned} \tag{B.3}$$

Since the trace and expectation are both linear operators, we can exchange them to obtain:

$$\begin{aligned}\mathbb{E}\left((x-m)^T\Sigma^{-1}(x-m)\right) &= \mathbb{E}\left(\mathrm{tr}\left((x-m)\otimes(x-m)^T\Sigma^{-1}\right)\right)\\ &= \mathrm{tr}\left(\mathbb{E}\left((x-m)\otimes(x-m)^T\Sigma^{-1}\right)\right)\\ &= \mathrm{tr}\left(\mathbb{E}\left((x-m)\otimes(x-m)^T\right)\Sigma^{-1}\right)\\ &= \mathrm{tr}\left(\Sigma\Sigma^{-1}\right)\\ &= \mathrm{tr}(I)\\ &= d\end{aligned} \tag{B.4}$$

Since both terms in the last line of (B.2) contain gradients of constants which do not depend on m, we therefore have $\nabla_m H(\varrho(x)) = 0$. □

Identity B.2. *In the context of Theorem 3.1 (specifically on p.22), we have for a Gaussian density ϱ:*

$$\nabla_m \varrho(x) = -\nabla_x \varrho(x) \tag{B.5}$$

Proof. We start by expanding the LHS of (B.5):

$$\nabla_m \varrho(x) = \nabla_m \left((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\right) \tag{B.6}$$

We can apply the chain rule (for vector calculus) to the term inside the exponential function to obtain:

$$\nabla_m \varrho(x) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\nabla_m\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right) \tag{B.7}$$

The terms to the left of ∇_m are equivalent to $\varrho(x)$. For the remaining terms, we can apply the identity $\frac{\partial}{\partial \mathbf{v}}(\mathbf{v^T A v}) = (\mathbf{A}+\mathbf{A}^T)\mathbf{v}$ ([85], eq. 86) along with another application of the chain rule to obtain:

$$\nabla_m \varrho(x) = \varrho(x)\left(-\frac{1}{2}(\Sigma^{-1}+(\Sigma^{-1})^T)(x-m)(-1)\right) \tag{B.8}$$

Meanwhile, let us note that $-\nabla_x\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right) = \frac{1}{2}(\Sigma^{-1}+(\Sigma^{-1})^T)(x-m)$. We thus have:

$$\nabla_m \varrho(x) = \varrho(x)\left(-\nabla_x\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\right). \tag{B.9}$$

Furthermore, it is clear from the product rule for scalar-by-vector derivatives that

$$\nabla_x \varrho(x) = \varrho(x)\nabla(x)\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right). \tag{B.10}$$

We therefore obtain:

$$\nabla_m \varrho(x) = -\nabla_x \varrho(x) \tag{B.11}$$

□

Identity B.3. *For a Gaussian density $\varrho = N(m,\Sigma)$, we have*

$$\nabla_\Sigma \varrho(x) = \frac{1}{2}\nabla_x^2 \varrho(x) \tag{B.12}$$

Proof. For convenience, let us first define the shorthand $\exp(...) := \exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)$. We can thus write:

$$\begin{aligned}\nabla_\Sigma \varrho(x) &= \nabla_\Sigma \left((2\pi)^{-\frac{d}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\right)\\ &= (2\pi)^{-\frac{d}{2}}\nabla_\Sigma\left((\det\Sigma)^{-\frac{1}{2}}\exp(...)\right)\end{aligned} \tag{B.13}$$

Both $(\det\Sigma)^{-\frac{1}{2}}$ and $\exp(...)$ are scalar functions of Σ; we shall thus apply the product rule and chain rule for scalar-by-matrix derivatives ([26], Equations C.53 and C.55) to obtain:

$$\begin{aligned}\nabla_\Sigma \varrho(x) &= (2\pi)^{-\frac{d}{2}}\left(\left(\nabla_\Sigma(\det\Sigma)^{-\frac{1}{2}}\right)\exp(...) + (\det\Sigma)^{-\frac{1}{2}}\left(\nabla_\Sigma\exp(...)\right)\right)\\ &= (2\pi)^{-\frac{d}{2}}\left(-\frac{1}{2}(\det\Sigma)^{-\frac{3}{2}}\left(\nabla_\Sigma\det\Sigma\right)\exp(...) + (\det\Sigma)^{-\frac{1}{2}}\exp(...)\left(-\frac{1}{2}\nabla_\Sigma(x-m)^T\Sigma^{-1}(x-m)\right)\right)\end{aligned} \tag{B.14}$$

By Jacobi's Formula, we have $\nabla_\Sigma\det\Sigma = \mathrm{adj}(\Sigma)^T = \mathrm{adj}(\Sigma^T) = \mathrm{adj}(\Sigma)$; furthermore, since Σ is positive-definite (and thus invertible), we have $\mathrm{adj}(\Sigma) = (\det\Sigma)\Sigma^{-1}$. For the second term on the RHS of (B.14), note that $(x-m)^T\Sigma^{-1}(x-m) = \mathrm{tr}((x-m)^T\Sigma^{-1}(x-m)) = \mathrm{tr}(\Sigma^{-1}(x-m)(x-m)^T)$; by [85] eq. (124) (setting $B \leftarrow (x-m)(x-m)^T$ and noting its symmetry), we have $\nabla_\Sigma\mathrm{tr}(\Sigma^{-1}(x-m)(x-m)^T) = -\Sigma^{-1}(x-m)(x-m)^T\Sigma^{-1}$. Combining these results yields:

$$\begin{aligned}\nabla_\Sigma\varrho(x) &= \left(-\frac{1}{2}\right)(2\pi)^{-\frac{d}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp(...)\left(-\Sigma^{-1}\left((x-m)(x-m)^T\right)\Sigma^{-1} + \Sigma^{-1}\right)\\ &= \frac{1}{2}p(x)\left(\Sigma^{-1}\left((x-m)(x-m)^T\right)\Sigma^{-1} - \Sigma^{-1}\right)\end{aligned} \tag{B.15}$$

Let us now consider the Hessian matrix $\nabla_x^2 p(x)$, which can be rewritten using (B.10) and eq. (85) from [85] to obtain:

$$\begin{aligned}\nabla_x^2 p(x) &= \nabla_x(\nabla_x p(x))\\ &= \nabla_x\left(p(x)\left(\nabla_x\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)\right)\right)\\ &= -\nabla_x\left(p(x)\left(\Sigma^{-1}(x-m)\right)\right)\end{aligned} \tag{B.16}$$

For a scalar-valued function $c(x)$ and a vector-valued function $b(x)$, we have the identity $\nabla_x c(x)b(x) = \nabla_x c(x)\otimes b(x) + c(x)\nabla_x b(x)$ (which can be checked by taking component-wise derivatives). Applying this result to (B.16) yields:

$$\begin{aligned}\nabla_x^2 p(x) &= -p(x)\left(-\Sigma^{-1}(x-m)\right)\left(\Sigma^{-1}(x-m)\right)^T - p(x)(\Sigma^{-1})\\ &= -p(x)\left(-\Sigma^{-1}(x-m)(x-m)^T\Sigma^{-1}\right) - p(x)\Sigma^{-1}\\ &= p(x)\left(\Sigma^{-1}\left((x-m)(x-m)^T\right)\Sigma^{-1} - \Sigma^{-1}\right)\\ &= 2\nabla_\Sigma p(x)\end{aligned} \tag{B.17}$$

□

Identity B.4. *For a Gaussian density $p = N(m, \Sigma)$ and the negative entropy $H(p) = \mathbb{E}(\log p(x))$, we have*

$$\nabla_\Sigma H(p) = -\frac{1}{2}\Sigma^{-1}. \tag{B.18}$$

Proof. We can rewrite (B.18) as

$$\nabla_\Sigma H(p) = \nabla_\Sigma \left(\log\left((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\right) - \frac{1}{2}\mathbb{E}\left((x-m)^T\Sigma^{-1}(x-m)\right)\right). \tag{B.19}$$

Using the trace trick (B.3), we can rewrite $\mathbb{E}\left((x-m)^T\Sigma^{-1}(x-m)\right) = \mathbb{E}\left((x-m)(x-m)^T\right)\Sigma^{-1} = I$. Furthermore, from Equations (B.14) and (B.15) we know that $\nabla_\Sigma(\det\Sigma)^{-1/2} = -\frac{1}{2}(\det\Sigma)^{-1/2}\Sigma^{-1}$. Therefore, applying the chain rule for scalar-by-matrix derivatives ([26], Eq. C.55) to the term $\nabla_\Sigma \log(\det\Sigma)^{-1/2}$ we obtain

$$\begin{aligned}\nabla_\Sigma H(p) &= 0 + \frac{1}{(\det\Sigma)^{-1/2}}\frac{-1}{2}(\det\Sigma)^{-1/2}\Sigma^{-1} + 0 \\ &= -\frac{1}{2}\Sigma^{-1}\end{aligned} \tag{B.20}$$

□

Identity B.5. *In the context of Theorem 3.1 (specifically on p.23), we have*

$$\nabla_\Sigma \mathbb{E}_p\left(\log\pi(x)\right) = \frac{1}{2}\mathbb{E}_p\left(\nabla_x^2 \log\pi(x)\right) \tag{B.21}$$

Proof. Using the Leibniz integral rule, we can write

$$\begin{aligned}\nabla_\Sigma \mathbb{E}_p\left(\log\pi(x)\right) &= \nabla_\Sigma \int_{\mathbb{R}^d} \log\pi(x)p(x)dx \\ &= \int_{\mathbb{R}^d} \log\pi(x)\nabla_\Sigma p(x)dx\end{aligned} \tag{B.22}$$

Applying Identity B.3, we obtain

$$\nabla_\Sigma \mathbb{E}_p\left(\log\pi(x)\right) = \frac{1}{2}\int_{\mathbb{R}^d} \log\pi(x)\nabla_x^2 p(x)dx \tag{B.23}$$

In [65], Lambert et al. claim a double application of "integration by parts" without providing further details. In practice, what is required here is component-wise integration by parts, applied in the same manner as with (3.9) in Subsection 3.2.2. This must be performed twice, with the first yielding integrals of the form $-\int_{\mathbb{R}}\left(\frac{\partial}{\partial x_i}p(x)\right)\left(\frac{\partial}{\partial x_i}(\log\pi(x))\right)dx_i$ and the second yielding $-\int_{\mathbb{R}} p(x)\left(\frac{\partial^2}{\partial x_i^2}(\log\pi(x))\right)dx_i$; in both cases, the terms extracted from these integrals vanish for the same reasons provided in Subsection 3.2.2 (the relative growth rates of $p(x)$ and $V(x)$). We thus obtain:

$$\begin{aligned}\nabla_\Sigma \mathbb{E}_p\left(\log\pi(x)\right) &= \frac{1}{2}\int_{\mathbb{R}^d}\left(\nabla_x^2\log\pi(x)\right)p(x)dx \\ &= \frac{1}{2}\mathbb{E}_p\left(\nabla_x^2\log\pi(x)\right)\end{aligned} \tag{B.24}$$

□

Identity B.6. *In the context of Theorem 3.1 (specifically on p.25), we have for* $p = N(m, \Sigma)$*:*

$$\mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\Sigma = \mathbb{E}_p\left(\nabla_x \log \pi(x) \otimes (x-m)\right). \tag{B.25}$$

Proof. Let us first write out expressions for $\nabla_x \log \pi(x)$ and $\nabla_x^2 \log \pi(x)$. Using the notation $\pi_i'(x) := \frac{\partial}{\partial x_i}\pi(x)$ and $\pi_{ij}''(x) := \frac{\partial^2}{\partial x_i \partial x_j}\pi(x)$ for clarity, we obtain:

$$\begin{aligned}\nabla_x \log \pi(x) &= \begin{pmatrix} \frac{\partial}{\partial x_1}\log \pi(x) \\ \vdots \\ \frac{\partial}{\partial x_d}\log \pi(x)\end{pmatrix} \\ &= \frac{1}{\pi(x)}\begin{pmatrix}\pi_1'(x) \\ \vdots \\ \pi_d'(x)\end{pmatrix}\end{aligned} \tag{B.26}$$

$$\begin{aligned}\nabla_x^2 \log \pi(x) &= \begin{pmatrix} \frac{\partial}{\partial x_1}\frac{\pi_1'(x)}{\pi(x)} & \cdots & \frac{\partial}{\partial x_1}\frac{\pi_d'(x)}{\pi(x)} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_d}\frac{\pi_1'(x)}{\pi(x)} & \cdots & \frac{\partial}{\partial x_d}\frac{\pi_d'(x)}{\pi(x)}\end{pmatrix} \\ &= \frac{1}{\pi(x)^2}\begin{pmatrix} \pi_{11}''(x)\pi(x) - \pi_1'(x)^2 & \cdots & \pi_{1d}''(x)\pi(x) - \pi_1'(x)\pi_d'(x) \\ \vdots & \ddots & \vdots \\ \pi_{1d}''(x)\pi(x) - \pi_1'(x)\pi_d'(x) & \cdots & \pi_{dd}''(x)\pi(x) - \pi_d'(x)^2\end{pmatrix}\end{aligned} \tag{B.27}$$

For all $i : 1 \leq i \leq d$, let us define the function $g_i(x) := \left(\nabla_x \log \pi(x)\right)_i$. Then:

$$\nabla_x g_i(x) = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_i}\log \pi(x) \\ \vdots \\ \frac{\partial^2}{\partial x_d \partial x_i}\log \pi(x)\end{pmatrix} \tag{B.28}$$

By the multivariate version of Stein's Lemma (referred to by Lambert et al. as "Gaussian integration by parts"), we have:

$$\mathbb{E}_p\left(g_i(x)(x-m)\right) = \Sigma \mathbb{E}_p\left(\nabla_x g_i(x)\right) \tag{B.29}$$

By setting $g(x) := \nabla_x \log \pi(x)$ and applying (B.29) across each row of the outer product $g(x) \otimes (x-m)$, we obtain the results:

$$\begin{aligned}\mathbb{E}_p\left(\nabla_x \log \pi(x) \otimes (x-m)\right) &= \left(\Sigma \mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\right)^T \\ &= \mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right)\Sigma\end{aligned} \tag{B.30}$$

By applying a transpose operation to (B.30), we obtain:

$$\mathbb{E}_p\left((x-m) \otimes \nabla_x \log \pi(x)\right) = \Sigma \mathbb{E}_p\left(\nabla_x^2 \log \pi(x)\right) \tag{B.31}$$

□

B.2. Identities for Chapter 4

Identity B.7. *For a Bures-JKO flow $\{p_t^L\}_{t\geq 0}$ and its associated velocity fields $\{v_t^L\}_{t\geq 0}$, we have*

$$||v_t^L||^2_{L^2(p_t^L)} = |a_t|^2 + \mathrm{tr}(S_t\Sigma_t^L S_t), \tag{B.32}$$

where $a_t = -\mathbb{E}_{p_t^L}\nabla V$ and $S_t = \mathbb{E}_{p_t^L}\nabla_x^2 V - (\Sigma_t^L)^{-1}$.

Proof. From (3.46), we know that $v_t^L(x) = a_t + S_t(x - m_t^L)$. Using $|v|^2 = v\cdot v$ to denote the (squared) Euclidean norm, we may write

$$\begin{aligned} |v_t^L(x)|^2 &= \big(a_t + S_t(x-m_t^L)\big)\cdot\big(a_t + S_t(x-m_t^L)\big) \\ &= |a_t|^2 + 2a_t\cdot S_t(x-m_t^L) + |S_t(x-m_t^L)|^2 \end{aligned} \tag{B.33}$$

The term $|S_T(x-m_t^L)|^2$ may be further rearranged as follows:

$$\begin{aligned} |S_t(x-m_t^L)|^2 &= S_t(x-m_t^L)\cdot S_t(x-m_t^L) \\ &= \big(S_t(x-m_t^L)\big)^T S_t(x-m_t^L) \\ &= (x-m_t^L)^T S_t^T S_t(x-m_t^L) \\ &= \mathrm{tr}\left((x-m_t^L)^T S_t S_t(x-m_t^L)\right) \\ &= \mathrm{tr}\left(S_t S_t(x-m_t^L)\otimes(x-m_t^L)^T\right) \end{aligned} \tag{B.34}$$

Note that $S_t^T = S_t$, since $S_t \in S^d$ (see Section 3.3 for details). Note also that $\mathbb{E}_{p_t^L}(a_t) = a_t$ and $\mathbb{E}_{p_t^L}(S_t) = S_t$, since a_t, S_t are themselves expectations of x (plus a constant term in the case of S_t). Returning to $||v_t^L||_{L^2(p_t^L)}$:

$$\begin{aligned} ||v_t^L||^2_{L^2(p_t^L)} &= \mathbb{E}_{p_t^L}\left(|v_t^L|^2\right) \\ &= \mathbb{E}_{p_t^L}\left(|a_t|^2 + 2a_t\cdot S_t(x-m_t^L) + \mathrm{tr}\left(S_tS_t(x-m_t^L)\otimes(x-m_t^L)^T\right)\right) \\ &= \mathbb{E}_{p_t^L}\left(|a_t|^2\right) + \mathbb{E}_{p_t^L}\left(2a_t\cdot S_t(x-m_t^L)\right) + \mathbb{E}_{p_t^L}\left(\mathrm{tr}\left(S_t^TS_t(x-m_t^L)\otimes(x-m_t^L)^T\right)\right) \\ &= |a_t|^2 + 2a_t\cdot S_t\mathbb{E}_{p_t^L}\left(x-m_t^L\right) + \mathrm{tr}\left(S_tS_t\mathbb{E}_{p_t^L}\left((x-m_t^L)\otimes(x-m_t^L)^T\right)\right) \\ &= |a_t|^2 + 2a_t\cdot 0 + \mathrm{tr}\left(S_tS_t\Sigma_t^L\right) \\ &= |a_t|^2 + \mathrm{tr}\left(S_t\Sigma_t^L S_t\right) \end{aligned} \tag{B.35}$$

□

Identity B.8. *In the context of Theorem 4.4, we observe the following:*

$$||v_t^L - v_t^*||^2_{L^2(p_t^L)} = \mathrm{tr}\left(\tilde{S}_t\Sigma_t^L\tilde{S}_t\right) + \mathbb{E}_{p_t^L}|\nabla V|^2 - |a_t|^2 \tag{B.36}$$

Proof. We follow a similar procedure to the proof of Identity B.7. Beginning with:

$$\begin{aligned} v_t^L - v_t^* &= a_t + S_t(x-m_t^L) - \nabla\log\frac{p_t^L(x)}{\pi(x)} \\ &= a_t + S_t(x-m_t^L) - \nabla_x(x-m_t^L)^T(\Sigma_t^L)^{-1}(x-m_t^L) - \nabla V \end{aligned} \tag{B.37}$$

Using $\nabla_x(x-m_t^L)^T(\Sigma_t^L)^{-1}(x-m_t^L) = 2(\Sigma_t^L)^{-1}(x-m_t^L)$ (Eq. (85) in [85]) and the definition $\tilde{S}_t := \mathbb{E}_{p_t^L}\nabla^2 V - 3(\Sigma_t^L)^{-1}$, we may write

$$\begin{aligned} v_t^L - v_t^* &= a_t + S_t(x-m_t^L) - 2(\Sigma_t^L)^{-1}(x-m_t^L) - \nabla V \\ &= a_t + \tilde{S}_t(x-m_t^L) - \nabla V \end{aligned} \tag{B.38}$$

Let us now consider $|v_t^L - v_t^*|^2 = (v_t^L - v_t^*)\cdot(v_t^L - v_t^*)$. By the bilinearity of the dot product over $\mathbb{R}^d$, we obtain

$$|v_t^L - v_t^*|^2 = |a_t|^2 + |\tilde{S}_t(x-m_t^L)|^2 + |\nabla V|^2 + 2a_t\cdot\tilde{S}_t(x-m_t^L) - 2a_t\cdot\nabla V - 2\tilde{S}_t(x-m_t^L)\cdot\nabla V \tag{B.39}$$

We are now ready to take the expectation $||v_t^L - v_t^*||^2_{L^2(p_t^L)} = \int |v_t^L - v_t^*|^2 dp_t^L$ using the separated expression (B.39). $|a_t|^2$ does not depend on x, as $a_t = \mathbb{E}_{p_t^L}\nabla V(x)$ is itself an expectation which uses p_t^L. Hence, $\mathbb{E}_{p_t^L}|a_t|^2 = |a_t|^2$. For $|\tilde{S}_t(x-m_t^L)|^2$, note that we may repeat the procedure used in (B.34) to obtain $|\tilde{S}_t(x-m_t^L)|^2 = \mathrm{tr}\left(\tilde{S}_t^2(x-m_t^L)\otimes(x-m_t^L)^T\right)$ and $\mathbb{E}_{p_t^L}|\tilde{S}_t(x-m_t^L)|^2 = \mathrm{tr}\left(\tilde{S}_t\Sigma_t^L\tilde{S}_t\right)$. When $\mathbb{E}_{p_t^L}$ is applied, the two dot product terms in (B.39) involving a single $\tilde{S}_t(x-m_t^L)$ vector will vanish, as $\tilde{S}_t$ is also comprised of terms which do not depend on x. Combining these observations yields

$$\mathbb{E}_{p_t^L}|v_t^L - v_t^*|^2 = |a_t|^2 + \mathrm{tr}\left(\tilde{S}_t\Sigma_t^L\tilde{S}_t\right) + \mathbb{E}_{p_t^L}|\nabla V|^2 - 2\mathbb{E}_{p_t^L}\left(a_t\cdot\nabla V\right) \tag{B.40}$$

Since, by the linearity of expectation, we have $\mathbb{E}_{p_t^L}\left(a_t\cdot\nabla V\right) = a_t\cdot\mathbb{E}_{p_t^L}\nabla V = a_t\cdot a_t$, we thus have

$$\mathbb{E}_{p_t^L}|v_t^L - v_t^*|^2 = \mathrm{tr}\left(\tilde{S}_t\Sigma_t^L\tilde{S}_t\right) + \mathbb{E}_{p_t^L}|\nabla V|^2 - |a_t|^2 \tag{B.41}$$

□

Identity B.9. *In the context of Conjecture 4.2: for a JKO flow $\{p_t^J\}_{t\geq 0} \in \mathcal{P}_2(\mathbb{R}^d)$ and its associated velocity fields $\{v_t^J\}_{t\geq 0} : v_t^J = -\nabla\log\frac{p_t^J}{\pi}$, we have*

$$\langle\nabla_2, v_t^J\rangle_t^J \leq W_2(p_t^L, p_t^J)\sqrt{I(p_t^J|\pi)} \tag{B.42}$$

‘

where $\nabla_2 = \nabla_{p_t^J}\frac{1}{2}W_2^2(p_t^L, p_t^J)$ for a fixed density p_t^L and $I(p_t^J|\pi)$ denotes the relative Fisher information between p_t^J and π.

Proof. First, by applying the Cauchy-Schwarz Inequality to the $\langle\cdot,\cdot\rangle_t^J$ inner product:

$$\begin{aligned} \langle\nabla_2, v_t^J\rangle_t^J &\leq |\langle\nabla_2, v_t^J\rangle_t^J| \\ &\leq ||\nabla_2||_{L^2(p_t^J)}||v_t^J||_{L^2(p_t^J)} \end{aligned} \tag{B.43}$$

We may rewrite $\nabla_2 = \nabla\psi$, where $\nabla\psi = x - T_t^{-1}(x)$ for the optimal W_2 transport map from p_t^L to p_t^J. For the term $||\nabla_2||_{L^2(P_t^J)}$, this yields

$$\begin{aligned} ||\nabla_2||_{L^2(P_t^J)} &= \left(\int|\nabla\psi|^2 p_t^J dx\right)^{1/2} \\ &= \left(\int|x - T_t^{-1}(x)|^2 p_t^J dx\right)^{1/2} \end{aligned} \tag{B.44}$$

By considering (MP), we see that this is nothing more than the Mongé formulation of $W_2(p_t^J, p_t^L)$. Meanwhile, let us also rewrite the term $||v_t^J||_{L^2(P_t^J)}$:

$$\begin{aligned} ||v_t^J||_{L^2(P_t^J)} &= \left(\int \left| -\nabla \log \frac{p_t^J}{\pi}(x) \right|^2 p_t^J dx \right)^{1/2} \\ &= \left(\int \left| \nabla \log \frac{p_t^J}{\pi}(x) \right|^2 p_t^J dx \right)^{1/2} \end{aligned} \tag{B.45}$$

This is precisely the square root of the relative Fisher information $I(p_t^J|\pi)$ (Eq. 8 in [79]). □

Identity B.10. *For a probability measure $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ and a Gaussian measure $\mu \in BW(\mathbb{R}^d)$:*

$$\underset{\mu \in \mathrm{BW}(\mathbb{R}^d)}{\operatorname{argmin}} \; KL(\pi||\mu) = N(m_\pi, \Sigma_\pi) \tag{B.46}$$

Proof. Let us first write out $KL(\pi||\mu)$. Note that all expectations in this proof are taken w.r.t. π.

$$\begin{aligned} KL(\pi||\mu) &= \mathbb{E} \log \pi - \mathbb{E} \log \mu \\ &= \mathbb{E} \log \pi - \mathbb{E} \left(\log \left((2\pi)^{-d/2} \right) + \log \left(|\Sigma_\mu|^{-1/2} \right) - \frac{1}{2} (x - m_\mu)^T \Sigma_\mu^{-1} (x - m_\mu) \right) \\ &= \mathbb{E} \log \pi + \frac{d}{2} \log 2\pi + \frac{1}{2} \log |\Sigma_\mu| + \frac{1}{2} \mathbb{E} \left((x - m_\mu)^T \Sigma_\mu^{-1} (x - m_\mu) \right) \end{aligned} \tag{B.47}$$

The expectation in the last term can be rearranged as follows:

$$\begin{aligned} \mathbb{E} \left((x - m_\mu)^T \Sigma_\mu^{-1} (x - m_\mu) \right) &= \mathbb{E} \left(\mathrm{tr} \left((x - m_\mu)^T \Sigma_\mu^{-1} (x - m_\mu) \right) \right) \\ &= \mathbb{E} \left(\mathrm{tr} \left((x - m_\mu)(x - m_\mu)^T \Sigma_\mu^{-1} \right) \right) \\ &= \mathrm{tr} \left(\mathbb{E} \left((x - m_\mu)(x - m_\mu)^T \Sigma_\mu^{-1} \right) \right) \\ &= \mathrm{tr} \left(\mathbb{E} \left((x - m_\mu)(x - m_\mu)^T \right) \Sigma_\mu^{-1} \right) \end{aligned} \tag{B.48}$$

Let us now examine the outer product $(x - m_\mu)(x - m_\mu)^T$ in more detail. Note that the outer product is a linear operation in both arguments, i.e. $(a-b)(c-d)^T = a(c-d)^T - b(c-d)^T = (a-b)c^T - (a-b)d^T = ac^T - bc^T - ad^T + bd^T$. Hence:

$$\begin{aligned} (x - m_\mu)(x - m_\mu)^T &= xx^T - m_\mu x^T - x m_\mu^T + m_\mu m_\mu^T \\ &= xx^T - m_\mu x^T - x m_\mu^T + m_\mu m_\mu^T + x m_\pi^T - x m_\pi^T + m_\pi (x - m_\pi)^T - m_\pi (x - m_\pi)^T \\ &= x(x - m_\pi)^T - m_\pi (x - m_\pi)^T + x(m_\pi - m_\mu)^T - m_\mu x^T + m_\mu m_\mu^T + m_\pi x^T - m_\pi m_\pi^T \\ &= (x - m_\pi)(x - m_\pi)^T - x(m_\pi - m_\mu)^T + (m_\pi - m_\mu) x^T + m_\mu m_\mu^T - m_\pi m_\pi^T \end{aligned} \tag{B.49}$$

Let us now introduce the expectation into the RHS of (B.49):

$$\begin{aligned} \mathbb{E} \left((x - m_\mu)(x - m_\mu)^T \right) &= \Sigma_\pi + m_\pi (m_\pi - m_\mu)^T + (m_\pi - m_\mu) m_\pi^T + m_\mu m_\mu^T - m_\pi m_\pi^T \\ &= \Sigma_\pi + m_\pi m_\pi^T - m_\pi m_\mu^T - m_\mu m_\pi^T + m_\mu m_\mu^T \end{aligned} \tag{B.50}$$

Reintroducing Σ_μ^{-1} and the trace operator to the above, we obtain:

$$\begin{aligned}\mathbb{E}\left((x-m_\mu)^T\Sigma_\mu^{-1}(x-m_\mu)\right) &= \operatorname{tr}\left(\left(\Sigma_\pi + m_\pi m_\pi^T - m_\pi m_\mu^T - m_\mu m_\pi^T + m_\mu m_\mu^T\right)\Sigma_\mu^{-1}\right)\\ &= \operatorname{tr}\left(\Sigma_\pi\Sigma_\mu^{-1}\right) + \operatorname{tr}\left(m_\pi m_\pi^T\Sigma_\mu^{-1}\right) - \operatorname{tr}\left(m_\pi m_\mu^T\Sigma_\mu^{-1}\right) - \operatorname{tr}\left(m_\mu m_\pi^T\Sigma_\mu^{-1}\right) + \operatorname{tr}\left(m_\mu m_\mu^T\Sigma_\mu^{-1}\right)\\ &= \operatorname{tr}\left(\Sigma_\pi\Sigma_\mu^{-1}\right) + \operatorname{tr}\left(m_\pi^T\Sigma_\mu^{-1}m_\pi\right) - \operatorname{tr}\left(m_\mu^T\Sigma_\mu^{-1}m_\pi\right) - \operatorname{tr}\left(m_\pi^T\Sigma_\mu^{-1}m_\mu\right) + \operatorname{tr}\left(m_\mu^T\Sigma_\mu^{-1}m_\mu\right)\\ &= \operatorname{tr}\left(\Sigma_\pi\Sigma_\mu^{-1}\right) + m_\pi^T\Sigma_\mu^{-1}m_\pi - m_\mu^T\Sigma_\mu^{-1}m_\pi - m_\pi^T\Sigma_\mu^{-1}m_\mu + m_\mu^T\Sigma_\mu^{-1}m_\mu\end{aligned} \tag{B.51}$$

By the symmetry of Σ_μ, we have that Σ_μ^{-1} is also symmetric and:

$$\mathbb{E}\left((x-m_\mu)^T\Sigma_\mu^{-1}(x-m_\mu)\right) = \operatorname{tr}\left(\Sigma_\pi\Sigma_\mu^{-1}\right) + m_\pi^T\Sigma_\mu^{-1}m_\pi - 2m_\mu^T\Sigma_\mu^{-1}m_\pi + m_\mu^T\Sigma_\mu^{-1}m_\mu \tag{B.52}$$

We can therefore write $KL(\pi||\mu)$ as:

$$KL(\pi||\mu) = \mathbb{E}\log\pi + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_\mu| + \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_\pi\Sigma_\mu^{-1}\right) + m_\pi^T\Sigma_\mu^{-1}m_\pi - 2m_\mu^T\Sigma_\mu^{-1}m_\pi + m_\mu^T\Sigma_\mu^{-1}m_\mu\right) \tag{B.53}$$

We can find a minimiser for $KL(\pi||\mu)$ using the standard first-order condition for the two parameters m_μ and Σ_μ. First, we take the derivative w.r.t. m_μ:

$$\frac{d}{dm_\mu}KL(\pi||\mu) = -\Sigma_\mu^{-1}m_\pi + \Sigma_\mu^{-1}m_\mu \tag{B.54}$$

By setting this to zero, we see that the optimal $\tilde{m_\mu} = m_\pi$. Returning to $KL(\pi||\mu)$, we can set $m_\mu = m_\pi$ to eliminate the terms $m_\pi^T\Sigma_\mu^{-1}m_\pi - 2m_\mu^T\Sigma_\mu^{-1}m_\pi + m_\mu^T\Sigma_\mu^{-1}m_\mu$. Then, we can obtain the derivative w.r.t. Σ_μ using standard matrix calculus identities ([85], eqs. (57) and (104)):

$$\frac{d}{d\Sigma_\mu}KL(\pi||\mu) = \frac{1}{2}\Sigma_\mu^{-1} + \frac{1}{2}\left(-\Sigma_\mu^{-1}\Sigma_\pi\Sigma_\mu^{-1}\right) \tag{B.55}$$

By setting this expression to zero and multiplying by Σ_μ on either the left or the right, we obtain that $\tilde{\Sigma_\mu} = \Sigma_\pi$. □