# GAUSSIAN MIXTURE VARIATIONAL INFERENCE WITH WASSERSTEIN–FISHER–RAO GRADIENT FLOWS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The choice of geometry affects the efficiency of gradient-based algorithms for variational inference (VI). We study Wasserstein–Fisher–Rao (WFR) gradient flows over Gaussians and their mixtures. For Gaussian VI with smooth, strongly log-concave targets, we prove exponential convergence of a WFR splitting scheme to the optimal Gaussian approximation. For general targets, we establish stationarity guarantees for Gaussian and mixture VI. The splitting scheme evaluates target statistics twice per iteration, once for each geometry. For efficiency, we instead develop a switched algorithm that chooses one suitable geometry per update. For Gaussian targets, we show that Wasserstein warm-up followed by Fisher–Rao refinement avoids the logarithmic initialization delay of Fisher–Rao and the slow convergence of Wasserstein on ill-conditioned targets. We extend this approach to Gaussian mixtures and propose adaptive rules for choosing each component's switching time. We prove descent and stationarity guarantees for the resulting algorithm. Across benchmarks in machine learning and PDE-based Bayesian inverse problems, the switched method is competitive with or better than competing methods and is substantially less sensitive to initialization.

## 1 INTRODUCTION

Sampling from and approximating probability distributions are fundamental tasks in Bayesian inverse problems and machine learning (Stuart, 2010; Blei et al., 2017). Variational inference (VI) provides a scalable approach by minimizing the reverse Kullback–Leibler (KL) divergence over a tractable family. Gaussian VI is a simple example, while normalizing flows offer more expressive approximations through invertible transformations (Rezende & Mohamed, 2015). We focus on Gaussians and their mixtures. Gaussian mixtures can represent multiple modes and balance computational tractability with expressivity (Lin et al., 2019; Arenz et al., 2023; Petit-Talamon et al., 2025). They remain competitive with more elaborate variational methods on posterior-inference benchmarks (Blessing et al., 2024).

Geometry has played an important role in the development of sampling and VI algorithms: it determines how a gradient method updates a probability distribution. Wasserstein geometry comes from optimal transport, describes the transport of mass, and gives the gradient-flow interpretation of Langevin sampling (Jordan et al., 1998). On Gaussians, it becomes Bures–Wasserstein geometry (Takatsu, 2011). Fisher–Rao geometry comes from information geometry and yields natural-gradient updates (Amari, 1998). The Gaussian Fisher–Rao flow is invariant under invertible affine changes of variables, which helps it adapt to target anisotropy (Chen et al., 2026). Wasserstein–Fisher–Rao (WFR) combines the two geometries (Chizat et al., 2018; Liero et al., 2018; 2026).

This paper provides a systematic study of WFR dynamics for Gaussian and mixture VI, with theoretical foundations and practical algorithmic insights. For smooth, strongly log-concave targets, Liero et al. (2026) prove exponential convergence of the Gaussian WFR flow. We prove exponential convergence of a splitting discretization to the optimal Gaussian approximation. For general smooth targets, we instead establish stationarity guarantees in both continuous and discrete time.

Splitting schemes that alternate the two geometries (Gallouët & Monsaingeon, 2017), however, recompute target statistics between the Wasserstein and Fisher–Rao updates. This doubles the evaluation cost per iteration relative to either update alone, and also obscures how each geometry contributes to the efficiency of the algorithm. Towards a more principled use of the two geometries, we



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ask when each geometry is useful, whether a single geometry per update suffices, and how to choose the geometry adaptively in practice. Guided by a clean analysis in the Gaussian setting, our main contribution is an efficient algorithm for Gaussian mixture VI that adaptively selects the geometry of each mixture component, using one round of target-statistic evaluations per iteration, together with descent and stationarity guarantees.

More precisely, our contributions are as follows:

- *WFR convergence.* For smooth, strongly log-concave targets, we prove exponential convergence of the Gaussian WFR splitting to the optimal Gaussian approximation. For general smooth targets, we prove $O(T^{-1})$ and $O(N^{-1})$ stationarity bounds for the flow and a covariance-constrained splitting, respectively.
- *Gaussian theory for choosing geometry.* For Gaussian targets, we show that Wasserstein warm-up removes the delay caused by small initial covariances, while Fisher–Rao refinement improves the dependence on the anisotropy of the target; thus a switching that goes from Wasserstein to Fisher–Rao dynamics achieves the best of both worlds. The covariance recovery result also extends to smooth log-concave targets.
- *Componentwise switching for Gaussian mixtures.* We extend switching to mixtures: each component chooses its geometry adaptively, and each iteration uses one round of target-statistic evaluations. Using a separable upper bound tangent to the objective, we prove descent for any, possibly adaptive, choice of component geometries and an $O(N^{-1})$ stationarity bound.
- *Numerical evaluation.* We test our algorithm on a 40-mode Gaussian mixture, a 196-dimensional learned digit distribution, and a 64-dimensional Darcy Bayesian inverse problem. The switched method achieves competitive performance across these tasks and is less sensitive to initialization.

**Related work.** Lambert et al. (2022) study Wasserstein gradient flows for Gaussian and Gaussian-mixture VI. Diao et al. (2023) give convergence guarantees for a Gaussian forward–backward algorithm. For Gaussian mixtures, Huix et al. (2024) analyze components with fixed covariances, while Petit-Talamon et al. (2025) study isotropic components.

Fisher–Rao geometry leads to natural-gradient methods for VI (Amari, 1998). Convergence of natural-gradient methods for Gaussian VI has been studied in both deterministic (Chen et al., 2026; Kumar et al., 2025) and stochastic settings (Wu & Gardner, 2024; Sun et al., 2025; Guilmeau et al., 2026). For Gaussian mixtures, natural-gradient methods update both component parameters and weights and achieve superior performance (Lin et al., 2019; Arenz et al., 2023); see also recent developments in derivative-free methods (Chen et al., 2024) and stable implementations (Che et al., 2025; 2026).

WFR can be discretized by alternating Wasserstein and Fisher–Rao updates (Gallouët & Monsaingeon, 2017). Crucinio & Pathiraja (2026) study how the order of the two flows affects convergence. Liero et al. (2026) derive Gaussian WFR dynamics and prove continuous-time convergence under strong log-concavity. In Gaussian examples, they also observe faster initial progress for Wasserstein and faster asymptotic convergence for Fisher–Rao. We quantify this behavior for Gaussian targets, and use it to design a componentwise switched method for Gaussian mixtures.

## 2 Variational inference and Wasserstein–Fisher–Rao gradient flows

### 2.1 Problem setup and gradient-flow dynamics

Let $\pi$ be a target density on $\mathbb{R}^d$ known up to its normalizing constant,

$$\pi(\theta) = Z^{-1}\exp(-V(\theta)), \qquad Z = \int_{\mathbb{R}^d} \exp(-V(\theta))\,\mathrm{d}\theta.$$

VI approximates $\pi$ within a tractable family $\mathcal{Q}$ by minimizing the reverse KL objective

$$\mathcal{F}(q) := \mathrm{KL}(q\|\pi) = \mathbb{E}_q[V] + \mathbb{E}_q[\log q] + \log Z, \qquad q \in \mathcal{Q},$$

whose minimization does not require the normalizing constant $Z$. We consider the Gaussian family $\mathcal{Q}_{\mathrm{G}} = \{\mathcal{N}(m, C) : m \in \mathbb{R}^d,\ C \in \mathbb{S}^d_{++}\}$ and Gaussian mixture family $\mathcal{Q}^K_{\mathrm{GM}} = \left\{\sum_{k=1}^K w_k \mathcal{N}(m_k, C_k) : w \in \Delta^\circ_K,\ m_k \in \mathbb{R}^d,\ C_k \in \mathbb{S}^d_{++}\right\}$, where $\Delta^\circ_K := \{w \in (0,1)^K : \sum_{k=1}^K w_k = 1\}$ is the interior of the probability simplex.

**Gaussian VI.** Let $q_t = \mathcal{N}(m_t, C_t)$. Define the gradient and Hessian

$$g_t := -\nabla_{m_t} \mathcal{F}(q_t) = \mathbb{E}_{q_t}\left[\nabla_\theta \log \frac{\pi}{q_t}\right], \qquad H_t := -2\nabla_{C_t} \mathcal{F}(q_t) = \mathbb{E}_{q_t}\left[\nabla^2_\theta \log \frac{\pi}{q_t}\right],$$

where the equalities follow from Gaussian integration by parts. Equipping the Gaussian family with the Wasserstein or Fisher–Rao geometry gives, respectively, the Wasserstein and Fisher–Rao gradient flows (Lambert et al., 2022; Diao et al., 2023; Chen et al., 2026),

$$\begin{aligned} &\text{Wasserstein}: && \dot{m}_t = g_t, && \dot{C}_t = C_t H_t + H_t C_t, \\ &\text{Fisher–Rao}: && \dot{m}_t = C_t g_t, && \dot{C}_t = C_t H_t C_t. \end{aligned} \tag{1}$$

Equivalently, the Riemannian gradients of $\mathcal{F}$ under the Wasserstein and Fisher–Rao geometries are

$$\mathrm{grad}_{\mathsf{W}} \mathcal{F}(q) = (-g, -(CH + HC)), \qquad \mathrm{grad}_{\mathsf{FR}} \mathcal{F}(q) = (-Cg, -CHC),$$

with squared Riemannian norms

$$\|\mathrm{grad}_{\mathsf{W}} \mathcal{F}(q)\|^2_{\mathsf{W}} = \|g\|^2 + \mathrm{Tr}(HCH), \qquad \|\mathrm{grad}_{\mathsf{FR}} \mathcal{F}(q)\|^2_{\mathsf{FR}} = g^\top C g + \frac{1}{2}\mathrm{Tr}(CHCH). \tag{2}$$

Wasserstein–Fisher–Rao (WFR) geometry combines these two gradient directions (Liero et al., 2018; Chizat et al., 2018; Liero et al., 2026). Allowing their strengths to vary through nonnegative schedules $\lambda^{\mathsf{W}}_t$ and $\lambda^{\mathsf{FR}}_t$, we obtain

$$\dot{m}_t = \left(\lambda^{\mathsf{W}}_t I + \lambda^{\mathsf{FR}}_t C_t\right) g_t, \qquad \dot{C}_t = \lambda^{\mathsf{W}}_t \left(C_t H_t + H_t C_t\right) + \lambda^{\mathsf{FR}}_t C_t H_t C_t. \tag{3}$$

The choices $(\lambda^{\mathsf{W}}_t, \lambda^{\mathsf{FR}}_t) = (1, 0)$ and $(0, 1)$ recover the pure Wasserstein and Fisher–Rao flows, while $(1, 1)$ gives the standard WFR combination.

Although $g_t$ and $H_t$ involve target gradients and Hessians, Stein's identity (Gaussian integration by parts) avoids Hessian evaluations and can even yield derivative-free implementations (Che et al., 2025; 2026).

**Gaussian mixture VI.** We next consider $q_t = \sum_{k=1}^K w_{t,k} q_{t,k}$ with $q_{t,k} = \mathcal{N}(m_{t,k}, C_{t,k})$. For each component, define

$$g_{t,k} := \mathbb{E}_{q_{t,k}}\left[\nabla_\theta \log \frac{\pi}{q_t}\right],\ H_{t,k} := \mathbb{E}_{q_{t,k}}\left[\nabla^2_\theta \log \frac{\pi}{q_t}\right],\ \ell_{t,k} := \mathbb{E}_{q_{t,k}}\left[\log \frac{\pi}{q_t}\right],\ \bar{\ell}_t := \sum_{j=1}^K w_{t,j} \ell_{t,j}.$$

We consider the following WFR gradient flow in the Gaussian mixture space:

$$\begin{aligned} \dot{m}_{t,k} &= \left(\lambda^{\mathsf{W}}_t I + \lambda^{\mathsf{FR}}_t C_{t,k}\right) g_{t,k}, \\ \dot{C}_{t,k} &= \lambda^{\mathsf{W}}_t \left(C_{t,k} H_{t,k} + H_{t,k} C_{t,k}\right) + \lambda^{\mathsf{FR}}_t C_{t,k} H_{t,k} C_{t,k}, \\ \dot{w}_{t,k} &= \lambda^{\mathsf{FR}}_t w_{t,k} (\ell_{t,k} - \bar{\ell}_t), \qquad k = 1, \ldots, K. \end{aligned} \tag{4}$$

This flow combines the Wasserstein (Lambert et al., 2022) and Fisher–Rao (Lin et al., 2019; Chen et al., 2024) dynamics for Gaussian mixtures: the Wasserstein part transports the Gaussian components while preserving their weights, whereas the Fisher–Rao part evolves both component parameters and weights. Setting either schedule to zero recovers the corresponding pure-geometry dynamics, and for $K = 1$ the system reduces to (3).

### 2.2 Convergence and discretization of Gaussian WFR

We first consider smooth, strongly log-concave targets. In this setting, the Gaussian VI objective is strongly geodesically convex on the Bures–Wasserstein Gaussian manifold and hence admits a unique minimizer (Lambert et al., 2022, Lemma 2 and Corollary 3). We show that the Gaussian WFR flow and its splitting discretization converge exponentially to this optimal Gaussian approximation. We then consider general smooth targets, where the corresponding guarantees weaken to first-order stationarity.

**Assumption 2.1** (Smoothness)**.** For some $\beta > 0$, the potential satisfies $\left\|\nabla^2 V(\theta)\right\|_{\mathrm{op}} \leq \beta$ for all $\theta \in \mathbb{R}^d$.

**Assumption 2.2** (Strong log-concavity)**.** For some $\alpha > 0$, the potential satisfies $\nabla^2 V(\theta) \succeq \alpha I$ for all $\theta \in \mathbb{R}^d$.

Let $q_\star$ be the optimal Gaussian approximation. The next result follows from the Gaussian Hellinger–Kantorovich energy-dissipation argument of Liero et al. (2026, Theorem 5.11); Appendix A.1 gives the specialization to our notation.

**Theorem 2.3** (Exponential convergence of Gaussian WFR gradient flow)**.** *Under Assumptions 2.1 and 2.2, suppose* $\mu_{0,\min} I \preceq C_0 \preceq \mu_{0,\max} I$*, and define* $\underline{\mu} := \min\left\{\mu_{0,\min}, \frac{1}{\beta}\right\}$ *and* $\overline{\mu} := \max\left\{\mu_{0,\max}, \frac{1}{\alpha}\right\}$*, then* $\underline{\mu} I \preceq C_t \preceq \overline{\mu} I$ *along (3) and*

$$\mathcal{F}(q_t) - \mathcal{F}(q_\star) \leq \exp\left(-\alpha \int_0^t \left(2\lambda_s^{\mathsf{W}} + \underline{\mu}\lambda_s^{\mathsf{FR}}\right) \mathrm{d}s\right) \left(\mathcal{F}(q_0) - \mathcal{F}(q_\star)\right).$$

**Splitting discretization.** We specialize the WFR flow with schedule $(\lambda_t^{\mathsf{W}}, \lambda_t^{\mathsf{FR}}) = (1, 1)$ and propose a discretization composing two forward–backward updates. The Wasserstein step is the deterministic FB–GVI update of Diao et al. (2023). For the Fisher–Rao step, we linearize the potential energy in $(m, C)$ and treat the entropy implicitly, using the Gaussian KL divergence as the proximal term; its closed-form update is derived in Appendix A.2. Given $q_n = \mathcal{N}(m_n, C_n)$, compute $g_n$ and $H_n$ at $q_n$ and set

$$\begin{aligned} m_{n+1/2} &:= m_n + h g_n, \qquad \widetilde{C}_{n+1/2} := \left[I + h(H_n - C_n^{-1})\right] C_n \left[I + h(H_n - C_n^{-1})\right], \\ C_{n+1/2} &:= \frac{1}{2}\left[\widetilde{C}_{n+1/2} + 2hI + \left(\widetilde{C}_{n+1/2}(\widetilde{C}_{n+1/2} + 4hI)\right)^{1/2}\right]. \end{aligned} \tag{5}$$

Writing $q_{n+1/2} = \mathcal{N}(m_{n+1/2}, C_{n+1/2})$, recompute $g_{n+1/2}$ and $H_{n+1/2}$ and set

$$m_{n+1} := m_{n+1/2} + h C_{n+1/2} g_{n+1/2}, \qquad C_{n+1} := \left(C_{n+1/2}^{-1} - \frac{h}{1+h} H_{n+1/2}\right)^{-1}. \tag{6}$$

We denote the corresponding forward–backward update maps by $\mathsf{W}_h(q; g, H)$ and $\mathsf{FR}_h(q; g, H)$, respectively. Hence the WFR splitting can be written compactly as

$$q_{n+1/2} = \mathsf{W}_h(q_n; g_n, H_n), \qquad q_{n+1} = \mathsf{FR}_h(q_{n+1/2}; g_{n+1/2}, H_{n+1/2}). \tag{7}$$

Since the Fisher–Rao step recomputes $g_{n+1/2}$ and $H_{n+1/2}$ after the Wasserstein step, each splitting iteration requires two rounds of target-statistic evaluations.

**Theorem 2.4** (Exponential convergence of Gaussian WFR splitting)**.** *Under Assumptions 2.1 and 2.2, let* $\underline{\mu}$ *and* $\overline{\mu}$ *be as in Theorem 2.3. If* $0 < h \leq \min\left\{\frac{1}{\beta}, \frac{1}{\beta\overline{\mu}}\right\}$*, then* $\underline{\mu} I \preceq C_n, C_{n+1/2} \preceq \overline{\mu} I$ *and*

$$\mathcal{F}(q_N) - \mathcal{F}(q_\star) \leq \exp\left(-\frac{\alpha \underline{\mu} N h}{8}\right) \left(\mathcal{F}(q_0) - \mathcal{F}(q_\star)\right).$$

**General smooth targets.** Without Assumption 2.2, convergence to the globally optimal Gaussian need not hold. Nevertheless, the continuous WFR flow and a covariance-constrained version of the splitting retain first-order stationarity guarantees. The constraint is needed because, without strong log-concavity, we no longer have uniform bounds on the covariance iterates; we therefore restrict the covariances to

$$\mathcal{C} := \{C \in \mathbb{S}^d_{++} : \mu_- I \preceq C \preceq \mu_+ I\}, \qquad 0 < \mu_- < \mu_+ < \infty,$$

and let $\mathsf{W}_h^{\mathcal{C}}(q; g, H)$ and $\mathsf{FR}_h^{\mathcal{C}}(q; g, H)$ denote the corresponding covariance-constrained update maps; their closed-form expressions are given in Lemma A.5 in Appendix A.4.

**Proposition 2.5** (Stationarity for general targets)**.** *Under Assumption 2.1, suppose the flow (3) with* $(\lambda_t^{\mathsf{W}}, \lambda_t^{\mathsf{FR}}) = (1, 1)$ *is well defined on* $[0, T]$*. Let* $\mathcal{F}_{\inf}^{\mathrm{G}} := \inf_{q \in \mathcal{Q}_{\mathrm{G}}} \mathcal{F}(q)$*. Then*

$$\min_{0 \leq t \leq T}\left[\|\mathrm{grad}_{\mathsf{W}}\, \mathcal{F}(q_t)\|_{\mathsf{W}}^2 + \|\mathrm{grad}_{\mathsf{FR}}\, \mathcal{F}(q_t)\|_{\mathsf{FR}}^2\right] \leq \frac{\mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathrm{G}}}{T}.$$

*Moreover, if $C_0 \in \mathcal{C}$ and $0 < h \le \min\{(2\beta)^{-1}, (4\beta\mu_+)^{-1}\}$, the covariance-constrained splitting satisfies*

$$\min_{0\le n<N} \mathcal{R}_n^{\mathsf{WFR}} \le \frac{4\left(\mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathcal{C}}\right)}{Nh} = O(N^{-1}), \quad \mathcal{R}_n^{\mathsf{WFR}} := \frac{W_2^2(q_{n+1/2}, q_n)}{h^2} + \frac{\mathrm{KL}(q_{n+1}\|q_{n+1/2})}{h^2},$$

*where $\mathcal{F}_{\inf}^{\mathcal{C}} := \inf\{\mathcal{F}(\mathcal{N}(m, C)) : m \in \mathbb{R}^d,\ C \in \mathcal{C}\}$. Moreover, $\mathcal{R}_n^{\mathsf{WFR}} = 0$ if and only if $q_n$ is first-order stationary for the covariance-constrained Gaussian VI problem.*

The proof is given in Appendix A.4. Altogether, these results provide theoretical foundations for using Wasserstein–Fisher–Rao gradient flows in Gaussian VI. On the other hand, the splitting scheme (7) treats the two geometries equally and needs two rounds of target-statistic evaluations per iteration. Section 3 studies when each geometry is preferable, which is the basis of the algorithm in Section 4 that selects one geometry per iteration.

## 3 Complementary roles of Wasserstein and Fisher–Rao

For simplicity and illustration of ideas, we consider a Gaussian target $\pi = \mathcal{N}(m_\star, C_\star)$ to isolate when each geometry is advantageous. We use $\mu(\cdot)$ to denote the eigenvalues and let $\mu_0 := \mu_{\min}(C_\star)$ and $\kappa := \mu_{\max}(C_\star)/\mu_0$. To compare the two flows on the same time scale, we set $(\lambda_t^{\mathsf{W}}, \lambda_t^{\mathsf{FR}}) = (\mu_0, 0)$ for Wasserstein and $(0, 1)$ for Fisher–Rao, so that their mean contraction rates agree along the $\mu_0$ direction at the target.

For Gaussian targets, both flows are explicitly solvable. The Fisher–Rao solution satisfies (Chen et al., 2026, Lemma E.3)

$$(C_t^{\mathsf{FR}})^{-1} = C_\star^{-1} + e^{-t}(C_0^{-1} - C_\star^{-1}), \qquad m_t^{\mathsf{FR}} - m_\star = e^{-t} C_t^{\mathsf{FR}} C_0^{-1}(m_0 - m_\star), \tag{8}$$

whereas solving the Wasserstein dynamics gives

$$m_t^{\mathsf{W}} - m_\star = e^{-\mu_0 t C_\star^{-1}}(m_0 - m_\star), \qquad C_t^{\mathsf{W}} = C_\star + e^{-\mu_0 t C_\star^{-1}}(C_0 - C_\star)e^{-\mu_0 t C_\star^{-1}}. \tag{9}$$

Theorem B.1 in Appendix B.1 gives quantitative convergence bounds for arbitrary Gaussian initialization. For illustration, the following theorem fixes $C_0 = \varepsilon C_\star$ and $C_0 = C_\star$ to isolate the covariance-recovery and anisotropic mean-refinement regimes, respectively.

**Theorem 3.1** (Complementary convergence regimes for Gaussian target $\pi = \mathcal{N}(m_\star, C_\star)$)**.**

1. ***Covariance recovery.*** *Suppose $C_0 = \varepsilon C_\star$ with $0 < \varepsilon < 1/2$. For $\mathrm{G} \in \{\mathsf{W}, \mathsf{FR}\}$, define $T_{\mathrm{G}}^{\mathrm{init}} := \inf\left\{t \ge 0 : C_t^{\mathrm{G}} \succeq \frac{\mu_0}{2} I\right\}$, then*

$$T_{\mathsf{W}}^{\mathrm{init}} \le \frac{\log 2}{2}, \qquad T_{\mathsf{FR}}^{\mathrm{init}} = \log\frac{1-\varepsilon}{\varepsilon}.$$

2. ***Anisotropic refinement.*** *Suppose $C_0 = C_\star$ and $m_0 \ne m_\star$. For $0 < \delta < 1$ and $\mathrm{G} \in \{\mathsf{W}, \mathsf{FR}\}$, define*

$$T_{\mathrm{G}}^{\mathrm{ref}}(\delta) := \inf\left\{t \ge 0 : (m_t^{\mathrm{G}} - m_\star)^\top C_\star^{-1}(m_t^{\mathrm{G}} - m_\star) \le \delta (m_0 - m_\star)^\top C_\star^{-1}(m_0 - m_\star)\right\},$$

*then*

$$\frac{1}{2}\log\frac{1}{\delta} \le T_{\mathsf{W}}^{\mathrm{ref}}(\delta) \le \frac{\kappa}{2}\log\frac{1}{\delta}, \qquad T_{\mathsf{FR}}^{\mathrm{ref}}(\delta) = \frac{1}{2}\log\frac{1}{\delta}.$$

Thus Wasserstein reaches the covariance-recovery scale in $O(1)$ time while Fisher–Rao requires $\Theta(\log \varepsilon^{-1})$; conversely, Fisher–Rao refinement is independent of target conditioning, whereas the worst-case Wasserstein refinement time scales linearly with $\kappa$.

**Corollary 3.2** (Convergence of Wasserstein-to-Fisher–Rao switching)**.** *Let $\pi = \mathcal{N}(m_\star, C_\star)$ and suppose $0 \prec C_0 \preceq C_\star$. Run the Wasserstein flow until time $(\log 2)/2$, and then switch to the Fisher–Rao flow. Then, for every $t \ge 0$,*

$$\mathrm{KL}\left(q_{\frac{\log 2}{2} + \log(2\kappa) + t} \,\middle\|\, \pi\right) \le e^{-2t}\left(\frac{1}{2}\left\|C_\star^{-1/2}(m_0 - m_\star)\right\|^2 + \frac{d}{4}\right).$$

*Consequently, the switched flow reaches KL error $\delta$ in $O(1 + \log\kappa + \log\delta^{-1})$ time. The additive $\log\kappa$ term accounts for the relative covariance mismatch remaining at the switch; thereafter the Fisher–Rao convergence rate is independent of $\kappa$.*

**Algorithm 1** Componentwise switched Gaussian-mixture WFR

**Require:** Iteration count $N$; initial mixture $q_0 = \sum_{k=1}^{K} w_{0,k}\mathcal{N}(m_{0,k}, C_{0,k})$ with $C_{0,k} \in \mathcal{C}$; step-sizes $\eta_{n,k}, \tau_{n,k}, \nu_n > 0$; switching rule

1: **for** $n = 0, \dots, N-1$ **do**
2: Compute $g_{n,k}$, $H_{n,k}$, $\ell_{n,k}$ at $q_n$ for $k = 1, \dots, K$.
3: Choose $G_{n,k} \in \{\mathsf{W}, \mathsf{FR}\}$ for each $k$.
4: Update weights: $w_{n+1,k} \propto w_{n,k} \exp(\nu_n \ell_{n,k})$.
5: Update components: $q_{n+1,k}$ via (10) for each $k$.
6: $q_{n+1} \leftarrow \sum_{k=1}^{K} w_{n+1,k}\, q_{n+1,k}$.
7: **end for**
8: **return** $q_N$.

The proofs are provided in Appendix B.1. Together, Theorem 3.1 and Corollary 3.2 advocate a one-way Wasserstein-to-Fisher–Rao switch: Wasserstein removes the small-covariance initialization bottleneck, after which Fisher–Rao avoids the worst-case linear dependence on $\kappa$. The covariance-recovery separation also persists in discrete time for smooth log-concave targets (Proposition B.2).

## 4 Switched WFR algorithm for Gaussian mixtures

We now carry the Wasserstein-to-Fisher–Rao switch over to Gaussian mixtures. Since different components may reach the covariance-recovery scale at different times, we let each component choose its own geometry at each iteration.

### 4.1 Switched component and weight updates

At iteration $n$, write $q_n = \sum_{k=1}^{K} w_{n,k} q_{n,k}$ with $q_{n,k} = \mathcal{N}(m_{n,k}, C_{n,k})$, and recall the component statistics $g_{n,k}$, $H_{n,k}$, and $\ell_{n,k}$ from Section 2.1. Each component selects a geometry $G_{n,k} \in \{\mathsf{W}, \mathsf{FR}\}$ and is updated by the corresponding covariance-constrained map,

$$q_{n+1,k} := \begin{cases} \mathsf{W}^{\mathcal{C}}_{\eta_{n,k}}(q_{n,k}; g_{n,k}, H_{n,k}), & G_{n,k} = \mathsf{W}, \\ \mathsf{FR}^{\mathcal{C}}_{\tau_{n,k}}(q_{n,k}; g_{n,k}, H_{n,k}), & G_{n,k} = \mathsf{FR}, \end{cases} \tag{10}$$

using the covariance box $\mathcal{C}$ of Section 2.2, with closed forms given in Lemma A.5. The mixture weights follow an entropic mirror step, equivalently an explicit-Euler discretization of the Fisher–Rao natural-gradient dynamics in logarithmic coordinates (Ying, 2020; Beck & Teboulle, 2003),

$$w_{n+1,k} := \frac{w_{n,k} \exp(\nu_n \ell_{n,k})}{\sum_{j=1}^{K} w_{n,j} \exp(\nu_n \ell_{n,j})}, \qquad k = 1, \dots, K.$$

All component updates use statistics evaluated at the current mixture $q_n$, so each iteration requires only one round of target-statistic evaluations. Algorithm 1 summarizes the procedure.

### 4.2 Descent and stationarity under componentwise switching

We first give guarantees for Algorithm 1 that hold for any switching rule, and discuss practical switching rules in Section 4.3. The main theoretical question is whether such componentwise switching is compatible with descent of $\mathcal{F}$, as the mixture entropy $\mathbb{E}_q[\log q]$ couples all components, so descent estimates for Wasserstein and Fisher–Rao updates cannot simply be applied componentwise and added together. We therefore introduce a separable upper bound on $\mathcal{F}$.

At the current mixture $q_n$, define $r_{n,k}(\theta) := w_{n,k} q_{n,k}(\theta)/q_n(\theta)$, and for any candidate mixture $q = \sum_{k=1}^{K} w_k q_k$ with $r_k(\theta) := w_k q_k(\theta)/q(\theta)$, let

$$\mathcal{Q}_n(q) := \sum_{k=1}^{K} w_k \log w_k + \sum_{k=1}^{K} w_k \mathbb{E}_{q_k}[\log q_k + V - \log r_{n,k}] + \log Z.$$

**Proposition 4.1** (Separable upper bound)**.** *For every positive-weight Gaussian mixture $q$,*

$$\mathcal{Q}_n(q) - \mathcal{F}(q) = \mathbb{E}_q\left[\sum_{k=1}^{K} r_k(\theta)\log\frac{r_k(\theta)}{r_{n,k}(\theta)}\right] \geq 0.$$

*Moreover, $\mathcal{Q}_n(q_n) = \mathcal{F}(q_n)$, their first derivatives with respect to the component parameters and mixture weights agree at $q_n$, and $\mathcal{Q}_n$ separates over the Gaussian components and mixture weights.*

The proposition isolates the coupling that prevents a direct componentwise descent argument. We then prove the descent of $\mathcal{Q}_n$, which leads to the descent of $\mathcal{F}$; see also Arenz et al. (2020; 2023). To make the resulting one-step bounds uniform over the iterates, we impose one additional boundedness condition beyond the smoothness assumption on the target.

**Assumption 4.2** (Bounded component means)**.** There exists $R < \infty$ such that the iterates generated by Algorithm 1 satisfy $\|m_{n,k}\| \leq R$ for all $n \geq 0$ and $k = 1, \ldots, K$.

For each iteration, let $\mathcal{I}_n^{\mathsf{W}} := \{k : G_{n,k} = \mathsf{W}\}$ and $\mathcal{I}_n^{\mathsf{FR}} := \{k : G_{n,k} = \mathsf{FR}\}$. Define the residual

$$\mathcal{R}_n := \sum_{k\in\mathcal{I}_n^{\mathsf{W}}} w_{n+1,k}\frac{W_2^2(q_{n+1,k}, q_{n,k})}{\eta_{n,k}^2} + \sum_{k\in\mathcal{I}_n^{\mathsf{FR}}} w_{n+1,k}\frac{\mathrm{KL}(q_{n+1,k}\|q_{n,k})}{\tau_{n,k}^2} + \frac{\mathrm{KL}(w_n\|w_{n+1})}{\nu_n^2}.$$

By Lemma C.4, $\mathcal{R}_n = 0$ if and only if all selected component maps and the weight map are fixed, which is equivalent to first-order stationarity of the covariance-constrained mixture objective.

**Theorem 4.3** (Descent and stationarity of Algorithm 1)**.** *Under Assumptions 2.1 and 4.2, there exist $\bar{\eta}, \bar{\tau} > 0$, depending only on $\beta$, $\mu_-$, $\mu_+$, $d$, and $R$, such that the following holds. If $0 < \eta_{n,k} \leq \bar{\eta}$, $0 < \tau_{n,k} \leq \bar{\tau}$, and $0 < \nu_n \leq 1$, then, for any choice of $G_{n,k} \in \{\mathsf{W}, \mathsf{FR}\}$, which may depend on $q_n$,*

$$\mathcal{F}(q_{n+1}) \leq \mathcal{F}(q_n) - \sum_{k\in\mathcal{I}_n^{\mathsf{W}}}\frac{w_{n+1,k}}{4\eta_{n,k}}W_2^2(q_{n+1,k}, q_{n,k}) - \sum_{k\in\mathcal{I}_n^{\mathsf{FR}}}\frac{w_{n+1,k}}{2\tau_{n,k}}\mathrm{KL}(q_{n+1,k}\|q_{n,k}) - \frac{\mathrm{KL}(w_n\|w_{n+1})}{\nu_n}.$$

*If, in addition, all active stepsizes are bounded below by $h > 0$ and $\mathcal{F}_{\inf}$ denotes the infimum of $\mathcal{F}$ over positive-weight mixtures with covariances in $\mathcal{C}$, then*

$$\min_{0\leq n<N}\mathcal{R}_n \leq \frac{4\big(\mathcal{F}(q_0) - \mathcal{F}_{\inf}\big)}{Nh} = O(N^{-1}).$$

The proof is given in Appendix C.4. Because the descent inequality holds for every admissible componentwise choice, the guarantee applies to the adaptive switching rule below. The analysis uses exact statistics $g_{n,k}$, $H_{n,k}$, and $\ell_{n,k}$; in the experiments they are estimated by Monte Carlo.

### 4.3 Practical switching rules

Motivated by the complementary Gaussian regimes (Theorem 3.1), we use a one-way rule: each component begins with Wasserstein and, once switched, remains with Fisher–Rao.

**Covariance-recovery score.** To detect whether a component is still underdispersed, we compare its current covariance with the local target curvature. For component $k$ at iteration $n$, let

$$z_i(C) := \mu_i\big(C\,\mathbb{E}_{q_{n,k}}[\nabla^2 V]\big), \qquad \mathcal{T}_{n,k}(C) := \max\Big\{0, -\log\min_i z_i(C)\Big\}, \tag{11}$$

where $\mu_i$ denotes the $i$-th eigenvalue; the score is defined when $\mathbb{E}_{q_{n,k}}[\nabla^2 V] \succ 0$. If the expected target Hessian is not positive definite, the score does not trigger a switch and Wasserstein is retained unless the cap $n_{\mathsf{W}}^{\max}$ below is reached. For $K = 1$ and $\pi = \mathcal{N}(m_\star, C_\star)$ with $C = \varepsilon C_\star$, one obtains $\mathcal{T} = \log\varepsilon^{-1}$, which matches the Fisher–Rao initialization delay of Theorem 3.1 as $\varepsilon \to 0$.

**Switching rule.** Each component starts with Wasserstein and switches permanently to Fisher–Rao at the first iteration at which it has taken $n_{\mathsf{W}}^{\max}$ Wasserstein steps, its score vanishes, or a Fisher–Rao step would lower the score at least as much as a Wasserstein step, $\mathcal{T}_{n,k}(\widehat{C}_{n,k}^{\mathsf{FR}}) \leq \mathcal{T}_{n,k}(\widehat{C}_{n,k}^{\mathsf{W}})$. Here $\widehat{C}_{n,k}^{\mathsf{W}}$ and $\widehat{C}_{n,k}^{\mathsf{FR}}$ are the candidate covariances from (10), computed with the current $g_{n,k}$ and $H_{n,k}$. The test requires no additional target evaluations, and its linear-algebra cost, of the same order as a component update, is incurred only until the component switches; it is secondary when target evaluations dominate the cost, as in PDE-constrained problems. More details are in Appendix E.

## 5 NUMERICAL EXPERIMENTS

We evaluate the switched WFR algorithm on three targets of different nature: a two-dimensional 40-mode Gaussian mixture, a 196-dimensional learned digit distribution, and a 64-dimensional PDE-constrained Bayesian inverse problem. Appendices I and J report three further posteriors.

**Baselines.** We compare against three Gaussian mixture VI methods. Isotropic Bures–Wasserstein (IBW; Petit-Talamon et al., 2025) restricts every component to an isotropic covariance and uses a Wasserstein proximal update. Gaussian mixture black-box variational inference (GMBBVI; Che et al., 2026) is a derivative-free method based on Fisher–Rao natural-gradient dynamics. SEMTRUX (Arenz et al., 2023) is a modular Fisher–Rao natural-gradient method that builds on Arenz et al. (2020) and Lin et al. (2019).

**Common protocol.** All methods use the same number $K$ of mixture components and the same initialization for each experiment. We denote by $J$ the Monte Carlo batch size per component used to estimate the required expectations. The full-covariance methods (GMBBVI, SEMTRUX, and WFR) share the same $J$ for each target, while IBW uses its own batch size, smaller than $J$ on NICE digits and Darcy, consistent with its lower-dimensional isotropic covariance parameterization. For methods requiring expected Hessians of the log-density ratio, we use Stein's identity to avoid explicit Hessian computation. In figures and tables, WFR denotes Algorithm 1 with the switching rule of Section 4.3. Implementation details and hyperparameters are in Appendices D and K.

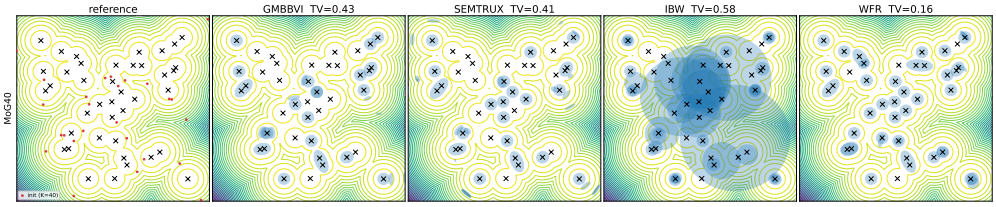
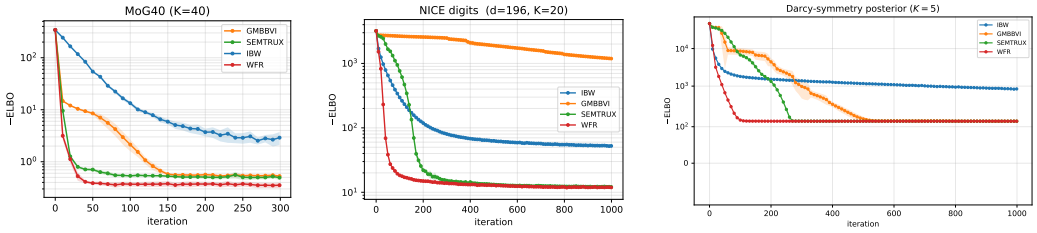



Figure 1: Final mixtures for MoG40 with $K = 40$ components. Red dots show the initial component means; black crosses mark the true mode means; blue ellipses show the fitted Gaussian components.




(a) MoG40. (b) NICE digits. (c) Darcy.

Figure 2: $-$ELBO against iteration on the three targets, mean over seeds.

**Two-dimensional mixture of 40 Gaussians.** We begin with the MoG40 benchmark (Midgley et al., 2023), fitting $K = 40$ components with batch size $J = 64$. All methods are initialized with means sampled from $\mathcal{N}(0, 40^2 I)$ to cover the mode grid, and a small covariance $C_i = 10^{-5} I$. This small initial covariance is challenging for Fisher–Rao-based methods: their mean update is preconditioned by the covariance and is nearly frozen until the covariance recovers, which by Theorem 3.1 takes time logarithmic in the inverse initial scale.

Figure 1 shows the fitted mixtures and the total variation of each method after 300 iterations. The convergence curves in Figure 2a confirm that WFR converges faster during both the warm-up phase and the subsequent refinement. The Wasserstein phase enables rapid covariance recovery in early iterations, while the Fisher–Rao switch accelerates convergence once the components reach high-probability regions. Appendix F compares all four methods under other initializations.

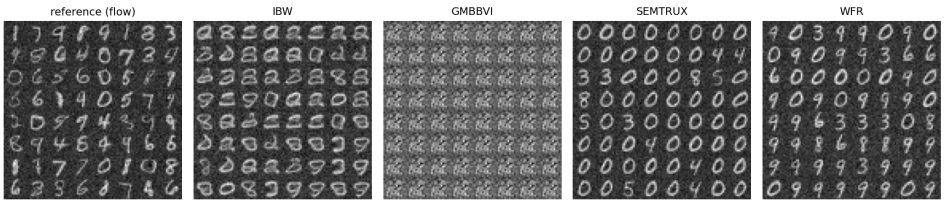



Figure 3: NICE digits: samples drawn from the reference flow and from each method's final mixture.

**High-dimensional NICE digits target.** We next approximate the log-density of a pretrained NICE normalizing flow (Dinh et al., 2014) fitted to $14 \times 14$ MNIST digits ($d = 196$). All methods use $K = 20$ components and run for 1000 iterations from the initialization $m_i \sim \mathcal{N}(0, I)$, $C_i = 10^{-4} I$. WFR drives the $-$ELBO down fastest (Figure 2b), and its final mixture produces recognizable digits from more classes than the other methods (Figure 3). Appendix G gives the remaining settings and a UMAP projection of the fitted mixtures.

**Darcy inverse problem.** Finally, we consider a Bayesian inverse problem for a 64-dimensional log-permeability field, inferred from 120 noisy pressure observations governed by the Darcy flow equation. Averaging observations at reflected sensor locations induces a reflection-symmetric posterior, so the true field and its mirror image are equally consistent with the data. All methods use $K = 5$ components and run for 1000 iterations. WFR reaches its $-$ELBO floor several times sooner than SEMTRUX and GMBBVI (Figure 2c), splits its mass between the two symmetric modes, and recovers both the true field and its mirror image (Figure 4), showing that the method applies to computationally demanding PDE-constrained inverse problems. Appendix H gives the setup and the full comparison.

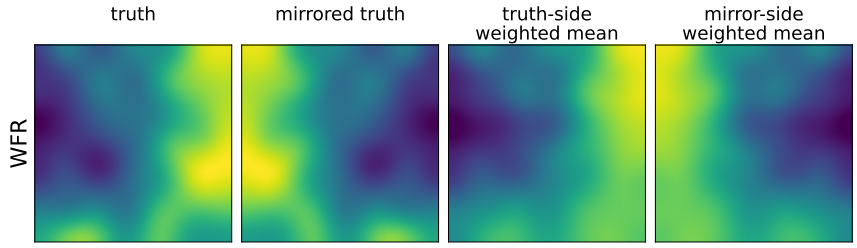



Figure 4: Darcy inverse problem. Left two figures: the true log-permeability field and its mirror image. Right two figures: the WFR mixture mean conditional on each side of the symmetry.

Across these experiments, part of the margin of WFR reflects the small initial covariances, which easily arise when the target scale is unknown and under which the switched method is more robust. With better initial covariances on MoG40, the gap narrows, and WFR still gives the best fit in our test (Appendix F).

## 6 Conclusions

This work takes a step toward Gaussian mixture VI that adapts the geometry of each mixture component to the target. We studied a one-way switch from Wasserstein to Fisher–Rao geometry; for Gaussian targets, the switched flow provably combines the advantages of both geometries. The resulting algorithm uses one round of target-statistic evaluations per iteration, satisfies descent and stationarity guarantees for general smooth targets under any switching rule, and is markedly more robust to initialization in experiments. These guarantees assume exact statistics, whereas the experiments use Monte Carlo estimates. Closing this gap, other ways of adapting the geometry, and quantifying the benefit of switching beyond Gaussian targets are left for future work.

### AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI's ChatGPT-5.6 Sol and Anthropic's Claude Fable 5.1, to assist with mathematical derivations and proof checking, provide feedback on the design and implementation of numerical experiments, and edit the manuscript for clarity and presentation. All AI-assisted mathematical arguments, experimental results, and cited references were independently checked by the authors. We take responsibility for the final content of this work, including all text, claims, proofs, code, and experimental results produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

The assumptions of our theoretical results are stated in Assumptions 2.1, 2.2, and 4.2, and the proofs are given in Appendices A, B, and C. Algorithm 1, Section 4.3, and Appendix E specify the method, including the switching rule, the Stein estimator, and the stepsizes. Appendix D describes the baselines and metrics, Appendices F–J the targets, and Appendix K all initializations and hyperparameters.

## REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Oleg Arenz, Mingjun Zhong, and Gerhard Neumann. Trust-region variational inference with gaussian mixture models. *Journal of Machine Learning Research*, 21:1–60, 2020.

Oleg Arenz, Philipp Dahlinger, Zihan Ye, Michael Volpp, and Gerhard Neumann. A unified perspective on natural gradient variational inference with Gaussian mixture models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=tLBjsX4tjs.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. Beyond ELBOs: A large-scale evaluation of variational methods for sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2406.07423.

Nawaf Bou-Rabee, Bob Carpenter, Tore Selland Kleppe, and Sifan Liu. The within-orbit adaptive leapfrog no-u-turn sampler. *Journal of Machine Learning Research*, 27(113):1–64, 2026.

Baojun Che, Yifan Chen, Zhenghao Huan, Daniel Zhengyu Huang, and Weijie Wang. Stable derivative-free Gaussian mixture variational inference for Bayesian inverse problems. *SIAM Journal on Scientific Computing*, 47(5):A2583–A2608, 2025. doi: 10.1137/25M1722548.

Baojun Che, Yifan Chen, Daniel Zhengyu Huang, Xinying Mao, and Weijie Wang. Adaptive exponential integration for stable Gaussian mixture black-box variational inference. *arXiv preprint arXiv:2601.14855*, 2026. URL https://arxiv.org/abs/2601.14855.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M. Stuart. Efficient, multimodal, and derivative-free Bayesian inference with Fisher–Rao gradient flows. *Inverse Problems*, 40(12):125001, 2024. doi: 10.1088/1361-6420/ad847b.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M. Stuart. Sampling via gradient flows in the space of probability measures. *Mathematics of Computation*, 2026. URL https://arxiv.org/abs/2310.03597.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018. doi: 10.1016/j.jfa.2018.01.008.

Francesca R. Crucinio and Sahani Pathiraja. An operator splitting analysis of Wasserstein–Fisher–Rao gradient flows. *arXiv preprint arXiv:2511.18060*, 2026. URL https://arxiv.org/abs/2511.18060.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward–backward Gaussian variational inference via JKO in the Bures–Wasserstein space. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7960–7991. PMLR, 2023. URL https://proceedings.mlr.press/v202/diao23a.html.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. doi: 10.48550/arXiv.1410.8516. URL https://arxiv.org/abs/1410.8516.

Thomas O. Gallouët and Léonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017. doi: 10.1137/16M106666X.

Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M Stuart. Interacting langevin diffusions: Gradient structure and ensemble kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Thomas Guilmeau, Hadrien Hendrikx, and Florence Forbes. Convergence of projected stochastic natural gradient variational inference for various step size and sample or batch size schedules. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=ErbBuMVhRi.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Tom Huix, Anna Korba, Alain Durmus, and Eric Moulines. Theoretical guarantees for variational inference with fixed-variance mixture of Gaussians. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20700–20721. PMLR, 2024. URL https://proceedings.mlr.press/v235/huix24a.html.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359.

Navish Kumar, Thomas Möllenhoff, Mohammad Emtiyaz Khan, and Aurelien Lucchi. Optimization guarantees for square-root natural-gradient variational inference. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=OMOFmb6ve7.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14434–14447, 2022. URL https://arxiv.org/abs/2205.15902.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy–transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211 (3):969–1117, 2018. doi: 10.1007/s00222-017-0759-8.

Matthias Liero, Alexander Mielke, Oliver Tse, and Jia-Jie Zhu. Evolution of Gaussians in the Hellinger–Kantorovich–Boltzmann gradient flow. *Communications on Pure and Applied Analysis*, 31:166–198, 2026. doi: 10.3934/cpaa.2025105. URL https://www.aimsciences.org/article/doi/10.3934/cpaa.2025105.

Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3992–4002. PMLR, 2019. URL https://proceedings.mlr.press/v97/lin19b.html.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XCTVFJwS9LJ.

Marguerite Petit-Talamon, Marc Lambert, and Anna Korba. Variational inference with mixtures of isotropic Gaussians. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4844. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d551343f85fcf5e1a230fd393406306e-Abstract-Conference.html.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538. PMLR, 2015. URL https://proceedings.mlr.press/v37/rezende15.html.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Andrew M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Fangyuan Sun, Ilyas Fatkhullin, and Niao He. Natural gradient VI: Guarantees for non-conjugate models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/9a07bb7288caaea2ecc4c367188bc6db-Abstract-Conference.html.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011. URL https://projecteuclid.org/euclid.ojm/1326291215.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993.

Kaiwen Wu and Jacob R. Gardner. Understanding stochastic natural gradient variational inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53398–53421. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24f.html.

Lexing Ying. Mirror descent algorithms for minimizing interacting free energy. *Journal of Scientific Computing*, 84(3):51, 2020. doi: 10.1007/s10915-020-01303-z.

The appendix is organized as follows. Appendices A–C contain the proofs: Appendix A proves the Gaussian WFR results of Section 2, namely exponential convergence of the flow and the splitting and stationarity for general smooth targets; Appendix B proves the complementary regimes of Section 3 and their discrete extension to log-concave targets; and Appendix C proves the descent and stationarity guarantees of Section 4. Appendix D describes the baselines and evaluation metrics, and Appendix E the switching rule and stepsizes used in the experiments. Appendices F–J report further experiments: MoG40 under different initializations with an ablation of the two geometries, details of the NICE digits and Darcy experiments, two two-dimensional posteriors with a comparison to MCMC, and Bayesian logistic regression. Appendix K lists all hyperparameters.

# A Theory for Gaussian Wasserstein–Fisher–Rao VI

## A.1 Continuous-time Gaussian WFR flow

The proof uses a uniform lower bound on the covariance to compare the Fisher–Rao and Wasserstein gradient norms. Together with the Wasserstein PL inequality, this converts the energy-dissipation identity below into exponential convergence.

Following the Gaussian Hellinger–Kantorovich energy-dissipation calculation of Liero et al. (2026), the Gaussian WFR flow satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(q_t) = -\lambda_t^{\mathsf{W}} \left\|\mathrm{grad}_{\mathsf{W}}\, \mathcal{F}(q_t)\right\|_{\mathsf{W}}^2 - \lambda_t^{\mathsf{FR}} \left\|\mathrm{grad}_{\mathsf{FR}}\, \mathcal{F}(q_t)\right\|_{\mathsf{FR}}^2 . \tag{12}$$

*Proof of Theorem 2.3.* Under Assumptions 2.1 and 2.2, the definition of $H_t$ gives

$$C_t^{-1} - \beta I \preceq H_t \preceq C_t^{-1} - \alpha I. \tag{13}$$

Applying (13) to (3) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_{\min}(C_t) \geq \left(2\lambda_t^{\mathsf{W}} + \lambda_t^{\mathsf{FR}}\mu_{\min}(C_t)\right)\left(1 - \beta\mu_{\min}(C_t)\right),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_{\max}(C_t) \leq \left(2\lambda_t^{\mathsf{W}} + \lambda_t^{\mathsf{FR}}\mu_{\max}(C_t)\right)\left(1 - \alpha\mu_{\max}(C_t)\right).$$

Scalar comparison therefore yields

$$\underline{\mu} I \preceq C_t \preceq \overline{\mu} I, \qquad t \geq 0.$$

Since $C_t \succeq \underline{\mu} I$, (2) gives

$$\left\|\mathrm{grad}_{\mathsf{FR}}\, \mathcal{F}(q_t)\right\|_{\mathsf{FR}}^2 \geq \frac{\underline{\mu}}{2} \left\|\mathrm{grad}_{\mathsf{W}}\, \mathcal{F}(q_t)\right\|_{\mathsf{W}}^2 .$$

The Gaussian Wasserstein PL inequality (Lambert et al., 2022, Appendix D) gives

$$\left\|\mathrm{grad}_{\mathsf{W}}\, \mathcal{F}(q_t)\right\|_{\mathsf{W}}^2 \geq 2\alpha \left(\mathcal{F}(q_t) - \mathcal{F}(q_\star)\right). \tag{14}$$

Combining the preceding inequalities yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{F}(q_t) - \mathcal{F}(q_\star)\right) \leq -\alpha \left(2\lambda_t^{\mathsf{W}} + \underline{\mu}\lambda_t^{\mathsf{FR}}\right)\left(\mathcal{F}(q_t) - \mathcal{F}(q_\star)\right),$$

and the conclusion follows from Grönwall's inequality. □

## A.2 Gaussian forward–backward maps

**Fisher–Rao KL-proximal forward–backward step.** Write

$$\mathcal{F}(q) = \mathcal{V}(q) + \mathcal{H}(q) + \log Z, \qquad \mathcal{V}(q) := \mathbb{E}_q[V], \qquad \mathcal{H}(q) := \mathbb{E}_q[\log q].$$

Given $q = \mathcal{N}(m, C)$, we linearize $\mathcal{V}$ in $(m, C)$, keep $\mathcal{H}$ implicit, and use the Gaussian KL divergence as the proximal term:

$$q^+ \in \operatorname*{arg\,min}_{q' = \mathcal{N}(m', C')} \left\{ \mathcal{H}(q') + \langle \nabla_m \mathcal{V}(q), m' - m\rangle + \langle \nabla_C \mathcal{V}(q), C' - C\rangle + \frac{1}{h}\mathrm{KL}(q'\|q) \right\}. \tag{15}$$

Since $\nabla_m \mathcal{V}(q) = -g$ and $\nabla_C \mathcal{V}(q) = \frac{1}{2}(C^{-1} - H)$, the first-order conditions give

$$m^+ = m + hCg, \qquad C^+ = \left(C^{-1} - \frac{h}{1+h}H\right)^{-1}. \tag{16}$$

Thus (6) is the Fisher–Rao KL-proximal forward–backward step evaluated at $q_{n+1/2}$. Moreover, as $h \to 0$,

$$m^+ = m + hCg + O(h^2), \qquad C^+ = C + hCHC + O(h^2),$$

so the update is a first-order discretization of the Gaussian Fisher–Rao gradient flow in (1).

For the covariance-constrained variants, recall the covariance box

$$\mathcal{C} := \{C \in \mathbb{S}^d_{++} : \mu_- I \preceq C \preceq \mu_+ I\}.$$

For $A = O\,\mathrm{diag}(a_1, \dots, a_d)O^\top$, define the spectral projection

$$\Pi_{\mathcal{C}}(A) := O\,\mathrm{diag}\left(\min\{\mu_+, \max\{\mu_-, a_1\}\}, \dots, \min\{\mu_+, \max\{\mu_-, a_d\}\}\right) O^\top.$$

**Wasserstein descent.** The following descent estimate follows from the deterministic FB–GVI analysis of Diao et al. (2023) and applies both with and without the covariance constraint.

**Lemma A.1** (Wasserstein one-step descent)**.** *Under Assumption 2.1, the Wasserstein forward–backward update, with or without the covariance constraint $\mathcal{C}$, satisfies*

$$\mathcal{F}(q^{\mathsf{W}}) \le \mathcal{F}(q) - \left(\frac{1}{2h} - \frac{\beta}{2}\right) W_2^2(q^{\mathsf{W}}, q). \tag{17}$$

*In particular, if $h \le 1/(2\beta)$, then*

$$\mathcal{F}(q^{\mathsf{W}}) \le \mathcal{F}(q) - \frac{1}{4h} W_2^2(q^{\mathsf{W}}, q).$$

*Proof.* The potential functional $q \mapsto \mathbb{E}_q[V]$ is $\beta$-smooth in Bures–Wasserstein geometry under Assumption 2.1. The deterministic forward–backward argument of Diao et al. (2023) therefore gives the first inequality. Restricting the covariance proximal problem to $\mathcal{C}$ does not change the argument, since the current covariance is feasible and the proof uses only comparison with the current iterate. The second inequality follows from $h \le 1/(2\beta)$. □

### A.3 Exponential convergence of the Gaussian WFR splitting

Both substeps preserve a common spectral interval for the covariance. On this interval, the Fisher–Rao step contracts the objective gap, while the Wasserstein step does not increase it. These estimates give the contraction of the full splitting step in Theorem 2.4.

**Lemma A.2.** *Under Assumptions 2.1 and 2.2, suppose $\underline{\mu} I \preceq C_n \preceq \overline{\mu} I$ and $0 < h \le 1/\beta$. Then the splitting (5)–(6) satisfies*

$$\underline{\mu} I \preceq C_{n+1/2}, C_{n+1} \preceq \overline{\mu} I.$$

*Proof.* For the Wasserstein forward step, (13) gives

$$(1 - h\beta)I \preceq I + h(H_n - C_n^{-1}) \preceq (1 - h\alpha)I,$$

and hence

$$(1 - h\beta)^2 \underline{\mu} I \preceq \widetilde{C}_{n+1/2} \preceq (1 - h\alpha)^2 \overline{\mu} I.$$

The scalar covariance proximal map

$$s \mapsto \frac{1}{2}\left(s + 2h + \sqrt{s(s+4h)}\right)$$

is increasing, has value $h$ at $s = 0$, and satisfies

$$\frac{1}{2}\left[c\left(1 - \frac{h}{c}\right)^2 + 2h + \sqrt{c\left(1 - \frac{h}{c}\right)^2 \left(c\left(1 - \frac{h}{c}\right)^2 + 4h\right)}\right] = c, \qquad 0 \le h \le c.$$

If $h \geq \underline{\mu}$, then $C_{n+1/2} \succeq hI \succeq \underline{\mu} I$. If $h < \underline{\mu}$, then $\underline{\mu} \leq 1/\beta$ implies

$$(1 - h\beta)^2 \underline{\mu} \geq \left(1 - \frac{h}{\underline{\mu}}\right)^2 \underline{\mu},$$

so monotonicity of the proximal map again gives $C_{n+1/2} \succeq \underline{\mu} I$. Similarly, $\overline{\mu} \geq 1/\alpha$ and $h \leq 1/\beta$ imply

$$(1 - h\alpha)^2 \overline{\mu} \leq \left(1 - \frac{h}{\overline{\mu}}\right)^2 \overline{\mu},$$

and therefore $C_{n+1/2} \preceq \overline{\mu} I$.

For the Fisher–Rao step, (16) and $H = C^{-1} - \mathbb{E}_q[\nabla^2 V]$ give

$$(C^+)^{-1} = \frac{1}{1+h}\left(C^{-1} + h\mathbb{E}_q[\nabla^2 V]\right).$$

Hence

$$\frac{1}{1+h}\left(C^{-1} + h\alpha I\right) \preceq (C^+)^{-1} \preceq \frac{1}{1+h}\left(C^{-1} + h\beta I\right).$$

Using $\underline{\mu} \leq 1/\beta$ and $\overline{\mu} \geq 1/\alpha$ gives

$$\overline{\mu}^{-1} I \preceq (C^+)^{-1} \preceq \underline{\mu}^{-1} I,$$

and therefore $\underline{\mu} I \preceq C^+ \preceq \overline{\mu} I$. □

**Fisher–Rao one-step contraction.** To control the Fisher–Rao step, we bound the error in the potential linearization by the Gaussian KL divergence.

**Lemma A.3** (Relative smoothness for log-concave targets)**.** *Suppose Assumptions 2.1 and 2.2 hold and $q_0 = \mathcal{N}(m_0, C_0)$ satisfies $C_0 \preceq \overline{\mu} I$. Then, for every Gaussian $q = \mathcal{N}(m, C)$,*

$$\mathbb{E}_q[V] \leq \mathbb{E}_{q_0}[V] + \langle \nabla_m \mathbb{E}_{q_0}[V], m - m_0 \rangle + \langle \nabla_C \mathbb{E}_{q_0}[V], C - C_0 \rangle + \beta\overline{\mu}\,\mathrm{KL}(q\|q_0).$$

*Proof.* Let $T(\theta) = m + B(\theta - m_0)$ be the optimal affine transport from $q_0$ to $q$, where $B = B^\top \succ 0$ and $BC_0B = C$. By $\beta$-smoothness of $V$,

$$\mathbb{E}_q[V] \leq \mathbb{E}_{q_0}[V] + \mathbb{E}_{q_0}\left[\nabla V(\theta)^\top (T(\theta) - \theta)\right] + \frac{\beta}{2} W_2^2(q, q_0).$$

Stein's identity and $\nabla^2 V \succeq 0$ give

$$\mathbb{E}_{q_0}\left[\nabla V(\theta)^\top (T(\theta) - \theta)\right] \leq \langle \nabla_m \mathbb{E}_{q_0}[V], m - m_0 \rangle + \langle \nabla_C \mathbb{E}_{q_0}[V], C - C_0 \rangle,$$

because the difference between the covariance linearization and the transport linearization equals

$$\frac{1}{2}\,\mathrm{Tr}\left(\mathbb{E}_{q_0}[\nabla^2 V](B - I)C_0(B - I)\right) \geq 0.$$

Since $C_0 \preceq \overline{\mu} I$, the Gaussian reference measure $q_0 = \mathcal{N}(m_0, C_0)$ is $(1/\overline{\mu})$-strongly log-concave. Therefore, by the Talagrand $T_2$ inequality (Villani, 2009, Theorem 22.10),

$$W_2^2(q, q_0) \leq 2\overline{\mu}\,\mathrm{KL}(q\|q_0),$$

which proves the claim. □

The contraction estimate also uses an exact identity for the proximal update. In this identity, the mean correction from the $(m, C)$ linearization cancels part of the reverse-KL mean term.

**Lemma A.4** (Fisher–Rao contraction)**.** *Under Assumptions 2.1 and 2.2, suppose $\underline{\mu} I \preceq C \preceq \overline{\mu} I$ and $0 < h \leq 1/(\beta\overline{\mu})$. Then the Fisher–Rao update (16) satisfies*

$$\mathcal{F}(q^+) - \mathcal{F}(q_\star) \leq \left(1 - \frac{\alpha\underline{\mu}h}{8}\right)\left(\mathcal{F}(q) - \mathcal{F}(q_\star)\right).$$

*Proof.* Set $G := \mathbb{E}_q[\nabla^2 V] = C^{-1} - H$ and $\Delta m := m^+ - m = hCg$. By (16),

$$(C^+)^{-1} = \frac{1}{1+h}\left(C^{-1} + hG\right).$$

A direct calculation using the Gaussian KL formula and (16) gives

$$\begin{aligned}\mathcal{H}(q) - \mathcal{H}(q^+) - \left\langle \nabla_m \mathcal{V}(q), m^+ - m\right\rangle - \left\langle \nabla_C \mathcal{V}(q), C^+ - C\right\rangle \\ = \left(1 + \frac{1}{h}\right)\mathrm{KL}(q\|q^+) + \frac{1}{h}\mathrm{KL}(q^+\|q) - \frac{1}{2}\Delta m^\top G \Delta m.\end{aligned}$$

Combining this identity with Lemma A.3 yields

$$\mathcal{F}(q) - \mathcal{F}(q^+) \geq \left(1 + \frac{1}{h}\right)\mathrm{KL}(q\|q^+) - \frac{1}{2}\Delta m^\top G \Delta m + \left(\frac{1}{h} - \beta\overline{\mu}\right)\mathrm{KL}(q^+\|q). \tag{18}$$

Since $h \leq 1/(\beta\overline{\mu})$, the last term is nonnegative.

For the mean contribution to the first two terms in (18),

$$\left(1 + \frac{1}{h}\right)\frac{1}{2}\Delta m^\top (C^+)^{-1}\Delta m - \frac{1}{2}\Delta m^\top G \Delta m = \frac{1}{2h}\Delta m^\top C^{-1}\Delta m = \frac{h}{2} g^\top C g \geq \frac{\underline{\mu} h}{2}\|g\|^2.$$

For the covariance contribution, let $s_i$ be the eigenvalues of $C^{1/2}GC^{1/2}$, so that $\alpha\underline{\mu} \leq s_i \leq \beta\overline{\mu}$, and let $r_i := \frac{1+hs_i}{1+h}$, which are the eigenvalues of $(C^+)^{-1}C$. Since $\beta\overline{\mu} \geq 1$, the quantity $R := \frac{1+h\beta\overline{\mu}}{1+h}$ satisfies $R \geq 1$ and $r_i \leq R$. The function $\varphi(r) = r - \log r - 1$ has $\varphi''(r) = r^{-2}$, hence

$$r_i - \log r_i - 1 \geq \frac{(r_i - 1)^2}{2R^2}.$$

Consequently,

$$\frac{1}{2}\left[\mathrm{Tr}((C^+)^{-1}C) - \log\det((C^+)^{-1}C) - d\right] \geq \frac{h^2}{4(1+h\beta\overline{\mu})^2}\mathrm{Tr}(CHCH).$$

It follows that

$$\begin{aligned}\left(1 + \frac{1}{h}\right)\frac{1}{2}\left[\mathrm{Tr}((C^+)^{-1}C) - \log\det((C^+)^{-1}C) - d\right] &\geq \frac{h(1+h)}{4(1+h\beta\overline{\mu})^2}\mathrm{Tr}(CHCH) \\ &\geq \frac{h}{16}\mathrm{Tr}(CHCH) \geq \frac{\underline{\mu} h}{16}\mathrm{Tr}(HCH),\end{aligned}$$

where the second inequality uses $h\beta\overline{\mu} \leq 1$, and the last uses $C \succeq \underline{\mu} I$.

Combining the mean and covariance estimates in (18) gives

$$\mathcal{F}(q) - \mathcal{F}(q^+) \geq \frac{\underline{\mu} h}{16}\left(\|g\|^2 + \mathrm{Tr}(HCH)\right).$$

The Wasserstein PL inequality (14) therefore gives

$$\mathcal{F}(q) - \mathcal{F}(q^+) \geq \frac{\alpha\underline{\mu} h}{8}\left(\mathcal{F}(q) - \mathcal{F}(q_\star)\right),$$

which is equivalent to the claimed contraction. □

*Proof of Theorem 2.4.* Lemma A.2 and induction give

$$\underline{\mu} I \preceq C_n, C_{n+1/2} \preceq \overline{\mu} I, \qquad n \geq 0.$$

Since $h \leq 1/\beta$, the Wasserstein FB–GVI result (17) gives $\mathcal{F}(q_{n+1/2}) \leq \mathcal{F}(q_n)$. Since $h \leq 1/(\beta\overline{\mu})$, Lemma A.4 applied at $q_{n+1/2}$ gives

$$\mathcal{F}(q_{n+1}) - \mathcal{F}(q_\star) \leq \left(1 - \frac{\alpha\underline{\mu} h}{8}\right)\left(\mathcal{F}(q_{n+1/2}) - \mathcal{F}(q_\star)\right) \leq \left(1 - \frac{\alpha\underline{\mu} h}{8}\right)\left(\mathcal{F}(q_n) - \mathcal{F}(q_\star)\right).$$

Iterating over $n$ gives

$$\mathcal{F}(q_N) - \mathcal{F}(q_\star) \leq \left(1 - \frac{\alpha\underline{\mu} h}{8}\right)^N \left(\mathcal{F}(q_0) - \mathcal{F}(q_\star)\right),$$

and the exponential bound follows from $1 - x \leq e^{-x}$. □

### A.4 STATIONARITY FOR GENERAL SMOOTH TARGETS

Proposition 2.5 requires only Assumption 2.1. The continuous-time estimate follows directly from energy dissipation. For the discrete scheme, we impose the covariance box $\mathcal{C}$ and bound the objective decrease by the displacement of each substep. Summing these estimates gives the stated residual bound.

**Continuous-time stationarity.**

*Proof of the continuous-time statement in Proposition 2.5.* Setting $(\lambda_t^{\mathsf{W}}, \lambda_t^{\mathsf{FR}}) = (1, 1)$ in (12) and integrating over $[0, T]$ gives

$$\int_0^T \left[ \|\mathrm{grad}_{\mathsf{W}} \mathcal{F}(q_t)\|_{\mathsf{W}}^2 + \|\mathrm{grad}_{\mathsf{FR}} \mathcal{F}(q_t)\|_{\mathsf{FR}}^2 \right] \mathrm{d}t = \mathcal{F}(q_0) - \mathcal{F}(q_T) \le \mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathrm{G}}.$$

Dividing by $T$ and bounding the minimum by the average yields

$$\min_{0 \le t \le T} \left[ \|\mathrm{grad}_{\mathsf{W}} \mathcal{F}(q_t)\|_{\mathsf{W}}^2 + \|\mathrm{grad}_{\mathsf{FR}} \mathcal{F}(q_t)\|_{\mathsf{FR}}^2 \right] \le \frac{\mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathrm{G}}}{T}.$$

□

**Constrained proximal maps.**

**Lemma A.5** (Covariance-constrained Gaussian forward–backward maps)**.** *Let $q = \mathcal{N}(m, C)$ with $C \in \mathcal{C}$, and let $g \in \mathbb{R}^d$ and $H \in \mathbb{S}^d$. For the Wasserstein map, set*

$$m^{\mathsf{W}} := m + hg, \qquad M := I + h(H - C^{-1}), \qquad \widetilde{C} := MCM,$$

*and*

$$C^{\mathsf{W}} := \Pi_{\mathcal{C}} \left\{ \frac{1}{2} \left[ \widetilde{C} + 2hI + \left( \widetilde{C}(\widetilde{C} + 4hI) \right)^{1/2} \right] \right\}.$$

*For the Fisher–Rao map, let*

$$(1 + h)C^{-1} - hH = O \,\mathrm{diag}(p_1, \ldots, p_d) O^\top,$$

*and set*

$$m^{\mathsf{FR}} := m + hCg, \qquad C^{\mathsf{FR}} := O \,\mathrm{diag}(c_1, \ldots, c_d) O^\top,$$

*where $c_i := \mu_+$ if $p_i \le 0$, and otherwise $c_i := \min\{\mu_+, \max\{\mu_-, (1 + h)/p_i\}\}$.*

*Proof.* For the Wasserstein step, let $\widetilde{C} = O \,\mathrm{diag}(\widetilde{c}_1, \ldots, \widetilde{c}_d) O^\top$. By the spectral characterization of the Bures–Wasserstein distance (Bhatia et al., 2019), the constrained entropy proximal problem separates over the eigenvalues. For each $\widetilde{c}_i \ge 0$, the unconstrained minimizer is

$$c_i = \frac{1}{2} \left[ \widetilde{c}_i + 2h + \sqrt{\widetilde{c}_i(\widetilde{c}_i + 4h)} \right].$$

Restricting $c_i$ to $[\mu_-, \mu_+]$ therefore gives the stated spectral projection.

For the Fisher–Rao step, the mean and covariance variables separate in the Gaussian KL proximal problem. The mean first-order condition gives $m^{\mathsf{FR}} = m + hCg$. Up to constants independent of $C'$, the covariance problem is

$$\min_{C' \in \mathcal{C}} \left\{ \mathrm{Tr}\left[ ((1 + h)C^{-1} - hH)C' \right] - (1 + h) \log\det C' \right\}.$$

Diagonalizing $(1 + h)C^{-1} - hH$ reduces the problem to $\min_{\mu_- \le c \le \mu_+} \{pc - (1 + h) \log c\}$. Its minimizer is $(1 + h)/p$ clipped to $[\mu_-, \mu_+]$ when $p > 0$, and $\mu_+$ when $p \le 0$. □

**Fisher–Rao descent for general smooth targets.** The Wasserstein one-step estimate is given by Lemma A.1. For the Fisher–Rao map, we first establish a KL-relative-smoothness bound that does not require convexity of $V$.

**Lemma A.6** (Fisher–Rao relative smoothness)**.** *Under Assumption 2.1, let $q = \mathcal{N}(m, C)$ and $q' = \mathcal{N}(m', C')$ with $C, C' \in \mathcal{C}$. Then*

$$\mathbb{E}_{q'}[V] \le \mathbb{E}_q[V] + \langle \nabla_m \mathbb{E}_q[V], m' - m\rangle + \langle \nabla_C \mathbb{E}_q[V], C' - C\rangle + 2\beta\mu_+ \mathrm{KL}(q'\|q).$$

*Proof.* Let $T(\theta) = m' + B(\theta - m)$ be the optimal affine transport from $q$ to $q'$, where $B = B^\top \succ 0$ and $BCB = C'$. By Assumption 2.1,

$$\mathbb{E}_{q'}[V] \le \mathbb{E}_q[V] + \mathbb{E}_q\left[\nabla V(\theta)^\top (T(\theta) - \theta)\right] + \frac{\beta}{2} W_2^2(q', q).$$

Writing $D := B - I$ and $G := \mathbb{E}_q[\nabla^2 V]$, Gaussian integration by parts gives

$$\mathbb{E}_q\left[\nabla V(\theta)^\top (T(\theta) - \theta)\right] = \langle \nabla_m \mathbb{E}_q[V], m' - m\rangle + \mathrm{Tr}(GDC).$$

Since $\nabla_C \mathbb{E}_q[V] = G/2$ and $C' - C = DC + CD + DCD$,

$$\langle \nabla_C \mathbb{E}_q[V], C' - C\rangle = \mathrm{Tr}(GDC) + \frac{1}{2}\mathrm{Tr}(GDCD).$$

Therefore, using $\|G\|_{\mathrm{op}} \le \beta$,

$$\mathbb{E}_q\left[\nabla V(\theta)^\top (T(\theta) - \theta)\right] \le \langle \nabla_m \mathbb{E}_q[V], m' - m\rangle + \langle \nabla_C \mathbb{E}_q[V], C' - C\rangle + \frac{\beta}{2}\mathrm{Tr}(DCD).$$

Since $\mathrm{Tr}(DCD)$ is the covariance contribution to $W_2^2(q', q)$,

$$\mathbb{E}_{q'}[V] \le \mathbb{E}_q[V] + \langle \nabla_m \mathbb{E}_q[V], m' - m\rangle + \langle \nabla_C \mathbb{E}_q[V], C' - C\rangle + \beta W_2^2(q', q).$$

Finally, $C \preceq \mu_+ I$ and the Gaussian Talagrand inequality (Villani, 2009, Theorem 22.10) give

$$W_2^2(q', q) \le 2\mu_+ \mathrm{KL}(q'\|q),$$

which proves the claim. □

**Lemma A.7** (Fisher–Rao one-step descent)**.** *Under Assumption 2.1, let $q^{\mathsf{FR}} = \mathsf{FR}_h^{\mathcal{C}}(q; g, H)$ with $C \in \mathcal{C}$. Then*

$$\mathcal{F}(q^{\mathsf{FR}}) \le \mathcal{F}(q) - \left(\frac{1}{h} - 2\beta\mu_+\right)\mathrm{KL}(q^{\mathsf{FR}}\|q).$$

*In particular, if $h \le 1/(4\beta\mu_+)$, then*

$$\mathcal{F}(q^{\mathsf{FR}}) \le \mathcal{F}(q) - \frac{1}{2h}\mathrm{KL}(q^{\mathsf{FR}}\|q).$$

*Proof.* By optimality of the covariance-constrained version of the KL-proximal problem (15), using $q$ as a feasible comparison point,

$$\mathbb{E}_{q^{\mathsf{FR}}}[\log q^{\mathsf{FR}}] - \mathbb{E}_q[\log q] + \left\langle \nabla_m \mathbb{E}_q[V], m^{\mathsf{FR}} - m\right\rangle + \left\langle \nabla_C \mathbb{E}_q[V], C^{\mathsf{FR}} - C\right\rangle + \frac{1}{h}\mathrm{KL}(q^{\mathsf{FR}}\|q) \le 0.$$

Combining this inequality with Lemma A.6 yields

$$\mathcal{F}(q^{\mathsf{FR}}) \le \mathcal{F}(q) - \left(\frac{1}{h} - 2\beta\mu_+\right)\mathrm{KL}(q^{\mathsf{FR}}\|q).$$

If $h \le 1/(4\beta\mu_+)$, then $1/h - 2\beta\mu_+ \ge 1/(2h)$, proving the second claim. □

**Fixed points of the constrained maps.** We next identify the stationarity condition detected by both constrained forward–backward maps.

**Lemma A.8** (Fixed points of the constrained Gaussian maps)**.** *Under the discrete stepsize conditions of Proposition 2.5, either constrained map $\mathsf{W}_h^{\mathcal{C}}$ or $\mathsf{FR}_h^{\mathcal{C}}$ leaves $q = \mathcal{N}(m, C)$ unchanged if and only if*

$$g = 0, \qquad \mathrm{Tr}[H(D - C)] \le 0 \quad \textit{for every } D \in \mathcal{C}.$$

*These are exactly the first-order stationarity conditions for minimizing $\mathcal{F}(\mathcal{N}(m, C))$ over $m \in \mathbb{R}^d$ and $C \in \mathcal{C}$.*

*Proof.* Since $\nabla_m \mathcal{F}(q) = -g$ and $\nabla_C \mathcal{F}(q) = -H/2$, the displayed conditions are precisely the first-order variational inequality for the covariance-constrained Gaussian VI problem.

For the Fisher–Rao map, the mean update is fixed exactly when $g = 0$, while the first-order condition of the constrained KL-proximal problem in Lemma A.5 gives the covariance variational inequality.

For the Wasserstein map, the mean update is again fixed exactly when $g = 0$. Moreover, Assumption 2.1 and $h \le 1/(2\beta)$ imply

$$I + h(H - C^{-1}) \succeq \frac{1}{2} I,$$

so this matrix is the optimal transport map from $C$ to the forward covariance. Differentiating the constrained Bures entropy proximal objective at $C$ gives the same covariance variational inequality $\mathrm{Tr}[H(D - C)] \le 0$ for every $D \in \mathcal{C}$. □

**Discrete-time stationarity.**

*Proof of the discrete-time statement in Proposition 2.5.* Apply Lemma A.1 to the constrained Wasserstein half-step and Lemma A.7 to the constrained Fisher–Rao half-step. Under the stated stepsize restriction,

$$\mathcal{F}(q_n) - \mathcal{F}(q_{n+1}) \ge \frac{1}{4h} W_2^2(q_{n+1/2}, q_n) + \frac{1}{2h} \mathrm{KL}(q_{n+1} \| q_{n+1/2}) \ge \frac{h}{4} \mathcal{R}_n^{\mathsf{WFR}}.$$

Summing from $n = 0$ to $N - 1$ gives

$$\frac{h}{4} \sum_{n=0}^{N-1} \mathcal{R}_n^{\mathsf{WFR}} \le \mathcal{F}(q_0) - \mathcal{F}(q_N) \le \mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathcal{C}}.$$

Therefore

$$\min_{0 \le n < N} \mathcal{R}_n^{\mathsf{WFR}} \le \frac{1}{N} \sum_{n=0}^{N-1} \mathcal{R}_n^{\mathsf{WFR}} \le \frac{4\left(\mathcal{F}(q_0) - \mathcal{F}_{\inf}^{\mathcal{C}}\right)}{Nh}.$$

Finally, Lemma A.8 shows that $\mathcal{R}_n^{\mathsf{WFR}} = 0$ if and only if the iterate is first-order stationary for the covariance-constrained Gaussian VI problem. □

# B Proofs for the complementary roles of Wasserstein and Fisher–Rao

## B.1 Gaussian targets

We first record convergence bounds for arbitrary Gaussian initialization using the explicit solutions (8)–(9). The Gaussian dynamics and their convergence rates are studied in Chen et al. (2026, Lemma E.3, Theorem 5.6, and Appendix E.5); the estimates below make the dependence on the initial covariance explicit and allow $C_0$ and $C_\star$ to be noncommuting.

**Theorem B.1** (Convergence bounds for Gaussian targets)**.** *We write $\|\cdot\|$ for the Euclidean norm and $\|x\|_{C_\star^{-1}}^2 := x^\top C_\star^{-1} x$. Suppose the target is $\pi = \mathcal{N}(m_\star, C_\star)$, then for the Wasserstein flow (9) and the Fisher–Rao flow (8), the following bounds hold for every $t \ge 0$:*

1. ***Wasserstein.*** *The mean and covariance satisfy*

$$e^{-t}\left\|m_0 - m_\star\right\| \le \left\|m_t^{\mathsf{W}} - m_\star\right\| \le e^{-t/\kappa}\left\|m_0 - m_\star\right\|, \tag{19}$$

$$\left\|C_t^{\mathsf{W}} - C_\star\right\|_{\mathrm{op}} \le e^{-2t/\kappa}\left\|C_0 - C_\star\right\|_{\mathrm{op}}, \tag{20}$$

*and*

$$C_t^{\mathsf{W}} \succeq \mu_0(1 - e^{-2t})I. \tag{21}$$

*The inequalities in (19) also hold with $\|\cdot\|$ replaced by $\|\cdot\|_{C_\star^{-1}}$. In particular, $C^{\mathsf{W}}_{(\log 2)/2} \succeq \mu_0 I/2$ for every $C_0 \succ 0$.*

2. ***Fisher–Rao.*** *The mean and covariance satisfy*

$$\left\|m_t^{\mathsf{FR}} - m_\star\right\| \le e^{-t}\max\{\|C_0\|_{\mathrm{op}}, \|C_\star\|_{\mathrm{op}}\}\left\|C_0^{-1}(m_0 - m_\star)\right\|, \tag{22}$$

$$\left\|C_t^{\mathsf{FR}} - C_\star\right\|_{\mathrm{op}} \le e^{-t}\max\{\|C_0\|_{\mathrm{op}}, \|C_\star\|_{\mathrm{op}}\}\left\|C_0^{-1}\right\|_{\mathrm{op}}\left\|C_0 - C_\star\right\|_{\mathrm{op}}. \tag{23}$$

*Proof.* For Wasserstein, $\kappa^{-1}I \preceq \mu_0 C_\star^{-1} \preceq I$, so the eigenvalues of $e^{-\mu_0 t C_\star^{-1}}$ lie in $[e^{-t}, e^{-t/\kappa}]$. Applying these spectral bounds to the mean formula in (9) proves (19). Since $e^{-\mu_0 t C_\star^{-1}}$ commutes with $C_\star^{-1/2}$, the same argument gives the bounds in the $C_\star^{-1}$ norm. The covariance estimate (20) follows from (9) and submultiplicativity of the operator norm.

To prove the covariance floor, rewrite the Wasserstein covariance solution as

$$C_t^{\mathsf{W}} = e^{-\mu_0 t C_\star^{-1}} C_0 e^{-\mu_0 t C_\star^{-1}} + 2\mu_0 \int_0^t e^{-2\mu_0 s C_\star^{-1}}\,\mathrm{d}s.$$

The first term is positive semidefinite. Moreover, $\mu_0 C_\star^{-1} \preceq I$ implies $e^{-2\mu_0 s C_\star^{-1}} \succeq e^{-2s} I$ by the spectral theorem. Integrating gives (21).

For Fisher–Rao, (8) expresses $(C_t^{\mathsf{FR}})^{-1}$ as a convex combination of $C_0^{-1}$ and $C_\star^{-1}$. Both matrices are bounded below by $\max\{\|C_0\|_{\mathrm{op}}, \|C_\star\|_{\mathrm{op}}\}^{-1} I$, and hence

$$\left\|C_t^{\mathsf{FR}}\right\|_{\mathrm{op}} \le \max\{\|C_0\|_{\mathrm{op}}, \|C_\star\|_{\mathrm{op}}\}.$$

Combining this bound with the mean formula in (8) proves (22). For the covariance, the same inverse-covariance formula yields

$$C_t^{\mathsf{FR}} - C_\star = C_t^{\mathsf{FR}}\big(C_\star^{-1} - (C_t^{\mathsf{FR}})^{-1}\big)C_\star = e^{-t} C_t^{\mathsf{FR}} C_0^{-1}(C_0 - C_\star).$$

Taking operator norms proves (23). These identities preserve the order of matrix multiplication and require no commutativity assumption. □

The Wasserstein covariance lower bound in (21) is independent of $C_0$. The Fisher–Rao bounds decay as $e^{-t}$, but their prefactors can diverge as $C_0$ approaches singularity. Theorem 3.1 compares the two flows under the initializations $C_0 = \varepsilon C_\star$ and $C_0 = C_\star$, which isolate covariance recovery and mean refinement, respectively.

*Proof of Theorem 3.1.* For covariance recovery, (21) gives $C^{\mathsf{W}}_{(\log 2)/2} \succeq \mu_0 I/2$, and therefore $T_{\mathsf{W}}^{\mathrm{init}} \le (\log 2)/2$. For Fisher–Rao, substituting $C_0 = \varepsilon C_\star$ into (8) gives

$$C_t^{\mathsf{FR}} = \frac{C_\star}{1 + (\varepsilon^{-1} - 1)e^{-t}}, \qquad \mu_{\min}(C_t^{\mathsf{FR}}) = \frac{\mu_0}{1 + (\varepsilon^{-1} - 1)e^{-t}}.$$

Thus $C_t^{\mathsf{FR}} \succeq \mu_0 I/2$ if and only if $(\varepsilon^{-1} - 1)e^{-t} \le 1$, which gives $T_{\mathsf{FR}}^{\mathrm{init}} = \log((1-\varepsilon)/\varepsilon)$.

For refinement, $C_0 = C_\star$ implies $C_t^{\mathsf{W}} = C_t^{\mathsf{FR}} = C_\star$. Squaring the $C_\star^{-1}$-norm version of (19) and using the definition of $T_{\mathsf{W}}^{\mathrm{ref}}(\delta)$ gives

$$\frac{1}{2}\log\frac{1}{\delta} \le T_{\mathsf{W}}^{\mathrm{ref}}(\delta) \le \frac{\kappa}{2}\log\frac{1}{\delta}.$$

The upper bound is attained when $m_0 - m_\star$ lies in the $\mu_{\max}(C_\star)$ eigenspace. For Fisher–Rao, (8) reduces to $m_t^{\mathsf{FR}} - m_\star = e^{-t}(m_0 - m_\star)$, so the squared $C_\star^{-1}$ norm decreases by the factor $e^{-2t}$ and $T_{\mathsf{FR}}^{\mathrm{ref}}(\delta) = \frac{1}{2}\log(1/\delta)$. □

*Proof of Corollary 3.2.* The Wasserstein phase gives a covariance lower bound at the switching time that is independent of $C_0$. Set $t_{\mathsf{W}} := (\log 2)/2$ and $D := C_\star^{-1/2} C_{t_{\mathsf{W}}} C_\star^{-1/2}$. Since $0 \prec C_0 \preceq C_\star$, (9) gives

$$D \preceq I, \qquad D \succeq I - \exp(-2\mu_0 t_{\mathsf{W}} C_\star^{-1}) \succeq \left(1 - 2^{-1/\kappa}\right) I \succeq \frac{1}{2\kappa} I.$$

Moreover, the Wasserstein mean solution in (9) is contractive in the $C_\star^{-1}$ norm, so

$$\left\| C_\star^{-1/2}(m_{t_{\mathsf{W}}} - m_\star) \right\| \le \left\| C_\star^{-1/2}(m_0 - m_\star) \right\|.$$

Now let $u \ge 0$ denote the Fisher–Rao time after the switch. By (8),

$$S_u := C_\star^{-1/2} C_{t_{\mathsf{W}}+u} C_\star^{-1/2} = \left[ I + e^{-u}(D^{-1} - I) \right]^{-1},$$

and, writing $y_u := C_\star^{-1/2}(m_{t_{\mathsf{W}}+u} - m_\star)$,

$$y_u = e^{-u} S_u D^{-1} y_0, \qquad y_0 := C_\star^{-1/2}(m_{t_{\mathsf{W}}} - m_\star).$$

Since $D^{-1} \preceq 2\kappa I$ and $S_u \preceq I$, taking $u = \log(2\kappa) + t$ gives

$$\|y_u\| \le e^{-t} \|y_0\| \le e^{-t} \left\| C_\star^{-1/2}(m_0 - m_\star) \right\|.$$

It remains to control the covariance contribution to the KL divergence. If $r \in [1/(2\kappa), 1]$ is an eigenvalue of $D$, the corresponding eigenvalue of $S_u$ is $(1+v)^{-1}$ with

$$v := e^{-u}(r^{-1} - 1) \le e^{-t}.$$

Using

$$\log(1+v) - \frac{v}{1+v} \le \frac{v^2}{2}, \qquad v \ge 0,$$

we obtain

$$\frac{1}{2}\left[ \mathrm{Tr}(S_u) - \log\det S_u - d \right] \le \frac{d}{4} e^{-2t}.$$

Combining this with the mean term in the Gaussian KL formula yields

$$\mathrm{KL}\left( q_{\frac{\log 2}{2} + \log(2\kappa) + t} \,\middle\|\, \pi \right) \le e^{-2t} \left( \frac{1}{2} \left\| C_\star^{-1/2}(m_0 - m_\star) \right\|^2 + \frac{d}{4} \right).$$

□

### B.2 Discrete covariance recovery beyond Gaussian targets

The covariance-recovery comparison also holds for forward–backward updates on smooth log-concave targets. The proof uses the Fisher–Rao covariance recursion and the lower bound imposed by the Wasserstein entropy proximal step.

**Proposition B.2** (Discrete covariance recovery for log-concave targets)**.** *Suppose $0 \preceq \nabla^2 V(\theta) \preceq \beta I$ for all $\theta \in \mathbb{R}^d$, and let $c_0 := \mu_{\min}(C_0) < 1/(2\beta)$.*

1. *Iterating only the Fisher–Rao forward–backward step with fixed stepsize $h > 0$, let $N_{\mathsf{FR}}$ be the first iteration for which $C_n \succeq (2\beta)^{-1} I$. Then*

$$N_{\mathsf{FR}} = \Theta\left( \frac{1}{\log(1+h)} \log \frac{1}{\beta c_0} \right) \qquad \text{as } \beta c_0 \downarrow 0.$$

2. *A single Wasserstein forward–backward step of size $h$ satisfies $C_1 \succeq hI$ for every $C_0 \succ 0$. In particular, $h = 1/(2\beta)$ gives $C_1 \succeq (2\beta)^{-1} I$ in one step.*

*Proof of Proposition B.2.* For pure Fisher–Rao, the covariance update (16) can be written as

$$C_{n+1} = (1+h)\left(C_n^{-1} + h\mathbb{E}_{q_n}[\nabla^2 V]\right)^{-1}.$$

Since $\mathbb{E}_{q_n}[\nabla^2 V] \succeq 0$, we have $C_{n+1} \preceq (1+h)C_n$, and hence $C_n \preceq (1+h)^n C_0$. Therefore, $C_n \succeq (2\beta)^{-1} I$ implies

$$\frac{1}{2\beta} \le \mu_{\min}(C_n) \le (1+h)^n c_0,$$

so

$$N_{\mathsf{FR}} \ge \left\lceil \frac{\log(1/(2\beta c_0))}{\log(1+h)} \right\rceil.$$

For the upper bound, define $c_{n+1} := \frac{(1+h)c_n}{1+h\beta c_n}$. We claim that $C_n \succeq c_n I$ for all $n$. Indeed, if $C_n \succeq c_n I$, then $\mathbb{E}_{q_n}[\nabla^2 V] \preceq \beta I$ gives

$$C_{n+1} \succeq \frac{(1+h)c_n}{1+h\beta c_n} I = c_{n+1} I.$$

The scalar recursion has the explicit solution

$$\frac{1}{\beta c_n} = 1 + \left(\frac{1}{\beta c_0} - 1\right)(1+h)^{-n}.$$

Hence $c_n \ge 1/(2\beta)$ whenever

$$n \ge \frac{\log\left((\beta c_0)^{-1} - 1\right)}{\log(1+h)},$$

and therefore

$$N_{\mathsf{FR}} \le \left\lceil \frac{\log\left((\beta c_0)^{-1} - 1\right)}{\log(1+h)} \right\rceil.$$

Combining the two bounds yields the stated $\Theta$ complexity.

For a Wasserstein forward–backward step, let $\widetilde{C} \succeq 0$ denote the covariance after the forward part. The entropy proximal step gives

$$C_1 = \frac{1}{2}\left[\widetilde{C} + 2hI + \left(\widetilde{C}(\widetilde{C} + 4hI)\right)^{1/2}\right].$$

The scalar map $x \mapsto \frac{1}{2}(x + 2h + \sqrt{x(x+4h)})$ is increasing on $[0,\infty)$ and equals $h$ at $x = 0$, so $C_1 \succeq hI$. Taking $h = 1/(2\beta)$ gives $C_1 \succeq \frac{1}{2\beta} I$. □

# C Theory for switched Gaussian-mixture updates

The proof of Theorem 4.3 combines component and weight descent through the upper bound $\mathcal{Q}_n$. We estimate its decrease by changing the weights with the components fixed, then changing the components with the new weights fixed:

$$\mathcal{F}(q_{n+1}) \le \mathcal{Q}_n(q_{n+1}) \le \mathcal{Q}_n(w_{n+1}, \{q_{n,k}\}) \le \mathcal{Q}_n(q_n) = \mathcal{F}(q_n).$$

The component decreases are therefore weighted by $w_{n+1,k}$, as in the residual $\mathcal{R}_n$.

## C.1 Separable and frozen component objectives

*Proof of Proposition 4.1.* For a candidate mixture $q = \sum_{k=1}^K w_k q_k$, define $P(k,\theta) := w_k q_k(\theta)$ and $R_n(k,\theta) := r_{n,k}(\theta)\pi(\theta)$. Their $\theta$-marginals are $q$ and $\pi$, while their conditional distributions over the component label are $r(\cdot \mid \theta)$ and $r_n(\cdot \mid \theta)$. The relative-entropy chain rule gives

$$\mathrm{KL}(P\|R_n) = \mathrm{KL}(q\|\pi) + \mathbb{E}_q\left[\mathrm{KL}\left(r(\cdot \mid \theta)\|r_n(\cdot \mid \theta)\right)\right].$$

Direct expansion also gives

$$\mathrm{KL}(P\|R_n) = \sum_{k=1}^{K} w_k \log w_k + \sum_{k=1}^{K} w_k \mathbb{E}_{q_k}\left[\log q_k + V - \log r_{n,k}\right] + \log Z = \mathcal{Q}_n(q).$$

This proves the upper bound. At $q = q_n$, the two conditional label distributions coincide, so the gap vanishes. Since the gap is nonnegative and differentiable with a minimum of zero at the current parameters, its first derivative also vanishes there. □

For the componentwise analysis, define $U_{n,k} := V - \log r_{n,k}$ and the corresponding Gaussian objective $\mathcal{J}_{n,k}(q) := \mathbb{E}_q[\log q + U_{n,k}]$. Up to terms involving only the mixture weights, $\mathcal{Q}_n$ is the weighted sum of these component objectives. At $q_{n,k}$, using $\mathbb{E}_{q_{n,k}}[\nabla \log q_{n,k}] = 0$ and $\nabla^2 \log q_{n,k} = -C_{n,k}^{-1}$,

$$g_{n,k} = -\mathbb{E}_{q_{n,k}}[\nabla U_{n,k}], \qquad H_{n,k} = C_{n,k}^{-1} - \mathbb{E}_{q_{n,k}}[\nabla^2 U_{n,k}]. \tag{24}$$

The component updates use the expected Hessian of the frozen potential, $\mathbb{E}_{q_{n,k}}[\nabla^2 U_{n,k}] = C_{n,k}^{-1} - H_{n,k}$, whereas the switching rule in Section 4.3 uses the expected target Hessian $\mathbb{E}_{q_{n,k}}[\nabla^2 V]$. These matrices satisfy

$$C_{n,k}^{-1} - H_{n,k} = \mathbb{E}_{q_{n,k}}[\nabla^2 V] + \left(C_{n,k}^{-1} + \mathbb{E}_{q_{n,k}}[\nabla^2 \log q_n]\right).$$

For $K = 1$, the bracketed term vanishes because $\nabla^2 \log q_n = -C_{n,k}^{-1}$, so the two expected Hessians coincide. For $K > 1$, the additional term accounts for interactions between mixture components; its growth is controlled in Appendix C.2.

### C.2 Uniform smoothness of the frozen component objectives

The mixture log density satisfies

$$\nabla^2 \log q_n(\theta) = -\sum_{j=1}^{K} r_{n,j}(\theta) C_{n,j}^{-1} + \mathrm{Cov}_{r_n(\cdot|\theta)}\left(-C_{n,j}^{-1}(\theta - m_{n,j})\right).$$

Under the bounded-mean and covariance assumptions, the component log densities grow at most linearly in $\theta$. Their conditional covariance therefore grows at most quadratically. Gaussian moment bounds then give uniform smoothness estimates for the expected frozen potentials.

**Lemma C.1** (Uniform component smoothness)**.** *Under Assumptions 2.1 and 4.2, with $C_{n,k} \in \mathcal{C}$, there exists a finite constant $L \geq \beta + \mu_-^{-1}$, depending only on $\beta$, $\mu_-$, $\mu_+$, $d$, and $R$, such that uniformly over $n$ and $k$,*

$$\mathbb{E}\left[(u + Dz)^\top \nabla^2 U_{n,k}(m + Sz)(u + Dz)\right] \leq L\left(\|u\|^2 + \|D\|_{\mathrm{F}}^2\right)$$

*for $z \sim \mathcal{N}(0, I)$, $u \in \mathbb{R}^d$, $D \in \mathbb{R}^{d\times d}$, and parameters satisfying $\|m\| \leq R$ and $\|S\|_{\mathrm{op}} \leq \sqrt{\mu_+}$. Moreover, for $q = q_{n,k} = \mathcal{N}(m, C)$ and $q' = \mathcal{N}(m', C')$ with $\|m'\| \leq R$ and $C' \in \mathcal{C}$,*

$$\mathbb{E}_{q'}[U_{n,k}] \leq \mathbb{E}_q[U_{n,k}] - g_{n,k}^\top(m' - m) + \frac{1}{2}\mathrm{Tr}\left[(C^{-1} - H_{n,k})(C' - C)\right] + 2L\mu_+ \mathrm{KL}(q'\|q).$$

*The stepsize bounds in Theorem 4.3 may be chosen as*

$$\bar{\eta} := \frac{1}{2L}, \qquad \bar{\tau} := \frac{1}{4L\mu_+}.$$

*Proof.* Let $R < \infty$ be a uniform bound on the component means. Since $C_{n,j} \in \mathcal{C}$,

$$C_{n,j}^{-1} \preceq \mu_-^{-1} I, \qquad C_{n,j}^{-1} \succeq \mu_+^{-1} I.$$

Using Assumptions 2.1 and 4.2 together with the preceding expression for $\nabla^2 \log q_n$, there exist finite constants $a_0, a_2 > 0$, depending only on $\beta$, $R$, $\mu_-$, and $\mu_+$, such that

$$\nabla^2 U_{n,k}(\theta) \preceq \left(a_0 + a_2 \|\theta\|^2\right) I$$

uniformly over $n$ and $k$. Indeed, the only unbounded term is the conditional covariance of the component scores, and

$$\left\|C_{n,j}^{-1}(\theta - m_{n,j})\right\|^2 \le \mu_-^{-2}(\|\theta\| + R)^2.$$

For $\theta = m + Sz$ with $\|m\| \le R$ and $\|S\|_{\mathrm{op}} \le \sqrt{\mu_+}$, Gaussian second and fourth moments therefore give a finite $L$ such that

$$\mathbb{E}\left[(u + Dz)^\top \nabla^2 U_{n,k}(m + Sz)(u + Dz)\right] \le L\left(\|u\|^2 + \|D\|_{\mathrm{F}}^2\right).$$

The covariance of the conditional log density is positive semidefinite, so the Hessian formula also gives $\nabla^2 U_{n,k}(\theta) \succeq -(\beta + \mu_-^{-1})I$. Enlarging $L$ if necessary ensures $L \ge \beta + \mu_-^{-1}$.

For the second claim, let $q = q_{n,k}$ and choose factors $C = SS^\top$ and $C' = S'S'^\top$ satisfying $\|S' - S\|_{\mathrm{F}}^2 = W_2^2(\mathcal{N}(0, C'), \mathcal{N}(0, C))$. Set $D := S' - S$. Along the segment joining $(m, S)$ and $(m', S')$, the mean norm is at most $R$ and the factor norm is at most $\sqrt{\mu_+}$. Taylor's theorem and the first bound therefore give a remainder of at most $LW_2^2(q', q)/2$ in factor coordinates. The identity $C' - C = SD^\top + DS^\top + DD^\top$ and the lower bound $\mathbb{E}_q[\nabla^2 U_{n,k}] \succeq -LI$ add at most $L\|D\|_{\mathrm{F}}^2/2$ when expressing the linear term in covariance coordinates. Hence

$$\mathbb{E}_{q'}[U_{n,k}] \le \mathbb{E}_q[U_{n,k}] - g_{n,k}^\top(m' - m) + \frac{1}{2}\operatorname{Tr}\left[(C^{-1} - H_{n,k})(C' - C)\right] + LW_2^2(q', q).$$

Since $C \preceq \mu_+ I$, the Gaussian Talagrand inequality (Villani, 2009, Theorem 22.10) gives $W_2^2(q', q) \le 2\mu_+ \mathrm{KL}(q'\|q)$, which proves the result. □

### C.3 One-step component and weight descent

The component descent estimates follow by applying the Gaussian proximal arguments to each frozen objective $\mathcal{J}_{n,k}$, using the smoothness bounds above. The weight estimate is obtained with the components fixed.

**Lemma C.2** (Component descent)**.** *Under the conditions of Theorem 4.3, a selected Wasserstein step satisfies*

$$\mathcal{J}_{n,k}(q_{n+1,k}) \le \mathcal{J}_{n,k}(q_{n,k}) - \frac{W_2^2(q_{n+1,k}, q_{n,k})}{4\eta_{n,k}},$$

*whereas a selected Fisher–Rao step satisfies*

$$\mathcal{J}_{n,k}(q_{n+1,k}) \le \mathcal{J}_{n,k}(q_{n,k}) - \frac{\mathrm{KL}(q_{n+1,k}\|q_{n,k})}{2\tau_{n,k}}.$$

*Proof.* For the Wasserstein step, let $C_{n,k} = SS^\top$, choose the forward factor

$$\widetilde{S} := \left[I + \eta_{n,k}(H_{n,k} - C_{n,k}^{-1})\right] S,$$

and let $S^+$ be the factor corresponding to the constrained Wasserstein entropy proximal step in Lemma A.5. Define

$$E^2 := \|m_{n+1,k} - m_{n,k}\|^2 + \left\|S^+ - S\right\|_{\mathrm{F}}^2.$$

The Gaussian coupling induced by the two factors gives $E^2 \ge W_2^2(q_{n+1,k}, q_{n,k})$. Optimality of the entropy proximal step, compared with the old component, yields

$$\mathbb{E}_{q_{n+1,k}}[\log q_{n+1,k}] \le \mathbb{E}_{q_{n,k}}[\log q_{n,k}] - \frac{E^2}{2\eta_{n,k}} + g_{n,k}^\top(m_{n+1,k} - m_{n,k}) + \operatorname{Tr}\left[(C_{n,k}^{-1} - H_{n,k})S(S - S^+)^\top\right].$$

The first part of Lemma C.1 gives

$$\mathbb{E}_{q_{n+1,k}}[U_{n,k}] \le \mathbb{E}_{q_{n,k}}[U_{n,k}] - g_{n,k}^\top(m_{n+1,k} - m_{n,k}) - \operatorname{Tr}\left[(C_{n,k}^{-1} - H_{n,k})S(S - S^+)^\top\right] + \frac{L}{2}E^2.$$

Adding the two inequalities and using $\eta_{n,k} \le \bar{\eta} = (2L)^{-1}$ gives a decrease of at least $E^2/(4\eta_{n,k})$, hence the Wasserstein claim.

For Fisher–Rao, optimality of the constrained KL-proximal problem defining $\mathsf{FR}^{\mathcal{C}}_{\tau_{n,k}}$ in Lemma A.5 gives

$$\mathbb{E}_{q_{n+1,k}}[\log q_{n+1,k}] - \mathbb{E}_{q_{n,k}}[\log q_{n,k}] - g_{n,k}^{\top}(m_{n+1,k} - m_{n,k}) + \frac{1}{2}\operatorname{Tr}\left[(C_{n,k}^{-1} - H_{n,k})(C_{n+1,k} - C_{n,k})\right] + \frac{\mathrm{KL}(q_{n+1,k}\|q_{n,k})}{\tau_{n,k}} \le 0.$$

Applying the second part of Lemma C.1 and adding yields

$$\mathcal{J}_{n,k}(q_{n+1,k}) \le \mathcal{J}_{n,k}(q_{n,k}) - \left(\frac{1}{\tau_{n,k}} - 2L\mu_+\right)\mathrm{KL}(q_{n+1,k}\|q_{n,k}).$$

The restriction $\tau_{n,k} \le \bar{\tau} = (4L\mu_+)^{-1}$ proves the claim. □

**Lemma C.3** (Weight descent)**.** *With the components fixed at their current values, the weight update in Algorithm 1 satisfies*

$$\mathcal{Q}_n(w_{n+1}, \{q_{n,k}\}) = \mathcal{Q}_n(w_n, \{q_{n,k}\}) - \frac{\mathrm{KL}(w_n\|w_{n+1})}{\nu_n} - \left(\frac{1}{\nu_n} - 1\right)\mathrm{KL}(w_{n+1}\|w_n).$$

*Hence* $0 < \nu_n \le 1$ *gives a decrease of at least* $\mathrm{KL}(w_n\|w_{n+1})/\nu_n$.

*Proof.* With the components fixed, the weight-dependent part of $\mathcal{Q}_n$ is negative entropy plus an affine function. The first-order condition of the entropic mirror step gives

$$\log\frac{w_{n+1,k}}{w_{n,k}} = \nu_n \ell_{n,k} + c_n,$$

where $c_n$ is independent of $k$ and enforces normalization. Taking the inner product with $w_{n+1} - w_n$ and using the Bregman identity for negative entropy gives the stated equality. □

### C.4 Fixed points and stationarity

The descent estimates control the update displacements. The next lemma identifies their common zeros with the first-order stationarity conditions for the covariance-constrained mixture problem.

**Lemma C.4** (Fixed points of the component updates)**.** *Let* $q_k = \mathcal{N}(m_k, C_k)$ *with* $C_k \in \mathcal{C}$. *Under the stepsize conditions of Theorem 4.3, either selected component map leaves* $q_k$ *unchanged if and only if*

$$g_k = 0, \qquad \operatorname{Tr}\left[H_k(D - C_k)\right] \le 0 \quad \textit{for every } D \in \mathcal{C}.$$

*If* $C_k$ *lies in the interior of* $\mathcal{C}$*, the second condition reduces to* $H_k = 0$. *The weight map is fixed if and only if* $\ell_1 = \cdots = \ell_K$.

*Proof.* Both mean maps are fixed exactly when $g_k = 0$. For the Wasserstein covariance map, write

$$M := I + \eta(H - C^{-1}), \qquad \widetilde{C} := MCM.$$

Lemma C.1 gives $C^{-1} - H \preceq LI$, so $\eta \le \bar{\eta} = (2L)^{-1}$ implies $M \succeq \frac{1}{2}I \succ 0$. By Lemma A.5, the updated covariance is the unique minimizer over $D \in \mathcal{C}$ of

$$-\frac{1}{2}\log\det D + \frac{1}{2\eta}W_2^2\left(\mathcal{N}(0, D), \mathcal{N}(0, \widetilde{C})\right).$$

The optimal transport map from $C$ to $MCM$ is $M$, so the derivative of the covariance Bures term at $D = C$ is $I - M$. Consequently, the derivative of the preceding objective at $C$ is $-H/2$. Since $\mathcal{C}$ is convex, $C$ is fixed exactly when $\operatorname{Tr}[H(D - C)] \le 0$ for every $D \in \mathcal{C}$.

For Fisher–Rao, the constrained covariance problem underlying Lemma A.5 is, up to a positive constant,

$$\operatorname{Tr}\left[\left((1+\tau)C^{-1} - \tau H\right)D\right] - (1+\tau)\log\det D.$$

Its derivative at $D = C$ is again proportional to $-H$, giving the same variational inequality. If $C$ is interior, both positive and negative covariance perturbations are feasible, so $H = 0$.

Finally, the weight update is fixed exactly when all exponential factors differ only by the common normalization, which is equivalent to $\ell_1 = \cdots = \ell_K$. □

*Proof of Theorem 4.3.* Apply Lemma C.3 with the components fixed at their current values, and then keep the new weights fixed while applying Lemma C.2 to each selected component. Proposition 4.1 gives $\mathcal{F}(q_{n+1}) \le \mathcal{Q}_n(q_{n+1})$ and $\mathcal{Q}_n(q_n) = \mathcal{F}(q_n)$, so

$$\mathcal{F}(q_n) - \mathcal{F}(q_{n+1}) \ge \sum_{k \in \mathcal{I}_n^{\mathsf{W}}} \frac{w_{n+1,k}}{4\eta_{n,k}} W_2^2(q_{n+1,k}, q_{n,k}) + \sum_{k \in \mathcal{I}_n^{\mathsf{FR}}} \frac{w_{n+1,k}}{2\tau_{n,k}} \mathrm{KL}(q_{n+1,k} \| q_{n,k}) + \frac{\mathrm{KL}(w_n \| w_{n+1})}{\nu_n}.$$

This proves the one-step descent statement. The estimate holds for either geometry in each component, so the geometry choices may depend on the current iterate.

Let $h_n := \min\big(\{\nu_n\} \cup \{\eta_{n,k} : k \in \mathcal{I}_n^{\mathsf{W}}\} \cup \{\tau_{n,k} : k \in \mathcal{I}_n^{\mathsf{FR}}\}\big)$ be the smallest active stepsize. Since $h_n \le \eta_{n,k}$ for $k \in \mathcal{I}_n^{\mathsf{W}}$, $h_n \le \tau_{n,k}$ for $k \in \mathcal{I}_n^{\mathsf{FR}}$, and $h_n \le \nu_n$, the definitions of $h_n$ and $\mathcal{R}_n$ give

$$\mathcal{F}(q_n) - \mathcal{F}(q_{n+1}) \ge \frac{h_n}{4} \mathcal{R}_n.$$

Summing from $n = 0$ to $N - 1$ gives

$$\sum_{n=0}^{N-1} h_n \mathcal{R}_n \le 4\left(\mathcal{F}(q_0) - \mathcal{F}(q_N)\right) \le 4\left(\mathcal{F}(q_0) - \mathcal{F}_{\inf}\right).$$

Therefore

$$\min_{0 \le n < N} \mathcal{R}_n \le \frac{4\left(\mathcal{F}(q_0) - \mathcal{F}_{\inf}\right)}{\sum_{n=0}^{N-1} h_n}.$$

Since $h_n \ge h$ for every $n$, this gives the bound stated in Theorem 4.3. The fixed-point characterization in Lemma C.4 gives the stationarity interpretation of $\mathcal{R}_n = 0$. □

# D Experiment setup

This appendix describes the baselines we compare with and the evaluation metrics.

## D.1 Baselines

We compare against three Gaussian-mixture VI methods, and against one sampler as an external reference. The three variational methods share the number of components $K$, the initial mixture, and the iteration budget on each target; where a method carries its own batch size, since its estimator consumes samples differently.

**IBW.** Isotropic Bures–Wasserstein (Petit-Talamon et al., 2025) restricts every component to an isotropic covariance $C_k = \varepsilon_k I$ and holds the weights uniform. IBW follows the Wasserstein gradient flows, and the variance update has the closed form $\varepsilon \leftarrow (1 - \text{step})^2 \varepsilon$ with a step proportional to the learning rate, the only parameter we tune. Its Monte Carlo batch size $B_G$ is set separately from $J$, since the isotropic parameterization estimates far fewer quantities per component.

**GMBBVI.** Gaussian mixture black-box VI (Che et al., 2026) is a derivative-free algorithm, which approximates the gradient and Hessian of the target through Stein's lemma (Stein, 1981). GMBBVI uses full covariances and free weights, and each iteration estimates a block-diagonal natural gradient from $J$ whitened samples per component, with a mean direction $b_k$ and a covariance direction $E_k$. The mean and weights take a forward-Euler step while the covariance takes an exponential-integrator step,

$$C_k \leftarrow L_k \exp\left(-\Delta t\, E_k\right) L_k^\top, \qquad L_k L_k^\top = C_k,$$

which keeps every covariance positive definite for any stepsize. The stepsize itself is adaptive,

$$\Delta t = \min\left\{\Delta t_{\max}\, \eta(n),\ \frac{\beta}{\max_k \|E_k\|}\right\},$$

where $\beta$ is a constant and $\eta(n)$ is a cosine-annealing schedule. We use the authors' default $\beta = 0.9$ except where Table 6 states otherwise, and set $\Delta t_{\max}$ per target.

**SEMTRUX.** SEMTRUX (Arenz et al., 2023) belongs to a modular family of natural-gradient Gaussian-mixture methods, built on Arenz et al. (2020) and Lin et al. (2019). Its seven-letter name is the codeword for seven independent design choices:

- **S** the natural gradient is estimated by Stein's lemma, a first-order estimator using the target score;
- **E** the number of components is fixed;
- **M** samples are drawn from the individual co**m**ponents, with no reuse across iterations;
- **T** each component is updated inside a KL **t**rust region;
- **R** the component stepsize adapts by an improvement criterion;
- **U** the weights take a direct natural-gradient **u**pdate;
- **X** the weight stepsize is fixed at 1.

For SEMTRUX, we set only $J$ and the component count, and we disable component adaptation so that all three variational methods hold the same $K$ throughout.

**WALNUTS.** Within-Orbit Adaptive Leapfrog No-U-Turn Sampler (Bou-Rabee et al., 2026, WALNUTS) generalizes the No-U-Turn Sampler (Hoffman & Gelman, 2014, NUTS) by adapting the leapfrog stepsize within an orbit instead of once per chain: at fixed intervals of simulated time it takes the largest stepsize from a dyadic schedule that keeps the energy error below a threshold, and it selects among states along the orbit by the same biased progressive rule as NUTS. To compare it with the variational methods, we score it by the maximum mean discrepancy and the entropic $W_2$ distance and place it on the same cost axis by counting the evaluated target's value–gradient pairs.

### D.2 Evaluation metrics

**Reverse KL and the ELBO.** For the target potential $V$ with normalizing constant $Z$, we compute the reverse Kullback–Leibler (KL) divergence and the (negative) evidence lower bound (Blei et al., 2017, ELBO) via

$$\mathrm{KL}(q\|\pi) = \mathbb{E}_q\left[\log q + V\right] + \log Z, \qquad -\mathrm{ELBO}(q) = \mathbb{E}_q\left[\log q + V\right] = \mathrm{KL}(q\|\pi) - \log Z.$$

Thus, $-\mathrm{ELBO}$ can be evaluated without knowing $Z$ and differs from the reverse KL only by the target-dependent constant $-\log Z$. We report $-\mathrm{ELBO}$ throughout; when $Z$ is known, the corresponding reverse KL can be recovered by adding $\log Z$.

**Total variation.** On the two-dimensional targets the reference density can be tabulated and we report the following total variation (TV) distance

$$\mathrm{TV}(q,\pi) = \tfrac{1}{2}\sum_{\theta\in\mathcal{G}} \left|\bar{q}(\theta) - \bar{\pi}(\theta)\right|,$$

where $\mathcal{G}$ is a regular grid over the plotting box and $\bar{q}, \bar{\pi}$ are the two densities on $\mathcal{G}$ normalized to unit mass.

**Entropic mode coverage.** For multimodal targets where the modes are enumerable, we quantify how $q$ spreads its mass over them using the entropic mode coverage (EMC) of Blessing et al. (2024). We partition $\mathbb{R}^d$ into descriptors $\xi_1, \ldots, \xi_M$ of the $M$ target modes and assign each draw to its nearest mode under the Mahalanobis metric defined by that mode's covariance. Let $p_i$ denote the fraction of $N$ draws from $q$ assigned to mode $i$. Following Blessing et al. (2024), we compute

$$\mathrm{EMC} := -\sum_{i=1}^{M} p_i \log_M p_i \in [0,1].$$

"Modes covered" counts the target modes with at least one component mean within Mahalanobis radius 3 of the mode center. It answers whether any component reached the mode.

**Maximum mean discrepancy.** Where the target comes with an exact sampler we also compare $n$ draws from $q$ with $n$ exact draws from $\pi$. The maximum mean discrepancy (Gretton et al., 2012) is the distance between the mean embeddings of the two laws in the reproducing kernel Hilbert space (RKHS) of $k(\theta, \theta') = \exp(-\|\theta - \theta'\|^2 / 2\sigma^2)$, which we estimate by the biased V-statistic

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2}\sum_{i,j} k(\theta_i, \theta_j) + \frac{1}{n^2}\sum_{i,j} k(\theta'_i, \theta'_j) - \frac{2}{n^2}\sum_{i,j} k(\theta_i, \theta'_j),$$

with $\theta_i \sim q$ and $\theta'_j \sim \pi$, and report its square root. The bandwidth $\sigma$ is the median of the pairwise distances among the target samples.

**Entropic $W_2$ distance.** The entropic $W_2$ is the square root of the entropy-regularized transport cost between the same two clouds,

$$W^2_{2,\varepsilon}(q, \pi) = \min_{P \in \Pi(a,b)} \left\{ \langle P, M \rangle + \varepsilon \,\mathrm{KL}\left(P \| a \otimes b\right) \right\}, \qquad M_{ij} = \left\| \theta_i - \theta'_j \right\|^2,$$

minimized over couplings of the two uniform clouds and computed by Sinkhorn iterations (Cuturi, 2013). The regularization is set relative to the target scale, $\varepsilon = \rho\, s_\pi$ with $s_\pi$ the median pairwise squared distance among target samples, and $\rho$ is chosen once per target as the smallest value at which the iterations converge for every method and seed. Our implementation builds on the OTT-JAX library (`https://github.com/ott-jax/ott`).

## E Practical schedule in WFR algorithm

This section describes how the curvature scales are estimated, and then presents the two curvature-based schedules: the switching rule that selects between Wasserstein and Fisher–Rao for each component, and the adaptive stepsizes in each branch.

**Curvature scales.** Recall from Section 4.3 the curvature scales $z_i(C_{n,k})$, the eigenvalues of

$$C_{n,k}\, \mathbb{E}_{q_{n,k}}\left[\nabla^2 V\right].$$

Since a Gaussian's covariance is the inverse of its negative log-density Hessian, $z_i$ is the ratio of the target curvature to the component's own curvature along each principal direction. Therefore $z_i$ measures local target curvature relative to the scale of the component.

To avoid the expensive computation of the Hessian, we estimate $\mathbb{E}_{q_{n,k}}[\nabla^2 V]$ by Stein's identity (Stein, 1981) from the same batch $\theta_{k,1}, \ldots, \theta_{k,J} \sim q_{n,k}$ that supplies $g_{n,k}$. Writing $s_j := \nabla V(\theta_{k,j})$ and $\bar{s}$ for their mean,

$$\mathbb{E}_{q_{n,k}}\left[\nabla^2 V\right] \approx \mathrm{Sym}\left( \frac{1}{J-1}\sum_{j=1}^{J} (s_j - \bar{s}) \left[C_{n,k}^{-1}(\theta_{k,j} - m_{n,k})\right]^\top \right),$$

so no Hessian of the target is required.

**Switching rule.** The motivation of the switching rule comes from the analysis of Fisher–Rao gradient flows. Consider the current mean $m_0$, a second-order expansion makes the target locally Gaussian, $\mathcal{N}(m_\star, \Lambda^{-1})$ with $\Lambda := \nabla^2 V(m_0) \succ 0$. The Fisher–Rao flow of a component started from $C_0 = \sigma_0 I$ is,

$$\dot{\sigma}_i = \sigma_i(1 - h_i\sigma_i), \qquad \dot{\tilde{m}}_i = -h_i\sigma_i\, \tilde{m}_i,$$

where $h_i$ are the eigenvalues of $\Lambda$ and $\tilde{m}_i(t)$ are the coordinates of $m_t - m_\star$ in this eigenbasis. With $z_i := \sigma_0 h_i$, the mean in direction $i$ solves to $\tilde{m}_i(t) = \tilde{m}_i(0)/\big(1 + z_i(e^t - 1)\big)$, which barely moves until $t \approx \log(1/z_i)$.

We therefore define $T = \max_i \log(1/z_i)$, which measures the time the slowest component needs to recover its covariance. By (9), Wasserstein is not trapped by the small covariance. We therefore use Wasserstein while $T$ is large and switch to Fisher–Rao once Fisher–Rao is more efficient. Now we explain how to construct the candidate covariance $\widehat{C}^{\mathsf{W}}_{n,k}$ and $\widehat{C}^{\mathsf{FR}}_{n,k}$, and how we choose the branch through comparison.

We can compute the candidate covariance $\widehat{C}^{\mathsf{W}}_{n,k}$ and $\widehat{C}^{\mathsf{FR}}_{n,k}$ following the corresponding gradient flows. However, to avoid the extra evaluation of the target, we freeze the curvature at its current estimate $\widehat{\Lambda}_{n,k}$. We then compute the remaining covariance recovery time $T$ for each candidate covariance, and choose the one with a smaller $T$.

The component turns to Fisher–Rao if the covariance is already recovered or Fisher–Rao reduces $T$ more than Wasserstein, i.e., $T(C_{n,k}) \le 0$ or $T(\widehat{C}^{\mathsf{FR}}_{n,k}) \le T(\widehat{C}^{\mathsf{W}}_{n,k})$.

Since the analysis is only valid and $T = \max_i \log(1/z_i)$ is only meaningful for target's positive eigenvalues, we only apply the switching rule while $\widehat{\Lambda}_{n,k}$ has no clearly negative direction. Otherwise, the component remains Wasserstein, since it is more stable for negative curvature.

**Wasserstein stepsize.** A component in the Wasserstein branch takes stepsize $\eta$,

$$\eta_{n,k} = \min\left\{\eta_{\max},\ \frac{1}{2\rho(\widehat{\Lambda}_{n,k})}\right\}, \qquad \rho(A) := \max_i |\mu_i(A)|,$$

where $\eta_{\max}$ is the upper bound for the Wasserstein stepsize, and $\mu$ is the eigenvalue.

The adaptive schedule for Wasserstein follows Proposition B.2 in Appendix B.2. In practice, we replace the global smoothness constant $\beta$ by the largest local curvature $\rho(\widehat{\Lambda}_{n,k})$ of the current component, so that we could adapt to the local curvature and accelerate the algorithm. The ceiling $\eta_{\max}$ binds when the local curvature is small.

**Fisher–Rao stepsize.** A component in the Fisher–Rao branch takes stepsize $\tau$,

$$E_{n,k} := -C_{n,k}H_{n,k}, \qquad \tau_{n,k} = \min\left\{\tau_{\max},\ \frac{1}{2\rho(E_{n,k})}\right\},$$

where $E_{n,k}$ measures the residual of the Bregman proximal update and vanishes at its fixed point. The ceiling $\tau_{\max}$ guards the Fisher–Rao stepsize.

## F MoG40 UNDER DIFFERENT INITIALIZATIONS

In this section, we test how sensitive the four methods (GMBBVI, SEMTRUX, IBW and WFR) are to different initializations, and present an ablation study for WFR.

Here we consider the MoG40 problem, and fix the mixture size $K = 40$, the number of iterations (300) and the batch size $J = 64$, and only the initialization changes. The component means are drawn as $m_i \sim \mathcal{N}(\bar{m}, s^2 I)$ with covariances $C_i = cI$, and the four methods share one initial mixture per setting. The 40 target modes are unit-scale and spread over the grid $[-40, 40]^2$. We use three initial settings:

| $\bar{m}$ | $s$ | $c$ | Initial mixture |
|---|---|---|---|
| 0 | 40 | $10^{-5}$ | Means cover the mode grid, but the initial covariance is small and ill-conditioned. This is the setting of Section 5. |
| 0 | 40 | 1 | Means cover the grid and the initial covariance is well-scaled. This is the easiest start. |
| $(30, 30)$ | 5 | $10^{-5}$ | The initial mixtures are concentrated in a corner of the grid, with ill-conditioned covariance. |

Every method is tuned separately for each setting. For each method and setting, we tune the parameters on ten seeds, and report the results from five new seeds. Table 7 lists the selected settings, and Appendix K gives the final values.

**Results.** Table 1 and Figure 6 show the outcome. WFR gives the best fit in all three settings in total variation, −ELBO, mode count and EMC, and its fit changes little from one setting to another. The result of the easier start setting shows that a better initial covariance narrows the gap between WFR and other methods.

Table 1: MoG40 under the three initializations after 300 iterations: $K = 40$, $J = 64$, mean $\pm$ 1 s.d. over 5 seeds.

| Initialization | Method | TV ↓ | −ELBO ↓ | Modes ↑ | EMC ↑ |
|---|---|---|---|---|---|
| $\bar{m} = 0$, $s = 40$, $c = 10^{-5}$ | GMBBVI | $0.3933 \pm 0.0224$ | $0.5106 \pm 0.0378$ | $25 \pm 1.1$ | $0.8723 \pm 0.0124$ |
| | SEMTRUX | $0.3815 \pm 0.0310$ | $0.5098 \pm 0.0679$ | $25 \pm 1.5$ | $0.8804 \pm 0.0150$ |
| | IBW | $0.5710 \pm 0.0139$ | $2.9018 \pm 0.8074$ | $21 \pm 0.7$ | $0.8470 \pm 0.0161$ |
| | WFR | $\mathbf{0.2147 \pm 0.0439}$ | $\mathbf{0.2277 \pm 0.0578}$ | $\mathbf{34 \pm 1.5}$ | $\mathbf{0.9538 \pm 0.0119}$ |
| $\bar{m} = 0$, $s = 40$, $c = 1$ | GMBBVI | $0.3751 \pm 0.0255$ | $0.4822 \pm 0.0470$ | $26 \pm 0.8$ | $0.8826 \pm 0.0102$ |
| | SEMTRUX | $0.3830 \pm 0.0281$ | $0.5741 \pm 0.0827$ | $26 \pm 1.9$ | $0.8758 \pm 0.0125$ |
| | IBW | $0.4466 \pm 0.0135$ | $0.7521 \pm 0.0360$ | $24 \pm 0.9$ | $0.8027 \pm 0.0088$ |
| | WFR | $\mathbf{0.3348 \pm 0.0145}$ | $\mathbf{0.4069 \pm 0.0249}$ | $\mathbf{27 \pm 0.8}$ | $\mathbf{0.8949 \pm 0.0084}$ |
| $\bar{m} = (30, 30)$, $s = 5$, $c = 10^{-5}$ | GMBBVI | $0.7954 \pm 0.0318$ | $1.6092 \pm 0.1346$ | $8 \pm 1.3$ | $0.5670 \pm 0.0408$ |
| | SEMTRUX | $0.8050 \pm 0.0112$ | $1.6368 \pm 0.0609$ | $8 \pm 0.4$ | $0.5565 \pm 0.0162$ |
| | IBW | $0.8500 \pm 0.0009$ | $2.8008 \pm 0.1604$ | $7 \pm 0.0$ | $0.4343 \pm 0.0025$ |
| | WFR | $\mathbf{0.2809 \pm 0.0479}$ | $\mathbf{0.3379 \pm 0.0852}$ | $\mathbf{30 \pm 1.5}$ | $\mathbf{0.9256 \pm 0.0156}$ |

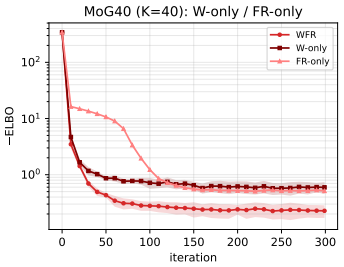



Figure 5: WFR Ablation Study

The corner start is hard for every baseline. For GMBBVI and SEMTRUX, the Fisher–Rao mean update is preconditioned by the component covariance, so from $C_i = 10^{-5} I$ the means barely move until the covariance has recovered (Theorem 3.1). IBW does not bound the covariance, so a large learning rate can expand it by many orders of magnitude in one step; tuning therefore forces a small learning rate (Table 7), and the means again move slowly. WFR recovers the covariance with Wasserstein steps, with the upper bound $\mu_+$ bounding the expansion of covariance. After the covariance recovery phase, the components shift to Fisher–Rao steps, and further refine, with a faster convergence speed due to the affine invariance of the Fisher–Rao geometry. This makes WFR robust to this challenging initialization.

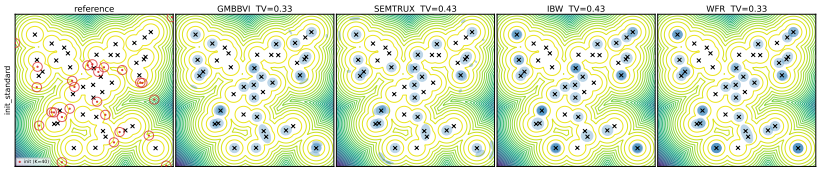
**Ablation study.** We isolate the choice of two geometries on the first initialization ($\bar{m} = 0$, $s = 40$, $c = 10^{-5}$). W-only and FR-only run the Wasserstein and Fisher–Rao flows throughout, respectively. Figure 5 separates the mechanisms cleanly. FR-only is held at the small initial covariance and barely descends at covariance recovery phase, while W-only leaves it at once but stalls well above WFR. This is consistent with the complementary regimes for Wasserstein and Fisher–Rao as studied in Section 3 and Appendix B, and hence switching between the two geometries makes Gaussian mixture VI faster and better.

## G THE NICE DIGITS TARGET

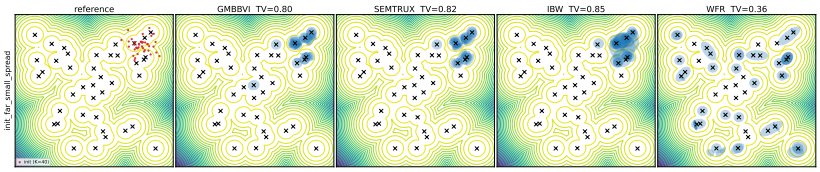
We give more details about the NICE digits target, which follows the benchmark of Blessing et al. (2024), and provide a UMAP projection to visualize the target and the sampled Gaussian mixtures.

A NICE flow (Dinh et al., 2014) maps a simple base density through an invertible network. The target considered here is the log-density of such a flow, pretrained on $14 \times 14$ MNIST digits. All methods use $K = 20$ components and run for 1000 iterations from the initialization $m_i \sim \mathcal{N}(0, I)$, $C_i = 10^{-4} I$, with batch size $J = 128$ for the full-covariance methods and $J = 64$ for IBW.

The UMAP projection (McInnes et al., 2018) in Figure 7 reveals that the initialization $m_i \sim \mathcal{N}(0, I)$ is far from the typical set of the NICE target. Combined with the small initial covariance, this puts the components far from any high-probability region, and they have to cover a long distance before they can start fitting the target. The generated digit samples in Figure 3 show that WFR captures most digit modes and produces recognizable samples, because the Wasserstein phase moves components toward high-probability regions quickly, and the Fisher–Rao phase refines the approximation once they arrive. GMBBVI and IBW fail to capture the main mode of the target. SEMTRUX captures the




(a) $\bar{m} = 0$, $s = 40$, $c = 1$.




(b) $\bar{m} = (30, 30)$, $s = 5$, $c = 10^{-5}$.

Figure 6: Final MoG40 mixtures after 300 iterations, with every method tuned per setting. Contours show the target, black crosses its 40 mode centres, red dots the shared initial means and blue ellipses the fitted components; each panel title gives the total variation to the target.

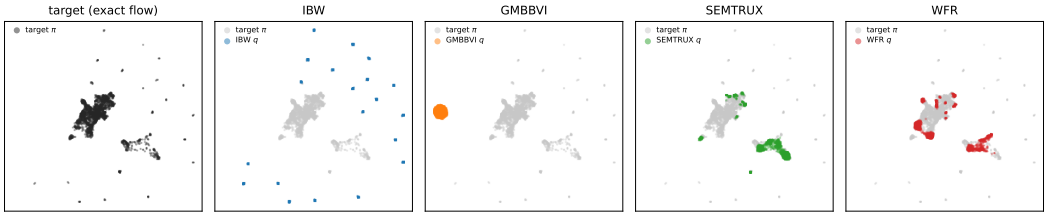



Figure 7: NICE digits: UMAP projection of target samples (grey) and mixture samples (colour), for each method's final mixture.

target but exhibits stronger mode-seeking behavior and converges more slowly than WFR under the ill-conditioned initialization, as the convergence curves in Figure 2b show.

## H THE DARCY-FLOW INVERSE PROBLEM

We describe the Darcy-flow benchmark used in our experiments, following Che et al. (2025); Garbuno-Inigo et al. (2020). The unknown is a log-permeability field parameterized by $\theta \in \mathbb{R}^{64}$. The observation operator is symmetrized so that the posterior is bimodal by construction.

**Unknown and prior.** The unknown is the log-permeability field on the unit square $D = (0, 1)^2$, parameterized by a truncated Karhunen–Loève expansion

$$\log \kappa(x; \theta) = \sum_l \theta_l \sqrt{\lambda_l}\, \varphi_l(x), \qquad \lambda_l = \left(\pi^2 |l|^2 + \tau^2\right)^{-2},$$

where $\{\varphi_l\}$ are the cosine eigenfunctions of the Laplacian. We sort the modes by increasing $|l|^2$ and keep the first $N_\theta = 64$, giving $\theta \in \mathbb{R}^{64}$. The prior is

$$\theta \sim \mathcal{N}(0, \sigma_0^2 I), \qquad \sigma_0 = \sqrt{5},$$

with $\tau = 3$. The truth field $\theta^\dagger$ is drawn at the wider scale $\sigma_{\text{tr}} = 5$ and held fixed, together with the data it generates.

**Forward model and symmetrised observations.** Given $\theta$, the pressure $h$ solves

$$-\nabla \cdot (\kappa(\cdot\,; \theta) \nabla h) = f \quad \text{in } D, \qquad h = 0 \quad \text{on } \partial D,$$

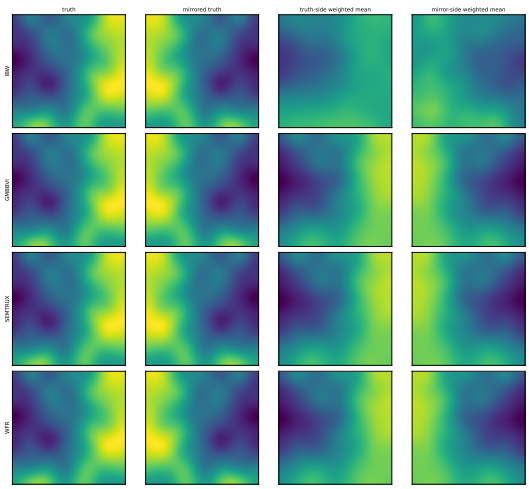



Figure 8: Darcy fields, one row per method. The first two columns show the true log-permeability field and its mirror image; the last two show the fitted mixture's mean conditional on each side of the symmetry.

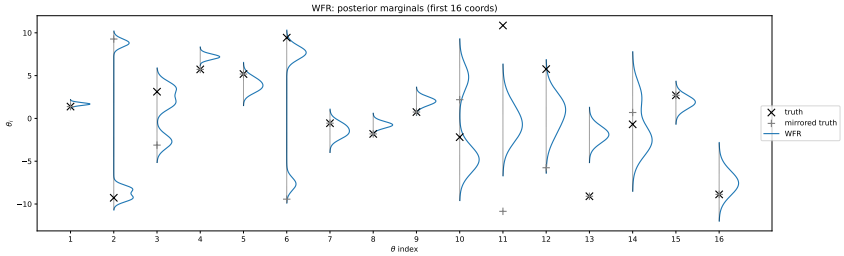



Figure 9: Darcy: marginals of the WFR mixture in the first 16 coordinates, against the truth field and its mirror image.

where the source $f$ depends only on the vertical coordinate, as in Che et al. (2025).

The pressure is observed on a $15 \times 15$ sensor lattice, with each measurement averaged with its reflection about $x_1 = 1/2$: $G_{ij}(\theta) = \frac{1}{2}[h(p_i, q_j; \theta) + h(1 - p_i, q_j; \theta)]$. The data are $y = G(\theta^\dagger) + \eta$ with the noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$ and $\sigma_\eta = 0.25$, giving the posterior

$$\pi(\theta \mid y) \propto \exp\left(-\Phi_R(\theta)\right), \qquad \Phi_R(\theta) = \frac{\|y - G(\theta)\|^2}{2\sigma_\eta^2} + \frac{\|\theta\|^2}{2\sigma_0^2}.$$

Because the symmetrised observation operator cannot distinguish a field from its left–right reflection, every mode of the posterior comes with a mirror twin, making the posterior at least bimodal. Figure 8 reports the mixture mean conditional on each side of the symmetry.

**Results.** All four methods fit $K = 5$ components for 1000 iterations with initial covariances $C_i = 10^{-4} I$. From the negative ELBO (Figure 2c), we observe that WFR converges fastest, SEMTRUX and GMBBVI take longer to converge to the same level, while IBW, using isotropic mixtures, gives a worse approximation.

Figure 8 shows the recovered fields. Conditioned on either side of the symmetry, the WFR mixture reproduces the truth field and its mirror image alike, as do SEMTRUX and GMBBVI. Figure 9 shows the marginals of the first 16 coordinates of WFR. For most of the marginals, WFR successfully captures the bimodal structure and WFR therefore can capture both modes of this posterior and recovers the field.

# I TWO-DIMENSIONAL POSTERIORS: THE CIRCLE AND THE ROSENBROCK RIDGE

Here we consider two two-dimensional posteriors. Both are least-squares targets $\pi(\theta) \propto \exp(-\frac{1}{2}\|F(\theta)\|^2)$: the first is shaped like a circle, the second defined by the Rosenbrock function, whose posterior concentrates along a strongly curved valley. Both fit $K = 40$ components with $J = 8$ samples per component from means $m_i \sim \mathcal{N}(0, s^2 I)$ and covariances $C_i = 10^{-5} I$.

**The circle.** Taking $F(\theta) = (1 - \|\theta\|^2)/0.3$ puts the mass on the unit circle, a one-dimensional manifold of modes. We set $s = 2$ so that most components begin away from the ring, and run for 500 iterations. Table 2 shows that WFR achieves the lowest TV and negative ELBO, and Figures 10 and 11 show that WFR converges quickly and successfully spreads its components along the ring.

**The Rosenbrock.** Taking $F(\theta) = \big(-10(\theta_2 - \theta_1^2),\, 1 - \theta_1\big)/\sqrt{10}$ gives a single mode on a strongly curved and ill-conditioned ridge. We set $s = 1$ and run for 600 iterations. With strongly curved geometry, the Rosenbrock target naturally favors samplers based on Fisher–Rao flows. Table 2 shows WFR is the best or second best for every metric, and successfully captures the full geometry of the target, shown in Figure 10b.

**Comparison with WALNUTS.** We also compare against WALNUTS (Bou-Rabee et al., 2026), a Hamiltonian Monte Carlo method that adapts the leapfrog stepsize along each orbit (Appendix D.1). Since IBW, SEMTRUX, WFR and WALNUTS all use only the target's value and gradient, we compare them on a common axis of value–gradient evaluations. Figure 12 traces the MMD and entropic $W_2$ along this axis. We notice that, though WALNUTS is given a larger budget, it does not overtake WFR on either target.

Table 2: The circle (500 iterations) and the Rosenbrock ridge (600 iterations). The best value is marked in bold and the second best is underlined.

| Target | Method | TV ↓ | −ELBO ↓ | MMD ↓ | entropic $W_2$ ↓ |
|---|---|---|---|---|---|
| Circle | GMBBVI | $0.1001 \pm 0.0271$ | $-0.8043 \pm 0.0384$ | $0.0221 \pm 0.0100$ | $\mathbf{0.2896 \pm 0.0033}$ |
| | SEMTRUX | $0.0442 \pm 0.0063$ | $\underline{-0.8524 \pm 0.0024}$ | $\mathbf{0.0108 \pm 0.0054}$ | $0.2937 \pm 0.0048$ |
| | IBW | $\underline{0.0317 \pm 0.0037}$ | $-0.8461 \pm 0.0124$ | $0.0174 \pm 0.0043$ | $\underline{0.2901 \pm 0.0023}$ |
| | WFR | $\mathbf{0.0251 \pm 0.0079}$ | $\mathbf{-0.8575 \pm 0.0029}$ | $\underline{0.0119 \pm 0.0012}$ | $0.2924 \pm 0.0024$ |
| Ridge | GMBBVI | $0.0847 \pm 0.0089$ | $-1.6136 \pm 0.0259$ | $0.1397 \pm 0.0160$ | $10.56 \pm 0.57$ |
| | SEMTRUX | $\mathbf{0.0404 \pm 0.0087}$ | $\mathbf{-1.7578 \pm 0.0151}$ | $\underline{0.0449 \pm 0.0070}$ | $\underline{6.05 \pm 0.55}$ |
| | IBW | $0.4792 \pm 0.0071$ | $-0.8901 \pm 0.0139$ | $0.4425 \pm 0.0025$ | $17.20 \pm 0.05$ |
| | WFR | $\underline{0.0509 \pm 0.0123}$ | $\underline{-1.7420 \pm 0.0467}$ | $\mathbf{0.0355 \pm 0.0095}$ | $\mathbf{4.70 \pm 0.85}$ |

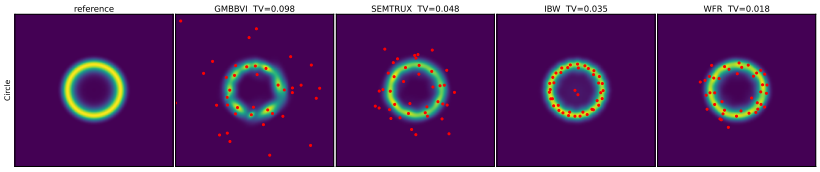
# J BAYESIAN LOGISTIC REGRESSION: THE BREAST-CANCER POSTERIOR

In Bayesian logistic regression, each observation is a feature vector $x_j$ with a binary label $y_j \in \{0, 1\}$, modelled as $y_j \mid \theta \sim \text{Bernoulli}\big(\sigma(\theta^\top x_j)\big)$ with $\sigma(t) = 1/(1 + e^{-t})$. Under a Gaussian prior, the posterior

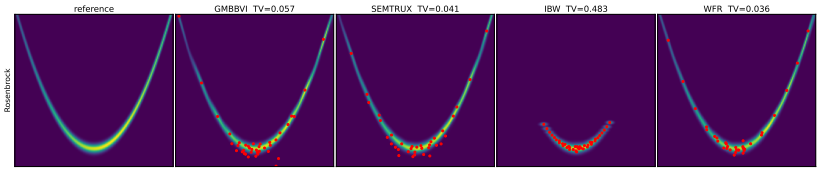
$$\pi(\theta) \propto \mathcal{N}(\theta; 0, 10^2 I) \prod_j \sigma(\theta^\top x_j)^{y_j} \big(1 - \sigma(\theta^\top x_j)\big)^{1 - y_j}$$

is smooth and log-concave. We use the Wisconsin Diagnostic Breast Cancer dataset (Wolberg et al., 1993), with 30 features scaled by their training standard deviations plus an intercept, so $d = 31$. We fit $K = 3$ components for 200 iterations with $J = 124$ samples per component, from means $m_i \sim \mathcal{N}(0, 4I)$ and covariances $C_i = 10^{-5} I$.

Table 3 and Figure 13 show the results. WFR converges fastest of the four methods, in both the −ELBO and the final accuracy. On this Bayesian posterior, WFR thus combines fast convergence with an accurate approximation of the posterior.

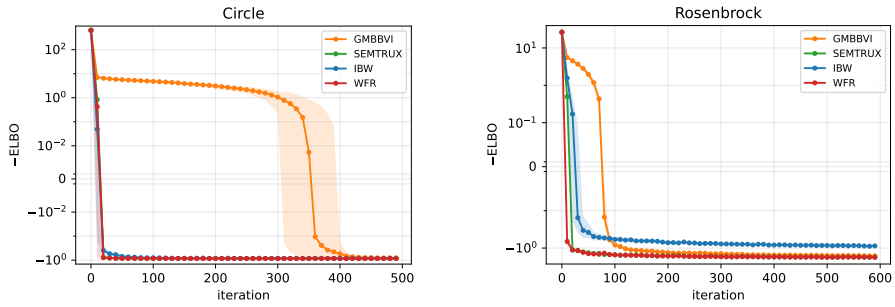



(a) The circle, $F(\theta) = (1 - \|\theta\|^2)/0.3$.




(b) The Rosenbrock ridge, $F(\theta) = \big(-10(\theta_2 - \theta_1^2),\ 1 - \theta_1\big)/\sqrt{10}$.

Figure 10: Final mixtures on the two-dimensional posteriors. Red dots mark the component means and each panel title carries the total variation to the reference.




(a) The circle, over 5 seeds.

(b) The Rosenbrock ridge, over 5 seeds.

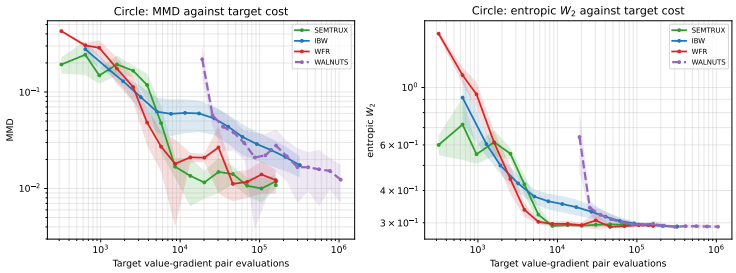
Figure 11: $-$ELBO against iteration on the two-dimensional posteriors, $K = 40$ components.

Table 3: Breast-cancer posterior over 5 seeds. Test accuracy is the posterior-predictive accuracy on the held-out half. The best value is marked in bold and the second best is underlined.

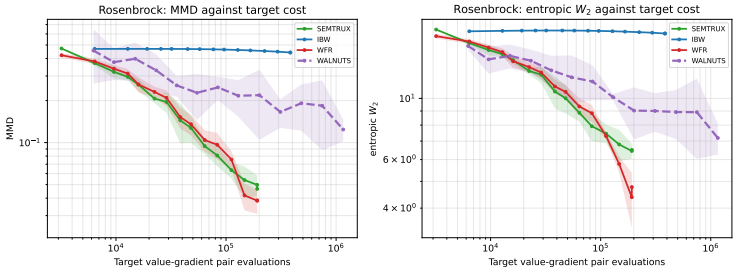
| Method | $-$ELBO ↓ | Test accuracy ↑ |
|---|---|---|
| GMBBVI | $519.65 \pm 20.10$ | $0.7930 \pm 0.0046$ |
| IBW | $208.71 \pm 4.30$ | $0.9430 \pm 0.0084$ |
| SEMTRUX | $\underline{51.31 \pm 0.15}$ | $\mathbf{0.9472 \pm 0.0000}$ |
| WFR | $\mathbf{50.93 \pm 0.03}$ | $\mathbf{0.9472 \pm 0.0000}$ |

## K Configurations and hyperparameters

We present the configuration and hyperparameters for reproducibility. Tables 4–6 give the target and initialization, the WFR hyperparameters, and the baseline hyperparameters. Throughout, each component mean is drawn as $m_i \sim \mathcal{N}(\bar{m}, s^2 I)$ with covariance $C_i = cI$. The initial mixture is drawn once per condition from a fixed seed and shared across all four methods. Here $N$ is the iteration count, $J$ the Monte Carlo batch size per component, and $[\mu_-, \mu_+]$ the spectral box $\mathcal{C}$ of Section 2.2 that constrains the component covariances.

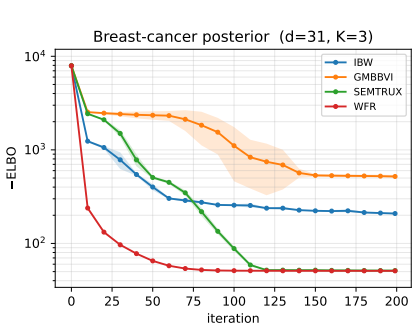



(a) The circle, over 5 seeds.

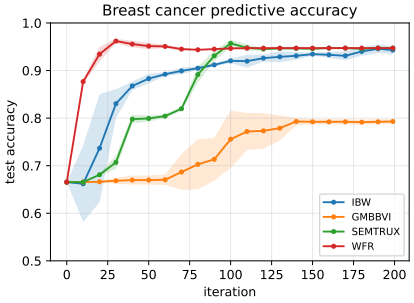



(b) The Rosenbrock ridge, over 5 seeds.

Figure 12: Maximum mean discrepancy and entropic $W_2$ between draws from the fitted mixture and exact draws from the target, against the number of target value–gradient pairs spent.




(a) $-$ELBO against iteration.




(b) Held-out predictive accuracy against iteration.

Figure 13: Breast-cancer posterior, $d = 31$, $K = 3$, over 5 seeds.

Table 4: Targets, iteration number and initializations. The three MoG40 rows are the conditions compared in Appendix F.

| Target | $d$ | $K$ | $N$ | $\bar{m}$ / $s$ | $c$ |
|---|---|---|---|---|---|
| MoG40 | 2 | 40 | 300 | 0 / 40 | $10^{-5}$ |
| | 2 | 40 | 300 | 0 / 40 | 1 |
| | 2 | 40 | 300 | $(30, 30)$ / 5 | $10^{-5}$ |
| Circle | 2 | 40 | 500 | 0 / 2 | $10^{-5}$ |
| Rosenbrock | 2 | 40 | 600 | 0 / 1 | $10^{-5}$ |
| Breast cancer | 31 | 3 | 200 | 0 / 2 | $10^{-5}$ |
| NICE digits | 196 | 20 | 1000 | 0 / 1 | $10^{-4}$ |
| Darcy | 64 | 5 | 1000 | 0 / $\sqrt{5}$ | $10^{-4}$ |

Table 5: WFR hyperparameters. $\tau$ and $\eta_w$ are stepsizes for Fisher–Rao steps and Wasserstein steps respectively. $n_{\mathrm{W}}^{\max}$ is the ceiling for Wasserstein steps. $[\mu_-, \mu_+]$ is the spectral box of covariance.

| Target | Step rule | $J$ | $\tau$ | $\eta_w$ | $n_{\mathrm{W}}^{\max}$ | $[\mu_-, \mu_+]$ |
|---|---|---|---|---|---|---|
| MoG40, $\bar{m} = 0$, $c = 10^{-5}$ | fixed | 64 | 0.15 | 0.25 | 25 | $[10^{-5}, 2 \times 10^{1}]$ |
| MoG40, $\bar{m} = 0$, $c = 1$ | fixed | 64 | 0.10 | 1.0 | 25 | $[10^{-5}, 2 \times 10^{1}]$ |
| MoG40, $\bar{m} = (30, 30)$, $c = 10^{-5}$ | fixed | 64 | 0.15 | 1.0 | 25 | $[10^{-5}, 2 \times 10^{1}]$ |
| Circle | adaptive | 8 | 0.25 | $2 \times 10^{-3}$ | 10 | $[10^{-5}, 10^{2}]$ |
| Rosenbrock | adaptive | 8 | 1.25 | $2 \times 10^{-3}$ | 10 | $[10^{-5}, 10^{2}]$ |
| Breast cancer | adaptive | 124 | 0.2 | $10^{-3}$ | 10 | $[10^{-5}, 1.5 \times 10^{2}]$ |
| NICE digits | adaptive | 128 | 0.3 | 0.25 | 25 | $[10^{-4}, 10^{3}]$ |
| Darcy | adaptive | 512 | 0.1 | 0.25 | 10 | $[10^{-4}, 10^{4}]$ |

Table 6: Baseline hyperparameters. For SEMTRUX, $\varepsilon$ is the per-component KL trust-region bound, set as both its initial and its maximum value. The GMBBVI damping $\beta$ is given in parentheses where it departs from its default of 0.9. $B_G$ is the batch size of the IBW.

| Target | GMBBVI<br>$J$ / $\Delta t_{\max}$ | SEMTRUX<br>$J$ / $\varepsilon$ | IBW<br>$B_G$ / learning rate |
|---|---|---|---|
| MoG40, $\bar{m} = 0$, $c = 10^{-5}$ | 64 / 1 ($\beta = 1.5$) | 64 / 1 | 64 / $2 \times 10^{-2}$ |
| MoG40, $\bar{m} = 0$, $c = 1$ | 64 / 0.5 ($\beta = 1.5$) | 64 / 2 | 64 / $5 \times 10^{-1}$ |
| MoG40, $\bar{m} = (30, 30)$, $c = 10^{-5}$ | 64 / 8 ($\beta = 1.5$) | 64 / 16 | 64 / $5 \times 10^{-3}$ |
| Circle | 8 / 0.5 ($\beta = 0.8$) | 8 / 1 | 16 / $2 \times 10^{-3}$ |
| Rosenbrock | 8 / 0.5 | 8 / 1 | 16 / $3 \times 10^{-3}$ |
| Breast cancer | 124 / 0.5 | 124 / 1 | 10 / $5 \times 10^{-4}$ |
| NICE digits | 128 / 0.5 | 128 / 1 | 64 / $2 \times 10^{-3}$ |
| Darcy | 512 / 0.5 | 512 / 0.1 | 64 / $10^{-4}$ |

Table 7: Per-setting tuning of the baselines on MoG40. Tuning uses seeds 5–14, disjoint from the reporting seeds 0–4. Against the baselines' defaults, the selected settings move each baseline by about one mode in every setting.

| Initialization | Method | Setting | Modes ↑ | −ELBO ↓ | TV ↓ |
|---|---|---|---|---|---|
| $\bar{m} = 0$<br>$s = 40$<br>$c = 10^{-5}$ | GMBBVI | default: $\Delta t_{\max} = 0.5$, $\beta = 0.9$ | $25.0 \pm 1.7$ | $0.515 \pm 0.075$ | 0.393 |
| | | selected: $\Delta t_{\max} = 1$, $\beta = 1.5$ | $25.9 \pm 1.0$ | $0.470 \pm 0.046$ | 0.370 |
| | SEMTRUX | default: step $= 1$, $\gamma_w = 1$, inc $= 1.15$ | $25.0 \pm 1.7$ | $0.503 \pm 0.074$ | 0.385 |
| | | selected: step $= 1$, $\gamma_w = 1$, inc $= 1.5$ | $25.5 \pm 1.4$ | $0.513 \pm 0.073$ | 0.381 |
| $\bar{m} = 0$<br>$s = 40$<br>$c = 1$ | GMBBVI | default: $\Delta t_{\max} = 0.5$, $\beta = 0.9$ | $25.7 \pm 1.2$ | $0.479 \pm 0.028$ | 0.375 |
| | | selected: $\Delta t_{\max} = 0.5$, $\beta = 1.5$ | $27.1 \pm 1.4$ | $0.432 \pm 0.035$ | 0.345 |
| | SEMTRUX | default: step $= 1$, $\gamma_w = 1$, inc $= 1.15$ | $24.8 \pm 0.8$ | $0.504 \pm 0.030$ | 0.389 |
| | | selected: step $= 2$, $\gamma_w = 1.5$, inc $= 1.3$ | $26.0 \pm 1.4$ | $0.518 \pm 0.068$ | 0.365 |
| $\bar{m} = (30, 30)$<br>$s = 5$<br>$c = 10^{-5}$ | GMBBVI | default: $\Delta t_{\max} = 0.5$, $\beta = 0.9$ | $7.2 \pm 0.4$ | $1.717 \pm 0.056$ | 0.820 |
| | | selected: $\Delta t_{\max} = 8$, $\beta = 1.5$ | $8.3 \pm 0.5$ | $1.588 \pm 0.041$ | 0.791 |
| | SEMTRUX | default: step $= 1$, $\gamma_w = 1$, inc $= 1.15$ | $7.1 \pm 0.3$ | $1.730 \pm 0.042$ | 0.822 |
| | | selected: step $= 16$, $\gamma_w = 1$, inc $= 1.15$ | $7.7 \pm 0.5$ | $1.650 \pm 0.065$ | 0.807 |
| | IBW | default: learning rate $= 0.05$ | $5.5 \pm 1.2$ | $6.407 \pm 1.331$ | 0.851 |
| | | selected: learning rate $= 0.005$ | $7.0 \pm 0.0$ | $2.896 \pm 0.267$ | 0.849 |