GAUSSIAN APPROXIMATIONS FOR PROBABILITY MEASURES ON $\mathbf{R}^d$

YULONG LU, ANDREW STUART, AND HENDRIK WEBER

Abstract. This paper concerns the approximation of probability measures on $\mathbf{R}^d$ with respect to the Kullback-Leibler divergence. Given an admissible target measure, we show the existence of the best approximation, with respect to this divergence, from certain sets of Gaussian measures and Gaussian mixtures. The asymptotic behavior of such best approximations is then studied in the frequently occuring small parameter limit where the measure concentrates; this asympotic behaviour is characterized using Γ-convergence. The theory developed is then applied to understanding the frequentist consistency of Bayesian inverse problems. For a fixed realization of noise, we show the asymptotic normality of the posterior measure in the small noise limit. Taking into account the randomness of the noise, we prove a Bernstein-Von Mises type result for the posterior measure.

MSC 2010: 28C20, 62F15, 60H07.

Keywords: Gaussian approximation, Kullback-Leibler divergence, Gamma-convergence, Bernstein-Von Mises Theorem.

1. Introduction

In this paper, we study the "best" approximation of a general finite dimensional probability measure, which could be non-Gaussian, from a set of simple probability measures, such as a single Gaussian measure or a Gaussian mixture family. We define "best" to mean the measure within the simple class which minimizes the Kullback-Leibler divergence between itself and the target measure. This type of approximation is central to many ideas currently used in machine learning [3], yet has not been the subject of any substantial systematic underpinning theory. The goal of this paper is to contribute to the development of such a theory in the concrete finite dimensional setting in two ways: (i) by establishing the existence of best approximations; (ii) by studying their asymptotic properties in a measure concentration limit of interest. The abstract theory is then applied to study frequentist consistency [8] of Bayesian inverse problems.

The idea of approximation for probability measures with respect to Kullback-Leibler divergence has been applied in a number of areas; see for example [17, 13, 16, 21]. Despite the wide usage of Kullback-Leibler approximation, systematic theoretical study has only been initiated recently. In [20], the measure approximation problem is studied from the calculus of variations point of view, and existence of minimizers established therein; the companion paper [19] proposed numerical algorithms for implementing Kullback-Leibler minimization in practice. In [16], Gaussian approximation is used as a new approach for identifying the most likely path between equilibrium states in molecular dynamics; furthermore, the asymptotic behavior of the Gaussian approximation in the small temperature limit is

YL is supported by EPSRC as part of the MASDOC DTC at the University of Warwick with grant No. EP/HO23364/1. AMS is supported by DARPA, EPSRC and ONR. HW is supported by the Royal Society through the University Research Fellowship UF140187.

analyzed via Γ-convergence. Here our interest is to develop the ideas in [16] in the context of a general class of measure approximation problems in finite dimensions.

To be concrete we consider approximation of a family of probability measures $\{\mu_\varepsilon\}_{\varepsilon>0}$ on $\mathbf{R}^d$ with (Lebesgue) density of the form

$$\mu_\varepsilon(dx) = \frac{1}{Z_{\mu,\varepsilon}} \exp\left(-\frac{1}{\varepsilon}V_1^\varepsilon(x) - V_2(x)\right) dx; \tag{1.1}$$

here $Z_{\mu,\varepsilon}$ is the normalization constant. A typical example of a measure μ_ε with this form arises in the study of Bayesian inference. The measure μ_ε is the posterior, the function $\varepsilon^{-1}V_1^\varepsilon$ is the negative log-likelihood, up to an additive constant, and V_2 is, up to an additive constant, the negative logarithm of the prior density. In addition, the parameter ε is associated with the number of observations or the noise level of the statistical experiment.

Given a measure μ_ε defined by (1.1), we find a measure ν from a set of simple measures which minimizes the Kullback-Leibler divergence $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$. In addition, we characterize the limiting behavior of the best approximation from the set of simple measures, as well as the limiting behaviour of the Kullback-Leibler divergence as $\varepsilon \downarrow 0$.

The rest of the paper is organized as follows. In Section 2 we set up various underpinning concepts which are used throughout the paper: in Subsections 2.1 and 2.2, we recall some basic facts on Kullback-Leibler divergence and Γ-convergence and in Subsections 2.3 and 2.4 we spell out the assumptions made and the notation used. Sections 3 and Section 4 concern the problem of approximation of the measure μ_ε by, respectively, a single Gaussian measure and a Gaussian mixture. In particular, the small ε asymptotics of the Gaussians (or Gaussian mixtures) are captured by using the framework of Γ-convergence. Finally, in Section 5, the theory which we have developed is applied to understand posterior consistency for Bayesian inverse problems.

2. Set-Up

2.1. Kullback-Leibler Divergence. Let ν and μ be two probability measures on $\mathbb{R}^d$ and assume that ν is absolutely continuous with resepct to μ. The Kullback-Leibler divergence, or relative entropy, of ν with respect to μ is

$$D_{\mathrm{KL}}(\nu||\mu) = \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu}\right).$$

If ν is not absolutely continuous with respect to μ, then the Kullback-Leibler divergence is defined as $+\infty$. By definition, the Kullback-Leibler divergence is non-negative but it is not a metric since it does not obey the triangle inequality and it is not symmetric in its two arguments. In this paper, we will consider minimizing $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$ with respect to ν, over a suitably chosen set of measures, and with μ_ε being the target measure defined in (1.1). Swapping the order of these two measures within the divergence is undesirable for our purposes. This is because minimizing $D_{\mathrm{KL}}(\mu_\varepsilon||\cdot)$ within the set of all Gaussian measures will lead to matching of moments [3]; this is inappropriate for multimodal measures where a more desirable outcome would be the existence of multiple local minimizers at each mode [20, 19].

Although the Kullback-Leibler divergence is not a metric, its information theoretic interpretation make it natural for approximate inference. Furthermore it is a convenient quantity to work with for at least two reasons. First the divergence provides useful upper bound for many metrics; in particular, one has the Pinsker inequality

$$d_{\mathrm{TV}}(\nu,\mu) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\nu||\mu)} \tag{2.1}$$

where d_{TV} denotes the total variation distance. Second the logarithmic structure of $D_{\mathrm{KL}}(\cdot||\cdot)$ allows us to carry out explicit calculations, and numerical computations, which are considerably more difficult when using the total variation distance directly.

subsec:gamma

2.2. **Γ-convergence.** We recall the definition and a basic result concerning Γ-convergence. This is a useful tool for studying families of minimization problems. In this paper we will use it to study the parametric limit $\varepsilon \to 0$ in our approximation problem.

d:gcc **Definition 2.1.** *Let $\mathcal{X}$ be a metric space and $E_\varepsilon : \mathcal{X} \to \mathbf{R}$ a family of functionals indexed by $\varepsilon > 0$. Then E_ε Γ-converges to $E : \mathcal{X} \to \mathbf{R}$ as $\varepsilon \to 0$ if the following conditions hold:*

(i) (liminf inequality) for every $u \in \mathcal{X}$, and for every sequence $u_\varepsilon \in \mathcal{X}$ such that $u_\varepsilon \to u$, it holds that $E(u) \le \liminf_{\varepsilon\downarrow 0} E_\varepsilon(u_\varepsilon)$;

(ii) (limsup inequality) for every $u \in \mathcal{X}$ there exists a recovery sequence $\{u_\varepsilon\}$ such that $u_\varepsilon \to u$ and $E(u) \ge \limsup_{\varepsilon\downarrow 0} E_\varepsilon(u_\varepsilon)$.

We say a sequence of functionals $\{E_\varepsilon\}$ is compact if $\limsup_{\varepsilon\downarrow 0} E_\varepsilon(u_\varepsilon) < \infty$ implies that there exists a subsequence $\{u_{\varepsilon_j}\}$ such that $u_{\varepsilon_j} \to u \in \mathcal{X}$.

The notion of Γ-convergence is useful because of the following fundamental theorem, which can be proved by similar methods as the proof of [4, Theorem 1.21].

thm:fgamma **Theorem 2.2.** *Let u_ε be a minimizer of E_ε with $\limsup_{\varepsilon\downarrow 0} E_\varepsilon(u_\varepsilon) < \infty$. If E_ε is compact and Γ-converges to E, then there exists a subsequence u_{ε_j} such that $u_{\varepsilon_j} \to u$ where u is a minimizer of E.*

Thus, when this theorem applies, it tells us that minimizers of E characterize the limits of convergent subsequences of minimizers of E_ε. In other words the $\Gamma-$limit captures the behavior of the minimization problem in the small ε limit.

ssec:A

2.3. **Assumptions.** Throughout the paper, we make the following assumptions on the potential functions V_1^ε and V_2 which define the target measure of interest.

assump **Assumptions 2.3.**

(A-1) *For any $\varepsilon > 0$, V_1^ε and V_2 are non-negative functions in the space $C^4(\mathbf{R}^d)$. Moreover, there exists constants $\varepsilon_0 > 0$ and $M_V > 0$ such that when $\varepsilon < \varepsilon_0$,*

$$|\partial_x^\alpha V_1^\varepsilon(x)| \vee |\partial_x^\alpha V_2(x)| \le M_V e^{|x|^2}$$

for all $|\alpha| \le 4$ and all $x \in \mathbf{R}^d$.

(A-2) *There exists $n > 0$ such that when $\varepsilon \ll 1$, the set of minimizers of V_1^ε is $\mathscr{E}^\varepsilon = \{x_\varepsilon^1, x_\varepsilon^2, \cdots, x_\varepsilon^n\}$ and $V_1^\varepsilon(x_\varepsilon^i) = 0, i = 1, \cdots, n$.*

(A-3) *There exists V_1 such that $V_1^\varepsilon \to V_1$ pointwise. The limit V_1 has n distinct global minimisers which are given by $\mathscr{E} = \{x^1, x^2, \cdots, x^n\}$. For each $i = 1, \ldots, n$ the Hessian $D^2V_1(x^i)$ is positive definite.*

(A-4) *The convergence $x_\varepsilon^i \to x^i$ holds.*

(A-5) *There exist constants $c_0, c_1 > 0$ and $\varepsilon_0 > 0$ such that when $\varepsilon < \varepsilon_0$,*

$$V_1^\varepsilon(x) \ge -c_0 + c_1|x|^2, x \in \mathbf{R}^d.$$

Remark 2.4. Conditions (A-2)-(A-4) mean that for sufficiently small $\varepsilon > 0$, the function V_1^ε behaves like a quadratic function in the neighborhood of the minimizers x_ε^i and of x^i. In particular, in conjunction with Condition (A-5) this implies that there exists $\delta > 0$ and $C_\delta > 0$ such that $\forall\, 0 \le \eta < \delta$,

eq:distV
$$\operatorname{dist}(x, \mathscr{E}) \ge \eta \Longrightarrow \liminf_{\varepsilon\downarrow 0} V_1^\varepsilon(x) \ge C_\delta |\eta|^2. \tag{2.2}$$

Remark 2.5. The local boundedness in C^4 (Assumption (A-1)) together with the pointwise convergence of V_1^ε to V_1 (Assumption (A-3)) imply the much stronger locally uniform convergence of derivatives up to order 3. Furthermore, (A-4) then implies that $V_1^\varepsilon(x_\varepsilon^i) \to V_1(x^i)$ and $D^2 V_1^\varepsilon(x_\varepsilon^i) \to D^2 V_1(x^i)$.

ssec:N

2.4. **Notation.** Throughout the paper, C and $\tilde{C}$ will be generic constants which are independent of the quantities of interest, and may change from line to line. Let $\mathcal{S}_{\geq}(\mathbf{R}, d)$ and $\mathcal{S}_{>}(\mathbf{R}, d)$ be the set of all $d \times d$ real matrices which are positive semi-definite or positive definite, respectively. Denote by $N(m, \mathbf{\Sigma})$ a Gaussian measure with mean m and covariance matrix $\mathbf{\Sigma}$. We use $|\mathbf{A}|$ to denote the Frobenius norm of the $d \times d$ matrix $\mathbf{A}$, namely $|\mathbf{A}| = \sqrt{\mathrm{Tr}(\mathbf{A}^T \mathbf{A})}$. We denote by $\lambda_{\min}(\mathbf{A})$ the smallest eigenvalue of $\mathbf{A}$. We let $B(x, r)$ denote a ball in $\mathbf{R}^d$ with center x and radius r. Given a random variable η, we use $\mathbb{E}^\eta$ and $\mathbb{P}^\eta$ when computing the expectation and the probability under the law of η respectively.

sec:SGaussian

3. Approximation by Single Gaussian measures

Let $\mathcal{A}$ be the set of Gaussian measures on $\mathbf{R}^d$, given by

$$\mathcal{A} = \{N(m, \mathbf{\Sigma}) : m \in \mathbf{R}^d, \mathbf{\Sigma} \in \mathcal{S}_{\geq}(\mathbf{R}, d)\}.$$

The set $\mathcal{A}$ is closed with respect to weak convergence of probability measures. Consider the variational problem

prob:var

$$\inf_{\nu \in \mathcal{A}} D_{\mathrm{KL}}(\nu || \mu_\varepsilon). \tag{3.1}$$

Given $\nu = N(m, \mathbf{\Sigma}) \in \mathcal{A}$, the Kullback-Leibler divergence $D_{\mathrm{KL}}(\nu || \mu_\varepsilon)$ can be calculated explicitly as

eq:dkl

$$\begin{aligned} D_{\mathrm{KL}}(\nu || \mu_\varepsilon) &= \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu_\varepsilon}\right) \\ &= \frac{1}{\varepsilon}\mathbb{E}^\nu V_1^\varepsilon(x) + \mathbb{E}^\nu V_2(x) - \log\sqrt{(2\pi)^d \det \mathbf{\Sigma}} - \frac{d}{2} + \log Z_{\mu,\varepsilon}. \end{aligned} \tag{3.2}$$

If $\mathbf{\Sigma}$ is non-invertible then $D_{\mathrm{KL}}(\nu || \mu_\varepsilon) = +\infty$. The term $-\frac{d}{2}$ comes from the expectation $\mathbb{E}^\nu \frac{1}{2}(x-m)^T \mathbf{\Sigma}(x-m)$ and is independent of $\mathbf{\Sigma}$. The term $-\log\sqrt{(2\pi)^d \det \mathbf{\Sigma}}$ prevents the measure ν from being too close to a Dirac measure. The following theorem shows that the problem (3.1) has a solution.

thm:exists1

Theorem 3.1. *Consider the measure μ_ε given by (1.1). For any $\varepsilon > 0$, there exists at least one probability measure $\overline{\nu}_\varepsilon \in \mathcal{A}$ solving the problem (3.1).*

Proof. We first show that the infimum of (3.1) is finite. In fact, consider $\nu^* = N(0, \frac{1}{4}\mathbf{I}_d)$. Under the Assumption 2.3 (A-1) we have that

$$\mathbb{E}^{\nu^*} V_1^\varepsilon(x) \vee \mathbb{E}^{\nu^*} V_2(x) \leq \frac{M_V}{\sqrt{(2\pi \times \frac{1}{4})^d}} \int_{\mathbf{R}^d} e^{-\frac{4}{2}|x|^2 + |x|^2} dx < \infty.$$

Note that the integral in the last expression is finite due to $-\frac{4}{2} + 1 < 0$. Hence we know from (3.2) that $\inf_{\nu \in \mathcal{A}} D_{\mathrm{KL}}(\nu || \mu_\varepsilon) < \infty$. Then the existence of minimizers follows from the fact that the Kullback-Leibler divergence has compact sub-level sets and the closedness of $\mathcal{A}$ with respect to weak convergence of probability measures; see e.g. [20, Corollary 2.2]. □

We aim to understand the asymptotic behavior of the minimizers $\overline{\nu}_\varepsilon$ of the problem (3.1) as $\varepsilon \downarrow 0$. Due to the factor $\frac{1}{\varepsilon}$ in front of V_1^ε in the definition of μ_ε, (1.1), we expect the typical size of fluctuations around the minimizers to be of order $\sqrt{\varepsilon}$ and we reflect that in our choice of scaling. More precisely, for $m \in \mathbf{R}^d$, $\mathbf{\Sigma} \in \mathcal{S}_{\geq}(\mathbf{R}, d)$ we define $\nu_\varepsilon = N(m, \varepsilon\mathbf{\Sigma})$ and set

$$F_\varepsilon(m, \mathbf{\Sigma}) := D_{\mathrm{KL}}(\nu_\varepsilon || \mu_\varepsilon).$$

Understanding the asymptotic behavior of minimizers $\overline{\nu}_\varepsilon$ in the small ε limit may be achieved by understanding Γ-convergence of the functional F_ε.

To that end, we define weights

$$\beta^i = \left(\det D^2 V_1(x^i)\right)^{-\frac{1}{2}} \cdot e^{-V_2(x^i)}, \qquad i = 1, \cdots, n,$$

and the counting probability measure on $\{1, \dots, n\}$ given by

$$\boldsymbol{\beta} := \frac{1}{\sum_{j=1}^n \beta^j}(\beta^1, \cdots, \beta^n).$$

Intuitively, as $\varepsilon \downarrow 0$, we expect the measure μ_ε to concentrate on the set $\{x^i\}$ with weights on each x^i given by $\boldsymbol{\beta}$; this intuition is reflected in the asymptotic behavior of the normalization constant $Z_{\mu,\varepsilon}$, as we now show. By definition,

$$Z_\varepsilon = \int_{\mathbf{R}^d} \exp\left(-\frac{1}{\varepsilon} V_1^\varepsilon(x) - V_2(x)\right) dx.$$

The following lemma follows from the Laplace approximation for integrals (see e.g. [12]) and Assumption 2.3 (A-4).

lem:normconst **Lemma 3.2.** *Let V_1^ε and V_2 satisfy the Assumptions 2.3. Then as $\varepsilon \downarrow 0$,*

$$Z_{\mu,\varepsilon} = \sqrt{(2\pi\varepsilon)^d} \cdot \left(\sum_{i=1}^n \beta^i\right) \cdot (1 + o(1))\,. \tag{3.3}$$

In view of the original expression (3.2) for $D_{\mathrm{KL}}(\nu, \mu_\varepsilon)$ as well as the specific scaling of $\nu_\varepsilon = N(m, \varepsilon\mathbf{\Sigma})$, Lemma 3.2 yields that

eq:dkl_asym1

$$F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \frac{1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) - \frac{d}{2} - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) + \log\left(\sum_{i=1}^n \beta^i\right) + o(1). \tag{3.4}$$

Armed with this analysis of the normalization constant we may now prove the following theorem which identifies the Γ-limit of F_ε. To this end we define

$$F_0(m, \mathbf{\Sigma}) := V_2(m) + \frac{1}{2}\mathrm{Tr}\left(D^2 V_1(m) \cdot \mathbf{\Sigma}\right) - \frac{d}{2} - \frac{1}{2}\log\det \mathbf{\Sigma} + \log\left(\sum_{i=1}^n \beta^i\right).$$

thm:gamma **Theorem 3.3.** *The Γ-limit of F_ε is*

eq:F

$$F(m, \mathbf{\Sigma}) := \begin{cases} F_0(m, \mathbf{\Sigma}) & \text{if } m \in \mathscr{E} \text{ and } \mathbf{\Sigma} \in \mathcal{S}_{>}(\mathbf{R}, d), \\ \infty & \text{otherwise.} \end{cases} \tag{3.5}$$

The following corollary follows directly from the Γ-convergence of F_ε.

cor:covmin **Corollary 3.4.** *Let $\{(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)\}$ be a family of minimizers of $\{F_\varepsilon\}$. Then there exists a subsequence $\{\varepsilon_k\}$ such that $(m_{\varepsilon_k}, \mathbf{\Sigma}_{\varepsilon_k}) \to (m, \mathbf{\Sigma})$ and $F_{\varepsilon_k}(m_{\varepsilon_k}, \mathbf{\Sigma}_{\varepsilon_k}) \to F(m, \mathbf{\Sigma})$. Moreover, $(m, \mathbf{\Sigma})$ is a minimizer of F.*

Before we give the proof of Theorem 3.3, let us first discuss the limit functional F as well as its minimization. We assume that $m = x^{i_0}$ for some $i_0 \in \{1, \dots, n\}$

and rewrite the definition of $F_0(x^{i_0}, \mathbf{\Sigma})$, by adding and subtracting $\log(\beta^{i_0}) = -V_2(x^{i_0}) - \frac{1}{2}\log\left(\left(\det D^2V_1(x^{i_0})\right)\right)$ and cancelling the terms involving $V_2(x^{i_0})$ as

eq:F0-1

$$F_0(x^{i_0}, \mathbf{\Sigma}) = \frac{1}{2}\mathrm{Tr}\left(D^2V_1(x^{i_0}) \cdot \mathbf{\Sigma}\right) - \frac{d}{2} - \frac{1}{2}\log\det(D^2V_1(x^{i_0}) \cdot \mathbf{\Sigma}) \\ + \log\left(\sum_{i=1}^{n} \beta^i\right) - \log\left(\beta^{i_0}\right). \tag{3.6}$$

Now it is interesting to see that the first line of (3.6) gives the Kullback-Leibler divergence $D_{\mathrm{KL}}\left(N(x^{i_0}, \mathbf{\Sigma}) \,||\, N(x^{i_0}, (D^2V_1(x^{i_0}))^{-1})\right)$. The second line of (3.6) is equal to the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mathbf{e^i} \,||\, \boldsymbol{\beta})$, for $\mathbf{e^i} := (\mathbf{0}, \cdots, \mathbf{1}, \cdots, \mathbf{0})$. In conclusion,

eq:F0-2

$$F_0(x^i, \mathbf{\Sigma}) = D_{\mathrm{KL}}\left(N(x^i, \mathbf{\Sigma}) \,||\, N(x^i, (D^2V_1(x^i))^{-1})\right) + D_{\mathrm{KL}}(\mathbf{e^i} \,||\, \boldsymbol{\beta}), \tag{3.7}$$

in other words, in the limit $\varepsilon \downarrow 0$, the Kullback-Leibler divergence between the best Gaussian measure ν_ε and the measure μ_ε consists of two parts: the first part is the relative entropy between the Gaussian measure with rescaled covariance $\mathbf{\Sigma}$ and the Gaussian measure with covariance determined by $(D^2V_1(x^i))^{-1}$; the second part is the relative entropy between the Dirac mass supported at x^i and a weighted sum of Dirac masses, with weights $\boldsymbol{\beta}$, at the $\{x^j\}_{j=1}^n$. Clearly, to minimize $F_0(m, \mathbf{\Sigma})$, on the one hand, we need to choose $m = x^i$ and $\mathbf{\Sigma} = D^2V_1(x^i))^{-1}$ for some $i \in 1, \cdots, n$; for this choice the first term on the right side of (3.6) vanishes. In order to minimize the second term we need to choose the minimum x^i with maximal weight β^i. In particular, the following corollary holds.

cor:sgaussian

Corollary 3.5. *The minimum of F_0 is zero when $n = 1$, but it is strictly positive when $n > 1$.*

Corollary 3.5 reflects the fact that, in the limit $\varepsilon \downarrow 0$, a single Gaussian measure is not the best choice for approximating a non-Gaussian measure with multiple modes; this motivates our study of Gaussian mixtures in Section 4.

The proofs of Theorem 3.3 and Corollary 3.4 are provided after establishing a sequence of lemmas. The following lemma shows that the functional F_ε is compact (recall Definition 2.1). It is well known, that the Kullback-Leibler divergence (with respect to a fixed reference μ) has compact sub-level sets with respect to weak convergence of probability measures. Here we prove a stronger statement, which is specific to the family of reference measures μ_ε, namely a uniform bound from above and below for the rescaled covariances, i.e. we prove a bound from above and below for $\mathbf{\Sigma}_\varepsilon$ if we control $F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$.

lem:compactness

Lemma 3.6. *Assume that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. Then*

eq:compactness

$$0 < \liminf_{\varepsilon\downarrow 0} \lambda_{\min}(\mathbf{\Sigma}_\varepsilon) < \limsup_{\varepsilon\downarrow 0} \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) < \infty \tag{3.8}$$

and $\mathrm{dist}(m_\varepsilon, \mathscr{E}) \downarrow 0$ as $\varepsilon \downarrow 0$. In particular, there exist common subsequences $\{m_k\}_{k\in\mathbb{N}}$ of $\{m_\varepsilon\}$, $\{\mathbf{\Sigma}_k\}_{k\in\mathbb{N}}$ of $\{\mathbf{\Sigma}_\varepsilon\}$ such that $m_k \to x_{i_0}$ with $1 \le i_0 \le n$ and $\mathbf{\Sigma}_k \to \mathbf{\Sigma} \in \mathcal{S}_>(\mathbf{R}, d)$.

Proof. Let $M := \limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. Since m_ε and $\mathbf{\Sigma}_\varepsilon$ are defined in finite dimensional spaces, we only need to show that both sequences are uniformly bounded. The proof consists of the following steps.

Step 1. We first prove the following rough bounds for $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)$: there exists positive constants C_1, C_2 such that when $\varepsilon \ll 1$,

eq:bdsigma0

$$C_1 \le \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) \le \frac{C_2}{\varepsilon}. \tag{3.9}$$

In fact, from the formula (3.4) and the assumption that V_1^ε and V_2 are non-negative, we can get that when $\varepsilon \ll 1$

eq:logdetsig
$$\log(\det \mathbf{\Sigma}_\varepsilon) \geq 2(C_V - M - 1) \tag{3.10}$$

where the constant

$$C_V := -\frac{d}{2} + \log\left(\sum_{i=1}^n \beta^i\right).$$

Then the lower bound of (3.9) follows from (3.10) and the arithmetic-geometric mean inequality

eq:trace-det
$$\det \mathbf{A} \leq \left(\frac{1}{d}\mathrm{Tr}(\mathbf{A})\right)^d \tag{3.11}$$

which holds for any positive definite $\mathbf{A}$. In addition, using the condition (A-5) for the potential V_1^ε, we obtain from (3.4) that when $\varepsilon \ll 1$,

eq:bdsigma1
$$\begin{aligned}
M &\geq F_\varepsilon(m_\varepsilon, \mathbf{A}_\varepsilon)\\
&\geq \mathbb{E}^{\nu_\varepsilon} V_2(x) + \frac{c_1}{\varepsilon}\mathbb{E}^{\nu_\varepsilon}|x|^2 - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) + C_V - 1\\
&= \mathbb{E}^{\nu_\varepsilon} V_2(x) + c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) + \frac{c_1|m_\varepsilon|^2}{\varepsilon} - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right) + C_V - 1\\
&\geq c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{c_0}{\varepsilon} - \frac{1}{2}\log\left(\left(\frac{1}{d}\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)\right)^d\right) + C_V - 1\\
&= c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{c_0}{\varepsilon} - \frac{d}{2}\log(\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)) + \frac{d\log d}{2} + C_V - 1,
\end{aligned} \tag{3.12}$$

where we have used the inequality (3.11) and the assumption that V_2 is non-negative. Dropping the non-negative terms on the right hand side we rewrite this expression as an estimate on $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)$,

$$c_1 \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) - \frac{d}{2}\log(\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon)) \leq M + \frac{c_0}{\varepsilon} + 1,$$

and conclude that there exists $C_2 > 0$ such that $\mathrm{Tr}(\mathbf{\Sigma}_\varepsilon) \leq C_2/\varepsilon$ by observing that for $x \gg 1$ we have $c_1 x - \frac{d}{2}\log x \geq \frac{c_1}{2}x$.

Step 2. In this step we show that for $\varepsilon \ll 1$ the mass of ν_ε concentrates near the minimizers. More precisely, we claim that there exist constants $R_1, R_2 > 0$, such that for every $\varepsilon \ll 1$ there exists an index $i_0 \in \{1, 2, \cdots, n\}$ such that

eq:bsigma2
$$\nu_\varepsilon\left(B\left(x_{i_0}, \sqrt{\varepsilon(R_1 + R_2\log\left(\det \mathbf{\Sigma}_\varepsilon\right))}\right)\right) \geq \frac{1}{2n}. \tag{3.13}$$

On the one hand, from the expression (3.4) and the assumption that for all $r \ll 1$ $\limsup_{\varepsilon \downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) \leq M$ we know that there exist $C_3, C_4 > 0$ such that when $\varepsilon \ll 1$

eq:bsigma3
$$\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) \leq \varepsilon\left(C_3 + C_4\log\left(\det \mathbf{\Sigma}_\varepsilon\right)\right). \tag{3.14}$$

On the other hand, it follows from (2.2) that for $\eta \ll 1$

$$\begin{aligned}
\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) &\geq \mathbb{E}^{\nu_\varepsilon}\left[V_1^\varepsilon(x)\mathbf{I}_{(\cup_{i=1}^n B(x_i,\eta))^c}(x)\right]\\
&\geq C_\delta \eta^2 \nu_\varepsilon(\cup_{i=1}^n B(x_i,\eta))^c,
\end{aligned} \tag{3.15}$$

which combines with (3.14) to

jji
$$\nu_\varepsilon(\cup_{i=1}^n B(x_i,\eta))^c) \leq \varepsilon\frac{(C_3 + C_4\log\left(\det \mathbf{\Sigma}_\varepsilon\right))}{C_\delta \eta^2}. \tag{3.16}$$

Now we choose $\eta = \eta_\varepsilon := \sqrt{2\varepsilon(C_3 + C_4 \log(\det \mathbf{\Sigma}_\varepsilon))/C_\delta}$ (by the rough bound (3.9) this η_ε tends to zero as $\varepsilon \to 0$, which permits to apply (2.2)). This implies (3.13) with $R_1 = \frac{2C_3}{C_\delta}$ and $R_2 = \frac{2C_4}{C_\delta}$, by passing to the complement and observing that

$$\sup_{i \in \{1,\dots,n\}} \nu_\varepsilon(B(x_i, \eta_\varepsilon)) \geq \frac{1}{n} \nu_\varepsilon \left(\cup_{i \in \{1,\dots,n\}} B(x_i, \eta_\varepsilon) \right).$$

Step 3. We prove the bounds (3.8). As in the previous step we set $\eta_\varepsilon = \sqrt{\varepsilon(R_1 + R_2 \log(\det \mathbf{\Sigma}_\varepsilon))}$. It follows from (3.13) that

eq:bsigma4

$$\begin{aligned}
\frac{1}{2n} &\leq \nu_\varepsilon(B(x_{i_0}, \eta_\varepsilon)) \\
&= \frac{1}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} \int_{B(x_{i_0}, \eta_\varepsilon)} \exp\left(-\frac{1}{2\varepsilon} \langle x - m_\varepsilon, \mathbf{\Sigma}_\varepsilon^{-1}(x - m_\varepsilon) \rangle \right) dx \\
&\leq \frac{1}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} |B(x_{i_0}, \eta_\varepsilon)| \\
&\leq C \frac{1}{\sqrt{\varepsilon^d \det \mathbf{\Sigma}_\varepsilon}} \eta_\varepsilon^d \leq C \sqrt{\frac{R_1 + R_2 \log(\det \mathbf{\Sigma}_\varepsilon)}{\det \mathbf{\Sigma}_\varepsilon}}.
\end{aligned} \tag{3.17}$$

This implies that $\limsup_{\varepsilon \downarrow 0} \det \mathbf{\Sigma}_\varepsilon < C$ for some $C > 0$. In order to get a lower bound on individual eigenvalues $\mathbf{\Lambda}_\varepsilon^{(i)}$ of $\mathbf{\Sigma}_\varepsilon$, we rewrite the same integral in a slightly different way. We use the change of coordinates $y = \frac{\mathbf{P}_\varepsilon^T(x-m)}{\sqrt{\varepsilon}}$, where $\mathbf{P}_\varepsilon$ is orthogonal and diagonalises Σ and observe that under this transformation $B(x^i, \eta_\varepsilon)$ is mapped into $B(\frac{x^i - m}{\sqrt{\varepsilon}}, \frac{\eta_\varepsilon}{\varepsilon}) \subseteq \{y \colon |y_j - \frac{(x^i - m)}{\sqrt{\varepsilon}}| \leq \frac{\eta_\varepsilon}{\sqrt{\varepsilon}} \quad \text{for } j = 1, \dots, n\}$. This yields

eq:logdet

$$\begin{aligned}
\frac{1}{2n} &\leq \frac{1}{\sqrt{(2\pi)^d \det \mathbf{\Sigma}_\varepsilon}} \int_{\{|y_j - \frac{(x^i - m)}{\sqrt{\varepsilon}}| \leq \frac{\eta_\varepsilon}{\sqrt{\varepsilon}}\}} \exp\left(-\frac{1}{2} \langle y_i, (\mathbf{\Lambda}_\varepsilon^{(i)})^{-1} y_i \rangle \right) dy \\
&\leq \frac{1}{\sqrt{(2\pi)^d \det \mathbf{\Sigma}_\varepsilon}} \left(\frac{2\eta_\varepsilon}{\sqrt{\varepsilon}} \right)^{d-1} \int_{\mathbf{R}} \exp\left(-\frac{|y_i|^2}{2\mathbf{\Lambda}_\varepsilon^{(i)}} \right) dy_i \\
&= \sqrt{\frac{\mathbf{\Lambda}_\varepsilon^{(i)}}{(2\pi)^d \det \mathbf{\Sigma}_\varepsilon}} \left(R_1 + R_2 \log(\det \mathbf{\Sigma}_\varepsilon) \right)^{\frac{d-1}{2}},
\end{aligned} \tag{3.18}$$

for any $i \in \{1, 2, \cdots, d\}$. Together with uniform boundedness of $\det \mathbf{\Sigma}_\varepsilon$ this implies that $\mathbf{\Lambda}_\varepsilon^{(i)} > C'$ for some $C' > 0$. Finally,

eq:upptracedet

$$\operatorname{Tr}(\mathbf{\Sigma}_\varepsilon) = \sum_{i=1}^d \mathbf{\Lambda}_\varepsilon^{(i)} = \sum_{i=1}^d \frac{\det(\mathbf{\Sigma}_\varepsilon)}{\prod_{j=1, j \neq i}^d \mathbf{\Lambda}_\varepsilon^{(j)}} \leq \frac{dC}{(C')^{d-1}} < \infty. \tag{3.19}$$

This proves (3.8).

Step 4. We show that $\operatorname{dist}(m_\varepsilon, \mathscr{E}) \downarrow 0$ as $\varepsilon \downarrow 0$. On the one hand, by the upper bound on the variance in (3.8) and standard Gaussian concentration, we see that there exists a constant $c > 0$, such that for $\varepsilon \ll 1$ we have $\nu_\varepsilon(B(m_\varepsilon, \sqrt{\varepsilon}c)) \geq \frac{3}{4}$. On the other hand, we had already seen in (3.16) that for $\eta = \eta_\varepsilon$ we have

$$\nu_\varepsilon(\cup_{i=1}^n B(x_i, \eta_\varepsilon))^c) \leq \frac{1}{2},$$

and hence $B(m_\varepsilon, \sqrt{\varepsilon}c)$ must intersect at least one of the $B(x_i, \eta_\varepsilon)$. This yields for this particular index i

$$|x_i - m_\varepsilon| \leq \eta_\varepsilon + \sqrt{\varepsilon}c,$$

and establishes the claim.

$\square$

lem:dkl-asym2 **Lemma 3.7.** *Let* $\{(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)\}$ *be a sequence such that* $\limsup_{\varepsilon\downarrow 0} |m_\varepsilon| =: C_1 < \infty$ *and*

$$0 < c_2 := \liminf_{\varepsilon\downarrow 0} \lambda_{\min}(\mathbf{\Sigma}_\varepsilon) < \limsup_{\varepsilon\downarrow 0} \operatorname{Tr}(\mathbf{\Sigma}_\varepsilon) =: C_2 < \infty.$$

Then as $\varepsilon \downarrow 0$*,*

eq:dkl-asym2 (3.20)

$$\begin{aligned} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} &+ V_2(m_\varepsilon) + \frac{1}{2}\operatorname{Tr}(D^2 V_1^\varepsilon(m_\varepsilon)\cdot \mathbf{\Sigma}_\varepsilon) - \frac{1}{2}\log\left((2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon\right)\\ &- \frac{d}{2} + \log Z_{\mu,\varepsilon} + r_\varepsilon \end{aligned}$$

where $|r_\varepsilon| \le C\varepsilon$ *with* $C = C(C_1, c_2, C_2, M_V)$ *(Recall that* M_V *is the constant defined in Assumptions 2.3 (A-1)).*

Proof. The lemma follows directly from the expression (3.2) and Taylor expansion. Indeed, we first expand V_2 near m_ε up to the first order and then take expectation to get

$$\mathbb{E}^{\nu_\varepsilon} V_2(x) = V_2(m_\varepsilon) + \mathbb{E}^{\nu_\varepsilon} R_\varepsilon(x)$$

with residual

$$R_\varepsilon(x) = \sum_{|\alpha|=2} \frac{(x-m_\varepsilon)^\alpha}{\alpha!}\int_0^1 \partial^\alpha V_2\left(\xi x + (1-\xi)m_\varepsilon\right)(1-\xi)^2 d\xi.$$

Thanks to the condition (A-1), one can obtain the bound

eq:res0 (3.21)

$$\begin{aligned} \mathbb{E}^{\nu_\varepsilon} R_\varepsilon(x) &\le \sum_{|\alpha|=2} \frac{1}{\alpha!}\max_{\xi\in[0,1]}\left\{\mathbb{E}^{\nu_\varepsilon}\left[|x-m_\varepsilon|^2 \partial^\alpha V_2\left(\xi x + (1-\xi)m_\varepsilon\right)\right]\right\}\\ &\le \frac{M_V}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} \max_{\xi\in[0,1]}\left\{\int_{\mathbf{R}^d} |x|^2 e^{(|x|+|m_\varepsilon|)^2}\cdot e^{-\frac{1}{2\varepsilon}x^T\mathbf{\Sigma}_\varepsilon^{-1}x}dx\right\}\\ &\le \frac{M_V}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon}} e^{2|m_\varepsilon|^2}\int_{\mathbf{R}^d}|x|^2 e^{-\frac{1}{2\varepsilon}x^T(\mathbf{\Sigma}_\varepsilon^{-1}-4\varepsilon\cdot\mathbf{I}_d)x}dx\\ &= \frac{M_V\varepsilon}{\sqrt{\det \mathbf{\Sigma}_\varepsilon}} e^{2|m_\varepsilon|^2}\det(\mathbf{\Sigma}_\varepsilon^{-1} - 4\varepsilon\cdot\mathbf{I}_d)^{-1}\\ &\le C\varepsilon, \end{aligned}$$

when $\varepsilon \ll 1$. Note that in the last inequality we have used the assumption that all eigenvalues of $\mathbf{\Sigma}_\varepsilon$ are bounded from above which ensures that for $\varepsilon \ll 1$ the matrix $\mathbf{\Sigma}_\varepsilon^{-1} - 4\varepsilon\cdot\mathbf{I}_d$ is positive definite. Hence

$$\mathbb{E}^{\nu_\varepsilon} V_2(x) = V_2(m_\varepsilon) + r_{1,\varepsilon}$$

with $r_{1,\varepsilon} \le C\varepsilon$ as $\varepsilon \downarrow 0$. Similarly, one can take the fourth order Taylor expansion for V_1^ε near m_ε and then take expectation to obtain that

$$\mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} + \frac{1}{2}\operatorname{Tr}\left(D^2 V_1^\varepsilon(m_\varepsilon)\cdot\mathbf{\Sigma}_\varepsilon\right) + r_{2,\varepsilon}$$

with $r_{2,\varepsilon} \le C\varepsilon$. Then (3.20) follows directly by inserting the above equations into the expression (3.2). □

The following corollary is a direct consequence of Lemma 3.2, Lemma 3.6 and Lemma 3.7, providing an asymptotic formula for $F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$ as $\varepsilon \downarrow 0$.

Corollary 3.8. *Assume that* $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$*. Then for* $\varepsilon \ll 1$

eq:dkl-asym22 (3.22)

$$\begin{aligned} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \frac{V_1^\varepsilon(m_\varepsilon)}{\varepsilon} &+ V_2(m_\varepsilon) + \frac{1}{2}\operatorname{Tr}(D^2 V(m_\varepsilon)\cdot\mathbf{\Sigma}_\varepsilon) - \frac{1}{2}\log\left(\det \mathbf{\Sigma}_\varepsilon\right)\\ &- \frac{d}{2} + \sum_{i=1}^n \beta^i + o(1). \end{aligned}$$

rem:3-9 *Remark* 3.9. We do not have a bound on the convergence rate for the residual expression (3.22), because Lemma 3.2 does not provide a convergence rate on the $Z_{\mu,\varepsilon}$. This is because we do not impose any rate of convergence for the convergence of the x^i_ε to x^i. The bound $|r_\varepsilon| \leq C\varepsilon$ in Lemma 3.7 will be used to prove the rate of convergence for the posterior measures that arise from Bayesian inverse problems; see Theorem 5.4 in Section 5, and its proof.

Proof of Theorem 3.3. We first prove the liminf inequality. Let $(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$ be such that $m_\varepsilon \to m$ and $\mathbf{\Sigma}_\varepsilon \to \mathbf{\Sigma}$. We want to show that $F(m, \mathbf{\Sigma}) \leq \liminf_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon)$. We may assume that $\liminf_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$ since otherwise there is nothing to prove. By Lemma 3.6 this implies that $m \in \mathscr{E}$ and that $\mathbf{\Sigma}$ is positive definite. Then the liminf inequality follows from (3.22) and the fact that $V_1^\varepsilon \geq 0$.

Next we show the limsup inequality is true. Given $m \in \mathscr{E}, \mathbf{\Sigma} \in \mathcal{S}_>(\mathbf{R}, d)$, we want to find recovery sequences $(m_k, \mathbf{\Sigma}_k)$ such that $(m_k, \mathbf{\Sigma}_k) \to (m, \mathbf{\Sigma})$ and $\limsup_k F_{\varepsilon_k}(m_k, \mathbf{\Sigma}_k) \leq F(m, \mathbf{\Sigma})$. In fact, we set $\mathbf{\Sigma}_k = \mathbf{\Sigma}$. Moreover, by Assumptions 2.3 (A-4), we can choose $\{m_k\}$ to be one of the zeros of $V_1^{\varepsilon_k}$ so that $V_1^{\varepsilon_k}(m_k) = 0$ and $m_k \to m \in \mathscr{E}$. This implies that $V_2(m_k) \to V_2(m)$. Then the limsup inequality follows from (3.22). ☐

Proof of Corollary 3.4. First we show that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(m_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. In fact, let $\tilde{m}_\varepsilon = x^1_\varepsilon$ and $\tilde{\mathbf{\Sigma}}_\varepsilon = D^2 V_1^\varepsilon(x^1_\varepsilon)$. It follows from (3.22) that $\limsup_{\varepsilon\downarrow 0} F_\varepsilon(\tilde{m}_\varepsilon, \tilde{\mathbf{\Sigma}}_\varepsilon) < \infty$. According to Theorem 2.2, the convergence of minima and minimizers is a direct consequence of Lemma 3.6 and Theorem 3.3. ☐

sec:gaussianmix

4. Approximation by Gaussian mixtures

In the previous section we demonstrated the approximation of the target measure (1.1) by a Gaussian. Corollary 3.5 shows that, when the measure has only one mode, this approximation is perfect in the limit $\varepsilon \to 0$: the limit KL-divergence tends to zero since both entropies in (3.7) tend to zero. However when multiple modes exist, and persist in the small ε limit, the single Gaussian is inadequate because the relative entropy term $D_{\mathrm{KL}}(\mathbf{e_i}||\boldsymbol{\beta})$ can not be small even though the relative entropy between Gaussians tends to zero. In this section we consider the approximation of the target measure μ_ε by Gaussian mixtures in order to overcome this issue. We show that in the case of n minimizers of V_1, the approximation with a mixture of n Gaussians is again perfect as $\varepsilon \to 0$. The Gaussian mixture model is widely used in the pattern recognition and machine learning community; see the relevant discussion in [3, Chapter 9].

Let $\triangle^n$ be the standard n-simplex, i.e.,

$$\triangle^n = \left\{ \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \cdots, \alpha^n) \in \mathbf{R}^n : \alpha^i \geq 0 \text{ and } \sum_{i=1}^n \alpha^i = 1 \right\}.$$

For $\xi \in (0,1)$, we define $\triangle^n_\xi = \{\boldsymbol{\alpha} = (\alpha^1, \alpha^2, \cdots, \alpha^n) \in \mathbf{R}^n : \alpha^i \geq \xi\}$.

Recall that $\mathcal{A}$ is the set of Gaussian measures and define the set of Gaussian mixtures

eq:Gaussmix
$$\mathcal{M}_n = \left\{ \nu = \sum_{i=1}^n \alpha^i \nu^i : \nu^i \in \mathcal{A}, \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \cdots, \alpha^n) \in \triangle^n \right\}. \tag{4.1}$$

Also, for a fixed $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0, \infty)$ we define the set

eq:Gaussmix-delta
$$\begin{aligned} \mathcal{M}_n^{\boldsymbol{\xi}} = \Big\{ \nu = \sum_{i=1}^n \alpha^i \nu^i : \nu^i = N(m^i, \mathbf{\Sigma}^i) \in \mathcal{A} \text{ with } \min_{i\neq j} |m^i - m^j| \geq \xi_2, \\ \boldsymbol{\alpha} = (\alpha^1, \alpha^2, \cdots, \alpha^n) \in \triangle^n_{\xi_1} \Big\}. \end{aligned} \tag{4.2}$$

While $\mathcal{M}_n$ is the set of all convex combinations of n Gaussians taken from $\mathcal{A}$; the set $\mathcal{M}_n^{\boldsymbol{\xi}}$ can be seen as an "effective" version of $\mathcal{M}_n$, in which each Gaussian component plays an active role, and no two Gaussians share a common center.

Consider the problem of minimizing $D_{\mathrm{KL}}(\nu||\mu_\varepsilon)$ within $\mathcal{M}_n$ or $\mathcal{M}_n^{\boldsymbol{\xi}}$. Since the sets $\mathcal{M}_n$ and $\mathcal{M}_n^{\boldsymbol{\xi}}$ are both closed with respect to weak convergence, we have the following existence result whose proof is similar to Theorem 3.1 and is omitted.

thm:exists2 **Theorem 4.1.** *Consider the measure μ_ε given by (1.1) with fixed $\varepsilon > 0$, and the problems of minimizing the functional*

eq:minfunc
$$\nu \mapsto D_{KL}(\nu||\mu_\varepsilon) \tag{4.3}$$

from the set $\mathcal{M}_n$, or from the set $\mathcal{M}_n^{\boldsymbol{\xi}}$ with some fixed $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0,\infty)$. In both cases, there exists at least one minimizer to the functional (4.3).

Now we continue to investigate the asymptotic behavior of the Kullback-Leibler approximations based on Gaussian mixtures. To that end, we again parametrize a measure ν in the set $\mathcal{M}_n$ or $\mathcal{M}_n^{\xi}$ by the weights $\boldsymbol{\alpha} = (\alpha^1, \alpha^2, \cdots, \alpha^n)$ as well as the n means as well as the n covariances matrices. Similar to the previous section we need to chose the right scaling in our Gaussian mixtures to reflect the typical size of fluctuations of μ_ε. Thus for $\mathbf{m} = (m^1, m^2, \cdots, m^n)$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^1, \boldsymbol{\Sigma}^2, \cdots, \boldsymbol{\Sigma}^n)$. we set

eq:nu-form
$$\nu_\varepsilon = \sum_{i=1}^n \alpha^i N(m^i, \varepsilon \boldsymbol{\Sigma}^i). \tag{4.4}$$

We can view $D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon)$ as a functional of $(\boldsymbol{\alpha}, \mathbf{m}, \boldsymbol{\Sigma})$ and study the Γ-convergence of the resulting functional. For that purpose, we need to restrict our attention to finding the best Gaussian mixtures within $\mathcal{M}_n^{\boldsymbol{\xi}}$ for some $\boldsymbol{\xi} \in (0,1) \times (0,\infty)$. The reasons are the following. First, we require individual Gaussian measures ν^i to be active (i.e. $\alpha^i > \xi_1 > 0$) because $D_{\mathrm{KL}}(\nu_\varepsilon, \mu_\varepsilon)$, as a family of functionals of $(\boldsymbol{\alpha}, \mathbf{m}, \boldsymbol{\Sigma})$ indexed by ε, is not compact if we allow some of the α^i to vanish. In fact, if $\alpha^i_\varepsilon = 0$ for some $i \in 1, 2, \cdots, n$, then $D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon)$ is independent of m^i_ε and $\boldsymbol{\Sigma}^i_\varepsilon$. In particular, if $|m^i_\varepsilon| \wedge |\boldsymbol{\Sigma}^i_\varepsilon| \to \infty$ while $|m^j_\varepsilon| \vee |\boldsymbol{\Sigma}^j_\varepsilon| < \infty$ for all the j's such that $j \neq i$, then it still holds that $\limsup_{\varepsilon \downarrow 0} D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon) < \infty$. Second, it makes more sense to assume that the individual Gaussian means stay apart from each other (i.e. $\min_{i \neq j} |m^i - m^j| \geq \xi_2 > 0$) since we primarily want to locate different modes of the target measure. Moreover, it seems impossible to identify a sensible Γ-limit without such an assumption; see Remark 4.7.

Recall that the measure ν has the form (4.4). Let $\boldsymbol{\xi} = (\xi_1, \xi_2) \in (0,1) \times (0,\infty)$ be fixed. In view of these considerations it is useful to define

$$S_{\boldsymbol{\xi}} = \{(\boldsymbol{\alpha}, \mathbf{m}) \in \triangle^n_{\xi_1} \times \mathbf{R}^{nd} : \min_{i \neq j} |m^i - m^j| \geq \xi_2.\}$$

We define the functional

eq:dkl-2
$$G_\varepsilon(\boldsymbol{\alpha}, \mathbf{m}, \boldsymbol{\Sigma}) := \begin{cases} D_{\mathrm{KL}}(\nu||\mu_\varepsilon) & \text{if } (\boldsymbol{\alpha}, \mathbf{m}) \in S_{\boldsymbol{\xi}}, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.5}$$

By the definition of the Kullback-Leibler divergence, if $(\boldsymbol{\alpha}, \mathbf{m}) \in S_{\boldsymbol{\xi}}$, then

eq:dkl-3
$$G_\varepsilon(\boldsymbol{\alpha}, \mathbf{m}, \boldsymbol{\Sigma}) = \int \rho(x) \log \rho(x) dx + \frac{1}{\varepsilon} \mathbb{E}^\nu V_1^\varepsilon(x) + \mathbb{E}^\nu V_2(x) + \log Z_{\mu,\varepsilon} \tag{4.6}$$

where ρ is the probability density function (p.d.f) of ν.

Recall the Γ-limit F defined in (3.5). Then we have the following Γ-convergence result.

thm:gamma-2 **Theorem 4.2.** *The* Γ*-limit of* G_ε *is*

eq:limit-G
$$G(\boldsymbol{\alpha},\mathbf{m},\boldsymbol{\Sigma}) := \sum_{i=1}^{n} \alpha^i D_{KL}\left(N(m^i,\boldsymbol{\Sigma}^i) \mid\mid N(m^i,(D^2V_1(m^i))^{-1})\right) + D_{KL}(\boldsymbol{\alpha} \mid\mid \boldsymbol{\beta}) \tag{4.7}$$

if $(\boldsymbol{\alpha},\mathbf{m}) \in S_{\boldsymbol{\xi}}$ *and* $m^i \in \mathscr{E}$*, and* ∞ *otherwise.*

Remark 4.3. The right hand side of G consists of two parts: the first part is a weighted relative entropy which measures the discrepancy between two Gaussians, and the second part is the relative entropy between sums of Dirac masses at $\{x^j\}_{j=1}^n$ with weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. This has the same spirit as the entropy splitting used in [18, Lemma 2.4].

Before we prove Theorem 4.2, we consider the minimization of the limit functional G. First let ξ_2 be such that $0 < \xi_2 \leq \min_{i\neq j}|x^i - x^j|$ where $\{x^i\}_{i=1}^n$ are the minimizers of V_1. To minimize G, without loss of generality, we may choose $m^i = \overline{m}^i := x^i$. Then the weighted relative entropy in the first term in the definition (4.7) of G vanishes if we set $\boldsymbol{\Sigma}^i = \overline{\boldsymbol{\Sigma}}^i := D^2V_1(x^i)^{-1}$. The relative entropy of the weights also vanishes if we choose the weight $\boldsymbol{\alpha} = \overline{\boldsymbol{\alpha}} := \boldsymbol{\beta}$. To summarize, the minimizer $(\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}})$ of G is given by

eq:minimizermix
$$\overline{m}^i = x^i,\ \overline{\boldsymbol{\Sigma}}^i = D^2V_1(x^i)^{-1}, \overline{\alpha}^i = \beta^i, \tag{4.8}$$

and $G(\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}}) = 0$. The following corollary is a direct consequence of the Γ-convergence of G_ε.

cor:convmin **Corollary 4.4.** *Let* $\{(\boldsymbol{\alpha}_\varepsilon,\mathbf{m}_\varepsilon,\boldsymbol{\Sigma}_\varepsilon)\}$ *be a family of minimizers of* $\{G_\varepsilon\}$*. Then there exists a subsequence* $\{\varepsilon_k\}$ *such that* $(\boldsymbol{\alpha}_{\varepsilon_k},\mathbf{m}_{\varepsilon_k},\boldsymbol{\Sigma}_{\varepsilon_k}) \to (\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}})$ *and that* $G_{\varepsilon_k}(\boldsymbol{\alpha}_{\varepsilon_k},\mathbf{m}_{\varepsilon_k},\boldsymbol{\Sigma}_{\varepsilon_k}) \to G(\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}})$*. Moreover,* $(\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}})$ *is a minimizer of* G *and* $G(\overline{\boldsymbol{\alpha}},\overline{\mathbf{m}},\overline{\boldsymbol{\Sigma}}) = 0$.

For a non-Gaussian measure μ_ε with multiple modes, i.e., $n > 1$ in the Assumption 2.3, we have seen in Remark 3.5 that the Kullback-Leibler divergence between μ_ε and the best Gaussian measure selected from $\mathcal{A}$ remains positive as $\varepsilon \downarrow 0$. However, this gap is filled by using Gaussian mixtures, namely, with ν_ε being chosen as the best Gaussian mixture, the Kullback-Leibler divergence $D_{\mathrm{KL}}(\nu_\varepsilon||\mu_\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$.

Similarly to the proof of Theorem 3.3, Theorem 4.2 follows directly from Corollary 4.8 below, the proof of which requires several lemmas. We first show the compactness of $\{G_\varepsilon\}$.

lem:compactness2 **Lemma 4.5.** *Let* G_ε *be defined by* (4.5)*. Let* $\{(\boldsymbol{\alpha}_\varepsilon,\mathbf{m}_\varepsilon,\boldsymbol{\Sigma}_\varepsilon)\}$ *be a sequence such that* $\limsup_{\varepsilon\downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon,\mathbf{m}_\varepsilon,\boldsymbol{\Sigma}_\varepsilon) < \infty$*. Then*

eq:compactness2
$$\liminf_{\varepsilon\downarrow 0}\min_i \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon^i) > 0,\ \limsup_{\varepsilon\downarrow 0}\max_i \mathrm{Tr}(\boldsymbol{\Sigma}_\varepsilon^i) < \infty \tag{4.9}$$

and dist$(m_\varepsilon^i,\mathscr{E}) \downarrow 0$ *as* $\varepsilon \downarrow 0$*. In particular, for any* i*, there exists* $j = j(i) \in \{1,2,\cdots,n\}$ *and a subsequence* $\{m_k^i\}_{k\in\mathbb{N}}$ *of* $\{m_\varepsilon^i\}$ *such that* $m_k \to x_j$ *as* $k \to \infty$*.*

Proof. We write $M = \limsup_{\varepsilon\downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon,\mathbf{m}_\varepsilon,\boldsymbol{\Sigma}_\varepsilon)$ and

$$\nu_\varepsilon = \sum_{i=1}^{n} \alpha_\varepsilon^i \nu_\varepsilon^i$$

where $\nu_\varepsilon^i = N(m_\varepsilon^i, \varepsilon \mathbf{\Sigma}_\varepsilon^i)$. Then we get

$$
\begin{aligned}
D_{KL}(\nu_\varepsilon || \mu_\varepsilon) &= \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log \left(\sum_i \alpha_\varepsilon^i \frac{d\nu_\varepsilon^i}{d\mu_\varepsilon} \right) \\
&\geq \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log \left(\alpha_\varepsilon^j \frac{d\nu_\varepsilon^j}{d\mu_\varepsilon} \right) \\
&= \sum_{j=1}^n \alpha_\varepsilon^j \log(\alpha_\varepsilon^j) + \sum_{j=1}^n \alpha_\varepsilon^j \, \mathbb{E}^{\nu_\varepsilon^j} \log \left(\frac{d\nu_\varepsilon^j}{d\mu_\varepsilon} \right) \\
&= \sum_{j=1}^n \alpha_\varepsilon^j \log(\alpha_\varepsilon^j) + \sum_{j=1}^n \alpha_\varepsilon^j D_{KL}(\nu_\varepsilon^j || \mu_\varepsilon)
\end{aligned}
$$

where the inequality follows simply from the monotonicity of the logarithm. As each of term $D_{KL}(\nu_\varepsilon^j || \mu_\varepsilon)$ is non-negative, this implies the bound

$$
D_{KL}(\nu_\varepsilon^j || \mu_\varepsilon) \leq \frac{1}{\alpha_\varepsilon^j} \left(M - n \min_{\alpha \in [0,1]} \alpha \log \alpha \right).
$$

Using the lower bound $\alpha_\varepsilon^j > \xi_1$ which holds by assumption we get a uniform upper bound on $D_{KL}(\nu_\varepsilon^j || \mu_\varepsilon)$ which in turn permits to invoke Lemma 3.6.

□

lem:dkl-asym3 **Lemma 4.6.** *Let $\{(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon)\}$ be such that $\boldsymbol{\alpha}_\varepsilon \in \triangle_{\xi_1}^n$, $\min_{i \neq j} |m_\varepsilon^i - m_\varepsilon^j| \geq \xi_2 > 0$ and that $c_1 \leq \liminf_{\varepsilon \downarrow 0} \min_i \lambda_{\min}(\mathbf{\Sigma}_\varepsilon^i) < \limsup_{\varepsilon \downarrow 0} \max_i |m_\varepsilon^i| \vee \mathrm{Tr}(\mathbf{\Sigma}_\varepsilon^i) \leq C_1 < \infty$. Then*

eq:dkl-asym3

$$
\begin{aligned}
&G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) \\
&= \sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + V_2(m_\varepsilon^i) + \frac{1}{2} \mathrm{Tr}(D^2 V_1^\varepsilon(m_\varepsilon^i) \cdot \mathbf{\Sigma}_\varepsilon^i) - \frac{1}{2} \log \left(\det \mathbf{\Sigma}_\varepsilon^i \right) \right) \\
&+ \sum_{i=1}^n \alpha_\varepsilon^i \log \alpha_\varepsilon^i - \frac{d}{2} + \log Z_{\mu,\varepsilon} + r_\varepsilon.
\end{aligned}
\tag{4.10}
$$

where $r_\varepsilon \leq C\varepsilon$ with $C = C(c_1, C_1, M_V, \xi_2)$.

Proof. By assumption, we we know from (4.6) that

$$
G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) = \int \rho_\varepsilon(x) \log \rho_\varepsilon(x) dx + \frac{1}{\varepsilon} \mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) + \log Z_{\mu,\varepsilon}
$$

where $\rho_\varepsilon = \sum_{i=1}^n \alpha_\varepsilon^i \rho_\varepsilon^i$ is the probability density of the measure ν_ε. First of all, applying the same Taylor expansion arguments used to obtain (3.20), one can deduce that

eq:expV

$$
\begin{aligned}
&\frac{1}{\varepsilon} \mathbb{E}^{\nu_\varepsilon} V_1^\varepsilon(x) + \mathbb{E}^{\nu_\varepsilon} V_2(x) \\
&= \sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + \frac{1}{2} \mathrm{Tr} \left(\nabla^2 V_1^\varepsilon(m_\varepsilon^i) \cdot \mathbf{\Sigma}_\varepsilon^i \right) + V_2(m_\varepsilon^i) \right) + r_{1,\varepsilon}
\end{aligned}
\tag{4.11}
$$

with $r_{1,\varepsilon} \leq C\varepsilon$ and $C = C(C_1, c_1, M_V)$. Next, we claim that the entropy of ρ_ε can be rewritten as

eq:rhologrho-0

$$
\int \rho_\varepsilon(x) \log \rho_\varepsilon(x) dx = \sum_{i=1}^n \alpha_\varepsilon^i \left(\int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i \right) + r_{2,\varepsilon}
\tag{4.12}
$$

where $r_{2,\varepsilon} \leq e^{-\frac{C}{\varepsilon}}$ with $C = C(C_1, c_2, \xi_2)$ when $\varepsilon \ll 1$. By definition,

$$\int \rho_\varepsilon(x) \log \rho_\varepsilon(x) dx = \sum_{i=1}^n \alpha_\varepsilon^i \int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx,$$

so it suffices to show that for each $i \in \{1, \ldots, n\}$ we have

eq:rhologrho-1

$$\int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx = \int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i + r_{2,\varepsilon} \tag{4.13}$$

with $r_{2,\varepsilon} \leq e^{-\frac{C}{\varepsilon}}$. Indeed, on the one hand, by monotonicity of the logarithm it is clear that

eq:rhologrho-2

$$\int \rho_\varepsilon^i(x) \log \left(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j(x) \right) dx \geq \int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx + \log \alpha_\varepsilon^i. \tag{4.14}$$

In order to show the matching lower bound we first recall that the means m_ε^i of the ν_ε^i are well separated by assumption, $\min_{j \neq i} |m_\varepsilon^i - m_\varepsilon^j| > \xi_2$. Let $\delta \ll \frac{\xi}{2}$ to be fixed below and set $B_\delta^i = B(m_\varepsilon^i, \delta)$ Then we write

eq:rhologrho-3

$$\begin{aligned} &\int \rho_\varepsilon^i \log \Big(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \Big) \\ &= \int \rho_\varepsilon^i \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) + \int_{B_\delta^i} \rho_\varepsilon^i \Big(\log \Big(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \Big) - \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) \Big) \\ &\quad + \int_{(B_\delta^i)^c} \rho_\varepsilon^i \Big(\log \Big(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \Big) - \log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) \Big) \\ &=: \Big(\int \rho_\varepsilon^i \log \rho_\varepsilon^i + \log \alpha_\varepsilon^i \Big) + E_\varepsilon^1 + E_\varepsilon^2. \end{aligned} \tag{4.15}$$

We first show that the error term E_ε^2 is exponentially small. To that end, we first drop the exponential term in the Gaussian density to obtain the crude bound

eq:bd1

$$\log \Big(\sum_{j=1}^n \alpha_\varepsilon^j \rho_\varepsilon^j \Big) \leq \log \Big(\sum_{j=1}^n \alpha_\varepsilon^j \frac{1}{\sqrt{(2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon^j}} \Big) \leq \frac{d}{2} \log \varepsilon^{-1} + C. \tag{4.16}$$

where in the second inequality we use the fact that $\det \mathbf{\Sigma}_\varepsilon^i$ is bounded away from zero, which has been established in (4.9). Moreover, by definition we have

eq:bd2

$$-\log \left(\alpha_\varepsilon^i \rho_\varepsilon^i \right) \leq \frac{d}{2} \log \varepsilon^{-1} + C + \frac{|x - m_\varepsilon^i|^2}{\varepsilon} \tag{4.17}$$

Plugging bounds (4.16) and (4.16) in and using Gaussian concentration as well as the lower bound on $\lambda_{\min}$ established in (4.5)

eq:rhologrho-4

(4.18)
$$E_\varepsilon^2 \leq \int_{(B_\delta^i)^c} \rho_\varepsilon^i(x) \Big(\frac{d}{2} \log \varepsilon^{-1} + C + \frac{|x - m_\varepsilon^i|^2}{\varepsilon} \Big) dx \leq C \big(\log \varepsilon^{-1} + \varepsilon^{-1} \big) e^{-\frac{C\delta}{\varepsilon}}$$

when $\varepsilon \ll 1$. Next, we want to bound E_ε^1. Notice that $m_\varepsilon^j \to m^j$ for $j = 1, \cdots, n$, hence if $x \in B_\delta^i$ and if $\delta < \xi_1$, then $|x - m_\varepsilon^j| > \xi_1 - \delta$ for any $j \neq i$ when $\varepsilon \ll 1$. As a consequence,

$$\int_{B_\delta^i} \sum_{j=1, j \neq i}^n \alpha_\varepsilon^j \rho_\varepsilon^j \leq C \varepsilon^{-\frac{d}{2}} e^{-\frac{C(\xi_1 - \delta)^2}{\varepsilon}}. \tag{4.19}$$

This together with the elementary inequality

$$\log(x+y) = \log(x) + \int_x^{x+y} \frac{1}{t} dt \le \log x + \frac{y}{x}$$

for $x, y > 0$ implies

eq:rhologrho-5

$$\begin{aligned} E_\varepsilon^1 &= \int_{B_\delta^i} \rho_\varepsilon^i \Big(\log \Big(\alpha_\varepsilon^i \rho_\varepsilon^i + \sum_{j=1, j\neq i}^n \alpha_\varepsilon^j \rho_\varepsilon^j \Big) - \log \Big(\alpha_\varepsilon^i \rho_\varepsilon^i \Big) \Big) \\ &\le \int_{B_\delta^i} \frac{\sum_{j=1, j\neq i}^n \alpha_\varepsilon^j \rho_\varepsilon^j}{\alpha_\varepsilon^i} \\ &\le C\delta^d \varepsilon^{-\frac{d}{2}} e^{-\frac{C(\xi_1-\delta)^2}{\varepsilon}}. \end{aligned} \tag{4.20}$$

where we used that α_ε^i is bounded below from zero. Hence (4.13) follows directly from (4.14)-(4.20).

Finally, (4.10) follows from combining (4.11), (4.12) and the identity

$$\int \rho_\varepsilon^i(x) \log \rho_\varepsilon^i(x) dx = -\frac{1}{2} \log \left((2\pi\varepsilon)^d \det \mathbf{\Sigma}_\varepsilon^i \right) - \frac{d}{2}.$$

□

rem:assum

Remark 4.7. The assumption that $\min_{j\neq i} |m_\varepsilon^i - m_\varepsilon^j| > \xi_2 > 0$ is the crucial condition that allows us to express the entropy of the Gaussian mixture in terms of the mixture of entropies of individual Gaussian (i.e. the equation (4.12)), leading to the asymptotic formula (4.10). Neither formula (4.12) nor (3.20) is likely to be true without such an assumption since the cross entropy terms are not negligible.

The following corollary immediately follows from Lemma 4.6 by plugging in the Laplace approximation of the normalization constant $Z_{\mu,\varepsilon}$ given in Lemma 3.2 and rearranging the terms.

cor:dkl-asym4

Corollary 4.8. *Assume that* $\limsup_{\varepsilon\downarrow 0} G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) < \infty$. *Then*

eq:dkl-asym4

$$\begin{aligned} &G_\varepsilon(\boldsymbol{\alpha}_\varepsilon, \mathbf{m}_\varepsilon, \mathbf{\Sigma}_\varepsilon) \\ &= \sum_{i=1}^n \alpha_\varepsilon^i \left(\frac{V_1^\varepsilon(m_\varepsilon^i)}{\varepsilon} + V_2(m_\varepsilon^i) - \frac{d}{2} + \frac{1}{2} \mathrm{Tr}(D^2 V_1^\varepsilon(m_\varepsilon^i) \cdot \mathbf{\Sigma}_\varepsilon^i) \right) \\ &+ \sum_{i=1}^n \alpha_\varepsilon^i \Big(\log \alpha_\varepsilon^i - \frac{1}{2} \log \left(\det \mathbf{\Sigma}_\varepsilon^i \right) + \log \Big(\sum_{j=1}^n \beta^j \Big) \Big) + o(1). \end{aligned} \tag{4.21}$$

rem:4-9

Remark 4.9. Similarly to the discussion in Remark 3.9, the residual in (4.21) is here demonstrated to be of order $o(1)$, but the quantitative bound that $|r_\varepsilon| \le C\varepsilon$ in Lemma 4.10 can be used to extract a rate of convergence. This can be used to study the limiting behaviour of posterior measures arising from Bayesian inverse problems when multiple modes are present; see the next section.

sec:app

5. Applications in Bayesian inverse problems

Consider the inverse problem of recovering $x \in \mathbf{R}^d$ from the noisy data $y \in \mathbf{R}^d$, where y and x are linked through the equation

$$y = G(x) + \eta. \tag{5.1}$$

Here G is called the forward operator which maps from $\mathbf{R}^d$ into itself, $\eta \in \mathbf{R}^d$ represents the observational noise. We take a Bayesian approach to solving the inverse problem. The main idea is to first model our knowledge about x with a prior probability distribution, leading to a joint distribution on (x, y) once the probabilistic structure on η is defined. We then update the prior based on the observed data y;

specifically we obtain the posterior distribution μ^y which is the conditional distribution of x given y, and is the solution to the Bayesian inverse problem. From this measure one can extract information about the unknown quantity of interest. We remark that since G is non-linear in general, the posterior is generally not Gaussian even when the noise and prior are both assumed to be Gaussian. A systematic treatment of the Bayesian approach to inverse problems may be found in [24].

In Bayesian statistics there is considerable interest in the study of the asymptotic performance of posterior measures from a frequentist perspective; this is often formalized as the *posterior consistency* problem. To define this precisely, consider a sequence of observations $\{y_j\}_{j\in\mathbb{N}}$, generated from the truth $x^\dagger$ via

eq:yn
$$y_j = G(x^\dagger) + \eta_j, \tag{5.2}$$

where $\{\eta_j\}_{j\in\mathbb{N}}$ is a sequence of random noises. This may model a statistical experiment with increasing amounts of data or with vanishing noise. In either case, posterior consistency refers to concentration of the posterior distribution around the truth as the data quality increases. For parametric statistical models, Doob's consistency theorem [8, Theorem 10.10] guarantees posterior consistency under the identifiability assumption about the forward model. For nonparametric models, in which the parameters of interest lie in infinite dimensional spaces, the corresponding posterior consistency is a much more challenging problem. Schwartz's theorem [22, 2] provides one of the main theoretical tools to prove posterior consistency in infinite dimensional space, which replaces identifiability by a stronger assumption on testability. The posterior contraction rate, quantifying the speed that the posterior contracts to the truth, has been determined in various Bayesian statistical models (see [10, 23, 6]). In the context of the Bayesian inverse problem, the posterior consistency problem has mostly been studied to date for linear inverse problems with Gaussian priors [14, 1]. The recent paper [25] studied posterior consistency for a specific nonlinear Bayesian inverse problem, using the stability estimate of the underlying inverse problem together with posterior consistency results for the Bayesian regression problem.

In this section, our main interest is not in the consistency of posterior distribution, but in characterizing in detail its asymptotic behavior. We will consider two limit processes in (5.2): the small noise limit and the large data limit. In the former case, we assume that the noise $\eta_i = \frac{1}{\sqrt{i}}\eta$ where η is distributed according to the standard normal $N(0, \mathbf{I}_d)$, and we consider the data $\mathbf{y}_N$ given by the most accurate observation, i.e. $\mathbf{y}_N = y_N$. In the later case, the sequence $\{\eta_i\}_{i\in\mathbb{N}}$ is assumed to be independent identically distributed according to the standard normal and we accumulate the observations so that the data $\mathbf{y}_N = \{y_1, y_2, \cdots, y_N\}$. In addition, assume that the prior distribution is μ_0 which has the density

$$\mu_0(dx) = \frac{1}{Z_0} e^{-V_0(x)} dx$$

with the normalization constant $Z_0 > 0$. Since the data and the posterior are fully determined by the noise $\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \eta$ or $\boldsymbol{\eta} = \{\eta_i\}_{i\in\mathbb{N}}$, we denote the posterior by $\mu_N^{\boldsymbol{\eta}}$ to indicate the dependence. By using Bayes's formula, we calculate the posterior distribution for both limiting cases below.

- Small noise limit

eq:postdist1
$$\begin{aligned} \mu_N^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_{N,1}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}\left|y_n - G(x)\right|^2\right)\mu_0(dx) \\ &= \frac{1}{Z_{N,1}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}\left|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta\right|^2\right)\mu_0(dx). \end{aligned} \tag{5.3}$$

- Large data limit

eq:postdist2

$$
\begin{aligned}
\mu_n^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} |y_i - G(x)|^2\right) \mu_0(dx) \\
&= \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \left|G(x^\dagger) - G(x) + \eta_i\right|^2\right) \mu_0(dx).
\end{aligned} \tag{5.4}
$$

In both cases, we are interested in the limiting behavior of the posterior distribution $\mu_N^{\boldsymbol{\eta}}$ as $N \to \infty$. We divide our discussion below according to whether the noise is fixed or is considered a random variable. For a fixed realization of noise $\boldsymbol{\eta} = \eta$, by applying the theory developed in the previous section, we show the asymptotic normality for $\mu_N^{\boldsymbol{\eta}}$ in the small noise limit. Furthermore, we obtain a Bernstein-Von Mises type theorem for $\mu_N^{\boldsymbol{\eta}}$ with respect to both limit processes, small noise and large data.

5.1. **Asymptotic Normality.** In this subsection, we assume that the data is generated from the truth $x^\dagger$ and a single realization of the Gaussian noise $\eta^\dagger$, i.e.

$$
y = G(x^\dagger) + \frac{1}{\sqrt{N}}\eta^\dagger.
$$

Then the resulting posterior distribution $\mu_N^{\boldsymbol{\eta}}$ has the density of the form

eq:muepsBIP

$$
\begin{aligned}
\mu_N^{\boldsymbol{\eta}}(dx) &= \frac{1}{Z_N^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}|y - G(x)|^2 - V_0(x)\right) dx \\
&= \frac{1}{Z_N^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta^\dagger|^2 - V_0(x)\right) dx
\end{aligned} \tag{5.5}
$$

where $Z_N^{\boldsymbol{\eta}}$ is the normalization constant. Notice that $\mu_N^{\boldsymbol{\eta}}$ has the same form as the measure defined in (1.1) with $\varepsilon = \frac{1}{N}, V_1^\varepsilon(x) = V_1^N(x) := \frac{1}{2}|G(x^\dagger) - G(x) + \frac{1}{\sqrt{N}}\eta^\dagger|^2$ and $V_2(x) = V_0(x)$.

Now we consider the asymptotics of the measure $\mu_N^{\boldsymbol{\eta}}$ in the limit $N \to \infty$, under the following two different assumptions on V_0 and G:

assum-bip

Assumptions 5.1. *(i) $V_0 \in C^2(\mathbf{R}^d;\mathbf{R})$, $G \in C^3(\mathbf{R}^d;\mathbf{R}^d)$ and $G(x) = G(x^\dagger)$ implies $x = x^\dagger$. Moreover, G is a homeomorphism in the neighborhood of $x^\dagger$.*

(ii) $V_0 \in C^2(\mathbf{R}^d;\mathbf{R})$, $G \in C^3(\mathbf{R}^d;\mathbf{R}^d)$ and the zero set of the equation $G(x) = G(x^\dagger)$ is $\{x_i^\dagger\}_{i=1}^n$. Moreover $x_1^\dagger = x^\dagger$ and G is a homeomorphism in the neighborhood of $x_i^\dagger$.

Clearly in either case above, the potentials V_1^ε and V_2 satisfy the Assumption 2.3. In particular, we have $V_1^\varepsilon(x) \to V_1(x) := \frac{1}{2}|G(x^\dagger) - G(x)|^2$ for any $x \in \mathbf{R}^d$ and that $D^2V_1(x_i^\dagger) = DG(x_i^\dagger)^T DG(x_i^\dagger)$. Recall the set of Gaussian measures $\mathcal{A}$ and the set of Gaussian mixtures $\mathcal{M}_n$ and $\mathcal{M}_n^{\boldsymbol{\xi}}$ (defined in (4.1) and (4.2)). Again, we set $\boldsymbol{\xi} = (\xi_1, \xi_2)$ such that $\xi_1 \in (0,1)$ and $\min_{i\neq j}|x^i - x^j| \geq \xi_2 > 0$.

The following theorem concerning the asymptotic normality of $\mu_N^{\boldsymbol{\eta}}$ is a direct consequence of Corollary 3.4 and Corollary 4.4.

thm:bip

Theorem 5.2.
(i) *Suppose Assumptions 5.1 (i) holds. Given any $N \in \mathbb{N}$, let $\nu_N = N(m_N, \frac{1}{N}\boldsymbol{\Sigma}_N) \in \mathcal{A}$ be a minimizer of the functional $\nu \mapsto D_{KL}(\nu||\mu_N^{\boldsymbol{\eta}})$ within $\mathcal{A}$. Then $D_{KL}(\nu_N||\mu_N^{\boldsymbol{\eta}}) \downarrow 0$ as $N \to \infty$. Moreover, $m_N \to x^\dagger$ and $\boldsymbol{\Sigma}_N \to \left(DG(x^\dagger)^T DG(x^\dagger)\right)^{-1}$.*

(ii) *Suppose Assumptions 5.1 (ii) holds. Given any* $N \in \mathbb{N}$*, let* $\nu_N \in \mathcal{M}_n^{\boldsymbol{\xi}}$ *be a minimizer of the functional* $\nu \mapsto D_{KL}(\nu || \mu_N^{\boldsymbol{\eta}})$ *within* $\mathcal{M}_n^{\boldsymbol{\xi}}$*. Let* $\nu_N = \sum_{i=1}^N \alpha_N^i \nu_N^i$ *with* $\nu_N^i = N(m_N^i, \frac{1}{N}\boldsymbol{\Sigma}_N^i)$*. Then it holds that as* $N \to \infty$

$$m_N^i \to x_i^\dagger, \boldsymbol{\Sigma}_N^i \to \Big(DG(x_i^\dagger)^T DG(x_i^\dagger)\Big)^{-1} \text{ and } \alpha_N^i \to \frac{\Big[\det DG(x_i^\dagger)\Big]^{-1} \cdot e^{-V_0(x_i^\dagger)}}{\sum_{j=1}^n \Big[\det DG(x_j^\dagger)\Big]^{-1} \cdot e^{-V_0(x_j^\dagger)}}.$$

Theorem 5.2 (i) states that the measure $\mu_N^{\boldsymbol{\eta}}$ is asymptotically Gaussian when certain uniqueness and stability properties hold in the inverse problem. Moreover, in this case, the asymptotic Gaussian distribution is fully determined by the truth and the forward map, and is independent of the prior. In the case where the uniqueness fails, but the data only corresponds to a finite number of unknowns, Theorem 5.2 (ii) demonstrates that the measure $\mu_N^{\boldsymbol{\eta}}$ is asymptotically a Gaussian mixture, with each Gaussian mode independent of the prior. However, prior beliefs affect the proportions of the individual Gaussian components within the mixture; more precisely, the un-normalized weights of each Gaussian mode are proportional to the value of the prior evaluated at the corresponding unknown.

Remark 5.3. In general, when $\{\eta_i\}_{i\in\mathbb{N}}$ is a sequence of fixed realizations of the normal distribution, Theorem 5.2 does not hold for the measure $\mu_N^{\boldsymbol{\eta}}$ defined in (5.4) in the large data case. However, we will show that $D_{\mathrm{KL}}(\nu_N || \mu_N^{\boldsymbol{\eta}})$ will converge to zero in some average sense; see Theorem 5.4.

5.2. **A Bernstein-Von Mises type result.** The asymptotic Gaussian phenomenon in Theorem 5.2 is very much in the same spirit as the celebrated Bernstein-Von Mises (BvM) theorem [8]. This theorem asserts that for a certain class of regular priors, the posterior distribution converges to a Gaussian distribution, independently of the prior, as the sample size tends to infinity. Let us state the Bernstein-Von Mises theorem more precisely in the i.i.d case. Consider observing a set of i.i.d samples $\mathbf{X}^N := \{X^1, X^2, \cdots, X^N\}$, where X^i is drawn from distribution P_θ, indexed by an unknown parameter $\theta \in \Theta$. Let P_θ^N be the law of $\mathbf{X}^N$. Let Π be the prior distribution on θ and denote by $\Pi(\cdot|\mathbf{X}^N)$ the resulting posterior distribution. The Bernstein-Von Mises Theorem is concerned with the behavior of the posterior $\Pi(\cdot|\mathbf{X}^N)$ under the frequentist assumption that X^i is drawn from some true model P_{θ_0}. A standard finite-dimensional BvM result (see e.g. [8, Theorem 10.1]) states that, under certain conditions on the prior Π and the model P_θ, as $N \to \infty$

eq:bvm1
$$d_{TV}\left(\Pi(\theta|\mathbf{X}^N), N\left(\hat{\theta}_N, \frac{1}{N} I_{\theta_0}^{-1}\right)\right) \xrightarrow{P_{\theta_0}^N} 0 \tag{5.6}$$

where $\hat{\theta}_N$ is an efficient estimator for θ, I_θ is the Fisher information matrix of P_θ and d_{TV} represents the total variation distance. As an important consequence of the BvM result, Bayesian credible sets are asymptotically equivalent to frequentist confidence intervals. Moreover, it has been proved that the optimal rate of convergence in the Bernstein-Von Mises theorem is $O(1/\sqrt{N})$; see, for instance, [5, 11]. This means that for any $\delta > 0$, there exists $M = M(\delta) > 0$ such that

eq:bvm2
$$P_{\theta_0}^N\left(\mathbf{X}^N : d_{TV}\left(\Pi(\theta|\mathbf{X}^N), N\left(\hat{\theta}_N, \frac{1}{N} I_{\theta_0}^{-1}\right)\right) \geq M\frac{1}{\sqrt{N}}\right) \leq \delta \tag{5.7}$$

Unfortunately, BvM results like (5.6) and (5.7) do not fully generalize to infinite dimensional spaces, see counterexamples in [9]. Regarding the asymptotic frequentist properties of posterior distributions in nonparametric models, various positive results have been obtained recently, see e.g. [10, 23, 14, 15, 6, 7]. For the convergence rate in the nonparametric case, we refer to [10, 23, 6].

In the remainder of the section, we prove a Bernstein-Von Mises type result for the posterior distribution $\mu_N^{\boldsymbol{\eta}}$ defined by (5.3) and (5.4). If we view the observational noise η and η_i appearing in the data as random variables, then the posterior measures appearing become random probability measures. Furthermore, exploiting the randomness of the η_i, we claim that the posterior distribution in the large date case can be rewritten in the form of the small noise case. Indeed, by completing the square, we can write the expression (5.4) as

$$\mu_N^{\boldsymbol{\eta}}(dx) = \frac{1}{Z_{N,2}^{\boldsymbol{\eta}}} \exp\left(-\frac{N}{2}|G(x^\dagger) - G(x) + \frac{1}{N}\sum_{i=1}^N \eta_i|^2\right) dx \tag{5.8}$$

Observe that $\mathcal{L}\left(\frac{1}{N}\sum_{i=1}^N \eta_i\right) = \mathcal{L}(\frac{1}{\sqrt{N}}\eta) = N(0, \frac{1}{N}\mathbf{I}_d)$ due to the normality assumptions on η and η_i. As a consequence it makes no difference which formulation is chosen when one is concerned with the statistical dependence of $\mu_N^{\boldsymbol{\eta}}$ on the law of $\boldsymbol{\eta}$. For this reason, we will only prove the Bernstein-Von Mises result for $\mu_N^{\boldsymbol{\eta}}$ given directly in the form (5.3).

For notational simplicity, we write the noise level $\sqrt{\varepsilon}$ in place of $\frac{1}{\sqrt{N}}$ and consider random observations $\{y_\varepsilon\}$, generated from a truth $x^\dagger$ and normal noise η, i.e.

$$y_\varepsilon = G(x^\dagger) + \sqrt{\varepsilon}\eta.$$

Given the same prior defined as before, we obtain the posterior distribution

$$\begin{aligned}\mu_\varepsilon^\eta(dx) &= \frac{1}{Z_{\mu,\varepsilon}^\eta} \exp\left(-\frac{1}{2\varepsilon}|y_\varepsilon - G(x)|^2 - V_0(x)\right) dx\\ &= \frac{1}{Z_{\mu,\varepsilon}^\eta} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right) dx.\end{aligned}$$

For any fixed η, let ν_ε^η be the best Gaussian measure which minimizes the Kullback-Leibler divergence $D_{\mathrm{KL}}(\nu||\mu_\varepsilon^\eta)$ over $\mathcal{A}$. For ease of calculations, from now on we only consider the rate of convergence under Assumption 5.1 (i); the other case can be dealt with in the same manner, see Remark 5.9. The main result is as follows.

thm:expdkl **Theorem 5.4.** *There exists $C > 0$ such that*

eq:expdkl
$$\mathbb{E}^\eta D_{KL}(\nu_\varepsilon^\eta||\mu_\varepsilon^\eta) \le C\varepsilon \tag{5.9}$$

as $\varepsilon \downarrow 0$.

With the help of Pinsker's inequality (2.1) as well as the Markov inequality, one can derive the following BvM-type result from Theorem 5.4.

Corollary 5.5. *For any $\delta > 0$, there exists a constant $M = M(\delta) > 0$ such that*

eq:bvm3
$$\mathbb{P}^\eta\left(\eta : d_{TV}(\mu_\varepsilon^\eta, \nu_\varepsilon^\eta) \ge M\sqrt{\varepsilon}\right) \le \delta \tag{5.10}$$

when $\varepsilon \downarrow 0$.

Remark 5.6. By comparing the classical BvM result (5.7) with our new BvM-type result (5.10), we see that the asymptotic Gaussian distribution $N(\hat{\theta}_N, \frac{1}{N}I_{\theta_0}^{-1})$ is replaced by the best (with respect to Kullback-Leibler minimization) Gaussian ν_ε^η and that the optimal convergence rate $\mathcal{O}(\sqrt{\varepsilon})$ is also achieved.

Remark 5.7. For fixed realization of the noise η, we have shown in Theorem 5.2 (i) that $D_{\mathrm{KL}}(\nu_N||\mu_N^{\boldsymbol{\eta}}) \downarrow 0$ as $N \to \infty$. In fact, by following the proof of the Laplace method, one can prove that $D_{\mathrm{KL}}(\nu_N||\mu_N^{\boldsymbol{\eta}}) = \mathcal{O}(1/\sqrt{N})$. However, we obtain higher convergence rate in (5.9) (with ε replacing $1/N$) mainly because of symmetric cancellations in the evaluation of Gaussian integrals.

We start the proof of Theorem 5.4 with an averaging estimate for the normalization constant $Z^\eta_{\mu,\varepsilon}$.

Lemma 5.8.

eq:explogZ

$$\mathbb{E}^\eta \log Z^\eta_{\mu,\varepsilon} \leq \frac{d}{2}\log(2\pi\varepsilon) - V_0(x^\dagger) + \log\det DG(x^\dagger) + r_\varepsilon \tag{5.11}$$

where $r_\varepsilon \leq C\varepsilon$ for some $C > 0$ independent of ε.

Proof. Take a constant $\gamma \in (0, \frac{1}{2})$. We write $\mathbb{E}^\eta \log Z^\eta_{\mu,\varepsilon}$ as the sum

$$\mathbb{E}^\eta \log Z^\eta_{\mu,\varepsilon} = \mathbb{E}^\eta\left(\log Z^\eta_{\mu,\varepsilon}\mathbf{1}_{|\eta|\leq\varepsilon^{-\gamma}}\right) + \mathbb{E}^\eta\left(\log Z^\eta_{\mu,\varepsilon}\mathbf{1}_{|\eta|\geq\varepsilon^{-\gamma}}\right) =: I_1 + I_2.$$

We first find an upper bound for I_2. By definition,

$$\begin{aligned} Z^\eta_{\mu,\varepsilon} &= \int_{\mathbf{R}^d} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right)dx \\ &\leq \int_{\mathbf{R}^d} e^{-V_0(x)}dx = Z_0. \end{aligned}$$

It follows that

$$I_2 \leq \log Z_0 \cdot P^\eta(\eta : |\eta| \geq \varepsilon^{-\gamma}) \leq \log Z_0 \cdot e^{-\varepsilon^{-2\gamma}}.$$

For I_1, we need to estimate $Z^\eta_{\mu,\varepsilon}$ under the assumption that $|\eta| \leq \varepsilon^{-\gamma}$. Thanks to the condition (i) on G, when $\varepsilon \ll 1$ there exists a unique $m^\dagger_{\varepsilon,\eta}$ such that $G(m^\dagger_{\varepsilon,\eta}) = G(x^\dagger) + \sqrt{\varepsilon}\eta$. Moreover, denoting by H the inverse of G in the neighborhood of $G(x^\dagger)$, we get from Taylor expansion that

eq:mdagger0

$$m^\dagger_{\varepsilon,\eta} = x^\dagger + DH(G(x^\dagger))\sqrt{\varepsilon}\eta + \varepsilon\sum_{|\alpha|=2}\partial_\alpha H(\xi G(x^\dagger) + (1-\xi)\sqrt{\varepsilon}\eta)\eta^\alpha \tag{5.12}$$

with some $\xi \in (0,1)$. Thanks to the smoothness assumption on G, the function H is at least three times differentiable and hence the coefficients in the summation are uniformly bounded. Moreover, noting that $DH(G(x^\dagger)) = DG(x^\dagger)^{-1}$, we obtain

eq:mdagger1

$$m^\dagger_{\varepsilon,\eta} = x^\dagger + DG(x^\dagger)^{-1}\sqrt{\varepsilon}\eta + \varepsilon R_\varepsilon(\eta) \tag{5.13}$$

where $\limsup_{\varepsilon\downarrow 0}|R_\varepsilon(\eta)| \leq C|\eta|^2$ for some positive C which is independent of ε and η. Next, according to the proof of Lemma 3.2, given any sufficiently small $\delta > 0$, we can write $Z^\eta_{\mu,\varepsilon} = I^{\delta,\eta}_\varepsilon + J^{\delta,\eta}_\varepsilon$ where $|J^{\delta,\eta}_\varepsilon| \leq Ce^{-\frac{C}{\varepsilon}}$ with some $C > 0$ independent of η and

$$I^{\delta,\eta}_\varepsilon = \int_{B^{\delta,\eta}_\varepsilon} \exp\left(-\frac{1}{2\varepsilon}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2 - V_0(x)\right)dx$$

with $B^{\delta,\eta}_\varepsilon := B(m^\dagger_{\varepsilon,\eta}, \delta)$. Now we seek bounds for $I^{\delta,\eta}_\varepsilon$. Thanks to Assumption 5.1 (i) and the fact that $m^\dagger_{\varepsilon,\eta} \to 0$, G is a homeomorphism in the neighborhood of $m^\dagger_{\varepsilon,\eta}$. Therefore there exist positive constants $\delta_1 < \delta_2$ depending only on δ such that $B\left(G(m^\dagger_{\varepsilon,\eta}), \delta_1\right) \subset G(B^{\delta,\eta}_\varepsilon) \subset B\left(G(m^\dagger_{\varepsilon,\eta}), \delta_2\right)$. After applying the transformation $x \mapsto H(x)$ in evaluation of the integral $I^{\delta,\eta}_\varepsilon$, we get

$$\tilde{I}^{\delta_1,\eta}_\varepsilon \leq I^{\delta,\eta}_{i,\varepsilon} \leq \tilde{I}^{\delta_2,\eta}_\varepsilon$$

where

$$\tilde{I}^{\delta,\eta}_\varepsilon := \int_{B(0,\delta)} \exp\left(-\frac{1}{2\varepsilon}|y|^2 - V_0 \circ H(y + G(m^\dagger_{\varepsilon,\eta}))\right)\det(DH(y + G(m^\dagger_{\varepsilon,\eta}))dy.$$

In order to estimate $\tilde{I}^{\delta,\eta}_\varepsilon$, we define two auxiliary functions in $B(0,\delta)$ for small δ. Let $f_{\varepsilon,\eta}(\cdot) := \exp(-V_0 \circ H(\cdot + G(m^\dagger_{\varepsilon,\eta})))\det(DH(\cdot + G(m^\dagger_{\varepsilon,\eta}))$ and let $L(\cdot) := \exp(-V_0 \circ H(G(\cdot)))\det(DH(G(\cdot)) = \exp(-V_0(\cdot))/\det(DG(\cdot))$. It is worthy of note that within the ball $B(0,\delta)$, all derivatives up to second order of $f_{\varepsilon,\eta}$ as well as of

L can be bounded uniformly with respect to sufficiently small ε and η such that $|\eta| \leq \varepsilon^{-\gamma}$. Taking the equation (5.13) into account, we can expand L near $m^\dagger$ to get that

eq:feps-eta1 (5.14)

$$\begin{aligned}
f_{\varepsilon,\eta}(0) &= L(m^\dagger_{\varepsilon,\eta})\\
&= L(x^\dagger) + \nabla L(x^\dagger)^T(m^\dagger_{\varepsilon,\eta} - x^\dagger) + \frac{1}{2}(m^\dagger_{\varepsilon,\eta} - x^\dagger)^T\nabla^2 L(\theta x^\dagger + (1-\theta)m^\dagger_{\varepsilon,\eta})(m^\dagger_{\varepsilon,\eta} - x^\dagger)\\
&= \frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \varepsilon^{\frac{1}{2}}\nabla L(x^\dagger)^T DG(x^\dagger)^{-1}\eta + r_{1,\varepsilon,\eta}
\end{aligned}$$

with some $\theta \in (0,1)$ and the residual $|r_{1,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$ for some $C > 0$. Moreover, for any $y \in B(0,\delta)$,

eq:feps-eta2

$$f_{\varepsilon,\eta}(y) = f_{\varepsilon,\eta}(0) + \nabla f_{\varepsilon,\eta}(0)^T y + \frac{1}{2}y^T\nabla^2 f_{\varepsilon,\eta}(\xi y)y \tag{5.15}$$

for some $\xi = \xi(y) \in (0,1)$. Then it follows from (5.14) and (5.15) that

(5.16)

$$\begin{aligned}
\tilde{I}^{\delta,\eta}_\varepsilon &= \int_{B(0,\delta)} \exp(-\frac{1}{2\varepsilon}|y|^2) f_{\varepsilon,\eta}(y)dy\\
&= \varepsilon^{\frac{d}{2}}\int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp(-\frac{1}{2}|y|^2) f_{\varepsilon,\eta}(\varepsilon^{\frac{1}{2}}y)dy\\
&= \varepsilon^{\frac{d}{2}}\left(f_{\varepsilon,\eta}(0)\int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp(-\frac{1}{2}|y|^2)dy + \frac{\varepsilon}{2}\int_{B(0,\varepsilon^{-\frac{1}{2}}\delta)} \exp(-\frac{1}{2}|y|^2)y^T\nabla^2 f_{\varepsilon,\eta}(\xi y)ydy\right)\\
&= (2\pi\varepsilon)^{\frac{d}{2}}\left(\frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \nabla L(x^\dagger)^T DG(x^\dagger)^{-1}\sqrt{\varepsilon}\eta + r_{2,\varepsilon,\eta}\right)
\end{aligned}$$

with $|r_{2,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. Notice that the linear term in the expansion (5.15) vanishes from the second line to the third line because the the domain of integration is symmetric about the origin; the final equality holds because we have counted the exponentially decaying Gaussian integral outside of the ball $B(0,\varepsilon^{-\frac{1}{2}}\delta)$ in the residual $r_{2,\varepsilon,\eta}$. Hence we obtain that for $|\eta| \leq \varepsilon^{-\gamma}$ and ε small enough

$$I^\delta_{\varepsilon,\eta} = (2\pi\varepsilon)^{\frac{d}{2}}\left(\frac{\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))} + \varepsilon^{\frac{1}{2}}\nabla L(x^\dagger)^T DG(x^\dagger)^{-1}\eta + r_{2,\varepsilon,\eta}\right)$$

with $|r_{2,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. As a result, $Z^\eta_{\mu,\varepsilon}$ satisfies the same bound as above. Then by using the Taylor expansion of the log function, one obtains that

$$\log Z^\eta_{\mu,\varepsilon} = \log\left(\frac{(2\pi\varepsilon)^{\frac{d}{2}}\exp(-V_0(x^\dagger))}{\det(DG(x^\dagger))}\right) + \varepsilon^{\frac{1}{2}}p^T\eta + r_{3,\varepsilon,\eta}$$

where p is vector depending only on L, G, V_0 and $x^\dagger$ and $|r_{3,\varepsilon,\eta}| \leq C\varepsilon|\eta|^2$. This implies that when ε is sufficiently small,

$$I_1 = \mathbb{E}^\eta\left(\log Z^\eta_{\mu,\varepsilon}\mathbf{1}_{|\eta|\leq\varepsilon^{-\gamma}}\right) = \frac{d}{2}\log(2\pi\varepsilon) - V_0(x^\dagger) + \log\det DG(x^\dagger) + r_\varepsilon.$$

with $|r_\varepsilon| \leq C\varepsilon$. Again the first order term $\varepsilon^{\frac{1}{2}}p^T\eta$ vanishes because of the symmetry in integration; the bound $|r_\varepsilon| \leq C\varepsilon$ follows from the bound for $r_{3,\varepsilon,\eta}$ and the Gaussian tail bound. This completes the proof. □

Proof of Theorem 5.4. We prove the theorem by constructing a family of Gaussian measures $\{\overline{\nu}^\eta_\varepsilon\}$ such that

eq:expdkl-1

$$\mathbb{E}^\eta D_{\mathrm{KL}}(\overline{\nu}^\eta_\varepsilon||\mu^\eta_\varepsilon) \leq C\varepsilon \tag{5.17}$$

for some $C > 0$. Then the theorem is proved by the optimality of $\nu_{\varepsilon,\eta}$. Recall that $m^\dagger_{\varepsilon,\eta}$ is defined by (5.12). Fixing $\gamma \in (0, \frac{1}{2})$, we define $\overline{\nu}^\eta_\varepsilon = N(\overline{m}_{\varepsilon,\eta}, \overline{\mathbf{\Sigma}}_{\varepsilon,\eta})$ with $\overline{m}_{\varepsilon,\eta}$ defined by

$$\overline{m}_{\varepsilon,\eta} = \begin{cases} m^\dagger_{\varepsilon,\eta} & \text{if } |\eta| \leq \varepsilon^{-\gamma}, \\ x^\dagger & \text{otherwise} \end{cases}$$

and that $\overline{\mathbf{\Sigma}}_{\varepsilon,\eta} = DG(\overline{m}_{\varepsilon,\eta})^{-1}$. Clearly, when ε is small enough, $\overline{m}_{\varepsilon,\eta}$ admits an expansion similar to (5.12). As a consequence, there exist positive constants C_1, c_2, C_2 which are independent of η, such that $\limsup_{\varepsilon\downarrow 0} |\overline{m}_{\varepsilon,\eta}| \leq C_1$ and $c_2 \leq \liminf_{\varepsilon\downarrow 0} \lambda_{\min}(\overline{\mathbf{\Sigma}}_{\varepsilon,\eta}) < \limsup_{\varepsilon\downarrow 0} \operatorname{Tr}(\mathbf{\Sigma}_\varepsilon) \leq C_2$ hold for all η. With the above choice for $(\overline{m}_{\varepsilon,\eta}, \overline{\mathbf{\Sigma}}_{\varepsilon,\eta})$, an application of Lemma 3.7 with $V^\varepsilon_1(x) = \frac{1}{2}|G(x^\dagger) - G(x) + \sqrt{\varepsilon}\eta|^2$ and $V_2(x) = V_0(x)$ yields that

eq:expdkl2

$$D_{\mathrm{KL}}(\overline{\nu}^\eta_\varepsilon || \overline{\mu}^\eta_\varepsilon) = V_0(\overline{m}_{\varepsilon,\eta}) - \frac{d}{2}\log(2\pi\varepsilon) + \frac{1}{2}\log\det DG(\overline{m}_{\varepsilon,\eta}) + \log Z^\eta_{\mu,\varepsilon} + r_\varepsilon \tag{5.18}$$

where $r_\varepsilon \leq C\varepsilon$ with $C = C(C_1, c_2, C_2, M_V)$. By the definition of $\overline{m}_{\varepsilon,\eta}$ and the expansion (5.12), it follows from the Taylor expansion for the function $x \mapsto V_0(x) + \frac{1}{2}\log\det DG(x)$ that when $|\eta| \leq \varepsilon^{-\gamma}$ and ε is small enough,

eq:rhologdet

$$V_0(\overline{m}_{\varepsilon,\eta}) + \frac{1}{2}\log\det DG(\overline{m}_{\varepsilon,\eta}) = V_0(x^\dagger) + \frac{1}{2}\log\det DG(x^\dagger) + \sqrt{\varepsilon} q^T\eta + \tilde{r}_{\varepsilon,\eta} \tag{5.19}$$

with some $q \in \mathbf{R}^d$ and $|\tilde{r}_{\varepsilon,\eta}| \leq C\varepsilon$ for some $C > 0$. Then the estimate (5.17) follows, by taking the expectation of (5.18) and using the equation (5.19) and Lemma 3.7. $\square$

rem:expdkl

Remark 5.9. Theorem 5.4 proves the rate of convergence with the assumption that G satisfies Assumption 5.1 (i). However, the convergence rate remains the same when Assumption 5.1 (ii) is fulfilled, and when the best Gaussian measure is replaced by the best Gaussian mixture.

6. Conclusions

We have studied a methodology widely used in applications, yet little analyzed, namely the approximation of a given target measure by a Gaussian, or by a Gaussian mixture. We have employed relative entropy as a measure of goodness of fit. Our theoretical framework demonstrates the existence of minimizers of the variational problem, and studies their asymptotic form in a relevant small parameter limit where the measure concentrates; the small parameter limit is studied by use of tools from Gamma convergence. In the case of a target with asymptotically unimodal distribution the Gamma limit demonstrates perfect reconstruction by the approximate single Gaussian method in the measure concentration limit; and in the case of multiple modes it quantifies the errors resulting from using a single mode fit. Furthermore the Gaussian mixture is shown to overcome the limitations of a single mode fit, in the case of target measure with multiple modes. These ideas are exemplified in the analysis of a Bayesian inverse problem in the small noise or large data set limits, and connections made to the Bernstein-Von Mises theory from asymptotic statistics.

A key conclusion of this work is that Γ-convergence is a very natural tool for the study of algorithms in machine learning and asymptotic statistics. Interesting future directions for the application of ideas from Γ-convergence include the study of limiting problems in which the target probability measure concentrates on a manifold, together with the study of inverse problems in infinite dimensional spaces. New ideas will be needed to tackle both of these cases.

References

ALS13 [1] S. Agapiou, S. Larsson, and A.M. Stuart. Posterior consistency of the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, 123(10):38283860, 2013.

BSW99 [2] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, **27**:536–561, 1999.

B06 [3] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st edition, 2006.

bra02a [4] A. Braides. Γ-*convergence for Beginners*. Oxford University Press, Oxford, 2002.

LeCam73 [5] L. M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, **1**:38–53, 1973.

CN13 [6] I. Castillo and R. Nickl. Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.*, **41**:1999–2028, 2013.

CN14 [7] I. Castillo and R. Nickl. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, **42**:1941–1969, 2014.

V00 [8] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

F99 [9] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, **27**:1119–1141, 1999.

GGV00 [10] S. Ghosal, J.K. Ghosal, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, **28**:500–531, 2000.

HM76 [11] C. Hipp and R. Michel. On the Bernstein-von Mises approximation of posterior distributions. *Ann. Statist.*, **4**:972–980, 1976.

Jensen [12] J. L. Jensen. *Saddlepoint approximations*, volume 16. Oxford statistical science series, 1995.

KP13 [13] MA Katsoulakis and P. Plechác. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *The Journal of Chemical Physics*, **139**:074115, 2013.

KVZ11 [14] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, **39**:2626–2657, 2011.

L11 [15] H. Leahu. On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Statist.*, **5**:373–404, 2011.

LSW16 [16] Y. Lu, A.M. Stuart, and H. Weber. Gaussian approximations for transition paths in molecular dynamics. *arXiv:1604.06594*, 2016.

MG11 [17] A. J. Majda and B. Gershgorin. Improving model fidelity and sensitivity for complex systems through empirical information theory. *Proceedings of the National Academy of Sciences*, **108**:10044–10049, 2011.

MS14 [18] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, **42**:1809–1884, 2014.

PSSW15b [19] 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.*, **37**:A2733–A2757, 2015.

PSSW15a [20] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber. Kullback-Leibler approximation for probability measures in infinite dimensional spaces. *SIAM J. Math. Anal*, **47**:4091–4122, 2015.

sanz2016gaussian [21] D. Sanz-Alonso and A.M. Stuart. Gaussian approximations of small noise diffusions in Kullback-Leibler divergence. Communications in Mathematical Sciences, To Appear.

S65 [22] L. Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, **4**, 1965.

SW01 [23] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, **29**:687–714, 2001.

S10a [24] A.M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

V13 [25] S. J. Vollmer. Posterior consistency for Bayesian inverse problems through stability and regression results. *Inverse Problems*, **29**:125011, 2013.

(Y. Lu) Mathematics Institute, University of Warwick, UK
E-mail address: `Yulong.Lu@warwick.ac.uk`

(A. M. Stuart) Computing and Mathematical Sciences, Caltech, USA
E-mail address: `astuart@caltech.edu`

(H. Weber) Mathematics Institute, University of Warwick, UK
E-mail address: `Hendrik.Weber@warwick.ac.uk`