Variational Inference via Radial Transport

Luca Ghafourpour[1,2] **Sinho Chewi**[3] **Alessio Figalli**[1] **Aram-Alexandre Pooladian**[3]

[1]ETH Zürich [2]Cambridge [3]Yale University

ldg34@cam.ac.uk afigalli@ethz.ch {sinho.chewi,aram-alexandre.pooladian}@yale.edu

Abstract

In variational inference (VI), the practitioner approximates a high-dimensional distribution π with a simple surrogate one, often a (product) Gaussian distribution. However, in many cases of practical interest, Gaussian distributions might not capture the correct radial profile of π, resulting in poor coverage. In this work, we approach the VI problem from the perspective of optimizing over these radial profiles. Our algorithm `radVI` is a cheap, effective add-on to many existing VI schemes, such as Gaussian (mean-field) VI and Laplace approximation. We provide theoretical convergence guarantees for our algorithm, owing to recent developments in optimization over the Wasserstein space—the space of probability distributions endowed with the Wasserstein distance—and new regularity properties of radial transport maps in the style of Caffarelli (2000).

1 INTRODUCTION

Variational inference (VI) is a fundamental optimization problem that takes place over subsets of probability distributions (Wainwright and Jordan, 2008; Blei et al., 2017). We consider a standard setup that arises in many applications, where the practitioner is given a high-dimensional posterior distribution $\pi \propto \exp(-V)$ and the goal is to solve

$$\pi^\star_{\mathcal{C}} := \operatorname*{arg\,min}_{\mu \in \mathcal{C}} \mathrm{KL}(\mu \,\|\, \pi)\,,$$

where $\mathcal{C} \subset \mathcal{P}(\mathbb{R}^d)$ is a fixed set of probability distributions. VI is a powerful computational stand-in for standard Markov Chain Monte Carlo (MCMC) methods for sampling from unnormalized posteriors π. Indeed, while MCMC methods require simulating Markov chains for prohibitively long periods of time, it might be possible to instead quickly learn a surrogate density that is a good enough approximation to the posterior for practical purposes; see the review by Blei et al. (2017) for more details.

In VI, the choice of $\mathcal{C} \subset \mathcal{P}(\mathbb{R}^d)$ is of the utmost importance. For example, the case where $\mathcal{C}$ is the set of all Gaussians (with positive definite covariance) is known as *Gaussian VI* (Barber and Bishop, 1997; Seeger, 1999; Opper and Archambeau, 2009). In large-scale machine learning applications, it is also common to optimize over the class of Gaussians with diagonal covariance, resulting in mean-field Gaussian VI. While these algorithms have a long history, a rigorous, theoretical analysis is only just emerging, based on the theory of optimal transport through Wasserstein gradient flows (Ambrosio et al., 2008). For example, the Gaussian case has been studied by Lambert et al. (2022); Diao et al. (2023); Kim et al. (2023). We note that it is possible to implement algorithms based on mixtures of Gaussians as outlined by Lambert et al. (2022); Petit-Talamon et al. (2025), though the mathematical analysis in this case is significantly more challenging. Separately, *Laplace approximation* is an alternative means of obtaining a surrogate measure to π, where one considers the following Gaussian approximation $\mathcal{N}(x^\star, (\nabla^2 V(x^\star))^{-1})$ where $x^\star = \operatorname{argmin} V$. The literature on Laplace approximations is vast; see e.g., Robert and Casella (2004).

Margossian and Saul (2025) highlight the strengths and weaknesses of existing VI-based algorithms. Notably, they provide some characterizations for when VI can hope to exactly recover the mean and correlation matrix of a target distribution π. While they are only interested in these particular statistics, a key detail of the paper is that the variational approximating family

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Correspondence to LG and AAP.

must be decided in advance, which leads to demonstrable shortcomings even in small-scale examples.

1.1 Contributions

To mitigate the issues brought about by these approximations, we study the VI problem over *radial profiles*. For fixed $m \in \mathbb{R}^d$ and $\Sigma \succ 0$, we consider the following variational family:

$$p_h(x) \propto \det(\Sigma)^{-1/2}\, h((x-m)^\top \Sigma^{-1}(x-m))\,,$$

as h ranges over non-negative functions on $[0,\infty)$. If m and Σ are known, or if estimates thereof can be imputed, then we can assume $m = 0$ and $\Sigma = I$ via a whitening procedure (see Section 4.5). We henceforth assume that this has been done, so that our variational family is the set $\mathcal{C}_{\mathrm{rad}}$ of *radially symmetric* distributions. This family encompasses the standard Gaussian via $h(y) = \exp(-y/2)$, but also Student-t distributions, the non-smooth Laplace distribution, the logistic distribution, among others.*

In this paper, we propose and analyze a tractable algorithm for solving

$$\pi^\star_{\mathrm{rad}} \coloneqq \operatorname*{arg\,min}_{\mu \in \mathcal{C}_{\mathrm{rad}}} \mathrm{KL}(\mu \,\|\, \pi)\,,$$

where $\pi \propto \exp(-V)$. Our contributions are of both theoretical and computational interest. We stress that our only assumptions throughout this work will be on the true posterior π, namely that π is log-smooth and strongly log-concave and centered at the origin. This pair of assumptions has been leveraged in nearly all works on the theory of sampling (Chewi, 2026) and in the theoretical and computational study of variational inference (Lambert et al., 2022; Arnese and Lacker, 2024; Lacker et al., 2024; Lavenant and Zanella, 2024; Jiang et al., 2025).

In Section 3, we prove existence and uniqueness of the radial minimizer $\pi^\star_{\mathrm{rad}}$, as well as establish regularity properties of said minimizer. For example, if π is log-smooth and strongly log-concave, Proposition 3.4 states that $\pi^\star_{\mathrm{rad}}$ is as well. We also prove Caffarelli-type contraction estimates (Caffarelli, 2000) for the corresponding optimal radial transport map $T^\star_{\mathrm{rad}}$ from the standard Gaussian $\rho = \mathcal{N}(0, I)$, say, to $\pi^\star_{\mathrm{rad}}$; see Theorem 3.5.

In Section 4, we embrace the conventional wisdom of "parametrizing then optimizing" in order to compute $\pi^\star_{\mathrm{rad}}$, leading to our proposed algorithm `radVI` (see Algorithm 1). Concretely, we make use of the representation of a given radial measure as the pushforward of the standard Gaussian by a radial map T_{rad}. Our approach is based on carefully parametrizing radial transport maps T_λ for $\lambda \in \mathbb{R}^{J+1}_+$ for some $J > 0$ where, if J is large enough, our parameterized set should encompass all possible radial maps (see Theorem 4.1). Then, writing our objective over the non-negative orthant as

$$\lambda \mapsto \mathcal{F}(\lambda) = \mathrm{KL}((T_\lambda)_\sharp \rho \,\|\, \pi) \quad (\lambda \in \mathbb{R}^{J+1}_+)$$

we show that standard Euclidean gradient descent converges to the *true* optimal radial transport map $T^\star_{\mathrm{rad}}$, i.e., $\|T_{\lambda^{(k)}} - T^\star_{\mathrm{rad}}\|_{L^2(\rho)} \le \varepsilon$ for all $k \ge \widetilde{\Omega}(\varepsilon^{-1})$ up to polynomial factors of the condition number of π—see Theorem 4.3. Notably, our convergence guarantees are effectively *dimension-free*.

In Section 5, we show how `radVI` can be (easily) used to improve existing VI methods on a collection of synthetic examples. In addition to recovering isotropic profiles, we also show how `radVI` can be used as a preconditioner: given any method of obtaining a mean-covariance proxy, such as through Laplace approximation or Gaussian VI, we show how `radVI` can often improve upon the existing approximation by better capturing the tail behavior of the posterior. As a final example, we turn to parameter estimation problems (estimating the second moment, or probability thresholds) given an unnormalized posterior. Focusing on the Neal's funnel distribution, we show how `radVI` on top of full-rank Gaussian VI can lead to substantially improved estimates of these quantities at virtually no added computational cost. We believe these findings merit further investigation.

Notation

We write $\mu \in \mathcal{P}(\mathbb{R}^d)$ if μ is a probability measure over $\mathbb{R}^d$; $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if μ has finite second moment; $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ if μ has a Lebesgue density; and $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) = \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_2(\mathbb{R}^d)$. For positive constants a and b, we write $a \lesssim b$ or $a = \mathcal{O}(b)$ (resp. $a \gtrsim b$ or $a = \Omega(b)$) to mean there exists a positive constant C such that $a \le Cb$ (resp. $a \ge Cb$). If both $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$. When omitting logarithmic factors in b, we write, e.g., $a \lesssim_{\log} b$ or $a = \widetilde{\mathcal{O}}(b)$ (analogously for lower bounds). Throughout, implied constants will always be independent of the dimension and other relevant problem parameters. We write the uniform distribution on the unit sphere (denoted S^{d-1}) in $\mathbb{R}^d$ as Unif. Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth and m-strongly convex if $0 \preceq mI \preceq \nabla^2 f \preceq MI$.

*On the other hand, if we fix h and vary (m, Σ), we end up at the theory of elliptical families, see e.g., Muirhead (1982, Section 1.4).

2 BACKGROUND

2.1 Primer on optimal transport

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the squared 2-*Wasserstein distance* between them is defined as

$$W_2^2(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \iint \|x-y\|^2 \,\mathrm{d}\pi(x,y)\,, \quad (1)$$

where $\Pi(\mu,\nu)$ is the set of joint measures with first marginal μ and second marginal equal to ν. For more details, see Villani (2009); Santambrogio (2015); Chewi et al. (2025).

The associated *optimal transport map* between μ and ν is given by

$$T^\star := \operatorname*{arg\,min}_{T \in \mathcal{T}(\mu,\nu)} \int \|x - T(x)\|^2 \,\mathrm{d}\mu(x)\,, \quad (2)$$

where $\mathcal{T}(\mu,\nu)$ is the set of admissible transport maps between μ and ν, which consist of vector-valued functions $T : \mathbb{R}^d \to \mathbb{R}^d$ such that for $X \sim \mu$, it holds that $T(X) \sim \nu$. Without further assumptions, it is possible that such a $T^\star$ need not exist. A theorem due to Brenier (1991) unifies both (1) and (2).

Theorem 2.1 (Brenier's theorem)**.** *Suppose $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Then (2) has a unique minimizer, with*

$$W_2^2(\mu,\nu) = \int \|x - T^\star(x)\|^2 \,\mathrm{d}\mu(x)\,,$$

and $T^\star = \nabla\varphi^\star$ for some convex function $\varphi^\star$.

2.2 Geodesic convexity via compatible families of optimal transport maps

We now consider the space of probability measures with densities endowed with the 2-Wasserstein distance, called the *Wasserstein space* $\mathbb{W}_2 := (\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), W_2)$. Our goal is to formulate the VI problem as an optimization problem over this space.

We let $\mathcal{T}_{\mathrm{rad}}$ denote the set of radial (transport) maps, which are vector-valued functions $T_{\mathrm{rad}} : \mathbb{R}^d \to \mathbb{R}^d$ with

$$x \mapsto T_{\mathrm{rad}}(x) = \Psi(\|x\|)\, x/\|x\|\,, \quad (3)$$

where $\Psi : [0,\infty) \to [0,\infty)$ is strictly monotone and continuous. One can check that $\mathcal{T}_{\mathrm{rad}}$ forms a compatible family of optimal transport maps (see Panaretos and Zemel, 2020, Section 2.3.2), and thus $(\mathcal{T}_{\mathrm{rad}})_\sharp \rho$ forms a geodesically convex subset of the Wasserstein space.

For an introduction to the space $\mathbb{W}_2$ focused on applications to statistics, see Panaretos and Zemel (2020); Chewi et al. (2025).

3 SETUP AND THEORETICAL RESULTS

Given a posterior $\pi \propto \exp(-V)$, our objective is to compute the following minimizer:

$$\pi_{\mathrm{rad}} \in \operatorname*{arg\,min}_{\mu \in \mathcal{C}_{\mathrm{rad}}} \mathrm{KL}(\mu \,\|\, \pi)\,, \quad (\mathsf{radVI})$$

where we recall that $\mathcal{C}_{\mathrm{rad}}$ is the set of radially symmetric measures. In this section, we first prove various theoretical properties surrounding π_{rad}, and we develop computational mechanisms in Section 4.

To this end, our sole assumption will be on the true posterior $\pi \propto \exp(-V)$. Namely, we will assume that π is *well-conditioned*: it is log-smooth and strongly log-concave, i.e., for $\ell_V, L_V > 0$,

$$0 \preceq \ell_V I \preceq \nabla^2 V \preceq L_V I\,, \quad (\mathsf{WC})$$

and minimized at the origin. These conditions are standard in the theoretical literature on log-concave sampling (Chewi, 2026) as well as in variational inference (see, e.g., Lambert et al., 2022; Arnese and Lacker, 2024; Lacker et al., 2024; Lavenant and Zanella, 2024; Jiang et al., 2025).

Remark 3.1. We note that the smoothness of V is not a particularly strong assumption as, e.g., Gaussian mixtures fall into this class. However, the strong convexity of V implies unimodality of the posterior, which is admittedly stringent. Despite the VI problem being well-posed even without this assumption (Proposition 3.2), we require it later in order to obtain guarantees for our Wasserstein gradient flow algorithm.

3.1 Regularity of radial minimizer

We first collect some basic properties of the optimal radial minimizer $\pi^\star_{\mathrm{rad}}$. Proofs of results from this subsection can be found in Appendix A.

We first state a result concerning the existence and uniqueness of the minimizer $\pi^\star_{\mathrm{rad}}$.

Proposition 3.2. *Assume that there exists $\mu \in \mathcal{C}_{\mathrm{rad}}$ with $\mathrm{KL}(\mu \,\|\, \pi) < \infty$. Then, there exists a unique minimizer to (radVI).*

Next, we explicitly characterize the stationary condition of the minimizer $\pi^\star_{\mathrm{rad}}$, where the proof is based on calculus of variations.

Proposition 3.3 (Stationary condition)**.** *Suppose $\pi \propto \exp(-V)$ and a minimizer $\pi^\star_{\mathrm{rad}}$ to (radVI) exists. Then it holds that $\pi^\star_{\mathrm{rad}} \propto \exp(-\overline{V})$ with*

$$\overline{V}(x) := \int_{S^{d-1}} V(\|x\|\theta)\,\mathrm{d}\mathsf{Unif}(\theta)\,, \quad (4)$$

where Unif *is the uniform measure on* S^{d-1}.

Using Proposition 3.3, we can leverage the existing regularity of π (through V) to show that $\pi^\star_{\mathrm{rad}}$ is also regular. In particular, log-smoothness and strong log-concavity are preserved.

Proposition 3.4 (Regularity of radial minimizer)**.** *Suppose* π *satisfies* (WC)*. The radial minimizer* $\pi^\star_{\mathrm{rad}}$ *also satisfies* (WC) *with the same parameters.*

3.2 Regularity of optimal radial maps

We now establish regularity properties of optimal radial (transport) maps from, say, the standard Gaussian to the minimizer of (radVI). Our main theorem below is essentially a specialization of Caffarelli's contraction theorem (Caffarelli, 2000) to the radial component of the optimal transport map.

Theorem 3.5 (Regularity of the optimal radial map)**.** *Suppose* π *satisfies* (WC) *and consider the corresponding minimizer to* (radVI)*, denoted* $\pi^\star_{\mathrm{rad}}$*. Let* $T^\star_{\mathrm{rad}}$ *denote the unique optimal transport map from* ρ *to* $\pi^\star_{\mathrm{rad}}$*, and write* $T^\star_{\mathrm{rad}}(x) = \Psi^\star(\|x\|)\, x/\|x\|$*. Writing* $r = \|x\|$*, the following regularity conditions hold:*

$$\frac{1}{\sqrt{L_V}} \le (\Psi^\star)'(r) \le \frac{1}{\sqrt{\ell_V}}\,,$$
$$|(\Psi^\star)''(r)| \lesssim \frac{\kappa}{\sqrt{\ell_V}}\left(1+\frac{d}{r^2}\right)\left(1+|r-\sqrt{d}|\right),$$

where $\kappa := L_V/\ell_V$.

The first result in Theorem 3.5 follows from a direct application of Caffarelli's contraction theorem; this is made possible by Proposition 3.4. To prove the second result, we differentiate the Monge–Ampère equation, and exploit the existing regularity of (the first derivative of) $\Psi^\star$. A key takeaway is that under ρ, the norm is tightly concentrated around $r = \sqrt{d} \pm O(1)$, and thus the bound on $(\Psi^\star)''$ is nearly dimension-free. While it is possible to differentiate the Monge–Ampère equation a second time and obtain *third*-order control on $\Psi^\star$, we omit this result as it is not necessary for our purposes. The full proof of Theorem 3.5 can be found in Appendix B.

4 PARAMETRIZE THEN OPTIMIZE

Recall that our goal is to learn the probability distribution $\pi^\star_{\mathrm{rad}}$ given query access to the (gradient of the) potential V of the unnormalized posterior π. Our proposed algorithm will directly learn the optimal transport map from ρ (an easy-to-sample reference measure) to the optimal radial distribution $\pi^\star_{\mathrm{rad}}$.

We first note the following equivalent optimization problems:

$$\min_{\mu \in \mathcal{C}_{\mathrm{rad}}} \mathrm{KL}(\mu \,\|\, \pi) = \min_{T \in \mathcal{T}_{\mathrm{rad}}} \mathrm{KL}(T_\sharp \rho \,\|\, \pi)\,,$$

where $\rho = \mathcal{N}(0, I)$. Indeed, any radial distribution $\mu \in \mathcal{C}_{\mathrm{rad}}$ can be expressed as the pushforward of the standard Gaussian with any other radial transport map in $\mathcal{T}_{\mathrm{rad}}$. Thus, in this section, we will be focusing on the optimization problem

$$T^\star_{\mathrm{rad}} = \operatorname*{arg\,min}_{T \in \mathcal{T}_{\mathrm{rad}}} \mathrm{KL}(T_\sharp \rho \,\|\, \pi)\,, \qquad \text{(radVI-T)}$$

from which we recover the optimal radial distribution.

We follow the approach put forth by Jiang et al. (2025), who suggest to appropriately *parametrize* the space of transport maps and then to *optimize* over it. Letting $\{\Psi_j\}_{j=0}^J$ be a fixed set of basis functions and $\alpha > 0$, we consider a parametrized subset of transport maps $\mathcal{T}_J \subseteq \mathcal{T}_{\mathrm{rad}}$, given by

$$\mathcal{T}_J := \left\{ \left(\alpha\,\|x\| + \sum_{j=0}^{J} \lambda_j \Psi_j(\|x\|)\right) \frac{x}{\|x\|} \;\middle|\; \lambda \in \mathbb{R}_+^{J+1} \right\}. \quad (5)$$

In the rest of this section, we will address the following questions, which naturally arise as a result of our choice of parametrization:

- What is the proposed basis $\{\Psi_j\}_{j=0}^J$? How large does J need to be such that $\mathcal{T}_J$ is a faithful approximation to $\mathcal{T}_{\mathrm{rad}}$?
- Does optimizing over $\mathcal{T}_J$ yield a map which is close to $T^\star_{\mathrm{rad}}$? Is it possible to obtain optimization guarantees?

4.1 Approximation guarantees

We first describe our choice of basis. As an arbitrary function Ψ is a continuous, strictly monotone function, we parametrize this class of functions by a piecewise linear monotone curve. Given a cutoff variable $R > 0$, we consider the following sequence of equi-spaced piecewise monotone functions on the interval $[\sqrt{d} - R,\, \sqrt{d} + R] \subset \mathbb{R}$

$$\Psi_j(r) := \Psi_{\mathrm{base}}(\delta_j^{-1}(r - a_j))\,, \quad (6)$$

where for $j \ge 1$, $\delta_j := \delta$ is the mesh size, $\{a_j\}_{j=1}^J$ are the knots; and for $j = 0$, $\delta_0 := \sqrt{d} - R$ and $a_0 := 0$. Here, $\Psi_{\mathrm{base}}(x) := 0 \vee (x \wedge 1)$. Ultimately, we will choose

$J = 2R/\delta + 1$ (we assume that R is divisible by δ and that $R \le \sqrt{d}$).

Our first result of this section establishes a "universality" property of the set $\mathcal{T}_J$ from an approximation perspective, and yields the choice of R and J required to complete our construction. Since we mainly care about high-dimensional approximation, henceforth we assume $d \ge 3$.

Theorem 4.1 (Universal approximation)**.** *Let $T^\star \in \mathcal{T}_{\mathrm{rad}}$ which satisfies Theorem 3.5, and let $\varepsilon \gg \exp(-\Omega(d))$. Define $\mathcal{T}_J$ with $\alpha = 1/\sqrt{L_V}$ and choose $R \asymp \sqrt{\log(d/\varepsilon)}$, $J = \widetilde{\Omega}(\sqrt{\kappa/\varepsilon})$ with the basis elements given by (6). Then there exists a $\hat{T} \in \mathcal{T}_J$, i.e., there exists $\hat{\lambda} \in \mathbb{R}_+^{J+1}$ with $\hat{T} = T_{\hat{\lambda}}$, such that*

$$\|T^\star - \hat{T}\|_{L^2(\rho)} \le \varepsilon/\ell_V^{1/2}\,,$$
$$\|D(T^\star - \hat{T})\|_{L^2(\rho)} \le \widetilde{\mathcal{O}}(\kappa^{1/2}\varepsilon^{1/2}/\ell_V^{1/2})\,.$$

4.2 Proposed algorithm: `radVI`

We now present our basic algorithm. Recall that our objective function, the KL divergence, is

$$\begin{aligned}\mathrm{KL}(\mu\,\|\,\pi) &= \mathcal{V}(\mu) + \mathcal{H}(\mu) + \log Z\,,\\ &:= \int V\,\mathrm{d}\mu + \int \log\mu\,\mathrm{d}\mu + \log Z\,,\end{aligned}$$

where $Z > 0$ is the unknown normalizing constant of π. If we write $\mu = T_\sharp\rho$, for some transport map T, then by a change-of-variables calculation, one obtains (up to omitted constants)

$$\mathrm{KL}(T_\sharp\rho\,\|\,\pi) = \int V \circ T\,\mathrm{d}\rho - \int \log\det DT\,\mathrm{d}\rho\,.$$

We now optimize over our prescribed parametrization. For $T_\lambda \in \mathcal{T}_J$, we see that we can write the KL divergence as a function over the non-negative orthant:

$$\lambda \mapsto \mathcal{F}(T_\lambda) := \mathrm{KL}((T_\lambda)\sharp\rho\,\|\,\pi) \quad (\lambda \in \mathbb{R}_+^{J+1})\,. \tag{7}$$

To continue, we require the following observation: there exists an isometry between the $L^2(\rho)$ distance on the transport maps $T_\lambda \in \mathcal{T}_J$ and a Euclidean distance over the weights $\lambda \in \mathbb{R}_+^{J+1}$. Indeed, for any two parameters $\lambda, \eta \in \mathbb{R}_+^{J+1}$, one readily computes

$$\begin{aligned}\|T_\lambda - T_\eta\|_{L^2(\rho)}^2 &= \|\textstyle\sum_{j=1}^J(\lambda_j - \eta_j)\,\Psi_j(\|x\|)\|_{L^2(\rho)}^2\\ &= (\lambda-\eta)^\top Q(\lambda-\eta)\,,\end{aligned}$$

where $Q \in \mathbb{S}_+^{J+1}$ is a Gram matrix with entries given by

$$Q_{i,j} := \mathbb{E}_{X\sim\rho}[\Psi_i(\|X\|)\,\Psi_j(\|X\|)]\,.$$

Algorithm 1 `radVI`

Input: Posterior $\pi \propto \exp(-V)$ with access to ∇V
Free parameters: Choose $K, h > 0$
Construct: Basis family $\{\Psi_j\}_{j=0}^J$ and matrix Q
Initialize: $\lambda^{(0)} \in \mathbb{R}_+^{J+1}$
while $k = 0, 1, \dots, K-1$ **do**

Compute stochastic gradient $\widehat{\nabla}_\lambda\mathcal{F}(T_{\lambda^{(k)}})$
$\lambda^{(k+1)} = \mathrm{Proj}_{\mathbb{R}_+^{J+1},\|\cdot\|_Q}\big(\lambda^{(k)} - hQ^{-1}\widehat{\nabla}_\lambda\mathcal{F}(T_{\lambda^{(k)}})\big)$

end while
Return: $T_{\lambda^{(K)}}$

We detail in Appendix C.2 how to compute the entries $Q_{i,j}$ via truncated moments of the chi distribution. Thus, convergence of the radial maps corresponds to convergence of the parameters in a *weighted* Euclidean space, namely $(\mathbb{R}_+^{J+1}, \|\cdot\|_Q)$. In other words, a discretization of a gradient flow of (7) would naturally correspond to *projected gradient descent* in the weighted metric $\|\cdot\|_Q$:

$$\lambda^{(k+1)} = \mathrm{Proj}_{\mathbb{R}_+^{J+1},\|\cdot\|_Q}\big(\lambda^{(k)} - hQ^{-1}\nabla_\lambda\mathcal{F}(T_{\lambda^{(k)}})\big)\,, \tag{8}$$

where $h > 0$ is the stepsize, and $\nabla_\lambda\mathcal{F}(T_\lambda)$ is the gradient of the objective function in the parameters λ.

Our complete algorithm, called `radVI`, is presented in Algorithm 1. Note that in practice, a stochastic estimate of the gradient will be used in place of the full gradient. Thus, `radVI` can be seen as a special instance of stochastic projected gradient descent (SPGD). We discuss this more in Section 4.4.

Remark 4.2. It is worth stressing that the choice of using gradient descent as a first-order algorithm was entirely arbitrary, and many other algorithms (e.g., Frank–Wolfe, mirror descent, etc.) are applicable within our framework. The functional of interest can be suitably arbitrary as well; see Jiang et al. (2025) for more details.

4.3 Optimization guarantees

The main result of this section is the following quantitative convergence of `radVI` to the optimal radial map.

Theorem 4.3 (Convergence of `radVI`)**.** *Suppose π satisfies (WC), and consider the family of transport maps constructed via (6). Then, the transport map $T_{\lambda^{(K)}}$ with $(\lambda^{(k)})_{k\ge0}$ given by (8) is ε-close to $T_{\mathrm{rad}}^\star$ (in $L^2(\rho)$) so long as $J = \widetilde{\Theta}(\kappa^2/\varepsilon)$, the step size $h = \widetilde{\Theta}(\varepsilon/(L_V\kappa^2))$ and for iterations*

$$K = \widetilde{\Omega}\big(\kappa^5\varepsilon^{-1}\log(\mathrm{KL}((T_{\lambda^{(0)}})_\sharp\rho\,\|\,\pi)/\varepsilon^2)\big)\,.$$

We outline a proof sketch, leaving the fine details to Appendix C.4. Let us first assume that the sequence $(\lambda^{(k)})_{k\geq 0}$ eventually converges to some optimal $\lambda^\star \in \mathbb{R}_+^{J+1}$. If π satisfies (WC), then the corresponding $T^\star_{\mathrm{rad}}$ satisfies the Theorem 3.5. Also, by Theorem 4.1, there exists $\hat{\lambda} \in \mathbb{R}_+^{J+1}$ such that $T_{\hat{\lambda}} \in \mathcal{T}_J$ is ε-close to $T^\star_{\mathrm{rad}}$ for J sufficiently large. By triangle inequality, we have

$$\begin{aligned}\|T_{\lambda^\star} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} &\lesssim \|T_{\lambda^\star} - T_{\hat{\lambda}}\|^2_{L^2(\rho)} + \|T_{\hat{\lambda}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \\ &\lesssim \|T_{\lambda^\star} - T_{\hat{\lambda}}\|^2_{L^2(\rho)} + \varepsilon^2 \,.\end{aligned}$$

Appealing to the strong convexity of the KL divergence along generalized geodesics, one can show that the remaining term can be bounded by precisely *both* terms from Theorem 3.5

$$\begin{aligned}\|T_{\lambda^\star} - T_{\hat{\lambda}}\|^2_{L^2(\rho)} \lesssim \kappa \,\|T_{\hat{\lambda}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \\ + \kappa^2 \,\|D(T_{\hat{\lambda}} - T^\star_{\mathrm{rad}})\|^2_{L^2(\rho)} \,.\end{aligned} \tag{9}$$

See Jiang et al. (2025, Appendix C) for a proof of this fact (in particular, the proof of their Theorem 5.9 and Corollary C.3). Thus, for J large enough, the minimizer over $\mathcal{T}_J$, i.e., $T_{\lambda^\star}$, can be made arbitrarily close to $T^\star_{\mathrm{rad}}$. To conclude, it remains to quantitatively assess that $T_{\lambda^{(k)}} \to T_{\lambda^\star}$, and then string everything together.

By the isometry property, it holds that

$$\|T_{\lambda^{(k)}} - T_{\lambda^\star}\|^2_{L^2(\rho)} = \|\lambda^{(k)} - \lambda^\star\|^2_Q \,,$$

where we recall that $Q \succ 0$ is the fixed Gram matrix defined from the basis functions. If $\lambda \mapsto \mathcal{F}(T_\lambda)$ is smooth and strongly convex (with respect to $\|\cdot\|_Q$), we can readily apply existing results for the convergence of first-order algorithms (see, e.g., Beck, 2017). Thankfully, the requisite properties of our functional can be verified; see Appendix C.3.

Proposition 4.4. *Suppose π satisfies (WC), and also consider $\mathcal{T}_J$ chosen as in Theorem 4.1. Then $\lambda \mapsto \mathcal{F}(T_\lambda)$ is ℓ_V-strongly convex and $\widetilde{O}(J^2 L_V)$-smooth (with respect to $\|\cdot\|_Q$).*

Remark 4.5. Note that the smoothness constant of $\lambda \mapsto \mathcal{F}(T_\lambda)$ explodes as $J \to \infty$. This is unsurprising, as the functional $\mu \mapsto \mathrm{KL}(\mu \,\|\, \pi)$ is not smooth over the space of all probability measures. In contrast, our parametrization creates a strict subset of all probability measures, over which the constant can be controlled.

4.4 Stochastic optimization

In practice, we use *stochastic* projected gradient descent (SPGD) to optimize $\lambda \mapsto \mathcal{F}(T_\lambda)$ over $(\mathbb{R}_+^{J+1}, \|\cdot\|_Q)$. Recall our full objective function (up to constants) is

$$\mathcal{F}(T_\lambda) = \int V(T_\lambda(x))\,\mathrm{d}\rho(x) - \int \log\det(DT_\lambda(x))\,\mathrm{d}\rho(x) \,.$$

We compute the gradient of the two terms separately. For the first term, we pass the gradient in the weights λ under the expectation and, due to the definition of T_λ, the inner gradient is simply

$$\nabla_\lambda V(T_\lambda(X)) = \vec{\Psi}(\|X\|)(X/\|X\|)^\top \nabla V(T_\lambda(X)) \,, \tag{10}$$

where $\vec{\Psi}(r) \coloneqq [\Psi_0(r), \ldots, \Psi_J(r)]$. Thus, to compute $\nabla_\lambda \mathbb{E}_{X\sim\rho}[V(T_\lambda(X))]$, it suffices to use a Monte Carlo approximation using i.i.d. draws $X_1, \ldots, X_N \sim \rho$.

For the second term, we first compute

$$\begin{aligned}\log\det(DT_\lambda(x)) = (d-1)\log(\alpha + \langle\lambda, \vec{\Psi}(\|x\|)\rangle/\|x\|) \\ + \log(\alpha + \langle\lambda, \vec{\Psi}'(\|x\|)\rangle) \,,\end{aligned}$$

where $\vec{\Psi}'(r) \coloneqq [\Psi_0'(r), \ldots, \Psi_J'(r)]$. It is straightforward to show that $\nabla_\lambda \log\det(DT_\lambda(x))$ has an analytical expression, and thus the integrated expression can again be computed via Monte Carlo integration. However, due to the precise nature of $\vec{\Psi}$ and $\vec{\Psi}'$, it is possible to evaluate our second integrated quantity in our objective using simple univariate numerical integration; we briefly touch on this in Appendix C.5.

The next theorem states that `radVI` still comes with convergence guarantees when using stochastic gradient estimates (e.g., Monte Carlo estimates) for (10). Unlike many works that use SPGD, we *prove* that under the well-conditioned assumption, we satisfy a classical *bounded variance* property which permits us to use existing theory; see Appendix C.6 for a proof.

Theorem 4.6 (Convergence of stochastic `radVI`)**.** *Assume that π is well-conditioned (WC) and consider the family of transport maps constructed via (6). Then, for all sufficiently small ε, the iterates of stochastic projected gradient descent yield a measure $\mu_{(t)}$ with the guarantee $\ell_V \mathbb{E}\|T_{\lambda^{(k)}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \leq \varepsilon^2$, with a number of iterations bounded by*

$$K = \widetilde{\Omega}\Big(d\kappa^2 J^3 L_V \varepsilon^{-2} \log(\|T_{\lambda^{(0)}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)}/\varepsilon)\Big) \,,$$

and step size $h = \widetilde{\Theta}(\varepsilon^2/(d\kappa^3 J^3 L_V))$. Moreover, if we choose the parameters of our dictionary as in Theorem 4.1, then we achieve the same ε-closeness above with

$$K = \widetilde{\Omega}\Big(d\kappa^{5/2} \varepsilon^{-5} \log(\|T_{\lambda^{(0)}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)}/\varepsilon)\Big) \,.$$

4.5 `radVI` with whitening

As we highlighted in the introduction, a strength of `radVI` is that it can be used in conjunction with other variational methods based on, say, Gaussian distributions, such as Gaussian VI, mean-field Gaussian VI,

Algorithm 2 `radVI` with whitening

Input: Posterior $\pi \propto \exp(-V)$ with access to ∇V
Whitening stage:
1. Fix $(m, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_+$
2. Define $x \mapsto T_{m,\Sigma}(x) := \Sigma^{1/2}x + m$
3. Modify posterior via $\widetilde{V} \leftarrow V \circ T_{m,\Sigma}$

Obtain: $\hat{T}_{\mathrm{rad}} \leftarrow \texttt{radVI}(\exp(-\widetilde{V}))$
Return: Composite map $T_{\mathrm{comp}} \leftarrow T_{m,\Sigma} \circ \hat{T}_{\mathrm{rad}}$

and Laplace approximation. Algorithm 2 illustrates how to use any of these off-the-shelf Gaussian approximation methods to whiten the target distribution and improve performance.

In summary, any Gaussian measure $\mathcal{N}(m, \Sigma)$ can be expressed as $(T_{m,\Sigma})_\sharp \rho$ where $\rho = \mathcal{N}(0, I)$ and $T_{m,\Sigma}(x) = \Sigma^{1/2}x + m$. Thus, given any Gaussian approximation to π, we can use the corresponding affine map to whiten the posterior $\pi \propto \exp(-V)$ by defining $\widetilde{V} := V \circ T_{m,\Sigma}$. Then, we run `radVI` on $\exp(-\widetilde{V})$, and output the composition of these two maps. Below, we briefly review two standard Gaussian approximation methods in the literature.

Laplace approximations (LA). The Laplace approximation method occurs in two stages. For a posterior $\pi \propto \exp(-V)$, we first compute the mode of the distribution, $x^\star \in \arg\min_{x \in \mathbb{R}^d} V(x)$, and then we compute $(\nabla^2 V(x^\star))^{-1}$. The final approximation to the posterior π is then

$$\pi_{\mathrm{LA}} = \mathcal{N}(x^\star, (\nabla^2 V(x^\star))^{-1})\,. \tag{11}$$

Note that this method fails if $\nabla^2 V(x^\star)$ is not invertible, and is known to be inaccurate when π is heavily skewed (when the mode is far away from the mean). See Katsevich (2023, 2024) for recent statistical developments.

Gaussian VI (GVI). In Gaussian VI, the practitioner replaces $\mathcal{C}_{\mathrm{rad}}$ in (radVI) with $\mathcal{N}$, the set of all normal distributions with positive definite covariance. The resulting optimization problem becomes

$$\pi_{\mathrm{GVI}} \in \arg\min_{\mu \in \mathcal{N}} \mathrm{KL}(\mu \,\|\, \pi)\,; \tag{12}$$

see Lambert et al. (2022); Diao et al. (2023); Kim et al. (2023). A major limitation to GVI is the storage and per-iteration complexity, as the running covariance estimate needs to be stored and inverted at each iteration, which is costly for $d \gg 1$. Optimizing over *product* Gaussian measures (Gaussian mean-field VI, or MFVI) can mitigate these numerical hurdles, reducing the per-iteration complexity from $\mathcal{O}(d^3)$ to $\mathcal{O}(d)$ at the cost of possibly being much farther from π.

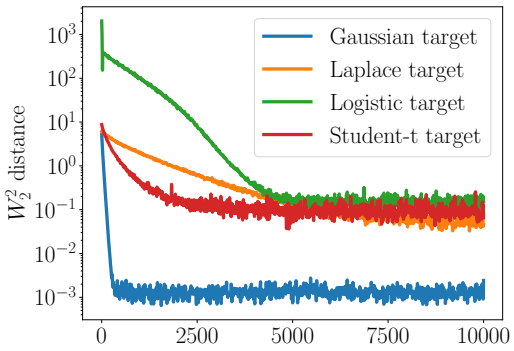

Figure 1: Convergence of `radVI` for various target distributions. See Table 1 for final-iterate comparisons between GVI and LA.

Remark 4.7. We briefly mention the work of Liang et al. (2022), which similarly tries to adjust the tail distributions of their approximated distribution. However, their approach is (i) parametric in nature, and (ii) builds off the mean-field variational inference perspective. Specifically, they only consider product measure approximations to the posteriors.

5 EXPERIMENTS

In all experiments, $d = 50$, we choose $\alpha = 0.01$ and $\lambda^{(0)} = \mathbf{1}_J$, $R = \sqrt{\log d}$, and $\delta = d^{-1/6}$ as parameters for constructing our dictionary. For LA, we use a standard optimization solver and closed-form expressions of the gradient and Hessian. For GVI, we use the Forward–Backward method of Diao et al. (2023).[†]

The candidate posteriors for the majority of this section are the Student-t, Laplace, and logistic distributions. We remark that none of these distributions fully satisfy our requirements. For instance, none of them are strongly convex (in fact, the Student-t distribution is non-log-concave), and the Laplace distribution is non-smooth. For Student-t, we use 10 degrees of freedom, which leads to significantly heavier tails than a Gaussian. Nevertheless, we are able to show that our scheme can recover the desired target distribution to better accuracy than standard VI methods. The precise definitions of these distributions, their potential functions, etc., can be found in Appendix D.2. In Section 5.3, we study parameter estimation from Neal's funnel distribution, a standard hierarchical prior.

[†]The code used for our implementation of `radVI` and the numerical experiments are made available at github.com/gluca99/radVI.

Method	Isotropic targets			
	Gaussian	Laplace	Logistic	Student-t
LA	2.45×10^{-4}	20.00	1.6×10^{3}	25.87
GVI	7.34×10^{-4}	8.24	3.96	1.99
`radVI`	$\mathbf{1.15 \times 10^{-4}}$	$\mathbf{5.37 \times 10^{-2}}$	$\mathbf{1.84 \times 10^{-1}}$	$\mathbf{1.19 \times 10^{-1}}$

Table 1: Estimated squared Wasserstein distance between various VI solutions for learning isotropic targets.

5.1 Learning isotropic radial families

We first investigate the case where the target measure is isotropic and radially symmetric. As $\pi = \pi^{\star}_{\mathrm{rad}} = (T^{\star}_{\mathrm{rad}})_{\sharp}\rho$, for any iterate $\lambda^{(k)}$ in our algorithm, we can approximate the L^2 distance between the maps via

$$\|T_{\lambda^{(k)}} - T^{\star}_{\mathrm{rad}}\|^2_{L^2(\rho)} \simeq \frac{1}{n}\sum_{i=1}^{n} \|T_{\lambda^{(k)}}(X_i) - T^{\star}_{\mathrm{rad}}(X_i)\|^2 ,$$

where $X_i \sim \rho$, and we use $n = 10^4$ to estimate all quantities. $T^{\star}_{\mathrm{rad}}$ is known in closed form for the Gaussian case and Student-t distribution, but must be solved numerically for the logistic and Laplace distributions; see Appendix D.2 for a short explanation. We compare the (squared) 2-Wasserstein distance between ground truth samples and those generated by LA and GVI. See Figure 1 for a plot comparing the convergence of these various methods. As a sanity check, our method recovers the Gaussian distribution with an error tolerance comparable to Gaussian methods. For the heavy-tailed Student-t distribution, however, we outperform these methods by wide margins. Repeating the experiment in $d = 100$ gives similar results (see Appendix E.1).

It is perhaps more informative to visually distinguish the approximation schemes. To this end, we plot the radial profiles of the various approximation methods. For instance, the top of Figure 2 compares the various learned profiles for the Student-t distribution. Unlike LA or GVI, `radVI` can find a close radial profile to the target. The corresponding figures for isotropic Laplace and logistic distributions appear in Appendix E.

Finally, we mention that we additionally performed a simple sensitivity analysis regarding our hyperparameter α similar to Figure 2 of Jiang et al. (2025). Focusing on the isotropic Gaussian case in which the ground-truth value is $\alpha = 1/\sqrt{L_V} = 1$, we run our algorithm where we vary $\alpha \in \{10^0, 10^{-1}, 10^{-2}\}$. Figure 3 shows that `radVI` converges to the optimal parameters exponentially fast, highlighting how our algorithm is robust to the choice of α.

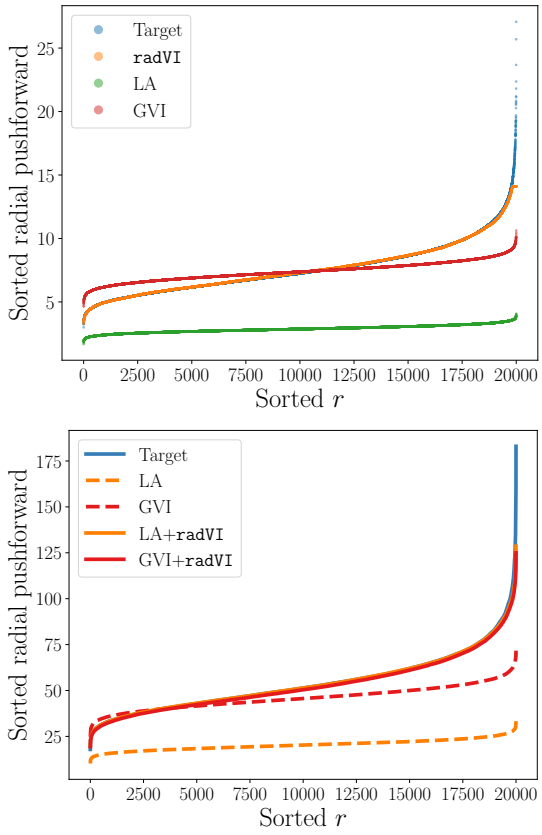

Figure 2: Comparing learned radial profiles of `radVI` versus other approximation methods for learning the Student-t distribution in the isotropic (**top**) and anisotropic case **(bottom)**.

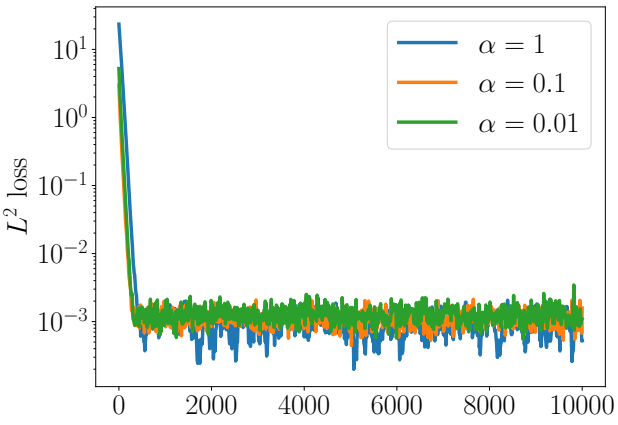

Figure 3: In the case where π is an isotropic Gaussian with $d = 50$, we verify that `radVI` is robust to the choice of α.

5.2 Learning anisotropic distributions

We now consider the anisotropic setting, where the distribution has a randomly generated covariance parameter $\Sigma = AA^{\top} + I$, where $A_{ij} \sim \mathcal{N}(0, 1)$ and $m = 0$. Following Algorithm 2, we first run either LA or GVI to create Gaussian approximations, and then we obtain our complete composite map, allowing us to draw as many

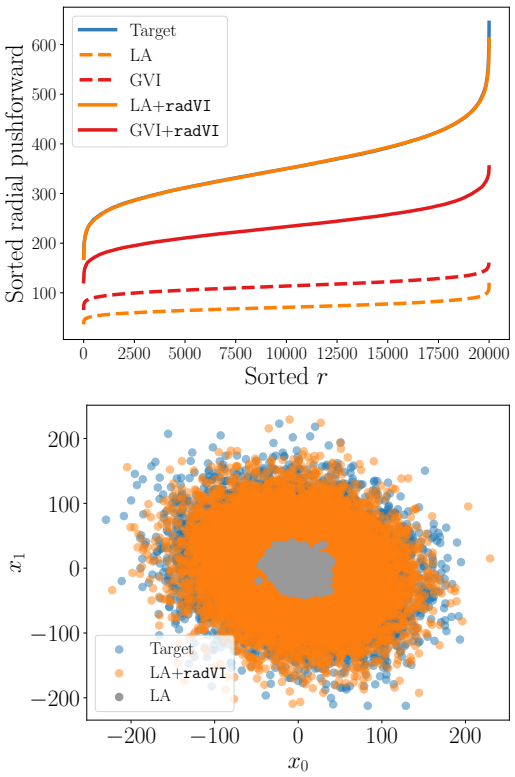

Figure 4: **Top:** Comparing whitening methods for learning the anisotropic logistic distribution, with and without `radVI`. **Bottom:** Visual comparison of true target samples, those generated by LA, and ours (LA+`radVI`).

samples as desired. Figure 4 showcases performance on an anisotropic logistic where we visualize generated samples. The Gaussian approximation methods fail to capture the correct tail behavior, while the whitened `radVI` approximations do. We observe the same performance for the anisotropic Student-t and Laplace distributions; see the bottom of Figure 2 and Appendix E.

5.3 `radVI` for parameter estimation

We now demonstrate how `radVI` can correct tail-underestimation for certain posteriors π. To illustrate this, first let $f : \mathbb{R}^d \to \mathbb{R}$ be a functional of interest and consider the problem of estimating $\mathbb{E}_{X\sim\pi}[f(X)]$.

Letting $\widehat{\pi}$ denote a generic variational approximation, note that we always have the approximation

$$\begin{aligned}\mathbb{E}_{X\sim\pi}[f(X)] &= \mathbb{E}_{Y\sim\widehat{\pi}}[f(Y)\pi(Y)/\widehat{\pi}(Y)] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}[f(Y_i)\pi(Y_i)/\widehat{\pi}(Y_i)]\,,\end{aligned}$$

where $Y_i \sim \widehat{\pi}$ are easily-drawn samples.

	$d = 25$		
Parameter	$\mathbb{E}[z^2] = 4$	$\mathbb{E}[x_1^2] \approx 7.389$	$\mathbb{P}(\lvert z\rvert > 2) \approx 0.317$
GVI	$0.274 \pm 4 \times 10^{-3}$	$1.12 \pm 1.4 \times 10^{-2}$	0
GVI+radVI	$\mathbf{2.41} \pm 2 \times 10^{-2}$	$\mathbf{5.96} \pm 1.0 \times 10^{-1}$	$\mathbf{0.214} \pm 3 \times 10^{-3}$
	$d = 50$		
Parameter	$\mathbb{E}[z^2] = 4$	$\mathbb{E}[x_1^2] \approx 7.389$	$\mathbb{P}(\lvert z\rvert > 2) \approx 0.317$
GVI	$0.328 \pm 5 \times 10^{-3}$	$1.61 \pm 2 \times 10^{-2}$	0
GVI+radVI	$\mathbf{2.51} \pm 4 \times 10^{-2}$	$\mathbf{6.48} \pm 2 \times 10^{-1}$	$\mathbf{0.19} \pm 5 \times 10^{-3}$

Table 2: In $d \in \{25, 50\}$, we compare the performance of GVI and GVI+`radVI` for parameter estimation.

To investigate how `radVI` can be used to improve GVI for parameter estimation, we study Neal's funnel distribution, a common example in the VI literature (Neal, 2003; Betancourt and Girolami, 2015). For $d > 1$, suppose $z \sim \mathcal{N}(0, 4)$ and $x_i \sim \mathcal{N}(0, e^z)$ for $i \in [d]$; one can explicitly write $\pi \in \mathcal{P}(\mathbb{R}^{d+1})$ and compute the corresponding log-density and its derivatives. We stress that this model is misspecified for radial distributions, and instead fall under the category of a *structured* posterior (Sheng et al., 2025). We follow related work and estimate the following quantities from the posterior: $\mathbb{E}[z^2] = 4$, $\mathbb{E}[x_1^2] \approx 7.389$, and $\mathbb{P}(|z| > 2) \approx 0.317$.

In Table 2, we report our results. We drew 2000 samples from both GVI and GVI+`radVI`, and reported the estimated parameters averaged over 1000 trials and also computed the standard error (averaged across the trials). As expected, the parameter estimates are significantly better when incorporating `radVI`, which we stress comes with minimal computational cost. For instance, standard Gaussian VI will report that the tail probability is identically zero whereas our estimate using `radVI` is significantly closer to the ground truth.

6 CONCLUSION

We propose and analyze a framework for variational inference over the space of *radial* transport maps, leading to our algorithm, `radVI`. Under standard assumptions, we prove convergence guarantees for our algorithm in both the deterministic and stochastic settings. Our analysis hinges on novel regularity properties of optimal transport maps between radially symmetric distributions. We demonstrate our ability to learn radial distributions in a suite of experiments where standard VI methods fail to capture the correct behavior. An interesting open question is to lift the log-concave assumptions and still obtain optimization guarantees in this setting, as many have in the sampling literature (Balasubramanian et al., 2022; Chewi, 2026).

Acknowledgements

AAP thanks the Yale Institute for the Foundations of Data Science for financial support.

References

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Arnese, M. and Lacker, D. (2024). Convergence of coordinate ascent variational inference for log-concave measures via optimal transport. *arXiv preprint arXiv:2404.08792*.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. (2022). Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR.

Barber, D. and Bishop, C. (1997). Ensemble learning for multi-layer networks. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4.

Blei, D., Kucukelbir, A., and McAuliffe, J. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112:859–877.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417.

Brent, R. P. (1973). Some efficient algorithms for solving systems of nonlinear equations. *SIAM Journal on Numerical Analysis*, 10(2):327–344.

Caffarelli, L. A. (2000). Monotonicity properties of optimal transportation and the FKG and related inequalities. *Communications in Mathematical Physics*, 214.

Chewi, S. (2026). Log-concave sampling. Available at https://chewisinho.github.io/.

Chewi, S., Niles-Weed, J., and Rigollet, P. (2025). *Statistical optimal transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham. École d'Été de Probabilités de Saint-Flour XLIX – 2019.

Chewi, S. and Pooladian, A.-A. (2023). An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *Comptes Rendus. Mathématique*, 361(G9):1471–1482.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. (2023). Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR.

Jiang, Y., Chewi, S., and Pooladian, A.-A. (2025). Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Foundations of Computational Mathematics*, pages 1–52.

Katsevich, A. (2023). The Laplace approximation accuracy in high dimensions: a refined analysis and new skew adjustment. *arXiv preprint arXiv:2306.07262*.

Katsevich, A. (2024). The Laplace asymptotic expansion in high dimensions. *arXiv preprint arXiv:2406.12706*.

Kim, K., Oh, J., Wu, K., Ma, Y., and Gardner, J. (2023). On the convergence of black-box variational inference. *Advances in Neural Information Processing Systems*, 36:44615–44657.

Lacker, D., Mukherjee, S., and Yeung, L. C. (2024). Mean field approximations via log-concavity. *International Mathematics Research Notices*, 2024:6008–6042.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35.

Lavenant, H. and Zanella, G. (2024). Convergence rate of random scan coordinate ascent variational inference under log-concavity. *SIAM Journal on Optimization*, 34(4):3750–3761.

Liang, F., Mahoney, M., and Hodgkinson, L. (2022). Fat–tailed variational inference with anisotropic tail adaptive flows. In *International Conference on Machine Learning*, pages 13257–13270. PMLR.

Margossian, C. and Saul, L. K. (2025). Variational inference in location-scale families: exact recovery of the mean and correlation matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 3466–3474. PMLR.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Comput.*, 21(3):786–792.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. Springer Nature.

Petit-Talamon, M., Lambert, M., and Korba, A. (2025). Variational inference with mixtures of isotropic Gaus-

sians. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer-Verlag.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham. Calculus of variations, PDEs, and modeling.

Seeger, M. (1999). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Sheng, S., Wu, B., Zhu, B., Chewi, S., and Pooladian, A.-A. (2025). Theory and computation for structured variational inference. *arXiv preprint arXiv:2511.09897*.

Van Handel, R. (2014). Probability in high dimension. *Lecture notes (Princeton University)*, 2(3):2–3.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

A Omitted proofs from Section 3.1

A.1 Proof of Proposition 3.2

The existence of a minimizer follows from standard arguments, since the KL divergence is weakly lower semicontinuous, has (weakly) compact sublevel sets, and $\mathcal{C}_{\mathrm{rad}}$ is (weakly) closed. Uniqueness follows from strict convexity of the KL divergence, together with convexity of $\mathcal{C}_{\mathrm{rad}}$: if $\mu, \nu \in \mathcal{C}_{\mathrm{rad}}$, then $\frac{1}{2}(\mu+\nu) \in \mathcal{C}_{\mathrm{rad}}$, since a mixture of two radial measures is radial.

A.2 Proof of Proposition 3.3

Recall that

$$\mathrm{KL}(\mu \,\|\, \pi) = \int V(y)\,\mathrm{d}\mu(y) + \int \log(\mu(y))\,\mathrm{d}\mu(y)\,,$$

where $\mu \in \mathcal{C}_{\mathrm{rad}}$. As μ is radial, we can express the above in polar coordinates as

$$s_d^{-1}\,\mathrm{KL}(\mu \,\|\, \pi) = \int_0^\infty \int_{S^{d-1}} V(r\theta)\,\mu(r)\,r^{d-1}\,\mathrm{d}\mathsf{Unif}(\theta)\,\mathrm{d}r + \int_0^\infty \log\mu(r)\,\mu(r)\,r^{d-1}\,\mathrm{d}r\,,$$

where $\mu(r) \coloneqq \mu(r\theta)$ for some (thus all) $\theta \in S^{d-1}$, and s_d denotes the volume of S^{d-1}. Taking the first variation in μ, one computes

$$\int_{S^{d-1}} V(r\theta)\,\mathrm{d}\mathsf{Unif}(\theta) + \log\mu(r) = \text{constant}\,.$$

Rearranging yields the claim.

A.3 Proof of Proposition 3.4

We first require the following computation.

Lemma A.1. *If V is L_V-smooth and ℓ_V-strongly convex, then $r \mapsto \overline{V}(r) \coloneqq \int_{S^{d-1}} V(r\theta)\,\mathrm{d}\mathsf{Unif}(\theta)$ is L_V-smooth and ℓ_V-strongly convex.*

Proof. We compute the first and second derivatives by two applications of the chain rule:

$$\overline{V}'(r) = \frac{\mathrm{d}}{\mathrm{d}r}\int_{S^{d-1}} V(r\theta)\,\mathrm{d}\mathsf{Unif}(\theta) = \int_{S^{d-1}} \langle \nabla V(r\theta), \theta\rangle\,\mathrm{d}\mathsf{Unif}(\theta)\,,$$
$$\overline{V}''(r) = \frac{\mathrm{d}}{\mathrm{d}r}\int_{S^{d-1}} \langle \nabla V(r\theta), \theta\rangle\,\mathrm{d}\mathsf{Unif}(\theta) = \int_{S^{d-1}} \langle \theta, \nabla^2 V(r\theta)\,\theta\rangle\,\mathrm{d}\mathsf{Unif}(\theta)\,.$$

As $0 \preceq \ell_V I \preceq \nabla^2 V \preceq L_V I$, the proof is complete (where we use that $\int_{S^{d-1}} \|\theta\|_2^2\,\mathrm{d}\mathsf{Unif}(\theta) = 1$). □

We now compute the Hessian of $x \mapsto \overline{V}(\|x\|)$:

$$\nabla\overline{V}(\|x\|) = \overline{V}'(\|x\|)\,\frac{x}{\|x\|}\,, \qquad \nabla\left(\overline{V}'(\|x\|)\,\frac{x}{\|x\|}\right) = \overline{V}''(\|x\|)\,\frac{xx^\top}{\|x\|^2} + \frac{\overline{V}'(\|x\|)}{\|x\|}\left(I - \frac{xx^\top}{\|x\|^2}\right).$$

By a rotation argument, we see that the Hessian can essentially be viewed as the following $d \times d$ diagonal matrix:

$$\mathrm{diag}(\overline{V}''(\|x\|), \overline{V}'(\|x\|)/\|x\|, \ldots, \overline{V}'(\|x\|)/\|x\|)\,.$$

By the fundamental theorem of calculus and since $\overline{V}'(0) = 0$ (as a consequence of $\nabla V(0) = 0$), we see that

$$\ell_V \le \overline{V}'(r)/r = (1/r)\int_0^r \overline{V}''(s)\,\mathrm{d}s \le L_V\,,$$

which concludes the proof.

B Proof of Theorem 3.5

Let $T^\star_{\mathrm{rad}}$ denote the optimal transport map from $\rho = \mathcal{N}(0, I)$ to $\pi^\star_{\mathrm{rad}}$. By design, this map should be of the form

$$T^\star_{\mathrm{rad}}(x) = \Psi^\star(\|x\|)\, x/\|x\| \coloneqq \psi^\star(\|x\|)\, x\,,$$

where ψ is some continuous, strictly monotone function.

As $\nabla^2(-\log\rho) = I$ and by Proposition 3.4, it holds by a two-sided version of Caffarelli's contraction theorem (see, e.g., Chewi and Pooladian, 2023) that

$$\frac{1}{\sqrt{L_V}}\, I \preceq DT^\star_{\mathrm{rad}}(x) \preceq \frac{1}{\sqrt{\ell_V}}\, I\,. \tag{13}$$

Computing $DT^\star_{\mathrm{rad}}$, we obtain

$$DT^\star_{\mathrm{rad}}(x) = (\Psi^\star)'(\|x\|)\, \frac{xx^\top}{\|x\|^2} + \psi^\star(\|x\|)\, (I - xx^\top/\|x\|^2)\,.$$

Combined with (13), this implies that

$$\begin{aligned} \frac{1}{\sqrt{L_V}} &\le (\Psi^\star)'(\|x\|) \le \frac{1}{\sqrt{\ell_V}}\,, \\ \frac{1}{\sqrt{L_V}} &\le \psi^\star(\|x\|) \le \frac{1}{\sqrt{\ell_V}}\,, \end{aligned} \tag{14}$$

and yields the first claim.

For the second, we use the log-Monge–Ampère equation between ρ and $\pi^\star_{\mathrm{rad}}$:

$$\log\rho(x) = \log\pi^\star_{\mathrm{rad}}(T^\star_{\mathrm{rad}}(x)) + \log\det(DT^\star_{\mathrm{rad}}(x))\,.$$

Plugging in the expressions for ρ, $T^\star_{\mathrm{rad}}$, and $\pi^\star_{\mathrm{rad}}$, one obtains

$$\overline{V}(\Psi^\star(\|x\|)) = \frac{\|x\|^2}{2} + (d-1)\log(\psi^\star(\|x\|)) + \log((\Psi^\star)'(\|x\|)) + \log c\,, \tag{15}$$

where c consists of normalizing constants. From here, we set $r \coloneqq \|x\|$ and rewrite (15), omitting the argument of $\Psi^\star$ for simplicity:

$$\overline{V} \circ \Psi^\star = \frac{r^2}{2} - (d-1)\log r + (d-1)\log\Psi^\star + \log(\Psi^\star)' + \log c\,. \tag{16}$$

Differentiating in r, one obtains

$$\frac{(\Psi^\star)''}{(\Psi^\star)'} = -r + \frac{d-1}{r} - (d-1)\,\frac{(\Psi^\star)'}{\Psi^\star} + \overline{V}'(\Psi^\star)\,(\Psi^\star)'\,.$$

Define the two functions

$$F_\rho(r) \coloneqq \frac{r^2}{2} - (d-1)\log r\,, \qquad F_\star(r) \coloneqq \overline{V}(r) - (d-1)\log r\,.$$

Let r_ρ, $r_\star$ denote the minimizers of F_ρ and $F_\star$ respectively; thus,

$$r_\rho = \sqrt{d-1}\,, \qquad \overline{V}'(r_\star) - \frac{d-1}{r_\star} = 0\,.$$

The intuition is that under ρ, the norm is concentrated around r_ρ, and similarly, under $\pi^\star_{\mathrm{rad}}$, the norm is concentrated around $r_\star$. We therefore expand (16) around $r \approx r_\rho$, $\Psi^\star(r) \approx r_\star$. We start by noting that

$$F''_\rho(r) = 1 + \frac{d-1}{r^2}\,, \qquad F''_\star(r) = \overline{V}''(r) + \frac{d-1}{r^2}\,.$$

Taylor expansion, together with $F_\rho'(r_\rho) = 0$, shows that

$$|F_\rho'(r)| = \Big|\int_{r_\rho}^{r} F_\rho''(s)\,\mathrm{d}s\Big| \le \Big(1 + \frac{d-1}{(r \wedge r_\rho)^2}\Big)\,|r - r_\rho| \lesssim \big(1 + \frac{d}{r^2}\big)\,|r - r_\rho|\,. \tag{17}$$

A similar argument yields

$$|F_\star'(r)| = \Big|\int_{r_\star}^{r} F_\star''(s)\,\mathrm{d}s\Big| \le \Big(L_V + \frac{d-1}{(r \wedge r_\star)^2}\Big)\,|r - r_\star|\,.$$

To simplify, we use

$$r_\star = \frac{d-1}{\overline{V}'(r_\star)} \ge \frac{d-1}{L_V r_\star}\,,$$

so $r_\star^2 \ge (d-1)/L_V$. Therefore,

$$|F_\star'(r)| \lesssim \big(L_V + \frac{d}{r^2}\big)\,|r - r_\star|\,. \tag{18}$$

Substituting (17) and (18) into (16), we see that

$$\frac{|(\Psi^\star)''|}{(\Psi^\star)'} \le |F_\rho'(r)| + |(\Psi^\star)'|\,|F_\star'(\Psi^\star)| \lesssim \big(1 + \frac{d}{r^2}\big)\,|r - r_\rho| + |(\Psi^\star)'|\,\big(L_V + \frac{d}{(\Psi^\star)^2}\big)\,|\Psi^\star - r_\star|\,. \tag{19}$$

The next step is to show that $\Psi^\star(r_\rho) \approx r_\star$, i.e., that the map $\Psi^\star$ approximately pushes forward the mode of the radial part of ρ to the radial part of $\pi^\star_{\mathrm{rad}}$.

Lemma B.1.

$$|\Psi^\star(r_\rho) - r_\star| \le \frac{2}{\sqrt{\ell_V}}\,.$$

Proof. We use the fact that for any α-strongly log-concave distribution π in dimension d with mode $x_\star$, it holds that $\int \|x - x_\star\|^2\,\mathrm{d}\pi(x) \le d/\alpha$ (c.f. Chewi, 2026, "basic lemma"). Let $X_\rho \sim \rho$, so that $\Psi^\star(\|X_\rho\|)$ is distributed according to the radial part of $\pi^\star_{\mathrm{rad}}$. The radial parts of ρ and of $\pi^\star_{\mathrm{rad}}$ are one-dimensional distributions, with potentials F_ρ and $F_\star$ respectively; this implies that they are both strongly log-concave, with respective parameters 1 and ℓ_V. Therefore,

$$\begin{aligned}
|\Psi^\star(r_\rho) - r_\star| &\le \mathbb{E}|\Psi^\star(r_\rho) - \Psi^\star(\|X_\rho\|)| + \mathbb{E}|\Psi^\star(\|X_\rho\|) - r_\star| \\
&\le \frac{1}{\sqrt{\ell_V}}\,\mathbb{E}\big|\,r_\rho - \|X_\rho\|\,\big| + \mathbb{E}|\Psi^\star(\|X_\rho\|) - r_\star| \\
&\le \frac{1}{\sqrt{\ell_V}} + \frac{1}{\sqrt{\ell_V}} = \frac{2}{\sqrt{\ell_V}}\,.
\end{aligned}$$

□

Continuing from (19), since $\Psi^\star \ge r/\sqrt{L_V}$,

$$\begin{aligned}
\frac{|(\Psi^\star)''|}{(\Psi^\star)'} &\lesssim \big(1 + \frac{d}{r^2}\big)\,|r - r_\rho| + L_V\,|(\Psi^\star)'|\,\big(1 + \frac{d}{r^2}\big)\,\big(|\Psi^\star - \Psi^\star(r_\rho)| + |\Psi^\star(r_\rho) - r_\star|\big) \\
&\lesssim \big(1 + \frac{d}{r^2}\big)\,\Big(|r - r_\rho| + L_V\,|(\Psi^\star)'|\,\big(\frac{|r - r_\rho|}{\sqrt{\ell_V}} + \frac{1}{\sqrt{\ell_V}}\big)\Big) \lesssim \kappa\,\big(1 + \frac{d}{r^2}\big)\,(1 + |r - r_\rho|)\,.
\end{aligned}$$

This proves the estimate

$$|(\Psi^\star)''| \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,\big(1 + \frac{d}{r^2}\big)\,(1 + |r - r_\rho|) \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,\big(1 + \frac{d}{r^2}\big)\,(1 + |r - \sqrt{d}|)\,.$$

C Omitted proofs from Section 4

C.1 Proof of Theorem 4.1

Our goal is to find a map $\hat{T} = T_{\hat{\lambda}}$ such that

$$\|T^\star - \hat{T}\|^2_{L^2(\rho)} \le \frac{\varepsilon^2}{\ell_V} \qquad \text{and} \qquad \|D(T^\star - \hat{T})\|^2_{L^2(\rho)} \le \frac{\varepsilon_1^2}{\ell_V}\,. \tag{20}$$

We will closely follow the proof strategy put forth by Jiang et al. (2025) with appropriate modifications throughout.

Step 1: Reformulate the problem in terms of Ψ. Let us write $\hat{\Psi}$ for the radial part of $\hat{T}$, i.e., $\|\hat{T}(x)\| = \hat{\Psi}(\|x\|)$. Immediately, we notice the following simplification can be made:

$$\|T^\star - \hat{T}\|^2_{L^2(\rho)} = \int \|T^\star(x) - \hat{T}(x)\|^2 \,\mathrm{d}\rho(x) = \int (\Psi^\star(\|x\|) - \hat{\Psi}(\|x\|))^2 \,\mathrm{d}\rho(x) = \|\Psi^\star - \hat{\Psi}\|^2_{L^2(\widetilde{\rho})}\,,$$

where $\widetilde{\rho} \coloneqq \mathrm{Law}(\|X\|)$ for $X \sim \rho = \mathcal{N}(0, I)$. Similarly,

$$\begin{aligned}
\|D(T^\star - \hat{T})\|^2_{L^2(\rho)} &= \int \Big\| (\Psi^\star - \hat{\Psi})'(\|x\|)\,\frac{xx^\top}{\|x\|^2} + (\psi^\star - \hat{\psi})(\|x\|)\,\big(I - \frac{xx^\top}{\|x\|^2}\big)\Big\|^2_{\mathrm{F}} \,\mathrm{d}\rho(x) \\
&\lesssim \int \big\{ |(\Psi^\star - \hat{\Psi})'(\|x\|)|^2 + d\,(\psi^\star - \hat{\psi})^2(\|x\|)\big\} \,\mathrm{d}\rho(x) \\
&\lesssim \int \big\{ |(\Psi^\star - \hat{\Psi})'(r)|^2 + d\,(\psi^\star - \hat{\psi})^2(r)\big\} \,\mathrm{d}\widetilde{\rho}(r) = \|(\Psi^\star - \hat{\Psi})'\|^2_{L^2(\widetilde{\rho})} + d\,\|\psi^\star - \hat{\psi}\|^2_{L^2(\widetilde{\rho})}\,.
\end{aligned}$$

From (14), we know that $1/\sqrt{L_V} \le (\Psi^\star)' \le 1/\sqrt{\ell_V}$.

Step 2: Remove the strongly convex part. Recall that by definition of $\mathcal{T}_J$ and $\alpha = 1/\sqrt{L_V}$,

$$\hat{\Psi}(r) = \frac{r}{\sqrt{L_V}} + \underbrace{\sum_{j=0}^{J} \hat{\lambda}_j \Psi_j(r)}_{=:\hat{\Psi}_\diamond(r)}\,.$$

Let us also write $\Psi^\star(r) \coloneqq r/\sqrt{L_V} + \Psi^\star_\diamond(r)$. Then, $\|\Psi^\star - \hat{\Psi}\|^2_{L^2(\widetilde{\rho})} = \|\Psi^\star_\diamond - \hat{\Psi}_\diamond\|^2_{L^2(\widetilde{\rho})}$, etc. Hence, our goal is equivalent to approximating $\Psi^\star_\diamond$ using a function of the form $\sum_{j=0}^J \hat{\lambda}_j \Psi_j$, where we know that $\Psi^\star_\diamond$ satisfies the bounds $0 \le (\Psi^\star_\diamond)' \le 1/\sqrt{\ell_V} - 1/\sqrt{L_V}$. For simplicity, we will replace the upper bound on $(\Psi^\star_\diamond)'$ by $1/\sqrt{\ell_V}$, which only makes our problem more difficult. Having reformulated our goal, we simply write $\hat{\Psi}$ for $\hat{\Psi}_\diamond$ and $\Psi^\star$ for $\Psi^\star_\diamond$ in the remaining steps to keep the notation concise.

Step 3: Truncate. Next, we consider the interval $\mathcal{I} \coloneqq [\sqrt{d} - R, \sqrt{d} + R]$, for some $R > 0$ that will be chosen later, and equally partition said interval into subintervals of length $\delta > 0$. This defines the dictionary $\{\Psi_j\}_{j=0}^J$. Next, our construction will ensure that $\hat{\Psi}(\sqrt{d} - R) = \Psi^\star(\sqrt{d} - R)$ and similarly $\hat{\Psi}(\sqrt{d} + R) = \Psi^\star(\sqrt{d} + R)$—outside this interval, the function $\hat{\Psi}$ is a constant. We want to bound

$$\|\Psi^\star - \hat{\Psi}\|^2_{L^2(\widetilde{\rho})} = (\mathsf{T1}) + (\mathsf{T2}) \coloneqq \int_{\mathcal{I}} |\Psi^\star(r) - \hat{\Psi}(r)|^2 \,\mathrm{d}\widetilde{\rho}(r) + \int_{\mathbb{R}\setminus\mathcal{I}} |\Psi^\star(r) - \hat{\Psi}(r)|^2 \,\mathrm{d}\widetilde{\rho}(r)\,.$$

We require the following lemma.

Lemma C.1. *For $r \notin \mathcal{I}$, then*

$$|\Psi^\star(r) - \hat{\Psi}(r)| \le |r - \sqrt{d}|/\sqrt{\ell_V}\,.$$

Proof. First take $r \ge \sqrt{d} + R$. Then, as $\Psi^\star, \hat{\Psi} \ge 0$ and by (14),

$$|\Psi^\star(r) - \hat{\Psi}(r)| = |\Psi^\star(r) - \Psi^\star(\sqrt{d} + R)| = \Psi^\star(r) - \Psi^\star(\sqrt{d} + R) \le (r - \sqrt{d} - R)/\sqrt{\ell_V} \le (r - \sqrt{d})/\sqrt{\ell_V}\,.$$

A symmetric argument completes the claim. □

Now, we bound via Cauchy–Schwarz,

$$(\mathsf{T2}) = \int_{\mathbb{R}\setminus\mathcal{I}} |\Psi^\star(r) - \hat{\Psi}(r)|^2 \,\mathrm{d}\widetilde{\rho}(r) \le \ell_V^{-1} \int_{\mathbb{R}\setminus\mathcal{I}} (r - \sqrt{d})^2 \,\mathrm{d}\widetilde{\rho}(r) \le \ell_V^{-1} \sqrt{\mathbb{E}\big[(\|X\| - \sqrt{d})^2\big]\,\mathbb{P}(\|X\| \notin \mathcal{I})}$$

where $X \sim \rho$. Using standard tools from high-dimensional probability pertaining to Gaussian tail bounds (Van Handel, 2014), the right-hand side is bounded by $O(\ell_V^{-1}\exp(-R^2/2))$. Therefore, if $R \gtrsim \sqrt{\log(1/\varepsilon)}$, then $(\mathsf{T1}) \le \varepsilon^2/\ell_V$.

For $r \notin \mathcal{I}$, $\hat{\Psi}'(r) = 0$, and $|(\Psi^\star)'| \le 1/\sqrt{\ell_V}$. Thus,

$$\int_{\mathbb{R}\setminus\mathcal{I}} |(\Psi^\star - \hat{\Psi})'|^2 \,\mathrm{d}\widetilde{\rho} \le \frac{1}{\ell_V}\,\mathbb{P}(\|X\| \notin \mathcal{I}) \lesssim \frac{\exp(-R^2/2)}{\ell_V}\,.$$

Next, for $r \le \sqrt{d} - R$, $\hat{\Psi}(r) = \Psi^\star(\sqrt{d} - R)$, and since $\psi^\star \le 1/\sqrt{\ell_V}$ everywhere,

$$\begin{aligned}
d\int_0^{\sqrt{d}-R} (\psi^\star - \hat{\psi})^2 &\lesssim d\int_0^{\sqrt{d}-R} \Big(\frac{\Psi^\star(\sqrt{d}-R)^2}{r^2} + \frac{1}{\ell_V}\Big)\,\mathrm{d}\widetilde{\rho}(r) \\
&\le d\Psi^\star(\sqrt{d}-R)^2\,\mathbb{E}[\|X\|^{-2}\,\mathbf{1}_{\|X\|\notin\mathcal{I}}] + \frac{d}{\ell_V}\,\mathbb{P}(\|X\| \notin \mathcal{I}) \\
&\le d\Psi^\star(\sqrt{d}-R)^2\,\mathbb{E}[\|X\|^{-2.5}]^{4/5}\,\mathbb{P}(\|X\| \notin \mathcal{I})^{1/5} + \frac{d}{\ell_V}\,\mathbb{P}(\|X\| \notin \mathcal{I}) \\
&\lesssim d\,\frac{\Psi^\star(\sqrt{d}-R)^2}{d}\,\mathbb{P}(\|X\| \notin \mathcal{I})^{1/5} + \frac{d}{\ell_V}\,\mathbb{P}(\|X\| \notin \mathcal{I}) \\
&\lesssim \frac{d}{\ell_V}\exp(-R^2/10)\,,
\end{aligned}$$

where we used $d \ge 3$. A similar argument controls the integral over $[\sqrt{d}+R, \infty)$. Hence, $R \gtrsim \sqrt{\log(d/\varepsilon)}$ implies $\int_{\|x\|\notin\mathcal{I}} \|D(T^\star - \hat{T})(x)\|_{\mathrm{F}}^2 \,\mathrm{d}\rho(x) \lesssim \varepsilon/\ell_V$.

Step 4: Control the approximation error on $\mathcal{I}$. Recall that our construction should satisfy $\hat{\Psi}(\sqrt{d}-R) = \Psi^\star(\sqrt{d}-R)$ and $\hat{\Psi}(\sqrt{d}+R) = \Psi^\star(\sqrt{d}+R)$. Since $\hat{\Psi}(\sqrt{d}-R) = \hat{\lambda}_0\Psi_0(\sqrt{d}-R) = \hat{\lambda}_0$, the first condition amounts to setting $\hat{\lambda}_0 = \Psi^\star(\sqrt{d}-R)$. For the second condition, we will in fact ensure that $\hat{\Psi}$ agrees with $\Psi^\star$ at the endpoints of every sub-interval $[a, a+\delta]$, for each knot a.

We now turn toward bounding $(\mathsf{T1})$. To do so, consider a sub-interval $[a, a+\delta]$ on which $\hat{\Psi}$ is affine. Since the graph of $\hat{\Psi}$ interpolates the points $(a, \Psi^\star(a))$ and $(a+\delta, \Psi^\star(a+\delta))$,

$$\hat{\Psi}(r) = \Psi^\star(a) + \delta^{-1}\,(\Psi^\star(a+\delta) - \Psi^\star(a))\,(r-a)\,.$$

On the other hand, by two applications of the mean value theorem, we have

$$\Psi^\star(r) = \Psi^\star(a) + (\Psi^\star)'(c_1)\,(r-a)\,, \quad \Psi^\star(a+\delta) = \Psi^\star(a) + (\Psi^\star)'(c_2)\,\delta\,,$$

for $c_1, c_2 \in [a, a+\delta]$. Thus,

$$|\Psi^\star(r) - \hat{\Psi}(r)| = |((\Psi^\star)'(c_1) - (\Psi^\star)'(c_2))\,(r-a)|\,.$$

Applying the mean value theorem again (as $\Psi^\star$ has two bounded derivatives), we have that

$$(\Psi^\star)'(c_1) - (\Psi^\star)'(c_2) = (\Psi^\star)''(c_3)\,(c_1 - c_2)\,,$$

for some $c_3 \in [c_1, c_2]$, and thus we have a bound

$$|\Psi^\star(r) - \hat{\Psi}(r)| = |(\Psi^\star)''(c_3)|\,|c_1 - c_2|\,|r-a| \le |(\Psi^\star)''(c_3)|\,\delta^2\,.$$

Using the gradient bound on $(\Psi^\star)''$ from Theorem 3.5, we obtain

$$|\Psi^\star(r) - \hat{\Psi}(r)| \lesssim \frac{\kappa\delta^2}{\sqrt{\ell_V}} \sup_{\xi\in[a,a+\delta]} \Big(1 + \frac{d}{\xi^2}\Big)\,(1 + |\xi - \sqrt{d}|) \lesssim \frac{\kappa\delta^2 R}{\sqrt{\ell_V}}\,,$$

provided $R \ll \sqrt{d}$. And so, over all of $\mathcal{I}$, this becomes

$$\sup_{r\in\mathcal{I}} |\Psi^\star(r) - \hat{\Psi}(r)| \lesssim \frac{\kappa\delta^2 R}{\sqrt{\ell_V}}\,.$$

We now prove the uniform bound for the gradient difference $D(T^\star - \hat{T})$. Since $\hat{\Psi}'(r) = (\Psi^\star(a+\delta) - \Psi^\star(a))/\delta = (\Psi^\star)'(c_2)$, it follows that

$$|(\Psi^\star - \hat{\Psi})'(r)| = |(\Psi^\star)'(r) - (\Psi^\star)'(c_2)| \lesssim \frac{\kappa\delta R}{\sqrt{\ell_V}}\,.$$

Similarly,

$$|\psi^\star(r) - \hat{\psi}(r)| = \frac{|\Psi^\star(r) - \hat{\Psi}(r)|}{r} \lesssim \frac{\kappa\delta^2 R}{\sqrt{d\ell_V}}\,.$$

Consequently, if we choose $\delta \asymp \varepsilon^{1/2}/(\kappa^{1/2}R^{1/2})$, then we can ensure

$$\int_{\|x\|\in\mathcal{I}} \|(T^\star - \hat{T})(x)\|^2 \,\mathrm{d}\rho(x) \lesssim \frac{\varepsilon^2}{\ell_V}\,, \qquad \int_{\|x\|\in\mathcal{I}} \|D(T^\star - \hat{T})(x)\|_{\mathrm{F}}^2 \,\mathrm{d}\rho(x) \lesssim \frac{\kappa R\varepsilon}{\ell_V}\,.$$

Step 5: Complete the proof. Combining all of the bounds, we set $R \asymp \sqrt{\log(d/\varepsilon)}$, $\delta \asymp \sqrt{\varepsilon/(\kappa\log(d/\varepsilon))}$. Since we require $R \ll \sqrt{d}$, this requires $d \gg \log(1/\varepsilon)$. With these choices, we can ensure

$$\|T^\star - \hat{T}\|_{L^2(\rho)}^2 \le \frac{\varepsilon^2}{\ell_V}\,, \qquad \|D(T^\star - \hat{T})\|_{L^2(\rho)}^2 \lesssim \frac{\kappa\varepsilon\sqrt{\log(d/\varepsilon)}}{\ell_V}\,.$$

This completes the proof.

C.2 Construction of the Gram matrix

In this section, we describe how to compute the Gram matrix $Q \in \mathbb{S}_+^{J+1}$ explicitly, where we recall that

$$Q_{i,j} \coloneqq \mathbb{E}_{X\sim\rho}[\Psi_i(\|X\|)\,\Psi_j(\|X\|)] = \int_0^\infty \Psi_i(r)\,\Psi_j(r)\,\mathrm{d}\widetilde{\rho}(r)\,,$$

where

$$\mathrm{d}\widetilde{\rho}(r) = \frac{1}{2^{d/2-1}\Gamma(d/2)} r^{d-1} e^{-r^2/2}\,\mathrm{d}r\,, \tag{21}$$

which is the distribution of $r = \|X\|$ for $X \sim \mathcal{N}(0, I)$ (also known as the chi distribution). Indeed, as as we'll see below, the piecewise linear nature of $\{\Psi_j\}_{j=0}^J$ will allow us to compute $Q_{i,j}$. To this end, we also require the following object, the n-th truncated moment of $\widetilde{\rho}$ over an interval $[a,b]$:

$$\mathcal{M}_n(a,b) \coloneqq \int_a^b r^n\,\mathrm{d}\widetilde{\rho}(r) = \frac{2^{n/2}}{\Gamma(d/2)}\left[\Gamma\left(\frac{n+d}{2}, \frac{a^2}{2}\right) - \Gamma\left(\frac{n+d}{2}, \frac{b^2}{2}\right)\right]\,,$$

where $\Gamma(s,x)$ is the upper incomplete gamma function.

To compute $Q_{i,j}$ for $i \le j$ (assuming $a_i \le a_j$), we decompose the integral based on the support of the basis functions. Namely, we consider a small sub-interval $[a_i, a_i+\delta]$ for Ψ_i and $[a_j, a_j+\delta]$ for Ψ_j.

Case 1: Disjoint Ramps ($a_j \ge a_i + \delta$). The rising part of Ψ_j starts after Ψ_i has already reached its plateau of 1. The integral splits into the ramp region of Ψ_j and the region where both functions are unity:

$$\begin{aligned} Q_{i,j} &= \int_{a_j}^{a_j+\delta} (1)\cdot\left(\frac{r-a_j}{\delta}\right)\mathrm{d}\widetilde{\rho}(r) + \int_{a_j+\delta}^\infty (1)\cdot(1)\,\mathrm{d}\widetilde{\rho}(r) \\ &= \frac{1}{\delta}\left[\mathcal{M}_1(a_j, a_j+\delta) - a_j\mathcal{M}_0(a_j, a_j+\delta)\right] + \mathcal{M}_0(a_j+\delta, \infty)\,. \end{aligned}$$

Case 2: Overlapping Ramps ($a_j < a_i + \delta$). In this setting, the domain of integration consists of three non-zero overlapping regions: (1) where both are ramps, (2) where Ψ_i saturates but Ψ_j is still rising, and (3) where both saturate:

$$Q_{i,j} = \int_{a_j}^{a_i+\delta} \left(\frac{r-a_i}{\delta}\right)\left(\frac{r-a_j}{\delta}\right) \mathrm{d}\widetilde{\rho}(r) + \int_{a_i+\delta}^{a_j+\delta} (1)\cdot\left(\frac{r-a_j}{\delta}\right) \mathrm{d}\widetilde{\rho}(r) + \int_{a_j+\delta}^{\infty} (1)\cdot(1)\,\mathrm{d}\widetilde{\rho}(r)\,.$$

Expanding the quadratic term in the first integral allows the entire expression to be computed as a sum of weighted moments $\mathcal{M}_n$.

C.3 Proof of Proposition 4.4

We follow the arguments of Jiang et al. (2025). By Jiang et al. (2025, Propositions 5.10 and 5.11), the objective $\lambda \mapsto \mathcal{F}(T_\lambda)$ is ℓ_V-strongly convex and $L_V\,(1+\Upsilon)$ smooth in the $\|\cdot\|_Q$ norm, where $\Upsilon > 0$ is the smallest positive constant such that $\|DT_\lambda - \alpha I\|^2_{L^2(\rho)} \le \Upsilon \,\|T_\lambda - \alpha\,\mathrm{id}\|^2_{L^2(\rho)}$ for all $\lambda \in \mathbb{R}^{J+1}$. Writing this out, we need

$$\begin{aligned}
&\int \Big\| \sum_{j=0}^{J} \lambda_j \left(\Psi_j'(\|x\|)\,\frac{xx^\top}{\|x\|^2} + \psi_j(\|x\|)\,\big(I - \frac{xx^\top}{\|x\|^2}\big)\right)\Big\|_{\mathrm{F}}^2 \mathrm{d}\rho(x) \\
&\quad = \int \Big\{ \Big(\sum_{j=0}^{J} \lambda_j \Psi_j'(\|x\|)\Big)^2 + (d-1)\Big(\sum_{j=0}^{J} \lambda_j \psi_j(\|x\|)\Big)^2 \Big\}\,\mathrm{d}\rho(x) \\
&\quad = \int \Big\{ \Big(\sum_{j=0}^{J} \lambda_j \Psi_j'(r)\Big)^2 + (d-1)\Big(\sum_{j=0}^{J} \lambda_j \psi_j(r)\Big)^2 \Big\}\,\mathrm{d}\widetilde{\rho}(r)
\end{aligned}$$

to be bounded by

$$\Upsilon \int \Big(\sum_{j=0}^{J} \lambda_j \Psi_j(r)\Big)^2 \mathrm{d}\widetilde{\rho}(r)\,.$$

It suffices to verify

$$\int_{a_k}^{a_k+\delta_k} \Big\{ \Big(\sum_{j=0}^{J} \lambda_j \Psi_j'(r)\Big)^2 + (d-1)\Big(\sum_{j=0}^{J} \lambda_j \psi_j(r)\Big)^2 \Big\}\,\mathrm{d}\widetilde{\rho}(r) \le \Upsilon \int_{a_k}^{a_k+\delta_k} \Big(\sum_{j=0}^{J} \lambda_j \Psi_j(r)\Big)^2 \mathrm{d}\widetilde{\rho}(r)$$

for each $k = 0, 1, \ldots, J$ separately. If we let $\bar{\Psi}_\lambda(\|x\|) \coloneqq \|T_\lambda(x) - \alpha x\|$, this reduces to

$$\int_{a_k}^{a_k+\delta_k} \Big\{ \underbrace{\lambda_k^2 \delta_k^{-2}}_{\mathsf{T}_1} + \underbrace{\frac{d-1}{r^2}\left(\bar{\Psi}_\lambda(a_k) + \lambda_k \delta_k^{-1}\,(r-a_k)\right)^2}_{\mathsf{T}_2} \Big\}\,\mathrm{d}\widetilde{\rho}(r) \le \Upsilon \int_{a_k}^{a_k+\delta_k} \left(\bar{\Psi}_\lambda(a_k) + \lambda_k \delta_k^{-1}\,(r-a_k)\right)^2 \mathrm{d}\widetilde{\rho}(r)\,.$$

Let us start with the case of $k \ge 1$. In this case, $r \gtrsim \sqrt{d}$, so the term T_2 is bounded, up to an absolute constant, by the integrand on the right-hand side. Thus, for term T_1, it suffices to prove $\int_{a_k}^{a_k+\delta} \mathrm{d}\widetilde{\rho}(r) \lesssim \Upsilon \inf_{\bar{r}\in\mathbb{R}} \int_{a_k}^{a_k+\delta} (r-\bar{r})^2\,\mathrm{d}\widetilde{\rho}(r)$, i.e., we need a lower bound on the variance of the distribution $\widetilde{\rho}$ restricted to the interval $[a_k, a_k+\delta]$. Note that $\widetilde{\rho} \propto \exp(-F_\rho)$ where F_ρ is defined in the proof of Theorem 3.5. Using the estimates from that proof, we know that F_ρ is $O(R)$-Lipschitz on the interval $[a_k, a_k+\delta]$, hence $\log\widetilde{\rho}$ only varies by $O(1)$ on this interval provided $\delta \lesssim 1/R$. The argument of Jiang et al. (2025, Lemma 5.13) now shows that the variance is lower bounded by $\Omega(\delta^2)$, and hence our desired estimate holds for $\Upsilon \asymp \delta^{-2}$.

Next, consider the case $k = 0$, so that $a_0 = 0$ and $\delta_0 = \sqrt{d} - R$. For the term T_2, we must prove $(d-1)\int_0^{\delta_0} \mathrm{d}\widetilde{\rho}(r) \lesssim \Upsilon \int_0^{\delta_0} r^2\,\mathrm{d}\widetilde{\rho}(r)$, which holds with $\Upsilon \asymp 1$: indeed, this follows from the fact that $\widetilde{\rho}|_{[0,\delta_0]}$ is 1-strongly log-concave with mode at $\delta_0 \asymp \sqrt{d}$, so $\int (r-\delta_0)^2\,\mathrm{d}\widetilde{\rho}|_{[0,\delta_0]}(r) \le 1$. The term T_1 is similar but easier.

All in all, this shows that we can take $\Upsilon \lesssim \delta^{-2} = \widetilde{\Theta}(J^2)$.

C.4 Proof of Theorem 4.3

Recall that for $\lambda \mapsto \mathcal{F}(T_\lambda)$, the smoothness constant is $\widetilde{O}(J^2 L_V)$. Choosing $h = 1/\widetilde{\Theta}(L_V J^2)$, we have by standard arguments for projected gradient descent for smooth and strongly convex functions (Beck, 2017, Theorem 10.29)

$$\|T_{\lambda^{(k)}} - T_{\lambda^\star}\|^2_{L^2(\rho)} \coloneqq \|\lambda^{(k)} - \lambda^\star\|^2_Q \le \varepsilon^2/\ell_V$$

so long as $k \gtrsim_{\log} \kappa J^2 \log(\mathrm{KL}((T_{\lambda^{(0)}})_\sharp \rho \,\|\, \pi)/\varepsilon^2)$. After two applications of triangle inequality, and invoking (9), we arrive at

$$\begin{aligned}
\|T_{\lambda^{(k)}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} &\lesssim \|T_{\lambda^{(k)}} - T_{\lambda^\star}\|^2_{L^2(\rho)} + \|T_{\lambda^\star} - T_{\hat\lambda}\|^2_{L^2(\rho)} + \|T_{\hat\lambda} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \\
&\lesssim \|\lambda^{(k)} - \lambda^\star\|^2_Q + \kappa \|T_{\hat\lambda} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} + \kappa^2 \|D(T_{\hat\lambda} - T^\star_{\mathrm{rad}})\|^2_{L^2(\rho)} \\
&\lesssim \varepsilon^2/\ell_V + \kappa \|T_{\hat\lambda} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} + \kappa^2 \|D(T_{\hat\lambda} - T^\star_{\mathrm{rad}})\|^2_{L^2(\rho)} .
\end{aligned}$$

For $\varepsilon_1 > 0$, Theorem 4.1 states that for $J = \widetilde{\Omega}(\sqrt{\kappa/\varepsilon_1})$, then we have

$$\|T_{\lambda^{(k)}} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \lesssim_{\log} \varepsilon^2/\ell_V + \kappa\varepsilon_1^2/\ell_V + \kappa^3\varepsilon_1/\ell_V .$$

Choosing $\varepsilon_1 \asymp_{\log} \varepsilon^2/\kappa^3$ (in the worst case), we conclude that the error tolerance is ε^2/ℓ_V. Then, $J = \widetilde{\Theta}(\kappa^2/\varepsilon)$, we have that the number of iterations required scales like $\kappa^5\varepsilon^{-1}$ up to logarithmic factors, and the required stepsize is $\varepsilon/(L_V\kappa^2)$ up to logarithmic factors.

C.5 Gradient of the objective

In the stochastic projected gradient descent algorithm, we require the gradient with respect to λ of

$$\mathcal{F}(T_\lambda) = \int V(T_\lambda(x))\,\mathrm{d}\rho(x) - \int \log\det DT_\lambda(x)\,\mathrm{d}\rho(x) .$$

Passing the gradient under the integral, we need to compute

$$\nabla_\lambda \mathcal{F}(T_\lambda) = \int \nabla_\lambda V(T_\lambda(x))\,\mathrm{d}\rho(x) - \int \nabla_\lambda \log\det DT_\lambda(x)\,\mathrm{d}\rho(x) \tag{22}$$

We first compute the gradient of the potential energy. Writing $r = \|x\|$ and letting $\{\Psi_j\}_{j=0}^J$ denote our basis, the chain rule gives for each coordinate $j \in \{0, 1, \dots, J\}$

$$\nabla_{\lambda_j} V(T_\lambda(x)) = \Psi_j(r)\langle x/r, \nabla V\big(T_\lambda(x)\big)\rangle.$$

Thus, the full gradient becomes

$$\nabla_\lambda V(T_\lambda(x)) = \vec{\Psi}(r)\langle x/r, \nabla V(T_\lambda(x))\rangle, \tag{23}$$

where $\vec{\Psi}(r) = (\Psi_0(r), \dots, \Psi_J(r)) \in \mathbb{R}_+^{J+1}$. For this term, we approximate the full gradient via Monte Carlo approximation, i.e.,

$$\nabla_\lambda \mathbb{E}_{X\sim\rho}[V(T_\lambda(X))] \simeq \frac{1}{n}\sum_{i=1}^n \vec{\Psi}(\|X_i\|)\langle X_i/\|X_i\|, \nabla V(T_\lambda(X_i))\rangle,$$

where $X_1, \dots, X_n \sim \rho$.

We next compute the gradient of the log-determinant term. For any $\lambda \in \mathbb{R}_+^{J+1}$, one can compute

$$DT_\lambda(x) = \frac{\alpha r + \sum_{j=0}^J \lambda_j \Psi_j(r)}{r} I + \Big(\sum_{j=0}^J \lambda_j \Psi_j'(r) - \frac{1}{r}\sum_{j=0}^J \lambda_j \Psi_j(r)\Big)\frac{xx^\top}{r^2} ,$$

and, upon rearranging

$$DT_\lambda(x) = \Big(\alpha + \sum_{j=0}^J \lambda_j \Psi_j'(r)\Big) xx^\top/r^2 + \Big(\alpha + \sum_{j=0}^J \lambda_j \Psi_j(r)/r\Big)(I - xx^\top/r^2) .$$

The eigenvalue in the radial direction x/r equals

$$\alpha + \sum_{j=0}^{J} \lambda_j \Psi_j'(r)\,,$$

while each of the remainig $d-1$ orthogonal directions has eigenvalue

$$\alpha + \sum_{j=0}^{J} \lambda_j \Psi_j(r)/r\,.$$

Accounting for the multiplicity of these eigenvalues, one arrives at

$$\log\det DT_\lambda(x) = (d-1)\log\big(\alpha + \langle\lambda, \vec{\Psi}(r)/r\rangle\big) + \log\big(\alpha + \langle\lambda, \vec{\Psi}'(r)\rangle\big)\,.$$

Writing $g_\lambda(r) = \alpha r + \langle\lambda, \Psi(r)\rangle$ and thus $g_\lambda'(r) = \alpha + \langle\lambda, \Psi'(r)\rangle$, differentiating with respect to λ yields

$$\nabla_{\lambda_j} \log\det DT_\lambda(x) = \frac{(d-1)\,\Psi_j(r)}{g_\lambda(r)} + \frac{\Psi_j'(r)}{g_\lambda'(r)}\,. \tag{24}$$

From (24) we have

$$\nabla_{\lambda_j}\, \mathbb{E}_{X\sim\rho}[\log\det DT_\lambda(X)] = \int_0^\infty \Big[\frac{(d-1)\,\Psi_j(r)}{g_\lambda(r)} + \frac{\Psi_j'(r)}{g_\lambda'(r)}\Big]\,\mathrm{d}\widetilde{\rho}(r) \tag{25}$$

where $\widetilde{\rho}$ is the radial law of $\|X\|$ under $\rho = \mathcal{N}(0, I)$. There are two approaches to computing this integral. One approach is via Monte Carlo. Below, we describe how to efficiently evaluate the integral deterministically for any $j \in \{0, \dots, J\}$.

To this end, define the sets

$$I_0 \coloneqq [0, a_1], \quad I_\ell \coloneqq [a_\ell, a_\ell + \delta]$$

for $\ell = 1, \dots, J$, and let $I_{J+1} = \{r : r \geq a_J + \delta\}$. Thus, we can partition the support of $\widetilde{\rho}$ as $[0,\infty) = \bigcup_{j=0}^{J+1} I_j$.

First, notice that $\Psi_j' = 1/\delta$ on the interval $\ell = j$ and zero otherwise. Furthermore, the derivative of all other basis functions evaluates to zero on this interval (they are only non-zero on their own interval). Thus, on the interval I_j, $g_\lambda'(r) = \alpha + \lambda_j \Psi_j'(r)$, and the second integral becomes

$$\int_{I_j} \frac{1/\delta}{\alpha + \lambda_j/\delta}\,\mathrm{d}\widetilde{\rho}(r) = \frac{1}{\alpha\delta + \lambda_j}\mathbb{P}(\|X\| \in I_j)\,.$$

This can be computed easily, and takes care of the second part of (25) for each component.

For the first part of (25), notice that for a given j index, the numerator of the integrand is zero on all I_ℓ for $\ell < j$. We arrive at

$$(d-1)\sum_{\ell\geq j}\int_{I_\ell} \frac{\Psi_j(r)}{\alpha r + \sum_k \lambda_k \Psi_k(r)}\,\mathrm{d}\widetilde{\rho}(r)\,.$$

However, in the denominator, for $k < \ell$, $\Psi_k = 1$ and, for $k > \ell$, $\Psi_k = 0$ and the integral simplifies to

$$(d-1)\sum_{\ell\geq j}\int_{I_\ell} \frac{\Psi_j(r)}{\alpha r + \sum_{k<\ell} \lambda_k + \lambda_\ell \Psi_\ell(r)}\,\mathrm{d}\widetilde{\rho}(r)\,.$$

The remaining integral can be easily evaluated using numerical integration, as the density of $\widetilde{\rho}$ is available in closed-form.

C.6 Proof of Theorem 4.6

To prove the main result, we follow the strategy of Jiang et al. (2025, Lemma 5.16). Setting up notation, recall that

$$\nabla_\lambda \mathcal{V}((T_\lambda)_\sharp \rho) = \nabla_\lambda \mathbb{E}_{X\sim\rho}[V(T_\lambda(X))] = \mathbb{E}_{X\sim\rho}\vec{\Psi}(\|X\|)\langle X/\|X\|, \nabla V(T_\lambda(X))\rangle$$
$$\hat{\nabla}_\lambda \mathcal{V}((T_\lambda)_\sharp \rho) = \vec{\Psi}(\|\hat{X}\|)\langle \hat{X}/\|\hat{X}\|, \nabla V(T_\lambda(\hat{X}))\rangle\,,$$

where $\hat{X}\sim\rho$. As the gradient is only stochastic in the potential term, we need to bound

$$\mathbb{E}\|Q^{-1}(\hat{\nabla}_\lambda \mathcal{F}(T_\lambda) - \nabla_\lambda \mathcal{F}(T_\lambda)\|_Q^2 = \mathbb{E}\|Q^{-1/2}(\hat{\nabla}_\lambda \mathcal{V}((T_\lambda)_\sharp\rho) - \nabla_\lambda \mathcal{V}((T_\lambda)_\sharp\rho)\|^2 \le c_0 + c_1 \mathbb{E}\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)}\,,$$

for constants $c_0, c_1 > 0$ which depend on the problem parameters (e.g., ℓ_V, L_V, d).

To start, note that it suffices to use the following bound

$$\begin{aligned}\operatorname{tr}\operatorname{Cov}(Q^{-1/2}\vec{\Psi}(\|X\|)\langle X/\|X\|, \nabla V(T_\lambda(X))\rangle) &= \mathbb{E}\langle Q^{-1}, \vec{\Psi}(\|X\|)\vec{\Psi}(\|X\|)^\top\rangle\langle X/\|X\|, \nabla V(T_\lambda(X))\rangle^2) \\ &\le J^3 \mathbb{E}_{X\sim\rho}\|\nabla V\circ T_\lambda(X)\|^2\,.\end{aligned}$$

where we bound $\langle Q^{-1}, \vec{\Psi}(\|x\|)\vec{\Psi}(\|x\|)^\top\rangle \lesssim J^3$. This follows by mimicking Lemma 5.15 by Jiang et al. (2025) but using the computations from the proof of Proposition 4.4; we omit this computation. We bound this last term as follows

$$\begin{aligned}\mathbb{E}_\rho\|\nabla V\circ T_\lambda\|^2 &\le 2\,\mathbb{E}_\rho\|\nabla V\circ T_\lambda - \nabla V\circ T_{\lambda^\star}\|^2 + 2\,\mathbb{E}_\rho\|\nabla V\circ T_{\lambda^\star}\|^2 \\ &\le 2L_V^2\,\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)} + 2\,\mathbb{E}_\rho\|\nabla V\circ T_{\lambda^\star}\|^2 \\ &\le 2L_V^2\,\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)} + 4L_V^2\,\|T_{\lambda^\star} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} + 4\mathbb{E}_\rho\|\nabla V\circ T^\star_{\mathrm{rad}}\|^2 \\ &\le 2L_V^2\,\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)} + 4L_V^2\,\|T_{\lambda^\star} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} + 4\kappa^2 L_V d\,,\end{aligned}$$

where the last inequality follows from Lemma C.2. Using a crude upper bound of $L_V\|T_{\lambda^\star} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} \le \kappa d$ (which can be derived via Theorem 4.1, we can simplify the bound to

$$\begin{aligned}\mathbb{E}\|Q^{-1}(\hat{\nabla}_\lambda \mathcal{F}(T_\lambda) - \nabla_\lambda \mathcal{F}(T_\lambda)\|_Q^2 &\lesssim J^3 L_V^2 \mathbb{E}\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)} + J^3\Big(L_V^2\|T_{\lambda^\star} - T^\star_{\mathrm{rad}}\|^2_{L^2(\rho)} + \kappa^2 L_V d\Big) \\ &\lesssim J^3 L_V^2 \mathbb{E}\|T_\lambda - T_{\lambda^\star}\|^2_{L^2(\rho)} + J^3\kappa^2 d L_V\,.\end{aligned}$$

The statement follows by employing Theorem 4.3 by Jiang et al. (2025).

Lemma C.2. *Let $\pi \propto \exp(-V)$ satisfy WC with V minimized at the origin, and let $\pi^\star_{\mathrm{rad}}$ be the optimal radial approximation. Then*

$$\mathbb{E}_{Y\sim\pi^\star_{\mathrm{rad}}}\|\nabla V(Y)\|^2 \le \kappa^2 L_V d\,.$$

Proof. By smoothness of V and strong convexity of $-\log(\pi^\star_{\mathrm{rad}})$

$$\mathbb{E}_{Y\sim\pi^\star_{\mathrm{rad}}}\|\nabla V(Y)\|^2 \le L_V^2\mathbb{E}_{Y\sim\pi^\star_{\mathrm{rad}}}\|Y\|^2 \le \kappa^2\mathbb{E}_{Y\sim\pi^\star_{\mathrm{rad}}}\|\nabla(-\log\pi^\star_{\mathrm{rad}})(Y)\|^2\,.$$

We conclude by invoking a standard fact about log-smooth measures (Chewi, 2026, "basic lemma"). □

Lemma C.3. *Let $\{\Psi_j\}_{j=0}^J$ be our dictionary of choice, and let $Q_{ij} = \mathbb{E}_{X\sim\rho}[\Psi_i(\|X\|)\Psi_j(\|X\|)]$. Then for any $x\in\mathbb{R}^d$, it holds that*

$$\langle Q^{-1}, \vec{\Psi}(x)\vec{\Psi}(x)^\top\rangle \lesssim M^3.$$

where $M = J+1$.

Proof. Our goal is to show that Q is lower bounded by some absolute constant $c_Q > 0$. This is sufficient, as

$$\langle Q^{-1}, \vec{\Psi}(x)\vec{\Psi}(x)^\top\rangle \le \frac{1}{c_Q}\operatorname{tr}(\vec{\Psi}(x)\vec{\Psi}(x)^\top) = \frac{1}{c_Q}\sum_{j=0}^J \Psi_j^2(\|x\|) \le \frac{M}{c_Q}\,,$$

where we used the trivial upper bound $\Psi_j \leq 1$. Ultimately, we want to show that $1/c_Q \lesssim M^2$. Taking $\lambda \in \mathbb{R}^M_+$, this is equivalent to proving the following lower bound

$$\lambda^\top Q \lambda = \int \Big(\sum_{j=0}^{J} \lambda_j \Psi_j(r)\Big)^2 \,\mathrm{d}\widetilde{\rho}(r) =: \int (\widetilde{T}_\lambda(r))^2 \,\mathrm{d}\widetilde{\rho}(r) \overset{!}{\gtrsim} M^{-2}\|\lambda\|^2 \,.$$

We can lower-bound this quantity term by term along the partitions $I_0 = [0, \sqrt{d} - R]$ and $I_\ell = [a_\ell, a_\ell + \delta]$. For $\ell \geq 1$,

$$\int_{a_\ell}^{a_\ell+\delta} (\widetilde{T}_\lambda(r))^2 \,\mathrm{d}\widetilde{\rho}(r) \geq \lambda_\ell^2 \inf_m \int_{a_\ell}^{a_\ell+\delta} ((r-a_\ell)/\delta - m)^2 \,\mathrm{d}\widetilde{\rho}(r) \gtrsim \lambda_\ell^2 \Upsilon^{-1} \int_{a_\ell}^{a_\ell+\delta} \mathrm{d}\widetilde{\rho}(r) \gtrsim \lambda_\ell^2 \Upsilon^{-1} \,,$$

where, over the interval, $\widetilde{\rho}$ is effectively constant (recall the arguments in Proposition 4.4). Also, for $\ell = 0$, we can use the arguments from Proposition 4.4) in this special case to prove that

$$\int_0^{\delta_0} (\widetilde{T}_\lambda(r))^2 \,\mathrm{d}\widetilde{\rho}(r) \geq \lambda_0^2 \delta_0^{-2} \int_0^{\delta_0} r^2 \,\mathrm{d}\widetilde{\rho}(r) \gtrsim \lambda_0^2 \,.$$

Adding up all the terms, this concludes the proof as $\Upsilon \asymp M^2$. □

D Experimental details

D.1 Hyperparameters

For `radVI`, our stochastic estimates use $n = 100$ samples. For the isotropic distributions: we use 10000 iterations for all experiments and the learning rate for Gaussian and Student-t was set to 7×10^{-3}, for the the logistic distribution, we used 5×10^{-2}, and 5×10^{-3} for Laplace. In the anisotropic case, we use 7×10^{-3} for all distributions and with 30000 iterations.

To obtain a Laplace approximation of a posterior, we use the "minimize" function in Scipy (Virtanen et al., 2020) with the BFGS optimizer. Our implementation of Forward-Backward VI is exactly as in the publicly available repository by Diao et al. (2023), and we use the same learning rate and number of iterations as our approach.

D.2 Synthetic distributions

We now go over the various synthetic distributions we considered throughout this work, omitting the trivial Gaussian case. We begin by defining the Mahalanobis distance $r(x) := \sqrt{(x-\mu)^\top \Sigma^{-1}(x-\mu)}$, where μ is the mean and Σ the covariance matrix.

D.2.1 Student-t distribution

Let π_{Stu} be the Student-t distribution with $\nu > 0$ degrees of freedom in $\mathbb{R}^d$, mean $\mu \in \mathbb{R}^d$, and scale matrix $\Sigma \in \mathbb{R}^{d\times d}$. We say $X \sim \pi_{\mathrm{Stu}}$ if $X = \mu + Z/\sqrt{W/\nu}$ where $Z \sim \mathcal{N}(0, \Sigma)$ and $W \sim \chi^2_\nu$ are independent. The density is given by

$$\pi_{\mathrm{Stu}}(x) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\,(\nu\pi)^{d/2}\,|\Sigma|^{1/2}} \Big(1 + \frac{r(x)^2}{\nu}\Big)^{-\frac{\nu+d}{2}} \,.$$

Writing $\pi_{\mathrm{Stu}} \propto \exp(-V_{\mathrm{Stu}})$, the potential function V is given by

$$V_{\mathrm{Stu}}(x) = \frac{\nu+d}{2} \log\Big(1 + \frac{r(x)^2}{\nu}\Big) + \frac{d}{2}\log(\nu\pi) + \frac{1}{2}\log|\Sigma| + \log\Gamma\Big(\frac{\nu}{2}\Big) - \log\Gamma\Big(\frac{\nu+d}{2}\Big) \,.$$

Closed-form radial transport map. Here we briefly derive the closed-form expression for the optimal transport map from the standard Gaussian to the Student-t distribution, written $T^\star(x) = \Psi^\star(\|x\|)\,\frac{x}{\|x\|}$. To compute $T^\star$, we invoke the principle of conservation of mass: under a transport T that pushes ρ to π, probability mass is preserved. Formally, for all Borel sets $A \subset \mathbb{R}^d$,

$$\pi(A) = ((T^\star)_\sharp \rho)(A) = \rho\big((T^\star)^{-1}(A)\big)\,.$$

In the radial setting, it suffices to enforce this on balls of radius r i.e., on sets of the form $\{x \in \mathbb{R}^d : \|x\| \le r\}$, yielding, for a random variable $X \sim \rho$ and $Y \sim \pi$,

$$\mathbb{P}(\|X\| \le r) = \mathbb{P}(\|Y\| \le \Psi^\star(r))\,.$$

For $\rho = \mathcal{N}(0, I_d)$ we have $\|X\|^2 \sim \chi^2_d$, hence

$$\mathbb{P}(\|X\| \le r) = \mathbb{P}(\|X\|^2 \le r^2) = F_{\chi^2_d}(r^2)\,,$$

and $\Psi^\star(r)$ is determined implicitly by solving

$$F_{\chi^2_d}(r^2) = \mathbb{P}(\|Y\| \le \Psi^\star(r))\,. \tag{26}$$

If $Y \sim \pi_{\mathrm{Stu}}$, then $\|Y\|^2/d \sim F_{d,\nu}$, where $F_{d,\nu}$ is the CDF of the F-distribution with (d, ν) degrees of freedom. Hence,

$$\mathbb{P}(\|Y\| \le s) = \mathbb{P}\Big(\frac{\|Y\|^2}{d} \le \frac{s^2}{d}\Big) = F_{d,\nu}\Big(\frac{s^2}{d}\Big)\,.$$

By (26),

$$F_{\chi^2_d}(r^2) = F_{d,\nu}\Big(\frac{\Psi^\star(r)^2}{d}\Big)\,,$$

and thus,

$$\Psi^\star(r) = \sqrt{d\,F^{-1}_{d,\nu}\big(F_{\chi^2_d}(r^2)\big)}\,.$$

D.2.2 Laplace distribution

Let π_{Lap} be the Laplace distribution in $\mathbb{R}^d$. To draw $X \sim \pi_{\mathrm{Lap}}$ with mean parameter μ and covariance parameter Σ, apply the transformation $X = \mu + \sqrt{Y}\,\Sigma^{1/2} Z$, where $Y \sim \mathrm{Exp}(1)$ $Z \sim \mathcal{N}(0, I_d)$ are independent.

The density of the d-dimensional symmetric multivariate Laplace distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is given by

$$\pi_{\mathrm{Lap}}(x) = \frac{2}{(2\pi)^{d/2}|\Sigma|^{1/2}} \left(\frac{r(x)^2}{2}\right)^{\nu/2} K_\nu\left(\sqrt{2}\,r(x)\right), \tag{27}$$

where $\nu = (2-d)/2$, and K_ν denotes the modified Bessel function of the second kind. Writing $\pi_{\mathrm{Lap}} \propto \exp(-V_{\mathrm{Lap}})$, we have

$$V_{\mathrm{Lap}}(x) = -\log\Big(\frac{2}{(2\pi)^{d/2}|\Sigma|^{1/2}}\Big) - \frac{\nu}{2}\log\Big(\frac{r(x)^2}{2}\Big) - \log K_\nu(\sqrt{2}r(x)). \tag{28}$$

Radial transport map. For $Y \sim \pi_{\mathrm{Lap}}$, we have that

$$\mathbb{P}(\|Y\| \le s) = \frac{\displaystyle\int_0^s \varphi(u)\,\mathrm{d}u}{\displaystyle\int_0^\infty \varphi(t)\,\mathrm{d}t}\,, \quad \varphi(s) = s^{\frac{d}{2}}\,K_{1-\frac{d}{2}}(\sqrt{2}\,s).$$

By (26), we can determine $\Psi^\star$ implicitly by solving

$$F_{\chi^2_d}(r^2) = \frac{\int_0^{\Psi^\star(r)} u^{\frac{d}{2}} K_{1-\frac{d}{2}}(\sqrt{2}\,u)\,\mathrm{d}u}{\int_0^{\infty} t^{\frac{d}{2}} K_{1-\frac{d}{2}}(\sqrt{2}\,t)\,\mathrm{d}t}.$$

Solving the implicit equation for the mapped radius $\Psi^\star(r)$ point-wise is computationally prohibitive. To circumvent this, we use an interpolation-based approach: the radial CDF of the target is pre-computed on a dense grid, and a linear interpolator of its inverse is constructed. The map then reduces to a direct, vectorized evaluation of this interpolator. Alternatively, the equation can be solved for each point using a root-finding algorithm, such as Brent's method (Brent, 1973), although this approach is less efficient for large datasets.

D.2.3 Logistic distribution

Let π_{Log} be the logistic distribution in $\mathbb{R}^d$ with mean μ, covariance Σ, and scale parameter $s > 0$. Unlike the other cases, we use rejection sampling here. The target distribution is elliptical, so it suffices to sample the Mahalanobis radius, whose density is

$$f_R(r) \propto r^{d-1}\,\frac{\exp(-r/s)}{(1+\exp(-r/s))^2}\,, \qquad r > 0\,.$$

As proposal we take $R \sim \mathrm{Gamma}(d, s)$, which has density $g(r) \propto r^{d-1}\exp(-r/s)$. The acceptance probability is

$$\alpha(r) = \frac{f_R(r)}{g(r)} = \frac{1}{(1+\exp(-r/s))^2}\,.$$

Accepted radii are combined with a uniform direction U on the unit sphere, yielding $X = \mu + \Sigma^{1/2}(RU)$.

The density of the d-dimensional multivariate logistic distribution with mean $\mu \in \mathbb{R}^d$, covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, and scale parameter $s > 0$ is

$$\pi_{\mathrm{Log}}(x) = \frac{1}{Z(d,\Sigma,s)}\,\frac{\exp(-r(x)/s)}{(1+\exp(-r(x)/s))^2}\,, \tag{29}$$

where

$$Z(d,\Sigma,s) = |\Sigma|^{1/2}\,\frac{2\pi^{d/2}}{\Gamma(d/2)}\,s^d\,\Gamma(d)\,(1-2^{-(d-2)})\,\zeta(d-1)\,.$$

Similarly as before, taking the negative logarithm of (29) realizes the potential V,

$$V_{\mathrm{Log}}(x) = \frac{r(x)}{s} + 2\log(1+\exp(-r(x)/s)) + \log Z(d,\Sigma,s). \tag{30}$$

Radial transport map. For $Y \sim \pi_{\mathrm{Log}}$, it holds that

$$\mathbb{P}(\|Y\| \le s) = \frac{\int_0^s \varphi(u)\,\mathrm{d}u}{\int_0^\infty \varphi(t)\,\mathrm{d}t}\,, \quad \varphi(s) = s^{d-1}\,\frac{e^{-s}}{(1+e^{-s})^2}\,.$$

By (26), we can determine $\Psi^\star$ implicitly by solving

$$F_{\chi^2_d}(r^2) = \frac{\int_0^{\Psi^\star(r)} u^{d-1}\,\frac{e^{-u}}{(1+e^{-u})^2}\,\mathrm{d}u}{\int_0^{\infty} t^{d-1}\,\frac{e^{-t}}{(1+e^{-t})^2}\,\mathrm{d}t}\,.$$

in the same way as in the case of the Laplace distribution.

E Auxiliary computational results

E.1 Higher dimensions

Here, we compare methods for isotropic distributions as in Table 1, but now with $d = 100$. We set the meshsize to $d^{-1/8}$ and double the number of iterations. All other parameters are left the same, and the results are reported below in Table 3.

	Isotropic targets			
Method	Gaussian	Laplace	Logistic	Student-t
LA	1.48×10^{-4}	43.04	7450	68.89
GVI	5.07×10^{-4}	18.34	8.67	5.21
`radVI`	3.71×10^{-4}	7.67×10^{-2}	1.96×10^{-1}	1.89×10^{-1}

Table 3: Estimated Wasserstein distance between various VI solutions for learning isotropic targets in $d = 100$.

E.2 Figures

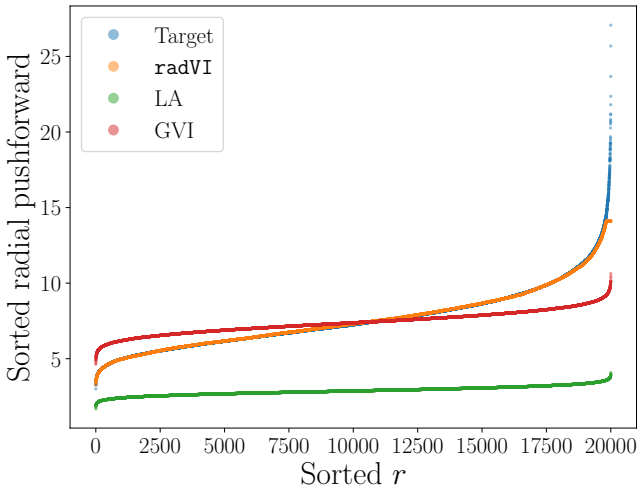

Figure 5: Comparing learned radial profiles of `radVI` versus other approximation methods for the isotropic Student-t distribution.

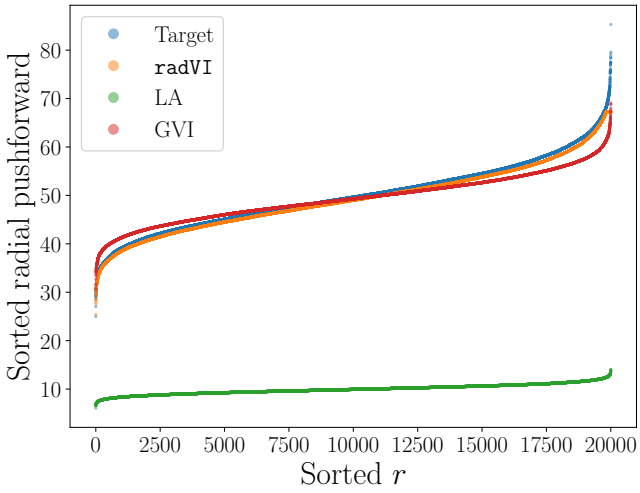

Figure 6: Comparing learned radial profiles of `radVI` versus other approximation methods for the isotropic logistic distribution.

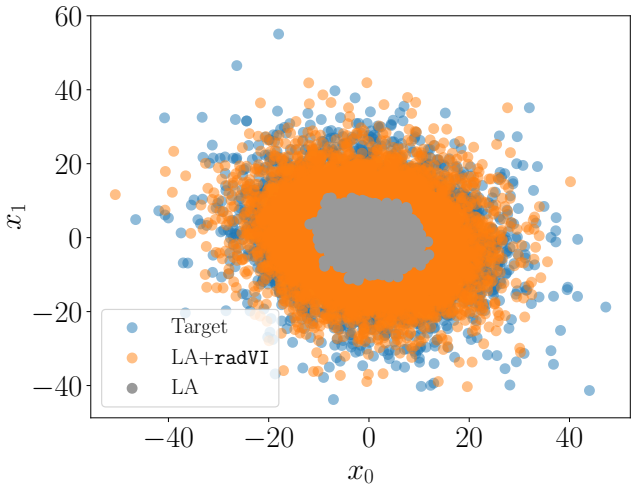

Figure 7: Visual comparison of true target samples, those generated by LA, and ours (LA+`radVI`), for learning the anisotropic Student-t distribution.

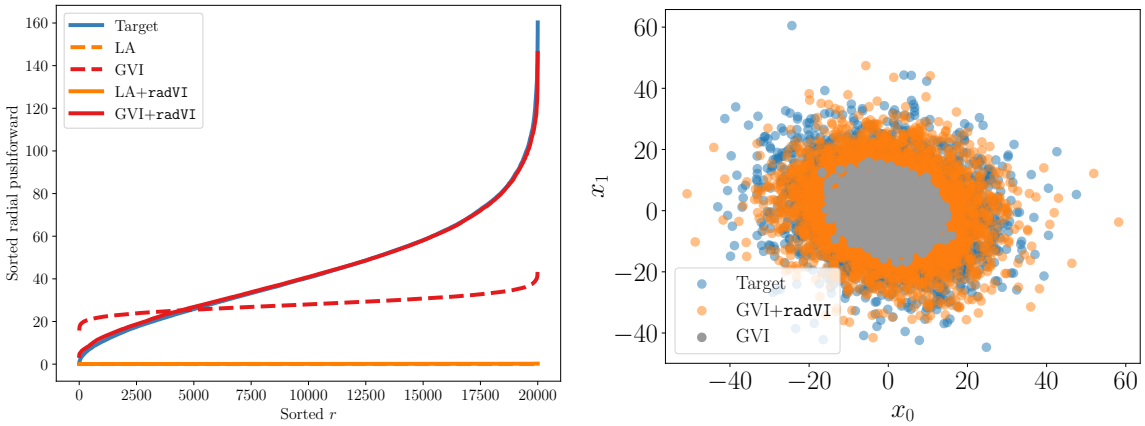

Figure 8: **Left:** Comparing whitening methods for learning the anisotropic Laplace distribution, with and without `radVI`. **Right:** Visual comparison of true target samples, those generated by GVI, and ours (GVI+`radVI`).