Theory and computation for structured variational inference

Shunan Sheng[1,*], Bohan Wu[1,*], Bennett Zhu[1]
Sinho Chewi[2], and Aram-Alexandre Pooladian[3]

[1]Department of Statistics, Columbia University
[2]Department of Statistics and Data Science, Yale University
[3]Foundations of Data Science, Yale University

{ss6574,bw2766,bz2500}@columbia.edu

{sinho.chewi,aram-alexandre.pooladian}@yale.edu

November 14, 2025

Abstract

Structured variational inference constitutes a core methodology in modern statistical applications. Unlike mean-field variational inference, the approximate posterior is assumed to have interdependent structure. We consider the natural setting of *star-structured* variational inference, where a root variable impacts all the other ones. We prove the first results for existence, uniqueness, and self-consistency of the variational approximation. In turn, we derive quantitative approximation error bounds for the variational approximation to the posterior, extending prior work from the mean-field setting to the star-structured setting. We also develop a gradient-based algorithm with provable guarantees for computing the variational approximation using ideas from optimal transport theory. We explore the implications of our results for Gaussian measures and hierarchical Bayesian models, including generalized linear models with location family priors and spike-and-slab priors with one-dimensional debiasing. As a by-product of our analysis, we develop new stability results for star-separable transport maps which might be of independent interest.

1 Introduction

Suppose we have a collection of observations $X_1, \ldots, X_n$ which are governed by latent parameters $(z_1, \ldots, z_d)$, where n and d are both potentially large; the joint distribution is

*Equal contribution.

written as $p(z, x)$. In Bayesian statistics, the practitioner wishes to perform inference on the *posterior* distribution

$$\pi(z) \propto p(z \mid X_1, \ldots, X_n)\,,$$

which is only known up to a normalizing constant.* See, for instance, Berger (1993); Hoff (2009); Gelman et al. (2014) and references therein. Inference on the latent parameters is performed by drawing samples from the unnormalized distribution π, which can be obtained through Markov Chain Monte Carlo (MCMC) methods. MCMC, a standard workhorse in Bayesian computation, has witnessed significant advancements over decades of research in methodological, statistical, and theoretical circles; see Hastings (1970); Gelfand and Smith (1990); Neal (1993); Robert and Casella (1999) for standard treatments. Unfortunately, MCMC can often be computationally expensive, since many iterations of a given chain are required to generate even a single approximate sample from the posterior π. Moreover, if the posterior respects a complex graphical structure, exact MCMC methods become intractable due to the curse of dimensionality (Wainwright and Jordan, 2008). Consequently, the computational burden often prevents practitioners from exploring different models within a reasonable time budget.

To solve this computational hurdle, there is a growing body of literature on *variational inference*, a classical method dating back to Parisi (1980); Hinton and Van Camp (1993); Jordan et al. (1999); Wainwright and Jordan (2008). In this setting, the statistician posits a tractable class $\mathcal{C}$ of probability measures, and computes the following projection in the sense of relative entropy, or Kullback–Leibler (KL) divergence:

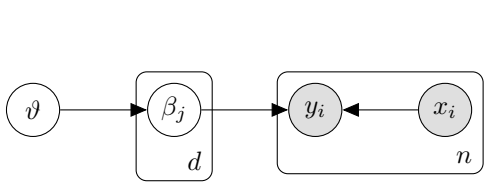
$$\pi_{\mathcal{C}} \in \operatorname*{arg\,min}_{\mu \in \mathcal{C}} \mathsf{KL}\left(\mu \parallel \pi\right) = \operatorname*{arg\,min}_{\mu \in \mathcal{C}} \int \log\left(\frac{\mu}{\pi}\right) \mathrm{d}\mu\,. \tag{1}$$

Variational inference has found applications in, for example, large-scale Bayesian inference problems such as topic modeling (Blei et al., 2003), deep generative modeling (Kingma and Welling, 2014; Lopez et al., 2018), robust Bayes (Wang and Blei, 2018), marketing (Braun and McAuliffe, 2010), and genome sequence modeling (Carbonetto and Stephens, 2012; Lopez et al., 2018; Wang et al., 2020; Kim et al., 2024). In these scenarios, obtaining an exact posterior sample may be intractable, whereas a well-calibrated approximation is sufficient for practical use.

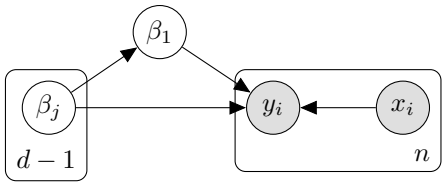
The choice of $\mathcal{C} \subset \mathcal{P}(\mathbb{R}^d)$ in the variational inference problem (1) is essential, as it dictates not only the computational complexity of the resulting optimization problem but also the bias introduced by approximation. One of the most popular choices for the constraint set is the space of product measures $\mathcal{C} = \mathcal{P}(\mathbb{R})^{\otimes d}$, known as mean-field VI (MFVI) in the literature (Blei et al., 2017). Due to its computational tractability, MFVI has been integrated into high-dimensional linear models (Carbonetto and Stephens, 2012; Wang et al., 2020; Kim et al., 2024), generative modeling (Kingma and Welling, 2014), and language modeling (Blei et al., 2003), among other problems in probabilistic machine learning (Murphy, 2023).

However, the minimizer obtained by choosing $\mathcal{C} = \mathcal{P}(\mathbb{R})^{\otimes d}$ can have low fidelity to the posterior π, as the latent variables are typically *not* independent. For Bayesian inference,

*Throughout, we use π to denote both the measure and its Lebesgue density.



(a) A parameter ϑ influences β, which generates the data. The posterior measure is over $(\vartheta, \beta) = (\vartheta, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$.



(b) A Bayesian GLM using *spike-and-slab* and *debiased* priors. The posterior is over the parameters $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$.

Figure 1: Bayesian GLMs with different structural dependencies on the model parameters.

MFVI fails to capture many standard statistical procedures; see Wang and Titterington (2004); Ghorbani et al. (2019). This weakness has spurred a line of work on *structured* VI (SVI), where dependencies are introduced among the variables of the variational distribution, dating back to Saul and Jordan (1995); Lauritzen (1996); Barber and Wiegerinck (1999). Often, the graphical structure of the family $\mathcal{C}$ is fixed in advance: let $\mathcal{G}$ denote a fixed tree graph[†]. We say that $\mu \in \mathcal{C}_\mathcal{G} \subset \mathcal{P}(\mathbb{R}^d)$ if

$$\mu(z_1, \ldots, z_d) = \prod_{(i,j) \in \mathcal{E}_\mathcal{G}} \varphi_{ij}(z_i, z_j)\,, \tag{2}$$

for some clique compatibility functions $\varphi_{ij} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ where $(i, j) \in \mathcal{E}_\mathcal{G}$ is an edge in the graph $\mathcal{G}$. The SVI problem is then

$$\min_{\mu \in \mathcal{C}_\mathcal{G}} \mathsf{KL}\left(\mu \parallel \pi\right). \tag{3}$$

MFVI is a special case where the set of edges of the tree graph is empty, i.e., $\mathcal{E}_\mathcal{G} = \emptyset$.

With the increasing adoption of SVI, key empirical insights into its statistical and computational behavior have emerged. As noted by Blei et al. (2017), structured variational families can *"potentially improve the fidelity of the approximation"*, while coming at the cost of a *"more difficult-to-solve variational optimization problem"*. Motivated by these insights, we investigate the two following questions:

(a) **Approximation Guarantees:** Under what conditions does SVI provide accurate approximations to the posterior, and how can we quantify the approximation error?

(b) **Computational Guarantees:** Can we design a polynomial-time algorithm for solving the SVI problem with provable computational guarantees?

To answer these questions, we focus on a particular structured family which we call the *star variational family*, denoted $\mathcal{C}_{\text{star}}$. Formally, $\mu \in \mathcal{C}_{\text{star}}$ if the measures admits the following decomposition

$$\mu(z_1, \ldots, z_d) = \mu_1(z_1) \prod_{j=2}^{d} \mu_j(z_j \mid z_1)\,. \tag{4}$$

[†]Precisely, this is a graph with vertex-edge structure $\mathcal{G} = (\{1, \ldots, d\}, \mathcal{E}_\mathcal{G})$.

In other words, we consider the case when the graph $\mathcal{G}$ is a star graph: one of the variables takes the role of a "root" variable, and the remaining ones are leaves in the graph (see Figure 1 for examples).

Star graphs are ubiquitous in Bayesian statistics. Indeed, the fundamental result of de Finetti (de Finetti, 1929, 2017) states that the law of any (infinitely) exchangeable sequence $z_1, z_2, \ldots$ can be uniquely represented as the marginal law of the leaf variables in a star graph, where the root variable corresponds to a latent variable, and the leaf variables are $z_1, z_2, \ldots$. This observation has since been adopted as a foundational principle in Bayesian modeling and inference (Rubin, 1978; Saarela et al., 2023). Figure 1(a) illustrates an example of a nested hierarchical model (Hoffman and Blei, 2015; Fasano et al., 2022; Loaiza-Maya et al., 2022; Goplerud et al., 2025). In this setting, the joint distribution of the observed variables $(x_{1:n}, y_{1:n})$, the local latent variables $\beta_{1:d}$, and the global latent variable ϑ is given by

$$p_0(\vartheta) \prod_{j=1}^{d} p_1(\beta_j \mid \vartheta) \prod_{i=1}^{n} p_{\mathcal{L}}(y_i \mid \beta_{1:d}, x_i), \tag{5}$$

where $\vartheta \sim p_0$, $\beta_1, \ldots, \beta_d \mid \vartheta \overset{\text{i.i.d.}}{\sim} p_1(\cdot \mid \vartheta)$, and $y_i \mid \beta_{1:d}, x_i \sim p_{\mathcal{L}}(\cdot \mid \beta_{1:d}, x_i)$ for all $i \in [n]$. Consequently, the posterior distribution of $(\vartheta, \beta_{1:d})$ factorizes according to a star graph, in which the root variable is ϑ and the leaf variables are $\beta_1, \ldots, \beta_d$.

Another notable example of the class of nested hierarchical models is the high-dimensional generalized linear mixed model. In this context, Goplerud et al. (2025) shows that applying SVI leads to improved uncertainty quantification and scalable computation while MFVI underestimates posterior uncertainty. More broadly, hierarchical priors of the form (5) play a central role in multilevel modeling; see Gelman and Hill (2007) for a comprehensive overview.

The star-graph setting is, in essence, one step above the mean-field setting, when $\mathcal{E}_{\mathcal{G}}$ is empty. Nevertheless, statistical and computational properties of this regime have not been effectively explored in the literature.

1.1 Contributions

In this work, we develop the first theoretical and computational properties of **S**tar-**S**tructured **V**ariational **I**nference (SSVI): For a general Gibbs measure of the form $\pi \propto e^{-V}$, where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential function, find

$$\pi^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) . \tag{6}$$

Our contributions are several-fold.

Existence and uniqueness of the SSVI minimizer First, we investigate when this infinite-dimensional optimization problem is well-defined, i.e., establish conditions under which a unique minimizer exists. Drawing inspiration from recent works which analyze the simpler mean-field setting (Arnese and Lacker, 2024; Lacker et al., 2024; Lavenant and Zanella, 2024; Jiang et al., 2025), we consider the case where the posterior is log-concave. We

show that such a condition is sufficient to obtain a unique minimizer to (6). By definition of $\mathcal{C}_{\text{star}}$, this minimizer is of the form

$$\pi^\star(z_1, \dots, z_d) = p^\star(z_1)\, q^\star(z_{-1} \mid z_1)\,, \tag{7}$$

where $p^\star$ is a univariate probability measure and $q^\star$ is a stochastic kernel mapping inputs $z_1 \in \mathbb{R}$ to product measures over $\mathbb{R}^{d-1}$. This is established via a dynamic programming principle; see Proposition 2.1, and the complete result is given by Theorem 2.2.

Regularity properties of minimizers Next, we characterize the approximation quality of $\pi^\star$, i.e., how close is this approximation to the true posterior π?

To this end, we establish novel *self-consistency equations* for the star-structured minimizer. These equations explicitly relate the components $p^\star$ and $q^\star$ to each other and to the posterior $\pi \propto \exp(-V)$:

$$\begin{aligned} p^\star(z_1) &\propto \exp\Big(-\int_0^{z_1} \int \partial_1 V(s, z_{-1})\; q^\star(\mathrm{d}z_{-1} \mid s)\, \mathrm{d}s\Big)\,, \\ q_i^\star(z_i \mid z_1) &\propto \exp\Big(-\int V(z_1, z_i, z_{-\{1,i\}}) \prod_{j \geq 2,\, j \neq i} q_j^\star(\mathrm{d}z_j \mid z_1)\Big)\,, \end{aligned} \tag{8}$$

for $i \in \{2, \dots, d\}$; see Theorem 2.4. If we additionally assume the potential V itself satisfies a *root domination criteria*:

$$\partial_{11} V - \frac{\| \sum_{j=2}^d (\partial_{1j} V)^2 \|_{L^\infty}}{\ell_V} > 0\,, \tag{9}$$

then we establish the following approximation bound:

$$\mathsf{KL}\left(\pi^\star \parallel \pi\right) \lesssim \sum_{i \geq 2} \sum_{j > i} \mathbb{E}_{\pi^\star}[(\partial_{ij} V)^2]\,, \tag{10}$$

where the underlying constant is explicitly characterized; see Theorem 2.9. The required assumption essentially asserts that V has more curvature along the root direction than along the root–leaf interactions.

Our approximation guarantee strictly generalizes that of Lacker et al. (2024), which only applies to the mean-field case. Indeed, bound (10) substantially improves upon the MFVI guarantee of Lacker et al. (2024) when the root–leaf interactions $(\partial_{1j} V)_{j \geq 2}$ dominate the inter-leaf interactions $(\partial_{ij} V)_{i \neq j,\; i,j > 1}$. We conclude Section 2 with several illustrative examples, including the Gaussian posterior (Section 2.3.1) and Bayesian generalized linear models (Section 2.3.2) with location-family priors and spike-and-slab priors, respectively. Notably, in the Gaussian case, we provide an explicit characterization of the SSVI minimizer and quantify the improvement of performing SSVI over MFVI.

Computational guarantees via (linearized) optimal transport In Section 3, we shift focus to computational aspects. We stress that the preceding developments do not naturally give rise to an optimization routine which is amenable to computational analysis. This is simply because, at first glance, the constraint set $\mathcal{C}_{\text{star}}$ is non-convex. To circumvent this, we

draw inspiration from the growing body of work that leverages *optimal transport theory* in statistical applications (Panaretos and Zemel, 2020; Chewi et al., 2025), and recast the SSVI problem into an optimization problem at the level of transport maps.

Taking $\rho = \mathcal{N}(0, I)$, for each measure in $\mathcal{C}_{\text{star}}$, we can parameterize it by a map T that belongs to a family of *star-separable transport maps* $\mathcal{T}_{\text{star}}$ (see Section 3.1 for a precise definition). These maps naturally form a convex set and thus, the SSVI problem (6) can be reformulated as a convex optimization problem over $\mathcal{T}_{\text{star}}$,

$$T^\star = \operatorname*{arg\,min}_{T \in \mathcal{T}_{\text{star}}} \mathsf{KL}\left(T_\# \rho \parallel \pi\right). \tag{11}$$

We note that the set $\mathcal{T}_{\text{star}}$ is closed under convergence in $L^2(\rho)$, a property that is intimately connected to the adapted Wasserstein distance (see Beiglböck et al., 2023, and references therein). This transport map perspective not only yields computational guarantees but also provides an alternative route to proving many of the results in Section 2 via direct analysis of (11), including the existence and uniqueness of the SSVI minimizer (Theorem 3.2).

To proceed, we borrow inspiration from the machine learning literature (Wang et al., 2013; Jiang et al., 2025) and exhibit an explicit parameterization of the transport maps in $\mathcal{T}_{\text{star}}$, giving rise to a finite-dimensional parameter space Θ such that optimization over Θ *preserves* the convexity of the underlying optimization problem. Concretely, we devise a (projected) gradient-descent algorithm over our parameterized space, and we show that the minimizers of the gradient-descent scheme are close to the ground-truth minimizer in (11). The general scheme is

$$\operatorname*{arg\,min}_{T \in \mathcal{T}_{\text{star}}} \mathsf{KL}\left(T_\# \rho \parallel \pi\right) = T^\star \simeq T^\star_\Theta = \operatorname*{arg\,min}_{T \in \mathcal{T}_\Theta} \mathsf{KL}\left(T_\# \rho \parallel \pi\right),$$

and we perform gradient descent with respect to the parameters $\theta \in \Theta$ to obtain $T^\star_\Theta$.

The aforementioned convergence properties are made possible by way of exploiting the regularity properties of the optimal star-structured map $T^\star : \mathbb{R}^d \to \mathbb{R}^d$ in (11), which, to be concrete, is of the form

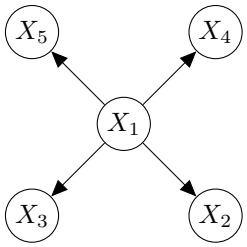
$$x \mapsto T^\star(x) = (z_1, T_2^\star(x_2 \mid z_1), \ldots, T_d^\star(z_d \mid z_1)), \quad z_1 = T_1^\star(x_1).$$

In order to develop our computational guarantees, we prove novel regularity estimates for $T^\star$ in the style of Caffarelli (2000). Under some assumptions (which are stronger than (9)), we prove, for instance, bounds on

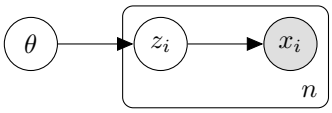
$$|\partial^2_{x_1} T_1^\star(x_1)|, \quad |\partial^2_{x_i} T_i^\star(x_i \mid z_1)|,$$

and the mixed derivatives. To the best of our knowledge, these are the first results that study the regularity of transport maps with respect to varying target measures supported on unbounded domains, and are of independent interest.

Going beyond the star graph Finally, in Section 4, we discuss the additional technical hurdles when going beyond the star-graph structure, which we leave for future work, and other open questions.



(a) A star-structured graph.



(b) A generic hierarchical model. The target measure is the posterior of $(z_{1:n}, \theta)$ given $x_{1:n}$, which has a star graphical structure.

Figure 2: A general purpose illustration of star-structured variational inference.

1.2 Related work

Structured variational inference There is a growing body of work on structured VI spanning decades of research (Saul and Jordan, 1995; Salimans and Knowles, 2013; Hoffman and Blei, 2015; Xiao and Su, 2024) with ample applications, such as time series models (Saul and Jordan, 1995; Salimans and Knowles, 2013; Tan and Nott, 2018; Frazier et al., 2023), and state-space modeling (Zhang and Gao, 2020a; Frazier et al., 2023).

The theory of VI has seen accelerated progress in recent years. Closely related to our work are results on the approximation accuracy of Gaussian VI (Katsevich and Rigollet, 2024) and MFVI (Lacker et al., 2024). Other lines of research have studied asymptotic normality (Hall et al., 2011; Bickel et al., 2013; Wang and Blei, 2019), posterior contraction rates (Zhang and Gao, 2020a), non-asymptotic risk bounds (Alquier et al., 2016; Alquier and Ridgway, 2020; Yang et al., 2020), and properties of variational inference in high-dimensional generalized linear models (Tan, 2021; Fasano et al., 2022; Mukherjee and Sen, 2022; Ray and Szabó, 2022; Mukherjee et al., 2023; Qiu, 2024).

A parallel line of work has focused on establishing computational guarantees for variational inference algorithms. More closely related is the work of Jiang et al. (2025), in which authors propose a polynomial-time algorithm for computing the MFVI minimizer by directly approximating and optimizing over the space of univariate optimal transport maps. Other results in this direction include convergence analyses for coordinate-ascent VI (Arnese and Lacker, 2024; Lavenant and Zanella, 2024; Bhattacharya et al., 2025), black-box variational inference (Domke et al., 2023; Kim et al., 2023), Gaussian VI (Lambert et al., 2022; Diao et al., 2023), particle-based algorithms (Du et al., 2024) and computational-statistical trade-offs (Bhatia et al., 2022; Wu and Blei, 2024). When the log-concavity assumption is violated, a recent result shows that MFVI suffers from mode collapse (Sheng et al., 2025a).We also mention the work of Zhang and Gao (2020b), which provides full statistical and computational guarantees for MFVI in stochastic block models.

Linearized optimal transport Our approach to recasting the SSVI problem as an optimization problem at the level of transport maps follows the general scheme of linearized optimal transport (Wang et al., 2013). Fixing an absolutely continuous probability measure $\rho \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ with finite second moment, one can represent any other probability measure

μ with finite second moment via the optimal transport map from ρ to μ. This perspective has been popularized in statistical applications. When $d = 1$, for example, Zhu and Müller (2023) use the Fréchet mean of a distributional time series as the base measure and propose an auto-regressive model defined directly on the transport maps. This viewpoint builds also on the classical work in distribution learning, such as Petersen and Müller (2016), where probability densities are transformed into a Hilbert space of functions via continuous, invertible maps, allowing tools from functional data analysis (Hsing and Eubank, 2015) to be applied. Similar ideas have also been explored by Bachoc et al. (2018) and Zemel and Panaretos (2019) in related problems involving learning of probability distributions. In higher dimensions, optimal transport maps have been used by Chernozhukov et al. (2017); Hallin et al. (2021); Ghosal and Sen (2022); Deb and Sen (2023) and others to define multivariate analogues of quantile functions and to perform rank-based inference and quantile regression.

1.3 Notation

We denote by $\mathbb{S}_+^d$ the set of positive definite $d \times d$ matrices. For any $A \in \mathbb{S}_+^d$, we write $A^{ij} := (A^{-1})_{ij}$, and $\|A\|_2$ for the operator norm of A. A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *α-convex* if $f - \frac{\alpha}{2}\|\cdot\|_2^2$ is convex, and *β-smooth* if $\nabla^2 f \preceq \beta I$. We use $\partial_{ij} f$ to denote the (i, j)-th entry of the Hessian $\nabla^2 f$. Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ and $\mathcal{P}_p(\mathbb{R}^d)$ the subspace of measures with finite p-moment for $p \geq 0$. For $\rho \in \mathcal{P}(\mathbb{R}^d)$, we say that a vector-valued function $T : \mathbb{R}^d \to \mathbb{R}^d$ belongs to $L^2(\rho)$ if the function $x \mapsto \|T(x)\|$ is in $L^2(\rho)$ and define $\|T\|_{L^2(\rho)} := \left(\int \|T(x)\|\, \rho(\mathrm{d}x)\right)^{1/2}$. We denote by $\|f\|_{L^\infty}$ the L^∞ norm of f under the Lebesgue measure. Given a measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, we write $T_{\#}\mu$ for the push-forward measure $\mu \circ T^{-1}$. If $\pi \in \mathcal{P}(\mathbb{R}^d)$ is absolutely continuous, we abuse notation and identify it with its Lebesgue density. If $\pi \propto \exp(-V)$, we say π is C^2 and *α-log-concave* if $V : \mathbb{R}^d \to \mathbb{R}$ is C^2 and α-convex; V is called the *potential* of π. Similarly, we say π is *β-log-smooth* if V is β-smooth.

We use $\mathbb{E}_\pi[f(Z)]$ to denote the expectation of f under π. The Kullback–Leibler (KL) divergence between $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$ is defined as $\mathsf{KL}(\mu \parallel \pi) := \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}\pi})\, \mathrm{d}\mu$ if μ is absolutely continuous with respect to π (denoted $\mu \ll \pi$) and $\mathsf{KL}(\mu \parallel \pi) = \infty$ otherwise. Here, $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ denotes the Radon–Nikodym derivative of μ with respect to π. The (negative) differential entropy is then defined as $\mathcal{H}(\mu) := \int \log \mu \, \mathrm{d}\mu$ if μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and ∞ otherwise. For any index set $I \subset [d] := \{1, \ldots, d\}$, we identify $z = (z_I; z_{-I}) \in \mathbb{R}^d$ and use π_I to denote the marginal density of the coordinates z_I. For $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the p-Wasserstein distance is given by $W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint \|x - y\|^p \, \pi(\mathrm{d}x, \mathrm{d}y)\right)^{1/p}$, where $\Pi(\mu, \nu)$ is the set of all couplings between μ and ν, i.e., the set of measures in $\mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals (μ, ν).

For quantities a, b, we write $a \lesssim b$ or $a = O(b)$ if there exists a constant $C > 0$ (which may depend on other parameters depending on context) such that $a \leq Cb$. If both $a \lesssim b$ and $b \lesssim a$, then we write that $a \asymp b$. We write $a = \tilde{O}(b)$ if $a = O(b \operatorname{polylog}(b))$, that is, $a \lesssim b$ up to a polylogarithmic factor.

Finally, throughout the rest of the paper, we assume that the potential V of the posterior

π satisfies the following growth condition: for any $c_2 > 0$, there exists $c_1 > 0$ such that

$$|V(z)| \leq c_1 e^{c_2 \|z\|^2} \qquad \text{for all } z \in \mathbb{R}^d \,. \tag{12}$$

This growth condition only requires that V grows slower than an exponential function, and it is satisfied when $\nabla^2 V$ is bounded, i.e., when V is smooth. We remark that essentially the same a similar growth condition is required by Lacker et al. (2024) to establish the uniqueness of the mean-field variational inference minimizer.

2 A first theory for star-structured variational inference

In this first section, we prove several novel properties for solutions to the star-structured variational inference (SSVI) problem:

$$\pi^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) . \tag{SSVI}$$

Recall that $\mu \in \mathcal{C}_{\text{star}} \subset \mathcal{P}(\mathbb{R}^d)$ is a probability measure that exhibits the following graphical structure:

$$\mu(z_1, \ldots, z_d) = \mu_1(z_1) \prod_{j=2}^{d} \mu_j(z_j \mid z_1) \,,$$

where $\mu_j(\cdot \mid z_1)$ is the conditional density of z_j given z_1 under μ.

Rather than studying specific models, we aim to identify general conditions under which one can provably solve the SSVI problem—for which a prerequisite is to have a unique minimizer. The core assumption we rely on throughout this work is log-concavity (Saumard and Wellner, 2014):

$$\pi \propto \exp(-V) \text{ is } C^2 \text{ and } \alpha\text{-log-concave, i.e., } 0 \prec \alpha I \preceq \nabla^2 V \,. \tag{\textbf{SLC}}$$

The assumption (**SLC**) is standard in the theoretical and computational study of sampling (Chewi, 2025) and variational inference (Lambert et al., 2022; Domke et al., 2023; Arnese and Lacker, 2024; Lavenant and Zanella, 2024; Jiang et al., 2025), as well as existing works which study regularity properties of the simpler mean-field setting (Lacker et al., 2024). Following recent trends which view sampling as optimization over the space of measures (Wibisono, 2018), we will see shortly that strong log-concavity allows us to speak of *unique* minimizers to the SSVI problem.

All remaining proofs from this section are deferred to Section A.

2.1 Existence and uniqueness of the minimizer

We first establish existence and uniqueness of the minimizer to the SSVI problem. For any $\mu \in \mathcal{C}_{\text{star}}$, we have by design the following decomposition via the chain rule for relative entropy:

$$\mathsf{KL}\left(\mu \parallel \pi\right) = \mathsf{KL}\left(\mu_1 \parallel \pi_1\right) + \int \underbrace{\mathsf{KL}\left(\mu_{-1}(\cdot \mid z_1) \parallel \pi_{-1}(\cdot \mid z_1)\right)}_{(*)} \mu_1(\mathrm{d}z_1) \,,$$

where, for fixed $z_1 \in \mathbb{R}$, we view $\pi_{-1}(\cdot \mid z_1)$ as a probability measure over $\mathbb{R}^{d-1}$. This formulation suggests that we can optimize over μ_1 and μ_{-1} separately. Above, the nested subproblem $(*)$ is itself a *mean-field* variational inference problem, since $\mu_{-1}(\cdot \mid z_1)$ is constrained to be a product measure, where the target measure is $\pi_{-1}(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R}^{d-1})$. We make use of the nested structure in our first result, which relates (SSVI) to a dynamic programming problem.

Proposition 2.1 (Dynamic programming)**.** *The following equivalence holds:*

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) = \inf_{\mu_1 \in \mathcal{P}(\mathbb{R})} \mathsf{KL}\left(\mu_1 \parallel \pi_1\right) + \int \inf_{\nu \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right) \mu_1(\mathrm{d}z_1) \tag{13}$$

The existence and uniqueness of the SSVI minimizer then follows as a consequence.

Theorem 2.2 (Existence and uniqueness of the SSVI minimizer)**.** *Under* (**SLC**)*,* (SSVI) *has a unique minimizer of the form*

$$\pi^\star(\mathrm{d}z) \coloneqq p^\star(\mathrm{d}z_1)\; q^\star(\mathrm{d}z_{-1} \mid z_1) \propto e^{-h_{\text{val}}(z_1)}\, \pi_1(\mathrm{d}z_1)\; q^\star(\mathrm{d}z_{-1} \mid z_1)\,,$$

where $q^\star(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$ *is the unique minimizer for the problem*

$$h_{\text{val}}(z_1) \coloneqq \inf_{\nu \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right)\,,$$

for π_1*-almost all* z_1*.*

Theorem 2.2 characterizes the structure of the SSVI minimizer by revealing a close connection to a conditional MFVI problem over the leaf nodes. The marginal distribution $p^\star$ of the root variable matches the exact marginal π_1 of the target posterior, up to a tilting factor h_{val} given by the approximation quality of the inner MFVI procedure. When the conditional $\pi_{-1}(\cdot \mid z_1)$ is exactly a product measure, the tilting vanishes, and the SSVI marginal recovers the true root marginal exactly.

We now sketch the strategy to construct the minimizer from Theorem 2.2. On the one hand, the first marginal of the SSVI minimizer is given by

$$p^\star(z_1) \propto e^{-h_{\text{val}}(z_1)}\, \pi_1(z_1)\,,$$

provided h_{val} is measurable in some sense. On the other hand, for each z_1, Assumption (**SLC**) and Lacker et al. (2024, Theorem 1.1) imply that there is a unique $q^\star(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$ attaining $h_{\text{val}}(z_1)$. Therefore, if $z_1 \mapsto q^\star(\cdot \mid z_1)$ induces a measurable stochastic kernel, Proposition 2.1 states that SSVI can be solved by "gluing" $p^\star$ with the stochastic kernel $z_1 \mapsto q^\star(\cdot \mid z_1)$. The right notion of measurability is *universal measurability*. We refer the reader to Bertsekas and Shreve (1978, §7) for the definition of this concept, and to the proof in Section A for the details of the argument. Notably, after modification on a null set under π_1, both h_{val} and the map $z_1 \mapsto q^\star(\cdot \mid z_1)$ can be made Borel measurable.

Recall that strong log-concavity is also assumed by Lacker et al. (2024) to show the existence and uniqueness of the solution in the simpler mean-field setting. However, the proof of existence in MFVI relies on the direct method using the topology of weak convergence, which does not suffice to establish existence of a solution in SSVI because conditional

independence is not preserved under weak convergence. For further discussion on this matter, see, e.g., Lauritzen (2024). In Theorem 3.2, we provide an alternative proof of the existence and uniqueness of the SSVI minimizer under (**SLC**), based on a reformulation of the problem under a different geometry, in which conditional independence is preserved after taking the limit.

Remark 2.3. We later use the characterization of SSVI as a dynamic program to show that when the posterior π is Gaussian, the SSVI minimizer remains Gaussian with a structured covariance; see Theorem 2.10 below.

2.2 Regularity and approximation quality via self-consistency equations

Theorem 2.2 shows that a unique minimizer exists under (**SLC**), but it does not directly yield important information of said minimizer. For instance, we do not glean any insights regarding its regularity properties or how good of an approximation it is to π.

To this end, we now establish *self-consistency equations* for the SSVI minimizer $\pi^\star$. These equations allow us to relate the components of $\pi^\star$ to the original posterior π, and thus transfer the regularity of π (namely, regularity of V) to regularity of $\pi^\star$.

Theorem 2.4 (Self-consistency equations for $\pi^\star$)**.** *Under* (**SLC**)*, any minimizer* $\pi^\star(\mathrm{d}z) = p^\star(\mathrm{d}z_1)\, q^\star(\mathrm{d}z_{-1} \mid z_1)$ *for* (SSVI) *with differentiable density satisfies the following equations*[‡]*:*

$$\begin{aligned} p^\star(z_1) &\propto \exp\Big(-\int_0^{z_1} \int \partial_1 V(s, z_{-1})\; q^\star(\mathrm{d}z_{-1} \mid s)\,\mathrm{d}s\Big)\,, \\ q_i^\star(z_i \mid z_1) &\propto \exp\Big(-\int V(z_1, z_i, z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z_1)\Big)\,, \qquad i \in [d] \setminus \{1\}\,. \end{aligned} \tag{14}$$

Our result is the first to extend the existing fixed-point characterization from the mean-field setting (Lacker et al., 2024, Theorem 1.1). Notably, when conditioning on the root variable z_1, the fixed-point equation in (14) for the conditional minimizer $q^\star(\cdot \mid z_1)$ recovers the fixed-point equation for the MFVI minimizer corresponding to the conditional target measure $\pi_{-1}(\cdot \mid z_1)$. We remark that the proof of Theorem 2.4 builds on the first-order optimality condition of SSVI in the Wasserstein geometry, which requires the SSVI minimizer $\pi^\star$ to be *a priori* differentiable. Since our focus here is to characterize the structure of the minimizer, we leave verification of this assumption to future work. Nonetheless, we will see in Theorem 2.8 that both $p^\star$ and $q^\star(\cdot \mid z_1)$ are *a posteriori* log-concave under further assumptions, ensuring their almost everywhere differentiability by Rockafellar (1970, Theorem 25.3). From this point onward, we will implicitly assume that $\pi^\star$ is differentiable in all subsequent results.

To proceed further, we require some assumptions on the Hessian of V. In light of the dynamic programming principle, it is natural to partition the assumptions between the contribution of the root of the Hessian, $\partial_{11} V$, and its principal minor $(\nabla^2 V)_{-1}$.[§] To this end, let $0 < \ell_V \le L_V$ and $L_V' > 0$; then, we assume

[‡]The integrals in (14) are well-defined due to the integrability of V and ∇V by (**SLC**) and the required growth condition on V in (12).

[§]Recall that this means we remove the first row and column of the Hessian.

$$\partial_{11} V \leq \frac{1}{2} L'_V \,, \tag{R}$$

$$0 \prec \ell_V I_{d-1} \preceq (\nabla^2 V)_{-1} \,, \tag{P1}$$

$$(\nabla^2 V)_{-1} \preceq L_V I_{d-1} \,. \tag{P2}$$

Proposition 2.1 tells us that the inner optimization problem is precisely a MFVI problem with respect to the leaf variables. The assumptions (**P1**) and (**P2**) therefore correspond to the assumptions in, e.g., Arnese and Lacker (2024); Lacker et al. (2024); Lavenant and Zanella (2024). Assumption (**R**), also standard, controls the curvature of V in the direction of the root variable. To complete our assumptions, we need to describe how these terms interact through a novel *root domination criteria*, which is necessary for SSVI. We require that

$$\partial_{11} V - \frac{\| \sum_{j=2}^d (\partial_{1j} V)^2 \|_{L^\infty}}{\ell_V} \geq \ell'_V > 0 \,. \tag{RD}$$

Geometrically, assumption (**RD**) guarantees that the potential V is more curved along the z_1 direction than it is "twisted" by the interactions between z_1 and $z_2, \ldots, z_d$. It can also be viewed as a uniform requirement on the Schur complement of $\nabla^2 V$ (see, e.g., Zhang, 2006, Theorem 1.12). Indeed, note that for V to be positive definite in general, the following criterion must hold[¶]

$$\partial_{11} V - (\nabla^2_{-1,1} V)^\top \left((\nabla^2 V)_{-1}\right)^{-1} (\nabla^2_{-1,1} V) > 0 \,.$$

Making use of (**P1**), it is clear that assumption (**RD**) is a quantitative version of this inequality, where we require uniform control on the lower bound.

Remark 2.5. When $V(x) = \sum_{i=1}^d V_i(x_i)$, i.e., π is a product measure, it is easy to see that (**RD**) is trivially satisfied, and $\ell'_V = \inf_{x \in \mathbb{R}^d} \partial_{11} V(x)$.

The following lemma establishes that, should a posterior π satisfy the aforementioned conditions, it is itself strongly log-concave.

Lemma 2.6. *If π satisfies (**P1**) and (**RD**), then π satisfies (**SLC**) with parameter $\ell_V \wedge \ell'_V$. Additionally, if (**R**) and (**P2**) hold, then π is $L_V \vee (L'_V/2)$-log-smooth.*

By Theorem 2.2, the strong log-concavity of π implies the existence of a unique star-structured solution.

Corollary 2.7 (Uniqueness)**.** *Suppose $\pi \propto \exp(-V)$ satisfies (**P1**) and (**RD**). Then there exists a unique star-structured minimizer of (SSVI).*

Our first main result of this section is the following theorem, which establishes strong log-concavity and log-smoothness of the marginals of the star-structured minimizer.

Theorem 2.8 (Log-concavity and log-smoothness of $\pi^\star$)**.** *Suppose $\pi \propto \exp(-V)$ satisfies (**P1**) and (**RD**). Then, for the SSVI minimizer $\pi^\star$:*

[¶]Here, $\nabla^2_{-1,1} V$ represents the first column of the Hessian $\nabla^2 V$ without the first entry.

(i) the marginal distribution $p^\star$ is ℓ'_V-log-concave,

(ii) for every $i \geq 2$, the conditional distribution $q_i^\star(\cdot \mid z_1)$ is ℓ_V-log-concave for any fixed $z_1 \in \mathbb{R}$.

In addition, if (**R**) *and* (**P2**) *hold, then $p^\star$ is L'_V-log-smooth and $q_i^\star(\cdot \mid z_1)$ is L_V-log-smooth for any fixed $z_1 \in \mathbb{R}$.*

Given the conditional independence structure, the properties of $q_i^\star(\cdot \mid z_1)$ for $i \geq 2$ can be derived from the self-consistency equations (14). However, analyzing $p^\star$ is substantially more intricate, as it depends on $q_i^\star(\cdot \mid z_1)$, which in turn depends on $p^\star$. To handle this, we employ a novel stability bound for mean-field variational inference developed by some of the authors (Sheng et al., 2025b). Combined with the self-consistency equation for $p^\star$, this yields the desired regularity result. In the special case where π is a product measure, we have $\ell'_V = \ell_V$, and our result recovers the properties of MFVI minimizer in Lacker et al. (2024). Aside from being of theoretical interest in its own right, Theorem 2.8 will later act as a cornerstone for our computational guarantees in Section 3.4.

Additionally, the aforementioned regularity results will allow us to establish the first approximation guarantees for the SSVI minimizer with respect to the true posterior.

Theorem 2.9 (Approximation guarantee for SSVI)**.** *Let $\pi \propto \exp(-V)$ be such that V satisfies* (**R**)*,* (**P1**)*, and* (**RD**)*. Then*

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) = \mathsf{KL}\left(\pi^\star \parallel \pi\right) \leq \frac{L'_V}{2\ell'_V \ell_V^2} \sum_{i \geq 2} \sum_{j > i} \mathbb{E}_{\pi^\star}\left[\left(\partial_{ij} V(Z)\right)^2\right]. \tag{15}$$

The proof of this result is based on delicate applications of functional inequalities for log-concave measures and is deferred to Section A. Before specializing this result to concrete examples, we present some general remarks here. Note that, as expected, when the target measure π already respects a star-graph structure, the partial derivatives $\partial_{ij} V$ on the right-hand side of (15) vanish.

It is natural to ask how the star-structured approximation compares to mean-field approximation. By Lemma 2.6, the target measure π is $(\ell_V \wedge \ell'_V)$-log-concave and, in this setting, Lacker et al. (2024, Theorem 1) establish an upper bound for the MFVI minimizer $\bar{\pi}$ which reads

$$\mathsf{KL}\left(\bar{\pi} \parallel \pi\right) \leq \frac{1}{(\ell_V \wedge \ell'_V)^2} \sum_{i \geq 1} \sum_{j > i} \mathbb{E}_{\bar{\pi}}\left[\left(\partial_{ij} V(Z)\right)^2\right].$$

Compared to our results, the mean-field bound introduces terms related to $\{\partial_{1i} V\}_{i=1}^d$. Consequently, when the association between the root and leaf nodes is more prominent than the inter-leaf associations, we expect a strong improvement when using SSVI compared to MFVI. The gap between the approximation errors of SSVI and MFVI is highlighted in the following Gaussian example.

2.3 Examples

We now instantiate our approximation guarantee from Theorem 2.9 on several examples that appear in the literature.

2.3.1 Gaussian posterior

We first consider the setting where π is itself a multivariate Gaussian distribution. Existing results, such as those by Lacker et al. (2024), state that the minimizer remains Gaussian when the variational family consists of product measures—the same is true in our setting.

Theorem 2.10. *Let $\pi = \mathcal{N}(m, \Sigma)$ with $m \in \mathbb{R}^d$ and $\Sigma \succ 0$, where the covariance has entries σ_{ij}. Then, (SSVI) admits a unique minimizer $\pi^\star(\mathrm{d}z) = p^\star(\mathrm{d}z_1)\, q^\star(\mathrm{d}z_{-1} \mid z_1)$, where $p^\star = \mathcal{N}(m_1, \sigma_{11})$ and $q^\star(\cdot \mid z_1) = \mathcal{N}(m^\star_{z_1}, \Sigma^\star_{|1})$ with*

$$\begin{aligned} m^\star_{z_1,i} &= m_i + (\sigma_{11})^{-1}\, \sigma_{1i}\, (z_1 - m_1)\,, \\ \Sigma^\star_{|1} &= \mathrm{diag}((\sigma^{22})^{-1}, \dots, (\sigma^{dd})^{-1})\,, \end{aligned}$$

where σ^{ij} is the (i,j)-th entry of Σ^{-1}. In other words, $\pi^\star = \mathcal{N}(m^\star, \Sigma^\star)$ where $\Sigma^\star = (\sigma^\star_{ij})_{1 \le i,j \le d} \in \mathbb{R}^{d \times d}$ satisfies

$$\begin{aligned} \sigma^\star_{1i} &= \sigma_{1i}\,, && i \in [d]\,, \\ \sigma^\star_{ii} &= (\sigma^{ii})^{-1} + \frac{\sigma^2_{1i}}{\sigma_{11}}\,, && i \in [d] \setminus \{1\}\,, \\ \sigma^\star_{ij} &= \frac{\sigma_{1i}\sigma_{1j}}{\sigma_{11}}\,, && i \neq j \in [d] \setminus \{1\}\,. \end{aligned} \tag{16}$$

In this case, recall that $\nabla^2 V = \Sigma^{-1}$ and thus $\partial_{ij} V = \sigma^{ij}$. The assumptions (**R**), (**P1**), and (**P2**) are trivially satisfied. Although assumption (**RD**) is stronger than the Schur complement condition for $\Sigma \succ 0$ and need not hold in general, $\pi^\star_1$ is still log-concave and log-smooth by Gaussianity.

We now offer a probabilistic interpretation of these results and how the SSVI minimizer compares structurally to the MFVI minimizer. When $m = 0$, for $(Z_1, \dots, Z_d) \sim \pi^\star$, Theorem 2.10 states that Z_i can be expressed as a linear combination of Z_1 and exogenous variables $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ independent of Z_1:

$$Z_1 = \sqrt{\sigma_{11}}\, \xi_1\,, \qquad Z_i = \frac{\sigma_{1i}}{\sigma_{11}}\, Z_1 + \frac{1}{\sqrt{\sigma^{ii}}}\, \xi_i\,, \quad \text{for } i \ge 2\,. \tag{17}$$

On the other hand, let $\bar{\pi}$ be the mean-field approximation to π with $(\bar{Z}_1, \dots, \bar{Z}_d) \sim \bar{\pi}$. Lacker et al. (2024, Theorem 1.1) show that each $\bar{Z}_i$ is precisely a rescaled version of this exogenous variable ξ_i, namely that

$$\bar{Z}_i = \frac{1}{\sqrt{\sigma^{ii}}}\, \xi_i\,, \qquad \text{for } i \in [d]\,.$$

SSVI improves upon MFVI by adding the correction term $(\sigma_{1i}/\sigma_{11})\, Z_1$ to the MFVI minimizer. The following corollary quantifies this improvement, showing that the gain is precisely given by the logarithmic ratio of the variances of the root variable under the respective minimizers.

Corollary 2.11. *In the setting of Theorem 2.10, it holds that*

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\,(\mu \parallel \pi) - \inf_{\mu \in \mathcal{P}(\mathbb{R})^{\otimes d}} \mathsf{KL}\,(\mu \parallel \pi) = -\frac{1}{2} \log\left(\frac{\sigma_{11}}{(\sigma^{11})^{-1}}\right) \le 0\,.$$

2.3.2 Bayesian generalized linear models with various priors

We now specialize our approximation results to a family of scenarios where the posterior is defined in terms of a generalized linear model (GLM). As GLMs are standard tools for statistical inference, we only briefly outline the basic principles and terminology required; see McCullagh and Nelder (1989) for more details.

Given a collection of covariate-response pairs $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathcal{Y}$, a GLM posits the following data-generating process

$$Y_i \mid \beta \sim \mathsf{EF}[X_i^\top \beta; \sigma]\,, \tag{18}$$

where $\mathsf{EF}[\cdot;\sigma]$ is defined as a probability measure on $\mathcal{Y}$, called an *exponential family*

$$\frac{\mathrm{d}\mathsf{EF}[X_i^\top \beta;\sigma]}{\mathrm{d}\mathsf{m}}(y) = \exp\!\left(\frac{yX_i^\top\beta - \psi(X_i^\top\beta)}{c(\sigma)}\right),$$

where ψ plays the role of the normalizing constant (called the log-partition function), and m can be discrete (when $\mathcal{Y}$ is discrete) or the Lebesgue measure (when $\mathcal{Y}$ is a subset of $\mathbb{R}$).

In the Bayesian version of GLMs, we further posit a prior on the coefficients $\beta \in \mathbb{R}^d$

$$\beta \sim \pi_0\,. \tag{19}$$

Conditioning on the covariates, (18)–(19) defines the data generative process for $y_{1:n}$. It is worth emphasizing that many standard regression models are special cases of GLMs.[‖]

In the analysis below, we consider two families of hierarchical priors for π_0: (i) location-scale priors, and (ii) spike-and-slab priors with one-dimensional debiasing.

Location-family prior Consider a location family of priors of the form:

$$\beta_1, \ldots, \beta_d \overset{\text{i.i.d.}}{\sim} \pi_0\left(\cdot \mid \vartheta\right) = \exp(-\varrho(\cdot - \vartheta))\,, \qquad \vartheta \sim \pi_0 = \exp(-\mathsf{g})\,, \tag{20}$$

where $\varrho : \mathbb{R} \to \mathbb{R}$ and $\mathsf{g} : \mathbb{R} \to \mathbb{R}$ are some potential functions. For example, $\varrho(t) = t^2/(2s^2)$ corresponds to a Gaussian distribution, while $\varrho(t) = t/s + 2\log\left(1 + \exp\left(-t/s\right)\right)$ corresponds to a logistic distribution.

Denote by $A \coloneqq \frac{1}{c(\sigma)}\boldsymbol{X}^\top\boldsymbol{X}$ and $w \coloneqq \frac{1}{c(\sigma)}\boldsymbol{X}^\top\mathbf{y}$, where $\boldsymbol{X} = [X_1, \ldots, X_n]^\top$ and $\mathbf{y} \coloneqq [Y_1, \ldots, Y_n]^\top$.

The full posterior $\pi_{\mathrm{loc}}(\mathrm{d}\vartheta, \mathrm{d}\beta) \propto \exp(-V_{\mathrm{loc}}(\vartheta,\beta))\,\mathrm{d}\vartheta\,\mathrm{d}\beta$ is given by:

$$V_{\mathrm{loc}}(\vartheta,\beta) = \mathsf{g}(\vartheta) - w^\top\beta + \sum_{i=1}^n \frac{\psi(X_i^\top\beta)}{c(\sigma)} + \sum_{j=1}^d \varrho(\beta_j - \vartheta)\,. \tag{21}$$

An example of a full generative process is shown in Figure 1(a) as a graphical model. The latent variable ϑ models a uniform shift in $\beta_{1:d}$ away from 0. While the prior for $(\vartheta, \beta_{1:d})$

[‖]For instance: Linear regression corresponds to the case $\mathcal{Y} = \mathbb{R}$, $\psi(t) = t^2/2$, and $c(\sigma) = \sigma^2$. Logistic regression corresponds to the binary class label in $\mathcal{Y} = \{0, 1\}$, $\psi(t) = \log(1 + \exp(t))$, $c(\sigma) = 1$. Poisson regression corresponds to $\mathcal{Y} = \mathbb{N}$, $\psi(t) = \exp(t)$, and $c(\sigma) = 1$.

follows a star graphical structure, the full posterior need not be a star graph. Nevertheless, we can compute a star-structured approximation and compute the approximation quality.

To state our result, we introduce a key definition. For $\beta \in \mathbb{R}^d$, denote the matrix $\mathcal{A}(\beta) := \frac{1}{c(\sigma)} \boldsymbol{X}^\top D(\beta) \boldsymbol{X}$ where $D(\beta)$ is a diagonal matrix with $D_{ii}(\beta) = \psi''(X_i^\top \beta)$ for $i = 1, \ldots, n$.

Theorem 2.12 (Approximation quality for location-family GLMs)**.** *Let $\pi_{\mathrm{loc}} \propto \exp(-V_{\mathrm{loc}})$ where V_{loc} is given by (21). Assume that $b \le \psi'' \le B$ for some $B \ge b > 0$ and $0 \le \varrho'' \le R$, and that the interaction matrix satisfies $A \succeq aI$ for some $a > 0$. Suppose that $\alpha \le \mathsf{g}'' \le L$ for some $\alpha \in \mathbb{R}$ such that $\ell_V' := \alpha - \frac{dR^2}{ab} > 0$. Then, under the model specified by (18)–(20), the star-structured variational approximation $\pi^\star_{\mathrm{loc}}$ satisfies the following bound:*

$$\mathsf{KL}\left(\pi^\star_{\mathrm{loc}} \parallel \pi_{\mathrm{loc}}\right) \le \frac{dR + L}{a^2 b^2 \ell_V'} \left\| \sum_{i=1}^d \sum_{j>i} \mathcal{A}_{ij}^2 \right\|_{L^\infty} .$$

It is known that the log-partition function ψ is convex in general exponential families (Wainwright and Jordan, 2008, Proposition 3.1). Theorem 2.12 further assumes ψ to be b-convex, which holds when the variance of the distribution $\mathsf{EF}(\cdot; \sigma)$ is uniformly lower bounded by b (Brown, 1986, Theorem 2.2).

Theorem 2.12 appears to be the first *quantitative* VI guarantee for GLMs. Our result establishes a star approximation guarantee of order $O\Big(\| \sum_{i=1}^d \sum_{j>i} \mathcal{A}_{ij}^2 \|_{L^\infty}\Big)$. This recovers the condition identified for the asymptotic correctness of mean-field approximation in canonical GLMs with compact support (see Definition 2 and Theorem 2 of Mukherjee et al., 2024). When the off-diagonal entries of $\mathcal{A}(\beta)$ are uniformly small across all β, the exact posterior of the GLM is accurately approximated by a star-structured variational distribution. In the case of a linear model, we obtain the simple form $\mathcal{A}(\beta) = A$ (see Proposition 2.13).

We next specialize Theorem 2.9 to a Bayesian linear model with Gaussian location prior:

$$\mathbf{y} = \boldsymbol{X}\beta + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)\,, \quad \beta_1, \ldots, \beta_d \sim \mathcal{N}(\vartheta, \tau^{-2})\,, \quad \vartheta \sim \exp(-\mathsf{g})\,. \tag{22}$$

Proposition 2.13. *Let $A := \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X}$. Assume that $aI \preceq A$ for some $a \ge 0$. Suppose that $\alpha \le \mathsf{g}'' \le L$ such that $\ell_V' = \alpha + \frac{da\tau^2}{a+\tau^2} > 0$. Under model (22), the SSVI minimizer $\pi^\star_{\mathrm{loc}}$ satisfies:*

$$\mathsf{KL}\left(\pi^\star_{\mathrm{loc}} \parallel \pi_{\mathrm{loc}}\right) \le \frac{d\tau^2 + L}{\ell_V' \, (a + \tau^2)^2} \sum_{i=1}^d \sum_{j>i} A_{ij}^2 \,. \tag{23}$$

Proposition 2.13 establishes a star approximation gap of order $O(\sum_{i=1}^d \sum_{j>i} A_{ij}^2)$. When the off-diagonal entries of A are sufficiently small, the exact posterior is accurately approximated by a star-structured variational distribution.

For Proposition 2.13 to hold, we do not require the interaction matrix A to be positive definite, nor g to be strongly convex. When $a = 0$, it suffices that the log-concavity parameter $\alpha > 0$ to obtain a non-trivial upper bound in (23). On the other hand, α could also be negative, provided that $\alpha > -\frac{da\tau^2}{a+\tau^2}$. Moreover, Proposition 2.13 yields a slightly stronger statement than Theorem 2.12: in Theorem 2.12 we only assume that ϱ is convex, whereas in Proposition 2.13 we use the fact that $\varrho'' = \tau^2$.

Spike-and-slab prior Another common setting in statistical inference is to impose a spike-and-slab prior on $\beta_{2:d}$ and a bias-correcting prior on β_1 (Castillo et al., 2024). To be precise, for $j \geq 2$, β_j is drawn from either a spike distribution ν_0 or a more diffuse slab distribution ν_1, with a mixture weight of $\eta \in [0,1]$:

$$\beta_2, \ldots, \beta_d \overset{\text{i.i.d.}}{\sim} \eta\nu_0 + (1-\eta)\nu_1 \,. \tag{24}$$

Following the approach in George and McCulloch (1993), we specify ν_0, ν_1 as normal distributions: $\nu_0 := \mathcal{N}(0, \tau_0^{-2})$, $\nu_1 := \mathcal{N}(0, \tau_1^{-2})$ where $\tau_1^2 < \tau_0^2$. Now, conditioned on $\beta_{2:d}$, we draw β_1 from the following prior:

$$\beta_1 \mid \beta_{2:d} \sim \exp\Big(-\mathsf{g}\Big(\cdot + \sum_{j=2}^{d} \gamma_j \beta_j\Big)\Big)\,, \tag{25}$$

where $\gamma_j := \boldsymbol{X}_1^\top \boldsymbol{X}_j / \|\boldsymbol{X}_1\|_2^2$ is chosen to be the rescaled correlation between covariates 1 and j; here $\boldsymbol{X}_j$ is the j^{th} column of $\boldsymbol{X}$. For linear regression, the resulting posterior is shown to debias β_1 (Yang, 2019; Castillo et al., 2024). For GLMs, a similar debiasing approach exists in the frequentist setting (van de Geer et al., 2014). Figure 1 (b) shows the full generative process, where the full posterior is given by $\pi_{\text{ss}} \propto \exp(-V_{\text{ss}})$ with

$$V_{\text{ss}}(\beta) = \sum_{i=1}^{n} \frac{\psi(X_i^\top \beta)}{c(\sigma)} - w^\top \beta - \sum_{j=2}^{d} \log\left(\eta\nu_0(\beta_j) + (1-\eta)\nu_1(\beta_j)\right) + \mathsf{g}\Big(\beta_1 + \sum_{j=2}^{d} \gamma_j \beta_j\Big)\,. \tag{26}$$

Comparison with Castillo et al. (2024). The previous work of Castillo et al. (2024) studies a high-dimensional linear model with a prior where the distribution of $\beta_1 \mid \beta_{2:d}$ is the same as (25) but replaces the continuous spike-and-slab prior (24) with a discrete spike-and-slab prior.

For inference, Castillo et al. (2024) applies MFVI to $\beta_{2:d}$ through a reparameterization scheme $\beta_1 \mapsto \beta_1' := \beta_1 + \sum_{j=2}^{d} \gamma_j \beta_j$ which decorrelates the posterior of β_1' and $\beta_{2:d}$. The approach, however, has two limitations: (1) the decorrelation approach fails in GLMs, where ψ induces non-linear dependencies between β_1 and $\beta_{2:d}$ that cannot be easily orthogonalized, and (2) MFVI restricts the marginal law of $\beta_{2:d}$ to be a product measure. These are both limitations which we sidestep with our theory.

Theorem 2.14 (Approximation quality for spike-and-slab GLMs)**.** *Let $\pi_{\text{ss}} \propto \exp(-V_{\text{ss}})$ where V_{ss} is given by (26). Assume that $b \leq \psi'' \leq B$ for some $B \geq b > 0$, and that the interaction matrix satisfies $a_d I \preceq A \preceq a_1 I$ for some $a_1 \geq a_d > 0$. Assume that $\alpha \leq \mathsf{g}'' \leq L$ such that:*

(i) $\ell_V' := bA_{11} + \alpha - \frac{2\sum_{j=2}^{d} \gamma_j^2 L^2}{ba_d} - \frac{2\,(Ba_1 - ba_d)\,(BA_{11} - ba_d)}{ba_d} > 0$; *and*

(ii) $\ell_V := ba_d + \tau_1^2 - 2(\tau_0^2 - \tau_1^2)\log\Big(1 + \frac{\eta\tau_0}{(1-\eta)e\tau_1}\Big) > 0.$

Then, under the model specified in (18)–(24)–(25), the star-structured variational approximation $\pi_{\text{ss}}^\star$ satisfies the following bound:

$$\mathsf{KL}\left(\pi_{\text{ss}}^\star \,\|\, \pi_{\text{ss}}\right) \leq \frac{2\,(BA_{11} + L)}{\ell_V^2 \ell_V'} \left(\Big\| \sum_{i=2}^{d} \sum_{j>i} \mathcal{A}_{ij}^2 \Big\|_{L^\infty} + L^2 \sum_{i=2}^{d} \sum_{j>i} \gamma_i^2 \gamma_j^2 \right).$$

The condition $\ell_V' > 0$ is readily satisfied when a_1 and a_d are close and α is sufficiently large. In the special case where $A_{1j} = 0$ for all $j \geq 2$, we have $\gamma_j = 0$ by the definition of γ_j, thus $\ell_V' > 0$ holds when we specify the prior g such that $\alpha > \frac{2\,(Ba_1 - ba_d)\,(BA_{11} - ba_d)}{ba_d} - bA_{11}$. The condition $\ell_V > 0$ is guaranteed when a_d grows with sample size so that it dominates the remainder term which only depends on the fixed prior. This behavior is typical in well-behaved models, as the next result (Proposition 2.15) illustrates.

We now specialize our results in the setting of Castillo et al. (2024) where

$$\mathbf{y} = \boldsymbol{X}\beta + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0, I)\,,$$

and each β_i, $i \geq 2$, is drawn i.i.d. from a continuous spike-and-slab prior:

$$\beta_1 \mid \beta_{2:d} \sim \mathcal{N}\Big(-\sum_{j=2}^{d} \gamma_j \beta_j, \tau^{-2}\Big), \quad \beta_2, \cdots, \beta_d \overset{\text{i.i.d.}}{\sim} \eta\nu_0 + (1-\eta)\nu_1\,.$$

We say that a random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is from the Σ-Gaussian ensemble if the rows x_i of $\boldsymbol{X}$ are drawn i.i.d. from the $\mathcal{N}(0, \Sigma)$ distribution. The following guarantee holds:

Proposition 2.15. *Let $A = \boldsymbol{X}^\top \boldsymbol{X}$ where $\boldsymbol{X}$ is from the Σ-Gaussian ensemble. Let λ_d, λ_1 denote the smallest and largest eigenvalues of Σ. Suppose that $d, n \to \infty$ and $d/n = o(1)$. Suppose further that $\lambda_1 = O(1)$, $\lambda_1/\lambda_d = o(\sqrt{n/d})$, and*

$$\ell_V' := \frac{n}{2}\left(\Sigma_{11} - \frac{8\sum_{j=2}^{d} |\Sigma_{1j}|^2}{\lambda_d}\right) > 0\,. \tag{27}$$

Then, the approximation error of $\pi_{\text{ss}}^\star$ satisfies

$$\mathsf{KL}\left(\pi_{\text{ss}}^\star \parallel \pi_{\text{ss}}\right) \leq \frac{8\,(A_{11} + \tau^2)}{n^2 \ell_V' \lambda_d^2} \sum_{i=2}^{d} \sum_{j>i} \Big(A_{ij}^2 + \tau^4 \gamma_i^2 \gamma_j^2\Big)\,,$$

with high probability.

Condition (27) elucidates the root domination criteria: being diagonally dominant in a Gaussian ensemble regression means that the marginal variance of the root node Σ_{11} dominates the marginal covariances between the root and leaf nodes, in the sense that $\Sigma_{11} > (4/\lambda_d)\sum_{j=2}^{d} |\Sigma_{1j}|^2$. A larger gap implies that the SSVI approximation is more precise.

3 Computational guarantees via optimal transport

We now turn our attention to computational aspects of solving (SSVI). We find it helpful to compare to the mean-field setting as in Section 2.

A popular algorithm for solving the mean-field variational inference problem is known as *coordinate-ascent variational inference*, or CAVI (Blei et al., 2017). For a given posterior π, this algorithm performs the following updates

$$\pi_j^{(k+1)} = \underset{\mu \in \mathcal{P}(\mathbb{R})}{\arg\min}\, \mathsf{KL}(\mu \otimes \pi_{-j}^{(k)} \parallel \pi)\,, \tag{28}$$

and we cycle through the coordinates $j \in \{1, \dots, d\}$. Despite forming the backbone of Bayesian computation, convergence guarantees for CAVI have only been recently established under the same assumptions as our work, namely log-smoothness and strong log-concavity of π (Arnese and Lacker, 2024; Lavenant and Zanella, 2024; Bhattacharya et al., 2025).[**] The iterates (28) are only available in closed form if we further restrict the optimization to take place over a conjugate families, which can be far from typical in practical applications (Wang and Blei, 2013).

The algorithmic story for structured variational inference is quite different. Even in the star-structured setting, the equivalence to a dynamic programming principle makes it difficult to derive an algorithm similar to the CAVI update scheme. For general structured VI, it is popular to perform gradient descent on a prespecified parametric family; see Hoffman and Blei (2015). In this section, we provide computational guarantees for computing the SSVI minimizer using ideas from optimal transport (Villani, 2009; Chewi et al., 2025) and non-parametric approximation theory (Wasserman, 2006; Tsybakov, 2009). In Section 3.1, we show that (SSVI) is equivalent to a convex optimization problem (T-SSVI) that occurs at the level of transport maps. This equivalence provides an alternative proof of the existence and uniqueness of the SSVI minimizer; see Section 3.2. In Section 3.3, we derive a Caffarelli-type regularity result for the optimizer of (T-SSVI), focusing in particular on its regularity with respect to the root variable. Section 3.4 is devoted to the computational method for finding a near-optimal transport map by approximating (T-SSVI) via a finite-dimensional optimization problem. We present both theoretical guarantees for the approximation and a projected gradient descent algorithm with convergence guarantees.

3.1 From star-structured measures to star-separable maps

Our approach is inspired by the popular technique of "normalizing flows" in the machine learning literature, which suggests that learning vector-valued functions between measures is easier than learning the density of a measure itself; see the review article by Kobyzev et al. (2020). We henceforth call $T : \mathbb{R}^d \to \mathbb{R}^d$ a transport map between two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ if, for $X \sim \mu$, $T(X) \sim \nu$. Our first objective is to rewrite (SSVI) as an optimization problem over suitable transport maps.

To this end, let $\rho = \bigotimes_{i=1}^d \rho_i \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R})^{\otimes d}$ be a fixed reference measure—we will always take $\rho = \mathcal{N}(0, I)$. A transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ emanating from ρ is called *star-separable* if there exists functions $T_1 : \mathbb{R} \to \mathbb{R}$ and $T_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (for $i \in \{2, \dots, d\}$) such that

$$x \mapsto T(x_1, \dots, x_d) = (z_1, T_2(x_2 \mid z_1), \dots, T_d(x_d \mid z_1)) \quad \text{and} \quad z_1 = T_1(x_1)\,. \tag{29}$$

Additionally, we require that T_1 (resp. T_i for $i \in \{2, \dots, d\}$) to be *(strictly)* increasing in x_1 (resp. in x_i for any fixed $z_1 \in \mathbb{R}$). Let $\mathcal{T}_{\mathrm{star}}$ be the collection of star-separable maps emanating from ρ.[††] The set $\mathcal{T}_{\mathrm{star}}$ is a special class of Knothe–Rosenblatt (KR) maps (Rosenblatt, 1952; Knothe, 1957), which are coordinate-wise transformations depending on the structure of preceding coordinates. We will explore this connection in greater detail in Section 4.

[**]Interestingly, Lavenant and Zanella (2024) establish improved convergence rates for CAVI under a random coordinate scan approach than a deterministic scheme.

[††]We omit the explicit dependence on ρ for ease of notation.

We claim that the original SSVI problem, an optimization over star-structured probability measures, is equivalent to the following optimization which takes place over star-separable maps:

$$\inf_{T \in \mathcal{T}_{\text{star}}} \mathsf{KL}\left(T_{\#}\rho \parallel \pi\right). \tag{T-SSVI}$$

In short, this equivalence implies that $\pi^\star$ is a solution to (SSVI) if and only if there exists $T^\star \in \mathcal{T}_{\text{star}}$ which solves (T-SSVI), whence

$$\pi^\star = (T^\star)_{\#}\rho\,.$$

Indeed, as $\pi^\star \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ by Theorem 2.8, we may restrict the minimization (SSVI) to the set $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) \cap \mathcal{C}_{\text{star}}$, and thus, without loss of generality, we may assume $\mathcal{C}_{\text{star}}$ is a subset of $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$. Therefore, Brenier's theorem (see, e.g., Villani, 2003, Theorem 2.32) implies $\mu \in \mathcal{C}_{\text{star}}$ if and only if $\mu = T_{\#}\rho$ for some $T \in \mathcal{T}_{\text{star}}$, which establishes the equivalence between (SSVI) and (T-SSVI).

This lifting yields a remarkable property: although $\mathcal{C}_{\text{star}}$ is not convex, its image $\mathcal{T}_{\text{star}}$ forms a convex subset of $L^2(\rho)$. We demonstrate this in the following proposition.

Proposition 3.1. *The set of star-separable transport maps $\mathcal{T}_{\text{star}}$ is convex.*

This result highlights an interesting computational paradigm. While Theorem 2.2 guarantees the existence of a unique minimizer to (SSVI), it does not provide a means for computing this minimizer. In fact, the set $\mathcal{C}_{\text{star}}$ is not convex under linear geometry — indeed, a mixture of two star-structured graphical models need not itself be star-structured. In contrast, the equivalence between (SSVI) and (T-SSVI), combined with Proposition 3.1, reveals the possibility of a convex optimization scheme over the space of star-separable transport maps. In the sequel, we will flesh out this idea by developing the theoretical properties of the class $\mathcal{T}_{\text{star}}$, and conclude by presenting a gradient-based algorithm that is computationally tractable and comes with convergence guarantees.

3.2 Existence and uniqueness of the minimizer via (strong) convexity

We now introduce a notion of distance over $\mathcal{T}_{\text{star}}$. Inspired by the literature on *linearized* optimal transport (Wang et al., 2013), a natural candidate is the $L^2(\rho)$ distance $\|T - \tilde{T}\|_{L^2(\rho)}$ between two maps $T, \tilde{T} \in \mathcal{T}_{\text{star}}$. This metric dominates the well-known adapted Wasserstein distance, which preserves conditional independence under limits (see Section B for a discussion). In contrast, conditional independence is not preserved under weak convergence; see Lauritzen (1996) for a counterexample. Thus, the topology induced by the $L^2(\rho)$ distance is particularly well-suited to our setting, as it aligns with the structural properties of conditional distributions that underlie SSVI.

The above preparation brings us to the following theorem, which highlights the requirements of strong log-concavity of π for minimizing (T-SSVI), just as it was necessary for (SSVI).

Theorem 3.2 (Existence and uniqueness)**.** *Suppose* (**SLC**) *holds, and let $\mathcal{T} \subseteq \mathcal{T}_{\text{star}}$ be a convex set. Then*

(i) the functional $T \mapsto \mathsf{KL}\left(T_{\#}\rho \parallel \pi\right)$ *is* α*-convex over* $\mathcal{T}$*, and*

(ii) there exists a unique minimizer to (T-SSVI).

To understand this claim, one can begin by expanding the objective (T-SSVI) for any $T \in \mathcal{T} \subseteq \mathcal{T}_{\text{star}}$,

$$\mathsf{KL}(T_{\#}\rho \parallel \pi) = \int V(T(x))\,\rho(\mathrm{d}x) - \int \log\det DT(x)\,\rho(\mathrm{d}x) + \text{CONST}\,, \tag{30}$$

where the constant consists of the entropy of ρ and the unknown normalizing constant of π. We see two terms: one involving the potential function V and the other involving the eigenvalues of the Jacobian of the transport map T. The following lemma explains the properties of the first term.

Lemma 3.3. *Let* $\mathcal{T} \subset \mathcal{T}_{\text{star}}$ *be a convex set and* $V : \mathbb{R}^d \to \mathbb{R}^d$ *which is* α*-convex and* β*-smooth in the usual sense. Then the functional* $T \mapsto \int V(T(x))\,\rho(\mathrm{d}x)$ *is* α*-convex and* β*-smooth over* $\mathcal{T}$*.*

For the second term, note that for any $T \in \mathcal{T} \subset \mathcal{T}_{\text{star}}$, the Jacobian DT is a lower triangular matrix with non-negative diagonals, thus

$$T \mapsto -\int_{\mathbb{R}^d} \log\det(DT(x))\,\rho(\mathrm{d}x)$$

is a *convex* functional. Combining these claims, one arrives at the first conclusion of Theorem 3.2 that $T \mapsto \mathsf{KL}\left(T_{\#}\rho \parallel \pi\right)$ is α-convex over $\mathcal{T}$. Rigorous proofs are found in Section B.

Remark that the statement of Theorem 3.2 holds true for arbitrary convex subsets $\mathcal{T} \subseteq \mathcal{T}_{\text{star}}$. This point will be crucial in Section 3.4, where we will choose a particular subset of the family of star-separable maps to derive a projected gradient method.

3.3 Theoretical properties of star-separable maps

Before diving into the final algorithmic details of our approach, we investigate some novel theoretical aspects of these transport maps, which are crucial for establishing the computational guarantees.

The regularity of the optimal star-separable map builds on a pointwise stability estimate of the map with respect to the root variable. This, in turn, requires a strengthened version of the root domination criteria (**RD**): for all $i \geq 2$,

$$\begin{aligned} &\Big\| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \Big\|_{L^\infty} < \ell_V\,, \qquad \|\partial_{1i} V\|_{L^\infty} < \infty\,, \qquad \text{and} \\ &\partial_{11} V - \frac{(d-1)\max_{i\geq 2} \|\partial_{1i} V\|_{L^\infty}^2}{\ell_V - \max_{i\geq 2} \| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \|_{L^\infty}} \geq \ell_V'\,. \end{aligned} \tag{**RD**+}$$

It is easy to see that (**RD**+) implies (**RD**) as

$$\frac{(d-1)\max_{i\geq 2} \|\partial_{1i} V\|_{L^\infty}^2}{\ell_V - \max_{i\geq 2} \| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \|_{L^\infty}} \geq \frac{\sum_{j=2}^d \|\partial_{1j} V\|_{L^\infty}^2}{\ell_V}\,.$$

Therefore, all results derived in Section 2 concerning the regularity of the SSVI minimizer $\pi^\star$ still hold under (**RD+**). Moreover, (**RD+**) enables us to derive *a priori* bounds on the derivatives of the density of the SSVI minimizer, as well as on those of the optimal star-separable map, as detailed in Lemma 3.4 and Theorem 3.5.

A fully rigorous analysis would require establishing the second-order differentiability of both $q^\star$ and $T^\star$, which requires arguments in the style of Caffarelli (1992, 1996) and lies beyond the scope of this work. In the present analysis, we assume that $q^\star$ and $T^\star$ are twice continuously differentiable, and focus on controlling the higher-order derivatives of the optimal star-separable map, which is crucial for establishing the desired computational guarantees in Section 3.4.

We begin with the following lemma, which provides a bound on the mixed derivative of $q_i^\star$ via a self-bounding argument under (**RD+**). This bound serves as the cornerstone for establishing control over the derivative of the optimal star-separable map with respect to the root variable.

Lemma 3.4 (Mixed derivative bound)**.** *Let* (**R**), (**P1**), (**P2**) *and* (**RD+**) *be satisfied. Let*

$$L := \max_{i\geq 2} \sup_{z_1, z_i} |\partial_1 \partial_i \log q_i^\star(z_i \mid z_1)| \,.$$

Then

$$L \leq \overline{L} := \frac{\ell_V \max_{i\geq 2} \|\partial_{1i} V\|_{L^\infty}}{\ell_V - \max_{i\geq 2} \| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \|_{L^\infty}} < +\infty \,. \tag{31}$$

Our main contributions in this subsection culminate in the following theorem, which states higher-order control on the optimal star-separable map; its proof appears in Section C.1.

Theorem 3.5 (Regularity for star-separable maps)**.** *Assume* (**R**), (**P1**), *and* (**P2**) *hold. Let* $T^\star$ *be the unique minimizer of* (T-SSVI) *over* $\mathcal{T}_{\mathrm{star}}$*, written*

$$x \mapsto T^\star(x) = (z_1, T_2^\star(x_2 \mid z_1), \ldots, T_d^\star(x_d \mid z_1)) \,, \qquad \text{with} \qquad z_1 = T_1^\star(x_1) \,.$$

If $T^\star$ *is twice-differentiable, then:*

(i) If (**RD**) *is satisfied, then for* $z_1 \in \mathbb{R}$ *and* $i \in \{2, \ldots, d\}$,

$$\sqrt{1/L_V'} \leq |\partial_{x_1} T_1^\star(\cdot)| \leq \sqrt{1/\ell_V'} \,, \qquad \sqrt{1/L_V} \leq |\partial_{x_i} T_i^\star(\cdot \mid z_1)| \leq \sqrt{1/\ell_V} \,,$$

and

$$|\partial_{x_1}^2 T_1^\star(\cdot)| \leq \frac{L_V'}{(\ell_V')^{3/2}} (1 + |\cdot|) \,, \qquad \text{and} \qquad |\partial_{x_i}^2 T_i^\star(\cdot \mid z_1)| \leq \frac{L_V}{\ell_V^{3/2}} (1 + |\cdot|) \,.$$

(ii) If, in addition, (**RD+**) *holds, then for all* $x_i \in \mathbb{R}$,

$$|\partial_{z_1} T_i^\star(x_i \mid \cdot)| \leq \frac{\overline{L}}{\ell_V} \,, \qquad \text{and} \qquad |\partial_{x_i} \partial_{z_1} T_i^\star(x_i \mid \cdot)| \lesssim \frac{\overline{L} L_V}{\ell_V^2} (1 + |x_i|) \,.$$

The first part of Theorem 3.5 can be viewed as an analogue of the celebrated Caffarelli contraction theorem (Caffarelli, 2000), specialized to star-separable transport maps. To establish higher-order regularity with respect to the leaf variable, we adopt a bootstrap strategy similar to that of Jiang et al. (2025).

The second and third parts of Theorem 3.5 concern the regularity of the optimal star-separable transport map with respect to the root variable. The first-order estimate in Theorem 3.5 (ii) builds upon the bound for the mixed derivative of $q_i^\star$ established in Lemma 3.4, and leverages recent advances in transport-information inequalities (Khudiakova et al., 2025). Higher-order estimates are then obtained via a bootstrap argument.

Since assumption (**RD+**) is stronger than (**RD**), the results in Theorem 3.5 (i) remain valid under (**RD+**).

3.4 A projected gradient descent algorithm for solving T-SSVI

In Section 3.1 and Section 3.2, we showed how to lift the infinite-dimensional space of star-structured probability measures to the space of star-separable transport maps, over which the objective becomes convex in the standard sense. However, the space $\mathcal{T}_{\text{star}}$ in (T-SSVI) is still prohibitively large to optimize over. We therefore approximate the set $\mathcal{T}_{\text{star}}$ by a finite-dimensional approximation, and show how to compute the minimizer over this set via first-order methods.

Existing gradient-based approaches to structured VI rely on parametrizing the variational distributions (Archer et al., 2015; Hoffman and Blei, 2015; Ranganath et al., 2016; Tan and Nott, 2018; Tan, 2021). To approximate $\pi^\star$, one posits a parametric family $\{\pi_\theta \mid \theta \in \Theta\} \subset \mathcal{C}_{\text{star}}$, often chosen as an exponential family, and optimizes the objective function $\theta \mapsto \mathsf{KL}(\pi_\theta \parallel \pi)$ using gradient descent. A key limitation is that the map $\theta \mapsto \mathsf{KL}(\pi_\theta \parallel \pi)$ is generally non-convex. As a result, despite empirical successes, such "black-box" VI methods lack theoretical guarantees and can be highly sensitive to initialization (Ranganath et al., 2014).

While existing algorithms face theoretical hurdles, we show that by carefully approximating the set of transport maps $\mathcal{T}_{\text{star}}$, one can retain convexity of the objective and derive convergence guarantees. To approximate the infinite-dimensional problem (T-SSVI) with a finite-dimensional one, we consider a parametrized set of star-separable transport maps $\mathcal{T}_\Theta \subset \mathcal{T}_{\text{star}}$ given by

$$\mathcal{T}_\Theta := \{T_\theta \mid \theta \in \Theta\},$$

where $\Theta \subseteq \mathbb{R}^p$ and $p \in \mathbb{N}$. The parametrized problem amounts to optimizing the following objective :

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \Theta} \mathsf{KL}\left((T_\theta)_{\#}\rho \parallel \pi\right). \tag{P-SSVI}$$

When Θ and $\mathcal{T}_\Theta$ are convex and the parametrization $\theta \mapsto T_\theta$ is affine, Theorem 3.2 implies that the parametrized problem (P-SSVI) is a strongly convex optimization problem over the finite-dimensional space Θ equipped with the norm $\|\theta\|_\Theta := \|T_\theta\|_{L^2(\rho)}$, and thus the optimizer $\hat{\theta}$ is uniquely defined.

The following theorem shows that one can find a parameter set Θ that is sufficiently large—yet still finite-dimensional—such that $T_{\hat{\theta}}$ approximates $T^\star$ up to the desired accuracy.

As mentioned in Section 3.2, we use the $L^2(\rho)$ norm as the evaluation metric, as it captures the graphical structure inherent to $\mathcal{C}_{\text{star}}$.

Theorem 3.6 (Existence of a scalable convex parametrization)**.** *Assume* (**R**), (**P1**), (**P2**) *and* (**RD+**) *hold. Assume further that*

$$|\partial_{z_1}^2 T_i^\star(x_i \mid \cdot)| \le L_{\mathsf{GR}} (1 + |x_i|)^\gamma , \quad \text{for some } L_{\mathsf{GR}}, \gamma > 0 . \tag{GR}$$

Then, for any $\epsilon > 0$, there exists a closed convex set Θ of dimension $O((d^2/\epsilon^2)\log(d/\epsilon^2))$ and an affine parametrization $\theta \mapsto T_\theta$, where the implied constant depends polynomially on $\ell_V, L_V, \ell_V', L_V', \overline{L}, L_{\mathsf{GR}}, M_{2\gamma}(\rho_1)$ and their inverses, where $M_{2\gamma}(\rho_1)$ is the 2γ-th absolute moment of ρ_1, such that the P-SSVI *minimizer $\hat{\theta}$ satisfies*

$$\|T_{\hat{\theta}} - T^\star\|_{L_2(\rho)} \le \epsilon .$$

Remark 3.7. In the theorem above, we assume that the problem parameters ℓ_V, L_V, etc. are dimension-free and we only report the final dimension dependence. This is purely done for the sake of exposition. Note that in order for, e.g., (**RD+**) to hold with a dimension-free constant, it requires $|\partial_{1j} V| = O(d^{-1/2})$, which can be restrictive. The complete dependence on the implied constants in Theorem 3.6, which is somewhat complicated, is explicitly derived in the proof.

The condition (**GR**) controls the growth rate of the second-order derivative of $T_i^\star$ with respect to the root variable. Notably, when $\pi_i^\star(\cdot \mid z_1) = \mathcal{N}(z_1, 1)$, then (**GR**) is satisfied with $\gamma = 0$ as $\partial_z^2 T_i^\star(x_i \mid z) = 0$. Whereas we have derived a priori bounds for the other derivatives of $T_i^\star$ in Theorem 3.5, the second derivative w.r.t. z is particularly difficult to control and hence we must adopt (**GR**) as an assumption; we leave it for future work to obtain an *a priori* bound for this quantity as well.

Theorem 3.6 is a direct consequence of Theorem C.15 in Section C.3, where we provide an explicit construction of the parameter set Θ based on a fixed dictionary of piecewise linear maps $\mathcal{M}$ and an explicit expression of the hidden constants. Our proposed dictionary $\mathcal{M}$ consists of piecewise linear maps of two variables, designed to approximate $T_i^\star$ for $i \ge 2$; see details in Section C.2. A similar dictionary construction for approximating univariate maps is introduced in Jiang et al. (2025).

The proof of Theorem C.15 builds on two non-trivial steps. First, we construct an "oracle" approximator $\widehat{T} \in \mathcal{T}_\Theta$ that is close to $T^\star$. To obtain such an approximation, we exploit the regularity results established in Theorem 3.5, particularly those concerning the dependence of $T^\star$ on the root variable. Second, we verify that the computed optimizer $T_{\hat{\theta}}$ is close to $\widehat{T}$. This is achieved by leveraging the geometry of the parametrized problem (P-SSVI) and the optimality of both $\pi^\star$ and $\pi_{\hat{\theta}}^\star$.

We now turn to the projected gradient descent algorithm for computing the minimizer $\hat{\theta}$. Starting from $\theta^{(0)}$, the iterates $\theta^{(t)}$, $t \in \mathbb{N}$, are defined recursively via

$$\theta^{(t+1)} = \mathrm{Proj}_{\Theta, \|\cdot\|_\Theta}\left(\theta^{(t)} - h \nabla_\theta \mathsf{KL}\left((T_{\theta^{(t)}})_\# \rho \parallel \pi\right)\right),$$

where $\mathrm{Proj}_{\Theta, \|\cdot\|_\Theta} : (\mathbb{R}^p, \|\cdot\|_2) \to (\Theta, \|\cdot\|_\Theta)$ is the projection operator and $h > 0$ is the constant step size for the detailed algorithm and the explicit formula for the gradient.

The following result provides the iterative complexity of the proposed scheme.

Theorem 3.8 (Iteration complexity of gradient descent)**.** *Let assumptions* (**R**), (**P1**), (**P2**), *and* (**RD+**) *be satisfied. Moreover, assume that* $\|\theta - \tilde{\theta}\|_\Theta \geq \lambda_\Theta \|\theta - \tilde{\theta}\|_2$ *for a conditioning number parameter* $\lambda_\Theta > 0$*. Let* $\{\theta^{(t)}\}_{t\in\mathbb{N}}$ *denote the iterates generated by the projected gradient descent algorithm. There exists* $\Upsilon > 0$ *depending linearly on* $L_V, L_V', 1/\lambda_\Theta$ *and quadratically on* d *such that for* $\kappa := \frac{L_V \vee (L_V'/2) + \Upsilon}{\ell_V \wedge \ell_V'} > 1$ *and step size* $h = \frac{1}{L_V \vee (L_V'/2) + \Upsilon}$*,*

(i) $\|\theta^{(t)} - \hat{\theta}\|_\Theta^2 \leq (1 - \kappa^{-1})^t \|\theta^{(0)} - \hat{\theta}\|_\Theta^2$.

(ii) $\mathsf{KL}(\pi_{\theta^{(t)}} \,\|\, \pi) - \mathsf{KL}(\pi_{\hat{\theta}} \,\|\, \pi) \leq (1 - \kappa^{-1})^t \frac{(L+\Upsilon)}{2} \|\theta^{(0)} - \hat{\theta}\|_\Theta^2$.

The conditioning number λ_Θ measures the distortion of $\|\cdot\|_\Theta$ compared to the standard Euclidean norm $\|\cdot\|_2$. In the detailed algorithm in Section C.4, we compute λ_Θ from the Gram matrix of the constructed dictionary for $\mathcal{T}_\Theta$.

The setting of Theorem 3.8 subsumes that of Theorem 3.6, and together these results provide a complete non-asymptotic, polynomial-time guarantee for solving the SSVI problem. Theorem 3.6 establishes an approximation bound—i.e., bias control—for the best approximation within a family whose size scales quadratically with d, while Theorem 3.8 shows that the projected gradient descent algorithm achieves an ϵ-accurate solution in $t \asymp \kappa \log(1/\epsilon)$ iterations.

The main technical challenge in proving Theorem 3.8 lies in establishing the Lipschitz smoothness of the objective function over the minimizing set. The smoothness is far from obvious due to the non-smoothness of the entropy H as a functional over the full Wasserstein space (noted in Diao et al., 2023; Chewi et al., 2025). Prior solutions have focused on establishing the smoothness of H over restricted subsets of the Wasserstein space (e.g., Lambert et al., 2022). Following Jiang et al. (2025), we explicitly characterize the additional smoothness parameter Υ of $\theta \mapsto \mathrm{KL}((T_\theta)_{\#}\rho \,\|\, \pi)$ over Θ in our analysis, where we use our explicit construction of the approximating family $\mathcal{T}_\Theta$; see details in Section C.4.

4 Going beyond the star graph

In this work, we inaugurated a rigorous study of star-structured variational inference. We established the existence and uniqueness of this minimization problem, and subsequent results pertaining to the regularity of said minimizers. By reformulating the problem in a more suitable geometry (one given by the theory of optimal transport), we provided a projected gradient descent algorithm which is able to approximate the minimizer to arbitrary precision. Here, we discuss possible extensions of star-structured variational inference when we decide to optimize over a fixed tree graph $\mathcal{G} = (\{1, \ldots, d\}, \mathcal{E}_\mathcal{G})$, and highlight the resulting technical challenges.

4.1 Difficulties in complete generality

Recall that a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is an element of $\mathcal{C}_\mathcal{G}$ if it can be represented in the following manner:

$$\mu(z_1, \ldots, z_d) = \prod_{(i,j)\in\mathcal{E}_\mathcal{G}} \varphi_{ij}(z_i, z_j)\,, \qquad \text{for} \qquad \varphi_{ij} : \mathbb{R}^2 \to [0, \infty)\,.$$

As before, we aim to extend our parameterization of such measures by representing them as pushforwards through suitable transport maps. Fixing $\rho = \mathcal{N}(0, I)$ (or more generally, any product measure with a density), we can associate a Knothe–Rosenblatt (KR) map to any distribution in $\mathcal{C}_\mathcal{G}$ via the following recursive representation for each $i \in \{1, \dots, d\}$:

$$z_i = T_i(x_i \mid z_{\mathrm{pa}(i)}),$$

where $\mathrm{pa}(i)$ denotes the parent node(s) of i. In analogy with Section 3.1, we denote by $\mathcal{T}_\mathcal{G}$ the set of KR maps induced by $\mathcal{C}_\mathcal{G}$.

However, for a general graph $\mathcal{G}$, the set $\mathcal{T}_\mathcal{G}$ may not be convex. Indeed, the argument in Proposition 3.1 relies on the injectivity of the map $\lambda T_1^1 + (1-\lambda)T_1^2$ for any two $T^1, T^2 \in \mathcal{T}_{\mathrm{star}}$ and $\lambda \in (0,1)$, which cannot be easily verified for general $\mathcal{G}$. For example, let $\mathcal{G}$ be a three-node Markov chain $z_1 \to z_2 \to z_3$, and let $\rho = \mathcal{N}(0, I_3)$. Define $T^1(x_1, x_2, x_3) \coloneqq \begin{pmatrix} x_1, & x_2, & x_3 \end{pmatrix}$ and $T^2(x_1, x_2, x_3) \coloneqq \begin{pmatrix} x_1, & x_1, & x_2 \end{pmatrix}$, then $T^1, T^2 \in \mathcal{T}_\mathcal{G}$, yet $\frac{1}{2}T^1 + \frac{1}{2}T^2 \notin \mathcal{T}_\mathcal{G}$. Moreover, obtaining regularity results for the minimizers is also challenging due to the complex structure of the problem. As a result, a principled approach to computation based on the polyhedral approximation is more costly, if not intractable, since the KR map T_i (as a function of x) depends on all nodes on the path to the tree root prior to node i.

4.2 A suitably nice subset of transport maps

A possible remedy is to restrict the minimization to a "nice" subset of $\mathcal{T}_\mathcal{G}$, making the problem more tractable. Define $\widetilde{\mathcal{T}}_\mathcal{G}$ as the set of transport maps with the following structure:

$$\widetilde{\mathcal{T}}_\mathcal{G} \coloneqq \left\{ T : x \mapsto (T_i(x_i; x_{\mathrm{pa}(i)}))_{i \in [d]} \;\middle|\; T_i(\cdot\,; x_{\mathrm{pa}(i)}) \text{ is strictly increasing, } \forall i \in [d] \right\}.$$

Recall that for any $T \in \mathcal{T}_\mathcal{G}$, the component maps T_i take the form $T_i(\cdot \mid z_{\mathrm{pa}(i)})$, where $z_{\mathrm{pa}(i)}$ depends on all the nodes lying on the path from node i to the root. Thus, we have $\widetilde{\mathcal{T}}_\mathcal{G} \subseteq \mathcal{T}_\mathcal{G}$, with equality achieved if and only if $\mathcal{G}$ is a star graph. We introduce the surrogate structured variational inference (Surro-SSVI) problem as follows:

$$T^\star \coloneqq \operatorname*{arg\,min}_{T \in \widetilde{\mathcal{T}}_\mathcal{G}} \mathsf{KL}\left(T_\# \rho \parallel \pi\right). \tag{Surro-SSVI}$$

Observe that $\widetilde{\mathcal{T}}_\mathcal{G}$ is convex. Thus, by the same argument as in Theorem 3.2, we can show that (Surro-SSVI) admits a unique solution under (**SLC**).

However, the minimizing set is less expressive than that of the full SSVI problem (T-SSVI), as it imposes additional graphical constraints on the distributions. For example, consider the case where the graph is a Markov chain given by $z_1 \to z_2 \to \cdots \to z_d$, and let $(Z_1, \dots, Z_d) \sim T_\# \rho$ for some $T \in \widetilde{\mathcal{T}}_\mathcal{G}$. Then, Z_1 is independent of Z_i for any $i \geq 3$. In particular, if we restrict $T \in \widetilde{\mathcal{T}}_\mathcal{G}$ to be a linear map, then $T_\# \rho$ is again a Gaussian distribution. Let $\Sigma \in \mathbb{R}^{d \times d}$ denote its covariance matrix, then $\Sigma_{ij} = 0$ whenever $|j - i| \geq 2$, implying that Σ is necessarily sparse. In contrast, the general structured variational inference framework typically assumes that the *precision* matrix Σ^{-1} is sparse, rather than Σ itself.

The extension of our results under (Surro-SSVI), and more generally to arbitrary tree graphs, is beyond the scope of this paper.

Acknowledgments

AAP thanks the Foundations of Data Science at Yale University for financial support. The authors thank Marcel Nutz for helpful discussions. In particular, Proposition 2.1 resulted from discussions with Marcel.

References

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *Ann. Statist.*, 48(3):1475–1497.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.*, 17:Paper No. 239, 41.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition.

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. (2015). Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*.

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

Bachoc, F. c., Gamboa, F., Loubes, J.-M., and Venet, N. (2018). A Gaussian process regression model for distribution inputs. *IEEE Trans. Inform. Theory*, 64(10):6620–6637.

Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., and Eder, M. (2020). All adapted topologies are equal. *Probab. Theory Related Fields*, 178(3-4):1125–1172.

Bakry, D., Gentil, I., and Ledoux, M. (2014). *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham.

Barber, D. and Wiegerinck, W. (1999). Tractable variational structures for approximating graphical models. In *Advances in Neural Information Processing Systems*, pages 183–189.

Barbie, M. and Gupta, A. (2014). The topology of information on the space of probability measures over Polish spaces. *J. Math. Econom.*, 52:98–111.

Beck, A. (2017). *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA.

Beiglböck, M., Pammer, G., and Posch, A. (2023). The Knothe–Rosenblatt distance and its induced topology. *arXiv preprint arXiv:2312.16515*.

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York. Corrected reprint of the second (1985) edition.

Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic optimal control: The discrete time case*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.

Bhatia, K., Kuang, N. L., Ma, Y.-A., and Wang, Y. (2022). Statistical and computational trade-offs in variational inference: a case study in inferential model selection. *arXiv preprint arXiv:2207.11208*.

Bhattacharya, A., Pati, D., and Yang, Y. (2025). On the convergence of coordinate ascent variational inference. *Ann. Statist.*, 53(3):929–962.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.*, 105(489):324–335.

Brown, L. D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA.

Caffarelli, L. A. (1992). Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math.*, 45(9):1141–1151.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2)*, 144(3):453–496.

Caffarelli, L. A. (2000). Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, 214(3):547–563.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, 7(1):73–107.

Carlen, E. A., Cordero-Erausquin, D., and Lieb, E. H. (2013). Asymmetric covariance estimates of Brascamp-Lieb type and related inequalities for log-concave measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(1):1–12.

Castillo, I., L'Huillier, A., Ray, K., and Travis, L. (2024). A variational bayes approach to debiased inference for low-dimensional parameters in high-dimensional linear regression. *arXiv preprint arXiv:2406.12659*.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45(1):223–256.

Chewi, S. (2025). 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, École d'Été de Probabilités de Saint-Flour.

Dalalyan, A. S., Karagulyan, A., and Riou-Durand, L. (2022). Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *J. Mach. Learn. Res.*, 23:Paper No. [235], 38.

de Finetti, B. (1929). Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190.

de Finetti, B. (2017). *Theory of probability*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, combined edition. A critical introductory treatment, Translated from the 1970 Italian original by Antonio Machí and Adrian Smith and with a preface by Smith.

Deb, N. and Sen, B. (2023). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Amer. Statist. Assoc.*, 118(541):192–207.

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.

Domke, J., Gower, R., and Garrigos, G. (2023). Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 66289–66327.

Du, Q., Wang, K., Zhang, E., and Zhong, C. (2024). A particle algorithm for mean-field variational inference. *arXiv preprint arXiv:2412.20385*.

Fasano, A., Durante, D., and Zanella, G. (2022). Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika*, 109(4):901–919.

Frazier, D. T., Loaiza-Maya, R., and Martin, G. M. (2023). Variational Bayes in state space models: inferential and predictive accuracy. *J. Comput. Graph. Statist.*, 32(3):793–804.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *J. Amer. Statist. Assoc.*, 88:881–889.

Ghorbani, B., Javadi, H., and Montanari, A. (2019). An instability in variational inference for topic models. In *International Conference on Machine Learning*, pages 2221–2231.

Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *Ann. Statist.*, 50(2):1012–1037.

Givens, C. R. and Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231 – 240.

Goplerud, M., Papaspiliopoulos, O., and Zanella, G. (2025). Partially factorized variational inference for high-dimensional mixed models. *Biometrika*, 112(2).

Hall, P., Pham, T., Wand, M. P., and Wang, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.*, 39(5):2502–2532.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: a measure transportation approach. *Ann. Statist.*, 49(2):1139–1165.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Texts in Statistics. Springer, New York.

Hoffman, M. D. and Blei, D. M. (2015). Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, pages 361–369.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Jiang, Y., Chewi, S., and Pooladian, A.-A. (2025). Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Found. Comput. Math.*

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Katsevich, A. and Rigollet, P. (2024). On the approximation accuracy of Gaussian variational inference. *Ann. Statist.*, 52(4):1384–1409.

Khudiakova, K. A., Maas, J., and Pedrotti, F. (2025). L^∞-optimal transport of anisotropic log-concave measures and exponential convergence in Fisher's infinitesimal model. *Ann. Appl. Probab.*, 35(3):1913–1940.

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

Kim, Y., Wang, W., Carbonetto, P., and Stephens, M. (2024). A flexible empirical Bayes approach to multiple linear regression, and connections with penalized regression. *J. Mach. Learn. Res.*, 25:Paper No. [185], 59.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*, volume 2.

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Math. J.*, 4:39–52.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: an introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

Lacker, D., Mukherjee, S., and Yeung, L. C. (2024). Mean field approximations via log-concavity. *Int. Math. Res. Not. IMRN*, (7):6008–6042.

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

Lauritzen, S. (2024). Total variation convergence preserves conditional independence. *Statist. Probab. Lett.*, 214:Paper No. 110200, 2.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

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

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2022). Fast and accurate variational inference for models with many latent variables. *J. Econometrics*, 230(2):339–362.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, second edition.

Mukherjee, S., Qiu, J., and Sen, S. (2024). On naive mean-field approximation for high-dimensional canonical GLMs. *arXiv preprint arXiv:2406.15247*.

Mukherjee, S., Sen, B., and Sen, S. (2023). A mean field approach to empirical Bayes estimation in high-dimensional linear regression. *arXiv preprint arXiv:2309.16843*.

Mukherjee, S. and Sen, S. (2022). Variational inference in high-dimensional linear regression. *J. Mach. Learn. Res.*, 23:Paper No. [304], 56.

Murphy, K. P. (2023). *Probabilistic machine learning: advanced topics*. MIT press.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto*.

Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham.

Parisi, G. (1980). Mean field theory for spin glasses. volume 67, pages 25–28. Common trends in particle and condensed matter physics (Proc. Winter Adv. Study Inst., Les Houches, 1980).

Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Statist.*, 44(1):183–218.

Qiu, J. (2024). Sub-optimality of the naive mean field approximation for proportional high-dimensional linear regression. In *Advances in Neural Information Processing Systems*, volume 36.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333. PMLR.

Ray, K. and Szabó, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *J. Amer. Statist. Assoc.*, 117(539):1270–1281.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York.

Rockafellar, R. T. (1970). *Convex analysis*, volume No. 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statistics*, 23:470–472.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, 6(1):34–58.

Saarela, O., Stephens, D. A., and Moodie, E. E. M. (2023). The role of exchangeability in causal inference. *Statist. Sci.*, 38(3):369–385.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Anal.*, 8(4):837–881.

Saul, L. and Jordan, M. (1995). Exploiting tractable substructures in intractable networks. volume 8.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Stat. Surv.*, 8:45–114.

Sheng, S., Wu, B., and González-Sanz, A. (2025a). Mode collapse of mean-field variational inference. *arXiv preprint arXiv:2510.17063*.

Sheng, S., Wu, B., González-Sanz, A., and Nutz, M. (2025b). Stability of mean-field variational inference. *arXiv preprint 2506.07856*.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statistics*, 21:124–127.

Tan, L. S. L. (2021). Use of model reparametrization to improve variational Bayes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(1):30–57.

Tan, L. S. L. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Stat. Comput.*, 28(2):259–275.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

Villani, C. (2009). *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Old and new.

Wainwright, M. J. (2019). *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. A non-asymptotic viewpoint.

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

Wang, B. and Titterington, D. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *J. Mach. Learn. Res.*, 14:1005–1031.

Wang, C. and Blei, D. M. (2018). A general method for robust Bayesian modeling. *Bayesian Anal.*, 13(4):1159–1187.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(5):1273–1300.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vis.*, 101(2):254–269.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.*, 114(527):1147–1161.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

Wibisono, A. (2018). Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Wu, B. and Blei, D. (2024). Extending mean-field variational inference via entropic regularization: theory and computation. *arXiv preprint arXiv:2404.09113*.

Xiao, J. and Su, Q. (2024). TreeVI: reparameterizable tree-structured variational inference for instance-level correlation capturing. In *Advances in Neural Information Processing Systems*, volume 37, pages 14509–14539.

Yang, D. (2019). Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression. *Electron. J. Stat.*, 13(2):3082–3094.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *Ann. Statist.*, 48(2):886–905.

Zemel, Y. and Panaretos, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976.

Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.

Zhang, F. and Gao, C. (2020a). Convergence rates of variational posterior distributions. *Ann. Statist.*, 48(4):2180–2207.

Zhang, F. and Gao, C. (2020b). Convergence rates of variational posterior distributions. *Ann. Statist.*, 48(4):2180–2207.

Zhu, C. and Müller, H.-G. (2023). Autoregressive optimal transport models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 85(3):1012–1033.

A Proofs from Section 2

We recall that a function $f : \mathbb{R} \to [0, \infty]$ is called *lower semianalytic* if the sets $\{f < c\}$ are analytic for all $c \in \mathbb{R}$, where a subset of $\mathbb{R}$ is called analytic if it is the image of a Borel subset of a Polish space under a Borel mapping. Any Borel function is lower semianalytic and a lower semianalytic function is universally measurable. We refer readers to Bertsekas and Shreve (1978, §7) for the terminology used in the following proof, including analytic sets, lower semianalytic functions, and universally measurable functions.

Proof of Proposition 2.1. For any $\mu \in \mathcal{C}_{\text{star}}$, $\mu(\mathrm{d}z) = \mu_1(\mathrm{d}z_1)\, \mu_{-1}(\mathrm{d}z_{-1} \mid z_1)$ with $\mu_{-1}(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$. The lower bound

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) \geq \inf_{\mu_1 \in \mathcal{P}(\mathbb{R})} \left\{ \mathsf{KL}\left(\mu_1 \parallel \pi_1\right) + \int \inf_{\nu \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right) \mu_1(\mathrm{d}z_1) \right\} =: V_{\text{iter}}$$

follows trivially from the KL chain rule.

To see the converse inequality, for every $\epsilon > 0$, we want to show that

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) \leq V_{\text{iter}} + \epsilon\,.$$

Consider

$$h_{\text{val}}(z_1) \coloneqq \inf_{\nu \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right)\,. \tag{32}$$

Then the minimizer to the outer problem is given by

$$p^\star(\mathrm{d}z_1) \propto e^{-h_{\text{val}}(z_1)}\, \pi_1(\mathrm{d}z_1)\,.$$

Since ν varies over a Borel subset of a Polish space, $z_1 \mapsto \pi_{-1}(\cdot \mid z_1)$ is Borel with values in a Polish space, and $\mathsf{KL}\left(\cdot \parallel \cdot\right)$ is jointly Borel and non-negative. This implies that h_{val} is lower semianalytic (Bertsekas and Shreve, 1978, Proposition 7.47). Then, the selection theorem for lower semianalytic functions (see, e.g., Bertsekas and Shreve, 1978, Proposition 7.50(a)) yields a universally measurable function $z_1 \mapsto \nu^\epsilon(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$ such that

$$\mathsf{KL}\left(\nu^\epsilon(\cdot \mid z_1) \parallel \pi_{-1}(\cdot \mid z_1)\right) \leq \inf_{\nu \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right) + \epsilon = h_{\text{val}}(z_1) + \epsilon\,.$$

Let $\pi^\epsilon(\mathrm{d}z) \coloneqq p_1^\star(\mathrm{d}z_1)\, \nu^\epsilon(\mathrm{d}z_{-1} \mid z_1)$. Then,

$$\inf_{\mu \in \mathcal{C}_{\text{star}}} \mathsf{KL}\left(\mu \parallel \pi\right) \leq \mathsf{KL}\left(\pi^\epsilon \parallel \pi\right) \leq V_{\text{iter}} + \epsilon\,.$$

As ϵ is chosen arbitrarily, we conclude the statement as desired. □

Proof of Theorem 2.2. Under (**SLC**), $h_{\text{val}}(z_1)$ attains a unique solution, denoted by $q^\star(\cdot \mid z_1)$; and $z_1 \mapsto q^\star(\cdot \mid z_1)$ is universally measurable (see, e.g., Bertsekas and Shreve, 1978, Proposition 7.50(b)). Therefore, after modifying on a null set with respect to π_1, $z_1 \mapsto q^\star(\cdot \mid z_1)$ is a Borel-measurable stochastic kernel (e.g., Bertsekas and Shreve, 1978, Lemma 7.28) so that

$$\pi^\star(\mathrm{d}z) \propto e^{-h_{\text{val}}(z_1)}\, \pi_1(\mathrm{d}z_1)\; q^\star(\mathrm{d}z_{-1} \mid z_1) \in \mathcal{C}_{\text{star}}\,.$$

This concludes the existence and uniqueness of the solution to (SSVI). □

Proof of Theorem 2.4. The SSVI problem is equivalent to finding the optimal star-separable map $T^\star$ in (T-SSVI). Recall that the Wasserstein gradient of $\mu \mapsto \mathsf{KL}(\mu \,\|\, \pi)$ at μ is $\nabla \log \frac{\mu}{\pi}$ (see e.g., Ambrosio et al., 2008; Chewi et al., 2025). This implies the first-order optimality condition

$$\mathbb{E}_{\pi^\star}\left\langle \nabla \log \frac{\pi^\star(Z)}{\pi(Z)},\, T(Z) - Z \right\rangle = 0\,, \tag{33}$$

for any map T such that $T_\# \pi^\star \in \mathcal{C}_{\mathrm{star}}$, i.e., for $T \in \mathcal{T}_{\mathrm{star}}$.

In particular, (33) holds for star-separable maps of the form

$$\mathbb{T}^i(z) = z + [0, \dots, 0, T_i(z_i; z_1) - z_i, 0, \dots, 0]$$

with the convention that $\mathbb{T}^1(z) = [T_1(z_1), 0, \dots, 0]$. Applying (33) to each $\mathbb{T}^i$,

$$\mathbb{E}_{\pi^\star}\left[\partial_1 \log \frac{\pi^\star(Z)}{\pi(Z)}\,(T_1(Z_1) - Z_1)\right] = 0\,, \tag{34}$$

and for $i = 2, \dots, d$

$$\mathbb{E}_{\pi^\star}\left[\partial_i \log \frac{\pi^\star(Z)}{\pi(Z)}\,(T_i(Z_i; Z_1) - Z_i)\right] = 0\,.$$

After decomposing $\pi^\star$, we have that for $i \neq 1$,

$$\mathbb{E}_{\pi^\star}\left[\left(\partial_i \log q_i^\star(Z_i \mid Z_1) + \partial_i V(Z)\right)(T_i(Z_i; Z_1) - Z_i)\right] = 0\,.$$

Since the function T_i is an arbitrary monotone function in the first argument and an arbitrary function in the second argument, the closed linear span of the family of functions $(x, y) \mapsto T_i(x; y) - x$ contains $C_{\mathrm{b}}(\mathbb{R}^2)$, the set of all bounded continuous functions. Therefore, we obtain a simple identity

$$\partial_i \log q_i^\star(z_i \mid z_1) + \mathbb{E}_{\pi^\star}\left[\partial_i V(Z) \mid Z_1 = z_1, Z_i = z_i\right] = 0\,,$$

for $\pi^\star$-a.e. (z_1, z_i). Integrating the above display on both sides with respect to z_i and noting that the order of differentiation and conditional expectation can be exchanged as V is convex, we have

$$q_i^\star(z_i \mid z_1) \propto \exp\left(-\mathbb{E}_{\pi^\star}\left[V(Z) \mid Z_1 = z_1, Z_i = z_i\right]\right).$$

By the same argument as above, (34) can be written as

$$\mathbb{E}_{\pi^\star}\left[\left(\partial_1 \log p^\star(Z_1) + \sum_{j \geq 2} \partial_1 \log q_j^\star(Z_j \mid Z_1) + \partial_1 V(Z)\right)(T_1(Z_1) - Z_1)\right] = 0\,.$$

Using the tower property, the above display becomes

$$\mathbb{E}_{\pi^\star}\left[\left(\partial_1 \log p^\star(Z_1) + \mathbb{E}_{\pi^\star}\left[\sum_{j \geq 2} \partial_1 \log q_j^\star(Z_j \mid Z_1) + \partial_1 V(Z) \,\middle|\, Z_1\right]\right)(T_1(Z_1) - Z_1)\right] = 0\,.$$

Hence,

$$\partial_1 \log p^\star(z_1) + \mathbb{E}_{\pi^\star}\left[\sum_{j \geq 2} \partial_1 \log q_j^\star(Z_j \mid Z_1) + \partial_1 V(Z) \,\middle|\, Z_1 = z_1\right] = 0\,, \tag{35}$$

for $p^\star$-a.e. z_1. To further simplify, we note that

$$\begin{aligned}\mathbb{E}_{\pi^\star}\left[\partial_1 \sum_{j\geq 2} \log q_j^\star(Z_j \mid Z_1) \,\middle|\, Z_1 = z_1\right] &= \sum_{j\geq 2} \mathbb{E}_{\pi^\star}\left[\partial_1 \log q_j^\star(Z_j \mid Z_1) \,\middle|\, Z_1 = z_1\right] \\ &= \sum_{j\geq 2} \int \frac{\partial_1 q_j^\star(z_j \mid z_1)}{q_j^\star(z_j \mid z_1)}\, q_j^\star(\mathrm{d}z_j \mid z_1) \\ &= \sum_{j\geq 2} \partial_1 \int q_j^\star(\mathrm{d}z_j \mid z_1) = 0\,.\end{aligned} \tag{36}$$

Therefore,

$$\partial_1 \log p^\star(z_1) + \mathbb{E}_{\pi^\star}\left[\partial_1 V(Z) \mid Z_1 = z_1\right] = 0\,,$$

for $p^\star$-a.e. z_1. Integrating both sides with respect to z_1 yields

$$p^\star(z_1) \propto \exp\left(-\int_0^{z_1} \mathbb{E}_{\pi^\star}\left[\partial_1 V(Z) \mid Z_1 = s\right] \mathrm{d}s\right), \tag{37}$$

as desired. □

Proof of Lemma 2.6. Since $L_V \vee (L_V'/2) I_{d-1} \succeq (\nabla^2 V)_{-1} \succeq \ell_V' \wedge \ell_V I_{d-1}$ and its Schur complement

$$\begin{aligned}L_V \vee (L_V'/2) &\geq \partial_{11} V - (\nabla^2_{-1,1} V)^\top \left((\nabla^2 V)_{-1}\right)^{-1} (\nabla^2_{-1,1} V) \\ &\geq \partial_{11} V - \frac{1}{\ell_V} \sum_{j=2}^d (\partial_{1j} V)^2 \geq \ell_V' \wedge \ell_V\,,\end{aligned}$$

where the last inequality follows from (**RD**). Applying the Schur's complement theorem (Zhang, 2006, Theorem 1.6) yields the desired result. □

We recall the following result by Sheng et al. (2025b).

Lemma A.1. *Let* (**R**), (**P1**), (**P2**), *and* (**RD**) *hold. For any* $z_1, z_1' \in \mathbb{R}$,

$$W_2\left(q^\star(\cdot \mid z_1),\, q^\star(\cdot \mid z_1')\right) \leq \frac{\left\|\sqrt{\sum_{i\geq 2} |\partial_{1i} V|^2}\right\|_{L^\infty}}{\ell_V} |z_1 - z_1'|\,.$$

Proof. The assumptions in Sheng et al. (2025b, Theorem 2.1) are satisfied by Lemma 2.6, and therefore

$$\begin{aligned}W_2\left(q^\star(\cdot \mid z_1),\, q^\star(\cdot \mid z_1')\right) &\leq \frac{1}{\ell_V} \left\|\nabla_{-1} V(z_1, \cdot) - \nabla_{-1} V(z_1', \cdot)\right\|_{L^2(q^\star(\cdot|z_1'))} \\ &\leq \frac{\left\|\sqrt{\sum_{i\geq 2} |\partial_{1i} V|^2}\right\|_{L^\infty}}{\ell_V} |z_1 - z_1'|\,.\end{aligned}$$

□

Proof of Theorem 2.8. Differentiating the self-consistency equation in Theorem 2.4 for $q_i^\star(z_i \mid z_1)$, we have

$$\begin{aligned}-\partial_i^2 \log q_i^\star\left(z_i \mid z_1\right) &= \partial_i \int \partial_i V(z_1, z_i, z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z_1) \\ &= \mathbb{E}_{\pi^\star}\left[\partial_{ii} V(Z) \mid Z_1 = z_1,\, Z_i = z_i\right].\end{aligned}$$

In the last line, we can interchange differentiation and integration because the map $z_i \mapsto \partial_i V(z_1, z_i, z_{-\{1,i\}})$ is Lipschitz, as $\ell_V \le \partial_{ii} V \le L_V$. Thus, $\ell_V \le -\partial_i^2 \log q_i^\star(z_i \mid z_1) \le L_V$.

On the other hand, by the self-consistency equation for $p^\star$ in Theorem 2.4, we obtain

$$-\partial_1 \log p^\star(z_1) = \mathbb{E}_{\pi^\star}\left[\partial_1 V(Z) \mid Z_1 = z_1\right],$$

and thus,

$$\begin{aligned}-\partial_1^2 \log p^\star(z_1) &= \partial_1 \int \partial_1 V(z_1, z_{-1})\, q^\star(\mathrm{d}z_{-1} \mid z_1) \\ &= \lim_{z_1' \to z_1} \frac{\int \partial_1 V(z_1', z_{-1})\, q^\star(\mathrm{d}z_{-1} \mid z_1') - \int \partial_1 V(z_1, z_{-1})\, q^\star(\mathrm{d}z_{-1} \mid z_1)}{|z_1 - z_1'|} \\ &= \lim_{z_1' \to z_1} \frac{\int \left(\partial_1 V(z_1', z_{-1}) - \partial_1 V(z_1, z_{-1})\right) q^\star(\mathrm{d}z_{-1} \mid z_1)}{|z_1 - z_1'|} \\ &\quad + \underbrace{\lim_{z_1' \to z_1} \frac{\int \partial_1 V(z_1', z_{-1})\left(q^\star(z_{-1} \mid z_1') - q^\star(z_{-1} \mid z_1)\right) \mathrm{d}z_{-1}}{|z_1 - z_1'|}}_{=:A} \\ &= \mathbb{E}_{\pi^\star}\left[\partial_{11} V(Z) \mid Z_1 = z_1\right] + A\,.\end{aligned}$$

As $z_{-1} \mapsto \partial_1 V(z_1', z_{-1})$ is $\|\sqrt{\sum_{i\ge 2} |\partial_{1i} V|^2}\|_{L^\infty}$-Lipschitz, the dual formulation for the 1-Wasserstein distance (see e.g., Villani, 2003) yields that for each $z_1' \in \mathbb{R}$,

$$\begin{aligned}&\left|\int \partial_1 V(z_1', z_{-1})\left(q^\star\left(z_{-1} \mid z_1'\right) - q^\star\left(z_{-1} \mid z_1\right)\right) \mathrm{d}z_{-1}\right| \\ &\qquad \le \left\|\Big(\sum_{i\ge 2} |\partial_{1i} V|^2\Big)^{1/2}\right\|_{L^\infty} W_1\left(q^\star(\cdot \mid z_1),\, q^\star(\cdot \mid z_1')\right).\end{aligned} \tag{38}$$

As $W_1 \le W_2$, we can apply Lemma A.1 and obtain

$$\begin{aligned}W_1\left(q^\star(\cdot \mid z_1),\, q^\star(\cdot \mid z_1')\right) &\le W_2\left(q^\star(\cdot \mid z_1),\, q^\star(\cdot \mid z_1')\right) \\ &\le \frac{1}{\ell_V} \left\|\nabla_{-1} V(z_1, \cdot) - \nabla_{-1} V(z_1', \cdot)\right\|_{L^2(q^\star(\cdot \mid z_1'))} \\ &\le \frac{\left\|\sqrt{\sum_{i\ge 2} |\partial_{1i} V|^2}\right\|_{L^\infty}}{\ell_V} |z_1 - z_1'|\,.\end{aligned}$$

Combining with (38) gives that

$$|A| \le \frac{\left\|\sum_{i\ge 2} |\partial_{1i} V|^2\right\|_{L^\infty}}{\ell_V}\,.$$

Therefore,

$$-\partial_1^2 \log p^\star(z_1) = \mathbb{E}_{\pi^\star}\left[\partial_{11} V(Z) \mid Z_1 = z_1\right] + A \geq \ell_V'$$

as we assumed that

$$\partial_{11} V - \frac{\left\| \sum_{i \geq 2} |\partial_{1i} V|^2 \right\|_{L^\infty}}{\ell_V} \geq \ell_V' \,.$$

Similarly,

$$-\partial_1^2 \log p^\star(z_1) \leq \mathbb{E}_{\pi^\star}\left[\partial_{11} V(Z) \mid Z_1 = z_1\right] + \frac{\left\| \sum_{i \geq 2} |\partial_{1i} V|^2 \right\|_{L^\infty}}{\ell_V} \leq L_V' \,.$$

□

Proof of Theorem 2.9. The proof builds on a collection of functional inequalities, such as the log-Sobolev inequality (Villani, 2003, Theorem 9.9), the Talagrand inequality (Otto and Villani, 2000, Theorem 1), and the Poincaré inequality (Bakry et al., 2014, Corollary 4.8.2).

First, by the optimal representation of $\pi^\star$ and (two applications of) the log-Sobolev inequality, we have

$$\begin{aligned} \mathsf{KL}\left(\pi^\star \parallel \pi\right) &\leq \mathsf{KL}\left(p^\star \parallel \pi_1\right) + \int \mathsf{KL}\left(q^\star(\cdot \mid z_1) \parallel \pi_{-1}(\cdot \mid z_1)\right) p^\star(\mathrm{d}z_1) \\ &\leq \frac{1}{2\ell_V'} \int_{\mathbb{R}} \left|\partial_1 \log p^\star(z_1) - \partial_1 \log \pi_1(z_1)\right|^2 p^\star(\mathrm{d}z_1) \\ &\quad + \frac{1}{2\ell_V} \sum_{i \geq 2} \int \left|-\partial_i \log q_i^\star(z_i \mid z_1) - \partial_i V(z)\right|^2 \pi^\star(\mathrm{d}z) \,. \end{aligned} \tag{39}$$

For each $i \geq 2$, the self-consistency equation (14) implies that

$$\mathbb{E}_{\pi^\star}[\partial_i V(Z) \mid Z_1, Z_i] = -\partial_i \log q_i^\star(Z_i \mid Z_1) \,.$$

Then as $q_i^\star(\mathrm{d}z_i \mid z_1)$ is ℓ_V-log-concave by Theorem 2.8 and $q^\star(\mathrm{d}z_{-1} \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$, the Poincaré inequality yields that

$$\begin{aligned} &\int \left|-\partial_i \log q_i^\star(z_i \mid z_1) - \partial_i V(z)\right|^2 \pi^\star(\mathrm{d}z) \\ &\quad = \int \mathrm{Var}_{\pi^\star}(\partial_i V(Z) \mid Z_1 = z_1, Z_i = z_i)\, \pi_{1i}^\star(\mathrm{d}z_1, \mathrm{d}z_i) \leq \frac{1}{\ell_V} \sum_{j \geq 2,\, j \neq i} \mathbb{E}_{\pi^\star}[|\partial_{ij} V(Z)|^2] \,. \end{aligned} \tag{40}$$

On the other hand, we note that

$$\partial_1 \log \pi_1(z_1) = \partial_1 \log \int \exp\left(-V(z)\right) \mathrm{d}z_{-1} = -\int \partial_1 V(z)\, \pi_{-1}(\mathrm{d}z_{-1} \mid z_1) \,.$$

Then, combining with the self-consistency equation (14), we obtain

$$\begin{aligned} &\int \left|\partial_1 \log p^\star(z_1) - \partial_1 \log \pi_1(z_1)\right|^2 p^\star(\mathrm{d}z_1) \\ &\quad = \int \left| \int \partial_1 V(z) \left(q^\star(\mathrm{d}z_{-1} \mid z_1) - \pi_{-1}(\mathrm{d}z_{-1} \mid z_1)\right) \right|^2 p^\star(\mathrm{d}z_1) \\ &\quad \leq \left\| \sum_{i \geq 2} |\partial_{1i} V|^2 \right\|_{L^\infty} \int W_1^2\left(q^\star(\mathrm{d}z_{-1} \mid z_1),\, \pi_{-1}(\mathrm{d}z_{-1} \mid z_1)\right) p^\star(\mathrm{d}z_1) \,. \end{aligned}$$

The last line uses the fact that $\partial_1 V(z_1, \cdot)$ is $\| \sum_{i\geq 2} |\partial_{1i} V|^2 \|_{L^\infty}^{1/2}$–Lipschitz.

By (**R**) and (**RD**), we have

$$L_V'/2 - \frac{1}{\ell_V} \sum_{j=2}^{d} (\partial_{1j} V)^2 \geq \partial_{11} V - \frac{1}{\ell_V} \sum_{j=2}^{d} (\partial_{1j} V)^2 \geq \ell_V' \,.$$

Hence, $\sum_{i\geq 2} (\partial_{1i} V)^2 \leq \frac{1}{2} \ell_V L_V' - \ell_V \ell_V' < \infty$.

Thus, by the T_1-transport inequality and then the log-Sobolev inequality,

$$\begin{aligned}
&\int |\partial_1 \log p^\star(z_1) - \partial_1 \log \pi_1(z_1)|^2 \; p^\star(\mathrm{d}z_1) \\
&\quad \leq \frac{\frac{1}{2}\ell_V L_V' - \ell_V \ell_V'}{\ell_V^2} \sum_{i\geq 2} \int |-\partial_i \log q_i^\star(z_i \mid z_1) - \partial_i V(z)|^2 \; \pi^\star(\mathrm{d}z) \\
&\quad \leq \frac{\frac{1}{2} L_V' - \ell_V'}{\ell_V^2} \sum_{i\geq 2} \sum_{j\geq 2,\, j\neq i} \mathbb{E}_{\pi^\star}[|\partial_{ij} V(Z)|^2] \,,
\end{aligned}$$

where the last line follows from (40).

Combining all the above bounds, we conclude that

$$\begin{aligned}
\mathsf{KL}\left(\pi^\star \parallel \pi\right) &\leq \Big(\frac{1}{2\ell_V^2} + \frac{L_V'}{4\ell_V' \ell_V^2} - \frac{1}{2\ell_V^2}\Big) \sum_{i\geq 2} \sum_{j\geq 2,\, j\neq i} \mathbb{E}_{\pi^\star}\Big[(\partial_{ij} V(Z))^2\Big] \\
&= \frac{L_V'}{2\ell_V' \ell_V^2} \sum_{i\geq 2} \sum_{j>i} \mathbb{E}_{\pi^\star}\Big[(\partial_{ij} V(Z))^2\Big] \,.
\end{aligned}$$

□

A.1 Proofs for Examples

We first recall the following lemma, which provides the explicit expression for the KL divergence between two Gaussian distributions.

Lemma A.2 (KL divergence between Gaussians)**.** *Consider $p_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, $p_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ with $\mu_0, \mu_1 \in \mathbb{R}^d$, $\Sigma_0, \Sigma_1 \in \mathbb{R}^{d\times d}$. Then,*

$$\mathsf{KL}\left(p_0 \parallel p_1\right) = \frac{1}{2} \log \frac{\det \Sigma_1}{\det \Sigma_0} + \frac{1}{2}(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) - d) + \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma_1^{-1}(\mu_0 - \mu_1) \,.$$

Proof of Theorem 2.10. By Proposition 2.1 and by Theorem 3.2, we know that (3) admits a unique optimizer given by

$$\pi^\star(z) \propto e^{-h(z_1)} \, \pi_1(z_1) \; q^\star(z_{-1} \mid z_1) \,,$$

where $q^\star(\cdot \mid z_1)$ is the minimizer for

$$h_{\mathrm{val}}(z_1) = \inf_{\nu \in \mathcal{P}_{ac}(\mathbb{R})^{\otimes(d-1)}} \mathsf{KL}\left(\nu \parallel \pi_{-1}(\cdot \mid z_1)\right) \,.$$

Since $\pi = \mathcal{N}(m, \Sigma)$, denote by $m_{-1} = (m_2, \ldots, m_d)^\top$, $\sigma_{ij} = \Sigma_{ij}$, $\Sigma_{-1,-1} = (\sigma_{ij})_{2 \le i,j \le d}$, and $\Sigma_{-1,1} = (\sigma_{12}, \ldots, \sigma_{1d})^\top$, we know that $\pi_{-1}(\cdot \mid z_1) = \mathcal{N}(m_{z_1}, \Sigma_{|1})$ where

$$
\begin{aligned}
m_{z_1} &= m_{-1} + (\sigma_{11})^{-1} \Sigma_{-1,1} (z_1 - m_1)\,, \\
\Sigma_{|1} &= \Sigma_{-1,-1} - (\sigma_{11})^{-1} \Sigma_{-1,1} \Sigma_{-1,1}^\top = ((\Sigma^{-1})_{-1,-1})^{-1}\,.
\end{aligned}
$$

Therefore, Lacker et al. (2024, Theorem 1.1) yields that $q^\star(\cdot \mid z_1) = \mathcal{N}(m_{z_1}, \Sigma^\star_{|1})$ where $\Sigma^*_{|1} = \operatorname{diag}(\Sigma_{|1}^{-1})^{-1}$; and by Lemma A.2, we derive that $h_{\mathrm{val}}(z_1) \equiv C$ for some $C > 0$. Thus, we may conclude that $\pi^\star$ is also Gaussian. Finally, (16) can be derived using the tower property of expectation. □

Proof of Corollary 2.11. By Theorem 2.10 and the results by Lacker et al. (2024), we know that both $\pi^\star$ and $\bar{\pi}$ are normal distributions with mean zero and covariance matrices $\Sigma^\star, \bar{\Sigma}$, respectively. Then Lemma A.2 implies that

$$
\mathsf{KL}\left(\bar{\pi} \parallel \pi\right) = \frac{1}{2} \log \frac{\det \bar{\Sigma}}{\det \Sigma} + \frac{1}{2} \left(\operatorname{tr}(\bar{\Sigma} \Sigma^{-1}) - d \right),
$$

and

$$
\mathsf{KL}\left(\pi^\star \parallel \pi\right) = \frac{1}{2} \log \frac{\det \Sigma^\star}{\det \Sigma} + \frac{1}{2} \left(\operatorname{tr}(\Sigma^\star \Sigma^{-1}) - d \right).
$$

Thus,

$$
\mathsf{KL}\left(\pi^\star \parallel \pi\right) - \mathsf{KL}\left(\bar{\pi} \parallel \pi\right) = \frac{1}{2} \log \frac{\det \Sigma^\star}{\det \bar{\Sigma}} + \frac{1}{2} \operatorname{tr}((\Sigma^\star - \bar{\Sigma}) \Sigma^{-1})\,. \tag{41}
$$

By (16), we can express the difference between $\Sigma^\star$ and $\bar{\Sigma}$ as follows:

$$
\begin{aligned}
\Sigma^\star - \bar{\Sigma} &= \begin{pmatrix}
\sigma_{11} - (\sigma^{11})^{-1} & \sigma_{12} & \cdots & \sigma_{1i} & \cdots & \sigma_{1d} \\
\sigma_{12} & \dfrac{\sigma_{12}^2}{\sigma_{11}} & \cdots & \dfrac{\sigma_{12}\sigma_{1i}}{\sigma_{11}} & \cdots & \dfrac{\sigma_{12}\sigma_{1d}}{\sigma_{11}} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
\sigma_{1i} & \dfrac{\sigma_{1i}\sigma_{12}}{\sigma_{11}} & \cdots & \dfrac{\sigma_{1i}^2}{\sigma_{11}} & \cdots & \dfrac{\sigma_{1i}\sigma_{1d}}{\sigma_{11}} \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
\sigma_{1d} & \dfrac{\sigma_{1d}\sigma_{12}}{\sigma_{11}} & \cdots & \dfrac{\sigma_{1d}\sigma_{1i}}{\sigma_{11}} & \cdots & \dfrac{\sigma_{1d}^2}{\sigma_{11}}
\end{pmatrix} \\
&=: \Sigma + \tilde{\Sigma}\,,
\end{aligned}
$$

where

$$
\tilde{\Sigma} := \begin{pmatrix}
-(\sigma^{11})^{-1} & 0 & \cdots & 0 & \cdots & 0 \\
0 & \dfrac{\sigma_{12}^2}{\sigma_{11}} - \sigma_{22} & \cdots & \dfrac{\sigma_{12}\sigma_{1i}}{\sigma_{11}} - \sigma_{2i} & \cdots & \dfrac{\sigma_{12}\sigma_{1d}}{\sigma_{11}} - \sigma_{2d} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
0 & \dfrac{\sigma_{1i}\sigma_{12}}{\sigma_{11}} - \sigma_{i2} & \cdots & \dfrac{\sigma_{1i}^2}{\sigma_{11}} - \sigma_{ii} & \cdots & \dfrac{\sigma_{1i}\sigma_{1d}}{\sigma_{11}} - \sigma_{id} \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
0 & \dfrac{\sigma_{1d}\sigma_{12}}{\sigma_{11}} - \sigma_{d2} & \cdots & \dfrac{\sigma_{1d}\sigma_{1i}}{\sigma_{11}} - \sigma_{di} & \cdots & \dfrac{\sigma_{1d}^2}{\sigma_{11}} - \sigma_{dd}
\end{pmatrix}.
$$

Then

$$\begin{aligned}\operatorname{tr}\Big((\Sigma^\star - \bar{\Sigma})\Sigma^{-1}\Big) &= \operatorname{tr}(I + \tilde{\Sigma}\Sigma^{-1}) = d + \operatorname{tr}(\tilde{\Sigma}\Sigma^{-1}) \\ &= d - (\sigma^{11})^{-1}\sigma^{11} + \sum_{i\geq 2}\sum_{j\geq 2}\Big(\frac{\sigma_{1i}\sigma_{1j}}{\sigma_{11}} - \sigma_{ij}\Big)\sigma^{ji} \\ &= d - 1 + \sum_{i\geq 2}\frac{\sigma_{1i}}{\sigma_{11}}\sum_{j\geq 2}\sigma_{1j}\sigma^{ji} - \sum_{i\geq 2}\sum_{j\geq 2}\sigma_{ij}\sigma^{ji}\,.\end{aligned}$$

By the fact that $\Sigma\Sigma^{-1} = I$, we have

$$\operatorname{tr}\Big((\Sigma^\star - \bar{\Sigma})\Sigma^{-1}\Big) = d - 1 + \sum_{i\geq 2}\frac{\sigma_{1i}}{\sigma_{11}}\left(-\sigma_{11}\sigma^{1i}\right) - \sum_{i\geq 2}(1 - \sigma_{i1}\sigma^{1i}) = 0\,.$$

To compute the determinant of $\Sigma^\star$, we observe that: letting $U_i \coloneqq \begin{pmatrix} 1 & v_i^\top \\ 0 & I_{d-1}\end{pmatrix}$, where $v_i \in \mathbb{R}^{d-1}$ with its $(i-1)$-th entry being $-\sigma_{11}^{-1}\sigma_{1i}$ and all the others being zero,

$$\Big(\prod_{i=2}^{d} U_i\Big)^\top \Sigma^\star \Big(\prod_{i=2}^{d} U_i\Big) = \operatorname{diag}(\sigma_{11}, (\sigma^{22})^{-1}, \ldots, (\sigma^{dd})^{-1})\,.$$

Since $\det U_i = 1$, we obtain that

$$\det \Sigma^\star = \sigma_{11}\prod_{i=2}^{d}(\sigma^{ii})^{-1}\,.$$

On the other hand, we know that $\det\bar{\Sigma} = (\prod_{i=1}^{d}\sigma^{ii})^{-1}$, thus

$$\frac{\det\Sigma^\star}{\det\bar{\Sigma}} = \frac{\sigma_{11}}{(\sigma^{11})^{-1}}\,. \tag{42}$$

Substitute (42) into (41), and by the fact that $\sigma_{11} \geq (\sigma^{11})^{-1}$, we obtain

$$\mathsf{KL}\left(\pi^\star \parallel \pi\right) - \mathsf{KL}\left(\bar{\pi} \parallel \pi\right) = -\frac{1}{2}\log\Big(\frac{\sigma_{11}}{(\sigma^{11})^{-1}}\Big) \leq 0\,.$$

□

Proof of Theorem 2.12. Recall that the target measure has the form

$$V_{\mathrm{loc}}(\vartheta, \beta) = \mathsf{g}(\vartheta) - w^\top\beta + \sum_{i=1}^{n}\frac{\psi(X_i^\top\beta)}{c(\sigma)} + \sum_{j=1}^{d}\varrho(\beta_j - \vartheta)\,.$$

To verify the conditions in Theorem 2.9, we compute $\partial_\vartheta^2 V_{\mathrm{loc}}$, the principal minor $\nabla_\beta^2 V_{\mathrm{loc}}$, and the cross terms. By the assumptions on g and ϱ, we have

$$\partial_\vartheta^2 V_{\mathrm{loc}}(\vartheta, \beta) = \sum_{j=1}^{d}\varrho''(\beta_j - \vartheta) + \mathsf{g}''(\vartheta) \leq dR + L\,.$$

Letting D be a $d \times d$ diagonal matrix with $D_{jj} = \varrho''(\beta_j - \vartheta) \geq 0$, the principal minor of $\nabla^2 V_{\mathrm{loc}}$ satisfies

$$\nabla^2_\beta V_{\mathrm{loc}}(\vartheta, \beta) = \sum_{i=1}^n \frac{X_i X_i^\top}{c(\sigma)} \psi''(X_i^\top \beta) + D \succeq bA \succeq baI\,,$$

as $A = \sum_{i=1}^n \frac{X_i X_i^\top}{c(\sigma)} \succeq aI$ and $\psi'' \geq b$. Finally, the cross-term is $\partial_{\vartheta\beta_j} V_{\mathrm{loc}} = -\varrho''(\beta_j - \vartheta)$. Then the root-domination criteria is satisfied as

$$\begin{aligned}
&\partial^2_\vartheta V_{\mathrm{loc}} - \frac{\left\| \sum_{j=1}^d \left(\partial_{\vartheta\beta_j} V_{\mathrm{loc}}\right)^2 \right\|_{L^\infty}}{ab} \\
&\qquad = \mathsf{g}''(\vartheta) + \sum_{j=1}^d \varrho''(\beta_j - \vartheta) - \frac{\left\| \sum_{j=1}^d \varrho''(\beta_j - \vartheta)^2 \right\|_{L^\infty}}{ab} \geq \alpha - \frac{dR^2}{ab}\,.
\end{aligned}$$

The computation above verifies assumptions (**R**), (**P1**), (**P2**), and (**RD**), with parameters given by $L'_V \leftarrow 2(dR + L)$, $\ell'_V \leftarrow \alpha - \frac{dR^2}{ab}$, $\ell_V \leftarrow ab$. Finally, by the definition of $\mathcal{A}(\beta)$, we have

$$\sum_{i=1}^d \sum_{j>i} \left(\partial_{ij} V_{\mathrm{loc}}\right)^2 = \sum_{i=1}^d \sum_{j>i} \mathcal{A}^2_{ij}(\beta) \;\leq\; \left\| \sum_{i=1}^d \sum_{j>i} \mathcal{A}^2_{ij}(\beta) \right\|_{L^\infty}.$$

Then, Theorem 2.9 implies the conclusion. □

Proof of Proposition 2.13. Note that $\varrho(t) = \frac{\tau^2}{2} t^2$, $\psi(t) = \frac{t^2}{2}$, $c(\sigma) = \sigma^2$. Then $\psi'' = 1$, $\varrho'' = \tau^2$ (note that this is stronger than the assumption $\varrho'' \geq 0$ in Theorem 2.12), $\ell_V = a + \tau^2$, and

$$\begin{aligned}
\partial^2_\vartheta V_{\mathrm{loc}} - \frac{\left\| \sum_{j=1}^d \left(\partial_{\vartheta\beta_j} V_{\mathrm{loc}}\right)^2 \right\|_{L^\infty}}{a + \tau^2} &= \mathsf{g}''(\vartheta) + d\tau^2 - \frac{d\tau^4}{a + \tau^2} \\
&= \mathsf{g}''(\vartheta) + \frac{ad\tau^2}{a + \tau^2} \geq \ell'_V > 0.
\end{aligned}$$

In addition, $L'_V = 2\,(d\tau^2 + L)$. Applying Theorem 2.9 yields the result. □

Lemma A.3 (Semi-log-concavity of a scale mixture of two Gaussians)**.** *Let ν_0, ν_1 be normal densities with $\nu_0 \coloneqq \mathcal{N}(0, \tau_0^{-2})$ and $\nu_1 \coloneqq \mathcal{N}(0, \tau_1^{-2})$, where $\tau_1^2 < \tau_0^2$. For $\eta \in [0, 1]$, define the function*

$$\xi(x) \;\coloneqq\; -\log\left(\eta\, \nu_0(x) + (1 - \eta)\, \nu_1(x)\right).$$

Then, its second derivative satisfies

$$\xi''(x) \geq \tau_1^2 \;-\; 2\,(\tau_0^2 - \tau_1^2) \log\!\left(1 + \frac{\eta\tau_0}{(1 - \eta)e\tau_1}\right).$$

Proof. Let $\zeta \coloneqq \frac{\eta}{1-\eta}$, and let $w \coloneqq \frac{\zeta\nu_0}{\zeta\nu_0 + \nu_1}$. Here we suppress the argument x for simplicity. For $k \in \{0, 1\}$, a calculation shows that

$$\nu'_k = -\tau_k^2 x \nu_k\,, \qquad \nu''_k = \tau_k^4 x^2 \nu_k - \tau_k^2 \nu_k\,.$$

Then, we have some useful identities:

$$(\nu_k')^2 - \nu_k \nu_k'' = \tau_k^2 \nu_k^2\,, \qquad 2\nu_0'\nu_1' - \nu_1\nu_0'' - \nu_0\nu_1'' = -(\tau_0^2-\tau_1^2)^2 x^2 \nu_0\nu_1 + (\tau_0^2+\tau_1^2)\nu_0\nu_1\,.$$

Therefore,

$$\begin{aligned}\left(\frac{\zeta\nu_0'+\nu_1'}{\zeta\nu_0+\nu_1}\right)^2 - \frac{\zeta\nu_0''+\nu_1''}{\zeta\nu_0+\nu_1} &= \frac{(\zeta\nu_0'+\nu_1')^2 - (\zeta\nu_0+\nu_1)(\zeta\nu_0''+\nu_1'')}{\left(\zeta\nu_0+\nu_1\right)^2} \\ &= \underbrace{\frac{\zeta^2\tau_0^2\nu_0^2 + \tau_1^2\nu_1^2 + \zeta(\tau_0^2+\tau_1^2)\nu_0\nu_1}{\left(\zeta\nu_0+\nu_1\right)^2}}_{:=A_1} - \underbrace{\frac{\zeta(\tau_0^2-\tau_1^2)^2x^2\nu_0\nu_1}{\left(\zeta\nu_0+\nu_1\right)^2}}_{:=A_2}\,.\end{aligned}$$

We lower bound A_1 by completing the square:

$$A_1 = \frac{\tau_0^2+\tau_1^2}{2} + \frac{(\tau_0^2-\tau_1^2)(\zeta^2\nu_0^2-\nu_1^2)}{2\left(\zeta\nu_0+\nu_1\right)^2} = \frac{\tau_0^2+\tau_1^2}{2} + \frac{(\tau_0^2-\tau_1^2)(\zeta\nu_0-\nu_1)}{2(\zeta\nu_0+\nu_1)}\,.$$

As $|\frac{\zeta\nu_0-\nu_1}{\zeta\nu_0+\nu_1}| < 1$, we can further bound A_1:

$$A_1 \geq \frac{\tau_0^2+\tau_1^2}{2} - \frac{\tau_0^2-\tau_1^2}{2} = \tau_1^2\,.$$

For A_2, we derive a bound in the following steps:

(a) With $f(t) \coloneqq \frac{t}{\zeta\tau_0+\tau_1\exp\left(\frac{1}{2}(\tau_0^2-\tau_1^2)t\right)}$, we have

$$A_2 \leq \frac{(\tau_0^2-\tau_1^2)^2x^2\zeta\nu_0}{\zeta\nu_0+\nu_1} = (\tau_0^2-\tau_1^2)^2\zeta\tau_0 f(x)\,.$$

(b) Note that $\lim_{t\to\infty} f(t) = 0$ and $f(0) = 0$. The maximum must be achieved at the stationary point $f'(t) = 0$. The first-order condition implies

$$\left(\frac{1}{2}(\tau_0^2-\tau_1^2)\,t - 1\right)\exp\left(\frac{1}{2}(\tau_0^2-\tau_1^2)\,t - 1\right) = \frac{\zeta\tau_0}{e\tau_1}\,.$$

Let $W(\cdot)$ denote the Lambert W function, defined by $W(a) = u$ when $u\exp(u) = a$. The optimal solution t^* solves

$$\frac{1}{2}(\tau_0^2-\tau_1^2)\,t^* - 1 = W\left(\frac{\zeta\tau_0}{e\tau_1}\right)\,.$$

Therefore, by the first-order condition, we have

$$\zeta\tau_0 + \tau_1\exp\left(\frac{1}{2}(\tau_0^2-\tau_1^2)\,t^*\right) = \frac{1}{2}(\tau_0^2-\tau_1^2)\,\tau_1 t^*\exp\left(\frac{1}{2}(\tau_0^2-\tau_1^2)\,t^*\right),$$

and

$$\tau_1 \exp\Big(\frac{1}{2}\,(\tau_0^2-\tau_1^2)\,t^*\Big) = \frac{\zeta\tau_0}{W\big(\frac{\zeta\tau_0}{e\tau_1}\big)}\,.$$

Plugging in t^* yields the maximum of f,

$$f(t^*) = \frac{2}{(\tau_0^2-\tau_1^2)\tau_1 \exp\left(\frac{1}{2}(\tau_0^2-\tau_1^2)t^*\right)} = \frac{2W\big(\frac{\zeta\tau_0}{e\tau_1}\big)}{(\tau_0^2-\tau_1^2)\zeta\tau_0}\,.$$

Thus, $A_2 \le (\tau_0^2-\tau_1^2)^2\zeta\tau_0 f(t^*) = 2(\tau_0^2-\tau_1^2)\,W\big(\frac{\zeta\tau_0}{e\tau_1}\big)$.

(c) Since $\exp(\ln(1+z))-1 = z$ and $\exp(u)-1 < u\exp(u)$, $W(z) \le \log(1+z)$. We conclude with the desired bound.

□

Proof of Theorem 2.14. We again compute the requisite partial derivatives. The second partial derivative of V_{ss} with respect to β_1 is given by

$$\partial_{11}V_{\mathrm{ss}}(\beta) = \left[\sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta)\right]_{11} + \mathsf{g}''\Big(\beta_1 + \sum_{j=2}^d \gamma_j\beta_j\Big)\,.$$

By our assumptions on ψ and g, we have the following uniform bounds:

$$bA_{11} + \alpha \le \partial_{11}V_{\mathrm{ss}}(\beta) \le BA_{11} + L\,.$$

Thus, (**R**) is satisfied with $L_V' = 2\,(BA_{11}+L)$.

To continue, define $\zeta \coloneqq \eta/(1-\eta)$ and $\xi(x) \coloneqq -\log\left(\eta\,\nu_0(x) + (1-\eta)\,\nu_1(x)\right)$. After some simplifications, one can compute the principal minor of the Hessian of V_{ss} to be

$$\begin{aligned}\left[\nabla^2 V_{\mathrm{ss}}(\beta)\right]_{-1} &= \left[\sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta)\right]_{-1} + \mathrm{diag}([\xi''(\beta_k)]_{k=2}^d) + [\gamma_i\gamma_j\,\mathsf{g}'']_{2\le i,j\le d} \\ &\succeq bA_{-1} + \mathrm{diag}([\xi''(\beta_k)]_{k=2}^d) \succeq \ell_V I_{d-1}\,,\end{aligned}$$

where the last two inequalities follow by our assumptions and Lemma A.3.

We now verify (**RD**). For each $i \ge 2$, the mixed derivative is

$$\partial_{1j}V_{\mathrm{ss}}(\beta) = \left[\sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta)\right]_{1j} + \gamma_j\,\mathsf{g}''\Big(\beta_1 + \sum_{j=2}^d \gamma_j\beta_j\Big)\,. \tag{43}$$

Since $ba_dI_d \preceq \sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta) \preceq Ba_1I_d$, the Schur complement theorem (Zhang, 2006, Theorem 1.6) implies

$$\left[\sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta)\right]_{11} - ba_d - \frac{\sum_{j=2}^d\Big(\Big[\sum_{i=1}^n \frac{X_iX_i^\top}{c(\sigma)}\,\psi''(X_i^\top\beta)\Big]_{1j}\Big)^2}{Ba_1 - ba_d} \ge 0\,. \tag{44}$$

For ease of notation, we omit β in the following expressions. Combining (43) and (44), we readily compute

$$\sum_{j=2}^{d} \left(\partial_{1j} V_{\mathrm{ss}}\right)^2 \le 2\left(Ba_1 - ba_d\right)\left(BA_{11} - ba_d\right) + 2\sum_{j=2}^{d} \gamma_j^2 L^2 \,,$$

thus

$$\begin{aligned}
\partial_{11} V_{\mathrm{ss}} - \frac{\left\|\sum_{j=2}^{d} \left(\partial_{1j} V_{\mathrm{ss}}\right)^2\right\|_{L^\infty}}{ba_d} \\
\ge bA_{11} + \alpha - \frac{2\sum_{j=2}^{d} \gamma_j^2 L^2}{ba_d} - \frac{2\left(Ba_1 - ba_d\right)\left(BA_{11} - ba_d\right)}{ba_d} = \ell_V' > 0 \,.
\end{aligned}$$

This verifies (**RD**). Now, computing the cross derivatives $\partial_{ij} V_{\mathrm{ss}}$ for $i \neq j \ge 2$, we find

$$\partial_{ij} V_{\mathrm{ss}}(\beta) = \sum_{k=1}^{n} \frac{\psi''(X_k^\top \beta)}{c(\sigma)} X_{ki} X_{kj} + \gamma_i \gamma_j g''\Big(\beta_1 + \sum_{j=2}^{d} \gamma_j \beta_j\Big).$$

Finally, using the definition of $\mathcal{A}(\beta)$ and $g'' \le L$, we have

$$\frac{1}{2}\sum_{i=2}^{d}\sum_{j>i} \left(\partial_{ij} V_{\mathrm{ss}}\right)^2 \le \Big\|\sum_{i=2}^{d}\sum_{j>i} \mathcal{A}_{ij}^2\Big\|_{L^\infty} + L^2 \sum_{i=2}^{d}\sum_{j>i} \gamma_i^2 \gamma_j^2 \,.$$

We apply Theorem 2.9 to yield the desired result. □

Proof of Proposition 2.15. The full posterior is given by $\pi_{\mathrm{ss}} \propto \exp(-V_{\mathrm{ss}})$ with

$$V_{\mathrm{ss}}(\beta) = \frac{\beta^\top A \beta}{2} - \sum_{j=2}^{d} \log\left(\eta\nu_0(\beta_j) + (1-\eta)\nu_1(\beta_j)\right) + \frac{\tau^2}{2}\left(\beta_1 + \sum_{j=2}^{d} \gamma_j \beta_j\right)^2 .$$

In the remainder of the proof, we verify that assumptions (**R**), (**P1**), and (**RD**) hold with high probability.

The second-order partial gradient of V_{ss} with respect to β_1 is $\partial_{11} V_{\mathrm{ss}}(\beta) = A_{11} + \tau^2$, thus (**R**) is satisfied with $L_V' = 2\left(A_{11} + \tau^2\right)$.

We now verify (**P1**). Let $\delta > 0$ be given. By Wainwright (2019, Theorem 6.1), with probability at least $1 - e^{-n\delta^2/2}$,

$$\lambda_{\min}(A) = \sigma_{\min}(\boldsymbol{X})^2 \ge \left(\sqrt{n\lambda_d}\,(1-\delta) - \sqrt{d\lambda_1}\right)_+^2 .$$

Take $\delta = 0.9$. Since $\lambda_1/\lambda_d = o(n/d)$ by assumption, for n sufficiently large, the lower bound is at least $0.8 n\lambda_d$. That is, $A \succeq 0.8\, n\lambda_d I_d$.

For the full potential, this further implies $\nabla_{-1}^2 V_{\mathrm{ss}} \succeq 0.8\, n\lambda_d I_{d-1} + D$, where D is a diagonal matrix that consists of the second derivatives of $-\log(\eta\nu_0(\beta_j) + (1-\eta)\nu_1(\beta_j))$ for $j = 2, \ldots, d$. By Lemma A.3, $D \succeq c_0 I_{d-1}$ for some constant c_0 independent of d and n. This implies that, for sufficiently large n, it holds that $\nabla_{-1}^2 V_{\mathrm{ss}} \succeq \frac{2n\lambda_d}{3} I_{d-1}$. Therefore, (**P1**) holds with the constant $\ell_V = \frac{2n\lambda_d}{3}$ with probability at least $1 - 2e^{-0.4n} = 1 - o(1)$.

It remains to verify (**RD**). As $\partial_{1j} V_{\mathrm{ss}}(\beta) = A_{1j} + \tau^2 \gamma_j$ for $j \geq 2$ and $(a+b)^2 \leq 2a^2 + 2b^2$, we know that

$$A_{11} + \tau^2 - \frac{\frac{3}{2}\sum_{j=2}^d (A_{1j} + \tau^2 \gamma_j)^2}{n\lambda_d} \geq \underbrace{A_{11} - \frac{3\sum_{j=2}^d A_{1j}^2}{n\lambda_d}}_{=:D_1} + \underbrace{\tau^2 - \frac{3\tau^4 \sum_{j=2}^d \gamma_j^2}{n\lambda_d}}_{=:D_2}. \tag{45}$$

We proceed by bounding D_1, D_2 separately.

By Young's inequality,

$$\begin{aligned}\sum_{j=2}^d A_{1j}^2 &= \sum_{j=2}^d \langle e_1, A\, e_j\rangle^2 \leq 1.1 \sum_{j=2}^d \langle e_1, n\Sigma\, e_j\rangle^2 + O\Big(\sum_{j=2}^d \langle e_1, (A - n\Sigma)\, e_j\rangle^2\Big) \\ &\leq 1.1 n^2 \sum_{j=2}^d |\Sigma_{1j}|^2 + O\Big(\|(A - n\Sigma)\, e_1\|_2^2\Big) \leq 1.1 n^2 \sum_{j=2}^d |\Sigma_{1j}|^2 + O(\|A - n\Sigma\|_2^2)\,.\end{aligned}$$

By Theorem 6.1 and Example 6.3 in Wainwright (2019),

$$\left\| n^{-1} A - \Sigma \right\|_2 \leq \lambda_1 \left(2\epsilon + \epsilon^2\right),$$

where $\epsilon := \sqrt{\frac{d}{n}} + \delta$, with probability at least $1 - 2e^{-n\delta^2/2}$. Then, setting $\delta = \sqrt{d/n}$, we see that with probability $1 - o(1)$,

$$\sum_{j=2}^d A_{1j}^2 \leq 1.1 n^2 \sum_{j=2}^d |\Sigma_{1j}|^2 + O(dn\lambda_1^2)\,. \tag{46}$$

On the other hand, A_{11} is a sum of n i.i.d. χ_1^2 random variables scaled by Σ_{11}, i.e., A_{11} is a χ_n^2 random variable scaled by Σ_{11}. By concentration of χ^2 random variables (Wainwright, 2019, Example 2.5), it follows that $A_{11} \geq \frac{2}{3} n\Sigma_{11}$ with high probability.

Combining with (46), we conclude that

$$D_1 \geq \frac{2}{3} n\Sigma_{11} - \frac{3.3 n^2 \sum_{j=2}^d |\Sigma_{1j}|^2 + O(dn\lambda_1^2)}{n\lambda_d}\,.$$

Since $d\lambda_1^2/\lambda_d \ll n\lambda_d \ll n\Sigma_{11}$, with probability $1 - o(1)$, we derive that

$$D_1 \geq \frac{n}{2}\left(\Sigma_{11} - \frac{8\sum_{j=2}^d |\Sigma_{1j}|^2}{\lambda_d}\right) = \ell_V'\,. \tag{47}$$

Moving on to D_2. Recall that $D_2 := \tau^2 - \frac{3\tau^4 \sum_{j=2}^d \gamma_j^2}{n\lambda_d}$. By the preceding arguments,

$$\sum_{j=2}^d \gamma_j^2 = \frac{\sum_{j=2}^d A_{1j}^2}{A_{11}^2} \leq \frac{1.1 n^2 \sum_{j=2}^d |\Sigma_{1j}|^2 + O(dn\lambda_1^2)}{4n^2\Sigma_{11}^2/9}\,.$$

Therefore, for $\lambda_1/\lambda_d = o(\sqrt{n/d})$, as $\ell'_V = \frac{n}{2}\left(\Sigma_{11} - \frac{8\sum_{j=2}^d |\Sigma_{1j}|^2}{\lambda_d}\right) > 0$.

$$D_2 \geq \tau^2 - \frac{8\tau^4 \sum_{j=2}^d |\Sigma_{1j}|^2}{n\lambda_d \Sigma_{11}^2} - O\Big(\frac{d\tau^4\lambda_1^2}{n^2\lambda_d^3}\Big) = \tau^2 + \frac{2\tau^4 \ell'_V}{n^2\Sigma_{11}^2} - \frac{\tau^4}{n\Sigma_{11}} - o\Big(\frac{\tau^4}{n\lambda_d}\Big)$$
$$\geq \tau^2 - O\Big(\frac{\tau^4}{n\lambda_d}\Big). \tag{48}$$

Hence, $D_2 \geq 0$ with probability $1 - o(1)$ provided $\tau^2 = o(n\lambda_d)$.

By combining (45), (47), and (48), (**RD**) is satisfied with the defined ℓ'_V with high probability. Finally, applying Theorem 2.9 with $L'_V = 2\,(A_{11} + \tau^2)$, $\ell_V = \frac{n\lambda_d}{2}$, and ℓ'_V, and noting that $\partial_{ij} V_{\mathrm{ss}}(\beta) = A_{ij} + \tau^2 \gamma_i \gamma_j$ for $i \neq j \geq 2$, we obtain

$$\mathsf{KL}\left(\pi^\star_{\mathrm{ss}} \parallel \pi_{\mathrm{ss}}\right) \leq \frac{8\,(A_{11} + \tau^2)}{n^2 \ell'_V \lambda_d^2} \sum_{i=2}^d \sum_{j>i} \left(A_{ij}^2 + \tau^4 \gamma_i^2 \gamma_j^2\right).$$

This concludes the proof. □

B Proofs from Section 3

Proof of Proposition 3.1. Consider two elements $T, \tilde{T} \in \mathcal{T}_{\mathrm{star}}$. We aim to show $\lambda T + (1-\lambda)\tilde{T} \in \mathcal{T}_{\mathrm{star}}$ for any $\lambda \in (0,1)$. Let $X = (X_1, \ldots, X_d) \sim \rho$ and denote $Y \coloneqq T(X)$, $Z \coloneqq \tilde{T}(X)$. It is equivalent to show that $\mathrm{Law}(\lambda Y + (1-\lambda)Z) \in \mathcal{C}_{\mathrm{star}}$.

Let $S_1, S_2 \subseteq \{2, \ldots, d\}$ be two sets such that $S_1 \cap S_2 = \emptyset$. As $\rho = \mathcal{N}(0, I)$, we know that $X_{S_1} \perp X_{S_2}$, where $X_{S_l} \coloneqq (X_j)_{j \in S_l}$ for $l \in \{1, 2\}$. Thus, as

$$\lambda Y_i + (1-\lambda) Z_i = \lambda T_i\left(X_i \mid T_1(X_1)\right) + (1-\lambda)\tilde{T}_i(X_i \mid \tilde{T}_1(X_1))$$

for every $i \in \{2, \ldots, d\}$, we conclude that

$$\{\lambda Y_i + (1-\lambda) Z_i\}_{i \in S_1} \perp \{\lambda Y_j + (1-\lambda) Z_j\}_{j \in S_2} \mid X_1\,. \tag{49}$$

Furthermore, since $\lambda T_1 + (1-\lambda)\tilde{T}_1$ is strictly increasing, the σ-algebra generated by $\lambda Y_1 + (1-\lambda)Z_1$ is the same as the one generated by X_1. Combining with (49) implies

$$\{\lambda Y_i + (1-\lambda) Z_i\}_{i \in S_1} \perp \{\lambda Y_j + (1-\lambda) Z_j\}_{j \in S_2} \mid (\lambda Y_1 + (1-\lambda) Z_1)\,.$$

Hence, applying the Hammersley–Clifford theorem (see, e.g., Wainwright, 2019, Theorem 11.8) shows the result. □

Before presenting the proof of Theorem 3.2, we provide an alternative proof of the existence and uniqueness of the solution to (6), relying on the convexity of the equivalent problem (T-SSVI) and the stability of conditional independence under convergence in $L^2(\rho)$ over $\mathcal{T}_{\mathrm{star}}$. A recent work by Beiglböck et al. (2023) discusses the close relation between the topology induced by star-separable maps (and more generally, Knothe–Rosenblatt maps) and the one induced by the adapted Wasserstein distance. The latter is finer than the information topology (Backhoff-Veraguas et al., 2020), which preserves conditional independence.

The *adapted* (2-)Wasserstein distance is defined as

$$AW_2^2(\mu,\nu) \coloneqq \inf_{\pi\in\Pi_{\mathrm{bc}}(\mu,\nu)} \int \|x-y\|^2 \,\pi(\mathrm{d}x,\mathrm{d}y)\,, \tag{50}$$

where we call $\Pi_{\mathrm{bc}}(\mu,\nu)$ the set of *bicausal* couplings between μ and ν. This set consists of $\pi \in \Pi(\mu,\nu)$ such that

$$\pi_k(\mathrm{d}x_k,\mathrm{d}y_k \mid x_{1:k-1},y_{1:k-1}) \in \Pi\left(\mu(\mathrm{d}x_k \mid x_{1:k-1}),\nu(\mathrm{d}y_k \mid y_{1:k-1})\right), \quad \pi\text{-a.s.}$$

for all $k \le d$, i.e.,

$$\pi(\mathrm{d}x_{1:d},\mathrm{d}y_{1:d}) = \pi_1(\mathrm{d}x_1,\mathrm{d}y_1)\,\pi_2(\mathrm{d}x_2,\mathrm{d}y_2 \mid x_1,y_1)\cdots\pi_d(\mathrm{d}x_d,\mathrm{d}y_d \mid x_{1:d-1},y_{1:d-1})\,.$$

Remark B.1. To connect with star-separable maps, we observe that for any $T_1, T_2 \in \mathcal{T}_{\mathrm{star}}$ with $T_{1\#}\rho = \mu$ and $T_{2\#}\rho = \nu$, where $\mathcal{T}_{\mathrm{star}}$ is defined in Section 3.1, then $(T_1,T_2)_{\#}\rho \in \Pi_{\mathrm{bc}}(\mu,\nu)$. Thus, $AW_2(\mu,\nu) \le \|T_1 - T_2\|_{L^2(\rho)}$.

Definition B.2 (Hellwig's information topology)**.** Define the family of maps $\{\mathcal{I}_k\}_{k=1}^{d-1}$ by

$$\begin{aligned}
\mathcal{I}_k &: \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^k \times \mathcal{P}(\mathbb{R}^{d-k}))\,,\\
\mathcal{I}_k(\mu) &\coloneqq \mathcal{K}^k_{\#}\mu\,,\\
\mathcal{K}^k(x_1,\dots,x_d) &\coloneqq (x_1,\dots,x_k,\mu(\mathrm{d}x_{k+1},\dots,\mathrm{d}x_d \mid x_1,\dots,x_k))\,.
\end{aligned}$$

Hellwig's information topology is the coarest topology such that $\mathcal{I}_k$ is continuous for all $k \le d-1$, where the space $\mathcal{P}(\mathbb{R}^k \times \mathcal{P}(\mathbb{R}^{d-k}))$ is endowed with the usual topology of weak convergence.

The following theorem says that the topology induced by the adapted Wasserstein distance is finer than the information topology.

Theorem B.3 (Backhoff-Veraguas et al. (2020, Theorem 1.2, Theorem 1.3, and Lemma 1.4))**.** *Convergence in the adapted Wasserstein distance is equivalent to convergence in the information topology plus convergence of the second moment.*

Finally, we recall the following fact which implies that convergence in the adapted Wasserstein distance preserves conditional independence.

Theorem B.4 (Barbie and Gupta (2014, Theorem 4.5))**.** *Let $\mathcal{X},\mathcal{Y},\mathcal{Z}$ be Polish spaces, $\{(\mu_n,\nu_n)\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \times \mathcal{P}(\mathcal{X}\times\mathcal{Z})$. Suppose that $\mu_n \to \mu$ in the information topology and $\nu_n \to \nu$ weakly. Let μ_n^x denote the transition kernel of μ_n w.r.t. the first marginal. Then,*

$$\nu_n \otimes \mu_n^x \to \nu \otimes \mu^x \quad \textit{weakly}\,.$$

Given $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, let $\pi_{xy} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ (resp. $\pi_{xz} \in \mathcal{P}(\mathcal{X}\times\mathcal{Z})$) denote the pushforward of π by the projection operator $(x,y,z) \mapsto (x,y)$ (resp. $(x,y,z) \mapsto (x,z)$).

Proposition B.5. *Given a sequence of probability measures $\{\pi_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ such that $\pi_n \to \pi$ weakly, suppose that $\pi_n = \pi_{n,xz} \otimes \pi^x_{n,xy}$, i.e., $\pi_n(\mathrm{d}y \mid x, z) = \pi_n(\mathrm{d}y \mid x)$, and $\pi_{n,xy} \to \pi_{xy}$ in the adapted Wasserstein distance. Then, $\pi = \pi_{xz} \otimes \pi^x_{xy}$.*

Proof. By Theorem B.3, $\pi_{n,xy} \to \pi_{xy}$ in the information topology. Applying Theorem B.4 to $\mu_n = \pi_{n,xy}$ and $\nu_n = \pi_{n,xz}$, the result follows from the uniqueness of the weak limit. □

Proof of Theorem 3.2. **Convexity:** This is similar to the proof that the KL divergence with respect to a strongly log-concave measure is strongly convex along generalized geodesics, see Ambrosio et al. (2008, Theorem 9.4.11).

Let $T_0, T_1 \in \mathcal{T} \subseteq \mathcal{T}_{\mathrm{star}}$, define $T_t \coloneqq (1-t)T_0 + tT_1$, and $\mu_t \coloneqq T_{t\#}\rho$. We now compute the second derivative of

$$t \mapsto \mathsf{KL}\left(\mu_t \parallel \pi\right) = \mathcal{V}(\mu_t) + \mathcal{H}(\mu_t) \coloneqq \int V \,\mathrm{d}\mu_t + \int \log \mu_t \,\mathrm{d}\mu_t \,.$$

The following two equalities can be easily verified (see, e.g., Villani, 2003, Chapter 5):

$$\frac{\mathrm{d}^2\mathcal{V}(\mu_t)}{\mathrm{d}t^2} = \int \left\langle \nabla^2 V(T_t)\,(T_1 - T_0),\, T_1 - T_0 \right\rangle \mathrm{d}\rho, \tag{51}$$

and

$$\frac{\mathrm{d}^2\mathcal{H}(\mu_t)}{\mathrm{d}t^2} = \int \|(DT_t)^{-1} D(T_1 - T_0)\|_{\mathrm{F}}^2 \,\mathrm{d}\rho \,. \tag{52}$$

It follows from (**SLC**) that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\,\mathsf{KL}(\mu_t \parallel \pi) \geq \alpha \,\|T_1 - T_0\|^2_{L^2(\rho)} \,,$$

which is what we wanted to show.

Existence and uniqueness: Let $\{\mu^n\}_{n\in\mathbb{N}} \subset \mathcal{C}_{\mathrm{star}}$ be a minimizing sequence, i.e., $\mathsf{KL}\left(\mu^n \parallel \pi\right) \to a \coloneqq \inf_{\mu\in\mathcal{C}_{\mathrm{star}}} \mathsf{KL}\left(\mu \parallel \pi\right)$ as $n \to \infty$. For every $m, n \in \mathbb{N}$, let $T^m \in \mathcal{T}_{\mathrm{star}}$ (resp. $T^n \in \mathcal{T}_{\mathrm{star}}$) denote the star-separable map from ρ to μ^m (resp. μ^n). By Proposition 3.1, $\mu^{m,n} \coloneqq (\frac{1}{2}T^m + \frac{1}{2}T^n)_{\#}\rho \in \mathcal{C}_{\mathrm{star}}$, thus $\mathsf{KL}\left(\mu^{m,n} \parallel \pi\right) \geq a$ and $\liminf_{m,n\to\infty} \mathsf{KL}\left(\mu^{m,n} \parallel \pi\right) \geq a$. By α-convexity of $T \mapsto \mathsf{KL}\left(T_{\#}\rho \parallel \pi\right)$, we have

$$\mathsf{KL}\left(\mu^{m,n} \parallel \pi\right) \leq \frac{1}{2}\,\mathsf{KL}\left(\mu^m \parallel \pi\right) + \frac{1}{2}\,\mathsf{KL}\left(\mu^n \parallel \pi\right) - \frac{\alpha}{8}\,\|T^m - T^n\|^2_{L^2(\rho)} \,.$$

As $\alpha > 0$, taking $\liminf_{m,n\to\infty}$ on both sides implies that

$$\limsup_{m,n\to\infty} \|T^m - T^n\|^2_{L^2(\rho)} = 0 \,.$$

Therefore, by completeness of $L^2(\rho)$, there exists $T^\star \in L^2(\rho)$ such that $T^n \to T^\star$ in $L^2(\rho)$. Denote by $\pi^\star \coloneqq T^\star_{\#}\rho$, then $\mu^n \to \pi^\star$ converges in Wasserstein distance and thus converges weakly. Combining with the weak lower semi-continuity of $\mu \mapsto \mathsf{KL}\left(\mu \parallel \pi\right)$, we derive that $\mathsf{KL}\left(\pi^\star \parallel \pi\right) \leq a$.

We claim that $\pi^\star \in \mathcal{C}_{\mathrm{star}}$, whence, we conclude that $\pi^\star$ is a minimizer to (3). Indeed, by the Hammersley–Clifford theorem (Wainwright, 2019, Theorem 11.8), it suffices to verify

that for any $1 < j \le d$, $\pi^\star_j(\mathrm{d}z_j \mid z_1, \dots, z_{j-1}) = \pi^\star_j(\mathrm{d}z_j \mid z_1)$. Since $T^n \to T^\star$ in $L^2(\rho)$, we know that $AW_2(\mu^n_{\{1,j\}}, \pi^\star_{\{1,j\}}) \to 0$ as the measurability is preserved under $L^2(\rho)$ convergence. Therefore, applying Proposition B.5 with $\pi_n = \mu^n_{[j]}$ and $\pi = \pi^\star_{[j]}$ yields the claim.

Finally, the uniqueness of $T^\star$ follows from the α-convexity and the convexity of $\mathcal{T}_{\mathrm{star}}$, which also shows the uniqueness of $\pi^\star$. □

In fact, the key is the convergence of T_n to $T^\star$ in $L^2(\rho)$, as this convergence preserves conditional independence. This also shows that the $L^2(\rho)$-closure of $\mathcal{T}_{\mathrm{star}}$ is also convex. Therefore, we can conclude that a minimizer exists over any set of structured distributions if the set of lifted maps is convex, which gives potential guidance for other structured VI problems.

C Computational guarantees

C.1 Regularity of star-separable maps

In this subsection, we provide the proofs of Lemma 3.4 and Theorem 3.5.

Lemma 3.4 establishes a bound on the mixed derivative of the SSVI minimizer. As this result is used repeatedly in the subsequent proofs, we restate it below for convenience.

Lemma C.1. *Let* (**R**), (**P1**), (**P2**), *and* (**RD+**) *be satisfied. Denote by*

$$L \coloneqq \max_{i \ge 2} \sup_{z_1, z_i \in \mathbb{R}} \left| \partial_1 \partial_i \log q^\star_i(z_i \mid z_1) \right|.$$

Then

$$L \le \overline{L} \coloneqq \frac{\ell_V \max_{i \ge 2} \|\partial_{1i} V\|_{L^\infty}}{\ell_V - \max_{i \ge 2} \left\| \sum_{j \ge 2,\, j \ne i} |\partial_{ij} V| \right\|_{L^\infty}} < +\infty .$$

Proof of Lemma 3.4. The proof uses the strategy of self-bounding. Denote $h_i(z_1, z_i) \coloneqq -\partial_1 \log q^\star_i(z_i \mid z_1)$. Theorem 2.4 yields that

$$\begin{aligned}
\partial_i h_i(z_1, z_i) &= \mathbb{E}_{\pi^\star}\left[\partial_{1i} V(Z) \mid Z_1 = z_1, Z_i = z_i\right] \\
&\quad + \int \partial_i V(z_1, z_i, z_{-\{1,i\}}) \sum_{j \ge 2,\, j \ne i} \partial_1 \log q^\star_j(z_j \mid z_1)\, q^\star_{-i}(\mathrm{d}z_{-\{1,i\}} \mid z_1) \\
&= \mathbb{E}_{\pi^\star}\left[\partial_{1i} V(Z) \mid Z_1 = z_1, Z_i = z_i\right] \\
&\quad - \mathbb{E}_{\pi^\star}\Big[\partial_i V(Z) \sum_{j \ge 2,\, j \ne i} h_j(Z_1, Z_j) \;\Big|\; Z_1 = z_1, Z_i = z_i\Big].
\end{aligned}$$

Therefore, as $\mathbb{E}_{\pi^\star}[h_j(Z_1, Z_j) \mid Z_1 = z_1, Z_i = z_i] = \mathbb{E}_{\pi^\star}[h_j(Z_1, Z_j) \mid Z_1 = z_1] = 0$ for all $j \ge 2$,

$i \neq j$, we have

$$\begin{aligned}
&|\partial_i h_i(z_1, z_i) - \mathbb{E}_{\pi^\star}\left[\partial_{1i} V(Z) \mid Z_1 = z_1, Z_i = z_i\right]| \\
&\quad = \left|\mathbb{E}_{\pi^\star}\Big[\partial_i V(Z) \sum_{j\geq 2,\, j\neq i} h_j(Z_1, Z_j) \;\Big|\; Z_1 = z_1, Z_i = z_i\Big]\right| \\
&\quad = \left|\sum_{j\geq 2,\, j\neq i} \mathbb{E}_{\pi^\star}\Big[\mathbb{E}_{\pi^\star}\left[\partial_i V(Z) \mid Z_1, Z_i, Z_j\right] h_j(Z_1, Z_j) \;\Big|\; Z_1 = z_1, Z_i = z_i\Big]\right| \\
&\quad = \left|\sum_{j\geq 2,\, j\neq i} \mathrm{Cov}_{\pi^\star}\Big(\mathbb{E}_{\pi^*}[\partial_i V(Z) \mid Z_1, Z_i, Z_j], h_j(Z_1, Z_j) \;\Big|\; Z_1 = z_1, Z_i = z_i\Big)\right|.
\end{aligned} \tag{53}$$

Using the asymmetric Brascamp–Lieb inequality (Carlen et al., 2013, Theorem 1.1) with $p = \infty$, $q = 1$, and noting that $-\partial_{jj} \log q_j^\star(z_j \mid z_1) = \mathbb{E}_{\pi^\star}[\partial_{jj} V(Z) \mid Z_1 = z_1, Z_j = z_j] \geq \ell_V$ by Theorem 2.4, we derive that for every $j \neq i \geq 2$,

$$\begin{aligned}
&\left|\mathrm{Cov}_{\pi^\star}\Big(\mathbb{E}_{\pi^*}[\partial_i V(Z) \mid Z_1, Z_i, Z_j], h_j(Z_1, Z_j) \;\Big|\; Z_1 = z_1, Z_i = z_i\Big)\right| \\
&\quad \leq \left(\int \left|\mathbb{E}_{\pi^*}\left[\partial_{ij} V(Z) \mid Z_1 = z_1, Z_i = z_i, Z_j = z_j\right]\right| q_j^\star(\mathrm{d}z_j \mid z_1)\right) \\
&\qquad \times \sup_{z_j \in \mathbb{R}} \left|\frac{\partial_j h_j(z_1, z_j)}{\mathbb{E}_{\pi^\star}\left[\partial_{jj} V(Z) \mid Z_1 = z_1, Z_j = z_j\right]}\right| \\
&\quad \leq \frac{L}{\ell_V}\, \mathbb{E}_{\pi^\star}\left[|\partial_{ij} V(Z)| \mid Z_1 = z_1, Z_i = z_i\right].
\end{aligned} \tag{54}$$

Combining (53) and (54), we obtain

$$\begin{aligned}
&|\partial_i h_i(z_1, z_i) - \mathbb{E}_{\pi^\star}\left[\partial_{1i} V(Z) \mid Z_1 = z_1, Z_i = z_i\right]| \\
&\qquad \leq \frac{L}{\ell_V} \sum_{j\geq 2,\, j\neq i} \mathbb{E}_{\pi^\star}\left[|\partial_{ij} V(Z)| \mid Z_1 = z_1, Z_i = z_i\right].
\end{aligned}$$

By the triangle inequality, we derive that

$$\begin{aligned}
&|\partial_i h_i(z_1, z_i)| \\
&\quad \leq \frac{L}{\ell_V} \sum_{j\geq 2,\, j\neq i} \mathbb{E}_{\pi^\star}\left[|\partial_{ij} V(Z)| \mid Z_1 = z_1, Z_i = z_i\right] + \mathbb{E}_{\pi^\star}\left[|\partial_{1i} V(Z)| \mid Z_1 = z_1, Z_i = z_i\right].
\end{aligned}$$

Taking the supremum over z_1, z_i and then maximum over $i \geq 2$ on both sides, we obtain

$$L - \max_{i\geq 2} \|\partial_{1i} V\|_{L^\infty} \leq \frac{L}{\ell_V} \max_{i\geq 2} \Big\| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \Big\|_{L^\infty}.$$

Under (**RD+**), the inequality above implies a bound on L:

$$L \leq \frac{\ell_V \max_{i\geq 2} \|\partial_{1i} V\|_{L^\infty}}{\ell_V - \max_{i\geq 2} \Big\| \sum_{j\geq 2,\, j\neq i} |\partial_{ij} V| \Big\|_{L^\infty}} = \overline{L}.$$

□

Lemma C.2. *Let* (R), (P1), (P2), *and* (RD+) *be satisfied. Let $T^\star$ be the optimal star-separable map from ρ to $\pi^\star$. Then, for any $z_1, \tilde{z}_1 \in \mathbb{R}$, we have*

$$\sum_{i=2}^{d} \int |T_i^\star(x_i \mid \tilde{z}_1) - T_i^\star(x_i \mid z_1)|^2 \, \rho_i(\mathrm{d}x_i) \le \frac{\left\| \sum_{i \ge 2} |\partial_{1i} V|^2 \right\|_{L^\infty}}{\ell_V^2} |\tilde{z}_1 - z_1|^2 . \tag{55}$$

Proof. Since $x_i \mapsto T_i^\star(x_i \mid z_1)$ is increasing for any $z_1 \in \mathbb{R}$, by Lemma A.1 we know that

$$\begin{aligned}
\sum_{i=2}^{d} \int |T_i^\star(x_i \mid \tilde{z}_1) - T_i^\star(x_i \mid z_1)|^2 \, \rho_i(\mathrm{d}x_i) &= \sum_{i=2}^{d} W_2^2(q_i^\star(\cdot \mid \tilde{z}_1), \, q_i^\star(\cdot \mid z_1)) \\
&= W_2^2(q^\star(\cdot \mid \tilde{z}_1), \, q^\star(\cdot \mid z_1)) \\
&\le \frac{\left\| \sum_{i=2}^{d} |\partial_{1i} V|^2 \right\|_{L^\infty}}{\ell_V^2} |\tilde{z}_1 - z_1|^2 ,
\end{aligned}$$

where the second line follows from the fact that $q^\star(\cdot \mid \tilde{z}_1), \, q^\star(\cdot \mid z_1) \in \mathcal{P}(\mathbb{R})^{\otimes(d-1)}$. □

Lemma C.3. *Assume* (R), (P1), (P2), *and* (RD+) *hold. For all $z_1, \tilde{z}_1 \in \mathbb{R}$, $i \ge 2$, $x_i \in \mathbb{R}$,*

$$|T_i^\star(x_i \mid z_1) - T_i^\star(x_i \mid \tilde{z}_1)| \le \frac{\overline{L}}{\ell_V} |\tilde{z}_1 - z_1| , \tag{56}$$

i.e., $z_1 \mapsto T_i^\star(x_i \mid z_1)$ is $\frac{\overline{L}}{\ell_V}$-Lipschitz in z_1 uniformly for all $x_i \in \mathbb{R}$.

Proof. We drop the subscripts in $x_i, z_1, \tilde{z}_1$ for brevity. We first make the following key observation:

$$|T_i^\star(x \mid z) - T_i^\star(x \mid \tilde{z})| \le W_\infty(q_i^\star(\cdot \mid z), \, q_i^\star(\cdot \mid \tilde{z})) \qquad \text{for all } x \in \mathbb{R} . \tag{57}$$

Here W_∞ denotes the ∞-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathbb{R}^2)$ defined by $W_\infty(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \|d\|_{L^\infty(\pi)}$, where $d(x, y) = \|x - y\|$ and $L^\infty(\pi)$ denotes the L^∞ norm under π.

To see this, we claim that the above inequality holds for ρ_1-almost every $x \in \mathbb{R}$, then by the continuity of $T_i^\star(\cdot \mid z)$ and $T_i^\star(\cdot \mid \tilde{z})$, the inequality holds for all $x \in \mathbb{R}$.

We now validate the claim using contradiction. Let $\delta > 0$. Suppose the inequality fails over a set $E \subset \mathbb{R}$ with $\rho_1(E) > 0$, such that

$$|T_i^\star(x \mid z) - T_i^\star(x \mid \tilde{z})| \ge W_\infty(q_i^\star(\cdot \mid z), \, q_i^\star(\cdot \mid \tilde{z})) + \delta , \qquad x \in E .$$

Then, because $(T_i^\star(\cdot \mid z), T_i^\star(\cdot \mid \tilde{z}))_{\#}\rho_1$ is a W_p-optimal coupling of $q_i^\star(\cdot \mid z)$ and $q_i^\star(\cdot \mid \tilde{z})$,

$$\begin{aligned}
W_p(q_i^\star(\cdot \mid z), q_i^\star(\cdot \mid \tilde{z})) &\ge \left(\int_E |T_i^\star(x \mid z) - T_i^\star(x \mid \tilde{z})|^p \, \rho_1(\mathrm{d}x) \right)^{1/p} \\
&\ge \left(W_\infty(q_i^\star(\cdot \mid z), \, q_i^\star(\cdot \mid \tilde{z})) + \delta \right) \rho_1(E)^{1/p} ,
\end{aligned} \tag{58}$$

where W_p is the p-Wasserstein distance. Letting $p \to \infty$, as $W_p(q_i^\star(\cdot \mid z), \, q_i^\star(\cdot \mid \tilde{z})) \to W_\infty(q_i^\star(\cdot \mid z), \, q_i^\star(\cdot \mid \tilde{z}))$ when $p \to \infty$ (see, e.g., Givens and Shortt, 1984, Proposition 3), we arrive at the contradiction.

Therefore, combining (57), Theorem 2.5 in Khudiakova et al. (2025), and Lemma 3.4, we have

$$\begin{aligned}
|T_i^\star(x \mid z) - T_i^\star(x \mid \tilde{z})| &\le W_\infty(q_i^\star(\cdot \mid z),\, q_i^\star(\cdot \mid \tilde{z})) \\
&\le \frac{1}{\ell_V}\, I_\infty(q_i^\star(\cdot \mid z) \,\|\, q_i^\star(\cdot \mid \tilde{z})) \\
&= \frac{1}{\ell_V}\, \|(\log q_i^\star)'(\cdot \mid z) - (\log q_i^\star)'(\cdot \mid \tilde{z})\|_{L^\infty(q_i^\star(\cdot|z))} \\
&\le \frac{\overline{L}}{\ell_V}\, |z - \tilde{z}|\,,
\end{aligned}$$

where for $\nu \ll \mu$, the L^∞ relative Fisher information is given by

$$I_\infty(\nu \,\|\, \mu) \coloneqq \Big\| \nabla \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big) \Big\|_{L^\infty(\mu)}\,.$$

□

Remark C.4. Lemma C.3 implies that $z_1 \mapsto T_i^\star(x_i \mid z_1)$ is differentiable in z_1 almost everywhere for all $x_i \in \mathbb{R}$.

We will make use of the following standard facts throughout the next part of the proof; see Jiang et al. (2025) for proofs.

Lemma C.5. *Let T denote the optimal transport map from $\rho = \mathcal{N}(0,1)$ to μ, and let m denote the mean of μ. If $T' \le \beta$, then $|T(0) - m| \le \sqrt{2/\pi}\,\beta$.*

Lemma C.6. *Let m and $\tilde{m}$ denote the mean and the mode of μ, respectively, where μ is ℓ_V-log-concave and univariate. Then, $|m - \tilde{m}| \le 1/\sqrt{\ell_V}$.*

Lemma C.7. *Assume (R), (P1), (P2), and (RD+) hold. For all $i \ge 2$, $x_i, z_1 \in \mathbb{R}$,*

$$|\partial_{z_1}\partial_{x_i} T_i^\star(x_i \mid z_1)| \lesssim \frac{L_V \overline{L}}{\ell_V^2}\,(1 + |x_i|)\,. \tag{59}$$

Proof of Lemma C.7. For ease of notation, we fix i and drop the subscripts in z_1, x_i. Denote by $T_z(x) \coloneqq T_i^\star(x \mid z)$ and q_z the density of $q_i^\star(\cdot \mid z)$. Taking the log of the change of variables formula gives

$$\log \partial_x T_z(x) = -\log\sqrt{2\pi} - \frac{x^2}{2} - \log q_z(T_z(x))$$

for all $x \in \mathbb{R}$. Then, taking the derivative w.r.t. z yields that

$$\frac{\partial_z\partial_x T_z(x)}{\partial_x T_z(x)} = -\frac{\partial_x q_z(T_z(x))\,\partial_z T_z(x)}{q_z(T_z(x))} - \frac{\partial_z q_z(T_z(x))}{q_z(T_z(x))} \qquad \text{for all } x \in \mathbb{R}\,.$$

Rearranging to isolate $\partial_z\partial_x T_z(x)$ yields

$$\partial_z\partial_x T_z(x) = \partial_x T_z(x)\left(-\partial_x \log q_z(T_z(x))\,\partial_z T_z(x) - \partial_z \log q_z(T_z(x))\right) \tag{60}$$

for all $x \in \mathbb{R}$. Here, $\partial_x \log q_z(T_z(x))$ should be understood as the derivative of $\log q_z(\cdot)$ evaluated at the point $T_z(x)$. By ℓ_V-log-concavity of q_z and Caffarelli's contraction theorem, T_z is $\sqrt{1/\ell_V}$-Lipschitz, i.e., $0 \le \partial_x T_z \le \frac{1}{\sqrt{\ell_V}}$. On the other hand, by Lemma C.3, $|\partial_z T_z(x)| \le \frac{\overline{L}}{\ell_V}$.

Therefore,

$$\begin{aligned}
|\partial_z \partial_x T_z(x)| &= \left|\partial_x T_z(x) \left(-\partial_x \log q_z(T_z(x))\, \partial_z T_z(x) - \partial_z \log q_z(T_z(x))\right)\right| \\
&\le \frac{1}{\sqrt{\ell_V}} \left(|\partial_x \log q_z(T_z(x))|\, \frac{\overline{L}}{\ell_V} + |\partial_z \log q_z(T_z(x))|\right).
\end{aligned}$$

Denote by $\tilde{m}_z$ and m_z the mode and mean of q_z, respectively. Since q_z is L_V-log-smooth by Theorem 2.8, and T_z is smooth (in x), we can apply Lemma C.5 and Lemma C.6 and obtain

$$\begin{aligned}
|\partial_x \log q_z(T_z(x))| &\le \underbrace{|\partial_x \log q_z(\tilde{m}_z)|}_{=0} + L_V\, |T_z(x) - \tilde{m}_z| \\
&\le L_V \left(|T_z(x) - T_z(0)| + |T_z(0) - m_z| + |m_z - \tilde{m}_z|\right) \\
&\le L_V \left(\frac{1}{\sqrt{\ell_V}}\, |x| + \sqrt{\frac{2}{\pi}}\, \frac{1}{\sqrt{\ell_V}} + \frac{1}{\sqrt{\ell_V}}\right) \\
&\lesssim \frac{L_V}{\sqrt{\ell_V}} \left(1 + |x|\right).
\end{aligned}$$

We proceed with bounding $\partial_z \log q_z(T_z(x))$. Note that by the fundamental theorem of calculus and as $|\partial_z \partial_x \log q_z(x)| \le \overline{L}$, we have (again, using Lemma C.5 and Lemma C.6)

$$\begin{aligned}
|\partial_z \log q_z(T_z(x))| &\le |\partial_z \log q_z(\tilde{m}_z)| + \overline{L} \left(|T_z(x) - T_z(0)| + |T_z(0) - m_z| + |m_z - \tilde{m}_z|\right) \\
&\lesssim |\partial_z \log q_z(\tilde{m}_z)| + \frac{\overline{L}}{\sqrt{\ell_V}} \left(1 + |x|\right).
\end{aligned}$$

By Theorem 2.4, the derivative $\partial_z \log q_z(z_i) = \partial_z \log q_i^\star(z_i \mid z)$ satisfies

$$\begin{aligned}
-\partial_z \log q_z(z_i) = \partial_z \Big(\int V(z, z_i, z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z)\Big) \\
+ \partial_z \log \int \exp\Big(-\int V(z, z_i', z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z)\Big)\, \mathrm{d}z_i'.
\end{aligned} \tag{61}$$

Note that

$$\begin{aligned}
&\partial_z \log \int \exp\Big(-\int V(z, z_i', z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z)\Big)\, \mathrm{d}z_i' \\
&\quad = -\int \partial_z \Big(\int V(z, z_i', z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z)\Big)\; q_i^\star(\mathrm{d}z_i' \mid z).
\end{aligned}$$

Denoting

$$F(z, z_i) \coloneqq \partial_z \Big(\int V(z, z_i, z_{-\{1,i\}})\; q_{-i}^\star(\mathrm{d}z_{-\{1,i\}} \mid z)\Big),$$

we can rewrite (61) as

$$-\partial_z \log q_z(z_i) = F(z, z_i) - \int F(z, z_i')\, q_i^\star(\mathrm{d}z_i' \mid z)$$
$$= \int \Big(F(z, z_i) - F(z, z_i')\Big)\, q_i^\star(\mathrm{d}z_i' \mid z)\,.$$

Moreover, by Lemma 3.4, we have

$$|\partial_{z_i} F(z, z_i)| = |\partial_{z_i}\partial_z \log q_i^\star(z_i \mid z)| \le \overline{L}\,.$$

Therefore, the mean value theorem and Dalalyan et al. (2022, Lemma 2) yield that

$$|\partial_z \log q_z(\tilde{m}_z)| \le \overline{L} \int |\tilde{m}_z - z_i'|\, q_i^\star(\mathrm{d}z_i' \mid z) \le \frac{\overline{L}}{\sqrt{\ell_V}}\,.$$

Combining the above displays, we conclude:

$$|\partial_z \partial_x T_z(x)| \lesssim \frac{1}{\sqrt{\ell_V}} \left(\frac{L_V \overline{L}}{\ell_V \sqrt{\ell_V}} (1 + |x|) + \frac{\overline{L}}{\sqrt{\ell_V}} (1 + |x|) + \frac{\overline{L}}{\sqrt{\ell_V}} \right) \lesssim \frac{L_V \overline{L}}{\ell_V^2} (1 + |x|)\,.$$

□

We conclude this subsection with the proof of Theorem 3.5.

Proof of Theorem 3.5. The first part of Theorem 3.5 follows along the lines of the proof of Jiang et al. (2025, Theorem 5.4). Then, the bound on $\partial_{z_1} T_i^\star(x_i \mid \cdot)$ is shown in Lemma C.3, and the bound on $\partial_{z_1}\partial_{x_i} T_i^\star(x_i \mid \cdot)$ is shown in Lemma C.7. □

C.2 Dictionary of maps

To continue the development from Section 3.4, we now design a parameterized family of star-separable transport maps $\mathcal{T}_\Theta$ to approximate the SSVI minimizer. The construction is based on a finite dictionary of star-separable maps $\mathcal{M}$, and we consider a polyhedral subset of $\mathcal{T}_{\mathrm{star}}$ defined by the conic hull $\mathrm{cone}(\mathcal{M})$. Following Jiang et al. (2025), the set $\mathrm{cone}(\mathcal{M})$ is constructed as

$$\mathrm{cone}(\mathcal{M}) := \left\{ \sum_{T \in \mathcal{M}} \lambda_T T \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \right\},$$

where $\mathbb{R}_+^{|\mathcal{M}|}$ is the non-negative orthant in $\mathbb{R}^{|\mathcal{M}|}$. Clearly, when $\mathcal{M} \subseteq \mathcal{T}_{\mathrm{star}}$, $\mathrm{cone}(\mathcal{M})$ is a convex cone in $\mathcal{T}_{\mathrm{star}}$.

We now construct the dictionary $\mathcal{M}$ from one-dimensional piecewise linear maps. Define $\psi : \mathbb{R} \to \mathbb{R}$ by

$$\psi(x) := \begin{cases} 0\,, & x \le 0\,, \\ x\,, & x \in [0, 1]\,, \\ 1\,, & x \ge 1\,. \end{cases}$$

The function ψ is the basic building block of our dictionary. Let $R > 0$, partition the interval $[-R, R]$ into $N = 2R/\delta$ sub-intervals[‡‡] of length $\delta > 0$, and let the set of endpoints be

[‡‡] We assume that R is a multiple of δ and omit the explicit dependence of $\mathcal{M}$ on δ and R for brevity.

$\{b_1, \dots, b_{N+1}\}$ with $b_1 = -R$ and $b_{N+1} = R$. For convenience, we denote by $\mathcal{B} = \{b_1, \dots, b_N\}$ and set $b_0 \coloneqq b_1$, $b_{N+2} \coloneqq b_{N+1}$. The function classes are given by

$$
\begin{aligned}
\mathcal{M}_0 &\coloneqq \left\{ x \mapsto \psi\Big(\frac{x_1 - b}{\delta}\Big)\, e_1 \;\middle|\; b \in \mathcal{B} \right\}, \\
\mathcal{M}_1 &\coloneqq \left\{ x \mapsto \psi\Big(\frac{x_i - b}{\delta}\Big)\psi\Big(\frac{x_1 - b'}{\delta}\Big)\mathbb{1}_{\{x_1 \in [b', b'+\delta)\}}\, e_i \;\middle|\; i \geq 2,\, b, b' \in \mathcal{B} \right\}, \\
\mathcal{M}_2 &\coloneqq \left\{ x \mapsto \psi\Big(\frac{x_i - b}{\delta}\Big)\psi\Big(1 - \frac{x_1 - b'}{\delta}\Big)\mathbb{1}_{\{x_1 \in [b', b'+\delta)\}}\, e_i \;\middle|\; i \geq 2,\, b, b' \in \mathcal{B} \right\}, \\
\mathcal{M}_3 &\coloneqq \left\{ x \mapsto \psi\Big(\frac{x_i - b}{\delta}\Big)\mathbb{1}_{\{x_1 \geq R\}}\, e_i \;\middle|\; i \geq 2,\, b \in \mathcal{B} \right\}, \\
\mathcal{M}_4 &\coloneqq \left\{ x \mapsto \psi\Big(\frac{x_i - b}{\delta}\Big)\mathbb{1}_{\{x_1 < -R\}}\, e_i \;\middle|\; i \geq 2,\, b \in \mathcal{B} \right\}, \\
\mathcal{M}_5 &\coloneqq \left\{ x \mapsto \pm\psi\Big(\frac{x_1 - b}{\delta}\Big)\, e_i \;\middle|\; i \geq 2,\, b \in \mathcal{B} \right\}, \\
\mathcal{M} &\coloneqq \mathcal{M}_0 \cup \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_5 \,.
\end{aligned}
\tag{62}
$$

We note that $\mathcal{M}_0$ consists of non-decreasing piecewise linear functions of x_1. The maps in $\mathcal{M}_1$ are piecewise linear maps depending jointly on x_i and x_1, which are non-decreasing over the set $(x_1, x_i) \in [b', b'+\delta) \times \mathbb{R}$. $\mathcal{M}_2$ is built on $\mathcal{M}_1$ by flipping the sign of x_1. Each element of $\mathcal{M}_2$ is a a monotonically increasing function in x_i but monotonically decreasing in x_1 for $x_1 \in [b', b'+\delta)$. Finally, $\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$ are designed to approximate $T_i^\star$ when $x_1, x_i \notin [-R, R]$.

We further enrich $\mathcal{M}$ with constant coordinate shifts $\pm e_i$, $i \in [d]$ to handle tail behavior. Equivalently, we define the *augmented cone*:

$$
\underline{\mathrm{cone}}(\mathcal{M}) \coloneqq \left\{ \sum_{T \in \mathcal{M}} \lambda_T T + v \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|},\, v \in \mathbb{R}^d \right\}.
$$

Finally, by Theorem 3.5, for $z_1 \in \mathbb{R}$ and $i \in \{2, \dots, d\}$,

$$
\sqrt{1/L_V'} \leq |\partial_{x_1} T_1^\star(\cdot)| \leq \sqrt{1/\ell_V'}\,, \qquad \sqrt{1/L_V} \leq |\partial_{x_i} T_i^\star(\cdot \mid z_1)| \leq \sqrt{1/\ell_V}\,.
$$

Define $\boldsymbol{\alpha} \coloneqq \big((L_V')^{-1/2}, L_V^{-1/2}, \dots, L_V^{-1/2}\big)$. We arrive at the *$\boldsymbol{\alpha}$-augmented-and-spiked cone*:

$$
\underline{\mathrm{cone}}(\mathcal{M};\, \boldsymbol{\alpha}\,\mathrm{id}) \coloneqq \boldsymbol{\alpha}\,\mathrm{id} + \underline{\mathrm{cone}}(\mathcal{M})\,. \tag{63}
$$

Thus, $\underline{\mathrm{cone}}(\mathcal{M};\, \boldsymbol{\alpha}\,\mathrm{id}) \subset \mathcal{T}_{\mathrm{star}}$.

C.2.1 The approximation procedure

We now show that the dictionary in the previous subsection is such that there exists an approximator $\widehat{T} \in \underline{\mathrm{cone}}\,(\mathcal{M};\, \boldsymbol{\alpha}\,\mathrm{id})$ which approximates $T^\star$ to arbitrary precision (see Corollary C.12).

Recall that the optimal star-separable map has the structure

$$
T^\star(x) = [T_1^\star(x_1), T_2^\star(x_2 \mid T_1^\star(x_1)), \dots, T_d^\star(x_d \mid T_1^\star(x_1))] \qquad \text{for all} \qquad x \in \mathbb{R}^d\,.
$$

For clarity, we write $T_i^\star(x_i; x_1) \coloneqq T_i^\star(x_i \mid T_1^\star(x_1))$ to emphasize the dependence of $T_i^\star$ on x_1.

To approximate $T^\star$, we first approximate $T_1^\star$ by a univariate piecewise linear function $\widehat{T}_1$. For $i \geq 2$, we approximate $T_i^\star(x_i; x_1)$ by decoupling its dependence on x_i and x_1 and constructing approximators for both parts separately.

We restrict (x_i, x_1) to $[-R, R] \times [-R, R]$ for all $i \geq 2$.

To approximate the map $T_1^\star$, we define

$$\widehat{T}_1(x_1) \coloneqq T_1^\star(-R) + \sum_{m=1}^{N} \lambda_m \, \psi\Big(\frac{x_1 - b_m}{\delta}\Big), \tag{64}$$

where the non-negative coefficients $\lambda_m \coloneqq T_1^\star(b_{m+1}) - T_1^\star(b_m) \geq 0$ encode the local increments of $T_1^\star$ along the partition grid.

We next construct an approximation for $T_i^\star$ with $i \geq 2$. For each endpoint $b_j \in \mathcal{B} \cup \{b_{N+1}\}$, define the localized approximation as

$$\widehat{T}_i(x_i; b_j) \coloneqq T_i^\star(-R; b_j) + \sum_{m=1}^{N} \lambda_{m,j} \, \psi\Big(\frac{x_i - b_m}{\delta}\Big), \tag{65}$$

where $\lambda_{m,j} \coloneqq T_i^\star(b_{m+1}; b_j) - T_i^\star(b_m; b_j) \geq 0$.

For $x_1 \in I_j = [b_j, b_{j+1})$, $j \in [N]$, we interpolate between adjacent approximation maps:

$$\begin{aligned}
\widehat{T}_i(x_i; x_1) &\coloneqq \frac{b_{j+1} - x_1}{\delta} \, \widehat{T}_i(x_i; b_j) + \frac{x_1 - b_j}{\delta} \, \widehat{T}_i(x_i; b_{j+1}) \\
&= \psi\Big(1 - \frac{x_1 - b_j}{\delta}\Big) \, \widehat{T}_i(x_i; b_j) + \psi\Big(\frac{x_1 - b_j}{\delta}\Big) \, \widehat{T}_i(x_i; b_{j+1}) \\
&= \widehat{T}_i(x_i; b_j) + \psi\Big(\frac{x_1 - b_j}{\delta}\Big) \Big(\widehat{T}_i(x_i; b_{j+1}) - \widehat{T}_i(x_i; b_j)\Big).
\end{aligned} \tag{66}$$

When $x_i \in I_j$, $j \in \{0, N+1\}$, we define

$$\widehat{T}_i(x_i; x_1) \coloneqq \widehat{T}_i(x_i; b_j).$$

Combining all the above, we conclude the construction of the approximator $\widehat{T}_i$:

$$\begin{aligned}
\widehat{T}_i(x_i; x_1) &= \widehat{T}_i(x_i; -R) \mathbb{1}_{\{x_1 < -R\}} + \widehat{T}_i(x_i; R) \mathbb{1}_{\{x_1 \geq R\}} \\
&\quad + \sum_{j=1}^{N} \left(\psi\Big(1 - \frac{x_1 - b_j}{\delta}\Big) \, \widehat{T}_i(x_i; b_j) + \psi\Big(\frac{x_1 - b_j}{\delta}\Big) \, \widehat{T}_i(x_i; b_{j+1}) \right) \mathbb{1}_{\{x_1 \in [b_j, b_{j+1})\}} \\
&= T_i^\star(-R, -R) + \sum_{j=1}^{N} \psi\Big(\frac{x_1 - b_j}{\delta}\Big) \Big(T_i^\star(-R; b_{j+1}) - T_i^\star(-R; b_j)\Big) \\
&\quad + \sum_{m=1}^{N} \Bigg\{ \lambda_{m,0} \, \psi\Big(\frac{x_i - b_m}{\delta}\Big) \mathbb{1}_{\{x_1 < -R\}} + \lambda_{m,N+1} \, \psi\Big(\frac{x_i - b_m}{\delta}\Big) \mathbb{1}_{\{x_1 \geq R\}} \\
&\qquad + \lambda_{m,j} \, \psi\Big(\frac{x_i - b_m}{\delta}\Big) \psi\Big(1 - \frac{x_1 - b_j}{\delta}\Big) \mathbb{1}_{\{x_1 \in [b_j, b_{j+1})\}} \\
&\qquad + \lambda_{m,j+1} \, \psi\Big(\frac{x_i - b_m}{\delta}\Big) \psi\Big(\frac{x_1 - b_j}{\delta}\Big) \mathbb{1}_{\{x_1 \in [b_j, b_{j+1})\}} \Bigg\}.
\end{aligned}$$

We remark that the proposed piecewise linear interpolation yields a continuous approximation in x_1 while preserving monotonicity in x_i. Moreover, it is straightforward to verify that $\widehat{T} \in \underline{\mathrm{cone}}(\mathcal{M})$.

Lemma C.8. *Assume* (R), (P1), (P2), *and* (RD+) *hold. For $i \geq 2$, the approximator $\widehat{T}_i$ defined under* (65)-(66) *satisfies the following:*

(i) When $x_1 \in I_0 \cup I_{N+1}$, $\partial_{x_1} \widehat{T}_i(x_i; x_1) = 0$.

(ii) When $x_i \in I_0 \cup I_{N+1}$, $\partial_{x_i} \widehat{T}_i(x_i; x_1) = 0$.

(iii) For all $x_i, x_1 \in \mathbb{R}$,

$$|\partial_{x_i} \widehat{T}_i(x_i; x_1)| \leq 1/\sqrt{\ell_V} \quad \text{and} \quad |\partial_{x_1} \widehat{T}_i(x_i; x_1)| \leq \frac{\overline{L}}{\ell_V \sqrt{\ell'_V}}\,.$$

Furthermore,

$$|\partial_{x_i} \partial_{x_1} \widehat{T}_i(x_i; x_1)| \leq \left(\frac{L_V}{\ell_V^{3/2}} + \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell'_V}} \right) (1 + |x_i| + \delta)\,.$$

Proof. (i) and (ii) hold trivially by construction. We now show that (iii) holds. Let $\eta_{k,j} \coloneqq T_i^\star(b_k; b_{j+1}) - T_i^\star(b_k; b_j)$. Note the following identity:

$$\lambda_{k,j} + \eta_{k+1,j} = T_i^\star(b_{k+1}; b_{j+1}) - T_i^\star(b_k; b_j) = \lambda_{k,j+1} + \eta_{k,j}\,.$$

Let $\Delta_{j,k} \coloneqq \lambda_{k,j+1} - \lambda_{k,j} = \eta_{k+1,j} - \eta_{k,j}$. For all $x_1 \in I_j$, $x_i \in I_k$, we first compute the partial gradients of $\widehat{T}_i$ w.r.t. x_1, x_i explicitly. Write $u_k(x_i) = \psi\left(\frac{x_i - b_k}{\delta}\right)$ and $w_j(x_1) = \psi\left(\frac{x_1 - b_j}{\delta}\right)$, then

$$\begin{aligned}
\partial_{x_1} \widehat{T}_i(x_i; x_1) &= \frac{\widehat{T}_i(x_i; b_{j+1}) - \widehat{T}_i(x_i; b_j)}{\delta} \\
&= (1 - u_k(x_i)) \, \frac{T_i^\star(b_k; b_{j+1}) - T_i^\star(b_k; b_j)}{\delta} \\
&\quad + u_k(x_i) \, \frac{T_i^\star(b_{k+1}; b_{j+1}) - T_i^\star(b_{k+1}; b_j)}{\delta}
\end{aligned}$$

and

$$\begin{aligned}
\partial_{x_i} \widehat{T}_i(x_i; x_1) &= (1 - w_j(x_1)) \, \partial_{x_i} \widehat{T}_i(x_i; b_j) + w_j(x_1) \, \partial_{x_i} \widehat{T}_i(x_i; b_{j+1}) \\
&= (1 - w_j(x_1)) \, \frac{T_i^\star(b_{k+1}; b_j) - T_i^\star(b_k; b_j)}{\delta} \\
&\quad + w_j(x_1) \, \frac{T_i^\star(b_{k+1}; b_{j+1}) - T_i^\star(b_k; b_{j+1})}{\delta}\,.
\end{aligned}$$

Using the definitions of $\lambda_{k,j}$ and $\eta_{k,j}$, we obtain the forms

$$\partial_{x_1} \widehat{T}_i(x_i; x_1) = \frac{\eta_{k,j}}{\delta} + u_k(x_i) \, \frac{\Delta_{j,k}}{\delta} \quad \text{and} \quad \partial_{x_i} \widehat{T}_i(x_i; x_1) = \frac{\lambda_{k,j}}{\delta} + w_j(x_1) \, \frac{\Delta_{j,k}}{\delta}\,. \tag{67}$$

By Theorem 3.5 and the mean value theorem, we derive that

$$0 \le \frac{\lambda_{k,j}}{\delta} \le \frac{1}{\sqrt{\ell_V}} \tag{68}$$

and for some $\eta_j \in [b_j, b_{j+1}]$,

$$\left|\frac{\eta_{k,j}}{\delta}\right| = \frac{1}{\delta}\left|T_i^\star(b_k; b_{j+1}) - T_i^\star(b_k; b_j)\right| \le \left|\partial_2 T_i^\star(b_k \mid T_1^\star(\eta_j))\right| |\partial_{x_1} T_1^\star(\eta_j)| \le \frac{\overline{L}}{\ell_V \sqrt{\ell_V'}}\,. \tag{69}$$

Similarly, for some $\xi_k, \chi_k \in I_k$, β_k between ξ_k, χ_k, and $\gamma_j \in [b_j, b_{j+1}]$,

$$\begin{aligned}
\left|\frac{\Delta_{j,k}}{\delta}\right| &= \left|\frac{\lambda_{k,j+1} - \lambda_{k,j}}{\delta}\right| = \left|\frac{T_i^\star(b_{k+1}; b_{j+1}) - T_i^\star(b_k; b_{j+1}) - (T_i^\star(b_{k+1}; b_j) - T_i^\star(b_k; b_j))}{\delta}\right| \\
&= |\partial_{x_i} T_i^\star(\xi_k; b_{j+1}) - \partial_{x_i} T_i^\star(\chi_k; b_j)| \\
&\le |\partial_{x_i} T_i^\star(\xi_k; b_{j+1}) - \partial_{x_i} T_i^\star(\chi_k; b_{j+1})| + |\partial_{x_i} T_i^\star(\chi_k; b_{j+1}) - \partial_{x_i} T_i^\star(\chi_k; b_j)| \\
&\le |\partial_{x_i}^2 T_i^\star(\beta_k; b_{j+1})|\,\delta + |\partial_{x_i}\partial_2 T_i^\star(\chi_k \mid T_1^\star(\gamma_j))|\,|\partial_{x_1} T_1^\star(\gamma_j)|\,\delta \\
&\le \frac{L_V}{\ell_V^{3/2}}\,(1 + |\beta_k|)\,\delta + \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\,(1 + |\chi_k|)\,\delta \\
&\le \frac{L_V}{\ell_V^{3/2}}\,(1 + |x_i| + \delta)\,\delta + \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\,(1 + |x_i| + \delta)\,\delta\,.
\end{aligned}$$

Thus,

$$\left|\frac{\Delta_{j,k}}{\delta^2}\right| \le \left(\frac{L_V}{\ell_V^{3/2}} + \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\right)(1 + |x_i| + \delta)\,. \tag{70}$$

By (67)–(70), we conclude that for all $x_i, x_1 \in \mathbb{R}$,

$$0 \le \partial_{x_i} \widehat{T}_i(x_i; x_1) \le \frac{1}{\delta}\max\{\lambda_{k,j}, \lambda_{k,j+1}\} \le \frac{1}{\sqrt{\ell_V}}\,,$$

$$|\partial_{x_1} \widehat{T}_i(x_i; x_1)| \le \frac{1}{\delta}\max\{|\eta_{k,j}|, |\eta_{k+1,j}|\} \le \frac{\overline{L}}{\ell_V \sqrt{\ell_V'}}\,,$$

and

$$|\partial_{x_i}\partial_{x_1} \widehat{T}_i(x_i; x_1)| = \left|\frac{\Delta_{j,k}}{\delta^2}\right| \le \left(\frac{L_V}{\ell_V^{3/2}} + \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\right)(1 + |x_i| + \delta)\,.$$

This concludes the proof. □

The next result states that $T_i^\star$ can be well approximated by $\widehat{T}_i$ when the leaf variable is fixed.

Lemma C.9. *Let* (**R**), (**P1**), *and* (**RD+**) *hold. The estimate* $\widehat{T}_1$ *of* $T_1^\star$ *defined in* (64) *satisfies*

$$\|\widehat{T}_1 - T_1^\star\|_{L^2(\rho)}^2 \lesssim \frac{1}{\ell_V'}\exp(-R^2/2) + \frac{(L_V')^2}{(\ell_V')^3}\,\delta^4\,,$$

and

$$\|(\widehat{T}_1 - T_1^\star)'\|_{L^2(\rho)}^2 \lesssim \frac{1}{\ell_V' R}\exp(-R^2/2) + \frac{(L_V')^2}{(\ell_V')^3}\,\delta^2\,.$$

If (**P2**) *also holds, then for any* $b \in \mathcal{B} \cup \{b_{N+1}\}$*, the estimate* $\widehat{T}_i^1(\cdot; b)$ *defined in* (64) *satisfies*

$$\|\widehat{T}_i(\cdot; b) - T_i^\star(\cdot; b)\|_{L^2(\rho)}^2 \lesssim \frac{1}{\ell_V}\exp(-R^2/2) + \frac{L_V^2}{\ell_V^3}\,\delta^4\,,$$

and

$$\|\partial_{x_i}\widehat{T}_i(\cdot; b) - \partial_{x_i}T_i^\star(\cdot; b)\|_{L^2(\rho)}^2 \lesssim \frac{1}{\ell_V R}\exp(-R^2/2) + \frac{L_V^2}{\ell_V^3}\delta^2\,.$$

Proof. It suffices to show the first pair of bounds; the second pair follows by the same argument. Following the proof of Theorem 5.6 in Jiang et al. (2025), the pointwise bounds hold for $|x_1| \geq R$:

$$|\widehat{T}_1(x_1) - T_1^\star(x_1)| \leq \frac{1}{\sqrt{\ell_V'}}\,(|x_1| - R)\,, \qquad |(\widehat{T}_1 - T_1^\star)'(x_1)| \leq \frac{1}{\sqrt{\ell_V'}}\,.$$

Combining these and integrating by parts yields

$$\int_{\mathbb{R}\setminus(-R,R)} |\widehat{T}_1 - T_1^\star|^2\,\mathrm{d}\rho_1 \lesssim \frac{1}{\ell_V'}\int_{\mathbb{R}\setminus(-R,R)} (|x_1| - R)^2\rho_1(\mathrm{d}x_1) \lesssim \frac{1}{\ell_V'}\,e^{-R^2/2}\,.$$

To bound the derivative term, we use the Mills ratio bound (see, e.g., Wainwright, 2019, Exercise 2.2) to get

$$\int_{\mathbb{R}\setminus(-R,R)} |(\widehat{T}_1 - T_1^\star)'|^2\,\mathrm{d}\rho_1 \lesssim \frac{1}{\ell_V'}\int_{\mathbb{R}\setminus(-R,R)} \rho_1(\mathrm{d}x_1) \lesssim \frac{1}{\ell_V' R}\,e^{-R^2/2}\,.$$

We now control the differences on the compact interval $[-R, R]$. By the definition of $\widehat{T}_1$ and the mean value theorem, for any $x_1 \in I_j$ with $j \in [N]$, there exist $\zeta_j, \eta_j, \chi_j \in [b_j, b_{j+1}]$ such that

$$\begin{aligned}
|\widehat{T}_1(x_1) - T_1^\star(x_1)|^2 &= \left|\frac{T_1^\star(b_{j+1}) - T_1^\star(b_j)}{\delta}\,(x_1 - b_j) + T_1^\star(b_j) - T_1^\star(x_1)\right|^2 \\
&= \left|(T_1^\star)'(\zeta_j)\,(x_1 - b_j) - (T_1^\star)'(\eta_j)\,(x_1 - b_j)\right|^2 \\
&= \left|(T_1^\star)''(\chi_j)\,(x_1 - b_j)\,(\zeta_j - \eta_j)\right|^2 \\
&\leq \delta^4\left|(T_1^\star)''(\chi_j)\right|^2.
\end{aligned}$$

Using the second derivative bound on $T_1^\star$ in Theorem 3.5,

$$\begin{aligned}
\sum_{j\in[N]}\int_{I_j} |\widehat{T}_1 - T_1^\star|^2\,\mathrm{d}\rho_1 &\leq \delta^4\sum_{j\in[N]}\int_{I_j}(1 + |\chi_j|)^2\,\frac{(L_V')^2}{(\ell_V')^3}\,\rho_1(\mathrm{d}x_1) \\
&\leq \delta^4\sum_{j\in[N]}\int_{I_j}(1 + |x_1| + \delta)^2\,\frac{(L_V')^2}{(\ell_V')^3}\,\rho_1(\mathrm{d}x_1) \\
&\lesssim \frac{(L_V')^2}{(\ell_V')^3}\,\delta^4\int(1 + |x_1|^2 + \delta^2)\,\rho_1(\mathrm{d}x_1) \lesssim \frac{(L_V')^2}{(\ell_V')^3}\,\delta^4\,.
\end{aligned}$$

Similarly, for the derivative difference, for some $\zeta_j, \beta_j \in [b_j, b_{j+1}]$ (depending on x_1),

$$\begin{aligned}
|(\widehat{T}_1 - T_1^\star)'(x_1)|^2 &= \left| \frac{T_1^\star(b_{j+1}) - T_1^\star(b_j)}{\delta} - (T_1^\star)'(x_1) \right|^2 \\
&= \left| (T_1^\star)'(\zeta_j) - (T_1^\star)'(x_1) \right|^2 \\
&= \left| (T_1^\star)''(\beta_j)\,(\zeta_j - x_1) \right|^2 \le \delta^2 \left| (T_1^\star)''(\beta_j) \right|^2 .
\end{aligned}$$

Applying the same second derivative bound,

$$\begin{aligned}
\sum_{j\in[N]} \int_{I_j} |(\widehat{T}_1 - T_1^\star)'|^2 \,\mathrm{d}\rho_1 &\le \delta^2 \sum_{j\in[N]} \int_{I_j} (1 + |\beta_j|)^2 \, \frac{(L_V')^2}{(\ell_V')^3} \, \rho_1(\mathrm{d}x_1) \\
&\le \delta^2 \sum_{j\in[N]} \int_{I_j} (1 + |x_1| + \delta)^2 \, \frac{(L_V')^2}{(\ell_V')^3} \, \rho_1(\mathrm{d}x_1) \\
&\lesssim \frac{(L_V')^2}{(\ell_V')^3} \, \delta^2 .
\end{aligned}$$

□

We now prove our main approximation result.

Lemma C.10. *Let* (**R**), (**P1**), (**P2**), *and* (**RD+**) *hold. Then the approximator* $\widehat{T}$ *defined in* (64)–(66) *satisfies*

$$\|\widehat{T} - T^\star\|_{L^2(\rho)}^2 \lesssim \left(\frac{1}{\ell_V'} + \frac{d}{\ell_V} + \frac{L_V'}{\ell_V \ell_V'} \right) \exp(-R^2/2) + \frac{d\overline{L}^2}{\ell_V^2 \ell_V'} \, \delta^2 + \left(\frac{(L_V')^2}{(\ell_V')^3} + \frac{dL_V^2}{\ell_V^3} \right) \delta^4 .$$

Assume further that (**GR**) *holds. Then,*

$$\begin{aligned}
\|D(\widehat{T} - T^\star)\|_{L^2(\rho)}^2 \lesssim{}& \left(\frac{1}{\ell_V'} + \frac{d}{\ell_V} + \frac{d\overline{L}^2}{\ell_V^2 \ell_V'} \right) \frac{\exp(-R^2/2)}{R} \\
&+ \left[\left(\frac{1}{d} + \frac{\overline{L}^2}{\ell_V^2} \right) \frac{(L_V')^2}{(\ell_V')^3} + \frac{L_V^2}{\ell_V^3} + \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + \frac{(1 + M_{2\gamma}(\rho_1))\, L_{\mathsf{GR}}^2}{(\ell_V')^2} \right] d\delta^2 .
\end{aligned}$$

Remark C.11. (a) The error bound for $\|\widehat{T} - T^\star\|_{L^2(\rho)}^2$ arises from three sources: (i) the error in approximating $T_1^\star$; (ii) the error in approximating the first argument of $T_i^\star(x_i; x_1)$; and (iii) the error in approximating the second argument of $T_i^\star(x_i; x_1)$.

(b) The error bound for $\|D(\widehat{T} - T^\star)\|_{L^2(\rho)}^2$ consists of: (i) the approximation error for $DT_1^\star$; (ii) the error in approximating the derivative $\partial_{x_i} T_i^\star(x_i; x_1)$ with respect to the first argument; and (iii) the error in approximating the derivative with respect to the root variable, i.e., $|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)|^2$.

(c) Under the setting of Theorem 2.10, the star-separable map $T^\star = (T_1^\star, T_2^\star, \ldots, T_d^\star)$ from $\mathcal{N}(0, I)$ to $\pi^\star = \mathcal{N}(m^\star, \Sigma^\star)$ has the explicit form

$$\begin{aligned}
T_1^\star(x_1) &= m_1 + \sqrt{\sigma_{11}}\, x_1 , \\
T_i^\star(x_i; x_1) &= m_i + \sigma_{i1}\, (\sigma_{11})^{-1/2}\, (T_1^\star(x_1) - m_1) + \sqrt{(\sigma^{ii})^{-1}}\, x_i , \qquad i \ge 2 .
\end{aligned}$$

Therefore, (**GR**) is satisfied, and the approximation is exact up to the truncation error.

Given Lemma C.10, we can establish the following result, which shows the *existence* of a map $\widehat{T}$ in our construction which satisfies the desired approximation guarantee.

Corollary C.12. *Assume* (**R**), (**P1**), (**P2**), (**RD+**), *and* (**GR**) *hold. Then, for any* $\epsilon > 0$, *there exists a closed convex set* Θ *of dimension* $O((d^2/\epsilon^2)\log(d/\epsilon^2))$ *and an affine parametrization* $\theta \mapsto T_\theta$, *where the implied constant depends polynomially on the constants* $\ell_V, L_V, \ell'_V, L'_V, \overline{L}, L_{\mathsf{GR}}, M_{2\gamma}(\rho_1)$ *and their inverses, where* $M_{2\gamma}(\rho_1)$ *is the* 2γ*-th absolute moment of* ρ_1, *such that there exists* $\hat{\theta} \in \Theta$ *satisfying*

$$\|T_{\hat{\theta}} - T^\star\|_{L_2(\rho)} \le \epsilon \qquad \textit{and} \qquad \|D(T_{\hat{\theta}} - T^\star)\|_{L^2(\rho)} \le \epsilon\,.$$

Remark C.13. (a) The proof of Lemma C.10 and Corollary C.12 can be adapted without changes to produce a map $\widehat{T} \in \underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$ that approximates the optimal star-separable map $T^\star$ and its gradient arbitrarily well in $L^2(\rho)$. Indeed, we can approximate $T^\star - \boldsymbol{\alpha}\,\mathrm{id}$ by some $\widehat{T}_0 \in \underline{\mathrm{cone}}(\mathcal{M})$ and set $\widehat{T} \coloneqq \widehat{T}_0 + \boldsymbol{\alpha}\,\mathrm{id}$.

(b) Notice that the error rates for approximating $T^\star$ and $DT^\star$ are both of order ϵ, which differs from the error rates in Jiang et al. (2025) for estimating univariate optimal transport maps. In Jiang et al. (2025), the functional value approximation bound scales as ϵ^2, while the derivative approximation scales as ϵ. The dominant error here comes from approximating each $T_i^\star$ in its second argument, i.e., $|T_i^\star(x_i; x_1) - T_i^\star(x_i; b_j)|$, which scales as $\delta \asymp \epsilon$.

Proof. Given R and δ, the size of $\mathcal{M}_0$ is $O(R/\delta)$, the sizes of $\mathcal{M}_1$, $\mathcal{M}_2$ are each at most $O((d-1)\,(R/\delta)^2)$, while the sizes of $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_5$ are each $O((d-1)\,(R/\delta))$. Therefore, the size of $\mathcal{M}$ is $|\mathcal{M}_0| + |\mathcal{M}_1| + |\mathcal{M}_2| + |\mathcal{M}_3| + |\mathcal{M}_4| + |\mathcal{M}_5| = O(d\,(R/\delta)^2)$. In what follows, we suppress polynomial dependence on L_V, ℓ_V, etc. in the asymptotic notation $\lesssim$, $\asymp$, keeping only the dependence on the dimension d and the accuracy ϵ.

Choose $R \asymp \sqrt{\log(d/\epsilon^2)}$ and $\delta \asymp \epsilon/\sqrt{d}$. Then, from the first bound of Lemma C.10, it is straightforward to see that $\|\widehat{T} - T^\star\|_{L^2(\rho)} \le \epsilon$. Moreover, from the second bound on Lemma C.10, these choices yield $\|D(\widehat{T} - T^\star)\|_{L^2(\rho)} \le \epsilon$ as well, which we will need for later reference. □

Finally, we introduce the following control on the second-order growth of $T_i^\star$ w.r.t. x_1, which serves as the final ingredient to establish Lemma C.10.

Lemma C.14. *Assume* (**R**), (**P1**), (**P2**), (**RD+**), *and* (**GR**) *hold. Then, for all* $i \ge 2$ *and* $x_i, x_1 \in \mathbb{R}$,

$$\partial_{x_1}^2 T_i^\star(x_i; x_1) \;\lesssim\; \frac{L_{\mathsf{GR}}}{\ell'_V}\,(1 + |x_i|)^\gamma + \frac{L'_V\,\overline{L}}{(\ell'_V)^{3/2}\,\ell_V}\,(1 + |x_1|)\,.$$

Proof. By the chain rule,

$$\partial_{x_1}^2 T_i^\star(x_i; x_1) = \partial_2^2 T_i^\star(x_i \mid T_1^\star(x_1))\,\big(\partial_{x_1} T_1^\star(x_1)\big)^2 + \partial_2 T_i^\star(x_i \mid T_1^\star(x_1))\,\partial_{x_1}^2 T_1^\star(x_1)\,.$$

By Theorem 3.5, the derivatives satisfy

$$\big|\partial_{x_1} T_1^\star(x_1)\big|^2 \le \frac{1}{\ell'_V}\,, \quad \big|\partial_2 T_i^\star(x_i \mid T_1^\star(x_1))\big| \le \frac{\overline{L}}{\ell_V}\,, \quad \big|\partial_{x_1}^2 T_1^\star(x_1)\big| \lesssim \frac{L'_V}{(\ell'_V)^{3/2}}\,(1 + |x_1|)\,.$$

In addition, (**GR**) implies that

$$\left|\partial_2^2 T_i^\star(x_i \mid T_1^\star(x_1))\right| \lesssim L_{\mathsf{GR}}\,(1+|x_i|)^\gamma\,.$$

Combining these bounds with Theorem 3.5 and applying the triangle inequality yields

$$\left|\partial_{x_1}^2 T_i^\star(x_i;x_1)\right| \;\lesssim\; \frac{L_{\mathsf{GR}}}{\ell_V'}\,(1+|x_i|)^\gamma + \frac{L_V'\,\overline{L}}{(\ell_V')^{3/2}\,\ell_V}\,(1+|x_1|)\,.$$

□

Now, we are ready to prove Lemma C.10.

Proof of Lemma C.10. **Approximation error of $\widehat{T}-T^\star$.** First, we bound the approximation error $\|\widehat{T}-T^\star\|_{L^2(\rho)}$. For every $i\in[d]$ and $j,k\in\{0,\dots,N+1\}$, define

$$S_{j,k}^{(i)} \coloneqq \{x\in\mathbb{R}^d : x_1\in I_j,\, x_i\in I_k\}\,,$$

where $I_j=[b_j,b_{j+1})$, $I_k=[b_k,b_{k+1})$ are the subintervals.

We adopt the shorthand: $\sum_j \coloneqq \sum_{j=0}^{N+1}$, $\sum_k \coloneqq \sum_{k=0}^{N+1}$, $\sum_l \coloneqq \sum_{l=0}^{N+1}$. Write ∂_1,∂_2 for differentiation with respect to the first and second arguments, respectively; e.g., $\partial_1 T_i^\star(x_i \mid z_1)=\partial_{x_i}T_i^\star(x_i\mid z_1)$ and $\partial_2 T_i^\star(x_i\mid z_1)=\partial_{z_1}T_i^\star(x_i\mid z_1)$. Recall $T_i^\star(x_i;x_1)\coloneqq T_i^\star(x_i\mid T_1^\star(x_1))$ and $\rho=\mathcal{N}(0,I)=\bigotimes_{i=1}^d\rho_i$ with $\rho_1=\cdots=\rho_d=\mathcal{N}(0,1)$.

We expand

$$\|\widehat{T}-T^\star\|_{L^2(\rho)}^2 = \int\left|\widehat{T}_1(x_1)-T_1^\star(x_1)\right|^2\rho_1(\mathrm{d}x_1) + \sum_{i=2}^d\int\left|\widehat{T}_i(x_i;x_1)-T_i^\star(x_i;x_1)\right|^2\rho(\mathrm{d}x)\,.$$

By the definition of $\widehat{T}_i$, we know that for $x_1\in I_j$,

$$\widehat{T}_i(x_i;x_1) = (1-w_j(x_1))\,\widehat{T}_i(x_i;b_j) + w_j(x_1)\,\widehat{T}_i(x_i;b_{j+1})\,,$$

where we recall that $w_j(x_1)=\psi\big(\frac{x_1-b_j}{\delta}\big)\in[0,1]$ and the convention that $b_0=b_1=-R$ and $b_{N+1}=b_{N+2}=R$. In particular, we observe that when $j\in\{0,N+1\}$, then $\widehat{T}_i(x_i;x_1)=\widehat{T}_i(x_i;b_j)$.

Applying the triangle inequality followed by Jensen's inequality yields

$$
\begin{aligned}
&\sum_{j,k} \int_{S_{j,k}^{(i)}} \left|\widehat{T}_i(x_i; x_1) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&\quad = \sum_{j,k} \int_{S_{j,k}^{(i)}} \left|(1 - w_j(x_1))\, \widehat{T}_i(x_i; b_j) + w_j(x_1)\, \widehat{T}_i(x_i; b_{j+1}) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&\quad \lesssim \sum_{j,k} \int_{S_{j,k}^{(i)}} w_j(x_1) \left|\widehat{T}_i(x_i; b_j) - T_i^\star(x_i; b_j)\right|^2 \rho(\mathrm{d}x) \\
&\qquad + \sum_{j,k} \int_{S_{j,k}^{(i)}} (1 - w_j(x_1)) \left|\widehat{T}_i(x_i; b_{j+1}) - T_i^\star(x_i; b_{j+1})\right|^2 \rho(\mathrm{d}x) \\
&\qquad + \sum_{j,k} \int_{S_{j,k}^{(i)}} w_j(x_1) \left|T_i^\star(x_i; b_j) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&\qquad + \sum_{j,k} \int_{S_{j,k}^{(i)}} (1 - w_j(x_1)) \left|T_i^\star(x_i; b_{j+1}) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&\quad \lesssim \sum_{j,k} \int_{S_{j,k}^{(i)}} \left|\widehat{T}_i(x_i; b_j) - T_i^\star(x_i; b_j)\right|^2 \rho(\mathrm{d}x) + \sum_{j,k} \int_{S_{j,k}^{(i)}} \left|T_i^\star(x_i; b_j) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x)\,.
\end{aligned}
$$

Combining the bounds above gives

$$
\begin{aligned}
\int \left|\widehat{T}_i(x_i; x_1) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \lesssim \underbrace{\sum_{j,k} \int_{S_{j,k}^{(i)}} \left|\widehat{T}_i(x_i; b_j) - T_i^\star(x_i; b_j)\right|^2 \rho(\mathrm{d}x)}_{C_i} \\
+ \underbrace{\sum_{j,k} \int_{S_{j,k}^{(i)}} \left|T_i^\star(x_i; b_j) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x)}_{D_i}\,.
\end{aligned}
$$

We bound C_i, D_i separately. For C_i, by Fubini's theorem and Lemma C.9,

$$
\begin{aligned}
C_i &= \sum_j \int_{\{x_1 \in I_j\}} \left|\widehat{T}_i(x_i; b_j) - T_i^\star(x_i; b_j)\right|^2 \rho(\mathrm{d}x) \\
&= \sum_j \rho_1(I_j) \int \left|\widehat{T}_i(x_i; b_j) - T_i^\star(x_i; b_j)\right|^2 \rho_i(\mathrm{d}x_i) \\
&\lesssim \left(\frac{1}{\ell_V} \exp(-R^2/2) + \frac{L_V^2}{\ell_V^3}\, \delta^4\right) \sum_j \rho_1(I_j) \\
&= \frac{1}{\ell_V} \exp(-R^2/2) + \frac{L_V^2}{\ell_V^3}\, \delta^4\,.
\end{aligned} \tag{71}
$$

For D_i, using Fubini's theorem and then Lemma C.2, we derive that

$$
\begin{aligned}
\sum_{i=2}^d D_i &= \sum_{i=2}^d \sum_j \int_{\{x_1 \in I_j\}} \left|T_i^\star(x_i; b_j) - T_i^\star(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&= \sum_j \int_{\{x_1 \in I_j\}} \left(\sum_{i=2}^d \int \left|T_i^\star(x_i \mid T_1^\star(b_j)) - T_i^\star(x_i \mid T_1^\star(x_1))\right|^2 \rho_i(\mathrm{d}x_i)\right) \rho_1(\mathrm{d}x_1) \\
&\le \frac{\left\|\sum_{i \ge 2} |\partial_{1i} V|^2\right\|_{L^\infty}}{\ell_V^2} \sum_j \int_{I_j} \left|T_1^\star(b_j) - T_1^\star(x_1)\right|^2 \rho_1(\mathrm{d}x_1)\,.
\end{aligned}
$$

Furthermore, since $T_1^\star$ is $(1/\sqrt{\ell'_V})$-Lipschitz by Theorem 3.5 (i), we conclude that

$$\sum_{i=2}^d D_i \le \frac{\left\|\sum_{i\ge 2} |\partial_{1i} V|^2\right\|_{L^\infty}}{\ell_V^2 \ell'_V} \sum_j \int_{I_j} |b_j - x_1|^2 \, \rho_1(\mathrm{d}x_1)\,. \tag{72}$$

Since the intervals I_j have length δ when $j \in [N]$,

$$\begin{aligned}\sum_j \int_{I_j} |b_j - x_1|^2 \, \rho_1(\mathrm{d}x_1) &\le \delta^2 + \int_{-\infty}^{-R} |x_1 + R|^2 \, \rho_1(\mathrm{d}x_1) + \int_R^\infty |x_1 - R|^2 \, \rho_1(\mathrm{d}x_1) \\ &\lesssim \delta^2 + \exp(-R^2/2)\,.\end{aligned}$$

Thus, (72) becomes

$$\sum_{i=2}^d D_i \lesssim \frac{\sum_{i\ge 2} \||\partial_{1i} V|^2\|_{L^\infty}}{\ell_V^2 \ell'_V} \left(\delta^2 + \exp(-R^2/2)\right) \lesssim \left(\frac{L'_V}{\ell_V \ell'_V} \wedge \frac{d\overline{L}^2}{\ell_V^2 \ell'_V}\right) \left(\delta^2 + \exp(-R^2/2)\right), \tag{73}$$

where the last step uses (**RD**) since $\frac{1}{2} L'_V - \frac{\left\|\sum_{i\ge 2} |\partial_{1i} V|^2\right\|_{L^\infty}}{\ell_V} \ge 0$, and $\overline{L} \ge \max_{i\ge 2} \|\partial_{1i} V\|_{L^\infty}$.

Putting together (71) and (73), and the bound on $\|\widehat{T}_1 - T_1^\star\|^2_{L^2(\rho)}$ from Lemma C.9, we obtain

$$\begin{aligned}\|\widehat{T} - T^\star\|^2_{L^2(\rho)} &\lesssim \int \left|\widehat{T}_1(x_1) - T_1^\star(x_1)\right|^2 \rho_1(\mathrm{d}x_1) + \sum_{i=2}^d C_i + \sum_{i=2}^d D_i \\ &\lesssim \frac{1}{\ell'_V} \exp(-R^2/2) + \frac{(L'_V)^2}{(\ell'_V)^3} \delta^4 + \left(\frac{L'_V}{\ell_V \ell'_V} \wedge \frac{d\overline{L}^2}{\ell_V^2 \ell'_V}\right) \left(\delta^2 + \exp(-R^2/2)\right) \\ &\quad + (d-1) \left(\frac{1}{\ell_V} \exp(-R^2/2) + \frac{L_V^2}{\ell_V^3} \delta^4\right).\end{aligned}$$

The right-hand side in the above display can be rearranged into

$$\|\widehat{T} - T^\star\|^2_{L^2(\rho)} \lesssim \left(\frac{1}{\ell'_V} + \frac{d}{\ell_V} + \frac{L'_V}{\ell_V \ell'_V}\right) \exp(-R^2/2) + \frac{d\overline{L}^2}{\ell_V^2 \ell'_V} \delta^2 + \left(\frac{(L'_V)^2}{(\ell'_V)^3} + \frac{dL_V^2}{\ell_V^3}\right) \delta^4,$$

as desired.

Approximation error of the gradients $D(\widehat{T} - T^\star)$. We decompose

$$\begin{aligned}\|D(\widehat{T} - T^\star)\|^2_{L^2(\rho)} = \int \left|\partial_{x_1} T_1^\star(x_1) - \partial_{x_1} \widehat{T}_1(x_1)\right|^2 \rho_1(\mathrm{d}x_1) \\ + \sum_{i=2}^d \underbrace{\int \left|\partial_{x_i} T_i^\star(x_i; x_1) - \partial_{x_i} \widehat{T}_i(x_i; x_1)\right|^2 \rho(\mathrm{d}x)}_{=: A_i} \\ + \sum_{i=2}^d \underbrace{\int \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)\right|^2 \rho(\mathrm{d}x)}_{=: B_i}.\end{aligned}$$

Bounding A_i. By (67), we derive for $x_1 \in I_j$,

$$\partial_{x_i}\widehat{T}_i(x_i;x_1) = (1-w_j(x_1))\,\partial_{x_i}\widehat{T}_i(x_i;b_j) + w_j(x_1)\,\partial_{x_i}\widehat{T}_i(x_i;b_{j+1})\,.$$

For each $i \geq 2$, applying the triangle inequality implies

$$\begin{aligned}
&\left|\partial_{x_i}T_i^\star(x_i;x_1) - \partial_{x_i}\widehat{T}_i(x_i;x_1)\right|^2 \\
&\quad\lesssim \left|\partial_{x_i}T_i^\star(x_i;x_1) - (1-w_j(x_1))\,\partial_{x_i}T_i^\star(x_i;b_j) - w_j(x_1)\,\partial_{x_i}T_i^\star(x_i;b_{j+1})\right|^2 \\
&\qquad + \left|(1-w_j(x_1))\left(\partial_{x_i}T_i^\star(x_i;b_j) - \partial_{x_i}\widehat{T}_i(x_i;b_j)\right)\right|^2 \\
&\qquad + \left|w_j(x_1)\left(\partial_{x_i}T_i^\star(x_i;b_{j+1}) - \partial_{x_i}\widehat{T}_i(x_i;b_{j+1})\right)\right|^2 \\
&\quad\lesssim \underbrace{\left|\partial_{x_i}T_i^\star(x_i;x_1) - \partial_{x_i}T_i^\star(x_i;b_j)\right|^2 + \left|\partial_{x_i}T_i^\star(x_i;x_1) - \partial_{x_i}T_i^\star(x_i;b_{j+1})\right|^2}_{a_{ij}^\star} \\
&\qquad + \underbrace{\left|\partial_{x_i}T_i^\star(x_i;b_j) - \partial_{x_i}\widehat{T}_i(x_i;b_j)\right|^2 + \left|\partial_{x_i}T_i^\star(x_i;b_{j+1}) - \partial_{x_i}\widehat{T}_i(x_i;b_{j+1})\right|^2}_{\hat{a}_{ij}}\,.
\end{aligned}$$

Therefore,

$$A_i \lesssim \sum_j \underbrace{\int_{\{x_1\in I_j\}} a_{ij}^\star\,\rho(\mathrm{d}x)}_{A_{ij}^\star} + \underbrace{\sum_j \int_{\{x_1\in I_j\}} \hat{a}_{ij}\,\rho(\mathrm{d}x)}_{\hat{A}_i}\,. \tag{74}$$

Thus, we can control A_i by the error $A_{ij}^\star$ and the approximation error $\hat{A}_i$.

By the chain rule and Theorem 3.5 (ii),

$$\left|\partial_{x_1}\partial_{x_i}T_i^\star(x_i;x_1)\right| = \left|\partial_{x_i}\partial_2 T_i^\star(x_i \mid T_1^\star(x_1))\right|\left|\partial_{x_1}T_1^\star(x_1)\right| \lesssim \frac{L_V\overline{L}}{\ell_V^2\sqrt{\ell_V'}}\,(1+|x_i|)\,, \tag{75}$$

where $\partial_2 T_i^\star(x_i \mid z_1) \coloneqq \partial_{z_1}T_i^\star(x_i \mid z_1)$. Therefore, using the mean value theorem, we know that for $x_1 \in I_j$, $j \in [N]$,

$$|\partial_{x_i}T_i^\star(x_i;x_1) - \partial_{x_i}T_i^\star(x_i;b_j)| \lesssim \frac{L_V\overline{L}}{\ell_V^2\sqrt{\ell_V'}}\,(1+|x_i|)\,|x_1-b_j| \leq \frac{L_V\overline{L}}{\ell_V^2\sqrt{\ell_V'}}\,(1+|x_i|)\,\delta\,.$$

The above bounds holds when we replace b_j with b_{j+1}. Therefore, Fubini's theorem gives

$$\begin{aligned}
A_{ij}^\star &\lesssim \left(\frac{L_V\overline{L}}{\ell_V^2\sqrt{\ell_V'}}\right)^2 \int_{\{x_1\in I_j\}} (1+|x_i|)^2\,\delta^2\,\rho(\mathrm{d}x) \\
&= \frac{L_V^2\overline{L}^2}{\ell_V^4\ell_V'}\int (1+|x_i|)^2\,\rho_i(\mathrm{d}x_i)\,\delta^2\rho_1(I_j) \lesssim \frac{L_V^2\overline{L}^2}{\ell_V^4\ell_V'}\,\delta^2\rho_1(I_j)\,.
\end{aligned}$$

On the other hand, Theorem 3.5 (i) and the Mills ratio bound imply that for $j \in \{0, N+1\}$,

$$\begin{aligned}
\int_{\{x_1\in I_j\}} \left|\partial_{x_i}T_i^\star(x_i;x_1) - \partial_{x_i}T_i^\star(x_i;b_j)\right|^2 \rho(\mathrm{d}x) &\leq \int_{\{x_1\in I_j\}} \left(\frac{1}{\sqrt{\ell_V}}\right)^2 \rho(\mathrm{d}x) \\
&\lesssim \frac{\rho_1([R,\infty))}{\ell_V} \lesssim \frac{\exp(-R^2/2)}{\ell_V R}\,.
\end{aligned}$$

Combining the previous two bounds yields that

$$\sum_j A^\star_{ij} \lesssim \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} \delta^2 \sum_{j=1}^N \rho_1(I_j) + \frac{\exp\left(-R^2/2\right)}{\ell_V R} \lesssim \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} \delta^2 + \frac{\exp\left(-R^2/2\right)}{\ell_V R} \,. \tag{76}$$

For $\hat{A}_i$, observe that

$$\begin{aligned}
\hat{A}_i &\lesssim \int \max_j \left|\partial_{x_i} T_i^\star(x_i; b_j) - \partial_{x_i} \widehat{T}_i(x_i; b_j)\right|^2 \rho_i(\mathrm{d}x_i) \\
&= \underbrace{\int_{-R}^R \max_j \left|\partial_{x_i} T_i^\star(x_i; b_j) - \partial_{x_i} \widehat{T}_i(x_i; b_j)\right|^2 \rho_i(\mathrm{d}x_i)}_{=:\hat{A}_{i1}} \\
&\quad + \underbrace{\int_{\mathbb{R}\setminus[-R,R]} \max_j \left|\partial_{x_i} T_i^\star(x_i; b_j) - \partial_{x_i} \widehat{T}_i(x_i; b_j)\right|^2 \rho_i(\mathrm{d}x_i)}_{=:\hat{A}_{i2}} \,.
\end{aligned}$$

From Lemma C.8, $|\partial_{x_i} \widehat{T}_i(\cdot \mid z_1)| \le 1/\sqrt{\ell_V}$. Using the Mills ratio and by Theorem 3.5 (i), we have

$$\hat{A}_{i2} \lesssim \frac{\rho\left([R,\infty)\right)}{\ell_V} \lesssim \frac{\exp\left(-R^2/2\right)}{\ell_V R} \,. \tag{77}$$

As Theorem 3.5 (i) shows $|\partial_{x_i}^2 T_i^\star(x_i; x_1)| \lesssim \frac{L_V}{\ell_V^{3/2}} (1 + |x_i|)$ for $x_i \in I_k$, $k \in [N]$, we have

$$\left|\partial_{x_i} T_i^\star(x_i \mid b_j) - \partial_{x_i} \widehat{T}_i(x_i \mid b_j)\right| \le \sup_{\tilde{x}_i \in I_k} \left|\partial_{x_i}^2 T_i^\star(\tilde{x}_i; b_j)\right| \delta \lesssim \frac{L_V}{\ell_V^{3/2}} (1 + |x_i| + \delta)\, \delta \,.$$

Hence,

$$\begin{aligned}
\hat{A}_{i1} &= \int_{-R}^R \max_j \left|\partial_{x_i} T_i^\star(x_i; b_j) - \partial_{x_i} \widehat{T}_i^1(x_i; b_j)\right|^2 \rho_i(\mathrm{d}x_i) \\
&\le \frac{L_V^2}{\ell_V^3} \delta^2 \int_{-R}^R (1 + |x_i| + \delta)^2 \, \rho_i(\mathrm{d}x_i) \lesssim \frac{L_V^2}{\ell_V^3} \delta^2 \,.
\end{aligned} \tag{78}$$

Combining (76), (77), and (78), we see that

$$A_i \lesssim \hat{A}_{i1} + \hat{A}_{i2} \lesssim \left(\frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + \frac{L_V^2}{\ell_V^3}\right) \delta^2 + \frac{\exp\left(-R^2/2\right)}{\ell_V R} \,. \tag{79}$$

Bounding B_i. We now bound B_i:

$$\begin{aligned}
B_i &= \int \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)\right|^2 \rho(\mathrm{d}x) \\
&= \underbrace{\sum_{j \in \{0, N+1\}} \sum_k \int_{S_{j,k}^{(i)}} \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)\right|^2 \rho(\mathrm{d}x)}_{=:B_{i1}} \\
&\quad + \underbrace{\sum_{j \in [N]} \sum_k \int_{S_{j,k}^{(i)}} \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)\right|^2 \rho(\mathrm{d}x)}_{=:B_{i2}} \,.
\end{aligned}$$

For $j \in \{0, N+1\}$, recall that $|\partial_{x_1} T_i^\star(x_i; x_1)| \le \frac{\overline{L}}{\ell_V \sqrt{\ell_V'}}$. In addition, Lemma C.8 shows $|\partial_{x_1} \widehat{T}_i(x_i; x_1)| \le \frac{\overline{L}}{\ell_V \sqrt{\ell_V'}}$. By the Mills ratio bound,

$$B_{i1} \lesssim \frac{\overline{L}^2}{\ell_V^2 \ell_V'} \int_{\mathbb{R}\setminus[-R,R]} \rho_1(\mathrm{d}x_1) \lesssim \frac{\overline{L}^2}{\ell_V^2 \ell_V' R} \exp(-R^2/2)\,. \tag{80}$$

Fix $j \in [N]$ and $k \in \{0, \dots, N+1\}$. By the mean value theorem, there exists $\zeta_j \in I_j$ such that $\partial_{x_1} T_i^\star(x_i; \zeta_j) = \delta^{-1}\left(T_i^\star(x_i; b_{j+1}) - T_i^\star(x_i; b_j)\right)$. Applying the mean value theorem again together with (67) with $\eta_{k,j} = T_i^\star(b_k; b_{j+1}) - T_i^\star(b_k; b_j)$, $\Delta_{j,k} = \eta_{k+1,j} - \eta_{k,j}$, we see that for some $\zeta_j \in I_j$,

$$\begin{aligned}
&\left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)\right| \\
&\quad \le \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} T_i^\star(b_k; \zeta_j) - (x_i - b_k)\,\frac{\Delta_{j,k}}{\delta^2}\right| \\
&\quad = \left|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} T_i^\star(b_k; x_1) + \partial_{x_1} T_i^\star(b_k; x_1) - \partial_{x_1} T_i^\star(b_k; \zeta_j) - (x_i - b_k)\,\frac{\Delta_{j,k}}{\delta^2}\right| \\
&\quad = \left|\left(\partial_{x_1}\partial_{x_i} T_i^\star(\beta_k; x_1) - \tfrac{\Delta_{j,k}}{\delta^2}\right)(x_i - b_k) + \partial_{x_1} T_i^\star(b_k; x_1) - \partial_{x_1} T_i^\star(b_k; \zeta_j)\right|
\end{aligned}$$

for some β_k between x_i and b_k. By Theorem 3.5 (ii),

$$\begin{aligned}
|\partial_{x_1}\partial_{x_i} T_i^\star(\zeta'; x_1)| &\le |\partial_2 \partial_{x_i} T_i^\star(\zeta' \mid T_1^\star(x_1))|\,|\partial_{x_1} T_1^\star(x_1)| \\
&\lesssim \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\,(1 + |\zeta'|) \le \frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\,(1 + |x_1| + \delta)\,.
\end{aligned}$$

In addition, Lemma C.14 implies

$$|\partial_{x_1} T_i^\star(b_k; x_1) - \partial_{x_1} T_i^\star(b_k; \zeta_j)| \lesssim \frac{L_{\mathsf{GR}}}{\ell_V'}\,(1 + |x_1| + \delta)^\gamma\,\delta + \frac{L_V'\,\overline{L}}{(\ell_V')^{3/2}\,\ell_V}\,(1 + |x_1| + \delta)\,\delta\,.$$

Combining the above bounds and applying the triangle inequality together with Lemma C.8, we have

$$\begin{aligned}
|\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)| \lesssim \Bigg[&\frac{L_V \overline{L}}{\ell_V^2 \sqrt{\ell_V'}}\,(1 + |x_i| + \delta) \\
&+ \frac{L_V'\,\overline{L}}{(\ell_V')^{3/2}\,\ell_V}\,(1 + |x_1| + \delta) + \frac{L_{\mathsf{GR}}}{\ell_V'}\,(1 + |x_1| + \delta)^\gamma\Bigg]\,\delta\,.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
B_{i2} &= \sum_{j,k=1}^{N} \int_{S_{j,k}^{(i)}} |\partial_{x_1} T_i^\star(x_i; x_1) - \partial_{x_1} \widehat{T}_i(x_i; x_1)|^2 \, \rho(\mathrm{d}x) \\
&\lesssim \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} \sum_{j,k=1}^{N} \int_{S_{j,k}^{(i)}} (1 + |x_i| + \delta)^2 \, \delta^2 \, \rho(\mathrm{d}x) \\
&\quad + \frac{(L_V')^2 \overline{L}^2}{(\ell_V')^3 \ell_V^2} \sum_{j,k=1}^{N} \int_{S_{j,k}^{(i)}} (1 + |x_1| + \delta)^2 \, \delta^2 \, \rho(\mathrm{d}x) \\
&\quad + \frac{L_{\mathsf{GR}}^2}{(\ell_V')^2} \sum_{j,k=1}^{N} \int_{S_{j,k}^{(i)}} (1 + |x_1| + \delta)^{2\gamma} \, \delta^2 \, \rho(\mathrm{d}x) \\
&\lesssim \delta^2 \left[\frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + \frac{(L_V')^2 \overline{L}^2}{(\ell_V')^3 \ell_V^2} + \frac{L_{\mathsf{GR}}^2 \, (1 + M_{2\gamma}(\rho_1))}{(\ell_V')^2} \right].
\end{aligned} \tag{81}
$$

Hence, combining (80) and (81) yields

$$
B_i \lesssim \frac{\overline{L}^2}{\ell_V^2 \ell_V' R} \exp(-R^2/2) + \delta^2 \left[\frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + \frac{(L_V')^2 \overline{L}^2}{(\ell_V')^3 \ell_V^2} + \frac{(1 + M_{2\gamma}(\rho_1)) \, L_{\mathsf{GR}}^2}{(\ell_V')^2} \right]. \tag{82}
$$

Finally, using Lemma C.9 together with (79) and (82),

$$
\begin{aligned}
\|D(\widehat{T} - T^\star)\|_{L^2(\rho)}^2 &\lesssim \left(\frac{1}{\ell_V'} + \frac{d-1}{\ell_V} + \frac{d\overline{L}^2}{\ell_V^2 \ell_V'} \right) \frac{\exp(-R^2/2)}{R} + \frac{(L_V')^2}{(\ell_V')^3} \, \delta^2 \\
&\quad + \left[\frac{L_V^2}{\ell_V^3} + \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + \frac{(L_V')^2 \overline{L}^2}{(\ell_V')^3 \ell_V^2} + \frac{(1 + M_{2\gamma}(\rho_1)) \, L_{\mathsf{GR}}^2}{(\ell_V')^2} \right] (d-1) \, \delta^2 \\
&\lesssim \left(\frac{1}{\ell_V'} + \frac{d}{\ell_V} + \frac{d\overline{L}^2}{\ell_V^2 \ell_V'} \right) \frac{\exp(-R^2/2)}{R} \\
&\quad + \left[\left(\frac{1}{d} + \frac{\overline{L}^2}{\ell_V^2} \right) \frac{(L_V')^2}{(\ell_V')^3} + \frac{L_V^2}{\ell_V^3} + \frac{L_V^2 \overline{L}^2}{\ell_V^4 \ell_V'} + + \frac{(1 + M_{2\gamma}(\rho_1)) \, L_{\mathsf{GR}}^2}{(\ell_V')^2} \right] d\delta^2 .
\end{aligned}
$$

□

C.3 The computational guarantees

In previous section, we explicitly construct a map $\widehat{T} \in \underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$ that approximate the optimal star-separable map $T^\star$ and its gradient arbitrarily well in $L^2(\rho)$. However, this construction requires knowledge of $T^\star$. In this section, we show that a good approximation to $T^\star$ can by computed by solving a convex optimization problem.

Given the dictionary $\mathcal{M} = \mathcal{M}_0 \cup \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_5$ defined in (62), we search for the minimizer $T_{\mathcal{M}}^\star$ of the KL divergence with respect to π over $\underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$, i.e.,

$$
T_{\mathcal{M}}^\star = \operatorname*{arg\,min}_{T \in \underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})} \mathsf{KL}\left(T_{\#}\rho \parallel \pi\right). \tag{SSVI-opt}
$$

The above optimization problem admits a unique minimizer. The computational guarantees in Theorem C.15 are shown without invoking the explicit form of the maps in $\underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$, as we choose to work with (SSVI-opt).

The proof of Theorem C.15 involves two steps. First, by Corollary C.12, we choose an "oracle" approximator $\widehat{T} \in \underline{\text{cone}}\,(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$ that approximates $T^\star$. Second, we show that $T^\star_{\mathcal{M}}$ is close to the approximating map $\widehat{T}$ in $L^2(\rho)$ using the geometry of the problem (SSVI-opt). The result then follows by combining the bounds from the two steps.

Theorem C.15. *Let* (**R**), (**P1**), (**P2**), (**RD+**), *and* (**GR**) *be satisfied. Then, for each* $\epsilon > 0$, *there exists a dictionary* $\mathcal{M}$ *of size* $O((d^2/\epsilon^2)\log(d/\epsilon^2))$ *such that*

$$\|T^\star_{\mathcal{M}} - T^\star\|_{L^2(\rho)} \lesssim \epsilon,$$

where $T^\star_{\mathcal{M}}$ *is the minimizer to* (SSVI-opt) *and the constants depend polynomially on* ℓ_V, L_V, ℓ'_V, L'_V, $\overline{L}$, L_{GR}, $M_{2\gamma}(\rho_1)$ *and their inverses.*

In the sequel, we will work with the following measures: $\pi^\star_{\mathcal{M}} \coloneqq (T^\star_{\mathcal{M}})_{\#}\rho$, $\hat{\pi} \coloneqq (\widehat{T})_{\#}\rho$. Define the finite-dimensional set

$$\mathcal{P}_{\mathcal{M}} \coloneqq \{T_{\#}\rho : T \in \underline{\text{cone}}\,(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})\} \subset \mathcal{T}_{\text{star}}\,.$$

The following result shows that the objective value of $\pi^\star_{\mathcal{M}}$ is close to that of $\hat{\pi}$, which is used to control the distance between $T^\star_{\mathcal{M}}$ and $\widehat{T}$ in Theorem C.15.

Lemma C.16. *Under the setting of Theorem C.15, there exists* $\kappa > 0$ *depending polynomially on* $\ell_V, \ell'_V, L_V, L'_V$ *such that*

$$\begin{aligned} 0 &\le \mathsf{KL}\,(\hat{\pi} \parallel \pi) - \mathsf{KL}\,(\pi^\star_{\mathcal{M}} \parallel \pi) \\ &\le \frac{1}{2}\left((L_V \vee L'_V)\,\|\widehat{T} - T^\star\|^2_{L^2(\rho)} + \kappa^2\,\|D(\widehat{T} - T^\star)\|^2_{L^2(\rho)}\right). \end{aligned}$$

Proof. The first inequality is straightforward since $\hat{\pi} \in \mathcal{P}_{\mathcal{M}}$. To show the second inequality, we use the fact that $\mathcal{P}_{\mathcal{M}} \subset \mathcal{C}_{\text{star}}$ to obtain

$$\mathsf{KL}\,(\hat{\pi} \parallel \pi) - \mathsf{KL}\,(\pi^\star_{\mathcal{M}} \parallel \pi) \le \mathsf{KL}\,(\hat{\pi} \parallel \pi) - \mathsf{KL}\,(\pi^\star \parallel \pi)\ .$$

To conclude, it suffices to bound the right-hand side above.

Recall that $T^\star$ (resp. $\widehat{T}$) denotes the star-separable map that pushes forward ρ to $\pi^\star$ (resp. $\hat{\pi}$). By definition, $T^\star, \widehat{T} \in \mathcal{T}_{\text{star}}$. Define $T_t \coloneqq (1-t)T^\star + t\widehat{T}$ and $\mu_t \coloneqq T_{t\#}\rho$, then $\mu_0 = \pi^\star$ and $\mu_1 = \hat{\pi}$. We note that $\mu_t \in \mathcal{C}_{\text{star}}$ as $\mathcal{T}_{\text{star}}$ is convex. Recall (51) and (52), with $T_0 = T^\star$ and $T_1 = \widehat{T}$.

By Lemma 2.6, $\nabla^2 V(T_t) \preceq [L_V \vee (L'_V/2)]\,I$, and thus the second derivative of the potential energy is bounded by

$$\frac{\mathrm{d}^2\mathcal{V}(\mu_t)}{\mathrm{d}t^2} \le (L_V \vee (L'_V/2))\,\|\widehat{T} - T^\star\|^2_{L^2(\rho)}\,. \tag{83}$$

By Theorem 3.5 and Lemma C.8, we apply Lemma C.17 with $L_1 = DT^\star$, $L_2 = D\widehat{T}$, $c_1 = \sqrt{1/L_V} \wedge \sqrt{1/L'_V} = \sqrt{1/(L_V \vee L'_V)}$, and $c_3 = \frac{\overline{L}}{\ell_V\sqrt{\ell'_V}}$ to derive that

$$\|(DT_t)^{-1}\|_2 \le \frac{c_1 + \sqrt{d-1}\,c_3}{c_1^2} \le \frac{1}{c_1} + \frac{L'_V}{2\sqrt{\ell'_V}\,c_1^2} =: \kappa\,,$$

where the last inequality follows from $\frac{L'_V}{2\sqrt{\ell'_V}} \geq \frac{(d-1)\overline{L}}{\ell_V\sqrt{\ell'_V}} = (d-1)\,c_3$ by (**RD+**) and $d \geq 2$. Therefore, by (52),

$$\frac{\mathrm{d}^2\mathcal{H}(\mu_t)}{\mathrm{d}t^2} \leq \kappa^2 \left\|D(\widehat{T} - T^\star)\right\|^2_{L^2(\rho)}. \tag{84}$$

Thus, combining (83) and (84) yields

$$\frac{\mathrm{d}^2\mathsf{KL}\left(\mu_t \parallel \pi\right)}{\mathrm{d}t^2} \leq \left(L_V \vee (L'_V/2)\right)\left\|\widehat{T} - T^\star\right\|^2_{L^2(\rho)} + \kappa^2\left\|D(\widehat{T} - T^\star)\right\|^2_{L^2(\rho)}. \tag{85}$$

Moreover, as

$$\begin{aligned}\mathsf{KL}\left(\hat{\pi} \parallel \pi\right) - \mathsf{KL}\left(\pi^\star \parallel \pi\right) = \left\langle[\nabla_W \mathsf{KL}\left(\cdot \parallel \pi\right)](\pi^\star),\, T - \mathrm{id}\right\rangle_{L^2(\pi^\star)} \\ + \int_0^1 \int_0^t \frac{\mathrm{d}^2}{\mathrm{d}s^2}\,\mathsf{KL}\left(\mu_s \parallel \pi\right)\mathrm{d}s\,\mathrm{d}t\,,\end{aligned}$$

the optimality of $\pi^\star$ and the derivative bound (85) imply that

$$\mathsf{KL}\left(\hat{\pi} \parallel \pi\right) - \mathsf{KL}\left(\pi^\star \parallel \pi\right) \leq \frac{1}{2}\left(\left(L_V \vee L'_V\right)\left\|\widehat{T} - T^\star\right\|^2_{L^2(\rho)} + \kappa^2\left\|D(\widehat{T} - T^\star)\right\|^2_{L^2(\rho)}\right).$$

□

Proof of Theorem C.15. Let $\widehat{T}$ be the approximation of $T^\star$ in $\underline{\mathrm{cone}}(\mathcal{M};\, \boldsymbol{\alpha}\,\mathrm{id})$, constructed in Corollary C.12 with $|\mathcal{M}| = O(\dim(\Theta)) = O\left((d^2/\epsilon^2)\log(d/\epsilon^2)\right)$. Then $T^\star_{\mathcal{M}}, \widehat{T} \in \mathcal{T}_{\mathrm{star}}$. By the triangle inequality,

$$\left\|T^\star_{\mathcal{M}} - T^\star\right\|_{L^2(\rho)} \leq \left\|T^\star_{\mathcal{M}} - \widehat{T}\right\|_{L^2(\rho)} + \left\|\widehat{T} - T^\star\right\|_{L^2(\rho)}.$$

The second term is controlled by Corollary C.12, recalling that

$$\left\|\widehat{T} - T^\star\right\|_{L^2(\rho)} \leq \epsilon\,. \tag{86}$$

It remains to bound the first term. By Lemma 2.6, π is $(\ell'_V \wedge \ell_V)$-log-concave. Then, by Theorem 3.2, the map $T \mapsto \mathrm{KL}(T_\#\rho \parallel \pi)$ is $(\ell'_V \wedge \ell_V)$-convex over the convex set $\underline{\mathrm{cone}}(\mathcal{M};\, \boldsymbol{\alpha}\,\mathrm{id}) \subset \mathcal{T}_{\mathrm{star}}$. Together with the optimality of $\pi^\star_{\mathcal{M}}$, this yields

$$\frac{\ell'_V \wedge \ell_V}{2}\left\|T^\star_{\mathcal{M}} - \widehat{T}\right\|^2_{L^2(\rho)} \leq \mathsf{KL}(\hat{\pi} \parallel \pi) - \mathsf{KL}(\pi^\star_{\mathcal{M}} \parallel \pi)\,. \tag{87}$$

In addition, by Lemma C.16, there exists a constant $\kappa > 0$ depending polynomially on $\ell_V, \ell'_V, L_V, L'_V$ such that

$$\mathsf{KL}\left(\hat{\pi} \parallel \pi\right) - \mathsf{KL}\left(\pi^\star \parallel \pi\right) \leq \frac{1}{2}\left(\left(L'_V \vee L_V\right)\left\|\widehat{T} - T^\star\right\|^2_{L^2(\rho)} + \kappa^2\left\|D(\widehat{T} - T^\star)\right\|^2_{L^2(\rho)}\right).$$

Together with (87), this gives

$$\left\|T^\star_{\mathcal{M}} - \widehat{T}\right\|^2_{L^2(\rho)} \leq \frac{L'_V \vee L_V}{\ell'_V \wedge \ell_V}\left\|\widehat{T} - T^\star\right\|^2_{L^2(\rho)} + \frac{\kappa^2}{\ell'_V \wedge \ell_V}\left\|D(\widehat{T} - T^\star)\right\|^2_{L^2(\rho)}. \tag{88}$$

The first term is already controlled in (86). On the other hand, by Corollary C.12,

$$\left\|D(\widehat{T} - T^\star)\right\|_{L^2(\rho)} \leq \epsilon\,.$$

The above bound and (86) imply that $\left\|T^\star_{\mathcal{M}} - T^\star\right\|_{L^2(\rho)} \lesssim \epsilon$. □

We conclude this section with a technical lemma used in establishing Lemma C.16.

Lemma C.17. *For $i = 0, 1$, let $L_i \coloneqq D_i + u_i e_1^\top$, where D_i is a diagonal matrix and $u_i = (0, u_{i2}, \dots, u_{id})^\top \in \mathbb{R}^d$. For $t \in [0,1]$, define $L_t \coloneqq (1-t)L_0 + tL_1$. Suppose further that $c_1 I \preceq D_0, D_1 \preceq c_2 I$ and $\max\{\|u_0\|_\infty, \|u_1\|_\infty\} \le c_3$ for some constants $c_1, c_2, c_3 > 0$. Then,*

$$\|L_t\|_2 \le c_2 + \sqrt{d-1}\, c_3 \qquad \text{and} \qquad \|L_t^{-1}\|_2 \le \frac{c_1 + \sqrt{d-1}\, c_3}{c_1^2}\,.$$

Proof. Denote $D_t \coloneqq (1-t)D_0 + tD_1$ and $u_t \coloneqq (1-t)u_0 + tu_1$. Then $L_t = D_t + u_t e_1^\top$ is invertible. Therefore,

$$\|L_t\|_2 \le \|D_t\|_2 + \|u_t e_1^\top\|_2 \le c_2 + \sqrt{d-1}\, c_3\,.$$

On the other hand, by the Sherman–Morrison formula (Sherman and Morrison, 1950),

$$L_t^{-1} = D_t^{-1} - \frac{D_t^{-1} u_t e_1^\top D_t^{-1}}{1 + e_1^\top D_t^{-1} u_t} = D_t^{-1} - D_t^{-1} u_t e_1^\top D_t^{-1}\,,$$

since $e_1^\top D_t^{-1} u_t = 0$ (the first coordinate of u_t is 0). Hence,

$$\|L_t^{-1}\|_2 \le \|D_t^{-1}\|_2 + \|D_t^{-1} u_t\|\, \|e_1^\top D_t^{-1}\| \le \frac{1}{c_1} + \frac{\sqrt{d-1}\, c_3}{c_1^2}\,,$$

because $D_t \succeq c_1 I$, $\|u_t\| \le \sqrt{d-1}\, c_3$, and $\|e_1^\top D_t^{-1}\| = \|D_t^{-1} e_1\| \le 1/c_1$. □

C.4 Algorithm details

This subsection provides the details for computing the minimizer $T^\star_{\mathcal{M}}$ of the problem (SSVI-opt) using the projected gradient descent algorithm. Recall that

$$T^\star_{\mathcal{M}} = \operatorname*{arg\,min}_{T \in \underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})} \mathsf{KL}\left(T_\# \rho \,\|\, \pi\right). \tag{89}$$

where $\underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$ denotes the augmented spiked cone defined in (63) with $\alpha_1 = (L'_V)^{-1/2}$ and $\alpha_2 = \cdots = \alpha_d = L_V^{-1/2}$.

To resolve this, we show that (SSVI-opt) can be instead reformulated as a *finite-dimensional*, *strongly convex* optimization problem. To this end, we introduce the following set: let $\mathcal{M}'_5$ be the set of functions in $\mathcal{M}_5$ with positive signs, that is

$$\mathcal{M}'_5 \coloneqq \left\{ x \mapsto \psi\Big(\frac{x_1 - b}{\delta}\Big)\, e_i \;\Big|\; i \ge 2,\, b \in \mathcal{B} \right\}. \tag{90}$$

Define $\mathcal{M}' \coloneqq \mathcal{M}_0 \cup \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_4 \cup \mathcal{M}'_5$. We require that $\int T(x)\, \rho(\mathrm{d}x) = 0$ for all $T \in \mathcal{M}'$. We can achieve this by centering each function and absorbing the constant shift in the vector v. Moreover, by construction, all functions in $\mathcal{M}'$ are linearly independent in the pointwise sense: no element of $\mathcal{M}'$ is a linear combination of the others almost everywhere with respect to ρ.

Let $\Theta := \left(\mathbb{R}_+^{|\cup_{i=0}^4 \mathcal{M}_i|} \times \mathbb{R}^{|\mathcal{M}_5'|}\right) \times \mathbb{R}^d$. Then, (SSVI-opt) is equivalent to the following finite-dimensional, $(\ell_V \wedge \ell_V')$-strongly convex optimization problem:

$$(\hat{\lambda}, \hat{v}) = \operatorname*{arg\,min}_{(\lambda,v)\in\Theta} \mathsf{KL}\left(\left(\boldsymbol{\alpha}\,\mathrm{id} + \sum_{T\in\cup_{i=0}^4\mathcal{M}_i} \lambda_T T + \sum_{\tilde{T}\in\mathcal{M}_5'} \lambda_{\tilde{T}}\tilde{T} + v\right)_{\#}\rho \,\middle\|\, \pi\right), \tag{91}$$

where Θ is equipped with the norm $\|(\lambda, v)\|_\Theta := \|\sum_{T\in\mathcal{M}'} \lambda_T T + v\|_{L^2(\rho)}$. To see the equivalence, we first note that by direct calculation

$$\|(\lambda, v)\|_\Theta^2 = \left\|\sum_{T\in\mathcal{M}'} \lambda_T T + v\right\|_{L^2(\rho)}^2 = \langle \lambda, Q\lambda\rangle + \langle v, v\rangle = \|\lambda\|_Q^2 + \|v\|^2,$$

where Q is a positive semi-definite Gram matrix with entries $Q_{T,\tilde{T}} := \langle T, \tilde{T}\rangle_{L_2(\rho)}$ for $T, \tilde{T} \in \mathcal{M}'$. In fact, Q is positive definite since functions in $\mathcal{M}'$ are linearly independent almost everywhere under ρ. As a result, $\|(\lambda, v)\|_\Theta = 0$ implies $(\lambda, v) = (0, 0)$, i.e., $\|\cdot\|_\Theta$ is indeed a norm.

Define $T_{\lambda,v} := \boldsymbol{\alpha}\,\mathrm{id} + \sum_{T\in\mathcal{M}'} \lambda_T T + v$. The set $\underline{\mathrm{cone}}\,(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})_{\#}\rho$ endowed with the linearized optimal transport distance admits a natural isometric embedding into the space Θ equipped with norm $\|\cdot\|_\Theta$, in the sense that

$$\begin{aligned}\mathrm{d}_{\mathrm{LOT},\rho}^2(\mu_{\eta,u}, \mu_{\lambda,v}) &= \left\|\sum_{T\in\mathcal{M}'} (\eta_T - \lambda_T)T + u - v\right\|_{L^2(\rho)}^2 \\ &= \|\eta - \lambda\|_Q^2 + \|u - v\|_2^2 = \|(\eta, u) - (\lambda, v)\|_\Theta^2.\end{aligned}$$

This equivalence of norms, together with the discussion in Section 3.1, establishes the claimed equivalence between the lifted and finite-dimensional formulations.

We can now employ tools from finite-dimensional convex optimization to solve (SSVI-opt) with convergence guarantees. In particular, we compute the minimizer using projected gradient descent (PGD) as described in Algorithm 1; see, for example, Section 10.2 of Beck (2017) for a detailed discussion of its convergence properties.

Algorithm 1 Projected gradient descent over $\underline{\mathrm{cone}}\,(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$

Input: Piecewise linear dictionary $\mathcal{M}'$, and projection cone $\mathcal{K} := \mathbb{R}_+^{|\cup_{i=0}^4 \mathcal{M}_i|} \times \mathbb{R}^{|\mathcal{M}_5'|}$. Constants $\alpha := \ell_V \wedge \ell_V'$, $L := L_V \vee (L_V'/2)$, $\Upsilon := 9\delta^{-2}\left((L_V')^{1/2} + (d-1)\,L_V^{1/2}\right)^2 \|Q^{-1}\|_2$.
Output: $(\hat{\lambda}, \hat{v})$.
Initialization: Initialize $(\lambda^{(0)}, v^{(0)}) \in \Theta$
for $t = 0, 1, 2, 3, \ldots$ **do**
  $\lambda^{(t+1)} = \mathrm{proj}_{\mathcal{K},Q}\left(\lambda^{(t)} - \frac{1}{L+\Upsilon}\, Q^{-1}\nabla_\lambda\, \mathsf{KL}(\mu_{\lambda^{(t)},v^{(t)}} \,\|\, \pi)\right)$
  $v^{(t+1)} = v^{(t)} - \frac{1}{L+\Upsilon}\,\nabla_v\, \mathsf{KL}(\mu_{\lambda^{(t)},v^{(t)}} \,\|\, \pi)$
end for

Theorem C.18 (Convergence results for Algorithm 1)**.** *Let Assumptions* (**R**), (**P1**), (**P2**), *and* (**RD+**) *be satisfied. Let* $\Upsilon := 9\delta^{-2}\left((L_V')^{1/2} + (d-1)\,L_V^{1/2}\right)^2 \|Q^{-1}\|_2$, *and let* $\{(\lambda^{(t)}, v^{(t)})\}_{t\in\mathbb{N}}$ *denote the iterates generated by Algorithm 1. Then for* $\kappa := \frac{L_V \vee (L_V'/2) + \Upsilon}{\ell_V \wedge \ell_V'}$,

(i) $\mathrm{d}^2_{\mathrm{LOT},\rho}(\mu_{\lambda(t),v(t)}, \mu_{\hat{\lambda},\hat{v}}) \le (1-\kappa^{-1})^t \, \mathrm{d}^2_{\mathrm{LOT},\rho}(\mu_{\lambda(0),v(0)}, \mu_{\hat{\lambda},\hat{v}})$.

(ii) $\mathsf{KL}(\mu_{\lambda(t),v(t)} \parallel \pi) - \mathsf{KL}(\mu_{\hat{\lambda},\hat{v}} \parallel \pi) \le (1-\kappa^{-1})^t \, \frac{(L+\Upsilon)}{2} \, \mathrm{d}^2_{\mathrm{LOT},\rho}(\mu_{\lambda(0),v(0)}, \mu_{\hat{\lambda},\hat{v}})$.

We first note that applying Theorem C.18 with $\theta^{(t)} \coloneqq (\lambda^{(t)}, v^{(t)})$ shows Theorem 3.8. The key step of the proof is to establish the *smoothness* property for the KL divergence over the finite-dimensional space $\underline{\mathrm{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$, which is non-trivial given that the entropy functional is non-smooth over the full Wasserstein space (see, e.g., Chewi, 2025, Section 6.2.2).

Computing the gradient. To apply the optimization algorithms, we compute the gradient of $\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ w.r.t. (λ, v) and the projection operator w.r.t. $\|\cdot\|_Q$. Under the change of variables formula,

$$\begin{aligned}\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = &- \int \log\det\Big(\mathrm{diag}(\boldsymbol{\alpha}) + \sum_{T\in\mathcal{M}'} \lambda_T DT(x)\Big)\,\rho(\mathrm{d}x) \\ &+ \int V\Big(\mathrm{diag}(\boldsymbol{\alpha})\,x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\,\rho(\mathrm{d}x) + \textsc{const}\,.\end{aligned}$$

Since $\mathrm{diag}(\boldsymbol{\alpha}) + \sum_{T\in\mathcal{M}'} \lambda_T DT(x)$ is always invertible, the partial gradients are computed as:

$$\nabla_v \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = \int \nabla V\Big(\mathrm{diag}(\boldsymbol{\alpha})\,x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\,\rho(\mathrm{d}x)\,, \tag{92}$$

and

$$\begin{aligned}\partial_{\lambda_T} \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = &- \int \Bigg[\mathrm{tr}\Big(\Big(\mathrm{diag}(\boldsymbol{\alpha}) + \sum_{T\in\mathcal{M}'} \lambda_T DT(x)\Big)^{-1} DT(x)\Big) \\ &+ \Big\langle T(x),\, \nabla V\Big(\mathrm{diag}(\boldsymbol{\alpha})\,x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\Big\rangle\Bigg]\,\rho(\mathrm{d}x)\,.\end{aligned} \tag{93}$$

To compute the first term in $\nabla_{\lambda_T} \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$, note that DT is the following lower triangular matrix with non-negative diagonals:

$$[DT(x)]_{ij} = \begin{cases} \partial_{x_i} T_i(x_i; x_1)\,, & \text{if } j = i\,, \\ \partial_{x_1} T_i(x_i; x_1)\,, & \text{if } j = 1\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

The inverse of $\mathrm{diag}(\boldsymbol{\alpha}) + \sum_{T\in\mathcal{M}'} \lambda_T DT(x)$ can be efficiently calculated using forward substitution with d linear equations, each with at most two non-zero terms per row. This gives a total computational cost of $O(d\,|\mathcal{M}'|)$ to compute all $DT(x)$ terms and $O(d)$ to invert it.

The terms involving ∇V can be approximated by taking the averages of ∇V evaluated at the samples drawn from the Gaussian distribution ρ.

Working toward proving Theorem C.18, we collect some key facts about the matrix $\sum_{T\in\mathcal{M}'} \lambda_T DT(x)$, including a bound on its spectral norm.

Lemma C.19. *Fix $x \in \mathbb{R}^d$. Define $M(\lambda) \coloneqq \sum_{T\in\mathcal{M}'} \lambda_T\, DT(x)$ for $\lambda \in \mathbb{R}_+^{|\cup_{i=0}^4 \mathcal{M}_i|} \times \mathbb{R}^{|\mathcal{M}'_5|}$. Then, $M(\lambda)$ is lower triangular with non-negative diagonal entries and if $\|\lambda\| = 1$, then $\|M(\lambda)\|_2^2 \le 9\,\delta^{-2}$.*

Proof. Recall that $\mathcal{M}' = \bigcup_{j=0}^{4} \mathcal{M}_j \cup \mathcal{M}'_5$. For each $T \in \mathcal{M}'$, the Jacobian matrix $DT(x)$ admits the following structure:

- If $T \in \mathcal{M}_0$, the $(1,1)$ entry of $DT(x)$ is $\delta^{-1}\mathbb{1}_{\{x_1 \in [b, b+\delta)\}}$, and all the others are zero.
- If $T \in \mathcal{M}_1$, the (i,i) entry of $DT(x)$ is $\delta^{-1}\psi\Big(\frac{x_1 - b'}{\delta}\Big)\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \in [b', b'+\delta)\}}$, and the $(i,1)$ entry is $\delta^{-1}\psi\Big(\frac{x_i - b}{\delta}\Big)\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \in [b', b'+\delta)\}}$. All other entries are zero.
- If $T \in \mathcal{M}_2$, the (i,i) entry of $DT(x)$ is $\delta^{-1}\psi\Big(1 - \frac{x_1 - b'}{\delta}\Big)\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \in [b', b'+\delta)\}}$, and the $(i,1)$ entry is $-\delta^{-1}\psi\Big(\frac{x_i - b}{\delta}\Big)\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \in [b', b'+\delta)\}}$. All other entries are zero.
- If $T \in \mathcal{M}_3$, the (i,i) entry of $DT(x)$ is $\delta^{-1}\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \geq R\}}$, and all the others are zero.
- If $T \in \mathcal{M}_4$, the (i,i) entry of $DT(x)$ is $\delta^{-1}\mathbb{1}_{\{x_i \in [b, b+\delta),\, x_1 \leq -R\}}$, and all the others are zero.
- If $T \in \mathcal{M}'_5$, the $(i,1)$ entry of $DT(x)$ is $\delta^{-1}\mathbb{1}_{\{x_1 \in [b, b+\delta)\}}$, and all the others are zero.

By the structure of the maps in $\mathcal{M}'$, we conclude that for any $x \in \mathbb{R}^d$,

$$[DT(x)]_{ii} \geq 0 \text{ for any } T \in \mathcal{M}', \quad \text{and} \quad \sum_{T \in \mathcal{M}'} [DT(x)]_{ii} \leq \delta^{-1}\,. \tag{94}$$

For any $\lambda \in \mathbb{R}_+^{|\cup_{j=0}^{4}\mathcal{M}_j|} \times \mathbb{R}^{|\mathcal{M}'_5|}$, it follows that $M \coloneqq \sum_{T \in \mathcal{M}'} \lambda_T DT(x)$ is a lower triangular matrix with $M_{ii} \geq 0$.

To control $\|M\|_2$, we follow the argument as in Lemma C.17 to write $M = \mathbf{D} + ve_1^\top$ where $\mathbf{D}$ is a diagonal matrix and $v = (0, v_2, \ldots, v_d)^\top$. Then,

$$\|M\|_2 \leq \|\mathbf{D}\|_2 + \|v\|\,. \tag{95}$$

By (94), we derive that for $\|\lambda\| = 1$,

$$0 \leq \mathbf{D}_{ii} \leq \sum_{T \in \mathcal{M}'} [DT(x)]_{ii} \leq \delta^{-1}\,.$$

It remains to compute $\|v\|$. For each $i \geq 2$, if $x_i \in [-R, R)$ and $x_1 \in [-R, R)$, there exist unique $T_1^i \in \mathcal{M}_1$, $T_2^i \in \mathcal{M}_2$, and $T_5^i \in \mathcal{M}'_5$ depending only on the sub-intervals of (x_1, x_i) such that

$$\max\Big\{|[DT_1^i(x)]_{i1}|,\, |[DT_2^i(x)]_{i1}|,\, |[DT_5^i(x)]_{i1}|\Big\} \leq \delta^{-1}\,,$$

and $[DT(x)]_{i1} = 0$ for any other $T \in \mathcal{M}'$. If $x_i \notin [-R, R)$ and $x_1 \in [-R, R)$, there exists $T_5^i \in \mathcal{M}'_5$ such that $|[DT_5^i(x)]_{i1}| \leq \delta^{-1}$ and $[DT(x)]_{i1} = 0$ for any other $T \in \mathcal{M}'$. Otherwise, if $x_1 \notin [-R, R)$, then $[DT(x)]_{i1} = 0$ for all $T \in \mathcal{M}'$. Therefore,

$$\begin{aligned}
\|v\|^2 &= \sum_{i=2}^{d}\Big(\sum_{T \in \mathcal{M}'} \lambda_T\,[DT(x)]_{i1}\Big)^2 \\
&\leq \sum_{i=2}^{d}\Big(\Big[|\lambda_{T_1^i} DT_1^i(x)| + |\lambda_{T_2^i} DT_2^i(x)| + |\lambda_{T_5^i} DT_5^i(x)|\Big]_{i1}\Big)^2 \leq 3\,\|\lambda\|^2\,\delta^{-2} = 3\delta^{-2}\,.
\end{aligned}$$

Hence, $\|M\|_2 \leq 3\delta^{-1}$. □

We now establish the geodesic smoothness result for the KL divergence over $\underline{\text{cone}}(\mathcal{M}; \boldsymbol{\alpha}\,\mathrm{id})$. Recall that $\Upsilon \coloneqq \left(L_V'^{1/2} + (d-1)\,L_V^{1/2}\right)^2 \frac{9}{\delta^2}\,\|Q^{-1}\|_2$ is defined in Theorem C.18.

Lemma C.20. *If π is L-log-smooth, the function $(\lambda, v) \mapsto \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ is $(L+\Upsilon)$-smooth over the space $(\Theta, \|\cdot\|_\Theta)$.*

Proof. Denote by $M \coloneqq \sum_{T\in\mathcal{M}'} \lambda_T DT(x)$. Recall the expressions (92) and (93) for the gradients of the KL divergence.

The second derivative of $\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ w.r.t. v is

$$\nabla_v^2 \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = \int \nabla^2 V\Big(\operatorname{diag}(\boldsymbol{\alpha})\, x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\, \rho(\mathrm{d}x)\,, \tag{96}$$

and the mixed partial derivatives w.r.t. (λ, v) are

$$\partial_{\lambda_T} \nabla_v \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = \int \nabla^2 V\Big(\operatorname{diag}(\boldsymbol{\alpha})\, x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\, T(x)\, \rho(\mathrm{d}x). \tag{97}$$

The Hessian of $\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ w.r.t. λ has entries:

$$\begin{aligned}\partial_{\lambda_T}\partial_{\lambda_{T'}} \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) = \int \Bigg[&\underbrace{\operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} DT'(x) \big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} DT(x)\Big)}_{=: H_{T,T'}} \\ &+ \Big\langle T(x), \nabla^2 V\Big(\operatorname{diag}(\boldsymbol{\alpha})\, x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\Big)\, T'(x)\Big\rangle \Bigg]\, \rho(\mathrm{d}x)\,.\end{aligned}$$

To bound the operator norm of the matrix $H = (H_{T,T'})_{T,T'\in\mathcal{M}'}$, consider an arbitrary unit vector $u \in \mathbb{R}^{|\mathcal{M}'|}$ with $\|u\| = 1$. We compute

$$\begin{aligned} u^\top H u &= \sum_{T,T'\in\mathcal{M}'} u_T \operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} DT'(x) \big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} DT(x)\Big) u_{T'} \\ &= \operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} \sum_{T'\in\mathcal{M}'} u_{T'} DT'(x) \big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} \sum_{T\in\mathcal{M}'} u_T DT(x)\Big)\,. \end{aligned}$$

For any lower triangular matrices $A, B \in \mathbb{R}^{d\times d}$ such that B has positive diagonal entries, the following hold $\operatorname{tr}(A^2) \le \operatorname{tr}(A)^2$ and $|\operatorname{tr}(AB)| \le \|A\|_2 \operatorname{tr}(B)$. Together with Lemma C.19, we derive that

$$\begin{aligned} u^\top H u &\le \operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1} \sum_{T\in\mathcal{M}'} u_T DT(x)\Big)^2 \\ &\le \operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + M\big)^{-1}\Big)^2 \frac{9}{\delta^2} \\ &= \operatorname{tr}\Big(\big(\operatorname{diag}(\boldsymbol{\alpha}) + \operatorname{diag}(M)\big)^{-1}\Big)^2 \frac{9}{\delta^2} \\ &\le \Big(\sum_{j=1}^d \alpha_j^{-1}\Big)^2 \frac{9}{\delta^2} = \left(L_V'^{1/2} + (d-1)\,L_V^{1/2}\right)^2 \frac{9}{\delta^2}\,, \end{aligned}$$

where the last two lines follow by the property of lower triangular matrices and the non-negativity of the diagonal entries of M. Therefore, we have shown that

$$H \preceq L_H I_{|\mathcal{M}'|}\,, \qquad \text{where } L_H \coloneqq \left(L_V'^{1/2} + (d-1)\, L_V^{1/2}\right)^2 \frac{9}{\delta^2}\,. \tag{98}$$

Denote by $\mathbf{T}_{\mathcal{M}'} = [T_1, \ldots, T_{|\mathcal{M}'|}] : \mathbb{R}^d \to \mathbb{R}^{d\times|\mathcal{M}'|}$ the matrix-valued function whose columns are transport maps in $\mathcal{M}'$. By (96)–(98), we derive that

$$\begin{aligned}
\nabla^2_{\lambda,v}\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right) &= \int \begin{bmatrix} \mathbf{T}_{\mathcal{M}'}(x) \\ I \end{bmatrix} \nabla^2 V\left(\operatorname{diag}(\boldsymbol{\alpha})\, x + \sum_{T\in\mathcal{M}'} \lambda_T T(x) + v\right) \left[\mathbf{T}_{\mathcal{M}'}(x), I\right] \rho(\mathrm{d}x) \\
&\quad + \begin{bmatrix} H & 0_{|\mathcal{M}'|\times d} \\ 0_{d\times|\mathcal{M}'|} & 0_{d\times d} \end{bmatrix} \\
&\preceq L \int \left[\mathbf{T}_{\mathcal{M}'}(x), I\right]^\top \left[\mathbf{T}_{\mathcal{M}'}(x), I\right] \rho(\mathrm{d}x) + L_H \begin{bmatrix} I_{|\mathcal{M}'|} & 0_{|\mathcal{M}'|\times d} \\ 0_{d\times|\mathcal{M}'|} & 0_{d\times d} \end{bmatrix}\,.
\end{aligned}$$

Under the geometry $(\Theta, \|\cdot\|_\Theta)$, the partial Hessian $\nabla^2_{\lambda,v}\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ needs to be rescaled by $Q^{-1/2}$. Concretely, for any linear operator $A : \mathbb{R}^{|\mathcal{M}'|+d} \to \mathbb{R}^{|\mathcal{M}'|+d}$, denote by $\|A\|_\Theta \coloneqq \sup\left\{\|A\theta\|^*_\Theta \,:\, \|\theta\|_\Theta = 1\right\}$ the operator norm of A where $\|\cdot\|^*_\Theta$ is the dual norm of $\|\cdot\|_\Theta$, then $\|A\|_\Theta = \|G^{-1/2} A G^{-1/2}\|_2$ with $G = \begin{bmatrix} Q & 0_{|\mathcal{M}'|\times d} \\ 0_{d\times|\mathcal{M}'|} & I \end{bmatrix}$. Thus,

$$\begin{aligned}
&\left\|\nabla^2_{\lambda,v}\mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)\right\|_\Theta \\
&\quad\le L \left\| \int \left[\mathbf{T}_{\mathcal{M}'}(x)\, Q^{-1/2},\, I\right]^\top \left[\mathbf{T}_{\mathcal{M}'}(x)\, Q^{-1/2},\, I\right] \rho(\mathrm{d}x) \right\|_2 + L_H \left\|Q^{-1}\right\|_2 \\
&\quad= L \left\| \begin{bmatrix} Q^{-1/2} \int \mathbf{T}_{\mathcal{M}'}(x)^\top \mathbf{T}_{\mathcal{M}'}(x)\, \rho(\mathrm{d}x)\, Q^{-1/2} & 0_{|\mathcal{M}'|\times d} \\ 0_{d\times|\mathcal{M}'|} & I \end{bmatrix} \right\|_2 + L_H \left\|Q^{-1}\right\|_2 \\
&\quad\le L + L_H \left\|Q^{-1}\right\|_2\,.
\end{aligned}$$

The second equality follows from the block structure of the matrix and the fact that $\int \mathbf{T}_{\mathcal{M}'}(x)\, \rho(\mathrm{d}x) = 0_{d\times|\mathcal{M}'|}$ after re-centering. □

Proof of Theorem C.18. First, we apply Lemma 2.6: as π satisfies (**P1**) and (**RD+**) implies (**RD**), π is $\ell'_V \wedge \ell_V$-log-concave. Since (**R**) and (**P2**) are also satisfied, π is $L_V \vee (L'_V/2)$-log-smooth.

By Lemma C.20, $(\lambda, v) \mapsto \mathsf{KL}\left(\mu_{\lambda,v} \parallel \pi\right)$ is $(\ell'_V \wedge \ell_V)$-strongly convex and $(L_V \vee (L'_V/2) + \Upsilon)$-smooth over $(\Theta, \|\cdot\|_\Theta)$. The guarantee now follows from (Beck, 2017, Theorem 10.29). □