SIAM J. MATH. ANAL.
Vol. 47, No. 6, pp. 4091–4122

KULLBACK–LEIBLER APPROXIMATION FOR PROBABILITY MEASURES ON INFINITE DIMENSIONAL SPACES*

F. J. PINSKI[†], G. SIMPSON[‡], A. M. STUART[§], AND H. WEBER[§]

Abstract. In a variety of applications it is important to extract information from a probability measure μ on an infinite dimensional space. Examples include the Bayesian approach to inverse problems and (possibly conditioned) continuous time Markov processes. It may then be of interest to find a measure ν, from within a simple class of measures, which approximates μ. This problem is studied in the case where the Kullback–Leibler divergence is employed to measure the quality of the approximation. A calculus of variations viewpoint is adopted, and the particular case where ν is chosen from the set of Gaussian measures is studied in detail. Basic existence and uniqueness theorems are established, together with properties of minimizing sequences. Furthermore, parameterization of the class of Gaussians through the mean and inverse covariance is introduced, the need for regularization is explained, and a regularized minimization is studied in detail. The calculus of variations framework resulting from this work provides the appropriate underpinning for computational algorithms.

Key words. Kullback–Leibler divergence, relative entropy, Gaussian measures

AMS subject classifications. 28C20, 49K40, 60-08, 60G15

DOI. 10.1137/140962802

1. Introduction.

This paper is concerned with the problem of minimizing the Kullback–Leibler divergence between a pair of probability measures, viewed as a problem in the calculus of variations. We are given a measure μ, specified by its Radon–Nikodym derivative with respect to a reference measure μ_0, and we find the closest element ν from a simpler set of probability measures. After an initial study of the problem in this abstract context, we specify to the situation where the reference measure μ_0 is Gaussian and the approximating set comprises Gaussians. It is necessarily the case that minimizers ν are then equivalent as measures to μ_0,[1] and we use the Feldman–Hajek theorem to characterize such ν in terms of their inverse covariance operators. This induces a natural formulation of the problem as minimization over the mean, from the Cameron–Martin space of μ_0, and over an operator from a weighted Hilbert–Schmidt space. We investigate this problem from the point of view of the calculus of variations, studying properties of minimizing sequences, regularization to improve the space in which operator convergence is obtained, and uniqueness under a slight strengthening of a log-convex assumption on the measure μ.

In the situation where the minimization is over a convex set of measures ν, the problem is classical and completely understood [10]; in particular, there is uniqueness

*Received by the editors March 31, 2014; accepted for publication (in revised form) March 23, 2015; published electronically November 5, 2015.
http://www.siam.org/journals/sima/47-6/96280.html

[†]Physics Department, University of Cincinnati, Cincinnati, OH 45221 (frank.pinski@uc.edu). This author's visit to Warwick was supported by the ERC, EPSRC, and ONR.

[‡]Department of Mathematics, Drexel University, Philadelphia, PA 19104 (simpson@math.drexel.edu). This author's research was supported by NSF PIRE grant OISE-0967140 and DOE grant DE-SC0002085. This author's visit to Warwick was supported by the ERC, EPSRC, and ONR.

[§]Mathematics Institute, Warwick University, Coventry CV4 7AL, UK (a.m.stuart@warwick.ac.uk, hendrik.weber@warwick.ac.uk). The third author's research was supported by the ERC, EPSRC, and ONR.

[1] Recall that probability measures μ_0 and ν are equivalent if they have the same sets of measure zero, or equivalently if $\mu_0 \ll \nu$ and $\nu \ll \mu_0$.

of minimizers. However, the emphasis in our work is on situations where the set of measures ν is not convex, such as the set of Gaussian measures, and in this context uniqueness cannot be expected in general. However, some of the ideas used in [10] are useful in our general developments, in particular methodologies to extract minimizing sequences converging in total variation. Furthermore, in the finite dimensional case the minimization problem at hand was studied by McCann [26] in the context of gas dynamics. He introduced the concept of "displacement convexity," which was one of the main ingredients for the recent developments in the theory of mass transportation (see, e.g., [1, 33]). Inspired by the work of McCann, we identify situations in which uniqueness of minimizers can occur even when approximating over nonconvex classes of measures. Nonetheless, we emphasize that a strong motivation for the methodology we adopt in this paper is that it allows the possibility of approximating measures which have several regions of high probability mass and that then nonuniqueness of the approximating measures is to be expected—we discuss this in the following section.

Approximation with respect to Kullback–Leibler divergence is not new and indeed forms a widely used tool in the field of machine learning [6], with motivation being the interpretation of Kullback–Leibler divergence as a measure of loss of information. Recently the methodology has been used for the coarse-graining of stochastic lattice systems [22], simple models for data assimilation [2, 3], the study of models in ocean-atmosphere science [25, 17], and molecular dynamics [21]. However, none of this applied work has studied the underlying calculus of variations problem, which is the basis for the algorithms employed. Understanding the properties of minimizing sequences is crucial for the design of good finite dimensional approximations (see, e.g., [4]), and this fact motivates the work herein. The companion paper [27] demonstrates the use of algorithms for Kullback–Leibler minimization which are informed by the analysis in this paper.

In section 2 we describe basic facts about Kullback–Leibler minimization in an abstract setting, and include an example illustrating our methodology, together with the fact that uniqueness is typically not to be expected when approximating within the Gaussian class; we also discuss briefly the infinite dimensional problems which have motivated our work. Section 3 then concentrates on the theory of minimization with respect to Gaussians. We demonstrate the existence of minimizers and then develop a regularization theory needed in the important case where the inverse covariance operator is parameterized via a Schrödinger potential. We also study the restricted class of target measures for which uniqueness can be expected, and we generalize the overall setting to the study of Gaussian mixtures. Proofs of all of our results are collected in section 4, whilst the appendices contain variants on a number of classical results which underlie those proofs.

2. General properties of Kullback–Leibler minimization.

In subsection 2.1 we present some basic background theory which underpins this paper. In subsection 2.2 we provide an explicit finite dimensional example which serves to motivate the questions we study in the remainder of the paper.

2.1. Background theory.

In this subsection we recall some general facts about Kullback-Leibler approximation on an arbitrary Polish space. Let $\mathcal{H}$ be a Polish space endowed with its Borel sigma algebra $\mathcal{F}$. Denote by $\mathcal{M}(\mathcal{H})$ the set of Borel probability measures on $\mathcal{H}$, and let $\mathcal{A} \subset \mathcal{M}(\mathcal{H})$. Our aim is to find the best approximation of a target measure $\mu \in \mathcal{M}(\mathcal{H})$ in the set $\mathcal{A}$ of "simpler" measures. As a measure for closeness we choose the Kullback–Leibler divergence, also known as the relative entropy. For any $\nu \in \mathcal{M}(\mathcal{H})$ that is absolutely continuous with respect to μ it is given by

$$D_{\mathrm{KL}}(\nu\|\mu) = \int_H \log\left(\frac{d\nu}{d\mu}(x)\right)\frac{d\nu}{d\mu}(x)\,\mu(dx) = \mathbb{E}^\mu\left[\log\left(\frac{d\nu}{d\mu}(x)\right)\frac{d\nu}{d\mu}(x)\right], \tag{2.1}$$

where we use the convention that $0\log 0 = 0$. If ν is not absolutely continuous with respect to μ, then the Kullback–Leibler divergence is defined as $+\infty$. The main aims of this paper are to discuss the properties of the minimization problem

$$\operatorname*{argmin}_{\nu\in\mathcal{A}} D_{\mathrm{KL}}(\nu\|\mu) \tag{2.2}$$

for suitable sets $\mathcal{A}$ and to create a mathematical framework appropriate for the development of algorithms to perform the minimization.

The Kullback–Leibler divergence is not symmetric in its arguments, and minimizing $D_{\mathrm{KL}}(\mu\|\nu)$ over ν for fixed μ in general gives a result different from (2.2). Indeed, if $\mathcal{H}$ is $\mathbb{R}^n$ and $\mathcal{A}$ is the set of Gaussian measures on $\mathbb{R}^n$, then minimizing $D_{\mathrm{KL}}(\mu\|\nu)$ yields for ν the Gaussian measure with the same mean and variance as μ; see [6, section 10.7]. Such an approximation is undesirable in many situations, for example if μ is bimodal; see [6, Figure 10.3]. We will demonstrate by example in subsection 2.2 that problem (2.2) is a more desirable minimization problem which can capture local properties of the measure μ such as individual modes. Note that the objective function in the minimization (2.2) can be formulated in terms of expectations only over measures from $\mathcal{A}$; if this set is simple, then this results in computationally expedient algorithms. Below we will usually choose for $\mathcal{A}$ a set of Gaussian measures, and hence these expectations are readily computable.

The following well-known result gives existence of minimizers for problem (2.2) as soon as the set $\mathcal{A}$ is closed under weak convergence of probability measures. For the reader's convenience we give a proof in Appendix A. We essentially follow the exposition in [14, Lemma 1.4.2]; see also [1, Lemma 9.4.3].

Proposition 2.1. *Let (ν_n) and (μ_n) be sequences in $\mathcal{M}(\mathcal{H})$ that converge weakly to $\nu_\star$ and $\mu_\star$. Then we have*

$$\liminf_{n\to\infty} D_{\mathrm{KL}}(\nu_n\|\mu_n) \geq D_{\mathrm{KL}}(\nu_\star\|\mu_\star).$$

Furthermore, for any $\mu \in \mathcal{M}(\mathcal{H})$ and for any $M < \infty$ the set

$$\{\nu \in \mathcal{M}(\mathcal{H})\colon D_{\mathrm{KL}}(\nu\|\mu) \leq M\}$$

is compact with respect to weak convergence of probability measures.

Proposition 2.1 yields the following immediate corollary, which, in particular, provides the existence of minimizers from within the Gaussian class.

Corollary 2.2. *Let $\mathcal{A}$ be closed with respect to weak convergence. Then, for given $\mu \in \mathcal{M}(\mathcal{H})$, assume that there exists $\nu \in \mathcal{A}$ such that $D_{\mathrm{KL}}(\nu\|\mu) < \infty$. It follows that there exists a minimizer $\nu \in \mathcal{A}$ solving problem* (2.2).

If we know in addition that the set $\mathcal{A}$ is *convex*, then the following classical stronger result holds.

Proposition 2.3 (see [10, Theorem 2.1]). *Assume that $\mathcal{A}$ is convex and closed with respect to* total variation *convergence. Assume, furthermore, that there exists a $\nu \in \mathcal{A}$ with $D_{\mathrm{KL}}(\nu\|\mu) < \infty$. Then there exists a* unique *minimizer $\nu \in \mathcal{A}$ solving problem* (2.2).

However, in most situations of interest in this paper, such as approximation by Gaussians, the set $\mathcal{A}$ is not convex. Moreover, the proof of Proposition 2.3 does not carry over to the case of nonconvex $\mathcal{A}$ and, indeed, uniqueness of minimizers is not expected in general in this case (see, however, the discussion of uniqueness in subsection 3.4). Still, the methods used in proving Proposition 2.3 do have the following interesting consequence for our setting. Before we state it we recall the definition of the total variation norm of two probability measures. It is given by

$$D_{\mathrm{tv}}(\nu,\mu) = \|\nu-\mu\|_{\mathrm{tv}} = \frac{1}{2}\int \left|\frac{d\nu}{d\lambda}(x) - \frac{d\mu}{d\lambda}(x)\right| \lambda(dx),$$

where λ is a probability measure on $\mathcal{H}$ such that $\nu \ll \lambda$ and $\mu \ll \lambda$.

LEMMA 2.4. *Let (ν_n) be a sequence in $\mathcal{M}(\mathcal{H})$, and let $\nu_\star \in \mathcal{M}(\mathcal{H})$ and $\mu \in \mathcal{M}(\mathcal{H})$ be probability measures such that for any $n \geq 1$ we have $D_{\mathrm{KL}}(\nu_n\|\mu) < \infty$ and $D_{\mathrm{KL}}(\nu_\star\|\mu) < \infty$. Suppose that the ν_n converge weakly to $\nu_\star$ and in addition that*

$$D_{\mathrm{KL}}(\nu_n\|\mu) \to D_{\mathrm{KL}}(\nu_\star\|\mu).$$

Then ν_n converges to $\nu_\star$ in total variation norm.

The proof of Lemma 2.4 can be found in subsection 4.1. Combining Lemma 2.4 with Proposition 2.1 implies in particular the following.

COROLLARY 2.5. *Let $\mathcal{A}$ be closed with respect to weak convergence and μ such that there exists a $\nu \in \mathcal{A}$ with $D_{\mathrm{KL}}(\nu\|\mu) < \infty$. Let $\nu_n \in \mathcal{A}$ satisfy*

$$D_{\mathrm{KL}}(\nu_n\|\mu) \to \inf_{\nu\in\mathcal{A}} D_{\mathrm{KL}}(\nu\|\mu). \tag{2.3}$$

Then, after passing to a subsequence, ν_n converges weakly to a $\nu_\star \in \mathcal{A}$ that realizes the infimum in (2.3). *Along the subsequence we have, in addition, that*

$$\|\nu_n - \nu_\star\|_{\mathrm{tv}} \to 0.$$

Thus, in particular, if $\mathcal{A}$ is the Gaussian class, then the preceding corollary applies.

2.2. A finite dimensional example. In this subsection we illustrate the minimization problem in the simplified situation where $\mathcal{H} = \mathbb{R}^n$ for some $n \geq 1$. In this situation it is natural to consider target measures μ of the form

$$\frac{d\mu}{d\mathcal{L}^n}(x) = \frac{1}{Z_\mu}\exp\big(-\Phi(x)\big) \tag{2.4}$$

for some smooth function $\Phi\colon \mathbb{R}^n \to \mathbb{R}_+$. Here $\mathcal{L}^n$ denotes the Lebesgue measure on $\mathbb{R}^n$. We consider the minimization problem (2.2) in the case where $\mathcal{A}$ is the set of all Gaussian measures on $\mathbb{R}^n$.

If $\nu = N(m,C)$ is a Gaussian on $\mathbb{R}^n$ with mean m and a nondegenerate covariance matrix C, we get

$$\begin{aligned} D_{\mathrm{KL}}(\nu\|\mu) &= \mathbb{E}^\nu\left[\Phi(x) - \frac{\langle x, C^{-1}x\rangle}{2}\right] - \frac{1}{2}\log\big(\det C\big) + \log\left(\frac{Z_\mu}{(2\pi)^{\frac{n}{2}}}\right) \\ &= \mathbb{E}^\nu\big[\Phi(x)\big] - \frac{1}{2}\log\big(\det C\big) - \frac{n}{2} + \log\left(\frac{Z_\mu}{(2\pi)^{\frac{n}{2}}}\right). \end{aligned} \tag{2.5}$$

The last two terms on the right-hand side of (2.5) do not depend on the Gaussian measure ν and can therefore be dropped in the minimization problem. In the case

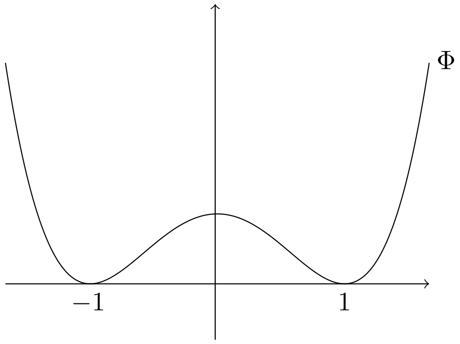

Fig. 1. *The double well potential* Φ.

where Φ is a polynomial, the expression $\mathbb{E}^{\nu}\big[\Phi(x)\big]$ consists of a Gaussian expectation of a polynomial and it can be evaluated explicitly.

To be concrete we consider the case where $n = 1$ and $\Phi(x) = \frac{1}{4\varepsilon}(x^2-1)^2$ so that the measure μ has two peaks: see Figure 1. In this one dimensional situation we minimize $D_{\mathrm{KL}}(\nu\|\mu)$ over all measures $N(m,\sigma^2)$, $m \in \mathbb{R}$, $\sigma \geq 0$. Dropping the irrelevant constants in (2.5), we are led to minimizing

$$\begin{aligned}\mathcal{D}(m,\sigma) &:= \mathbb{E}^{N(m,\sigma^2)}\big[\Phi(x)\big] - \log(\sigma)\\ &= \left(\Phi(m) + \frac{\sigma^2}{2}\Phi''(m) + \frac{3\sigma^4}{4!}\Phi^{(4)}(m)\right) - \log(\sigma)\\ &= \frac{1}{\varepsilon}\left(\frac{1}{4}(m^2-1)^2 + \frac{\sigma^2}{2}(3m^2-1) + \frac{3\sigma^4}{4}\right) - \log(\sigma)\end{aligned}$$

over $(m,\sigma) \in \mathbb{R}\times[0,\infty)$.

In the limit $\varepsilon \to 0$, there are critical points at $m = \pm1, 0$ and $\sigma = 0$ and a perturbation expansion demonstrates that these minimizers deform into two different Gaussian approximations, centered near ±1, for small ε. Numerical solution of the critical points of $\mathcal{D}$ (see Figure 2) illustrates this fact. Furthermore, we see the existence of three, then five, and finally one critical point as ε increases. For small ε the two minima near $x = \pm1$ are the global minimizers, whilst for larger ε the minimizer at the origin is the global minimizer.

2.3. Infinite dimensional motivation. For us there are two primary motivations for the theory that we have developed here. The first concerns the study of inverse problems in partial differential equations when given a Bayesian formulation [32]. This results in the need to determine properties of a probability measure μ on an infinite dimensional space, the posterior distribution, defined via its density with respect to another probability measure μ_0, the prior distribution. It is common to use Gaussian random field priors, in which case μ_0 is a Gaussian measure. Examples include determination of the initial condition for the Navier–Stokes equation from Eulerian or Lagrangian data at later times [8] and determination of the permeability in a Darcy groundwater flow model from hydraulic head measurements [13]. It is important to appreciate that the probability distribution μ can have multiple modes, as in the preceding finite dimensional example. One commonly adopted approach to Bayesian inverse problems, which is capable of capturing such multiple modes, is to find the maximum a posteriori (MAP) estimator. This corresponds to identifying

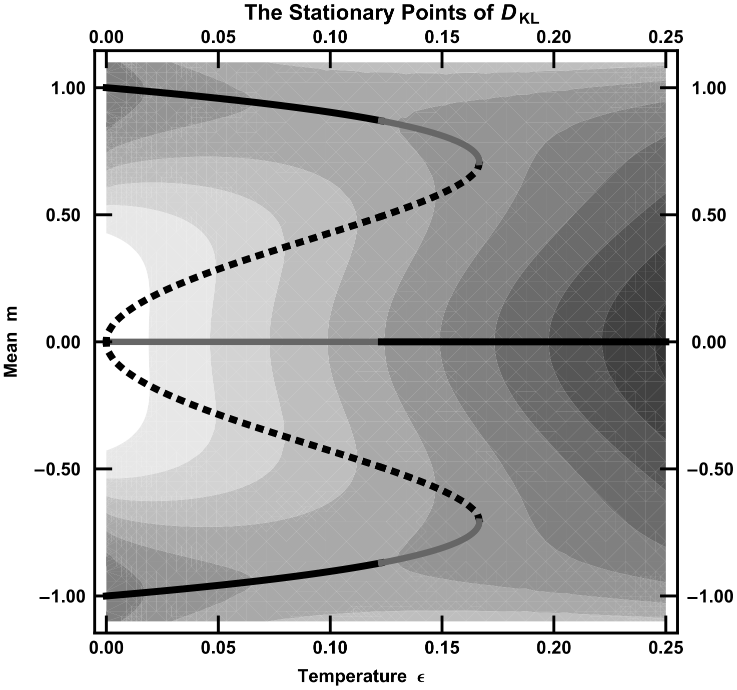

FIG. 2. *The solid lines denote minima, with the darker line used for the absolute minimum at the given temperature ε. The dotted lines denote maxima. At $\varepsilon = 1/6$, two stationary points annihilate one another at a fold bifurcation and only the symmetric solution, with mean $m = 0$, remains. However, even for $\varepsilon > 0.122822$, the symmetric mean zero solution is the global minimum.*

the center of balls of maximal probability in the limit of vanishingly small radius [12, 20]; this is linked to the classical theory of Tikhonov–Phillips regularization of inverse problems [15]. Another commonly adopted approach is to employ Monte Carlo Markov chain (MCMC) methods [24] to sample the probability measure of interest. The method of MAP estimation can be computationally tractable, but it loses important probabilistic information—it corresponds to finding the best Dirac measure approximations to the measure and does not include uncertainty around the point at which the Diracs are centered. In contrast, MCMC methods can, in principle, determine accurate probabilistic information but may be very expensive. The approach we advocate here, when adapted to find the best Gaussian approximation ν of the posterior measure μ, captures not only points of high probability but also the spread (uncertainty) around them; and the best Gaussian approximation may also be used to create improved MCMC methods which converge more quickly. These two ideas are illustrated in the companion paper [27]. We also highlight the fact, which is clearly illustrated in the preceding finite dimensional example, that minimizing $D_{\text{KL}}(\nu\|\mu)$ over Gaussian ν allows for the capture of several distinct modes, whilst minimizing $D_{\text{KL}}(\mu\|\nu)$ over Gaussian ν corresponds to moment matching (see [6, section 10.7]) and will hence combine multiple modes into a single Gaussian approximation. For inverse problems with multiple modes, the latter is clearly undesirable and the former becomes a preferred methodology.

The second primary motivation concerns the study of conditioned diffusion processes [19]. The paper [28] studies a variety of examples of Brownian dynamics models (gradient flow in a potential with additive white noise) which are conditioned to make transitions between critical points of the energy. The Onsager–Machlup functional (the analogue of the MAP estimator in this conditioned diffusion process setting) is studied, and the Γ-limit is identified in the zero temperature limit. Unfortunately, because the Onsager–Machlup functional does not capture entropic fluctuations, the predictions made by the Γ-limit functional can be physically unrealistic; in particular, when two modes are present in the probability measure, the minimizer of the Onsager–Machlup functional may incorrectly predict the preferred path between two critical points. The idea of studying best Gaussian approximations to the measure, rather than the Onsager–Machlup functional, holds the potential for removing this undesirable effect.

3. Kullback–Leibler minimization over Gaussian classes.

The previous subsection shows that the class of Gaussian measures is a natural one over which to minimize, although uniqueness cannot, in general, be expected. In this section we therefore study approximation within Gaussian classes and variants on this theme. Furthermore, we will assume that the measure of interest, μ, is equivalent (in the sense of measures) to a Gaussian $\mu_0 = N(m_0, C_0)$ on the separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$, with $\mathcal{F}$ the Borel σ-algebra.

More precisely, let $X \subseteq \mathcal{H}$ be a separable Banach space which is continuously embedded in $\mathcal{H}$, where X is measurable with respect to $\mathcal{F}$ and satisfies $\mu_0(X) = 1$. We also assume that $\Phi : X \to \mathbb{R}$ is continuous in the topology of X and that $\exp(-\Phi(x))$ is integrable with respect to μ_0.[2] Then the target measure μ is defined by

$$\frac{d\mu}{d\mu_0}(x) = \frac{1}{Z_\mu}\exp\big(-\Phi(x)\big), \tag{3.1}$$

where the normalization constant is given by

$$Z_\mu = \int_{\mathcal{H}} \exp\big(-\Phi(x)\big)\,\mu_0(dx) =: \mathbb{E}^{\mu_0}\big[\exp\big(-\Phi(x)\big)\big].$$

Here and below we use the notation $\mathbb{E}^{\mu_0}$ for the expectation with respect to the probability measure μ_0, and we also use similar notation for the expectation with respect to other probability measures. Measures of the form (3.1) with μ_0 Gaussian occur in the Bayesian approach to inverse problems with Gaussian priors and in the pathspace description of (possibly conditioned) diffusions with additive noise.

In subsection 3.1 we recall some basic definitions concerning Gaussian measure on Hilbert space and then state a straightforward consequence of the theoretical developments of the previous section for $\mathcal{A}$ comprising various Gaussian classes. Then, in subsection 3.2, we discuss how to parameterize the covariance of a Gaussian measure, introducing Schrödinger potential-type parameterizations of the precision (inverse covariance) operator. By example we show that whilst Gaussian measures within this parameterization may exhibit well-behaved minimizing sequences, the potentials themselves may behave badly along minimizing sequences, exhibiting oscillations or singularity formation. This motivates subsection 3.3, where we regularize the minimization to prevent this behavior. In subsection 3.4 we give conditions on Φ which result in uniqueness of minimizers, and in subsection 3.5 we make some remarks on generalizations of approximation within the class of Gaussian mixtures.

[2] In fact, continuity is only used in subsection 3.4; measurability will suffice in much of the paper.

3.1. Gaussian case. We start by recalling some basic facts about Gaussian measures. A probability measure ν on a separable Hilbert space $\mathcal{H}$ is Gaussian if for any ϕ in the dual space $\mathcal{H}^\star$ the push-forward measure $\nu \circ \phi^{-1}$ is Gaussian (where Dirac measures are viewed as Gaussians with variance 0) [11]. Furthermore, recall that ν is characterized by its mean and covariance, defined via the following (in the first case Bochner) integrals: the mean m is given by

$$m := \int_{\mathcal{H}} x \, \nu(dx) \in \mathcal{H},$$

and its covariance operator $C\colon \mathcal{H} \to \mathcal{H}$ satisfies

$$\int_{\mathcal{H}} \langle x, y_1 \rangle \langle x, y_2 \rangle \, \nu(dx) = \langle y_1, C y_2 \rangle$$

for all $y_1, y_2 \in \mathcal{H}$. Recall that C is a nonnegative, symmetric, trace-class operator, or equivalently $\sqrt{C}$ is a nonnegative, symmetric Hilbert–Schmidt operator. In what follows we will denote by $\mathcal{L}(\mathcal{H})$, $\mathcal{TC}(\mathcal{H})$, and $\mathcal{HS}(\mathcal{H})$ the spaces of linear, trace-class, and Hilbert–Schmidt operators on $\mathcal{H}$. We denote the Gaussian measure with mean m and covariance operator C by $N(m, C)$. We have collected some additional facts about Gaussian measures in Appendix B.

From now on, we fix a Gaussian measure $\mu_0 = N(m_0, C_0)$. We always assume that C_0 is a strictly positive operator. We denote the image of $\mathcal{H}$ under $C_0^{\frac{1}{2}}$, endowed with the scalar product $\langle C_0^{-\frac{1}{2}} \cdot, C_0^{-\frac{1}{2}} \cdot \rangle$, by $\mathcal{H}^1$, noting that this is the *Cameron–Martin space* of μ_0; we denote its dual space by $\mathcal{H}^{-1} = (\mathcal{H}^1)^\star$. We will make use of the natural finite dimensional projections associated to the operator C_0 in several places in what follows, and so we introduce notation associated to this for later use. Let $(e_\alpha, \alpha \geq 1)$ be the basis of $\mathcal{H}$ consisting of eigenfunctions of C_0, and let $(\lambda_\alpha, \alpha \geq 1)$ be the associated sequence of eigenvalues. For simplicity we assume that the eigenvalues are in nonincreasing order. Then for any $\gamma \geq 1$ we will denote $\mathcal{H}_\gamma := \operatorname{span}(e_1, \dots, e_\gamma)$ and the orthogonal projection onto $\mathcal{H}_\gamma$ by

$$\pi_\gamma \colon \mathcal{H} \to \mathcal{H}, \qquad x \mapsto \sum_{\alpha=1}^{\gamma} \langle x, e_\alpha \rangle \, e_\alpha. \tag{3.2}$$

Given such a measure μ_0 we assume that the target measure μ is given by (3.1).

For $\nu \ll \mu$, expression (2.1) can be rewritten, using (3.1) and the equivalence of μ and μ_0, as

$$\begin{aligned} D_{\mathrm{KL}}(\nu \| \mu) &= \mathbb{E}^\nu \left[\log \left(\frac{d\nu}{d\mu}(x) \right) \mathbf{1}_{\{\frac{d\nu}{d\mu} \neq 0\}} \right] \\ &= \mathbb{E}^\nu \left[\log \left(\frac{d\nu}{d\mu_0}(x) \times \frac{d\mu_0}{d\mu}(x) \right) \mathbf{1}_{\{\frac{d\nu}{d\mu_0} \neq 0\}} \right] \\ &= \mathbb{E}^\nu \left[\log \left(\frac{d\nu}{d\mu_0}(x) \right) \mathbf{1}_{\left\{\frac{d\nu}{d\mu_0} \neq 0\right\}} \right] + \mathbb{E}^\nu \left[\Phi(x) \right] + \log(Z_\mu). \end{aligned} \tag{3.3}$$

This calculation is classical and can be found, for example, in [1]. The expression in the first line shows that in order to evaluate the Kullback–Leibler divergence it is sufficient to compute an expectation with respect to the approximating measure $\nu \in \mathcal{A}$ and not with respect to the target μ.

The same expression shows positivity. To see this, decompose the measure μ into two nonnegative measures $\mu = \mu^{\parallel} + \mu^{\perp}$, where $\mu^{\parallel}$ is equivalent to ν and $\mu^{\perp}$ is singular with respect to ν. Then we can write with the Jensen inequality

$$\begin{aligned} D_{\mathrm{KL}}(\nu\|\mu) &= -\,\mathbb{E}^{\nu}\left[\log\left(\frac{d\mu^{\parallel}}{d\nu}(x)\right)\mathbf{1}_{\{\frac{d\nu}{d\mu}\neq 0\}}\right] \geq -\log\mathbb{E}^{\nu}\left[\frac{d\mu^{\parallel}}{d\nu}(x)\right] \\ &= -\log\mu^{\parallel}(\mathcal{H}) \geq 0. \end{aligned}$$

This establishes the general fact that relative entropy is nonnegative for our particular setting.

Finally, the expression in the third line of (3.3) shows that the normalization constant Z_μ enters into D_{KL} only as an additive constant that can be ignored in the minimization procedure. If we assume, furthermore, that the set $\mathcal{A}$ consists of Gaussian measures, Lemma 2.4 and Corollary 2.5 imply the following result.

THEOREM 3.1. *Let μ_0 be a Gaussian measure with mean $m_0 \in \mathcal{H}$ and covariance operator $C_0 \in \mathcal{TC}(\mathcal{H})$, and let μ be given by* (3.1)*. Consider the following choices for $\mathcal{A}$:*

1. *$\mathcal{A}_1 = \{$Gaussian measures on $\mathcal{H}\}$.*
2. *$\mathcal{A}_2 = \{$Gaussian measures on $\mathcal{H}$ equivalent to $\mu_0\}$.*
3. *For a fixed covariance operator $\hat{C} \in \mathcal{TC}(\mathcal{H})$,*

$$\mathcal{A}_3 = \{\textit{Gaussian measures on } \mathcal{H} \textit{ with covariance } \hat{C}\}.$$

4. *For a fixed mean $\hat{m} \in \mathcal{H}$,*

$$\mathcal{A}_4 = \{\textit{Gaussian measures on } \mathcal{H} \textit{ with mean } \hat{m}\}.$$

In each of these situations, as soon as there exists a single $\nu \in \mathcal{A}_i$ with $D_{\mathrm{KL}}(\nu\|\mu) < \infty$, there exists a minimizer of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ in $\mathcal{A}_i$. Furthermore, ν is necessarily equivalent to μ_0 in the sense of measures.

Remark 3.2. Even in the case $\mathcal{A}_1$ the condition that there exists a single ν with finite $D_{\mathrm{KL}}(\nu\|\mu)$ is not always satisfied. For example, if $\Phi(x) = \exp\left(\|x\|^4_{\mathcal{H}}\right)$, then for any Gaussian measure ν on $\mathcal{H}$ we have, using the identity (3.3), that

$$D_{\mathrm{KL}}(\nu\|\mu) = D_{\mathrm{KL}}(\nu\|\mu_0) + \mathbb{E}^{\nu}\big[\Phi(x)\big] + \log(Z_\mu) = +\infty.$$

In the cases $\mathcal{A}_1, \mathcal{A}_3$, and $\mathcal{A}_4$ such a ν is necessarily absolutely continuous with respect to μ and hence equivalent to μ_0; this equivalence is encapsulated directly in $\mathcal{A}_2$. The conditions for this to be possible are stated in the Feldman–Hajek theorem, Proposition B.2.

3.2. Parameterization of Gaussian measures.

When solving the minimization problem (2.2) it will usually be convenient to parameterize the set $\mathcal{A}$ in a suitable way. In the case where $\mathcal{A}$ consists of all Gaussian measures on $\mathcal{H}$ the first choice that comes to mind is to parameterize it by the mean $m \in \mathcal{H}$ and the covariance operator $C \in \mathcal{TC}(\mathcal{H})$. In fact it is often convenient, for both computational and modeling reasons, to work with the inverse covariance (precision) operator, which, because the covariance operator is strictly positive and trace-class, is a densely defined unbounded operator.

Recall that the underlying Gaussian centered reference measure μ_0 has covariance C_0. We will consider covariance operators C of the form

$$C^{-1} = C_0^{-1} + \Gamma \tag{3.4}$$

for suitable operators Γ. We will characterize below precisely the conditions on Γ which are necessary and sufficient for the resulting class of approximating Gaussian measures to be equivalent to μ_0. From an applications perspective it is interesting to consider the case where $\mathcal{H}$ is a function space and Γ is a multiplication operator. Then Γ has the form $\Gamma u = v(\cdot)u(\cdot)$ for some fixed function v which we refer to as a *potential* in analogy with the Schrödinger setting. In this case parameterizing the Gaussian family $\mathcal{A}$ by the pair of functions (m, v) comprises a considerable dimension reduction over parameterization by the pair (m, C), since C is an operator. We develop the theory of the minimization problem (2.2) in terms of Γ and extract results concerning the potential v as particular examples.

The end of Remark 3.2 shows that, without loss of generality, we can always restrict ourselves to covariance operators C corresponding to Gaussian measures which are equivalent to μ_0. In general the inverse C^{-1} of such an operator and the inverse C_0^{-1} of the covariance operator of μ_0 do not have the same *operator* domain. Indeed, see Example 3.8 below for an example of two equivalent centered Gaussian measures whose inverse covariance operators have different domains. But item 1 in the Feldman–Hajek theorem (Proposition B.2) implies that the domains of $C^{-\frac{1}{2}}$ and $C_0^{-\frac{1}{2}}$, i.e. the *form domains* of C^{-1} and C_0^{-1}, coincide. Hence, if we view the operators C^{-1} and C_0^{-1} as symmetric quadratic forms on $\mathcal{H}^1$ or as operators from $\mathcal{H}^1$ to $\mathcal{H}^{-1}$, it makes sense to add and subtract them. In particular, we can interpret (3.4) as

$$\Gamma := C^{-1} - C_0^{-1} \in \mathcal{L}(\mathcal{H}^1, \mathcal{H}^{-1}). \tag{3.5}$$

Actually, Γ is not only bounded from $\mathcal{H}^1$ to $\mathcal{H}^{-1}$. Item 3 in Proposition B.2 can be restated as

$$\|\Gamma\|^2_{\mathcal{HS}(\mathcal{H}^1,\mathcal{H}^{-1})} := \|C_0^{\frac{1}{2}} \Gamma C_0^{\frac{1}{2}}\|^2_{\mathcal{HS}(\mathcal{H})} < \infty; \tag{3.6}$$

here $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ denotes the space of Hilbert–Schmidt operators from $\mathcal{H}^1$ to $\mathcal{H}^{-1}$. The space $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ is continuously embedded into $\mathcal{L}(\mathcal{H}^1, \mathcal{H}^{-1})$.

Conversely, it is natural to ask whether condition (3.6) alone implies that Γ can be obtained from the covariance of a Gaussian measure as in (3.5). The following lemma states that this is indeed the case as soon as one has an additional positivity condition; the proof is left to section 4.2.

LEMMA 3.3. *For any symmetric Γ in $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ the quadratic form given by*

$$Q_\Gamma(u, v) = \langle u, C_0^{-1} v\rangle + \langle u, \Gamma v\rangle$$

is bounded from below and closed on its form domain $\mathcal{H}^1$. Hence it is associated to a unique self-adjoint operator which we will also denote by $C_0^{-1} + \Gamma$. The operator $(C_0^{-1} + \Gamma)^{-1}$ is the covariance operator of a Gaussian measure on $\mathcal{H}$ which is equivalent to μ_0 if and only if Q_Γ is strictly positive.

Lemma 3.3 shows that we can parameterize the set of Gaussian measures that are equivalent to μ_0 by their mean and by the operator Γ. For fixed $m \in \mathcal{H}$ and $\Gamma \in \mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ we write $N_{P,0}(m, \Gamma)$ for the Gaussian measure with mean m and covariance operator $C^{-1} = C_0^{-1} + \Gamma$, where the suffix $(P, 0)$ is to denote the specification via the shift in precision operator from that of μ_0. We use the convention to set $N_{P,0}(m, \Gamma) = \delta_m$ if $C_0^{-1} + \Gamma$ fails to be positive. Then we set

$$\mathcal{A} := \{N_{P,0}(m, \Gamma) \in \mathcal{M}(\mathcal{H}) \colon m \in \mathcal{H},\, \Gamma \in \mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})\}. \tag{3.7}$$

Lemma 3.3 shows that the subset of $\mathcal{A}$ in which Q_Γ is strictly positive comprises Gaussian measures absolutely continuous with respect to μ_0. Theorem 3.1, with the choice $\mathcal{A} = \mathcal{A}_2$, implies immediately the existence of a minimizer for problem (2.2) for this choice of $\mathcal{A}$.

Corollary 3.4. *Let μ_0 be a Gaussian measure with mean $m_0 \in \mathcal{H}$ and covariance operator $C_0 \in \mathcal{TC}(\mathcal{H})$, and let μ be given by* (3.1)*. Consider $\mathcal{A}$ given by* (3.7)*. Provided there exists a single $\nu \in \mathcal{A}$ with $D_{\mathrm{KL}}(\nu\|\mu) < \infty$, then there exists a minimizer of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ in $\mathcal{A}$. Furthermore, ν is necessarily equivalent to μ_0 in the sense of measures.*

However, this corollary does not tell us much about the manner in which minimizing sequences approach the limit. With some more work we can actually characterize the convergence more precisely in terms of the parameterization.

Theorem 3.5. *Let μ_0 be a Gaussian measure with mean $m_0 \in \mathcal{H}$ and covariance operator $C_0 \in \mathcal{TC}(\mathcal{H})$, and let μ be given by* (3.1)*. Consider $\mathcal{A}$ given by* (3.7)*. Let $N_{P,0}(m_n, \Gamma_n)$ be a sequence of Gaussian measures in $\mathcal{A}$ that converge weakly to $\nu_\star$ with*

$$D_{\mathrm{KL}}(\nu_n\|\mu) \to D_{\mathrm{KL}}(\nu_\star\|\mu).$$

Then $\nu_\star = N_{P,0}(m_\star, \Gamma_\star)$ and

$$\|m_n - m_\star\|_{\mathcal{H}^1} + \big\|\Gamma_n - \Gamma_\star\big\|_{\mathcal{HS}(\mathcal{H}^1,\mathcal{H}^{-1})} \to 0.$$

Proof. Lemma B.1 shows that $\nu_\star$ is Gaussian, and Theorem 3.1 shows that in fact $\nu_\star = N_{P,0}(m_\star, \Gamma_\star)$. It follows from Lemma 2.4 that ν_n converges to $\nu_\star$ in total variation. Lemma B.4, which follows, shows that

$$\big\|C_\star^{\frac{1}{2}}(C_n^{-1} - C_\star^{-1})C_\star^{\frac{1}{2}}\big\|_{\mathcal{HS}(\mathcal{H})} + \|m_n - m_\star\|_{\mathcal{H}^1} \to 0.$$

By the Feldman–Hajek theorem (Proposition B.2, item 1) the Cameron–Martin spaces $C_\star^{\frac{1}{2}}\mathcal{H}$ and $C_0^{\frac{1}{2}}\mathcal{H}$ coincide with $\mathcal{H}^1$, and hence, since $C_n^{-1} - C_\star^{-1} = \Gamma_n - \Gamma_\star$, the desired result follows. ▯

The following example concerns a subset of the set $\mathcal{A}$ given by (3.7) found by writing Γ as a multiplication of the identity I by a constant. This structure is useful for numerical computations, for example if μ_0 represents Wiener measure (possibly conditioned) and we seek an approximation ν to μ with a mean m and covariance of Ornstein–Uhlenbeck type (again possibly conditioned).

Example 3.6. Let $C^{-1} = C_0^{-1} + \beta I$ so that

$$C = (I + \beta C_0)^{-1} C_0. \tag{3.8}$$

Let $\mathcal{A}'$ denote the set of Gaussian measures on $\mathcal{H}$ which have covariance of the form (3.8) for some constant $\beta \in \mathbb{R}$. This set is parameterized by the pair $(m, \beta) \in \mathcal{H} \times \mathbb{R}$. The operator C^{-1} and the associated form are strictly positive if and only if $\beta \in \mathsf{I} = (-\lambda_1^{-1}, \infty)$; recall that λ_1, defined above (3.2), is the largest eigenvalue of C_0. Therefore, by Lemma 3.3, C is the covariance of a Gaussian equivalent to μ_0 if and only if β satisfies this condition. Note also that the covariance C satisfies $C^{-1} = C_0^{-1} + \beta$, and so $\mathcal{A}'$ is a subset of $\mathcal{A}$ given by (3.7) arising where Γ is multiplication by a constant.

Now consider minimizing sequences $\{\nu_n\}$ from $\mathcal{A}'$ for $D_{\mathrm{KL}}(\nu\|\mu)$. Any weak limit $\nu_\star$ of a sequence $\nu_n = N\big(m_n, (I + \beta_n C_0)^{-1} C_0\big) \in \mathcal{A}'$ is necessarily Gaussian by

Lemma B.1, item 1, and we denote it by $N(m_\star, C_\star)$. By item 2 of the same lemma we deduce that $m_n \to m_\star$ strongly in $\mathcal{H}$ and by item 3 that $(I+\beta_n C_0)^{-1}C_0 \to C_\star$ strongly in $\mathcal{L}(\mathcal{H})$. Thus, for any $\alpha \geq 1$, and recalling that e_α are the eigenvectors of C_0, $\|C_\star e_\alpha - (1+\beta_n\lambda_\alpha)^{-1}\lambda_\alpha e_\alpha\| \to 0$ as $n \to \infty$. Furthermore, necessarily $\beta_n \in \mathsf{I}$ for each n. We now argue by contradiction that there are no subsequences $\beta_{n'}$ converging to either $-\lambda_1^{-1}$ or ∞. For contradiction assume first that there is a subsequence converging to $-\lambda_1^{-1}$. Along this subsequence we have $(1+\beta_n\lambda_1)^{-1} \to \infty$, and hence we deduce that $C_\star e_1 = \infty$, so that $C_\star$ cannot be trace-class, a contradiction. Similarly assume for contradiction that there is a subsequence converging to ∞. Along this subsequence we have $(1+\beta_n\lambda_\alpha)^{-1} \to 0$ and hence that $C_\star e_\alpha = 0$ for every α. In this case $\nu_\star$ would be a Dirac measure and hence not equivalent to μ_0 (recall our assumption that C_0 is a strictly positive operator). Thus there must be a subsequence converging to a limit $\beta \in \mathsf{I}$ and we deduce that $C_\star e_\alpha = (1+\beta\lambda_\alpha)^{-1}\lambda_\alpha e_\alpha$, proving that $C_\star = (I+\beta C_0)^{-1}C_0$ as required.

Another class of Gaussians which is natural in applications, and in which the parameterization of the covariance is finite dimensional, is as follows.

Example 3.7. Recall the notation π_γ for the orthogonal projection onto $\mathcal{H}_\gamma :=$ span$(e_1, \dots, e_\gamma)$, the span of the first γ eigenvalues of C_0. We seek C in the form

$$C^{-1} = \big((I-\pi_\gamma)C_0(I-\pi_\gamma)\big)^{-1} + \Gamma,$$

where

$$\Gamma = \sum_{i,j\leq N} \gamma_{ij} e_i \otimes e_j.$$

It then follows that

$$C = (I-\pi_\gamma)C_0(I-\pi_\gamma) + \Gamma^{-1}, \tag{3.9}$$

provided that Γ is invertible. Let $\mathcal{A}'$ denote the set of Gaussian measures on $\mathcal{H}$ which have covariance of the form (3.9) for some operator Γ invertible on $\mathcal{H}_\gamma$. Now consider minimizing sequences $\{\nu_n\}$ from $\mathcal{A}'$ for $D_{\mathrm{KL}}(\nu\|\mu)$ with mean m_n and covariance $C_n = (I-\pi_\gamma)C_0(I-\pi_\gamma) + \Gamma_n^{-1}$. Any weak limit $\nu_\star$ of the sequence $\nu_n \in \mathcal{A}'$ is necessarily Gaussian by Lemma B.1, item 1, and we denote it by $N(m_\star, C_\star)$. As in the preceding example, we deduce that $m_n \to m_\star$ strongly in $\mathcal{H}$. Similarly we also deduce that Γ_n^{-1} converges to a nonnegative matrix. A simple contradiction shows that, in fact, this limiting matrix is invertible since otherwise $N(m_\star, C_\star)$ would not be equivalent to μ_0. We denote the limit by $\Gamma_\star^{-1}$. We deduce that the limit of the sequence ν_n is in $\mathcal{A}'$ and that $C_\star = (I-\pi_\gamma)C_0(I-\pi_\gamma) + \Gamma_\star^{-1}$.

3.3. Regularization for parameterization of Gaussian measures. The previous section demonstrates that parameterization of Gaussian measures in the set $\mathcal{A}$ given by (3.7) leads to a well-defined minimization problem (2.2) and that, furthermore, minimizing sequences in $\mathcal{A}$ will give rise to means m_n and operators Γ_n converging in $\mathcal{H}^1$ and $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$, respectively. However, convergence in the space $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ may be quite weak and unsuitable for numerical purposes; in particular, if $\Gamma_n u = v_n(\cdot)u(\cdot)$, then the sequence (v_n) may behave quite badly, even though (Γ_n) is well-behaved in $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$. For this reason we consider, in this subsection, regularization of the minimization problem (2.2) over $\mathcal{A}$ given by (3.7). But before doing so we provide two examples illustrating the potentially undesirable properties of convergence in $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$.

Example 3.8 (compare [30, Example 3 in Chapter X.2]). Let $C_0^{-1} = -\partial_t^2$ be the negative Dirichlet–Laplace operator on $[-1, 1]$ with domain $H^2([-1, 1]) \cap H_0^1([-1, 1])$, and let $\mu_0 = N(0, C_0)$; i.e., μ_0 is the distribution of a Brownian bridge on $[-1, 1]$. In this case $\mathcal{H}^1$ coincides with the Sobolev space H_0^1. We note that the measure μ_0 assigns full mass to the space X of continuous functions on $[-1, 1]$, and hence all integrals with respect to μ_0 in what follows can be computed over X. Furthermore, the centered unit ball in X,

$$B_X(0; 1) := \left\{ x \in X : \sup_{t\in[-1,1]} |x(t)| \le 1 \right\},$$

has positive μ_0 measure.

Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a standard mollifier; i.e., $\phi \in \mathcal{C}^\infty$, $\phi \ge 0$, ϕ is compactly supported in $[-1, 1]$, and $\int_{\mathbb{R}} \phi(t)\, dt = 1$. Then for any n define $\phi_n(t) = n\phi(tn)$, together with the probability measures $\nu_n \ll \mu_0$ given by

$$\frac{d\nu_n}{d\mu_0}(x(\cdot)) = \frac{1}{Z_n} \exp\left(-\frac{1}{2} \int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \right),$$

where

$$Z_n := \mathbb{E}^{\mu_0} \exp\left(-\frac{1}{2} \int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \right).$$

The ν_n are also Gaussian, as Lemma C.1 shows. Using the fact that $\mu_0(X) = 1$ it follows that

$$\exp(-1/2)\mu_0\big(B_X(0; 1)\big) \le Z_n \le 1.$$

Now define probability measure $\nu_\star$ by

$$\frac{d\nu_\star}{d\mu_0}(x(\cdot)) = \frac{1}{Z_\star} \exp\left(-\frac{x(0)^2}{2} \right),$$

noting that

$$\exp(-1/2)\mu_0\big(B_X(0; 1)\big) \le Z_\star \le 1.$$

For any $x \in X$ we have

$$\int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \to x(0)^2.$$

An application of the dominated convergence theorem shows that $Z_n \to Z_\star$ and hence that $Z_n^{-1} \to Z_\star^{-1}$ and $\log(Z_n) \to \log(Z_\star)$.

Further applications of the dominated convergence theorem show that the ν_n converge weakly to $\nu_\star$, which is also then Gaussian by Lemma B.1, and that the Kullback–Leibler divergence between ν_n and $\nu_\star$ satisfies

$$\begin{aligned} D_{\mathrm{KL}}(\nu_n \| \nu_\star) = \frac{1}{Z_n} \mathbb{E}^{\mu_0} \Bigg[&\exp\left(-\frac{1}{2} \int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \right) \\ &\times \frac{1}{2}\left(x(0)^2 - \int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \right) \Bigg] + \big(\log(Z_\star) - \log(Z_n) \big) \qquad \to 0. \end{aligned}$$

Lemma C.1 shows that ν_n is the centered Gaussian with covariance C_n given by $C_n^{-1} = C_0^{-1} + \phi_n$. Formally, the covariance operator associated to $\nu_\star$ is given by $C_0^{-1} + \delta_0$, where δ_0 is the Dirac δ function. Nonetheless, the implied multiplication operators converge to a limit in $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$. In applications such limiting behavior of the potential in an inverse covariance representation, to a distribution, may be computationally undesirable.

Example 3.9. We consider a second example in a similar vein but linked to the theory of averaging for differential operators. Choose μ_0 as in the preceding example, and now define $\phi_n(\cdot) = \phi(n\cdot)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a positive smooth 1-periodic function with mean $\overline{\phi}$. Define C_n by $C_n^{-1} = C_0^{-1} + \phi_n$ similarly as before. It follows, as in the previous example, by use of Lemma C.1, that the measures ν_n are centered Gaussians with covariance C_n and are equivalent to μ_0 and that

$$\frac{d\nu_n}{d\mu_0}(x(\cdot)) = \frac{1}{Z_n} \exp\bigg(-\frac{1}{2} \int_{-1}^{1} \phi_n(t)\, x(t)^2\, dt \bigg).$$

By the dominated convergence theorem, as in the previous example, it also follows that the ν_n converge weakly to $\nu_\star$ with

$$\frac{d\nu_\star}{d\mu_0}(x(\cdot)) = \frac{1}{Z_\star} \exp\bigg(-\frac{1}{2}\overline{\phi} \int_{-1}^{1} x(t)^2\, dt \bigg).$$

Again using Lemma C.1, $\nu_\star$ is the centered Gaussian with covariance $C_\star$ given by $C_\star^{-1} = C_0^{-1} + \overline{\phi}$. The Kullback–Leibler divergence satisfies $D_{\mathrm{KL}}(\nu_n \| \nu_\star) \to 0$, also by application of the dominated convergence theorem as in the previous example. Thus minimizing sequences may exhibit multiplication functions which oscillate with increasing frequency whilst the implied operators Γ_n converge in $\mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$. Again this may be computationally undesirable in many applications.

The previous examples suggest that, in order to induce improved behavior of minimizing sequences related to the operators Γ, in particular when Γ is a multiplication operator, it may be useful to regularize the minimization in problem (2.2). To this end, let $\mathcal{G} \subseteq \mathcal{HS}(\mathcal{H}^1, \mathcal{H}^{-1})$ be a Hilbert space of linear operators. For fixed $m \in \mathcal{H}$ and $\Gamma \in \mathcal{G}$ we write $N_{P,0}(m, \Gamma)$ for the Gaussian measure with mean m and its covariance operator given by (3.5). We now make the choice

$$\mathcal{A} := \{N_{P,0}(m, \Gamma) \in \mathcal{M}(\mathcal{H})\colon m \in \mathcal{H},\, \Gamma \in \mathcal{G}\}. \tag{3.10}$$

Again, we use the convention $N_{P,0}(m, \Gamma) = \delta_0$ if $C_0^{-1} + \Gamma$ fails to be positive. Then, for some $\delta > 0$ we consider the modified minimization problem

$$\operatorname*{argmin}_{\nu \in \mathcal{A}} \Big(D_{\mathrm{KL}}(\nu, \mu) + \delta \|\Gamma\|_{\mathcal{G}}^2 \Big). \tag{3.11}$$

We have existence of minimizers for problem (3.11) under very general assumptions. In order to state these assumptions, we introduce auxiliary interpolation spaces. For any $s > 0$, we denote by $\mathcal{H}^s$ the domain of $C_0^{-\frac{s}{2}}$ equipped with the scalar product $\langle \cdot, C_0^{-s} \cdot \rangle$ and define $\mathcal{H}^{-s}$ by duality.

THEOREM 3.10. *Let μ_0 be a Gaussian measure with mean $m_0 \in \mathcal{H}$ and covariance operator $C_0 \in \mathcal{TC}(\mathcal{H})$, and let μ be given by* (3.1)*. Consider $\mathcal{A}$ given by* (3.10)*. Suppose that the space $\mathcal{G}$ consists of symmetric operators on $\mathcal{H}$ and embeds compactly into the space of bounded linear operators from $\mathcal{H}^{1-\kappa}$ to $\mathcal{H}^{-(1-\kappa)}$ for some $0 < \kappa < 1$.*

Then, provided that $D_{\mathrm{KL}}(\mu_0\|\mu) < \infty$, there exists a minimizer $\nu_\star = N_{P,0}(m_\star, \Gamma_\star)$ for problem (3.11).

Furthermore, along any minimizing sequence $\nu(m_n, \Gamma_n)$ there is a subsequence $\nu(m_{n'}, \Gamma_{n'})$ along which $\Gamma_{n'} \to \Gamma_\star$ strongly in $\mathcal{G}$ and $\nu(m_{n'}, \Gamma_{n'}) \to \nu(m_\star, \Gamma_\star)$ with respect to the total variation distance.

Proof. The assumption $D_{\mathrm{KL}}(\mu_0\|\mu) < \infty$ implies that the infimum in (3.11) is finite and nonnegative. Let $\nu_n = N_{P,0}(m_n, \Gamma_n)$ be a minimizing sequence for (3.11). As both $D_{\mathrm{KL}}(\nu_n\|\mu)$ and $\|\Gamma_n\|_{\mathcal{G}}^2$ are nonnegative, this implies that $D_{\mathrm{KL}}(\nu_n\|\mu)$ and $\|\Gamma_n\|_{\mathcal{G}}^2$ are bounded along the sequence. Hence, by Proposition 2.1 and by the compactness assumption on $\mathcal{G}$, after passing to a subsequence twice we can assume that the measures ν_n converge weakly as probability measures to a measure $\nu_\star$ and the operators Γ_n converge weakly in $\mathcal{G}$ to an operator $\Gamma_\star$; furthermore, the Γ_n also converge in the operator norm of $\mathcal{L}(\mathcal{H}^{1-\kappa}, \mathcal{H}^{-(1-\kappa)})$ to $\Gamma_\star$. By the lower semicontinuity of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ with respect to weak convergence of probability measures (see Proposition 2.1) and by the lower semicontinuity of $\Gamma \mapsto \|\Gamma\|_{\mathcal{G}}^2$ with respect to weak convergence in $\mathcal{G}$ we can conclude that

$$\begin{aligned} D_{\mathrm{KL}}(\nu_\star\|\mu) + \delta\|\Gamma_\star\|_{\mathcal{G}}^2 &\leq \liminf_{n\to\infty} D_{\mathrm{KL}}(\nu_n\|\mu) + \liminf_{n\to\infty} \delta\|\Gamma_n\|_{\mathcal{G}}^2 \\ &\leq \lim_{n\to\infty}\Big(D_{\mathrm{KL}}(\nu_n\|\mu) + \delta\|\Gamma_n\|_{\mathcal{G}}^2\Big) \\ &= \inf_{\nu\in\mathcal{A}}\Big(D_{\mathrm{KL}}(\nu\|\mu) + \delta\|\Gamma\|_{\mathcal{G}}^2\Big). \end{aligned} \tag{3.12}$$

By Lemma B.1, $\nu_\star$ is a Gaussian measure with mean $m_\star$ and covariance operator $C_\star$ and we have

$$\|m_n - m_\star\|_{\mathcal{H}} \to 0 \qquad \text{and} \qquad \|C_n - C_\star\|_{\mathcal{L}(\mathcal{H})} \to 0. \tag{3.13}$$

We want to show that $C_\star = (C_0 + \Gamma_\star)^{-1}$ in the sense of Lemma 3.3. In order to see this, note that $\Gamma_\star \in \mathcal{L}(\mathcal{H}^{1-\kappa}, \mathcal{H}^{-(1-\kappa)})$, which implies that for $x \in \mathcal{H}^1$ we have for any $\lambda > 0$

$$\begin{aligned} \langle x, \Gamma_\star x\rangle &\leq \big\|\Gamma_\star\big\|_{\mathcal{L}(\mathcal{H}^{1-\kappa},\mathcal{H}^{-(1-\kappa)})}\big\|x\big\|_{\mathcal{H}^{1-\kappa}}^2 \\ &\leq \big\|\Gamma_\star\big\|_{\mathcal{L}(\mathcal{H}^{1-\kappa},\mathcal{H}^{-(1-\kappa)})}\Big(\lambda(1-\kappa)\big\|x\big\|_{\mathcal{H}^1}^2 + \lambda^{-\frac{1-\kappa}{\kappa}}\kappa\big\|x\big\|_{\mathcal{H}}^2\Big). \end{aligned}$$

Hence, $\Gamma_\star$ is infinitesimally form-bounded with respect to C_0^{-1} (see, e.g., [30, Chapter X.2]). In particular, by the KLMN theorem (see [30, Theorem X.17]) the form $\langle x, C_0^{-1}x\rangle + \langle x, \Gamma_\star x\rangle$ is bounded from below and closed. Hence there exists a unique self-adjoint operator denoted by $C_0^{-1} + \Gamma_\star$ with form domain $\mathcal{H}^1$ which generates this form.

The convergence of $C_n = (C_0^{-1} + \Gamma_n)^{-1}$ to $C_\star$ in $\mathcal{L}(\mathcal{H})$ implies, in particular, that the C_n are bounded in the operator norm, and hence the spectra of the $C_0^{-1} + \Gamma_n$ are away from zero from below, uniformly. This implies that

$$\inf_{\|x\|_{\mathcal{H}}=1}\Big(\langle x, C_0^{-1}x\rangle + \langle x, \Gamma_\star x\rangle\Big) \geq \liminf_{n\to\infty}\inf_{\|x\|_{\mathcal{H}}=1}\Big(\langle x, C_0^{-1}x\rangle + \langle x, \Gamma_n x\rangle\Big) > 0$$

so that $C_0^{-1} + \Gamma_\star$ is a positive operator and, in particular, invertible and so is $(C_0^{-1} + \Gamma_\star)^{\frac{1}{2}}$. As $(C_0^{-1} + \Gamma_\star)^{\frac{1}{2}}$ is defined on all of $\mathcal{H}^1$, its inverse maps onto $\mathcal{H}^1$, and hence the closed graph theorem implies that $C_0^{-\frac{1}{2}}(C_0^{-1}+\Gamma_\star)^{-\frac{1}{2}}$ is a bounded operator on $\mathcal{H}$. From this we can conclude that for all $x \in \mathcal{H}^1$

$$\begin{aligned}
&\big|\langle x,(C_0^{-1}+\Gamma_n)x\rangle - \langle x,(C_0^{-1}+\Gamma_\star)x\rangle\big| \\
&\quad\leq \big\|\Gamma_n-\Gamma_\star\big\|_{\mathcal{L}(\mathcal{H}^1,\mathcal{H}^{-1})}\big\|x\big\|_{\mathcal{H}^1}^2 \\
&\quad\leq \big\|\Gamma_n-\Gamma_\star\big\|_{\mathcal{L}(\mathcal{H}^1,\mathcal{H}^{-1})}\big\|C_0^{-\frac{1}{2}}(C_0^{-1}+\Gamma_\star)^{-\frac{1}{2}}\big\|_{\mathcal{L}(\mathcal{H})}^2\big\|(C_0^{-1}+\Gamma_\star)^{\frac{1}{2}}x\big\|_{\mathcal{H}}^2.
\end{aligned}$$

By [31, Theorem VIII.25] this implies that $C_0^{-1}+\Gamma_\star$ converges to $C_0^{-1}+\Gamma_\star$ in the strong resolvent sense. As all operators are positive and bounded away from zero by [31, Theorem VIII.23], we can conclude that the inverses $(C_0^{-1}+\Gamma_n)^{-1}$ converge to $(C_0^{-1}+\Gamma_\star)^{-1}$. By (3.13) this implies that $C_\star=(C_0^{-1}+\Gamma_\star)^{-1}$ as desired.

We can conclude that $\nu_\star = N_{P,0}(m_\star,\Gamma_\star)$ and hence that

$$D_{\mathrm{KL}}(\nu_\star\|\mu)+\delta\|\Gamma_\star\|_{\mathcal{G}}^2 \geq \inf_{\nu\in\mathcal{A}}\Big(D_{\mathrm{KL}}(\nu\|\mu)+\delta\|\Gamma\|_{\mathcal{G}}^2\Big),$$

implying from (3.12) that

$$\begin{aligned}
D_{\mathrm{KL}}(\nu_\star\|\mu)+\delta\|\Gamma_\star\|_{\mathcal{G}}^2 &= \liminf_{n\to\infty} D_{\mathrm{KL}}(\nu_n\|\mu)+\liminf_{n\to\infty}\delta\|\Gamma_n\|_{\mathcal{G}}^2 \\
&= \lim_{n\to\infty}\Big(D_{\mathrm{KL}}(\nu_n\|\mu)+\delta\|\Gamma_n\|_{\mathcal{G}}^2\Big) \\
&= \inf_{\nu\in\mathcal{A}}\Big(D_{\mathrm{KL}}(\nu\|\mu)+\delta\|\Gamma\|_{\mathcal{G}}^2\Big).
\end{aligned}$$

Hence we can deduce using the lower semicontinuity of $\Gamma \mapsto \|\Gamma\|_{\mathcal{G}}^2$ with respect to weak convergence in $\mathcal{G}$ that

$$\begin{aligned}
\limsup_{n\to\infty} D_{\mathrm{KL}}(\nu_n\|\mu) &\leq \lim_{n\to\infty}\Big(D_{\mathrm{KL}}(\nu_n\|\mu)+\delta\|\Gamma_n\|_{\mathcal{G}}^2\Big)-\liminf_{n\to\infty}\delta\|\Gamma_n\|_{\mathcal{G}}^2 \\
&\leq \Big(D_{\mathrm{KL}}(\nu_\star\|\mu)+\delta\|\Gamma_\star\|_{\mathcal{G}}^2\Big)-\delta\|\Gamma_\star\|_{\mathcal{G}}^2 \\
&= D_{\mathrm{KL}}(\nu_\star\|\mu),
\end{aligned}$$

which implies that $\lim_{n\to\infty} D_{\mathrm{KL}}(\nu_n\|\mu)=D_{\mathrm{KL}}(\nu_\star\|\mu)$. In the same way it follows that $\lim_{n\to\infty}\|\Gamma_n\|_{\mathcal{G}}^2=\|\Gamma_\star\|_{\mathcal{G}}^2$. By Lemma 2.4 we can conclude that $\|\nu_n-\nu_\star\|_{\mathrm{tv}}\to 0$. For the operators Γ_n we note that weak convergence together with convergence of the norm implies strong convergence. □

Example 3.11. The first example we have in mind is the case where, as in Example 3.8, $\mathcal{H}=L^2([-1,1])$, C_0^{-1} is the negative Dirichlet–Laplace operator on $[-1,1]$, $\mathcal{H}^1=H_0^1$, and $m_0=0$. Thus the reference measure is the distribution of a centered Brownian bridge. By a slight adaptation of the proof of [18, Theorem 6.16] we have that, for $p\in(2,\infty]$, $\|u\|_{L^p}\leq C\|u\|_{\mathcal{H}^s}$ for all $s>\frac{1}{2}-\frac{1}{p}$, and we will use this fact in what follows. For Γ we choose multiplication operators with suitable functions $\hat{\Gamma}\colon[-1,1]\to\mathbb{R}$. For any $r>0$ we denote by $\mathcal{G}^r$ the space of multiplication operators with functions $\hat{\Gamma}\in H^r([-1,1])$ endowed with the Hilbert space structure of $H^r([-1,1])$. In this notation, the compact embedding of the spaces $H^r([-1,1])$ into $L^2([-1,1])$ can be rephrased as a compact embedding of the space $\mathcal{G}^r$ into the space $\mathcal{G}^0$, i.e., the space of $L^2([-1,1])$ functions, viewed as multiplication operators. By the form of Sobolev embedding stated above we have that for $\kappa < \frac{3}{4}$ and any[3] $x \in \mathcal{H}^{1-\kappa}$

$$\langle x, \Gamma x\rangle = \int_{-1}^{1} \hat{\Gamma}(t) x(t)^2 dt \le \|\hat{\Gamma}\|_{L^2([-1,1])} \|x\|_{L^4}^2 \lesssim \|\hat{\Gamma}\|_{L^2([-1,1])} \|x\|_{\mathcal{H}^{1-\kappa}}^2. \tag{3.14}$$

Since this shows that

$$\|\Gamma\|_{\mathcal{L}(\mathcal{H}^{1-\kappa},\mathcal{H}^{-(1-\kappa)})} \lesssim \|\hat{\Gamma}\|_{L^2([-1,1])},$$

it demonstrates that $\mathcal{G}_0$ embeds continuously into the space $\mathcal{L}(\mathcal{H}^{1-\kappa}, \mathcal{H}^{-(1-\kappa)})$, and hence the spaces $\mathcal{G}^r$, which are compact in $\mathcal{G}_0$, satisfy the assumption of Theorem 3.10 for any $r > 0$.

Example 3.12. Now consider μ_0 to be a Gaussian field over a space of dimension 2 or more. In this case we need to take a covariance operator that has a stronger regularizing property than the inverse Laplace operator. For example, if we denote by Δ the Laplace operator on the n-dimensional torus $\mathbb{T}^n$, then the Gaussian field with covariance operator $C_0 = (-\Delta + I)^{-s}$ takes values in $L^2(\mathbb{T}^n)$ if and only if $s > \frac{n}{2}$. In this case, the space $\mathcal{H}^1$ coincides with the fractional Sobolev space $H^s(\mathbb{T}^n)$. Note that the condition $s > \frac{n}{2}$ precisely implies that there exists a $\kappa > 0$ such that the space $\mathcal{H}^{1-\kappa}$ embeds into $L^\infty(\mathbb{T}^n)$ and in particular into $L^4[0,T]$. As above, denote by $\mathcal{G}^r$ the space of multiplication operators on $L^2(\mathbb{T}^n)$ with functions $\hat{\Gamma} \in H^r(\mathbb{T}^n)$. Then the same calculation as (3.14) shows that the conditions of Theorem 3.10 are satisfied for any $r > 0$.

3.4. Uniqueness of minimizers.

As stated above in Proposition 2.3, the minimization problem (2.2) has a unique minimizer if the set $\mathcal{A}$ is convex. Unfortunately, in all of the situations discussed in this section, $\mathcal{A}$ is not convex, so this criterion does not apply.

In general we do not expect minimizers to be unique; the example in subsection 2.2 illustrates nonuniqueness. There is, however, one situation in which we have uniqueness for all of the choices of $\mathcal{A}$ discussed in Theorem 3.1, namely the case where instead of $\mathcal{A}$ the measure μ satisfies a convexity property. Let us first recall the definition of λ-convexity.

DEFINITION 3.13. *Let $\Phi\colon \mathcal{H}^1 \to \mathbb{R}$ be a function. For a $\lambda \in \mathbb{R}$ the function Φ is* λ-convex *with respect to $\mathcal{H}^1$ if*

$$\mathcal{H}^1 \ni x \mapsto \frac{\lambda}{2}\langle x, x\rangle_{\mathcal{H}^1} + \Phi(x) \tag{3.15}$$

is convex on $\mathcal{H}^1$.

Remark 3.14. Equation (3.15) implies that for any $x_1, x_2 \in \mathcal{H}^1$ and for any $t \in (0,1)$ we have

$$\Phi((1-t)x_1 + tx_2) \le (1-t)\Phi(x_1) + t\Phi(x_2) + \lambda\frac{t(1-t)}{2}\|x_1 - x_2\|_{\mathcal{H}^1}^2. \tag{3.16}$$

Equation (3.16) is often taken to define λ-convexity because it gives useful estimates even when the distance function does not come from a scalar product. For Hilbert spaces both definitions are equivalent.

[3] Throughout the paper we write $a \lesssim b$ to indicate that there exists a constant $c > 0$ independent of the relevant quantities such that $a \le cb$.

The following theorem implies uniqueness for the minimization problem (2.2) as soon as Φ is $(1-\kappa)$-convex for a $\kappa > 0$ and satisfies a mild integrability property. The proof is given in section 4.

Theorem 3.15. *Let μ be as in* (3.1), *and assume that there exists a $\kappa > 0$ such that Φ is $(1-\kappa)$-convex with respect to $\mathcal{H}^1$. Assume that there exist constants $0 < c_i < \infty$, $i = 1, 2, 3$, and $\alpha \in (0, 2)$ such that for every $x \in X$ we have*

$$-c_1\|x\|_X^\alpha \leq \Phi(x) \leq c_2 \exp\big(c_3\|x\|_X^\alpha\big). \tag{3.17}$$

Let $\nu_1 = N(m_1, C_1)$ and $\nu_2 = N(m_2, C_2)$ be Gaussian measures with $D_{\mathrm{KL}}(\nu_1\|\mu) < \infty$ and $D_{\mathrm{KL}}(\nu_2\|\mu) < \infty$. For any $t \in (0, 1)$ there exists an interpolated measure $\nu_t^{1\to 2} = N(m_t, C_t)$ which satisfies $D_{\mathrm{KL}}(\nu_t^{1\to 2}\|\mu) < \infty$. Furthermore, as soon as $\nu_1 \neq \nu_2$ there exists a constant $K > 0$ such that for all $t \in (0, 1)$

$$D_{\mathrm{KL}}(\nu_t^{1\to 2}\|\mu) \leq (1-t)D_{\mathrm{KL}}(\nu_1\|\mu) + tD_{\mathrm{KL}}(\nu_1\|\mu) - \frac{t(1-t)}{2}K.$$

Finally, if we have $m_1 = m_2$, then $m_t = m_1$ holds as well for all $t \in (0, 1)$, and in the same way, if $C_1 = C_2$, then $C_t = C_1$ for all $t \in (0, 1)$.

The measures $\nu_t^{1\to 2}$ introduced in Theorem 3.15 are a special case of geodesics on Wasserstein space first introduced in [26] in a finite dimensional situation. In addition, the proof shows that the constant K appearing in the statement is κ times the square of the Wasserstein distance between ν_1 and ν_2 with respect to the $\mathcal{H}^1$ norm. See [1, 16] for a more detailed discussion of mass transportation on infinite dimensional spaces. The following is an immediate consequence of Theorem 3.15.

Corollary 3.16. *Assume that μ is a probability measure given by* (3.1), *that there exists a $\kappa > 0$ such that Φ is $(1-\kappa)$ convex with respect to $\mathcal{H}^1$, and that Φ satisfies the bound* (3.17). *Then for any of the four choices of sets $\mathcal{A}_i$ discussed in Theorem* 3.1 *the minimizer of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ is unique in $\mathcal{A}_i$.*

Remark 3.17. The assumption that Φ is $(1-\kappa)$-convex for a $\kappa > 0$ implies in particular that μ is *log-concave* (see [1, Definition 9.4.9]). It can be viewed as a quantification of this log-concavity.

Example 3.18. As in Examples 3.8 and 3.9 above, let μ_0 be a centered Brownian bridge on $[-\frac{L}{2}, \frac{L}{2}]$. As above we have $\mathcal{H}^1 = H_0^1([-\frac{L}{2}, \frac{L}{2}])$ equipped with the *homogeneous* Sobolev norm and $X = C([-\frac{L}{2}, \frac{L}{2}])$.

For some $\mathcal{C}^2$ function $\phi\colon \mathbb{R} \to \mathbb{R}_+$ set $\Phi\big(x(\cdot)\big) = \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi(x(s))\, ds$. The integrability condition (3.17) translates immediately into the growth condition $-c_1'|x|^\alpha \leq \phi(x) \leq c_2' \exp(c_3'|x|^\alpha)$ for $x \in \mathbb{R}$ and constants $0 < c_i' < \infty$ for $i = 1, 2, 3$. Of course, the convexity assumption of Theorem 3.15 is satisfied if ϕ is convex. But we can allow for some nonconvexity. For example, if $\phi \in \mathcal{C}^2(\mathbb{R})$ and ϕ'' is uniformly bounded from below by $-K \in \mathbb{R}$, then we get for $x_1, x_2 \in \mathcal{H}^1$

$$\begin{aligned}
&\Phi((1-t)x_1 + tx_2\big)\\
&\quad = \int_{-\frac{L}{2}}^{\frac{L}{2}} \phi\big((1-t)x_1(s) + tx_2(s)\big)\, ds\\
&\quad \leq \int_{-\frac{L}{2}}^{\frac{L}{2}} (1-t)\phi\big((x_1(s)\big) + t\phi\big(x_2(s)\big) + \frac{1}{2}t(1-t)K\big|x_1(s) - x_2(s)\big|^2\, ds\\
&\quad = (1-t)\Phi(x_1) + t\Phi(x_2) + \frac{Kt(1-t)}{2}\int_{-\frac{L}{2}}^{\frac{L}{2}} \big|x_1(s) - x_2(s)\big|^2 ds.
\end{aligned}$$

Using the estimate

$$\int_{-\frac{L}{2}}^{\frac{L}{2}} \left|x_1(s) - x_2(s)\right|^2 ds \le \left(\frac{L}{\pi}\right)^2 \|x_1 - x_2\|_{\mathcal{H}^1}^2$$

we see that Φ satisfies the convexity assumption as soon as $K < \left(\frac{\pi}{L}\right)^2$. The proof of Theorem 3.15 is based on the influential concept of displacement convexity, introduced by McCann in [26], and heavily inspired by the infinite dimensional exposition in [1]. It can be found in subsection 4.3.

3.5. Gaussian mixtures. We have demonstrated a methodology for approximating measure μ given by (3.1) by a Gaussian ν. If μ is multimodal, then this approximation can result in several local minimizers centered on the different modes. A potential way to capture all modes at once is to use Gaussian mixtures, as explained in the finite dimensional setting in [6]. We explore this possibility in our infinite dimensional context: in this subsection we show existence of minimizers for problem (2.2) in the situation when we are minimizing over a set of convex combinations of Gaussian measures.

We start with a basic lemma for which we do not need to assume that the mixture measure comprises Gaussians.

LEMMA 3.19. *Let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{M}(\mathcal{H})$ be closed under weak convergence of probability measures. Then so is*

$$\mathcal{C} := \{\mu := p^1\nu^1 + p^2\nu^2 \colon 0 \le p^i \le 1,\, i = 1, 2; \quad p^1 + p^2 = 1;\, \nu^1 \in \mathcal{A};\, \nu^2 \in \mathcal{B}\}.$$

Proof. Let $(\nu_n) = (p_n^1\nu_n^1 + p_n^2\nu_n^2)$ be a sequence of measures in $\mathcal{C}$ that converges weakly to $\mu_\star \in \mathcal{M}(\mathcal{H})$. We want to show that $\mu_\star \in \mathcal{C}$. It suffices to show that a subsequence of the ν_n converges to an element in $\mathcal{C}$. After passing to a subsequence we can assume that for $i = 1, 2$ the p_n^i converge to $p_\star^i \in [0, 1]$ with $p_\star^1 + p_\star^2 = 1$. Let us first treat the case where one of these $p_\star^i$ is zero, say $p_\star^1 = 0$ and $p_\star^2 = 1$. In this situation we can conclude that the ν_n^2 converge weakly to $\mu_\star$ and hence $\mu_\star \in \mathcal{B} \subseteq \mathcal{C}$. Therefore, we can assume $p_\star^i \in (0, 1)$. After passing to another subsequence we can furthermore assume that the p_n^i are uniformly bounded from below by a positive constant $\hat{p} > 0$. As the sequence ν_n converges weakly in $\mathcal{M}(\mathcal{H})$, it is tight. We claim that this implies automatically the tightness of the sequences ν_n^i. Indeed, for a $\delta > 0$ let $K_\delta \subseteq \mathcal{H}$ be a compact set with $\nu_n(K_\delta) \le \delta$ for any $n \ge 1$. Then we have for any n and for $i = 1, 2$ that

$$\nu_n^i(K_\delta) \le \frac{1}{\hat{p}}\nu(K_\delta) \le \frac{\delta}{\hat{p}}.$$

After passing to yet another subsequence, we can assume that the ν_n^1 converge weakly to $\nu_\star^1 \in \mathcal{A}$ and the ν_n^2 converge weakly to $\nu_\star^2 \in \mathcal{B}$. In particular, along this subsequence the ν_n converge weakly to $p_\star^1\nu_\star^1 + p_\star^2\nu_\star^2 \in \mathcal{C}$. □

By a simple recursion, Lemma 3.19 extends immediately to sets $\mathcal{C}$ of the form

$$\tilde{\mathcal{C}} := \left\{\nu := \sum_{i=1}^{N} p^i\nu^i \colon 0 \le p^i \le 1,\, \sum_{i=1}^{N} p^i = 1,\, \nu^i \in \mathcal{A}_i\right\}$$

for fixed N and sets $\mathcal{A}_i$ that are all closed under weak convergence of probability measures. Hence we get the following consequence from Corollary 2.2 and Lemma B.1.

THEOREM 3.20. *Let μ_0 be a Gaussian measure with mean $m_0 \in \mathcal{H}$ and covariance operator $C_0 \in \mathcal{TC}(\mathcal{H})$, and let μ be given by* (3.1)*. For any fixed N and for any choice* *of set $\mathcal{A}$ as in Theorem* 3.1 *consider the following choice for $\mathcal{C}$:*

$$\mathcal{C} := \left\{ \mu := \sum_{i=1}^{N} p^i \nu^i \colon 0 \le p^i \le 1,\, \sum_{i=1}^{N} p^i = 1,\, \nu^i \in \mathcal{A} \right\}.$$

Then as soon as there exists a single $\nu \in \mathcal{A}$ with $D_{\mathrm{KL}}(\nu\|\mu) < \infty$ there exists a minimizer of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ in $\mathcal{C}$. This minimizer ν is necessarily equivalent to μ_0 in the sense of measures.

4. Proofs of main results.

Here we gather the proofs of various results used in this paper. The proofs may be of independent interest, but their inclusion in the main text would break from the flow of ideas related to Kullback–Leibler minimization.

4.1. Proof of Lemma 2.4.

The following "parallelogram identity" (see [10, equation (2.2)]) is easy to check: for any n, m

$$\begin{aligned} &D_{\mathrm{KL}}(\nu_n\|\mu) + D_{\mathrm{KL}}(\nu_m\|\mu) \\ &\quad = 2D_{\mathrm{KL}}\left(\frac{\nu_n+\nu_m}{2}\middle\|\mu\right) + D_{\mathrm{KL}}\left(\nu_n\middle\|\frac{\nu_n+\nu_m}{2}\right) + D_{\mathrm{KL}}\left(\nu_m\middle\|\frac{\nu_n+\nu_m}{2}\right). \end{aligned} \tag{4.1}$$

By assumption the left-hand side of (4.1) converges to $2D_{\mathrm{KL}}(\nu_\star\|\mu)$ as $n, m \to \infty$. Furthermore, the measure $1/2(\nu_n + \nu_m)$ converges weakly to $\nu_\star$ as $n, m \to \infty$ and by the lower semicontinuity of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ we have

$$\liminf_{n,m\to\infty} 2D_{\mathrm{KL}}\left(\frac{\nu_n+\nu_m}{2}\middle\|\mu\right) \ge 2D_{\mathrm{KL}}(\nu_\star\|\mu).$$

By the nonnegativity of D_{KL} this implies that

$$D_{\mathrm{KL}}\left(\nu_m\middle\|\frac{\nu_n+\nu_m}{2}\right) \to 0 \quad \text{and} \quad D_{\mathrm{KL}}\left(\nu_n\middle\|\frac{\nu_n+\nu_m}{2}\right) \to 0. \tag{4.2}$$

As we can write

$$\|\nu_n - \nu_m\|_{\mathrm{tv}} \le \left\|\nu_n - \frac{\nu_n+\nu_m}{2}\right\|_{\mathrm{tv}} + \left\|\nu_m - \frac{\nu_n+\nu_m}{2}\right\|_{\mathrm{tv}},$$

equations (4.2) and the Pinsker inequality

$$\|\nu - \mu\|_{\mathrm{tv}} \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\nu\|\mu)}$$

(a proof of which can be found in [9]) imply that the sequence is Cauchy with respect to the total variation norm. By assumption, the ν_n converge *weakly* to $\nu_\star$ and this implies convergence in total variation norm.

4.2. Proof of Lemma 3.3.

Recall $(e_\alpha,\, \lambda_\alpha, \alpha \ge 1)$, the eigenfunction/eigenvalue pairs of C_0 introduced above (3.2). For any α, β we write

$$\Gamma_{\alpha,\beta} = \langle e_\alpha, \Gamma e_\beta \rangle.$$

Then (3.6) states that

$$\sum_{1\le\alpha,\beta<\infty} \lambda_\alpha\, \lambda_\beta \Gamma_{\alpha,\beta}^2 < \infty.$$

Define $\mathbb{N}_0 = \mathbb{N}^2 \setminus \{1, \dots, N_0\}^2$. Then the preceding display implies that for any $\delta > 0$ there exists an $N_0 \geq 0$ such that

$$\sum_{(\alpha,\beta)\in\mathbb{N}_0} \lambda_\alpha \, \lambda_\beta \Gamma_{\alpha,\beta}^2 < \delta^2. \tag{4.3}$$

This implies that for $x = \sum_\alpha x_\alpha e_\alpha \in \mathcal{H}^1$ we get

$$\begin{aligned} \langle x, \Gamma x\rangle &= \sum_{1\leq\alpha,\beta<\infty} \Gamma_{\alpha,\beta} x_\alpha x_\beta \\ &= \sum_{1\leq\alpha,\beta\leq N_0} \Gamma_{\alpha,\beta} x_\alpha x_\beta + \sum_{(\alpha,\beta)\in\mathbb{N}_0} \Gamma_{\alpha,\beta} x_\alpha x_\beta. \end{aligned} \tag{4.4}$$

The first term on the right-hand side of (4.4) can be bounded by

$$\left| \sum_{1\leq\alpha,\beta\leq N_0} \Gamma_{\alpha,\beta} x_\alpha x_\beta \right| \leq \max_{1\leq\alpha,\beta\leq N_0} \left|\Gamma_{\alpha,\beta}\right| \|x\|_{\mathcal{H}}^2. \tag{4.5}$$

For the second term we get using the Cauchy–Schwarz inequality and (4.3)

$$\begin{aligned} \left| \sum_{(\alpha,\beta)\in\mathbb{N}_0} \Gamma_{\alpha,\beta} x_\alpha x_\beta \right| &= \left| \sum_{(\alpha,\beta)\in\mathbb{N}_0} \sqrt{\lambda_\alpha\lambda_\beta}\Gamma_{\alpha,\beta} \, \frac{x_\alpha x_\beta}{\sqrt{\lambda_\alpha\lambda_\beta}} \right| \\ &\leq \delta \langle x, C_0^{-1} x\rangle. \end{aligned} \tag{4.6}$$

We can conclude from (4.4), (4.5), and (4.6) that Γ is infinitesimally form-bounded with respect to C_0^{-1} (see, e.g.. [30, Chapter X.2]). In particular, by the KLMN theorem (see [30, Theorem X.17]) the form Q_Γ is bounded from below and closed, and there exists a unique self-adjoint operator denoted by $C_0^{-1} + \Gamma$ with form domain $\mathcal{H}^1$ that generates Q_Γ.

If Q_Γ is strictly positive, then so is $C_0^{-1} + \Gamma$ and its inverse $(C_0^{-1} + \Gamma)^{-1}$. As $C_0^{-1} + \Gamma$ has form domain $\mathcal{H}^1$, the operator $(C_0^{-1} + \Gamma)^{-\frac{1}{2}} C_0^{-\frac{1}{2}}$ is bounded on $\mathcal{H}$ by the closed graph theorem and it follows that, as the composition of a trace-class operator with two bounded operators,

$$(C_0^{-1} + \Gamma)^{-1} = \big((C_0^{-1} + \Gamma)^{-\frac{1}{2}} C_0^{-\frac{1}{2}}\big) C_0 \big((C_0^{-1} + \Gamma)^{-\frac{1}{2}} C_0^{-\frac{1}{2}}\big)^\star$$

is a trace-class operator. It is hence the covariance operator of a centered Gaussian measure on $\mathcal{H}$. It satisfies the conditions of the Feldman–Hajek theorem by assumption.

If Q_Γ is not strictly positive, then the intersection of the spectrum of $C_0^{-1} + \Gamma$ with $(-\infty, 0]$ is not empty, and hence it cannot be the inverse covariance of a Gaussian measure.

4.3. Proof of Theorem 3.15.

We start the proof of Theorem 3.15 with the following lemma.

Lemma 4.1. *Let $\nu = N(m, C)$ be equivalent to μ_0. For any $\gamma \geq 1$ let $\pi_\gamma \colon \mathcal{H} \to \mathcal{H}$ be the orthogonal projector on the space $\mathcal{H}_\gamma$ introduced in* (3.2). *Furthermore, assume that $\Phi \colon X \to \mathbb{R}_+$ satisfies the second inequality in* (3.17). *Then we have*

$$\lim_{\gamma\to\infty} \mathbb{E}^\nu\big[\Phi(\pi_\gamma x)\big] = \mathbb{E}^\nu\big[\Phi(x)\big]. \tag{4.7}$$

Proof. It is a well-known property of the white noise/Karhunen–Loève expansion (see, e.g., [11, Theorem 2.12]) that $\|\pi_\gamma x - x\|_X \to 0$ μ_0-almost surely and, as ν is equivalent to μ_0, also ν-almost surely. Hence, by continuity of Φ on X, $\Phi(\pi_\gamma x)$ converges ν-almost surely to $\Phi(x)$.

As $\nu(X) = 1$, there exists a constant $0 < K_\infty < \infty$ such that $\nu(\|x\|_X \geq K_\infty) \leq \frac{1}{8}$. On the other hand, by the ν-almost sure convergence of $\|\pi_\gamma x - x\|_X$ to 0 there exists a $\gamma_\infty \geq 1$ such that for all $\gamma > \gamma_\infty$ we have $\nu\big(\|\pi_\gamma x - x\|_X \geq 1\big) \leq \frac{1}{8}$, which implies that

$$\nu\big(\|\pi_\gamma x\| \geq K_\infty + 1\big) \leq \frac{1}{4} \qquad \text{for all } \gamma \geq \gamma_\infty.$$

For any $\gamma \leq \gamma_\infty$ there exists another $0 < K_\gamma < \infty$ such that $\nu\big(\|\pi_\gamma x\| \geq K_\gamma\big) \leq \frac{1}{4}$, and hence if we set $K = \max\{K_1, \ldots, K_{\gamma_\infty}, K_\infty + 1\}$, we get

$$\nu\big(\|\pi_\gamma x\| \geq K\big) \leq \frac{1}{4} \qquad \text{for all } \gamma \geq 1.$$

By Fernique's theorem (see, e.g., [11, Theorem 2.6]) this implies the existence of a $\lambda > 0$ such that

$$\sup_{\gamma \geq 1} \mathbb{E}^\nu \big[\exp\big(\lambda \|\pi_\gamma x\|^2\big)\big] < \infty.$$

Then the desired statement (4.7) follows from the dominated convergence theorem observing that (3.17) implies the pointwise bound

$$\Phi(x) \leq c_2 \exp(c_3 \|x\|_X^\alpha) \leq c_4 \exp(\lambda \|x\|_X^2) \tag{4.8}$$

for $0 < c_4 < \infty$ sufficiently large. □

Let us also recall the following property.

Proposition 4.2 (see [1, Lemma 9.4.5]). *Let $\mu, \nu \in \mathcal{M}(\mathcal{H})$ be a pair of arbitrary probability measures on $\mathcal{H}$, and let $\pi \colon \mathcal{H} \to \mathcal{H}$ be a measurable mapping. Then we have*

$$D_{\mathrm{KL}}(\nu \circ \pi^{-1} \| \mu \circ \pi^{-1}) \leq D_{\mathrm{KL}}(\nu \| \mu). \tag{4.9}$$

Proof of Theorem 3.15. As above (3.2), let $(e_\alpha, \alpha \geq 1)$ be the basis $\mathcal{H}$ consisting of eigenvalues of C_0 with the corresponding eigenvalues $(\lambda_\alpha, \alpha \geq 1)$. For $\gamma \geq 1$ let $\pi_\gamma \colon \mathcal{H} \to \mathcal{H}$ be the orthogonal projection on $\mathcal{H}_\gamma := \mathrm{span}(e_1, \ldots, e_\gamma)$. Furthermore, for $\alpha \geq 1$ and $x \in \mathcal{H}$ let $\xi_\alpha(x) = \langle x, e_\alpha \rangle_{\mathcal{H}}$. Then we can identify $\mathcal{H}_\gamma$ with $\mathbb{R}^\gamma$ through the bijection

$$\mathbb{R}^\gamma \ni \Xi_\gamma = (\xi_1, \ldots, \xi_\gamma) \mapsto \sum_{\alpha=1}^{\gamma} \xi_\alpha e_\alpha. \tag{4.10}$$

The identification (4.10) in particular gives a natural way to define the γ-dimensional Lebesgue measure $\mathcal{L}^\gamma$ on $\mathcal{H}_\gamma$.

Denote by $\mu_{0;\gamma} = \mu_0 \circ \pi_\gamma^{-1}$ the projection of μ_0 on $\mathcal{H}_\gamma$. We also define μ_γ by

$$\frac{d\mu_\gamma}{d\mu_{0;\gamma}}(x) = \frac{1}{Z_\gamma} \exp\big(-\Phi(x)\big),$$

where $Z_\gamma = \mathbb{E}^{\mu_{0,\gamma}}\big[\exp\big(-\Phi(x)\big)\big]$. Note that in general μ_γ does not coincide with the measure $\mu \circ \pi_\gamma$. The Radon–Nikodym density of μ_γ with respect to $\mathcal{L}^\gamma$ is given by

$$\frac{d\mu_\gamma}{d\mathcal{L}^\gamma}(x) = \frac{1}{\tilde{Z}_\gamma}\exp\big(-\Psi(x)\big),$$

where $\Psi(x) = \Phi(x) + \frac{1}{2}\langle x, x\rangle_{\mathcal{H}^1}$ and the normalization constant is given by

$$\tilde{Z}_\gamma = Z_\gamma (2\pi)^{\frac{\gamma}{2}} \prod_{\alpha=1}^{\gamma} \sqrt{\lambda_\alpha}.$$

According to the assumption the function $\Psi(x) - \frac{\kappa}{2}\langle x, x\rangle_{\mathcal{H}^1}$ is convex on $\mathcal{H}_\gamma$, which implies that for any $x_1, x_2 \in \mathcal{H}_\gamma$ and for $t \in [0,1]$ we have

$$\Psi\big((1-t)x_1 + tx_2\big) \leq (1-t)\Psi(x_1) + t\Psi(x_2) - \kappa\frac{t(1-t)}{2}\|x_1 - x_2\|^2_{\mathcal{H}^1}.$$

Let us also define the projected measures $\nu_{i;\gamma} := \nu_i \circ \pi_\gamma^{-1}$ for $i = 1, 2$. By assumption the measures ν_i are equivalent to μ_0 and therefore the projections $\nu_{i;\gamma}$ are equivalent to $\mu_{0;\gamma}$. In particular, the $\nu_{i;\gamma}$ are nondegenerate Gaussian measures on $\mathcal{H}_\gamma$. Their covariance operators are given by $C_{i;\gamma} := \pi_\gamma C_i \pi_\gamma$ and the means by $m_{i;\gamma} = \pi_\gamma m_i$.

There is a convenient coupling between the $\nu_{i;\gamma}$. Indeed, set

$$\Lambda_\gamma = C_{2;\gamma}^{\frac{1}{2}}\big(C_{2;\gamma}^{\frac{1}{2}} C_{1;\gamma} C_{2,\gamma}^{\frac{1}{2}}\big)^{-\frac{1}{2}} C_{2;\gamma}^{\frac{1}{2}} \in \mathcal{L}(\mathcal{H}_\gamma, \mathcal{H}_\gamma). \tag{4.11}$$

The operator Λ_γ is symmetric and strictly positive on $\mathcal{H}_\gamma$. Then define for $x \in \mathcal{H}_\gamma$

$$\tilde{\Lambda}_\gamma(x) := \Lambda_\gamma(x - m_{1;\gamma}) + m_{2;\gamma}. \tag{4.12}$$

Clearly, if $x \sim \nu_{1,\gamma}$, then $\tilde{\Lambda}_\gamma(x) \sim \nu_{2,\gamma}$. Now for any $t \in (0,1)$ we define the interpolation $\tilde{\Lambda}_{\gamma,t}(x) = (1-t)x + t\tilde{\Lambda}_\gamma(x)$ and the approximate interpolating measures $\nu_{t;\gamma}^{1\to2}$ for $t \in (0,1)$ as push-forward measures

$$\nu_{t;\gamma}^{1\to2} := \nu_{1,\gamma} \circ \tilde{\Lambda}_{\gamma,t}^{-1}. \tag{4.13}$$

From the construction it follows that the $\nu_{t;\gamma}^{1\to2} = N(m_{t,\gamma}, C_{t,\gamma})$ are nondegenerate Gaussian measures on $\mathcal{H}_\gamma$. Furthermore, if the means m_1 and m_2 coincide, then we have $m_{1,\gamma} = m_{2,\gamma} = m_{t,\gamma}$ for all $t \in (0,1)$, and in the same way, if the covariance operators C_1 and C_2 coincide, then we have $C_{1,\gamma} = C_{2,\gamma} = C_{t,\gamma}$ for all $t \in (0,1)$.

As a next step we will establish that for any γ the function

$$t \mapsto D_{\mathrm{KL}}(\nu_{t;\gamma}^{1\to2} \| \mu_\gamma)$$

is convex. To this end it is useful to write

$$D_{\mathrm{KL}}(\nu_{t;\gamma}^{1\to2} \| \mu_\gamma) = \mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to2}) + \mathscr{F}_\gamma(\nu_{t;\gamma}^{1\to2}) + \log(\tilde{Z}_\gamma), \tag{4.14}$$

where $\mathscr{F}_\gamma(\nu_{t;\gamma}^{1\to2}) = \mathbb{E}^{\nu_{t;\gamma}^{1\to2}}\big[\Psi(x)\big]$ and

$$\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to2}) = \int_{\mathcal{H}_\gamma} \frac{d\nu_{t;\gamma}^{1\to2}}{d\mathcal{L}^\gamma}(x) \log\left(\frac{d\nu_{t;\gamma}^{1\to2}}{d\mathcal{L}^\gamma}(x)\right) d\mathcal{L}^\gamma(x).$$

Note that $\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to 2})$ is completely independent of the measure μ_0. Also note that $\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to 2})$, the entropy of $\nu_{t;\gamma}^{1\to 2}$, can be negative because the Lebesgue measure is not a probability measure.

We will treat the terms $\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to 2})$ and $F_\gamma(\nu_{t;\gamma}^{1\to 2})$ separately. The treatment of F_γ is straightforward using the $(-\kappa)$-convexity of Ψ and the coupling described above. Indeed, we can write

$$
\begin{aligned}
\mathscr{F}_\gamma(\nu_{t;\gamma}^{1\to 2}) &= \mathbb{E}^{\nu_{t;\gamma}^{1\to 2}}\big[\Psi(x)\big] \\
&= \mathbb{E}^{\nu_{1,\gamma}}\big[\Psi\big((1-t)x + t\tilde{\Lambda}_\gamma(x)\big)\big] \\
&\le (1-t)\mathbb{E}^{\nu_{1,\gamma}}\big[\Psi(x)\big] + t\mathbb{E}^{\nu_{1,\gamma}}\big[\Psi(\tilde{\Lambda}_\gamma(x))\big] - \kappa\frac{t(1-t)}{2}\mathbb{E}^{\nu_{1,\gamma}}\|x - \tilde{\Lambda}_\gamma(x)\|_{\mathcal{H}^1}^2 \\
&\le (1-t)\mathscr{F}_\gamma\big(\nu_{1,\gamma}\big) + t\mathscr{F}_\gamma\big(\nu_{2,\gamma}\big) - \kappa\frac{t(1-t)}{2}\mathbb{E}^{\nu_{1,\gamma}}\|x - \tilde{\Lambda}_\gamma(x)\|_{\mathcal{H}^1}^2. \qquad (4.15)
\end{aligned}
$$

Note that this argument does not make use of any specific properties of the mapping $x \mapsto \tilde{\Lambda}_\gamma(x)$, except that it maps $\mu_{1;\gamma}$ to $\mu_{2;\gamma}$. The same argument would work for different mappings with this property.

To show the convexity of the functional $\mathscr{H}_\gamma$ we will make use of the fact that the matrix Λ_γ is symmetric and strictly positive. For convenience, we introduce the notation

$$
\rho(x) = \frac{\nu_{1;\gamma}}{d\mathcal{L}^\gamma}(x), \qquad \rho_t(x) := \frac{d\nu_{t;\gamma}^{1\to 2}}{d\mathcal{L}^\gamma}(x).
$$

Furthermore, for the moment we write $F(\rho) = \rho\log(\rho)$. By the change of variables formula we have

$$
\rho_t(\tilde{\Lambda}_\gamma(x)) = \frac{\rho(x)}{\det\big((1-t)\operatorname{Id}_\gamma + t\Lambda_\gamma\big)},
$$

where we denote by Id_γ the identity matrix on $\mathbb{R}^\gamma$. Hence we can write

$$
\begin{aligned}
\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to 2}) &= \int_{\mathcal{H}_\gamma} F\big(\rho_t(x)\big)d\mathcal{L}^\gamma(x) \\
&= \int_{\mathcal{H}_\gamma} F\left(\frac{\rho(x)}{\det\big((1-t)\operatorname{Id}_\gamma + t\Lambda_\gamma\big)}\right)\det\big((1-t)\operatorname{Id}_\gamma + t\Lambda_\gamma\big)d\mathcal{L}^\gamma(x).
\end{aligned}
$$

For a diagonalizable matrix Λ with nonnegative eigenvalues the mapping $[0,1] \ni t \mapsto \det((1-t)\operatorname{Id} + t\Lambda_\gamma)^{\frac{1}{\gamma}}$ is concave, and as the map $s \mapsto F(\rho/s^d)s^d$ is nonincreasing, the resulting map is convex in t. Hence we get

$$
\begin{aligned}
\mathscr{H}_\gamma(\nu_{t;\gamma}^{1\to 2}) &\le (1-t)\int_{\mathcal{H}_\gamma} F\big(\rho(x)\big)d\mathcal{L}^\gamma(x) + t\int_{\mathcal{H}_\gamma} F\left(\frac{\rho(x)}{\Lambda_\gamma}\right)\det\big(\Lambda_\gamma\big)d\mathcal{L}^\gamma(x) \\
&= (1-t)\mathscr{H}_\gamma\big(\nu_{1;\gamma}\big) + tH_\gamma\big(\nu_{2;\gamma}\big). \qquad (4.16)
\end{aligned}
$$

Therefore, combining (4.14), (4.15), and (4.16) we obtain for any γ that

$$
\begin{aligned}
D_{\mathrm{KL}}\big(\nu_{t;\gamma}^{1\to 2}\big\|\mu_\gamma\big) \le (1-t)D_{\mathrm{KL}}\big(\nu_{1,\gamma}\big\|\mu_\gamma\big) + tD_{\mathrm{KL}}\big(\nu_{2,\gamma}\big\|\mu_\gamma\big)& \\
- \kappa\frac{t(1-t)}{2}\mathbb{E}^{\nu_{1,\gamma}}\|x - \tilde{\Lambda}_\gamma(x)\|_{\mathcal{H}^1}^2.& \qquad (4.17)
\end{aligned}
$$

It remains to pass to the limit $\gamma \to \infty$ in (4.17). First we establish that for $i = 1, 2$ we have $D_{\mathrm{KL}}\big(\nu_{i,\gamma} \big\| \mu_\gamma\big) \to D_{\mathrm{KL}}\big(\nu_i \big\| \mu\big)$. In order to see that we write

$$D_{\mathrm{KL}}\big(\nu_{i,\gamma} \big\| \mu_\gamma\big) = D_{\mathrm{KL}}\big(\nu_{i,\gamma} \big\| \mu_{0,\gamma}\big) + \mathbb{E}^{\nu_{i,\gamma}}\big[\Phi(x)\big] + \log(Z_\gamma), \tag{4.18}$$

and a similar identity holds for $D_{\mathrm{KL}}\big(\nu_i \big\| \mu\big)$. The Gaussian measures $\nu_{i,\gamma}$ and $\mu_{0,\gamma}$ are projections of the measures ν_i and μ_0, and hence they converge weakly as probability measures on $\mathcal{H}$ to these measures as $\gamma \to \infty$. Hence the lower semicontinuity of the Kullback–Leibler divergence (Proposition 2.1) implies that for $i = 1, 2$

$$\liminf_{\gamma\to\infty} D_{\mathrm{KL}}\big(\nu_{i,\gamma} \big\| \mu_\gamma\big) \geq D_{\mathrm{KL}}\big(\nu_i \big\| \mu_0\big).$$

On the other hand, the Kullback–Leibler divergence is monotone under projections (Proposition 4.2), and hence we get

$$\limsup_{\gamma\to\infty} D_{\mathrm{KL}}\big(\nu_{i,\gamma} \big\| \mu_\gamma\big) \leq D_{\mathrm{KL}}\big(\nu_i \big\| \mu_0\big),$$

which established the convergence of the first term in (4.18). The convergence of the $Z_\gamma = \mathbb{E}^{\mu_{0,\gamma}}\big[\exp\big(-\Phi(x)\big)\big]$ and of the $\mathbb{E}^{\nu_{i,\gamma}}\big[\Phi(x)\big]$ follow from Lemma 4.1 and the integrability assumption (3.17).

In order to pass to the limit $\gamma \to \infty$ on the left-hand side of (4.17) we note that for fixed $t \in (0, 1)$ the measures $\nu_{t;\gamma}^{1\to2}$ form a tight family of measures on $\mathcal{H}$. Indeed, by weak convergence the families of measures $\nu_{1,\gamma}$ and $\nu_{2,\gamma}$ are tight on $\mathcal{H}$. Hence, for every $\varepsilon > 0$ there exist compact in $\mathcal{H}$ sets K_1 and K_2 such that for $i = 1, 2$ and for any γ we have $\nu_{i,\gamma}(K_i^c) \leq \varepsilon$. For a fixed $t \in (0, 1)$, the set

$$K_t := \{x = (1-t)x_1 + tx_2 \colon \quad x_1 \in K_1,\, x_2 \in K_2\}$$

is compact in $\mathcal{H}$ and we have, using the definition of $\nu_{t;\gamma}^{1\to2}$, that

$$\nu_{t;\gamma}^{1\to2}(K_t^c) \leq \nu_{1,\gamma}(K_1^c) + \nu_{2,\gamma}(K_2^c) \leq 2\varepsilon,$$

which shows the tightness. Hence we can extract a subsequence that converges to a limit $\nu_t^{1\to2}$. This measure is Gaussian by Lemma B.1, and by construction its mean coincides with m_1 if $m_1 = m_2$, and in the same way its covariance coincides with C_1 if $C_1 = C_2$. By the lower semicontinuity of the Kullback–Leibler divergence (Proposition 2.1) we get

$$D_{\mathrm{KL}}\big(\nu_t^{1\to2} \big\| \mu\big) \leq \liminf_{\gamma\to\infty} D_{\mathrm{KL}}\big(\nu_{t;\gamma}^{1\to2} \big\| \mu_\gamma\big). \tag{4.19}$$

Finally, we have

$$\limsup_{\gamma\to\infty} \mathbb{E}^{\nu_{1,\gamma}} \|x - \tilde{\Lambda}_\gamma(x)\|_{\mathcal{H}^1}^2 := K > 0. \tag{4.20}$$

In order to see this note that the measures $\rho_\gamma := \nu_{1,\gamma}[\mathrm{Id} + \tilde{\Lambda}_\gamma]^{-1}$ form a tight family of measures on $\mathcal{H} \times \mathcal{H}$. Denote by ρ a limiting measure. This measure is a coupling of ν_1 and ν_2, and hence if these measures do not coincide, we have

$$\mathbb{E}^\rho \|x - y\|_{\mathcal{H}^1}^2 > 0.$$

Hence, the desired estimate (4.20) follows from Fatou's lemma. This finishes the proof. ▯

Appendix A. Proof of Proposition 2.1. For completeness we give a proof of the well-known proposition, Proposition 2.1, following closely the exposition in [14, Lemma 1.4.2]; see also [1, Lemma 9.4.3].

We start by recalling the Donsker–Varadhan variational formula

$$D_{\mathrm{KL}}(\nu\|\mu) = \sup_{\Theta} \mathbb{E}^{\nu}\Theta - \log \mathbb{E}^{\mu} e^{\Theta}, \tag{A.1}$$

where the supremum can be taken either over all bounded continuous functions or all bounded measurable functions $\Theta\colon \mathcal{H} \to \mathbb{R}$. Note that as soon as ν and μ are equivalent, the supremum is realized for $\Theta = \log\left(\frac{d\nu}{d\mu}\right)$.

We first prove the lower semicontinuity. For any bounded and *continuous* $\Theta\colon \mathcal{H} \to \mathbb{R}$ the mapping $(\nu,\mu) \mapsto \mathbb{E}^{\nu}\Theta - \log \mathbb{E}^{\mu} e^{\Theta}$ is continuous with respect to weak convergence of ν and μ. Hence, by (A.1) the mapping $(\nu,\mu) \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ is lower semicontinuous as the pointwise supremum of continuous mappings.

We now prove the compactness of sublevel sets. By the lower semicontinuity of $\nu \mapsto D_{\mathrm{KL}}(\nu\|\mu)$ and Prokohorov's theorem [5] it is sufficient to show that for any $M < \infty$ the set $\mathcal{B} := \{\nu\colon D_{\mathrm{KL}}(\nu\|\mu) \le M\}$ is tight. The measure μ is inner regular, and therefore for any $0 < \delta \le 1$ there exists a compact set K_δ such that $\mu(K_\delta^c) \le \delta$. Then choosing $\Theta = \mathbf{1}_{K_\delta^c} \log\left(1+\delta^{-1}\right)$ in (A.1) we get, for any $\nu \in \mathcal{B}$,

$$\begin{aligned}
\log\left(1+\delta^{-1}\right)\nu(K_\delta^c) &= \mathbb{E}^{\nu}\Theta \\
&\le M + \log(\mathbb{E}^{\mu} e^{\Theta}) \\
&= M + \log\Big(\mu(K_\delta) + \mu(K_\delta^c)\left(1+\delta^{-1}\right)\Big) \\
&\le M + \log\Big(1 + \left(\delta+1\right)\Big).
\end{aligned}$$

Hence, if for $\varepsilon > 0$ we choose δ small enough to ensure that

$$\frac{M + \log(3)}{\log\left(1+\delta^{-1}\right)} \le \varepsilon,$$

we have, for all $\nu \in \mathcal{B}$, that $\nu(K_\delta^c) \le \varepsilon$.

Appendix B. Some properties of Gaussian measures. The following lemma summarizes some useful facts about the weak convergence of Gaussian measures.

Lemma B.1. *Let ν_n be a sequence of Gaussian measures on $\mathcal{H}$ with mean $m_n \in \mathcal{H}$ and covariance operators C_n.*

1. *If the ν_n converge weakly to $\nu_\star$, then $\nu_\star$ is also Gaussian.*
2. *If $\nu_\star$ is Gaussian with mean $m_\star$ and covariance operator $C_\star$, then ν_n converges weakly to $\nu_\star$ if and only if the following conditions are satisfied:*
   (a) *$\|m_n - m_\star\|_{\mathcal{H}}$ converges to 0.*
   (b) *$\|\sqrt{C_n} - \sqrt{C_\star}\|_{\mathcal{HS}(\mathcal{H})}$ converges to 0.*
3. *Condition* (b) *can be replaced by the following condition:*
   (b') *$\|C_n - C_\star\|_{\mathcal{L}(\mathcal{H})}$ and $\mathbb{E}^{\nu_n}\|x\|_{\mathcal{H}}^2 - \mathbb{E}^{\nu_\star}\|x\|_{\mathcal{H}}^2$ converge to 0 .*

Proof. Point 1. Assume that ν_n converges weakly to ν. Then for any continuous linear functional $\phi\colon \mathcal{H} \to \mathbb{R}$ the push-forward measures $\nu_n \circ \phi^{-1}$ converge weakly to $\nu \circ \phi^{-1}$. The measures $\nu_n \circ \phi^{-1}$ are Gaussian measures on $\mathbb{R}$. For one dimensional

Gaussians a simple calculation with the Fourier transform (see, e.g., [23, Proposition 1.1]) shows that weak limits are necessarily Gaussian and weak convergence is equivalent to the convergence of mean and variance. Hence $\nu \circ \phi^{-1}$ is Gaussian, which in turn implies that ν is Gaussian. Points 2 and 3 are established in [7, Chapter 3.8]. □

As a next step we recall the Feldman–Hajek theorem as proved in [11, Theorem 2.23].

Proposition B.2. *Let $\mu_1 = N(m_1, C_1)$ and $\mu_2 = N(m_2, C_2)$ be two Gaussian measures on $\mathcal{H}$. The measures μ_1 are either singular or equivalent. They are equivalent if and only if the following three assumptions hold:*

1. *The Cameron–Martin spaces $C_1^{\frac{1}{2}}\mathcal{H}$ and $C_2^{\frac{1}{2}}\mathcal{H}$ are norm equivalent spaces with, in general, different scalar products generating the norms; we denote the space by $\mathcal{H}^1$.*
2. *The means satisfy $m_1 - m_2 \in \mathcal{H}^1$.*
3. *The operator $\big(C_1^{\frac{1}{2}} C_2^{-\frac{1}{2}}\big)\big(C_1^{\frac{1}{2}} C_2^{-\frac{1}{2}}\big)^\star - \mathrm{Id}$ is a Hilbert–Schmidt operator on $\mathcal{H}$.*

Remark B.3. Actually, in [11], item 3 is stated as $\big(C_2^{-\frac{1}{2}} C_1^{\frac{1}{2}}\big)\big(C_2^{-\frac{1}{2}} C_1^{\frac{1}{2}}\big)^\star - \mathrm{Id}$ is a Hilbert–Schmidt operator on $\mathcal{H}$. We find the formulation in item 3 more useful, and the fact that it is well-defined follows since $C_1^{\frac{1}{2}} C_2^{-\frac{1}{2}}$ is the adjoint of $C_2^{-\frac{1}{2}} C_1^{\frac{1}{2}}$. The two conditions are shown to be equivalent in [7, Lemma 6.3.1(ii)].

The methods used within the proof of the Feldman–Hajek theorem, as given in [11, Theorem 2.23], are used below to prove the following characterization of convergence with respect to total variation norm for Gaussian measures.

Lemma B.4. *For any $n \geq 1$ let ν_n be a Gaussian measure on $\mathcal{H}$ with covariance operator C_n and mean m_n, and let $\nu_\star$ be a Gaussian measure with covariance operator $C_\star$ and mean $m_\star$. Assume that the measures ν_n converge to $\nu_\star$ in total variation. Then we have*

$$\big\| C_\star^{\frac{1}{2}} \big(C_n^{-1} - C_\star^{-1}\big) C_\star^{\frac{1}{2}} \big\|_{\mathcal{HS}(\mathcal{H})} \to 0 \quad \textit{and} \quad \|m_n - m_\star\|_{\mathcal{H}^1} \to 0. \tag{B.1}$$

In order to prove Lemma B.4 we recall that for two probability measures ν and μ the *Hellinger distance* is defined as

$$D_{\mathrm{hell}}(\nu;\mu)^2 = \frac{1}{2}\int \left(\sqrt{\frac{d\nu}{d\lambda}(x)} - \sqrt{\frac{d\mu}{d\lambda}(x)} \right)^2 d\lambda(dx),$$

where λ is a probability measure on $\mathcal{H}$ such that $\nu \ll \lambda$ and $\mu \ll \lambda$. Such a λ always exists (average ν and μ, for example), and the value does not depend on the choice of λ.

For this we need the Hellinger integral

$$H(\nu;\mu) = \int \sqrt{\frac{d\mu}{d\lambda}(x)} \sqrt{\frac{d\nu}{d\lambda}(x)}\, \lambda(dx) = 1 - D_{\mathrm{hell}}(\nu;\mu)^2. \tag{B.2}$$

We recall some properties of $H(\nu;\mu)$.

Lemma B.5 (see [11, Proposition 2.19]).

1. *For any two probability measures ν and μ on $\mathcal{H}$ we have $0 \leq H(\nu;\mu) \leq 1$. We have $H(\nu;\mu) = 0$ if and only if μ and ν are singular, and $H(\nu;\mu) = 1$ if and only if $\mu = \nu$.*
2. *Let $\tilde{\mathcal{F}}$ be a sub-σ-algebra of $\mathcal{F}$, and denote by $H_{\tilde{\mathcal{F}}}(\nu,\mu)$ the Hellinger integrals of the restrictions of ν and μ to $\tilde{\mathcal{F}}$. Then we have*

$$H_{\tilde{\mathcal{F}}}(\nu,\mu) \geq H(\nu;\mu). \tag{B.3}$$

Proof of Lemma B.4. Before commencing the proof we demonstrate the equivalence of the Hellinger and total variation metrics. On the one hand the elementary inequality $(\sqrt{a}-\sqrt{b})^2 \le |a-b|$ which holds for any $a, b \ge 0$ immediately yields that

$$D_{\text{hell}}(\nu;\mu)^2 \le \frac{1}{2}\int \left|\frac{d\nu}{d\lambda}(x) - \frac{d\mu}{d\lambda}(x)\right| \lambda(dx) = D_{\text{tv}}(\nu,\mu).$$

On the other hand the elementary equality $(a-b) = (\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$, together with the Cauchy–Schwarz inequality, yields

$$\begin{aligned} D_{\text{tv}}(\nu;\mu) &= \frac{1}{2}\int \left|\frac{d\nu}{d\lambda}(x) - \frac{d\mu}{d\lambda}(x)\right| \lambda(dx) \\ &\le D_{\text{hell}}(\nu;\mu)\int \left(\sqrt{\frac{d\nu}{d\lambda}(x)} + \sqrt{\frac{d\mu}{d\lambda}(x)}\right)^2 \lambda(dx) \le 4D_{\text{hell}}(\nu;\mu). \end{aligned}$$

This justifies study of the Hellinger integral to prove total variation convergence.

We now proceed with the proof. We first treat the case of centered measures; i.e., we assume that $m_n = m_\star = 0$. For n large enough, ν_n and $\nu_\star$ are equivalent, and therefore their Cameron–Martin spaces coincide as sets, and in particular the operators $C_\star^{-\frac{1}{2}} C_n^{\frac{1}{2}}$ are defined on all of $\mathcal{H}$ and are invertible. By Proposition B.2 they are invertible bounded operators on $\mathcal{H}$. Denote by R_n the operator $(C_\star^{-\frac{1}{2}} C_n^{\frac{1}{2}})(C_\star^{-\frac{1}{2}} C_n^{\frac{1}{2}})^\star$. This shows, in particular, that the expression (B.1) makes sense, as it can be rewritten as

$$\left\| R_n^{-1} - \text{Id} \right\|^2_{\mathcal{HS}(\mathcal{H})} \to 0.$$

Denote by $(e_\alpha, \alpha \ge 1)$[4] the orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of the operator $C_\star$ and by $(\lambda_\alpha, \alpha \ge 1)$ the corresponding sequence of eigenvalues. For any n the operator R_n can be represented in the basis (e_α) by the matrix $(r_{\alpha,\beta;n})_{1\le\alpha,\beta<\infty}$, where

$$r_{\alpha,\beta;n} = \frac{\langle C_n e_\alpha, e_\beta\rangle}{\sqrt{\lambda_\alpha \lambda_\beta}}.$$

For any $\alpha \ge 1$ define the linear functional

$$\xi_\alpha(x) = \frac{\langle x, e_\alpha\rangle}{\sqrt{\lambda_\alpha}}, \qquad x \in \mathcal{H}. \tag{B.4}$$

By definition, we have for all α, β that

$$\begin{aligned} &\mathbb{E}^{\nu_\star}\big[\xi_\alpha(x)\big] = 0, && \mathbb{E}^{\nu_n}\big[\xi_\alpha(x)\big] = 0, \\ &\mathbb{E}^{\nu_\star}\big[\xi_\alpha(x)\xi_\beta(x)\big] = \delta_{\alpha,\beta}, \qquad \text{and} && \mathbb{E}^{\nu_n}\big[\xi_\alpha(x)\xi_\beta(x)\big] = r_{\alpha,\beta;n}. \end{aligned} \tag{B.5}$$

For any $\gamma \ge 1$ denote by $\mathcal{F}_\gamma$ the σ-algebra generated by $(\xi_1, \dots, \xi_\gamma)$. Furthermore, denote by $R_{\gamma;n}$ and I_γ the matrices $(r_{\alpha,\beta;n})_{1\le\alpha,\beta\le\gamma}$ and $(\delta_{\alpha,\beta})_{1\le\alpha,\beta\le\gamma}$. With this notation (B.5) implies that we have

$$\frac{d\nu_n\big|_{\mathcal{F}_\gamma}}{d\nu_\star\big|_{\mathcal{F}_\gamma}} = \frac{1}{\sqrt{\det(R_{\gamma;n})}} \exp\bigg(-\frac{1}{2}\sum_{\alpha,\beta\le\gamma} \xi_\alpha\xi_\beta\big(\big(R_{\gamma;n}^{-1}\big)_{\alpha,\beta} - \delta_{\alpha,\beta}\big)\bigg),$$

[4] Use of the same notation as for the eigenfunctions and eigenvectors of C_0 elsewhere should not cause confusion.

and in particular we get the Hellinger integrals

$$H_{\mathcal{F}_\gamma}\big(\nu_n;\nu_\star\big) = \frac{(\det R_{\gamma,n}^{-1})^{\frac{1}{4}}}{\Big(\det\Big(\frac{I_\gamma + R_{\gamma;n}^{-1}}{2}\Big)\Big)^{\frac{1}{2}}}.$$

Denoting by $\big(\lambda_{\alpha;\gamma;n}, \alpha = 1,\dots,\gamma\big)$ the eigenvalues of $R_{\gamma,n}^{-1}$ this expression can be rewritten as

$$-\log\big(H_{\mathcal{F}_\gamma}\big(\nu_n;\nu_\star\big)\big) = \frac{1}{4}\sum_{\alpha=1}^{\gamma}\log\frac{(1+\lambda_{\alpha;\gamma;n})^2}{4\lambda_{\alpha;\gamma;n}} \le -\log\big(H(\nu_n;\nu_\star)\big), \tag{B.6}$$

where we have used (B.3). The right-hand side of (B.6) goes to zero as $n \to \infty$, and in particular it is bounded by 1 for n large enough, say for $n \ge n_0$. Hence there exist constants $0 < K_1, K_2 < \infty$ such that for all $n \ge n_0$ and all γ, α we have $K_1 \le \lambda_{\alpha;\gamma;n} \le K_2$. There exists a third constant $K_3 > 0$ such that for all $\lambda \in [K_1, K_2]$ we have

$$(1-\lambda)^2 \le \frac{K_3}{4}\log\frac{(1+\lambda)^2}{4\lambda}.$$

Hence, we can conclude that for $n \ge n_0$

$$\big\|R_{\gamma,n}^{-1} - I_\gamma\big\|_{\mathcal{HS}(\mathbb{R}^\gamma)}^2 = \sum_{\alpha=1}^{\gamma}\big|\lambda_{\alpha;\gamma;n} - 1\big|^2 \le -K_3\log\big(H(\nu_n;\nu_\star)\big).$$

As this bound holds uniformly in γ, the claim is proved in the case $m_n = m_\star = 0$.

As a second step let us treat the case where m_n and $m_\star$ are arbitrary but the covariance operators coincide, i.e., for all $n \ge 1$ we have $C_n = C_\star =: C$. As above, denote by $(e_\alpha, \alpha \ge 1)$ the orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of the operator C and by $(\lambda_\alpha, \alpha \ge 1)$ the corresponding sequence of eigenvalues. Furthermore, define the random variable ξ_α as above in (B.4). Then we get the identities

$$\begin{aligned} \mathbb{E}^{\nu_\star}\big[\xi_\alpha(x)\big] &= \frac{m_{\star;\alpha}}{\sqrt{\lambda_\alpha}}, & \mathbb{E}^{\nu_n}\big[\xi_\alpha(x)\big] &= \frac{m_{n;\alpha}}{\sqrt{\lambda_\alpha}},\\ \operatorname{cov}^{\nu_\star}\big(\xi_\alpha(x), \xi_\beta(x)\big) &= \delta_{\alpha,\beta}, \quad \text{and} & \operatorname{cov}^{\nu_n}\big(\xi_\alpha(x)\xi_\beta(x)\big) &= \delta_{\alpha,\beta}, \end{aligned}$$

where $\operatorname{cov}^{\nu_\star}$ and $\operatorname{cov}^{\nu_n}$ denote the covariances with respect to the measures $\nu_\star$ and ν_n. Here we have set $m_{\star;\alpha} := \langle m_\star, e_\alpha\rangle$ and $m_{n;\alpha} := \langle m_n, e_\alpha\rangle$. Denoting as above by $\mathcal{F}_\gamma$ the σ-algebra generated by $(\xi_1,\dots,\xi_\gamma)$ we get for any $\gamma \ge 1$

$$H_{\mathcal{F}_\gamma}\big(\nu_n;\nu_\star\big) = \exp\bigg(-\frac{1}{8}\sum_{\alpha=1}^{\gamma}\frac{1}{\lambda_\alpha}\big|m_{\star;\alpha} - m_{n;\alpha}\big|^2\bigg). \tag{B.7}$$

Noting that $\|m_n - m_\star\|_{\mathcal{H}^1}^2 = \sum_{\alpha\ge1}\frac{1}{\lambda_\alpha}\big|m_{n;\alpha} - m_{\star;\alpha}\big|^2$ and reasoning as above in (B.6) we get that $\|m_n - m_\star\|_{\mathcal{H}^1}^2 \to 0$.

The general case of arbitrary m_n, $m_\star$, C_n, and $C_\star$ can be reduced to the two cases above. Indeed, assume that ν_n converges to $\nu_\star$ in total variation. After a translation which does not change the total variation distance, we can assume that $m_\star = 0$. Furthermore, by symmetry, if the measures $N(m_n, C_n)$ converge to $N(0, C_\star)$ in total variation, then so do the measures $N(-m_n, C_n)$. A coupling argument, which we now give, shows that then the Gaussian measures $N(0, 2C_n)$ converge to $N(0, 2C_\star)$, also in total variation. Let (X_1, Y_1) be random variables with $X_1 \sim N(m_n, C_n)$ and $Y_1 \sim N(0, C_\star)$ and $\mathbb{P}(X_1 \neq Y_1) = \|N(m_n, C_n) - N(0, C_\star)\|_{\mathrm{tv}}$, and in the same way let (X_2, Y_2) be independent from (X_1, Y_1) and such that $X_2 \sim N(-m_n, C_n)$ and $Y_2 \sim N(0, C_\star)$ with $\mathbb{P}(X_2 \neq Y_2) = \|N(-m_n, C_n) - N(0, C_\star))\|_{\mathrm{tv}}$. Then we have $X_1 + X_2 \sim N(0, 2C_n)$, $Y_1 + Y_2 \sim N(0, 2C_\star)$, and

$$\begin{aligned}\|N(0, 2C_n) - N(0, 2C_\star)\|_{\mathrm{tv}} &= \mathbb{P}(X_1 + X_2 \neq Y_1 + Y_2)\\ &\leq \mathbb{P}(X_1 \neq Y_1) + \mathbb{P}(X_2 \neq Y_2)\\ &= 2\|N(m_n, C_n) - N(0, C_\star)\|_{\mathrm{tv}}.\end{aligned}$$

Hence we can apply the first part of the proof to conclude that the desired conclusion concerning the covariances holds.

We now turn to the means. From the fact that $N(m_n, C_n)$ and $N(0, C_n)$ converge to $N(0, C_\star)$ in total variation we can conclude by the triangle inequality that $\|N(m_n, C_n) - N(0, C_n)\|_{\mathrm{tv}} \to 0$ and hence $\log H(N(m_n, C_n), N(0, C_n)) \to 0$. By (B.7) this implies that

$$\|C_n^{-\frac{1}{2}} m_n\|_{\mathcal{H}} \leq 8 \log H(N(m_n, C_n), N(0, C_n)) \to 0.$$

Furthermore, the convergence of

$$\left\| C_\star^{\frac{1}{2}} (C_n^{-1} - C_\star^{-1}) C_\star^{\frac{1}{2}} \right\|_{\mathcal{HS}(\mathcal{H})} = \left\| (C_\star^{\frac{1}{2}} C_n^{-\frac{1}{2}})(C_\star^{\frac{1}{2}} C_n^{-\frac{1}{2}})^\star - \mathrm{Id} \right\|_{\mathcal{HS}(\mathcal{H})}$$

implies that $\sup_{n \geq 1} \|C_\star^{-\frac{1}{2}} C_n^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{H})} < \infty$. So we can conclude that, as desired,

$$\|m_n\|_{\mathcal{H}^1} \leq \left(\sup_{n \geq 1} \|C_\star^{-\frac{1}{2}} C_n^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{H})} \right) \|C_n^{-\frac{1}{2}} m_n\|_{\mathcal{H}} \to 0. \qquad \square$$

Appendix C. Characterization of Gaussian measures via precision operators.

LEMMA C.1. *Let $C_0 = (-\partial_t^2)^{-1}$ be the inverse of the Dirichlet Laplacian on $[-1, 1]$ with domain $H^2([-1, 1]) \cap H_0^1([-1, 1])$. Then $\mu_0 = N(0, C_0)$ is the distribution of a homogeneous Brownian bridge on $[-1, 1]$. Consider measure $\nu \ll \mu_0$ defined by*

$$\frac{d\nu}{d\mu_0}(x(\cdot)) = \frac{1}{Z} \exp\left(-\frac{1}{2} \int_{-1}^{1} \theta(t)\, x(t)^2\, dt \right), \tag{C.1}$$

where θ is a smooth function with infimum strictly larger than $-\frac{\pi^2}{4}$ on $[-1, 1]$. Then ν is a centered Gaussian $N(0, C)$ with $C^{-1} = C_0^{-1} + \theta$.

The following proof closely follows techniques introduced to prove Theorem 2.1 in [29].

Proof. As above, denote $\mathcal{H} = L^2([-1, 1])$ and $\mathcal{H}^1 = H_0^1([-1, 1])$. Furthermore, let $(e_\alpha, \lambda_\alpha, \alpha \geq 1)$ be the eigenfunction/eigenvalue pairs of C_0 ordered by decreasing eigenvalues. For any $\gamma \geq 1$ let π_γ be the orthogonal projection on $\mathcal{H}$ onto $\mathcal{H}_\gamma = \mathrm{span}(e_1, \ldots, e_\gamma)$. Denote $\mathcal{H}_\gamma^\perp = (\mathrm{Id} - \pi_\gamma)\mathcal{H}$.

For each $\gamma \geq 1$ define the measure $\nu_\gamma \ll \mu_0$ by

$$\frac{d\nu_\gamma}{d\mu_0}(x(\cdot)) = \frac{1}{Z_\gamma} \exp\left(-\frac{1}{2} \int_{-1}^{1} \theta(t) \left(\pi_\gamma x(t)\right)^2 dt \right).$$

We first show that the ν_γ are centered Gaussian, and we characterize their covariance. To see this note that μ_0 factors as the independent product of two Gaussians on $\mathcal{H}_\gamma$ and $\mathcal{H}_\gamma^\perp$. Since the change of measure defining ν_γ depends only on $\pi_\gamma x \in \mathcal{H}_\gamma$, it follows that ν_γ also factors as an independent product. Furthermore, the factor on $\mathcal{H}_\gamma^\perp$ coincides with the projection of μ_0 and is Gaussian. On $\mathcal{H}_\gamma$, which is finite dimensional, it is clear that ν_γ is also Gaussian because the change of measure is defined through a finite dimensional quadratic form. This Gaussian is centered and has inverse covariance (precision) given by $\pi_\gamma(C_0^{-1}+\theta)\pi_\gamma = \pi_\gamma C^{-1}\pi_\gamma$. Hence ν_γ is also Gaussian; denote its covariance operator by C_γ.

A straightforward dominated convergence argument shows that ν_γ converges weakly to ν as a measure on $\mathcal{H}$, and it follows that ν is a centered Gaussian by Lemma B.1; we denote the covariance by Σ. It remains to show that $\Sigma = C$. On the one hand, we have by Lemma B.1, item 3, that C_γ converges to Σ in the operator norm. On the other hand, we have for any $x \in \mathcal{H}^1$ and for $\gamma \geq 1$ that

$$\begin{aligned}\big|\langle x, C_\gamma^{-1}x\rangle - \langle x, C^{-1}x\rangle\big| &= \int_{-1}^{1}\theta(t)\big((\mathrm{Id}-\pi_\gamma)x(t)\big)^2\,dt \leq \|\theta\|_{L^\infty}\|(\mathrm{Id}-\pi_\gamma)x(t)\|_{L^2}^2\\ &\leq \|\theta\|_{L^\infty}\lambda_\gamma^2\|x(t)\|_{H_0^1}^2.\end{aligned}$$

As the $\lambda_\gamma \to 0$ for $\gamma \to \infty$ and as the operator $C^{\frac{1}{2}}C_0^{-\frac{1}{2}}$ is a bounded invertible operator on $\mathcal{H}^1$, this implies the convergence of C_γ^{-1} to C^{-1} in the strong resolvent sense by [31, Theorem VIII.25]. The conclusion then follows as in the proof of Theorem 3.10. ∎

Acknowledgment. The third author is grateful to Colin Fox for fruitful discussions on related topics.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed., Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2008.

[2] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, *Gaussian process approximations of stochastic differential equations*, J. Mach. Learn. Res., 1 (2007), pp. 1–16.

[3] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, *Variational inference for diffusion processes*, in Advances in Neural Information Processing Systems, Vol. 20, MIT Press, Cambridge, MA, 2008, pp. 17–24.

[4] J. M. Ball and G. Knowles, *A numerical method for detecting singular minimizers*, Numer. Math., 51 (1987), pp. 181–197.

[5] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, 2009.

[6] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Vol. 1, Springer, New York, 2006.

[7] V. I. Bogachev, *Gaussian Measures*, Math. Surveys Monogr. 62, AMS, Providence, RI, 1998.

[8] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M Stuart, *Bayesian inverse problems for functions and applications to fluid mechanics*, Inverse Problems, 25 (2009), 115008.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 2012.

[10] I. Csiszár, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probability, 3 (1975), pp. 146–158.

[11] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK, 1992.

[12] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss, *Map estimators and posterior consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 29 (2013), 095017.

[13] M. Dashti and A. M. Stuart, *Uncertainty quantification and weak approximation of an elliptic inverse problem*, SIAM J. Numer. Anal., 49 (2011), pp. 2524–2542.

[14] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Wiley Ser. Probab. Statist. Probab. Statist., John Wiley & Sons, New York, 1997.

[15] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Springer Science+Business Media, New York, 1996.

[16] D. Feyel and A. S. Üstünel, *Monge-Kantorovitch measure transportation and Monge-Ampère equation on Wiener space*, Probab. Theory Related Fields, 128 (2004), pp. 347–385.

[17] D. Giannakis and A. J. Majda, *Quantifying the predictive skill in long-range forecasting. Part* I: *Coarse-grained predictions in a simple ocean model*, J. Climate, 25 (2012), pp. 1793–1813.

[18] M. Hairer, *An Introduction to Stochastic PDEs*, preprint, arXiv:0907.4178, 2009.

[19] M. Hairer, A. M. Stuart, and J. Voss, *Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods*, in The Oxford Handbook of Nonlinear Filtering, D. Crisan and B. Rozovsky, eds., Oxford University Press, Oxford, UK, 2011, pp. 833–873.

[20] J. P. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer, New York, 2005.

[21] M. A. Katsoulakis and P. Plecháč, *Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems*, J. Chem. Phys., 139 (2013), 074115.

[22] M. A. Katsoulakis, P. Plecháč, L. Rey-Bellet, and D. K. Tsagkarogiannis, *Coarse-graining schemes and a posteriori error estimates for stochastic lattice systems*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 627–660.

[23] J.-F. Le Gall, *Mouvement Brownien, Martingales et Calcul Stochastique*, Math. Appl. (Berlin) 71, Springer, Heidelberg, 2013.

[24] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2008.

[25] A. J. Majda and B. Gershgorin, *Improving model fidelity and sensitivity for complex systems through empirical information theory*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 10044–10049.

[26] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[27] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions*, SIAM J. Sci. Comput., to appear.

[28] F. J. Pinski, A. M. Stuart, and F. Theil, Γ*-limit for transition paths of maximal probability*, J. Stat. Phys., 146 (2012), pp. 955–974.

[29] Y. Pokern, A. M. Stuart, and J. H. Van Zanten, *Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs*, Stochastic Process. Appl., 123 (2012), pp. 603–628.

[30] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* II. *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.

[31] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* I. *Functional Analysis,* 2nd ed., Academic Press, New York, 1980.

[32] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[33] C. Villani, *Optimal Transport: Old and New*, Grundlehren Math. Wiss. 338, Springer-Verlag, Berlin, 2009.