# Even More Guarantees for Variational Inference in the Presence of Symmetries

**Lena Zellinger**
School of Informatics
University of Edinburgh
l.zellinger@sms.ed.ac.uk

**Antonio Vergari**
School of Informatics
University of Edinburgh
avergari@ed.ac.uk

## Abstract

When approximating an intractable density via variational inference *(VI)* the variational family is typically chosen as a simple parametric family that very likely does not contain the target. This raises the question: *Under which conditions can we recover characteristics of the target despite misspecification?* In this work, we extend previous theoretical results on robust VI with location-scale families under target symmetries in two substantial ways: (1) We open them up to a wider range of divergences by providing sufficient conditions for exact recovery of the target mean and correlation matrix when using the forward Kullback-Leibler divergence and $\alpha$-divergences. (2) By doing so, we find that we can drop the restrictive assumption of a log-concave target made in previous work, allowing us to give guarantees for a wider range of targets, including multi-modal ones. In our experiments, we show how our guarantees can serve as guidelines for the choice of the variational family and $\alpha$-value and we illustrate on a diverse set of examples how and why optimization can fail in the absence of our sufficient conditions.

## 1 Introduction

Many tasks in machine learning (ML) and statistics revolve around inferring intractable statistical quantities of a given target distribution $p$, such as expected predictions [35, 61, 13, 73] and moments [67, 10]. Variational inference [VI, 62, 9] recasts the problem of performing these intractable inference tasks over $p$ as optimization: Given a variational family $\mathcal{Q}$, we find a surrogate $q^\star \in \mathcal{Q}$ by minimizing a divergence of choice $D$, and then use $q^\star$ to perform tractable inference. Typical choices for $\mathcal{Q}$ include simple parametric distributions [68, 53] and finite mixture models [54, 3, 31, 78].

In practice, $\mathcal{Q}$ is therefore often *misspecified*, i.e., $p \notin \mathcal{Q}$ [71]. In such cases, the choice of $\mathcal{Q}$ and $D$ can greatly affect the resulting approximation of the target (see Fig. 1 for examples). While $\mathcal{Q}$ is not expressive enough to fully capture $p$, we might still be able to capture informative summary statistics of the target, such as its mean and correlations between the variables. This raises the question: *Under which conditions on $p$, $\mathcal{Q}$, and $D$ are we **guaranteed** to capture relevant characteristics of the target despite misspecification?*

Most existing results studying the quality of VI approximations are *asymptotic* [29, 79, 72, 76, 20, 4, 58], relying, for instance, on the Bernstein-von-Mises theorem [69], or providing post-hoc diagnostics [77, 32]. Guarantees for *exact* recovery of target statistics are still sparse and often assume *Gaussian* approximations: Wainwright and Jordan [70] show that the optimal Gaussian approximation w.r.t. the FKL exactly matches the target mean and covariance. Katsevich and Rigollet [34] provide error bounds on the mean and covariance for Gaussian approximations, but do not guarantee exact recovery. Odgers et al. [56] study the quality of predictive uncertainty in conjugate Bayesian linear regressions for Gaussian mean-field VI. Domke [19] exploits smoothness of the target to establish a lower bound on the eigenvalues of the optimal covariance matrix in VI with general location-scale

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

families. Studying variational families beyond Gaussians is relevant in practice since surrogates with heavier tails can be preferable for expectation estimation via importance sampling [77, 17, 37].

Recently, Margossian and Saul [46] exploited *symmetries in the target distribution* to provide sufficient conditions for the *exact recovery of the target mean and correlation matrix* when $\mathcal{Q}$ is a location-scale family. Symmetries can arise naturally in Bayesian posteriors when either the likelihood or the prior exhibits symmetries and dominates the other (for instance, when only little data is available) [46]. Their initial results considered the *reverse Kullback-Leibler divergence (RKL)*, which is a standard objective in VI [9], but known for its mode-seeking behavior, which can lead to poor target coverage [26, 48]. Margossian et al. [48] subsequently extended their analysis to $f$-divergences [63], which generalize other popular divergences beyond the RKL, such as the *forward KL divergence (FKL)* [41, 70, 33, 12] and *$\alpha$-divergence ($\alpha$-DIV)* [14, 43, 16, 28, 27]. Unfortunately, only the RKL was shown to provably attain a unique minimizer under their conditions, while the remaining divergences are only guaranteed to reach a stationary point of the location parameter at the target mean. This begs the question: *Can we retrieve similar guarantees for VI when using the FKL and $\alpha$-DIV?*

We fill this gap by extending the analysis of Margossian and Saul [46, 47] and providing the following contributions: **(1)** In Sec. 3.1, we derive *complementary sufficient conditions* for the exact recovery of the mean when optimizing the FKL and $\alpha$-DIV. We find that we can guarantee exact recovery of the mean with the FKL when placing mild assumptions on the base distribution of the location-scale family. For the $\alpha$-DIV, we establish a more fine-grained criterion that additionally depends on the value of $\alpha$. Notably, our analysis, unlike previous results [46, 47], does not rely on log-concavity of the target, and hence also applies to multi-modal targets, such as mixture distributions [55, 50], allowing us to provide guarantees for a wider range of realistic targets. **(2)** In Sec. 3.2, we provide sufficient conditions for simultaneously recovering the correlation matrix of the target. Akin to Margossian and Saul [46, 47], we place additional symmetry requirements on the target for correlation recovery, but still alleviate the assumption of log-concavity. **(3)** Lastly, in Sec. 4, we discuss how our guarantees can inform the choice of the variational family in practice, and show how optimization can fail to recover target characteristics when our sufficient conditions are not met.

## 2 Location-scale VI under symmetries

**Notation.** Throughout the paper, we denote vectors by bold lowercase letters. For our analysis, $S^{\frac{1}{2}}$ refers to the principal square root of a positive definite matrix $S \in \mathbb{S}^d_{++}$, and we use $\text{supp}(p) = \{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) > 0\}$ to denote the support of a density $p$.

We consider a standard variational inference setting, where given a target $p$, we aim to find the best tractable surrogate within the variational family $q^\star \in \mathcal{Q}$ by optimizing a divergence of choice $D$:

$$q^\star = \text{argmin}_{q \in \mathcal{Q}} D(p, q). \tag{1}$$

Our analysis focuses on $\alpha$-DIVs [14], defined as

$$D_\alpha(p, q) := \frac{1}{\alpha(\alpha - 1)} \int \left[ \left( \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right)^\alpha - 1 \right] q(\boldsymbol{x}) d\boldsymbol{x}, \tag{$\alpha$-DIV}$$

with $\alpha > 0$ and $\alpha \neq 1$. The $\alpha$-DIV is a flexible family of divergences: For $\alpha = \frac{1}{2}$, we obtain a divergence that is proportional to the squared Hellinger distance [42], while $\alpha = 2$ recovers the $\chi^2$-divergence [15, 18]. In the limiting case of $\alpha \to 0$, Eq. ($\alpha$-DIV) recovers the RKL, $D_{\text{RKL}}(p, q) = \int \log\left(q(\boldsymbol{x})/p(\boldsymbol{x})\right) q(\boldsymbol{x}) d\boldsymbol{x}$, while $\alpha \to 1$ gives the FKL, $D_{\text{FKL}}(p, q) = \int \log\left(p(\boldsymbol{x})/q(\boldsymbol{x})\right) p(\boldsymbol{x}) d\boldsymbol{x}$. Following Margossian and Saul [46, 47], we study *location-scale families* as variational families.

**Definition 1** (Location-Scale Families)**.** *A member of a location-scale family can be expressed as*

$$q_{\boldsymbol{\nu}, S}(\boldsymbol{x}) = q_{\mathbf{0}}\left(S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu})\right) |S|^{-\frac{1}{2}}, \tag{2}$$

*where $q_{\mathbf{0}}$ is the base distribution, $\boldsymbol{\nu} \in \mathbb{R}^d$ is the location parameter and $S \in \mathbb{S}^d_{++}$ is a positive definite scale matrix. A* location family*, denoted by $q_{\boldsymbol{\nu}}$, only varies $\boldsymbol{\nu}$ while $S$ is fixed and equal for all members of the family.*

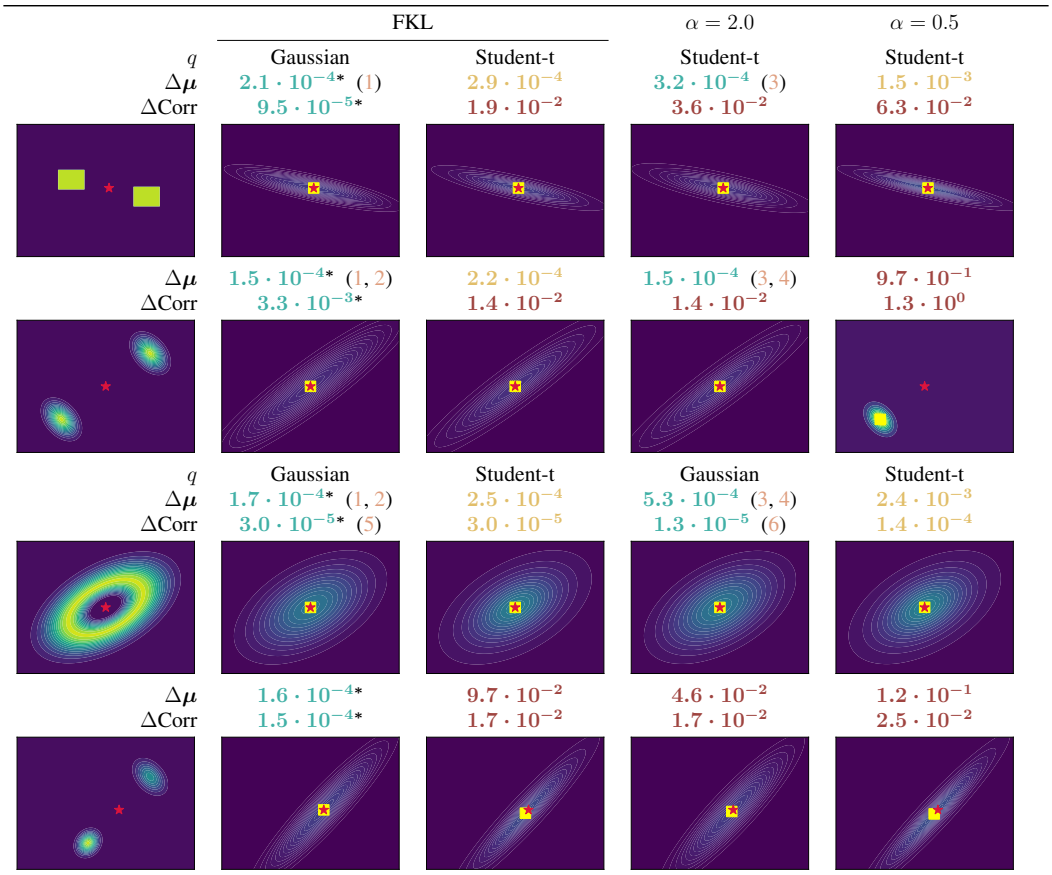



Figure 1: **We provide theoretical guarantees to recover the location and correlation of multi-modal targets which correspond to low error in practice**. The mean of the target $p$ is indicated via a red star. The mean of $q$ is denoted by a yellow square. The errors $\Delta\boldsymbol{\mu}$ and $\Delta$Corr are measured as described in Sec. 4.2. We distinguish 3 cases (1) **low error** ($< 10^{-2}$) **as predicted by our guarantees** (see corresponding theorems in parentheses) or previous results (indicated via *), (2) **no guarantee but low error in practice**, (3) **no guarantee and high error** ($\geq 10^{-2}$).

Location-scale families include many prevalent distributions in statistics, such as the Gaussian, Laplace, and Student-t families. For most location-scale families, $\boldsymbol{\nu}$ corresponds to the mean of $q_{\boldsymbol{\nu}}$, but this is not always the case. For instance, the mean of a Cauchy distribution is undefined, and $\boldsymbol{\nu}$ corresponds to its median (and mode) instead. Akin to Margossian and Saul [46, 47], our analysis places symmetry conditions on $q_0$ and the target $p$, which we illustrate in Fig. 2, and formally introduce next.

**Definition 2** (Even Symmetry)**.** *We say a function $h : \mathbb{R}^d \to \mathbb{R}$ is **even symmetric** around $\boldsymbol{\nu} \in \mathbb{R}^d$ if for all $\boldsymbol{x} \in \mathbb{R}^d$, $h(\boldsymbol{\nu} + \boldsymbol{x}) = h(\boldsymbol{\nu} - \boldsymbol{x})$.*

Fig. 2 illustrates even symmetry on the example of a mixture distribution. For every point, there exists a "mirrored" point on the opposite side of the mean with the exact same density value. We have now introduced all ingredients for (A1), the core assumptions underlying the results by Margossian and Saul [47] and our analysis for ***exact recovery of the target mean***. The result of Margossian and Saul [47] is stated for general $f$-divergences, defined as $D_f(p, q) = \int f(p(\boldsymbol{x})/q(\boldsymbol{x}))q(\boldsymbol{x})d\boldsymbol{x}$, which subsume the RKL, FKL, and $\alpha$-DIV for particular choices of the convex function $f$ (see App. A.2). Under (A1) and mild regularity conditions on $p$ and $q$, Margossian and Saul [47] arrive at Thm. 00.

**(A1).** *Let $q_{\boldsymbol{\nu}}$ be a location family with fixed scale matrix $S$ and let $p$ be the target. We make the following assumptions: The base distribution $q_0$ is even symmetric around $\mathbf{0}$. The target $p$ is even symmetric around its mean $\boldsymbol{\mu}$. Lastly, $D_f(p, q_{\boldsymbol{\nu}})$ is well-defined for all values of $\boldsymbol{\nu}$, and $\boldsymbol{\mu}$ is finite.*

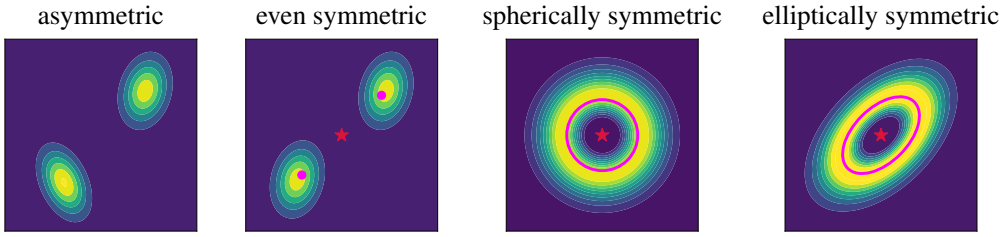



Figure 2: ***Illustration of the symmetries enabling our theoretical results***. A point of *even symmetry* is illustrated via a red star. The *spherically symmetric* ring was transformed into the *elliptically symmetric* ellipse via Eq. (3). Points (and sets of points) in pink attain the same density values within the respective figure.

**Theorem 00** (Mean recovery with the RKL [47])**.** *Under (A1), a stationary point of* $D_f(p, q_{\boldsymbol{\nu}})$ *occurs at* $\boldsymbol{\nu} = \boldsymbol{\mu}$*. Furthermore, if* $\varphi(\boldsymbol{v}) := f \circ \exp(\boldsymbol{v})$ *is convex and strictly decreasing, and* $p$ *somewhere-strictly log concave over* $\mathbb{R}^d$*, then* $\boldsymbol{\nu} = \boldsymbol{\mu}$ *is a unique minimizer of* $D_f(p, q_{\boldsymbol{\nu}})$*.*

Thm. 00 guarantees a *stationary point* of $D_f(p, q_{\boldsymbol{\nu}})$ at $\boldsymbol{\nu} = \boldsymbol{\mu}$. Notably, this partial result poses no additional requirements on $\varphi$, and hence applies to general $f$-divergences, such as the FKL and $\alpha$-DIV. The sufficient conditions guaranteeing a *unique minimizer* are more restrictive: On the one hand, (somewhere strict) log-concavity of $p$ implies a unimodal target. Moreover, the requirements on $\varphi$ are only fulfilled by the RKL among the $f$-divergences considered by Margossian and Saul [47]. Nevertheless, they found that the remaining $f$-divergences often located the correct mean in practice, even for targets that violate log-concavity, such as the Student-t distribution (cf. Fig. 2 in Margossian and Saul [47]). In Sec. 3.1, we provide *complementary sufficient conditions for exact recovery of the mean with the FKL and* $\alpha$*-DIV*. Notably, our results do not assume $p$ to be log-concave.

In order to additionally guarantee ***exact recovery of correlations***, we need to place a stronger set of assumptions on both the target and the variational family, which we summarize as **(A2)**. First, we introduce the concepts of ***spherical and elliptical symmetry***.

**Definition 3** (Spherical and Elliptical Symmetry)**.** *We say that a density* $p$ *is* ***spherically symmetric*** *if* $p(\boldsymbol{x}) = p(\boldsymbol{x}')$ *whenever* $||\boldsymbol{x}|| = ||\boldsymbol{x}'||$*. Further,* $p$ *is* ***elliptically symmetric*** *if there exists a positive definite matrix* $M \in \mathbb{S}^d_{++}$*, such that the density of* $\boldsymbol{\zeta} = M^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu})$ *is spherically symmetric.*

Conveniently, whenever $p$ is elliptically symmetric, we can express it in terms of a spherically symmetric base density $p_{\mathbf{0}}$ and a positive definite matrix $M \in \mathbb{S}^d_{++}$, which is directly related to the correlation matrix of $p$ via $\text{Corr}_p[x_i, x_j] = M_{ij}/\sqrt{M_{ii}M_{jj}}$,

$$p(\boldsymbol{x}) = p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu})\right)|M^{-\frac{1}{2}}|. \tag{3}$$

Moreover, since $p_{\mathbf{0}}$ is spherically symmetric, there exists a function $g : \mathbb{R}^+_0 \to \mathbb{R}^+_0$, such that $g(||\boldsymbol{x}||) = p_{\mathbf{0}}(\boldsymbol{x})$. Note that elliptical symmetry around a point $\boldsymbol{\mu}$ implies even symmetry around $\boldsymbol{\mu}$. Hence, **(A2)** below entails **(A1)**. Elliptical symmetry is, for instance, present in multivariate Gaussian and Student-t distributions, but we can also construct more intricate shapes fulfilling the property, such as the elliptic density in Fig. 2, by choosing a more expressive $p_{\mathbf{0}}$. By exploiting Eq. (3), Margossian and Saul [47] arrive at Thm. 0 under conditions **(A2)**.

**(A2).** *Let* $q_{\boldsymbol{\nu},S}$ *be a member of a location-scale family and let* $p$ *be a target with finite first and second moments. We make the following assumptions: The base distribution* $q_{\mathbf{0}}$ *is spherically symmetric. The target* $p$ *is elliptically symmetric with* $M \in \mathbb{S}^d_{++}$ *and mean* $\boldsymbol{\mu} \in \mathbb{R}^d$*. Lastly,* $D_f(p, q_{\boldsymbol{\nu},S})$ *is well-defined and finite for all considered values of* $\boldsymbol{\nu}$ *and* $S$*.*

**Theorem 0** (Mean and correlation recovery with the RKL [47])**.** *Assume (A2) and* ***additionally*** *that (i)* $p$ *is somewhere strictly log-concave over* $\mathbb{R}^d$ *and (ii)* $\varphi(\boldsymbol{u}) = f \circ \exp(\boldsymbol{u})$ *is convex and strictly decreasing and (iii)* $g(\boldsymbol{x}) = p_{\mathbf{0}}(||\boldsymbol{x}||)$ *is everywhere continuously differentiable with* $|g'(0)| < \infty$*. Then,* $D_f(p, q_{\boldsymbol{\nu},S})$ *has a unique minimizer with respect to the location-scale parameters* $(\boldsymbol{\nu}, S)$ *at* $\boldsymbol{\nu} = \boldsymbol{\mu}$ *and* $S = \gamma^2 M$ *for some* $\gamma > 0$*.*

First, we remark that despite the similarities of Eq. (3) and Eq. (2), Thm. 0 allows for misspecification of the variational family since the base distributions $q_{\mathbf{0}}$ and $p_{\mathbf{0}}$ can differ. Once again, however,

Thm. 0 is limited to the RKL and log-concave targets. In Sec. 3.2, we provide sufficient conditions on $p$ and $\mathcal{Q}$ under which we can guarantee simultaneous exact recovery of the target mean and correlations with the FKL and $\alpha$-DIV.

## 3 Guarantees for the FKL and $\alpha$-divergence

We now state our core theoretical results. The full proofs are given in App. B. Notably, none of our theorems requires log-concavity of $p$, hence, they apply to *multi-modal* and *heavy-tailed* targets. Our conditions primarily concern the *base distribution* of the variational family (Def. 1). See Tab. 1 for a summary of common location-scale distributions, and whether their base distributions fulfill properties related to our sufficient conditions.

### 3.1 Exact recovery of the mean

In this section, we state our results for exact mean recovery for the FKL and $\alpha$-DIV for location families. Throughout the analysis, we treat the scale parameter $S$ as fixed. When the conditions in our theorems are met, the target mean will be recovered for *any* assignment of $S$. As a result, our guarantees for exact mean recovery also apply when the variational family is further restricted to a simple *mean-field approximation* [9, 75], which is widely used in practice for computational efficiency. We begin by analyzing the **FKL**. First, note that optimizing the FKL w.r.t. the location parameter $\boldsymbol{\nu}$ results in the following objective:

$$D_{\text{FKL}}(p, q_{\boldsymbol{\nu}}) = \int \log\left(\frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu}}(\boldsymbol{x})}\right) p(\boldsymbol{x}) d\boldsymbol{x} = \int -\log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}\left(\boldsymbol{x}-\boldsymbol{\nu}\right)\right)\right) p(\boldsymbol{x}) d\boldsymbol{x} + C, \quad \text{(FKL)}$$

where we plug in the definition of a member of a location family (Def. 1) and $C$ is an additive constant in $\boldsymbol{\nu}$ (see App. A.2 for details). Viewing the FKL in this form gives rise to our Thm. 1.

**Theorem 1.** *Under (A1), $D_{FKL}(p, q_{\boldsymbol{\nu}})$ has a unique minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $-\log q_{\mathbf{0}}$ is strictly convex.*

Intuitively, the above condition ensures that $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ is strictly convex in $\boldsymbol{\nu}$. Combined with the insight from Thm. 00 that $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ has a stationary point at $\boldsymbol{\nu} = \boldsymbol{\mu}$, we conclude that $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ has a unique minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$. Interestingly, Margossian and Saul [46] already hinted at the strict convexity of general integral expressions akin to Eq. (FKL) in their Proposition 12, but to the best of our knowledge never used this result to state the existence of a unique minimizer for divergences besides the RKL. Thm. 1 also entails the classical result that the optimal Gaussian approximation w.r.t. the FKL will match the mean of the target [70]. However, it is more widely applicable: For instance, also other location-scale families such as the *logistic distribution* are covered by Thm. 1 (see App. A.1 for the definition and a visualization). Next, we state another result for the FKL that allows us to *relax strict convexity* of $-\log q_{\mathbf{0}}$ at the cost of placing a condition on the support of $p$.

**Theorem 2.** *Under (A1), $D_{FKL}(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if (1) $-\log q_{\mathbf{0}}$ is convex and strictly increasing in $||\boldsymbol{x}||$ and (2) $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure.*

Compared to Thm. 1, Thm. 2 places different assumptions on both $p$ and $q$: A Laplace distribution can fit into the framework of Thm. 2 but not Thm. 1. However, Thm. 2 indicates that VI with a Laplace family can fail to obtain a unique minimizer at $\boldsymbol{\mu}$ when $p$ is not supported around $\boldsymbol{\mu}$. In Sec. 4, we illustrate this phenomenon on a concrete example targeting a mixture of uniforms with disjoint support.

The condition that $-\log q_{\mathbf{0}}$ is strictly increasing in $||\boldsymbol{x}||$ is fulfilled by most standard location-scale families, as shown in Fig. 5. We note however that it is possible to construct functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $f$ is even symmetric and convex, but *not* strictly increasing in $||\boldsymbol{x}||$, such as $f(x_1, x_2) = x_1^4 + x_2^6$. At the same time, not every spherically symmetric function is strictly increasing in $||\boldsymbol{x}||$. For instance, the *uniform distribution* is not covered by our guarantees. We now provide intuition as to why Thm. 2 holds.

**Intuition for Theorem 2.** To prove Thm. 2 in App. B, we show directly that every alternative assignment to $\boldsymbol{\nu} = \boldsymbol{\mu}$ must result in a strictly larger value of $D_{\text{FKL}}$. For brevity, let $w := -\log q_{\mathbf{0}}$, $\boldsymbol{\nu}' \neq \mathbf{0}$ and $\boldsymbol{\tau} := \boldsymbol{x} - \boldsymbol{\mu}$. Consider the *difference in $D_{FKL}$* (Eq. (FKL)) *for alternative assignments of the location parameter $\boldsymbol{\nu} = \boldsymbol{\mu}$ and $\boldsymbol{\nu} = \boldsymbol{\mu} + \boldsymbol{\nu}'$*:

$$\Delta_{\text{FKL}} := D_{\text{FKL}}(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}) - D_{\text{FKL}}(p, q_{\boldsymbol{\mu}}) = \int \left[w(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')) - w(S^{-\frac{1}{2}}\boldsymbol{\tau})\right] p(\boldsymbol{\mu} + \boldsymbol{\tau}) d\boldsymbol{\tau}. \quad (4)$$

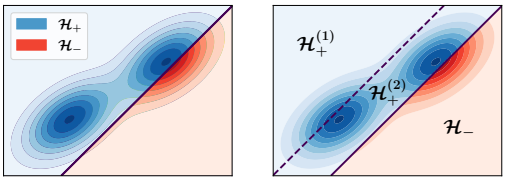



Figure 3: For Thm. 2 and Thm. 4, we exploit the symmetry of $p(\boldsymbol{\mu}+\boldsymbol{\tau})$ by splitting its domain into $\mathcal{H}_+$ (blue) and $\mathcal{H}_-$ (red), where $\Delta_{\text{FKL}}$ (Eq. (4)) *increases* over the first and *decreases* over the second (L). We further partition $\mathcal{H}_+$ into $\mathcal{H}_+^{(1)}$ and $\mathcal{H}_+^{(2)}$, where $\mathcal{H}_+^{(1)}$ mirrors $\mathcal{H}_-$ (R). We set $\boldsymbol{\nu}' = (1.5, -0.9)$ and $q_{\mathbf{0}}$ as a standard Gaussian.

Remember that $w(\boldsymbol{z})$ is assumed to be strictly increasing in $||\boldsymbol{z}||$. Therefore, we can distinguish points that increase $\Delta_{\text{FKL}}$ from those that decrease $\Delta_{\text{FKL}}$ via the condition $||S^{-\frac{1}{2}}(\boldsymbol{\tau}+\boldsymbol{\nu}')|| > ||S^{-\frac{1}{2}}\boldsymbol{\tau}||$. We denote the resulting sets of points, which form half-spaces for strictly radially increasing $w$, by $\mathcal{H}_+$ and $\mathcal{H}_-$ respectively (see Fig. 3 (L)). We further split $\mathcal{H}_+$ into disjoint sets $\mathcal{H}_+^{(1)}$ and $\mathcal{H}_+^{(2)}$, such that $\mathcal{H}_+^{(1)}$ "mirrors" $\mathcal{H}_-$: for every $\boldsymbol{\tau} \in \mathcal{H}_+^{(1)}$, it holds that $-\boldsymbol{\tau} \in \mathcal{H}_-$, and vice versa (see Fig. 3 (R)). Moreover, by even symmetry of $p$ around $\boldsymbol{\mu}$, we have $p(\boldsymbol{\mu}+\boldsymbol{\tau}) = p(\boldsymbol{\mu}-\boldsymbol{\tau})$. When convexity is not strict, it can happen that the contributions of $\mathcal{H}_+^{(1)}$ and $\mathcal{H}_-$ to $\Delta_{\text{FKL}}$ cancel. In such cases, we need to rely on $\mathcal{H}_+^{(2)}$ to ensure $\Delta_{\text{FKL}} > 0$. When $p$ is not supported around $\boldsymbol{\mu}$ on a set of positive measure, the contribution of $\mathcal{H}_+^{(2)}$ to $\Delta$FKL can be exactly 0, and hence the minimizer might not be unique. In Fig. 4, we show a practical example of this case, where a Laplace variational family (for which $-\log q_{\mathbf{0}}$ is convex but not strictly) fails to distinguish between alternative assignments of $\boldsymbol{\nu}$ due to a lack of support around the true mean.

Our construction also provides a different way to prove the existence of stationary points than Thm. 00. Moreover, it sheds further light on Thm. 1: When $-\log q_{\mathbf{0}}$ is strictly convex, the contribution of $\mathcal{H}_+^{(1)}$ to $\Delta_{\text{FKL}}$ will be strictly greater than the decrease induced by $\mathcal{H}_-$. Hence, we do not need to rely on $\mathcal{H}_+^{(2)}$ to ensure a strictly positive value of $\Delta_{\text{FKL}}$ in Thm. 1.

By the same line of reasoning, we arrive at analogous results for the **$\boldsymbol{\alpha}$-DIV** (Eq. ($\alpha$-DIV)) which for fixed $S$ simplifies to

$$D_\alpha(p, q_{\boldsymbol{\nu}}) = |S|^{-\frac{1-\alpha}{2}} \int \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}\Big(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\Big) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} + C, \tag{5}$$

where $\alpha > 0$ and $\alpha \neq 1$ and $C$ again denotes a constant in $\boldsymbol{\nu}$. Note that $|S|^{-\frac{1-\alpha}{2}}$ is positive and fixed, and $p^\alpha$ preserves non-negativity, support, and even symmetry of $p$ around $\boldsymbol{\mu}$. Hence, Eq. (5) closely resembles Eq. (FKL) in structure. In fact, we find that replacing the requirements on $-\log q_{\mathbf{0}}$ in Thms. 1 and 2 with the same conditions on $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ gives rise to more general statements for the $\alpha$-DIV as detailed below.

**Theorem 3.** *Under (A1), $D_\alpha(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is strictly convex.*

**Theorem 4.** *Under (A1), $D_\alpha(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is convex, strictly increasing in $||\boldsymbol{x}||$ and (2) $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure.*

In practice, we find that the requirements on $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ are governed by the chosen $\alpha$ more so than $q_{\mathbf{0}}$. Fig. 4 gives examples where the same location-scale family succeeds and fails to recover the target mean depending on the value of $\alpha$. As a general rule of thumb, $\alpha > 1$ is favorable for exact mean recovery. Fig. 6 shows $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ for varying values of $\alpha$ for common location-scale families. Notably, even for distributions for which $q_{\mathbf{0}}$ itself is not log-concave, such as the Student-t, it can be possible to find values of $\alpha$, such that our sufficient conditions are fulfilled. In App. A.1, we discuss precise guidelines for the value of $\alpha$ for common location-scale families.

### 3.2 Exact recovery of correlations

We now present our results for exact recovery of the correlation matrix. The statements closely resemble the results of Margossian and Saul [46] for the RKL, but we are once again able to alleviate

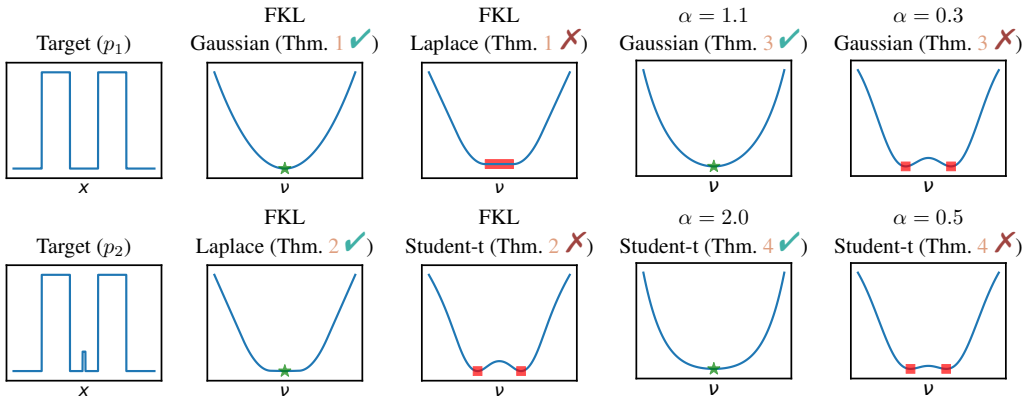



Figure 4: **Our sufficient conditions guarantee a unique global optimum at the true mean of the target. When they are violated, optimization may fail to locate the correct mean.** Illustration of settings that *comply with* (✔) and *violate* (✗) the sufficient conditions provided by our theorems for mean recovery (Sec. 3.1). The first figure depicts the target density $p$. The remaining figures show the divergence between $p$ and $q_{\boldsymbol{\nu}}$ as a function of $\boldsymbol{\nu}$. When our sufficient conditions are met, we expect a unique global minimizer of the divergence at the targets true mean (*indicated via a green star*). If our conditions are violated, we cannot guarantee the former (*indicated via a red square/bar*). The concrete settings are described in Sec. 4.1. We use 5 degrees of freedom for the Student-t.

the assumption that $p$ is (somewhere strictly) log-concave. While we still make the assumption that $p$ is elliptically symmetric, dropping log-concavity opens up a wider range of densities, including heavy-tailed targets or targets with deep valleys around their mean (Fig. 1, third row). This allows us to give guarantees for a wider range of misspecifications. The full proofs are given in App. B.5. We state our guarantees in terms of the matrix $M$ as used in Eq. (3) which is directly related to the correlation matrix of $p$.

**Theorem 5.** *Assume (A2). Moreover, assume that all conditions for either Thm. 1 or Thm. 2 are satisfied, and hence we are guaranteed to recover the target mean. If additionally* $g(||\boldsymbol{x}||) = \log q_{\mathbf{0}}(\boldsymbol{x})$ *is almost everywhere continuously differentiable with* $|g'(0)| < \infty$ *or* $\lim_{u \to 0^+} -g'(u) \geq 0$*, then* $D_{FKL}(p, q_{\boldsymbol{\nu},S})$ *has a unique minimizer at* $(\boldsymbol{\mu}, \gamma^2 M)$ *for some* $\gamma > 0$*.*

**Theorem 6.** *Assume (A2). Moreover, assume that all conditions for either Thm. 3 or Thm. 4 are satisfied, and hence we are guaranteed to recover the target mean. If additionally (i)* $g(||\boldsymbol{x}||) = q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{x})$ *is almost everywhere continuously differentiable with* $|g'(0)| < \infty$ *or* $\lim_{u \to 0^+} g'(u) \geq 0$*, (ii)* $\alpha > 1$*, and (iii)* $-\log q_{\mathbf{0}}$ *is convex, then* $D_\alpha(p, q_{\boldsymbol{\nu},S})$ *has a unique minimizer at* $(\boldsymbol{\mu}, \gamma^2 M)$ *for some* $\gamma > 0$*.*

Notably, our result for the **FKL** Thm. 5 does not require us to put strong additional assumptions on the variational family. The conditions on $g$ are fulfilled for most common location scale families. Note that we replace the condition $|g'(0)| < \infty$ with $\lim_{u \to 0^+} -g'(u) \geq 0$ for densities that do not have a well-defined derivative at the origin. Since our analysis relies on the derivative being increasing and non-negative, the singularity at the origin does not affect our results.

For the **$\alpha$-DIV**, we again obtain a more fine-grained criterion. However, compared to Thm. 3 and Thm. 4, we now explicitly decouple conditions on $\alpha$ and $q_{\mathbf{0}}$. Note that these conditions are consistent with the requirements for mean recovery, albeit stronger: While an appropriately high value of $\alpha$ can guarantee exact mean recovery for a Student-t (as we also show empirically in Fig. 4), we do not provide a guarantee for exact correlation recovery with the Student-t.

## 4 What happens in practice

To illustrate the practical implications of our derived sufficient conditions, we now present examples for which our theoretical results can aid the choice of variational family and value of $\alpha$. Experimental details and target specifications are in App. C. The code to reproduce our experiments can be found at https://github.com/april-tools/vi-under-symmetry.

### 4.1 Symbolic integration for mean recovery

We first consider scenarios for which the exact computation of the divergence is tractable. We construct the following *targets*: We define $p_1$ as a one-dimensional, symmetric mixture of uniform distributions $0.5 \cdot \text{Unif}(-9, -3) + 0.5 \cdot \text{Unif}(3, 9)$. The mean of this distribution ($\mu = 0$) is not contained in the support of $p$, i.e., $\mu \notin \text{supp}(p)$. To create a target $p_2$ for which $\mu \in \text{supp}(p)$, we mix $p_1$ with an additional uniform centered at 0, resulting in $0.99 \cdot p_1(x) + 0.01 \cdot \text{Unif}(-0.3, 0.3)$. Our results are summarized in Fig. 4.

**Thm. 1.** We are guaranteed to obtain a unique minimizer of the FKL at $\nu = \mu$ *if* we choose our variational family such that $-\log q_0$ is strictly convex. The variational family $\mathcal{Q}_1 = \mathcal{N}(\nu, 4)$ complies with this condition, and we obtain a unique minimizer at $\nu = \mu$ in practice. Next, we *violate* the sufficient condition in Thm. 1 by choosing $\mathcal{Q}_2 = \text{Laplace}(\nu, 4)$. We do not obtain a *unique* global minimizer. In fact, the objective obtains the exact same value for every $\nu \in [-3, 3]$. App. D.1 provides the full mathematical explanation for this phenomenon. Intuitively, for $\nu \in [-3, 3]$, we are in the setting where the integral over $\mathcal{H}_+^{(2)}$ (see Fig. 3) equals 0.

**Thm. 2.** As $p_2$ is supported around $\mu$, we do not require strict convexity of $-\log q_0$. As a result, we can reuse $\mathcal{Q}_2 = \text{Laplace}(\nu, 4)$, which indeed gives a unique minimizer of the FKL for $p_2$. To showcase that we *cannot* guarantee a unique minimizer when $-\log q_0$ is not convex, we next use a Student-t variational family with 5 degrees of freedom $\mathcal{Q}_3 = \text{Student}_5(\nu, 1)$. We observe that the FKL has a stationary point at $\mu = 0$ as per Thm. 00 [47], but it is a local maximizer.

**Thm. 3.** We now move to our guarantees for $\alpha$-DIVs. We use $\mathcal{Q}_1 = \mathcal{N}(\nu, 4)$. Note that $q_0$ *itself* being strictly log-concave does not guarantee a unique minimizer; the criterion depends on the interplay between $q_0$ and the chosen value of $\alpha$. We compare $\alpha = 1.1$, for which $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ is strictly convex and optimization succeeds, and $\alpha = 0.3$, which results in a function $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ that is not convex. The second case is not covered by any of our guarantees, and we obtain a local maximizer at $\mu = 0$. Results for further values of $\alpha$ and visualizations of $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ are in App. D.2.1.

**Thm. 4.** Lastly, we discuss Thm. 4, which relaxes Thm. 3 in the sense that $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ is required to be convex, but not necessarily strictly, while imposing the additional condition that $p$ is supported around $\mu$ on a set of positive measure. We find that for common location families, it is usually not the case that $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ is convex but not strictly convex (see Fig. 6). We nevertheless provide an example on the target $p_2$ with $\mathcal{Q}_4 = \text{Student-t}_5(\nu, 4)$ to show that the Student-t can succeed to recover the mean with the $\alpha$-DIV when our conditions are satisfied, despite not being covered by any guarantee when used with the FKL. We show $\alpha = 2.0$ and $\alpha = 0.5$, which correspond to the $\chi^2$-divergence and squared Hellinger distance respectively. As expected, we obtain a unique minimizer at the correct mean for $\alpha = 2.0$. For $\alpha = 0.5$, we instead observe a a local maximum at $\mu$.

### 4.2 Learning under misspecification

We conduct experiments using a standard VI setup based on stochastic gradient-based optimization. The aim of these experiments is to show that whenever our sufficient conditions are satisfied, we can recover the corresponding target statistics with low error, despite the noise introduced by stochastic optimization. We optimize the location and scale parameters simultaneously for all experiments.

**Setup.** All of our experiments in 2D use the Adam optimizer [36] with default parameters and learning rate $10^{-3}$. For the RKL and $\alpha$-DIV, we learn via reparameterization [40] when the target has full support and use REINFORCE [74] otherwise. For the *Ellipse*, we report the RKL with REINFORCE as it was more stable during optimization. For optimizing the FKL, we use i.i.d. samples from $p$ (see App. C for the used sampling schemes). We train for a maximum of $10^4$ steps using $10^5$ samples for each update. For the Student-t, we fix the degrees of freedom to 5. The reported error metrics are computed w.r.t. the checkpoint that achieved the lowest training divergence. Following Margossian and Saul [46], we measure the error of the mean estimate as $\Delta\boldsymbol{\mu} := \frac{1}{d} \sum_{i=1}^{d} |\nu_i - \mu_i| / \max(\sqrt{\mathbb{V}_p[x_i]}, |\mu_i|)$ and the error of the correlation matrix as $\Delta\text{Corr} := \frac{1}{d(d-1)} \sum_{i=1}^{d} \sum_{j \neq i} |\text{Corr}_q[x_i, x_j] - \text{Corr}_p[x_i, x_j]|$.

Below, we introduce a diverse set of targets for which we study VI under misspecification. Similar to Sec. 4.1, we present examples in pairs of cases which are (1) (partially) covered by our guarantees

and (2) not covered by any of our guarantees. Fig. 1 summarizes our insights. We provide results for additional combinations of divergences and variational families in App. D.2.2, including results for the RKL. Tab. 2 gives an overview of our targets and their compatibility with our theorems.

**Results.** On a high level, we observe that whenever our theorems guarantee exact recovery of the mean or correlation matrix in theory, we also observe low error in practice. At the same time, we find that when our conditions are not satisfied, optimization might either succeed or fail to capture the target characteristics. This is not at odds with our statements since we provide sufficient conditions. We now give a more fine-grained discussion of interesting cases for each target.

**MoU.** First, we discuss a two-dimensional mixture of uniforms with disjoint support, akin to the targets used in Sec. 4.1. For the **FKL**, we assess whether (1) a Gaussian correctly locates the mean as indicated via Thm. 1 and (2) whether a Student-t succeeds in recovering the mean despite the lack of a guarantee. Interestingly, we find that both is the case. For the $\boldsymbol{\alpha}$**-DIV**, we compare a Student-t surrogate learned with $\alpha = 2.0$, which complies with Thm. 3, and $\alpha = 0.5$, for which we do not have a guarantee. In both cases, we get a similar fit to the FKL and a good match for the mean.

**GMM.** Next, we target an *even symmetric* GMM, which has support on all of $\mathbb{R}^2$. The target is not elliptically symmetric, and hence we do not expect correlation recovery, except when we optimize a Gaussian with the **FKL** [70]. We observe that also a Student-t distribution optimized with the FKL recovers the mean well, despite the lack of a formal guarantee. Moreover, we find that a Student-t distribution learned via the $\boldsymbol{\alpha}$**-DIV** with $\alpha = 2.0$ closely matches the target mean, as guaranteed by Thms. 3 and 4, while for $\alpha = 0.5$, it fails to capture the mean.

**Ellipse.** Next, we discuss an *elliptically symmetric* target constructed via a squared subtractive mixture model [44, 78]. In this setting, we are guaranteed a unique global minimizer at the *correct mean and correlation* with the **FKL** when optimizing a Gaussian family via Thms. 1 and 2 and Thm. 5. We indeed observe low error in practice. At the same time, also a Student-t distribution learned via the FKL captures the characteristics well, despite the lack of a guarantee. For the $\boldsymbol{\alpha}$**-DIV**, we guarantee exact recovery of mean and correlation with a Gaussian for $\alpha = 2.0$ via Thms. 3 and 4 and Thm. 6. The approximation indeed achieves low error in practice. Interestingly, also pairing $\alpha = 0.5$ with a Student-t recovers the statistics well, despite not fulfilling our sufficient conditions.

**What happens on asymmetric targets?** Lastly, we construct an *asymmetric GMM target* which is neither even symmetric nor elliptically symmetric and therefore is not covered by any of our guarantees. Indeed, we find that all configurations shown in Fig. 1 fail to match the target mean and correlation, except for the Gaussian fit with the FKL, for which we expect mean and correlation recovery regardless of symmetry [70].

**What happens on higher-dimensional targets?** In App. D.2.2, we provide additional results for higher-dimensional versions of the targets *GMM* and *Ellipse* for $d \in \{4, 8, 16\}$. Our main observations prevail: When our sufficient conditions are satisfied, we recover the corresponding target characteristics with low error. We note however that optimizing the $\alpha$-DIV with $\alpha = 2.0$ was challenging in higher dimensions. We found that estimating an informative gradient for the $\alpha$-DIV in the early stages of training was difficult, causing learning to stagnate unless carefully initialized. This is a known problem [23], and how to improve the stability of $\alpha$-DIV optimization in practice is an interesting open question.

## 5 Discussion & future work

In this work, we have extended previous guarantees for robust VI with location-scale families under target symmetries to the FKL and $\alpha$-divergence. Notably, our analysis relaxes the assumption that the target needs to be log-concave, which allows us to provide guarantees for multi-modal targets. We demonstrate the practical impact of our guarantees on a diverse set of target densities with varying support and symmetries, and show how and why optimization can fail when our conditions are violated. Below, we discuss open questions and limitations of our analysis.

**Which target statistics should we match?** Throughout the paper, we have taken for granted that matching the target mean and correlations is desirable. This is true in many practical settings: Under the squared error loss, the posterior mean is the Bayes estimator [25] and hence of interest by itself if the main goal is parameter inference. Moreover, estimating the posterior mean correctly is crucial for

good predictions in Bayesian linear models where exactly capturing the posterior mean ensures that the approximate posterior predictive mean matches the true posterior predictive mean [8].

However, mean-matching solutions can be suboptimal for Bayesian neural networks (BNNs): In multi-modal BNN posteriors, matching the posterior mean can result in an interpolation between equivalent modes with overall low posterior density, resulting in poor predictive performance [24]. In such cases, it might be preferable to capture a posterior mode instead. Lastly, the importance of posterior correlations in BNN posteriors has, for instance, been highlighted by MacKay [45]. However, it is still largely unclear how correlation recovery in parameter space impacts predictive uncertainty [56]. Gaining a deeper understanding of how guarantees in parameter space relate to downstream predictions and decisions is an interesting future direction.

**Generalizing symmetries.** A natural future direction is to analyze more expressive variational families such as mixtures of location-scale families. Another promising opportunity is to investigate broader classes of symmetries that arise in practice: Concurrently to our work, Marks et al. [51] studied statistic recovery under general symmetry groups *under the assumption that the objective has a unique minimizer*. It would be interesting to understand under which conditions these unique minimizers exist and when they can be learned efficiently via stochastic optimization in polynomial time. Lastly, Pozza et al. [60] recently proposed to adapt an existing symmetric VI solution with a skew factor computed from the posterior to reduce the divergence to the target. Further connecting these results to statistic recovery could pave the way for better variational approximations in practice.

**Limitations.** We acknowledge that our analysis has the following limitations: First of all, the conditions we provide are only *sufficient* - investigating whether they are also necessary is interesting future work. Secondly, our guarantees cover the FKL and $\alpha$-DIV, which can be hard to optimize in practice [33, 23]. As we discuss in Sec. 4.2, even when we are guaranteed a unique minimizer of the objective for the $\alpha$-DIV via our theorems, the Monte Carlo estimate of the $\alpha$-DIV can have high variance, especially when the variational approximation is poorly initialized. Moreover, optimizing the FKL w.r.t. an unnormalized target (and without access to i.i.d. target samples) requires the use of self-normalized importance sampling [57] or Markov Chain Monte Carlo (MCMC) [64], which results in a biased estimator. It is unclear whether our guarantees also apply to this setting. Lastly, our guarantees for the $\alpha$-DIV focus on $\alpha > 1$. Concurrently to our work, Margossian et al. [49] have updated their results to provide guarantees for unimodal targets for various $f$-divergences, including the $\alpha$-DIV with $\alpha \in (0, 1)$. Finding guarantees for the $\alpha$-DIV with $\alpha \in (0, 1)$ for more complex targets is, to the best of our knowledge, still an open question.


### Acknowledgements

LZ and AV were supported by the "UNREAL: Unified Reasoning Layer for Trustworthy ML" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC. The authors are grateful to Nicola Branchini and Maximilian Pixner for useful feedback on the draft.


### Contributions

LZ had the initial idea and is responsible for all theoretical derivations, figures and experiments in the paper. LZ wrote the paper with help from AV, who supervised all stages of the project.

### Declaration of LLM usage

We used LLM assistance for coding and debugging. LLM assistance was also useful for literature search and finding references for the convexity arguments used in the paper. All proofs and experiments were designed and executed by the authors.

## References


[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

[2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine

learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM international conference on architectural support for programming languages and operating systems, volume 2*, pages 929–947, 2024.

[3] Oleg Arenz, Philipp Dahlinger, Zihan Ye, Michael Volpp, and Gerhard Neumann. A unified perspective on natural gradient variational inference with gaussian mixture models. *Transactions on Machine Learning Research*, 2023.

[4] François Bertholom, Randal Douc, and François Roueff. Asymptotics of alpha-divergence variational inference algorithms with exponential families. *Advances in Neural Information Processing Systems*, 37:106237–106263, 2024.

[5] Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

[6] Anna Bignami and A De Matteis. A note on sampling from combinations of distributions. *IMA Journal of Applied Mathematics*, 8(1):80–81, 1971.

[7] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006.

[9] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[10] Sam Bowyer, Thomas Heap, and Laurence Aitchison. Using autodiff to estimate posterior moments, marginals and samples. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 394–417. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/bowyer24a.html.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[12] Trevor Campbell and Xinglong Li. Universal boosting variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

[14] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[15] Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[16] Kamélia Daudel and Randal Douc. Mixture weights optimisation for alpha-divergence variational inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4397–4408. Curran Associates, Inc., 2021.

[17] Bernard Delyon and François Portier. Safe adaptive importance sampling. *The Annals of Statistics*, 49(2):885–917, 2021.

[18] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via $\chi$ upper bound minimization. *Advances in Neural Information Processing Systems*, 30, 2017.

[19] Justin Domke. Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning*, pages 2587–2596. PMLR, 2020.

[20] Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. *Advances in neural information processing systems*, 36: 66289–66327, 2023.

[21] Torbjørn Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[22] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

[23] Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3650–3659. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/geffner21a.html.

[24] Yoav Gelberg, Tycho F. A. van der Ouderaa, Mark van der Wilk, and Yarin Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for bayesian neural networks. In *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pages 233–248. PMLR, 29 Jul 2024. URL https://proceedings.mlr.press/v251/gelberg24a.html.

[25] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. chapman and hall/cRc, 1995.

[26] Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness, and variational bayes. *Journal of machine learning research*, 19(51):1–49, 2018.

[27] Thomas Guilmeau, Nicola Branchini, Emilie Chouzenoux, and Víctor Elvira. Adaptive importance sampling for heavy-tailed distributions via $\alpha$-divergence minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3871–3879. PMLR, 2024.

[28] Thomas Guilmeau, Emilie Chouzenoux, and Víctor Elvira. On variational inference and maximum likelihood estimation with the $\lambda$-exponential family. *Foundations of Data Science*, 6 (1):85–123, 2024. doi: 10.3934/fods.2024011.

[29] Peter Hall, Tung Pham, Matt P Wand, and Shen SJ Wang. Asymptotic normality and valid inference for gaussian variational approximation. 2011.

[30] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[31] Alexandra Hotti, Oskar Kviman, Ricky Molén, Víctor Elvira, and Jens Lagergren. Efficient mixture learning in black-box variational inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18972–18991. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hotti24a.html.

[32] Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

[33] Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A Heller, Yian Ma, and Michael I Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence. In *Uncertainty in Artificial Intelligence*, pages 1819–1829. PMLR, 2021.

[34] Anya Katsevich and Philippe Rigollet. On the approximation accuracy of gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409, 2024.

[35] Pasha Khosravi, Yitao Liang, YooJung Choi, and Guy Van den Broeck. What to expect of classifiers? reasoning about logistic regression with missing features. *arXiv preprint arXiv:1903.01620*, 2019.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[37] Anna Korba and François Portier. Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR, 2022.

[38] Samuel Kotz, Tomasz Kozubowski, and Krzystof Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

[39] Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991.

[40] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 18(14):1–45, 2017.

[41] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[42] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[43] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

[44] Lorenzo Loconte, M. Sladek Aleksanteri, Stefan Mengel, Martin Trapp, Arno Solin, Nicolas Gillis, and Antonio Vergari. Subtractive mixture models via squaring: Representation and learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[45] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[46] Charles Margossian and Lawrence K. Saul. Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3466–3474. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/margossian25a.html.

[47] Charles C Margossian and Lawrence K Saul. Generalized guarantees for variational inference in the presence of even and elliptical symmetry. *arXiv preprint arXiv:2511.01064*, 2025.

[48] Charles C Margossian, Loucas Pillaud-Vivien, and Lawrence K Saul. Variational inference for uncertainty quantification: An analysis of trade-offs. *Journal of Machine Learning Research*, 26(202):1–41, 2025.

[49] Charles C Margossian, Isaac E Rankin, and Lawrence K Saul. Generalized guarantees for variational inference in the presence of even and elliptical symmetry. *arXiv preprint arXiv:2511.01064*, 2025.

[50] Jean-Michel Marin, Kerrie Mengersen, and Christian P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

[51] Daniel Marks, Dario Paccagnan, and Mark van der Wilk. Symmetry guarantees statistic recovery in variational inference. *arXiv preprint arXiv:2604.18310*, 2026.

[52] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL https://doi.org/10.7717/peerj-cs.103.

[53] Chirag Modi, Robert Gower, Charles Margossian, Yuling Yao, David Blei, and Lawrence Saul. Variational inference with gaussian score matching. *Advances in Neural Information Processing Systems*, 36:29935–29950, 2023.

[54] Warren Morningstar, Sharad Vikram, Cusuh Ham, Andrew Gallagher, and Joshua Dillon. Automatic differentiation variational inference with mixtures. In *International Conference on Artificial Intelligence and Statistics*, pages 3250–3258. PMLR, 2021.

[55] Radford M Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods: Seattle, 1991*, pages 197–211. Springer, 1992.

[56] James Odgers, Ben Riegler, Siddharth Swaroop, and Vincent Fortuin. Gaussian mean field variational inference can overestimate predictive variance. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 94144–94165. PMLR, 06–11 Jul 2026. URL https://proceedings.mlr.press/v306/odgers26a.html.

[57] Art B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

[58] Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588. PMLR, 2018.

[59] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[60] Francesco Pozza, Daniele Durante, and Botond Szabo. Skew-symmetric approximations of posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf082, 2026.

[61] Tom Rainforth, Adam Golinski, Frank Wood, and Sheheryar Zaidi. Target–aware bayesian inference: how to beat optimal conventional estimators. *Journal of Machine Learning Research*, 21(88):1–54, 2020.

[62] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[63] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

[64] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 2004.

[65] Walter Rudin. Real and complex analysis. 1974.

[66] The april Lab. cirkit: a python framework to build, learn and reason about probabilistic circuits and tensor networks, October 2024. URL https://github.com/april-tools/cirkit.

[67] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[68] Marcin Tomczak, Siddharth Swaroop, and Richard Turner. Efficient low rank gaussian variational inference for neural networks. *Advances in Neural Information Processing Systems*, 33: 4610–4622, 2020.

[69] AW vd Vaart. Asymptotic statistics. *Cambridge series in statistical and probabilistic mathematics*, 1998.

[70] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[71] Yixin Wang and David Blei. Variational bayes under model misspecification. *Advances in Neural Information Processing Systems*, 32, 2019.

[72] Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

[73] Larry Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.

[74] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[75] Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*, 2012.

[76] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. $\alpha$-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

[77] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

[78] Lena Zellinger, Nicola Branchini, Lennert De Smet, Víctor Elvira, Nikolay Malkin, and Antonio Vergari. How to approximate inference with subtractive mixture models. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[79] Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, 2020.

Table 1: Base distributions of common location-scale families. For the multivariate Laplace distribution, we report its symmetric formulation [21]. $K_b$ denotes the modified Bessel function of the second kind [1, Ch. 9], $\Gamma$ denotes the Gamma function, and DoF refers to degrees of freedom. The column "sufficient $\alpha$" gives a value of $\alpha$ such that $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is convex.

| | $q_{\mathbf{0}}(\boldsymbol{x})$ | $-\log q_{\mathbf{0}}$ convex | $-\log q_{\mathbf{0}}$ strictly convex | $-\log q_{\mathbf{0}}$ str. incr. in $\|\|\boldsymbol{x}\|\|$ | $q_{\mathbf{0}}$ spher. symm. | sufficient $\alpha$ |
|---|---|---|---|---|---|---|
| Gaussian | $(2\pi)^{-d/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)$ | ✔ | ✔ | ✔ | ✔ | $\alpha > 1$ |
| Logistic ($d=1$) | $\frac{\exp(-x)}{(1+\exp(-x))^2}$ | ✔ | ✔ | ✔ | ✔ | $\alpha > 1$ |
| Laplace ($d=1$) | $\frac{1}{2}\exp(-\|x\|)$ | ✔ | ✗ | ✔ | ✔ | $\alpha > 1$ |
| Laplace ($d>1$) | $\frac{2}{(2\pi)^{d/2}} \left(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)^{\frac{b}{2}} K_b\left(\sqrt{2\boldsymbol{x}^T\boldsymbol{x}}\right)$, with $b=\frac{2-d}{2}$ | ✗ | ✗ | ✔ | ✔ | / |
| Student-t (DoF $= v$) | $\frac{\Gamma((v+d)/2)}{\Gamma(v/2)v^{d/2}\pi^{d/2}} \left[1+\frac{1}{v}\boldsymbol{x}^T\boldsymbol{x}\right]^{-(v+d)/2}$ | ✗ | ✗ | ✔ | ✔ | $\alpha \geq 1 + 2/(v+d)$ |
| Cauchy | $\frac{\Gamma((1+d)/2)}{\Gamma(1/2)\pi^{d/2}} \left[1+\boldsymbol{x}^T\boldsymbol{x}\right]^{-(1+d)/2}$ | ✗ | ✗ | ✔ | ✔ | $\alpha \geq 1 + 2/(1+d)$ |
| Uniform ($d=1$) | $\begin{cases}\frac{1}{2}, x \in [-1,1] \\ 0, \text{else}\end{cases}$ | ✔ | ✗ | ✗ | ✔ | / |

# A Definitions

## A.1 Common location-scale families

Below, we provide definitions for common location-scale families. Remember that, by definition, members of a location-scale family can be expressed as

$$q_{\boldsymbol{\nu},S}(\boldsymbol{x}) = q_{\mathbf{0}}\left(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\right)|S|^{-\frac{1}{2}}, \tag{6}$$

where $\boldsymbol{\nu} \in \mathbb{R}^d$ is the location parameter and $S \in \mathbb{S}^d_{++}$ is a positive definite scale matrix.

Tab. 1 gives an overview of the base distributions $q_{\mathbf{0}}$ of commonly used location-scale families and whether they fulfill properties related to our theorems. In Fig. 5, we show how $-\log q_{\mathbf{0}}$, which is central to our Thm. 1 and Thm. 2, behaves for common location families. We fix the scale to 1. In Fig. 6, we provide illustrations of $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$, as used in our Thm. 3 and Thm. 4, for varying values of $\alpha$. We find that choosing a high values of $\alpha$ results in functions $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ that comply with our theorems for mean recovery. We derive precise values of $\alpha$ that result in convex $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ below. The corresponding values can be found in the column "sufficient $\alpha$" in Tab. 1.

**Sufficient value of $\alpha$.** For families that have log-concave base distributions, such as the **Gaussian**, **Logistic** ($d=1$), and **Laplace** ($d=1$), it is sufficient to choose $\alpha > 1$. First, note that $\frac{1}{\alpha(\alpha-1)}$ is positive in this case, while $1-\alpha$ is negative. Hence, we need to check the behavior of $q_{\mathbf{0}}^{1-\alpha}$ for a negative exponent. We know that $\log q_{\mathbf{0}}^{1-\alpha} = (1-\alpha)\log q_{\mathbf{0}}$ is convex. As a result, also $q_{\mathbf{0}}^{1-\alpha}$ is convex since every log-convex function is convex [11, Ch. 3.5.1]. The **Student-t family**, which corresponds to the the **Cauchy** for $v=1$, is interesting since we need to "counteract" its heavy tails by choosing a sufficiently large $\alpha$. It is easy to see that it suffices to choose $\alpha \geq 1 + \frac{2}{v+d}$. Notably, the setting $\alpha = 1 + \frac{2}{v+d}$ has also been studied for the Student-t family by Guilmeau et al. [28].

Consider that the base distribution of the Student-t distribution (up to constants) amounts to

$$q_{\mathbf{0}}(\boldsymbol{x}) \propto \left[1 + \frac{1}{v}\boldsymbol{x}^T\boldsymbol{x}\right]^{-(v+d)/2} \tag{7}$$

Taking the above expression to the power $1-\alpha$ with $\alpha = 1 + \frac{2}{v+d}$ results in

$$\left(q_{\mathbf{0}}(\boldsymbol{x})\right)^{1-\alpha} \propto 1 + \frac{1}{v}\boldsymbol{x}^T\boldsymbol{x}, \tag{8}$$

which is strictly convex in $\boldsymbol{z}$ since $v > 0$. Moreover, the additional multiplicative factor $\frac{1}{\alpha(\alpha-1)}$ is positive for $1 + \frac{2}{v+d}$, ensuring that $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is strictly convex.

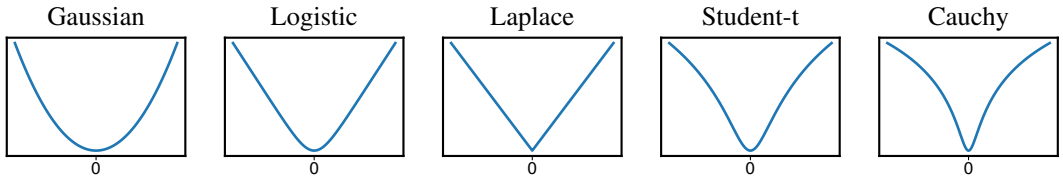



Figure 5: Visualization of $w(x) = -\log q_{\mathbf{0}}(x)$ for common location families. We depict the Student-t distribution with $5$ degrees of freedom.

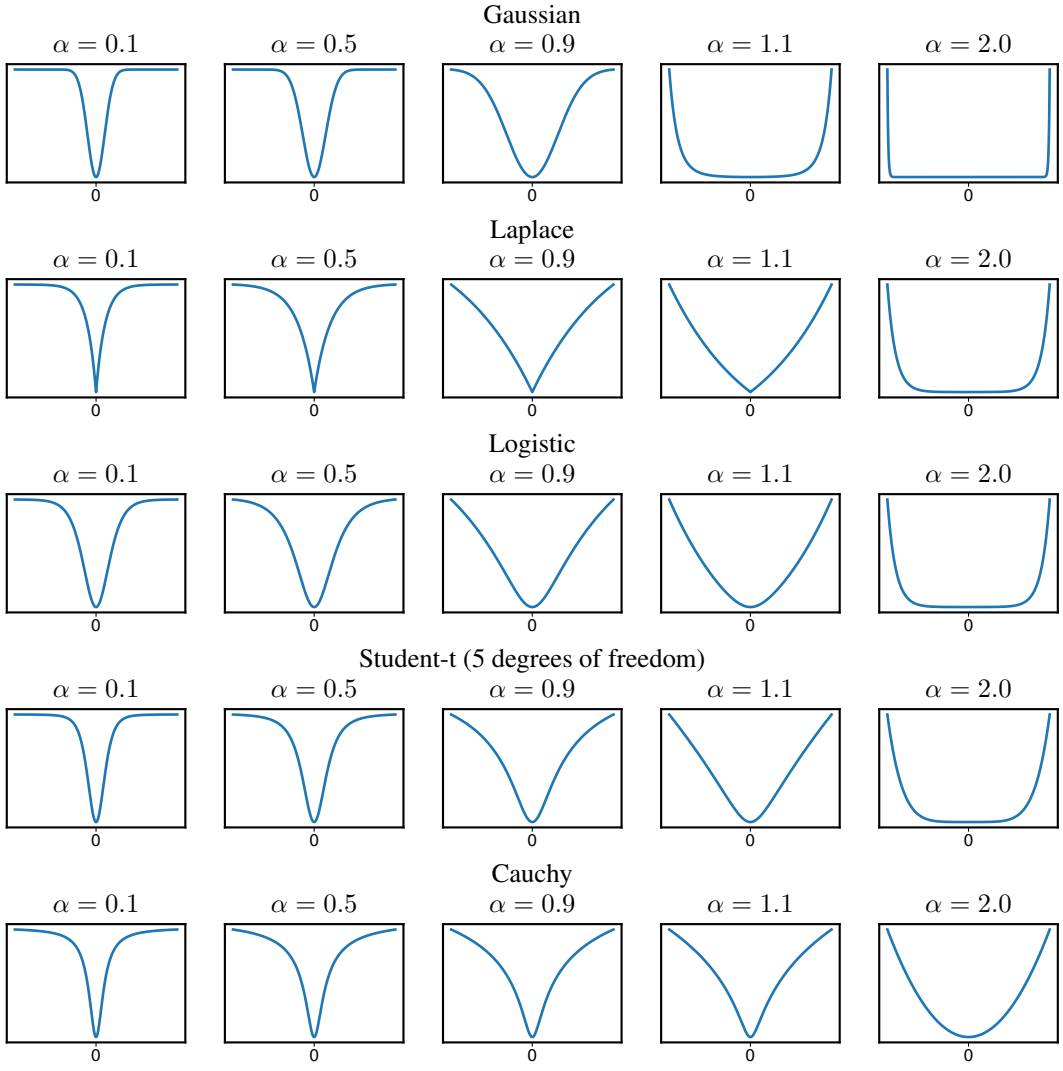



Figure 6: Visualization of $w(x) = \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}(x)$ for common location families. Choosing $\alpha > 1$ tends to result in strictly convex $w$, complying with Thm. 3 (except for the heavy-tailed Student-t and Cauchy, which require an even larger value of $\alpha$). For $\alpha$ close to $0$, the resulting weighting functions are not convex and hence do not comply with our theorems. We do not observe a clear case where $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is convex but not strictly.

### A.2 $f$-divergences

An $f$-divergence [63] is given as

$$D_f(p,q) = \int f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) q(\boldsymbol{x}) d\boldsymbol{x} \tag{9}$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex, and $f(1) = 0$, and $\lim_{u\to 0^+} f(u) = f(0)$. Many common divergences can be expressed as an $f$-divergence, including the RKL, FKL, and $\alpha$-DIV (see [47] for more examples). Beyond standard assumptions on $f$, the analysis by Margossian and Saul [47] requires $f$ and $\varphi$ to be differentiable.

#### A.2.1 Reverse KL divergence

Choosing $f(\boldsymbol{u}) = -\log(\boldsymbol{u})$ recovers the RKL, given as

$$D_{\text{RKL}}(p,q) = \int \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) q(\boldsymbol{x}) d\boldsymbol{x}. \tag{RKL}$$

Notably, the $f$ defining the RKL, and its associated $\varphi(\boldsymbol{u}) = -\boldsymbol{u}$, fulfill the conditions for a unique minimizer specified in Thm. 00 (under the additional condition that **(A1)** holds and the target is somewhere strictly log-concave).

#### A.2.2 Forward KL divergence

Choosing $f(\boldsymbol{u}) = \boldsymbol{u}\log(\boldsymbol{u})$ recovers the FKL, given as

$$D_{\text{FKL}}(p,q) = \int \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) p(\boldsymbol{x}) d\boldsymbol{x}. \tag{FKL}$$

#### A.2.3 $\alpha$-divergence

Choosing $f(\boldsymbol{u}) = \frac{\boldsymbol{u}^\alpha - 1}{\alpha(\alpha-1)}$ recovers the $\alpha$-divergence [14],

$$D_\alpha(p,q) = \frac{1}{\alpha(\alpha-1)} \int \left[\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^\alpha - 1\right] q(\boldsymbol{x}) d\boldsymbol{x}, \tag{$\alpha$-DIV}$$

where $\alpha > 0$ and $\alpha \neq 1$.

## B Proofs

In this section, we present the proofs for our theorems given in Sec. 3. First, we prove some useful lemmas (App. B.1) to simplify the proofs of our main results.

### B.1 Lemmas

**Lemma 1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strictly convex. Further, let $g(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$ be an injective affine mapping with $A \in \mathbb{R}^{d\times d}$, $\boldsymbol{b} \in \mathbb{R}^d$. Then $f \circ g$ is strictly convex.*

**Proof.** As convexity is preserved under composition with an affine mapping [11, Ch. 3.2.2.], we have $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d, \boldsymbol{x}_1 \neq \boldsymbol{x}_2, \forall \lambda \in (0,1)$:

$$f(\lambda g(\boldsymbol{x}_1) + (1-\lambda) g(\boldsymbol{x}_2)) \leq \lambda f(g(\boldsymbol{x}_1)) + (1-\lambda) f(g(\boldsymbol{x}_2)). \tag{10}$$

Further, $f$ is strictly convex, i.e., $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d, \boldsymbol{x}_1 \neq \boldsymbol{x}_2, \forall \lambda \in (0,1)$:

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) < \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2). \tag{11}$$

Moreover, $g$ is injective, i.e., $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d : \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \implies g(\boldsymbol{x}_1) \neq g(\boldsymbol{x}_2)$. Hence, the inequality in Eq. (10) is strict.

**Lemma 2.** *Let $f_{\boldsymbol{x}} : \mathbb{R}^d \to \mathbb{R}$ be strictly convex in $\boldsymbol{\nu}$. Further, let $p(\boldsymbol{x}) \geq 0 \; \forall \; \boldsymbol{x} \in \mathbb{R}^d$. Then,*

$$h(\boldsymbol{\nu}) = \int_{\mathbb{R}^d} f_{\boldsymbol{x}}(\boldsymbol{\nu}) p(\boldsymbol{x}) d\boldsymbol{x}$$

*is strictly convex in $\boldsymbol{\nu}$.*

**Proof.** First of all, note that it suffices to take the integral over the support of $p$, as $p(\boldsymbol{x}) = 0$ otherwise by definition. Hence, we have

$$h(\boldsymbol{\nu}) = \int_{\mathbb{R}^d} f_{\boldsymbol{x}}(\boldsymbol{\nu}) p(\boldsymbol{x}) d\boldsymbol{x} = \int_{\text{supp}(p)} f_{\boldsymbol{x}}(\boldsymbol{\nu}) p(\boldsymbol{x}) d\boldsymbol{x}. \tag{12}$$

By strict convexity of $f_{\boldsymbol{x}}$ in $\boldsymbol{\nu}$ we have $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathbb{R}^d, \boldsymbol{\nu}_1 \neq \boldsymbol{\nu}_2, \forall \lambda \in (0,1)$:

$$f_{\boldsymbol{x}}(\lambda \boldsymbol{\nu}_1 + (1-\lambda)\boldsymbol{\nu}_2) < \lambda f_{\boldsymbol{x}}(\boldsymbol{\nu}_1) + (1-\lambda) f_{\boldsymbol{x}}(\boldsymbol{\nu}_2). \tag{13}$$

Now, for $\boldsymbol{x} \in \text{supp}(p)$, it holds that $p(\boldsymbol{x}) > 0$, and hence multiplying Eq. (10) with $p(\boldsymbol{x})$ on both sides preserves strict convexity, i.e., $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathbb{R}^d, \boldsymbol{\nu}_1 \neq \boldsymbol{\nu}_2, \forall \lambda \in (0,1)$:

$$f_{\boldsymbol{x}}(\lambda \boldsymbol{\nu}_1 + (1-\lambda)\boldsymbol{\nu}_2) p(\boldsymbol{x}) < \lambda f_{\boldsymbol{x}}(\boldsymbol{\nu}_1) p(\boldsymbol{x}) + (1-\lambda) f_{\boldsymbol{x}}(\boldsymbol{\nu}_2) p(\boldsymbol{x}). \tag{14}$$

Integrating both sides of Eq. (14), we obtain $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathbb{R}^d, \boldsymbol{\nu}_1 \neq \boldsymbol{\nu}_2, \forall \lambda \in (0,1)$:

$$\begin{aligned}
h(\lambda \boldsymbol{\nu}_1 + (1-\lambda)\boldsymbol{\nu}_2) &= \int_{\text{supp}(p)} f_{\boldsymbol{x}}(\boldsymbol{\nu}_1 + (1-\lambda)\boldsymbol{\nu}_2) p(\boldsymbol{x}) d\boldsymbol{x} && (15)\\
&< \lambda \int_{\text{supp}(p)} f_{\boldsymbol{x}}(\boldsymbol{\nu}_1) p(\boldsymbol{x}) d\boldsymbol{x} + (1-\lambda) \int_{\text{supp}(p)} f_{\boldsymbol{x}}(\boldsymbol{\nu}_2) p(\boldsymbol{x}) d\boldsymbol{x} && (16)\\
&= \lambda h(\boldsymbol{\nu}_1) + (1-\lambda) h(\boldsymbol{\nu}_2), && (17)
\end{aligned}$$

which proves strict convexity of $h$ in $\boldsymbol{\nu}$.

### B.2 Mean recovery

In this section, we present the proofs for our guarantees for exact mean recovery.

We now restate our core assumptions.

**(A1).** *Let $q_{\boldsymbol{\nu}}$ be a location family with fixed scale matrix $S$ and let $p$ be the target. We make the following assumptions: The base distribution $q_{\mathbf{0}}$ is even symmetric around $\mathbf{0}$. The target $p$ is even symmetric around its mean $\boldsymbol{\mu}$. Lastly, $D_f(p, q_{\boldsymbol{\nu}})$ is well-defined for all values of $\boldsymbol{\nu}$, and $\boldsymbol{\mu}$ is finite.*

Throughout our proofs, we use that we can simplify both the FKL and $\alpha$-DIV when treating $S$ as a constant, as detailed below.

$$\begin{aligned}
D_{\text{FKL}}(p, q_{\boldsymbol{\nu}}) &= \int \log\Big(\frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu}}(\boldsymbol{x})}\Big) p(\boldsymbol{x}) d\boldsymbol{x} && (18)\\
&\sim \int -\log q_{\boldsymbol{\nu}}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} && (\text{Drop } -H(p))\\
&= \int -\log\Big(q_{\mathbf{0}}\Big(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\Big)|S|^{-\frac{1}{2}}\Big) p(\boldsymbol{x}) d\boldsymbol{x} && (\text{Plug in } q_{\boldsymbol{\nu}})\\
&= \int -\log\Big(q_{\mathbf{0}}\Big(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\Big)\Big) p(\boldsymbol{x}) d\boldsymbol{x} - \log|S|^{-\frac{1}{2}} && (\text{Pull out } -\log|S|^{-\frac{1}{2}})\\
&\sim \int w\Big(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\Big) p(\boldsymbol{x}) d\boldsymbol{x}. && (\text{Drop } -\log|S|^{-\frac{1}{2}} \text{ and set } w(\boldsymbol{z}) := -\log q_{\mathbf{0}}(\boldsymbol{z}))
\end{aligned}$$

where $w(\boldsymbol{z}) := -\log q_{\mathbf{0}}(\boldsymbol{z})$.

$$
\begin{aligned}
D_\alpha(p, q_{\boldsymbol{\nu}}) &= \frac{1}{\alpha(\alpha-1)} \int \Big[ \Big( \frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu}}(\boldsymbol{x})} \Big)^\alpha - 1 \Big] q_{\boldsymbol{\nu}}(\boldsymbol{x}) d\boldsymbol{x} && (19) \\
&= \int \frac{1}{\alpha(\alpha-1)} q_{\boldsymbol{\nu}}^{1-\alpha}(\boldsymbol{x}) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} - \frac{1}{\alpha(\alpha-1)} && (20) \\
&\sim \int \frac{1}{\alpha(\alpha-1)} q_{\boldsymbol{\nu}}^{1-\alpha}(\boldsymbol{x}) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} && (\text{Drop } -\tfrac{1}{\alpha(\alpha-1)}) \\
&= \int \frac{1}{\alpha(\alpha-1)} \Big( q_{\mathbf{0}} \Big( S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu}) \Big) |S|^{-\frac{1}{2}} \Big)^{1-\alpha} p^\alpha(\boldsymbol{x}) d\boldsymbol{x} && (\text{Plug in } q_{\boldsymbol{\nu}}) \\
&= |S|^{-\frac{1-\alpha}{2}} \int \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha} \Big( S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu}) \Big) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} && (21) \\
&\sim \int \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha} \Big( S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu}) \Big) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} && (\text{Drop } |S|^{-\frac{1-\alpha}{2}} > 0) \\
&:= \int w_\alpha(S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu})) p^\alpha(\boldsymbol{x}) d\boldsymbol{x}, && (\text{Set } w_\alpha(\boldsymbol{z}) := \tfrac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{z}))
\end{aligned}
$$

where $w_\alpha(\boldsymbol{z}) := \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{z})$. Note that we do not omit $\frac{1}{\alpha(\alpha-1)}$, as it can be positive or negative depending on the chosen value of $\alpha$.

We now prove our main theorems presented in Sec. 3.1. We assume throughout that **(A1)** holds.

### B.3 Proof of Theorem 1 and Theorem 3

We first prove Thm. 1 and Thm. 3 as they share analogous reasoning.

#### B.3.1 Proof of Theorem 1

**Theorem 1.** *Under (A1), $D_{FKL}(p, q_{\boldsymbol{\nu}})$ has a unique minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $-\log q_{\mathbf{0}}$ is strictly convex.*

**Proof.** Thm. 00 tells us that $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ has a stationary point at $\boldsymbol{\nu} = \boldsymbol{\mu}$. Therefore, if $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ (Eq. (FKL)) is strictly convex in $\boldsymbol{\nu}$, the minimizer will be unique and located at $\boldsymbol{\mu}$. Note that $w(\boldsymbol{z}) := -\log(q_{\mathbf{0}}(\boldsymbol{z}))$ is strictly convex in the setting of Thm. 1 as this is our additional sufficient condition. Moreover, $g_{\boldsymbol{x}}(\boldsymbol{\nu}) := S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu})$ is an affine mapping of $\boldsymbol{\nu}$. Since $S \in \mathbb{R}^{d \times d}$ is positive definite (see Def. 1), so is $S^{-\frac{1}{2}}$, and hence $g$ is bijective [5, Ch. 4]. As a result, $w(g_{\boldsymbol{x}}(\boldsymbol{\nu}))$ is strictly convex in $\boldsymbol{\nu}$ by Lemma 1. Moreover, since $p$ is a density, we have $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ and hence $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}}) = \int_{\mathbb{R}^d} w(g_{\boldsymbol{x}}(\boldsymbol{\nu})) p(\boldsymbol{x}) d\boldsymbol{x} + C$ is strictly convex in $\boldsymbol{\nu}$ by Lemma 2. As a result, we have a unique global minimizer of $D_{\text{FKL}}(p, q_{\boldsymbol{\nu}})$ at $\boldsymbol{\nu} = \boldsymbol{\mu}$.

#### B.3.2 Proof of Theorem 3

**Theorem 3.** *Under (A1), $D_\alpha(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ is strictly convex.*

**Proof.** We can use the same reasoning as for proving Thm. 1. Thm. 00 tells us that $D_\alpha(p, q_{\boldsymbol{\nu}})$ has a stationary point at $\boldsymbol{\nu} = \boldsymbol{\mu}$. The function $w(\boldsymbol{z}) := \frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{z})$ is strictly convex in the setting of Thm. 1 as this is our additional sufficient condition. Moreover, $g_{\boldsymbol{x}}(\boldsymbol{\nu}) := S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu})$ is an affine mapping of $\boldsymbol{\nu}$. Since $S \in \mathbb{R}^{d \times d}$ is positive definite (see Def. 1), so is $S^{-\frac{1}{2}}$, and hence $g$ is bijective [5, Ch. 4]. Hence, we conclude $w(g_{\boldsymbol{x}}(\boldsymbol{\nu}))$ is strictly convex in $\boldsymbol{\nu}$ by applying Lemma 1. Moreover, since $p$ is a density, we have $p^\alpha(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \text{supp}(p)$ and for any $\alpha \in \mathbb{R}$, and hence $D_\alpha(p, q_{\boldsymbol{\nu}}) \sim \int_{\mathbb{R}^d} w(g_{\boldsymbol{x}}(\boldsymbol{\nu})) p^\alpha(\boldsymbol{x}) d\boldsymbol{x}$ is strictly convex in $\boldsymbol{\nu}$ by Lemma 2. As a result, we have a unique global minimizer of $D_\alpha(p, q_{\boldsymbol{\nu}})$ at $\boldsymbol{\nu} = \boldsymbol{\mu}$.

### B.4 Proof of Theorem 2 and Theorem 4

Next, we provide a proof for and Thm. 2 and Thm. 4. We prove the following more general statement, which is agnostic to the specific form of $w$ and hence can be applied to both the FKL and $\alpha$-DIV.

**Theorem 7.** *Let $q_{\boldsymbol{\nu}}$ be a location family with a base distribution that is even symmetric around the origin (in line with (A1)). Moreover, let $I : \mathbb{R}^d \to \mathbb{R}_0^+$ be even symmetric around $\boldsymbol{\mu}$. Let $w = h \circ q_0$ with $h : \mathbb{R}_0^+ \to \mathbb{R}$. Lastly, assume $D(p, q_{\boldsymbol{\nu}})$ is well-defined for all values of $\boldsymbol{\nu}$, and $\boldsymbol{\mu}$ is finite. Then, $D(p, q_{\boldsymbol{\nu}}) = \int w(S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu}))I(\boldsymbol{x})d\boldsymbol{x}$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if **additionally***

- *$w$ is convex and strictly increasing in $||\boldsymbol{x}||$*
- *$I$ is supported around $\boldsymbol{\mu}$ on a set of positive measure*

**Proof.** Let $\boldsymbol{\nu}' \in \mathbb{R}^d$ and $\boldsymbol{\nu}' \neq \mathbf{0}$. We want to show that $\forall \boldsymbol{\nu}' : D(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}) - D(p, q_{\boldsymbol{\mu}}) > 0$.

Let $\boldsymbol{\tau} := \boldsymbol{x} - \boldsymbol{\mu}$ and hence $\boldsymbol{x} = \boldsymbol{\tau} + \boldsymbol{\mu}$.

$$D\left(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}\right) - D\left(p, q_{\boldsymbol{\mu}}\right) = \int_{\mathbb{R}^d} \left[w\left(S^{-\frac{1}{2}}\left(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{\nu}'\right)\right) - w\left(S^{-\frac{1}{2}}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right)\right] I\left(\boldsymbol{x}\right) d\boldsymbol{x} \tag{22}$$

$$= \int_{\mathbb{R}^d} \left[w\left(S^{-\frac{1}{2}}\left(\boldsymbol{\tau} - \boldsymbol{\nu}'\right)\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I\left(\boldsymbol{\tau} + \boldsymbol{\mu}\right) d\boldsymbol{\tau}. \tag{23}$$

Under the assumption that $w$ is strictly increasing in $||\boldsymbol{x}||$, it holds that

$$w\left(S^{-\frac{1}{2}}\left(\boldsymbol{\tau} - \boldsymbol{\nu}'\right)\right) \geq w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \text{ iff } ||S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')|| \geq ||S^{-\frac{1}{2}}\boldsymbol{\tau}||.$$

This condition corresponds to a half-space defined w.r.t. the hyperplane

$$||S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')|| - ||S^{-\frac{1}{2}}\boldsymbol{\tau}|| = \mathbf{0}.$$

Let $(\cdot)$ denote the inner product. It holds that

$$||S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')|| \geq ||S^{-\frac{1}{2}}\boldsymbol{\tau}|| \iff \boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' \leq \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}.$$

The above follows immediately from the properties of the inner product [39, Chapter 3]:

$$||S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')||^2 \geq ||S^{-\frac{1}{2}}\boldsymbol{\tau}||^2 \tag{24}$$

$$||S^{-\frac{1}{2}}\boldsymbol{\tau}||^2 - 2\boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' + ||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2 \geq ||S^{-\frac{1}{2}}\boldsymbol{\tau}||^2 \quad \text{(expanded quadratic form)}$$

$$-2\boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' \geq -||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2 \tag{25}$$

$$\boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' \leq \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}. \tag{26}$$

We now define a partition of the integration domain, given by

$$\mathcal{H}_1 := \left\{\boldsymbol{\tau} : \boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' \leq \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}\right\}, \text{ corresponding to } w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \geq 0$$

$$\mathcal{H}_2 := \left\{\boldsymbol{\tau} : \boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' > \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}\right\}, \text{ corresponding to } w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) < 0.$$

Fig. 7 provides an illustration. Using this partition, we obtain

$$D\left(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}\right) - D\left(p, q_{\boldsymbol{\mu}}\right) = \int_{\mathbb{R}^d} \left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau} \tag{27}$$

$$= \int_{\mathcal{H}_1} \left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau} \tag{28}$$

$$+ \int_{\mathcal{H}_2} \left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau}. \tag{29}$$

For Thm. 7 to hold, we hence need to show that

$$\int_{\mathcal{H}_1} \left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau} \tag{30}$$

$$\geq -\int_{\mathcal{H}_2} \left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau}. \tag{31}$$

We now further split $\mathcal{H}_1$ into disjoint regions defined as

$$\mathcal{H}_3 := \left\{ \boldsymbol{\tau} : \boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' < -\frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2} \right\}$$

$$\mathcal{H}_4 := \left\{ \boldsymbol{\tau} : -\frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2} \leq \boldsymbol{\tau} \cdot S^{-1}\boldsymbol{\nu}' \leq \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2} \right\}.$$

The rightmost plot in Fig. 7 illustrates the resulting partition. Using the above, we arrive at the following condition:

$$\int_{\mathcal{H}_3} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau} \tag{32}$$

$$+ \int_{\mathcal{H}_4} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau} \tag{33}$$

$$\geq -\int_{\mathcal{H}_2} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu}) d\boldsymbol{\tau}. \tag{34}$$

Note that $\int_{\mathcal{H}_4} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau} \geq 0$ since $\mathcal{H}_4 \subset \mathcal{H}_1$. Hence, it suffices to show that

$$\int_{\mathcal{H}_3} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau} \tag{35}$$

$$\geq -\int_{\mathcal{H}_2} \left[ w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau} \tag{36}$$

$$= \int_{\mathcal{H}_2} \left[ -w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\nu}')\right) + w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right) \right] I(\boldsymbol{\tau}+\boldsymbol{\mu})\, d\boldsymbol{\tau}. \tag{37}$$

To see why the above holds, pick any point $\boldsymbol{\eta} \in \mathcal{H}_3$. By definition of $\mathcal{H}_3$, $\boldsymbol{\eta} \cdot S^{-1}\boldsymbol{\nu}' < -\frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}$. As a result, it holds for the mirrored point $\boldsymbol{\omega} := -\boldsymbol{\eta}$ that $\boldsymbol{\omega} \cdot S^{-1}\boldsymbol{\nu}' > \frac{||S^{-\frac{1}{2}}\boldsymbol{\nu}'||^2}{2}$, which in turn implies $\boldsymbol{\omega} \in \mathcal{H}_2$. Moreover, by even symmetry of $I$ around $\boldsymbol{\mu}$, $I(\boldsymbol{\mu}+\boldsymbol{\eta}) = I(\boldsymbol{\mu}-\boldsymbol{\eta}) = I(\boldsymbol{\mu}+\boldsymbol{\omega})$.

To prove the inequality in Eq. (36) above, we show that

$$w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}-\boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\eta}\right) \geq -w\left(S^{-\frac{1}{2}}(-\boldsymbol{\eta}-\boldsymbol{\nu}')\right) + w\left(-S^{-\frac{1}{2}}\boldsymbol{\eta}\right). \tag{38}$$

We use the even symmetry of $q_{\mathbf{0}}$ around the origin to rewrite

$$w\left(S^{-\frac{1}{2}}(-\boldsymbol{\eta}-\boldsymbol{\nu}')\right) = w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}+\boldsymbol{\nu}')\right)$$

and

$$w\left(-S^{-\frac{1}{2}}\boldsymbol{\eta}\right) = w\left(S^{-\frac{1}{2}}\boldsymbol{\eta}\right).$$

Using the above, we can express Eq. (38) as

$$w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}-\boldsymbol{\nu}')\right) + w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}+\boldsymbol{\nu}')\right) \geq 2w\left(S^{-\frac{1}{2}}\boldsymbol{\eta}\right). \tag{39}$$

Finally, to show Eq. (39), we use the assumed convexity of $w$: $\forall \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^d, \forall \lambda \in [0,1]$ :

$$w(\lambda \boldsymbol{z}_1 + (1-\lambda)\boldsymbol{z}_2) \leq \lambda w(\boldsymbol{z}_1) + (1-\lambda) w(\boldsymbol{z}_2).$$

Setting $\lambda = 0.5$, $\boldsymbol{z}_1 = S^{-\frac{1}{2}}(\boldsymbol{\eta}-\boldsymbol{\nu}')$, $\boldsymbol{z}_2 = S^{-\frac{1}{2}}(\boldsymbol{\eta}+\boldsymbol{\nu}')$, we obtain

$$w\left(\frac{S^{-\frac{1}{2}}(\boldsymbol{\eta}-\boldsymbol{\nu}') + S^{-\frac{1}{2}}(\boldsymbol{\eta}+\boldsymbol{\nu}')}{2}\right) = w\left(S^{-\frac{1}{2}}\boldsymbol{\eta}\right) \leq \frac{1}{2}w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}-\boldsymbol{\nu}')\right) + \frac{1}{2}w\left(S^{-\frac{1}{2}}(\boldsymbol{\eta}+\boldsymbol{\nu}')\right),$$

which directly proves Eq. (39). We have hence shown that $\forall\, \boldsymbol{\nu}' : D(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}) - D(p, q_{\boldsymbol{\mu}}) \geq 0$.

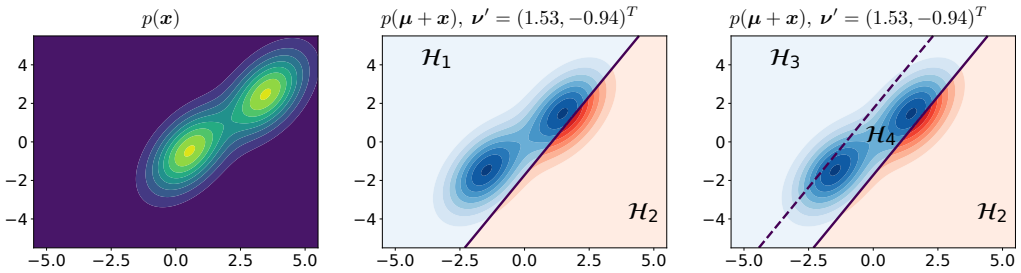



Figure 7: Illustration of the domain partitioning used for proving Thm. 7. **(1)** A bimodal target density $p$ that is even symmetric around $\boldsymbol{\mu} = (2, 1)^T$. **(2)** The shifted target $p(\boldsymbol{\mu} + \boldsymbol{x})$, which is even symmetric around $\mathbf{0}$. Areas where $w\left(S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu}')\right) \geq w\left(S^{-\frac{1}{2}}\boldsymbol{x}\right)$ are depicted in blue, the rest in red. These areas correspond to $\mathcal{H}_1$ and $\mathcal{H}_2$ in the proof. **(3)** Additional partition of $\mathcal{H}_1$ into $\mathcal{H}_3$ and $\mathcal{H}_4$. In the above examples, $w$ is the negative log-density of a multivariate standard Gaussian.

It remains to show that this inequality is strict under the additional condition that $I$ is supported around $\boldsymbol{\mu}$ on a set of positive measure. In particular, this condition ensures that $\int_{\mathcal{H}_4}\left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau} > 0$. First note that $I(\boldsymbol{\tau} + \boldsymbol{\mu})$ is centered at the origin. Hence, if $I$ is supported around $\boldsymbol{\mu}$, $I(\boldsymbol{\tau} + \boldsymbol{\mu})$ is supported around $\mathbf{0}$. Moreover, for any assignment of $\boldsymbol{\nu}'$, we have $\mathbf{0} \in \mathcal{H}_4$ by definition. We therefore conclude that $\int_{\mathcal{H}_4}\left[w\left(S^{-\frac{1}{2}}(\boldsymbol{\tau} - \boldsymbol{\nu}')\right) - w\left(S^{-\frac{1}{2}}\boldsymbol{\tau}\right)\right] I(\boldsymbol{\tau} + \boldsymbol{\mu})d\boldsymbol{\tau} > 0$.

We have shown that $\forall\, \boldsymbol{\nu}' : D\left(p, q_{\boldsymbol{\mu}+\boldsymbol{\nu}'}\right) - D\left(p, q_{\boldsymbol{\mu}}\right) > 0$, which completes the proof. □

### B.4.1 Proof of Theorem 2

We now use Thm. 7 to prove Thm. 2.

**Theorem 2.** *Under (A1), $D_{FKL}(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if (1) $-\log q_{\mathbf{0}}$ is convex and strictly increasing in $||\boldsymbol{x}||$ and (2) $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure.*

**Proof.** Let $w(\boldsymbol{x}) : -\log q_{\mathbf{0}}(\boldsymbol{x})$ and $I(\boldsymbol{x}) := p(\boldsymbol{x})$. Since Thm. 2 assumes that $w$ is convex and strictly increasing in $||\boldsymbol{x}||$, we can apply Thm. 7, which proves Thm. 2. □

### B.4.2 Proof of Theorem 4

We now use Thm. 7 to prove Thm. 4.

**Theorem 4.** *Under (A1), $D_\alpha(p, q_{\boldsymbol{\nu}})$ has a unique global minimizer at $\boldsymbol{\nu} = \boldsymbol{\mu}$ if $\frac{1}{\alpha(\alpha-1)}q_{\mathbf{0}}^{1-\alpha}$ is convex, strictly increasing in $||\boldsymbol{x}||$ and (2) $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure.*

**Proof.** Let $w(\boldsymbol{x}) := \frac{1}{\alpha(\alpha-1)}q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{x})$ and $I(\boldsymbol{x}) := p^\alpha(\boldsymbol{x})$. Note that $\forall\alpha \in \mathbb{R}$, $p^\alpha(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \text{supp}(p)$ and $\text{supp}(p) = \text{supp}(p^\alpha)$. Since Thm. 2 assumes that $w$ is convex and strictly increasing in $||\boldsymbol{x}||$, we can apply Thm. 7, which proves Thm. 4. □

**Notation.** The proof of Thm. 4 completes the proofs for our theorems for mean recovery. To simplify the following analysis, we define two settings, which correspond to Thms. 1 and 3 and Thms. 2 and 4 respectively:

**(S1).** *$w(\boldsymbol{x})$ is strictly convex in $\boldsymbol{x}$.*

**(S2).** *$w(\boldsymbol{x})$ is convex in $\boldsymbol{x}$, strictly increasing in $||\boldsymbol{x}||$, $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure.*

## B.5 Recovery of correlations

We begin by restating the assumptions used throughout our analysis of exact recovery for the correlation matrix.

**(A2).** *Let $q_{\boldsymbol{\nu},S}$ be a member of a location-scale family and let $p$ be a target with finite first and second moments. We make the following assumptions: The base distribution $q_{\mathbf{0}}$ is spherically symmetric. The target $p$ is elliptically symmetric with $M \in \mathbb{S}^d_{++}$ and mean $\boldsymbol{\mu} \in \mathbb{R}^d$. Lastly, $D_f(p, q_{\boldsymbol{\nu},S})$ is well-defined and finite for all considered values of $\boldsymbol{\nu}$ and $S$.*

**Theorem 5.** *Assume (A2). Moreover, assume that all conditions for either Thm. 1 or Thm. 2 are satisfied, and hence we are guaranteed to recover the target mean. If additionally $g(||\boldsymbol{x}||) = \log q_{\mathbf{0}}(\boldsymbol{x})$ is almost everywhere continuously differentiable with $|g'(0)| < \infty$ or $\lim_{u\to 0^+} -g'(u) \geq 0$, then $D_{FKL}(p, q_{\boldsymbol{\nu},S})$ has a unique minimizer at $(\boldsymbol{\mu}, \gamma^2 M)$ for some $\gamma > 0$.*

**Theorem 6.** *Assume (A2). Moreover, assume that all conditions for either Thm. 3 or Thm. 4 are satisfied, and hence we are guaranteed to recover the target mean. If additionally (i) $g(||\boldsymbol{x}||) = q_{\mathbf{0}}^{1-\alpha}(\boldsymbol{x})$ is almost everywhere continuously differentiable with $|g'(0)| < \infty$ or $\lim_{u\to 0^+} g'(u) \geq 0$, (ii) $\alpha > 1$, and (iii) $-\log q_{\mathbf{0}}$ is convex, then $D_\alpha(p, q_{\boldsymbol{\nu},S})$ has a unique minimizer at $(\boldsymbol{\mu}, \gamma^2 M)$ for some $\gamma > 0$.*

Our proofs for exact recovery of correlations use the concept of ***odd symmetry***.

**Definition 4** (Odd Symmetry)**.** *We say a function $h : \mathbb{R}^d \to \mathbb{R}$ is **odd symmetric** around $\boldsymbol{\nu} \in \mathbb{R}^d$ if for all $\boldsymbol{x} \in \mathbb{R}^d$, $h(\boldsymbol{\nu} + \boldsymbol{x}) = -h(\boldsymbol{\nu} - \boldsymbol{x})$.*

We now provide proofs for *exact recovery of the correlation matrix* for the FKL and $\alpha$-DIVs under our sufficient conditions. Our proofs closely follow the proof provided by Margossian and Saul [46, 47] for the RKL. We structure each proof in two parts: In **Part 1**, we show that the objective has a unique minimizer in $S$. In **Part 2**, we prove that this unique minimizer corresponds to the correct correlation matrix.

### B.5.1 FKL

We now provide the proof for *exact recovery of the correlation matrix* for the FKL (Thm. 5).

**Part 1.** First, we simplify $D_{\text{FKL}}$ as follows:

$$D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S}) := \int \log\left(\frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu},S}(\boldsymbol{x})}\right) p(\boldsymbol{x}) d\boldsymbol{x} \tag{40}$$

$$= -H(p) - \int \log\left(q_{\boldsymbol{\nu},S}(\boldsymbol{x})\right) p(\boldsymbol{x}) d\boldsymbol{x} \tag{41}$$

$$= -H(p) - \int \log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}\left(\boldsymbol{x} - \boldsymbol{\nu}\right)\right)|S|^{-\frac{1}{2}}\right) p(\boldsymbol{x}) d\boldsymbol{x} \tag{42}$$

$$= C \underbrace{- \log\left(|S^{-\frac{1}{2}}|\right)}_{:=h(S^{-\frac{1}{2}})} + \int \underbrace{- \log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}\left(\boldsymbol{x} - \boldsymbol{\nu}\right)\right)\right)}_{:=-\log(q_{\mathbf{0}}(a(S^{-\frac{1}{2}})))} p(\boldsymbol{x}) d\boldsymbol{x}, \tag{43}$$

where $H(p) = -\int \log p(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$ denotes the differential entropy of $p$ and $C$ summarizes constants in $\boldsymbol{\nu}$ and $S$.

Note that for both (S1) and (S2) $-\log q_{\mathbf{0}}$ is convex, and $D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S})$ has a unique minimizer w.r.t. $\boldsymbol{\nu}$ for any $S$, as guaranteed by our theorems for mean recovery (Thm. 1, Thm. 2). Moreover, $S$ is positive definite, and therefore so is $S^{-\frac{1}{2}}$ [5]. We will now show that $D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S})$ is strictly convex in $S^{-\frac{1}{2}}$, which implies a unique minimizer in $S$ since $S$ is invertible.

Consider the functions $h$ and $a$ as defined in Eq. (43). It is well-known that $h(A) = -\log(|A|)$ is strictly convex on $A \in \mathbb{S}^d_{++}$ [30, Ch. 7]. Hence, $h(S^{-\frac{1}{2}}) = -\log(|S^{-\frac{1}{2}}|)$ is strictly convex in $S^{-\frac{1}{2}}$. Now define $a : \mathbb{S}^d_{++} \to \mathbb{R}^d$ as the affine map $a(S^{-\frac{1}{2}}) := S^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\nu})$. Since convexity is preserved under composition with affine maps [11, Ch. 3.2.2], we conclude that $-\log q_{\mathbf{0}}(a(S^{-\frac{1}{2}}))$ is convex in $S^{-\frac{1}{2}}$. Further, $\int -\log q_{\mathbf{0}}(a(S^{-\frac{1}{2}})) p(\boldsymbol{x}) d\boldsymbol{x}$ preserves convexity in $S^{-\frac{1}{2}}$ via the argument presented in Lemma 2. Since Eq. (43) is a sum of (1) a strictly convex function in $S^{-\frac{1}{2}}$ and a (2) convex function in $S^{-\frac{1}{2}}$, the overall objective is strictly convex in $S^{-\frac{1}{2}}$. Since $S^{-\frac{1}{2}}$ is invertible, this corresponds to a unique minimizer in $S$ at its stationary point. Note that this partial result assumes neither log-concavity of $p$ nor elliptical symmetry of $p$.

**Part 2.** It remains to show that the unique minimizer of $D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S})$ w.r.t. $S$ recovers the correct correlation matrix, i.e., $S = \gamma^2 M$ for some $\gamma > 0$. Below, we exploit *elliptical symmetry* of $p$, using that under this assumption, $p$ can be written as in Eq. (3).

$$D_{\text{FKL}}(p, q) = \int \log\left(\frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu},S}(\boldsymbol{x})}\right) p(\boldsymbol{x}) d\boldsymbol{x} \tag{44}$$

$$= \int \log\left(p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)|M^{-\frac{1}{2}}|\right) p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)|M^{-\frac{1}{2}}| d\boldsymbol{x} \tag{45}$$

$$- \int \log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\right)|S|^{-\frac{1}{2}}\right) p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)|M^{-\frac{1}{2}}| d\boldsymbol{x} \tag{46}$$

$$= \log\left(|M^{-\frac{1}{2}}|\right) + \int \log\left(p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right) p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)|M^{-\frac{1}{2}}| d\boldsymbol{x} \tag{47}$$

$$- \log\left(|S|^{-\frac{1}{2}}\right) - \int \log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\right)\right) p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right)|M^{-\frac{1}{2}}| d\boldsymbol{x} \tag{48}$$

We now set $\boldsymbol{\zeta} := M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})$, and hence $\boldsymbol{x} = M^{\frac{1}{2}}\boldsymbol{\zeta} + \boldsymbol{\mu}$. Moreover, since we are guaranteed to recover the mean for any $S$ by either Thm. 1 or Thm. 2, we will assume $\boldsymbol{\nu} = \boldsymbol{\mu}$ from now on. Therefore, Eq. (47) simplifies to

$$\begin{aligned}
(47) &= \log\left(|M^{-\frac{1}{2}}|\right) + \int \log\left(p_{\mathbf{0}}\left(M^{-\frac{1}{2}}(M^{\frac{1}{2}}\boldsymbol{\zeta}+\boldsymbol{\mu}-\boldsymbol{\mu})\right)\right) p_{\mathbf{0}}(\boldsymbol{\zeta})|M^{-\frac{1}{2}}||M^{\frac{1}{2}}| d\boldsymbol{\zeta} \\
&= \log\left(|M^{-\frac{1}{2}}|\right) + \int \log\left(p_{\mathbf{0}}\left(\boldsymbol{\zeta}\right)\right) p_{\mathbf{0}}\left(\boldsymbol{\zeta}\right) d\boldsymbol{\zeta} \\
&= \log\left(|M^{-\frac{1}{2}}|\right) - H(p_{\mathbf{0}}).
\end{aligned}$$

Similarly, for Eq. (48), we obtain

$$(48) = -\log\left(|S|^{-\frac{1}{2}}\right) - \int \log\left(q_{\mathbf{0}}\left(S^{-\frac{1}{2}}M^{\frac{1}{2}}\boldsymbol{\zeta}\right)\right) p_{\mathbf{0}}(\boldsymbol{\zeta}) d\boldsymbol{\zeta}.$$

We now define $J := S^{-\frac{1}{2}}M^{\frac{1}{2}}$. Note that

$$\begin{aligned}
\log\left(|M^{-\frac{1}{2}}|\right) - \log\left(|S|^{-\frac{1}{2}}\right) &= \log\left(|M|^{-\frac{1}{2}}|S|^{\frac{1}{2}}\right) \\
&= -\log\left(|M^{\frac{1}{2}}S^{-\frac{1}{2}}|\right) \\
&= -\log\left(|S^{-\frac{1}{2}}M^{\frac{1}{2}}|\right) \\
&= -\log\left(|J|\right).
\end{aligned}$$

Summing Eq. (47) and Eq. (47), and further plugging in the above, we obtain

$$D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S}) = -H(p_{\mathbf{0}}) - \log(|J|) - \int \log q_{\mathbf{0}}\left(J\boldsymbol{\zeta}\right) p_{\mathbf{0}}\left(\boldsymbol{\zeta}\right) d\boldsymbol{\zeta}. \tag{49}$$

Remember that $M^{\frac{1}{2}}$ is fixed and invertible. We want to prove that we have a stationary point at $J = \gamma I$, for some $\gamma > 0$, since this implies $S^{\frac{1}{2}} = \frac{1}{\gamma}M^{\frac{1}{2}}$ and we have successfully recovered the correlation matrix. Since $p_{\mathbf{0}}$ and $\log q_{\mathbf{0}}$ are spherically symmetric, we can define them via functions depending on the norm of their inputs, $g : \mathbb{R}_0^+ \to \mathbb{R}$ and $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$, given as

$$\begin{aligned}
g(||J\boldsymbol{\zeta}||) &= \log q_{\mathbf{0}}(J\boldsymbol{\zeta}) \\
f(||\boldsymbol{\zeta}||) &= p_{\mathbf{0}}(\boldsymbol{\zeta}).
\end{aligned}$$

Plugging this into Eq. (49), we arrive at

$$D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S}) = -H(p_{\mathbf{0}}) - \log(|J|) - \int g\left(||J\boldsymbol{\zeta}||\right) f\left(||\boldsymbol{\zeta}||\right) d\boldsymbol{\zeta}. \tag{50}$$

Differentiating w.r.t. $J$ and using $\nabla_J |J| = |J| J^{-T}$ as well as $\nabla_J ||J\boldsymbol{\zeta}|| = \frac{J\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||J\boldsymbol{\zeta}||}$ [59, 46], we obtain

$$\nabla_J D_{\text{FKL}}(p, q) = -J^{-T} - \int g'(||J\boldsymbol{\zeta}||) \frac{J\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||J\boldsymbol{\zeta}||} f(||\boldsymbol{\zeta}||)\, d\boldsymbol{\zeta}. \tag{51}$$

We want this gradient to be 0 when setting $J = \gamma I$ for some $\gamma > 0$, i.e.,

$$0 = -(\gamma I)^{-1} - \int g'(||\gamma\boldsymbol{\zeta}||) \frac{\gamma\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||\gamma\boldsymbol{\zeta}||} f(||\boldsymbol{\zeta}||)\, d\boldsymbol{\zeta}, \qquad (J = \gamma I)$$

or equivalently,

$$\gamma^{-1} I = -\int f(||\boldsymbol{\zeta}||)\, g'(||\gamma\boldsymbol{\zeta}||) \frac{\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||\boldsymbol{\zeta}||} d\boldsymbol{\zeta}. \tag{52}$$

We need to show that the r.h.s. exhibits the desired diagonal structure implied by $\gamma^{-1} I$. Looking at the $i$-th component of integration, we have

$$-\int d\boldsymbol{\zeta}_{\setminus\{i\}} \int d\zeta_i \; \left[ f(||\boldsymbol{\zeta}||)\, g'(||\gamma\boldsymbol{\zeta}||) ||\boldsymbol{\zeta}||^{-1} \right] \zeta_i \zeta_j. \tag{53}$$

The above closely matches Eq. 26 in [46] (with the main difference being that $f$ and $g$ are "swapped" since $p$ and $q$ take different places in the FKL and the RKL). We can apply the same reasoning: For $i \neq j$, the inner integrand is odd-symmetric in $\zeta_i$ and hence the integral vanishes. If $i = j$, the spherical symmetry of the bracketed term ensures that we obtain the same result for every $i$. Hence, the overall result will be a matrix with (1) all zeros off-diagonal and (2) all equal values on the diagonal.

It remains to show that there is a solution for $\gamma$ that fulfills Eq. (52). Following [46], we show this by equating the traces the l.h.s. and r.h.s. of Eq. (52):

$$d\gamma^{-1} = -\int \left[ f(||\boldsymbol{\zeta}||)\, g'(\gamma||\boldsymbol{\zeta}||) \right] ||\boldsymbol{\zeta}|| d\boldsymbol{\zeta}. \tag{54}$$

Now, let $r := ||\boldsymbol{\zeta}||$ and $\boldsymbol{\zeta} = r\boldsymbol{w}$ with $\boldsymbol{w} \in \mathcal{S}^{d-1}$ denoting a *direction in the unit sphere*. We now change the variable of integration to $r$ on the r.h.s. of Eq. (54) [22, Ch. 2]. Below, $\sigma(\boldsymbol{w})$ denotes the surface measure on the unit sphere.

$$-\int \left[ f(||\boldsymbol{\zeta}||)\, g'(\gamma||\boldsymbol{\zeta}||) \right] ||\boldsymbol{\zeta}|| d\boldsymbol{\zeta} = -\int_{r\in(0,\infty)} \int_{\boldsymbol{w}\in\mathcal{S}^{d-1}} [f(r)\, g'(\gamma r)] r r^{d-1} d\sigma(\boldsymbol{w}) dr \tag{55}$$

$$= -\int_{r\in(0,\infty)} \int_{\boldsymbol{w}\in\mathcal{S}^{d-1}} [f(r)\, g'(\gamma r)] r^d d\sigma(\boldsymbol{w}) dr \tag{56}$$

$$= -\frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_{r\in(0,\infty)} [f(r)\, g'(\gamma r)] r^d dr, \tag{57}$$

where Eq. (57) uses that $\int_{\boldsymbol{w}\in\mathcal{S}^{d-1}} d\sigma(\boldsymbol{w}) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$.

We hence need to show that there exists a solution for $\gamma$ in Eq. (58) given below:

$$d\gamma^{-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_{r\in(0,\infty)} [-g'(\gamma r) f(r)] r^d dr. \tag{58}$$

Note that the l.h.s. is a non-negative, strictly decreasing function in $\gamma$, since $\gamma > 0$. For a unique solution of $\gamma$ to exist, we need the r.h.s. to be positive and increasing in $\gamma$. We now show that this is the case for **(S1)** and **(S2)**.

**(S1)**. Remember that $-g(\gamma r) = -\log q_{\mathbf{0}}(\gamma\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} = r\boldsymbol{w}$. By assumption of **(S1)**, $-\log q_{\mathbf{0}}$ is strictly convex in its argument. This implies that $-g'$ is strictly increasing [65, Chapter 3]. For $\gamma > 0$ and a fixed $r > 0$, $-g'(\gamma r)$ is hence strictly increasing in $\gamma$. Moreover, all involved terms are positive: We have $-g'(0) \geq 0$ by spherical symmetry around the origin and convexity (or at least $\lim_{u\to 0^+} -g'(u) \geq 0$ when the derivative at the origin is ill-defined), and we just established that $-g'(\gamma r)$ is strictly increasing on $\gamma \in (0, \infty)$ for $r > 0$. Moreover, $f(r) \geq 0$ and $r^d > 0$. Hence, the

r.h.s. is indeed a positive, increasing function in $\gamma$. We hence conclude that there is a unique solution for $\gamma$ w.r.t. Eq. (58).

(S2). We can apply the same reasoning for (S2). The main difference in this setting is that $-g'$ is only increasing, not necessarily strictly. Note that we also assume in (S2) that $-\log q_{\mathbf{0}}$ is strictly increasing in $||\boldsymbol{x}||$, hence $g'$ is not 0 everywhere. Therefore, the same argument still holds.

Since we can find a solution for $\gamma$, we conclude that under both (S1) and (S2), the unique minimizer of $D_{\text{FKL}}(p, q_{\boldsymbol{\nu},S})$ w.r.t. $S$ established in **Part 1** accurately captures the correlations of the target.

### B.5.2 $\alpha$-DIV

We now proceed to prove exact recovery of the target correlations under our conditions when optimizing the $\alpha$-DIV (Thm. 6). We again start by simplifying the objective.

$$D_\alpha(p, q_{\boldsymbol{\nu},S}) = \frac{1}{\alpha(\alpha-1)} \int \left[ \left( \frac{p(\boldsymbol{x})}{q_{\boldsymbol{\nu},S}(\boldsymbol{x})} \right)^\alpha - 1 \right] q_{\boldsymbol{\nu},S}(\boldsymbol{x}) d\boldsymbol{x} \tag{59}$$

$$= \int \frac{1}{\alpha(\alpha-1)} q_{\boldsymbol{\nu},S}^{1-\alpha}(\boldsymbol{x}) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} - \frac{1}{\alpha(\alpha-1)} \tag{60}$$

$$= \int \frac{1}{\alpha(\alpha-1)} \left( q_{\mathbf{0}}(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu}))|S|^{-\frac{1}{2}} \right)^{1-\alpha} p^\alpha(\boldsymbol{x}) d\boldsymbol{x} - \frac{1}{\alpha(\alpha-1)} \tag{61}$$

$$= \int \underbrace{\frac{1}{\alpha(\alpha-1)} \left( q_{\mathbf{0}}(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})) \right)^{1-\alpha}}_{:=w(S^{-\frac{1}{2}})} \underbrace{|S^{-\frac{1}{2}}|^{1-\alpha}}_{:=h(S^{-\frac{1}{2}})} p^\alpha(\boldsymbol{x}) d\boldsymbol{x} + C, \tag{62}$$

where $C$ summarizes constants in $\boldsymbol{\nu}$ and $S$.

**Part 1.** We first show that $D_\alpha(p, q_{\boldsymbol{\nu},S})$ has a unique minimizer w.r.t. $S$. For simplicity, we define

$$w(S^{-\frac{1}{2}}) := \frac{1}{\alpha(\alpha-1)} \left( q_{\mathbf{0}} \left( S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu}) \right) \right)^{1-\alpha}$$
$$h(S^{-\frac{1}{2}}) := |S^{-\frac{1}{2}}|^{1-\alpha}$$
$$A := S^{-\frac{1}{2}}.$$

We aim to show that $w(A)h(A)$ is strictly convex in $A$, which implies that $D_\alpha(p, q_{\boldsymbol{\nu},S})$ is strictly convex in $S^{-\frac{1}{2}}$ via Lemma 2.

First, consider $h(A)$. We know that $\log|A|$ is strictly concave in $A$ for $A \in \mathbb{S}^d_{++}$ [30, Ch. 7]. As a result, $(1-\alpha)\log|A|$ is strictly convex for $\alpha > 1$. Since strict convexity is preserved under composition with $\exp(\cdot)$ [11, Ch. 3.5.1], we conclude that $h(A) = |A|^{1-\alpha} = \exp((1-\alpha)\log|A|)$ is strictly convex in $A$ for $\alpha > 1$. Turning to $w(A)$, we know that it is convex in $A$ for both (S1) and (S2), since $A(\boldsymbol{x}-\boldsymbol{\nu})$ is affine in $A$ (same reasoning as in App. B.5.1).

We still need to show that the product $w(A)h(A)$ is strictly convex in $A$ for $\alpha > 1$. Note that $w(A) > 0$ for $\alpha > 1$. Since we assume $\alpha > 1$ (and therefore $1-\alpha < 0$) and further require that $-\log q_{\mathbf{0}}$ is convex, $w(A)$ is log-convex. Since $h(A)$ is strictly log-convex, as we established in the previous paragraph, $w(A)h(A)$ is also strictly log-convex, as (strict) log-convexity is preserved under multiplication [11, Ch. 3.5.2.]. Moreover, every strictly log-convex function is strictly convex. Hence, $w(A)h(A)$ is strictly convex and strict convexity of the integral is preserved via Lemma 2. Note that this result is valid for both (S1) and (S2) as we do not require strict convexity of $w(A)$. We conclude that under our conditions, $D_\alpha(p, q_{\boldsymbol{\nu},S})$ is strictly convex in $S^{-\frac{1}{2}}$ and hence has a unique minimizer in $S$.

**Part 2.** Akin to the FKL, we now show that $D_\alpha(p, q_{\boldsymbol{\nu},S})$ has a stationary point at $S = \gamma M$ for some $\gamma > 0$, such that we correctly recover correlations. Note that optimizing $D_\alpha$ as in Eq. (59) is equivalent to optimizing the *Rényi* $\alpha$-divergence [63, 43], given as

$$D_\alpha^R(p, q_{\boldsymbol{\nu},S}) = \frac{1}{\alpha-1} \log \int p^\alpha(\boldsymbol{x}) q_{\boldsymbol{\nu},S}^{1-\alpha}(\boldsymbol{x}) d\boldsymbol{x} \tag{63}$$

$$= \frac{1}{\alpha-1} \log\left(1 + \alpha(\alpha-1) D_\alpha(p, q_{\boldsymbol{\nu},S})\right). \tag{64}$$

Therefore, we will below consider optimization of the Rényi $\alpha$-divergence, as it facilitates exposition. The proof closely follows the steps performed for the FKL (App. B.5.1). We start by rewriting $D_\alpha^R$ using Eq. (3).

$$\begin{aligned} D_\alpha^R(p, q_{\boldsymbol{\nu},S}) &= \frac{1}{(\alpha-1)} \log \int q_{\boldsymbol{\nu},S}^{1-\alpha}(\boldsymbol{x}) p^\alpha(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{(\alpha-1)} \log \int \left(q_{\mathbf{0}}(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu}))|S^{-\frac{1}{2}}|\right)^{1-\alpha} \left(p_{\mathbf{0}}(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu}))|M^{-\frac{1}{2}}|\right)^{\alpha} d\boldsymbol{x} \\ &= \frac{1}{(\alpha-1)} \log \int \left(q_{\mathbf{0}}(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu}))\right)^{1-\alpha} \left(p_{\mathbf{0}}(M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu}))\right)^{\alpha} |S^{-\frac{1-\alpha}{2}} M^{-\frac{\alpha}{2}}| d\boldsymbol{x}. \end{aligned}$$

We now set $\boldsymbol{\zeta} := M^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})$, and hence $\boldsymbol{x} = M^{\frac{1}{2}}\boldsymbol{\zeta} + \boldsymbol{\mu}$. Moreover, since we are guaranteed to recover the mean for any $S$, we will assume $\boldsymbol{\nu} = \boldsymbol{\mu}$ from now on.

$$D_\alpha^R(p, q) = \frac{1}{(\alpha-1)} \log \int \left(q_{\mathbf{0}}(S^{-\frac{1}{2}} M^{\frac{1}{2}} \boldsymbol{\zeta})\right)^{1-\alpha} (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha |S^{-\frac{1-\alpha}{2}} M^{-\frac{\alpha}{2}}||M^{\frac{1}{2}}| d\boldsymbol{\zeta} \tag{65}$$

$$= \frac{1}{(\alpha-1)} \log \int \left(q_{\mathbf{0}}(S^{-\frac{1}{2}} M^{\frac{1}{2}} \boldsymbol{\zeta})\right)^{1-\alpha} (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha |S^{-\frac{1-\alpha}{2}} M^{\frac{1-\alpha}{2}}| d\boldsymbol{\zeta} \tag{66}$$

$$= \frac{1}{(\alpha-1)} \log \int (q_{\mathbf{0}}(J\boldsymbol{\zeta}))^{1-\alpha} (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha |J|^{1-\alpha} d\boldsymbol{\zeta} \qquad (J := S^{-\frac{1}{2}} M^{\frac{1}{2}})$$

$$= \frac{1}{(\alpha-1)} \log \left(|J|^{1-\alpha} \int (q_{\mathbf{0}}(J\boldsymbol{\zeta}))^{1-\alpha} (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha d\boldsymbol{\zeta}\right) \tag{67}$$

$$= \frac{1}{(\alpha-1)} \log \left(|J|^{1-\alpha}\right) + \frac{1}{(\alpha-1)} \log \int (q_{\mathbf{0}}(J\boldsymbol{\zeta}))^{1-\alpha} (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha d\boldsymbol{\zeta} \tag{68}$$

Since $p_{\mathbf{0}}^\alpha$ and $q_{\mathbf{0}}^{1-\alpha}$ are spherically symmetric, we can define them via functions depending on the norm of their inputs, i.e.,

$$\begin{aligned} g(||J\boldsymbol{\zeta}||) &= (q_{\mathbf{0}}(J\boldsymbol{\zeta}))^{1-\alpha} \\ f(||\boldsymbol{\zeta}||) &= (p_{\mathbf{0}}(\boldsymbol{\zeta}))^\alpha. \end{aligned}$$

Plugging this into Eq. (68), and taking the derivative w.r.t $J$ [59], we obtain

$$\nabla_J D_\alpha(p, q) = \frac{1}{(\alpha-1)} \nabla_J \log \left(|J|^{1-\alpha}\right) + \frac{1}{(\alpha-1)} \nabla_J \log \int g(||J\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta} \tag{69}$$

$$= \frac{1-\alpha}{\alpha-1} J^{-T} + \frac{1}{(\alpha-1)} \frac{\int g'(||J\boldsymbol{\zeta}||) \frac{J\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||J\boldsymbol{\zeta}||} f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}}{\int g(||J\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}} \tag{70}$$

Setting $J = \gamma I$ for some $\gamma > 0$ and using that the gradient should be 0 for this assignment of $J$, we obtain

$$-(1-\alpha)\gamma^{-1} I = \frac{\int g'(||\gamma\boldsymbol{\zeta}||) \frac{\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{||\boldsymbol{\zeta}||} f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}}{\int g(||\gamma\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}}. \tag{71}$$

We again need to show that the r.h.s. exhibits proper diagonal structure. First of all, note that the numerator on the r.h.s. has the same structure as for the FKL (App. B.5.1, Eq. (52)). Hence, it gives the desired isotropic diagonal structure. Moreover, the denominator $\int g(\gamma||\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}$ shares the same spherical property, and hence ensures that all diagonal entries obtain the same value. Hence, we maintain the required structure.

It remains to show that there exists a solution with $\gamma > 0$ for equation Eq. (58). Taking the trace on both sides of Eq. (71), we obtain

$$-(1-\alpha) d\gamma^{-1} = \frac{\int g'(\gamma||\boldsymbol{\zeta}||) ||\boldsymbol{\zeta}|| f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}}{\int g(\gamma||\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}}. \tag{72}$$

First of all, note that under the assumption that $\alpha > 1$, the l.h.s. is positive and strictly decreasing in $\gamma \in (0, \infty)$. Hence, we need to ensure that the r.h.s. is positive and increasing in $\gamma$ to show that a solution exists.

Table 2: A summary of characteristics of the targets used in Sec. 4. The last column highlights the corresponding applicable theorems. Note that for the condition $\boldsymbol{\mu} \in \text{supp}(p)$, we mean $p$ is supported around $\boldsymbol{\mu}$ on a set of positive measure, as stated in our theorems.

| | $d$ | Even symm. | Elliptically symm. | $\boldsymbol{\mu} \in \text{supp}(p)$ | Log-concave | Theorems |
|---|---|---|---|---|---|---|
| *Symbolic Integration* | | | | | | |
| $p_1$ (MoU) | 1 | ✔ | ✔ | ✗ | ✗ | Thms. 1 and 3 |
| $p_2$ (MoU) | 1 | ✔ | ✔ | ✔ | ✗ | Thms. 1 to 4 |
| *Stochastic Optimization* | | | | | | |
| MoU | 2 | ✔ | ✗ | ✗ | ✗ | Thms. 1 and 3 |
| GMM | 2 | ✔ | ✗ | ✔ | ✗ | Thms. 1 to 4 |
| Ellipse | 2 | ✔ | ✔ | ✔ | ✗ | Thms. 1 to 6 |
| A-GMM | 2 | ✗ | ✗ | ✔ | ✗ | / |
| GMM ($4D$) | 4 | ✔ | ✗ | ✔ | ✗ | Thms. 1 to 4 |
| GMM ($8D$) | 8 | ✔ | ✗ | ✔ | ✗ | Thms. 1 to 4 |
| GMM ($16D$) | 16 | ✔ | ✗ | ✔ | ✗ | Thms. 1 to 4 |
| Ellipse ($4D$) | 4 | ✔ | ✔ | ✔ | ✗ | Thms. 1 to 6 |
| Ellipse ($8D$) | 8 | ✔ | ✔ | ✔ | ✗ | Thms. 1 to 6 |
| Ellipse ($16D$) | 16 | ✔ | ✔ | ✔ | ✗ | Thms. 1 to 6 |

Note that the denominator on the r.h.s. is positive by the definitions of $g$ and $f$. For the numerator, we need to analyze $g'$. We know that $g$ corresponds to $q_{\mathbf{0}}^{1-\alpha}$ which for $\alpha > 1$ is convex under both **(S1)** and **(S2)** by assumption. Hence, $g'$ is increasing, and we know $g'(0) \geq 0$ (or at least $\lim_{u \to 0^+} g'(u) \geq 0$ when the derivative at the origin is ill-defined). Therefore, we have $g'(\gamma||\boldsymbol{\zeta}||) \geq 0$. Note that $g'$ cannot be 0 everywhere since **(S1)** requires $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ strictly convex and **(S2)** requires $\frac{1}{\alpha(\alpha-1)} q_{\mathbf{0}}^{1-\alpha}$ to be strictly increasing in $||\boldsymbol{x}||$. For $\alpha > 1$, also the multiplicative constant $\frac{1}{\alpha(\alpha-1)}$ is positive. We hence conclude that the r.h.s. of Eq. (72) is positive.

To see why the r.h.s. is increasing in $\gamma$, note that

$$\frac{\int g'(\gamma||\boldsymbol{\zeta}||)||\boldsymbol{\zeta}|| f(||\boldsymbol{\zeta}||)\, d\boldsymbol{\zeta}}{\int g(\gamma||\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta}} = \frac{\partial}{\partial \gamma} \left[ \log \int g(\gamma||\boldsymbol{\zeta}||) f(||\boldsymbol{\zeta}||) d\boldsymbol{\zeta} \right] \tag{73}$$

$$= \frac{\partial}{\partial \gamma} \left[ \log \int (q_{\mathbf{0}}(J\boldsymbol{\zeta}))^{1-\alpha} (p(\boldsymbol{\zeta})^{\alpha}) d\boldsymbol{\zeta} \right]_{J=\gamma I}. \tag{74}$$

We have established in **Part 1** that the integral on the r.h.s. is log-convex in $S^{-\frac{1}{2}}$ [11], which implies it is log-convex in $J = S^{-\frac{1}{2}} M^{\frac{1}{2}}$. As a result, the derivative of the r.h.s. is increasing in $\gamma \in (0, \infty)$. We conclude that the r.h.s. of Eq. (72) is increasing in $\gamma$.

Since the r.h.s. of Eq. (72) is positive and increasing in $\gamma$, we conclude that there exists a solution for $\gamma$, and hence $D_\alpha^R$ has a stationary point w.r.t. $S$ that fulfills $S^{\frac{1}{2}} = \frac{1}{\gamma} M^{\frac{1}{2}}$. This stationary point corresponds to the unique minimizer of $D_\alpha(p, q_{\boldsymbol{\nu},S})$ w.r.t. $S$, and hence our sufficient conditions allow for exact recovery of correlations with the $\alpha$-DIV.

# C Experimental setup

In this section, we describe the experimental setup used for Sec. 4.

## C.1 Targets

Below, we describe the targets used in our experiments with stochastic optimization. Tab. 2 gives an overview of all targets used throughout the paper and highlights how they relate to our theoretical results. We report the target statistics rounded to two decimal places.

### C.1.1 Two-dimensional targets

**MoU.** We define an *even symmetric* two-dimensional mixture of uniforms with *disjoint support*, akin to the targets used in our symbolic experiments.

$$0.5 \cdot \text{Unif}(x_1; (5.75, 9.75)) \cdot \text{Unif}(x_2; (-1.72, 2.28)) \\ + 0.5 \cdot \text{Unif}(x_1; (-5.75, -1.75)) \cdot \text{Unif}(x_2; (1.72, 5.72)),$$

$$\begin{aligned} \boldsymbol{\mu} &= (2, 2)^T \\ \text{Corr}_p &= \begin{pmatrix} 1 & -0.81 \\ -0.81 & 1 \end{pmatrix} \\ \text{Cov}_p &= \begin{pmatrix} 34.36 & -9.9 \\ -9.9 & 4.30 \end{pmatrix}. \end{aligned}$$

The target is not log-concave and not elliptically symmetric. We sample this target via ancestral sampling [8]. For this target, the RKL is not well-defined for our variational families. We use the REINFORCE gradient estimator instead of reparameterization for this target.

**GMM.** We define a GMM target that is *even symmetric* around its mean, given as

$$0.5 \cdot \mathcal{N}\left((-3, -3)^T, \Sigma\right) + 0.5 \cdot \mathcal{N}\left((7, 7)^T, \Sigma\right),$$

with $\Sigma = \begin{pmatrix} 1 & -0.57 \\ -0.57 & 2 \end{pmatrix}$.

For the full mixture, we have

$$\begin{aligned} \boldsymbol{\mu} &= (2, 2)^T \\ \text{Corr}_p &= \begin{pmatrix} 1 & 0.92 \\ 0.92 & 1 \end{pmatrix} \\ \text{Cov}_p &= \begin{pmatrix} 26 & 24.43 \\ 24.43 & 27 \end{pmatrix}. \end{aligned}$$

The target is not log-concave and not elliptically symmetric. We sample this target via standard ancestral sampling [8].

**Ellipse.** We use a *squared subtractive mixture model (SMM)* [44, 78] to represent an elliptic target with a valley in its center. For a two-component GMM, where each component has mean 0, *squaring* an SMM results in

$$\frac{1}{Z}\left(w_1 \mathcal{N}((0, 0)^T, \Sigma_1) - w_2 \mathcal{N}((0, 0)^T, \Sigma_2)\right)^2, \tag{75}$$

where $Z$ is the tractable normalizing constant of the mixture.

The base distribution $p_0$ of our *Ellipse* target is given as Eq. (75), parameterized with where $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$, $w_1 = 0.75$, $w_2 = 0.25$. The resulting density is visualized as the *spherically symmetric* density in Fig. 2. We transform this base distribution into an ellipse via Eq. (3) with $M = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ and shifting the target with $\boldsymbol{\mu} = (1, 1)^T$.

For the full mixture, we have

$$\begin{aligned} \boldsymbol{\mu} &= (1, 1)^T \\ \text{Corr}_p &= \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \\ \text{Cov}_p &= \begin{pmatrix} 3.70 & 1.85 \\ 1.85 & 3.70 \end{pmatrix}. \end{aligned}$$

The resulting target is *elliptically symmetric* (and hence even symmetric) but not log-concave. To sample from this target, we use rejection sampling based on the positively weighted components of the base distribution [6, 78] and transform the resulting samples:

$$\boldsymbol{x} \sim p_{\mathbf{0}} \text{ (via rejection)}$$
$$\boldsymbol{x} = M^{\frac{1}{2}}\boldsymbol{x} + \boldsymbol{\mu}.$$

We use the cirkit library [66] (released under the GPL-3.0 license) for implementing this target, as well as the *Ellipse* targets in higher dimensions.

**A-GMM.** Lastly, we construct an *asymmetric GMM* which is neither elliptically symmetric nor even symmetric. The target is given as

$$0.5 \cdot \mathcal{N}\left((5,4)^T, \Sigma_1\right) + 0.5 \cdot \mathcal{N}\left((-2,-6)^T, \Sigma_2\right),$$

with $\Sigma_1 = \begin{pmatrix} 1 & -0.57 \\ -0.57 & 2 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 0.5 & 0.14 \\ 0.14 & 1 \end{pmatrix}$

For the full mixture, we have

$$\boldsymbol{\mu} = (1.5, -1)^T$$
$$\text{Corr}_p = \begin{pmatrix} 1 & 0.93 \\ 0.93 & 1 \end{pmatrix}$$
$$\text{Cov}_p = \begin{pmatrix} 13.0 & 17.29 \\ 17.29 & 26.5 \end{pmatrix}.$$

We sample this target via ancestral sampling [8].

### C.1.2 Higher-dimensional targets

**GMM (d-dimensional).** We define a $d$-dimensional *even symmetric* GMM target. Let $\mathbf{1}_d$ denote a vector with all entries equal to 1 in $d$ dimensions.

$$0.5 \cdot \mathcal{N}\left(0 \cdot \mathbf{1}_d, \Sigma\right) + 0.5 \cdot \mathcal{N}\left(2 \cdot \mathbf{1}_d, \Sigma\right),$$

with an isotropic $\Sigma$ with marginal variance 1. The overall target has a mean of $\boldsymbol{\mu} = \mathbf{1}_d$, correlation of $0.5$ on all off-diagonals and the covariance matrix has marginal variance 2 and covariance 1. We sample this target via standard ancestral sampling [8].

**Ellipse (d-dimensional).** We create a higher-dimensional version of the two-dimensional squared SMM. The general setup is the same as for $d = 2$. The main differences is that we vary the mixture weight of the negatively weighted component *before squaring* (see [44]) with $d$ according to $w_1 = 1 - 1/d^2$ and $w_2 = 1/d^2$. In all scenarios, we have $\boldsymbol{\mu} = \mathbf{1}^d$ and the correlation between all pairs of distinct variables is $0.5$. The covariance matrices vary for each $d$, but the marginal variances ae generally close to 2 for all $d$, while off-diagonal entries of the covariance matrix are roughly 1. We use rejection sampling for this target as described in the two-dimensional case.

## C.2 Optimization details

Below, we describe our setup for both symbolic integration and stochastic optimization.

**Hardware.** We run experiments with symbolic integration on CPU (11th Gen Intel(R) Core(TM) i5-1145G7 @ 2.60GHz) and experiments with stochastic optimization on a cluster with 8 NVIDIA RTX A6000 (48GiB VRAM) GPUs and 2 AMD EPYC 7452 32-Core Processors.

**Symbolic integration.** For the experiments shown in Fig. 4, we use symbolic integration with SymPy [52] to compute the divergence. We evaluate $\nu$ on a grid with step size $0.01$, in the range $\nu \in [-15, 15]$.

**Stochastic optimization.** For experiments involving stochastic optimization, we use the Adam optimizer [36] with $lr = 10^{-3}$. We learn on two-dimensional targets for a maximum of $10^4$ steps with $10^5$ samples per update. On higher-dimensional targets, we train for $5 \cdot 10^4$ steps with $2 \cdot 10^4$ samples per update. For the 4-dimensional *Ellipse*, we instead use a learning rate of $0.01$ as we observed this configuration to be more stable. For the *Ellipse* targets, the reported RKL results were learned with the REINFORCE gradient estimator as opposed to reparameterization because we observed improved convergence using REINFORCE.[1] For the two-dimensional MoU target, optimization was generally stable. However, due to noise in the loss estimate, suboptimal model checkpoints were selected for several runs, despite target statistics still noticeably improving until the end of training. We therefore report the last checkpoint after the maximum number of steps for this target for all divergences. Our implementation is based on Pytorch [2] (version 2.9.1). In App. C.2.1, we provide implementation details for our variational families. Tab. 3 describes the initializations used for each target density.

### C.2.1 Implementation details of variational families

Below we describe core implementation details of the variational families used in our experiments.

**Gaussian.** We base our implementation on torch.distributions [2].

**Laplace (2D).** For the two-dimensional Laplace distribution [21], we use its symmetric formulation:

$$q^{\text{Laplace}}_{\boldsymbol{\nu},S}(\boldsymbol{x}) = |S^{-\frac{1}{2}}| q^{\text{Laplace}}_{\mathbf{0}}\left(S^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\nu})\right), \text{with}$$
$$q^{\text{Laplace}}_{\mathbf{0}}(\boldsymbol{z}) = \frac{1}{\pi} K_0\left(\sqrt{2\boldsymbol{z}^T\boldsymbol{z}}\right)$$

where $K_0$ is the modified Bessel function of the second kind [1, Ch. 9],

$$K_0(z) = \int_0^\infty \cos(z \sinh(t)) dt.$$

To sample from the two-dimensional Laplace distribution, we follow the procedure described in Kotz et al. [38] (Chapter 5.1.4) which exploits the fact that the multivariate Laplace distribution can be represented as scale-mixture of Gaussians. Let $\text{Exp}(1)$ denote a one-dimensional exponential distribution with rate parameter $\lambda = 1$ and let $I^{2\times 2}$ be the two-dimensional identity matrix. We generate samples from $q^{\text{Laplace}}_{\boldsymbol{\nu},S}$ as detailed below:

$$z \sim \text{Exp}(1)$$
$$\boldsymbol{w} \sim \mathcal{N}(0, I^{2\times 2})$$
$$\boldsymbol{x} = \boldsymbol{\nu} + \sqrt{z} S^{\frac{1}{2}} \boldsymbol{w},$$

where now $\boldsymbol{x} \sim \text{Laplace}(\boldsymbol{\nu}, S)$. Since the multivariate Laplace is not readily available in any major deep learning library, we manually implement the two-dimensional Laplace using built-in functions from PyTorch [2]. The two-dimensional Laplace distribution described above is not not log-concave.

**Student-t.** We use the implementation provided in Pyro [7]. In our experiments, we use $5$ degrees of freedom for two-dimensional targets and for the higher-dimensional *GMM* targets. For the higher-dimensional targets *Ellipse*, we use $v = 8$ since this improved stability during optimization.

**Parameterization.** We parameterize the scale of the variational approximation via an unconstrained lower-triangular matrix $L$. To ensure that the resulting covariance matrix is positive definite, we define a matrix $L'$ as follows:

$$L'_{ij} = \begin{cases} L_{ij} & \text{if } i \neq j \\ \exp(L_{ij}) & \text{if } i = j. \end{cases}$$

The covariance of $q$ is then defined as $\text{Cov}_q = L'(L')^T$.

[1] It has for instance been noted by Domke et al. [20] that for certain target geometries REINFORCE can indeed be preferable over reparameterization.

Table 3: Initializations for experiments with stochastic optimization. We initialize means as 0-vectors in $d$ dimensions. For the *Ellipse* in 16 dimensions, we defined separate initializations for the Student-t and Gaussian to improve learning stability.

| Target | d | Location ($\boldsymbol{\nu}$) | Marginal Variance | Off-diagonal Covariance |
|---|---|---|---|---|
| MoU | 2 | 0 | 25 | 0 |
| GMM | 2 | 0 | 25 | 0 |
| Ellipse | 2 | 0 | 30 | 25 |
| A-GMM | 2 | 0 | 25 | 0 |
| GMM | 4 | 0 | 4 | 3 |
| GMM | 8 | 0 | 4 | 3 |
| GMM | 16 | 0 | 4 | 3 |
| Ellipse | 4 | 0 | 30 | 25 |
| Ellipse | 8 | 0 | 30 | 25 |
| Ellipse (Gaussian $q$) | 16 | 0 | 3 | 1.8 |
| Ellipse (Student-t $q$) | 16 | 0 | 3 | 1 |

### C.2.2 Initializations

Tab. 3 summarizes our initialization scheme for each target. For repeated runs we *noise* the values reported in Tab. 3 as described below. We draw $\boldsymbol{\epsilon}_{\boldsymbol{\nu}}$ from a $d$-dimensional standard Gaussian and $\boldsymbol{\epsilon}_L$ denotes a $d \times d$ lower-triangular matrix whose non-zero entries are similarly sampled from a standard Gaussian. For $d < 16$, we then set

$$\begin{aligned}\boldsymbol{\nu}_{\text{init}} &= \boldsymbol{\nu}_{\text{init}} + \boldsymbol{\epsilon}_{\boldsymbol{\nu}} \\ L_{\text{init}} &= L_{\text{init}} + 0.2 \cdot \boldsymbol{\epsilon}_L.\end{aligned}$$

For $d = 16$, we found the $\alpha$-divergence with $\alpha = 2.0$ to be instable unless carefully initialized. Hence, we use softer noising for $d = 16$, as described below.

$$\begin{aligned}\boldsymbol{\nu}_{\text{init}} &= \boldsymbol{\nu}_{\text{init}} + 0.1 \cdot \boldsymbol{\epsilon}_{\boldsymbol{\nu}} \\ L_{\text{init}} &= L_{\text{init}} + 0.2 \cdot \boldsymbol{\epsilon}_L.\end{aligned}$$

## D Additional results and analysis

### D.1 Failure of the FKL with the Laplace variational family on $p_1$

Below, we formalize why the Laplace distribution in Fig. 4 fails to distinguish between different assignments of the location parameter $\nu' \in [-3, 3]$ when approximating $p_1$ with the FKL. The target is given as $0.5 \cdot \text{Unif}(-9, -3) + 0.5 \cdot \text{Unif}(3, 9)$, and the variational family is $\mathcal{Q} = \text{Laplace}(\nu, 4)$. We have $\text{supp}(p_1) = [-9, -3] \cup [3, 9]$ by definition of $p_1$. As shown in Fig. 4, optimizing $\nu$ in this setup results in a plateau of the divergence around the true mean $\mu = 0$.

The members of our variational family take the following form:

$$q_\nu(x) = \frac{1}{2 \cdot 4} e^{-\frac{|x-\nu|}{4}} = \frac{1}{4} q_0\Big(\frac{1}{4}(x - \nu)\Big), \quad \text{with} \quad q_0(x) = \frac{1}{2} e^{-|x|}. \tag{76}$$

Hence, we have $w(x) := -\log(q_0(x)) = -\log \frac{1}{2} + |x|$. Now, consider the difference in $w$ at alternative function inputs $\frac{1}{4}(x - \mu - \nu')$ and $\frac{1}{4}(x - \mu)$, noting that $\mu = 0$ by definition of $p_1$:

$$w\left(\frac{1}{4}(x - \nu')\right) - w\left(\frac{1}{4}x\right) = \frac{|x - \nu'|}{4} - \frac{|x|}{4}. \tag{77}$$

It is easy to see that Eq. (77) $\geq 0$ whenever $|x - \nu'| \geq |x|$. Consider the case $\nu' > 0$. Then, we have

$$\frac{|x - \nu'|}{4} - \frac{|x|}{4} = \begin{cases} \frac{\nu'}{4} & \text{for } x \leq 0 \\ \frac{-2x + \nu'}{4} & \text{for } 0 < x < \nu' \\ -\frac{\nu'}{4} & \text{for } x \geq \nu'. \end{cases} \tag{78}$$

As a result, in line with the proof of Thm. 7, we can define

$$\mathcal{H}_1 := \left(-\infty, \frac{\nu'}{2}\right] \text{ (indicating Eq. (78)} \geq 0) \tag{79}$$

$$\mathcal{H}_2 := \left(\frac{\nu'}{2}, \infty\right) \text{ (indicating Eq. (78)} < 0). \tag{80}$$

We further split $\mathcal{H}_1$ to obtain

$$\mathcal{H}_3 := \left(-\infty, -\frac{\nu'}{2}\right) \text{ (with Eq. (78)} \geq 0) \tag{81}$$

$$\mathcal{H}_4 := \left[-\frac{\nu'}{2}, \frac{\nu'}{2}\right] \text{ (with Eq. (78)} \geq 0). \tag{82}$$

Remember that $\text{supp}(p_1) = [-9, -3] \cup [3, 9]$. Now pick any $\nu' \in (0, 3]$: For this range of $\nu'$, $p_1$ is not supported on $\mathcal{H}_4$. Hence, $\mathcal{H}_4$ does not contribute to the difference in FKL values.

Further, consider that $\mathcal{H}_3' := \mathcal{H}_3 \cap \text{supp}(p_1) \subset [-9, -3]$ and $\mathcal{H}_2' := \mathcal{H}_2 \cap \text{supp}(p_1) \subset [3, 9]$. On these intervals, for $\nu' \in (0, 3]$, we have Eq. (78) $= \frac{\nu'}{4}$ for $x \in \mathcal{H}_3'$ and Eq. (78) $= -\frac{\nu'}{4}$ for $x \in \mathcal{H}_2'$. Combined with the even symmetry of $p_1$, this confirms that we cannot distinguish any $\nu' \in (0, 3]$ from the true mean, since the contributions of $\mathcal{H}_2$ and $\mathcal{H}_3$ to $\Delta$FKL cancel exactly. Analogous reasoning can be applied for $\nu' \in [-3, 0)$.

### D.2 Additional results

In this section, we provide additional empirical results supporting our main theorems.

#### D.2.1 Additional results for symbolic integration with the $\alpha$-DIV

In Fig. 8, we present additional results for the $\alpha$-DIV on the target $p_1$, as used in Sec. 4.1, with a Gaussian variational family. Following the setup in Sec. 4.1, we fix the standard deviation of the variational family to 4. We find that choosing $\alpha > 1$, which results in a strictly convex $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$, consistently guarantees a unique minimizer at the true mean which agrees with our Thm. 3. On the other hand, we observe that lower values of $\alpha$, which do not result in a strictly convex $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$, can be sufficient.

#### D.2.2 Additional results with stochastic optimization

In Tab. 4, we provide additional results with stochastic optimization, complementing Fig. 1. The metrics are defined in Sec. 4.2. For the experimental setup, please refer to App. C.

We find that, whenever our theorems guarantee a unique minimizer corresponding to the correct mean or correlation we observe low error in practice, given that optimization is stable. For the $\alpha$-divergence with $\alpha = 2.0$, for numbers denoted in black and bold in the table, learning was instable leading to non-negligible error in practice despite a theoretical guarantee of exact recovery.

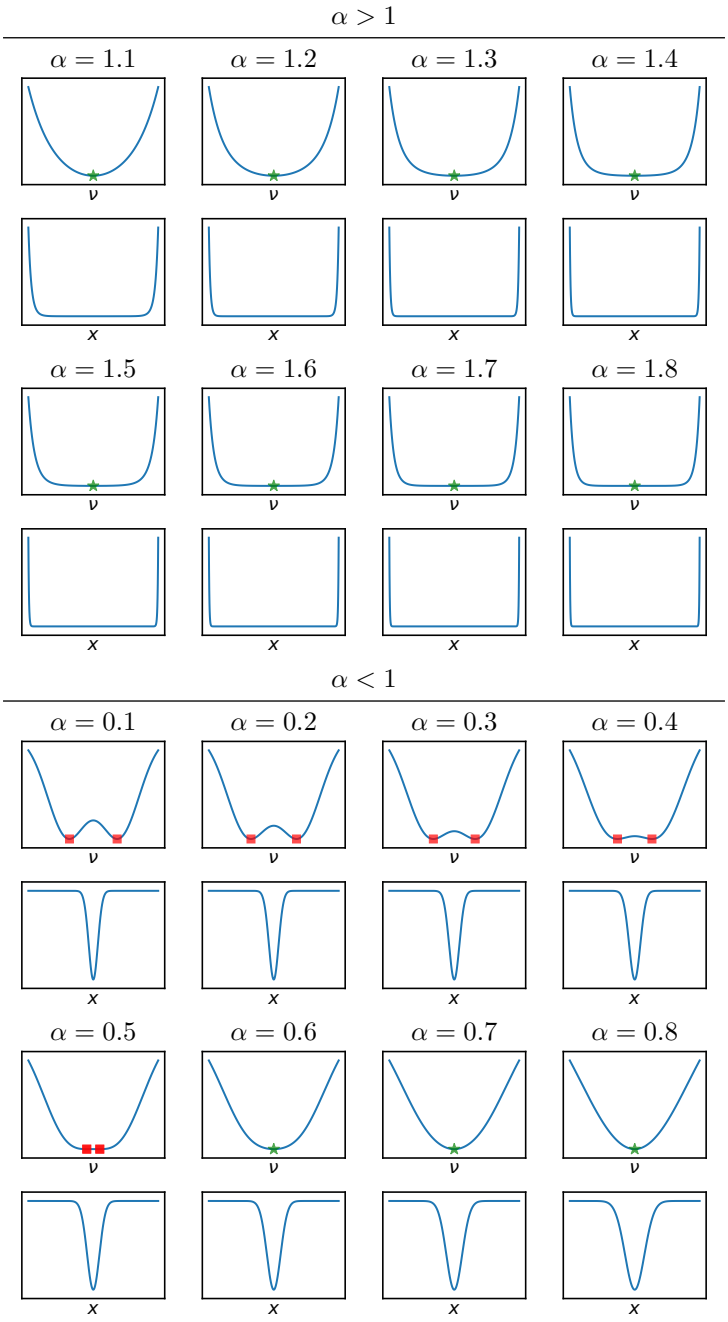



Figure 8: Additional results on $p_1$ with a Gaussian variational family using a fixed standard deviation of 4 (cf. Sec. 4.1, Thm. 3). The first row for each $\alpha$ shows the divergence, the second row shows the associated function $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$ on which we place our sufficient conditions. For the Gaussian variational family, choosing $\alpha > 1$ results in a strictly convex $\frac{1}{\alpha(\alpha-1)} q_0^{1-\alpha}$.

Table 4: **Whenever our sufficient conditions are satisfied, we recover the target characteristics with low error in practice.** We distinguish 3 cases (1) **low error in practice with guarantees (by our theorems or previous results)**, (2) **no guarantee but low error ($< 10^{-2}$) in practice**, (3) **no guarantee and high error ($\geq 10^{-2}$) in practice**. For higher-dimensional targets, the $\alpha$-divergence with $\alpha = 2.0$ occasionally **failed to optimize properly due to instability**, despite our guarantees. We denote these entries in black and bold. We highlight rows of approximations learned with the FKL and a Gaussian proposal via an asterisk (*) as mean and correlation recovery is guaranteed (regardless of target symmetry) via previous results by Wainwright and Jordan [70]. The metrics are defined in Sec. 4.2. We average results across 5 runs.

| | | RKL | | FKL | | $\alpha$-DIV ($\alpha = 0.5$) | | $\alpha$-DIV ($\alpha = 2.0$) | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $q$ | $\Delta\mu$ | $\Delta$Corr | $\Delta\mu$ | $\Delta$Corr | $\Delta\mu$ | $\Delta$Corr | $\Delta\mu$ | $\Delta$Corr |
| MoU | Gaussian | / | / | $2.9 \cdot 10^{-4} \pm 1.7 \cdot 10^{-4}$* | $1.3 \cdot 10^{-4} \pm 7.4 \cdot 10^{-5}$* | $5.8 \cdot 10^{-4} \pm 2.4 \cdot 10^{-4}$ | $4.1 \cdot 10^{-2} \pm 2.9 \cdot 10^{-4}$ | $1.8 \cdot 10^{-4} \pm 9.9 \cdot 10^{-5}$ | $5.0 \cdot 10^{-2} \pm 6.6 \cdot 10^{-4}$ |
| | Laplace | / | / | $2.1 \cdot 10^{-3} \pm 2.8 \cdot 10^{-3}$ | $3.8 \cdot 10^{-2} \pm 8.9 \cdot 10^{-5}$ | $9.1 \cdot 10^{-1} \pm 2.1 \cdot 10^{-4}$ | $8.0 \cdot 10^{-1} \pm 5.5 \cdot 10^{-3}$ | $5.0 \cdot 10^{-4} \pm 2.9 \cdot 10^{-4}$ | $2.1 \cdot 10^{-2} \pm 5.0 \cdot 10^{-4}$ |
| | Student-t | / | / | $3.4 \cdot 10^{-4} \pm 1.1 \cdot 10^{-4}$ | $1.9 \cdot 10^{-2} \pm 1.9 \cdot 10^{-4}$ | $1.3 \cdot 10^{-3} \pm 7.4 \cdot 10^{-4}$ | $6.3 \cdot 10^{-2} \pm 4.0 \cdot 10^{-4}$ | $4.7 \cdot 10^{-4} \pm 2.5 \cdot 10^{-4}$ | $3.7 \cdot 10^{-2} \pm 3.2 \cdot 10^{-4}$ |
| GMM | Gaussian | $9.7 \cdot 10^{-1} \pm 3.9 \cdot 10^{-4}$ | $1.3 \cdot 10^{0} \pm 7.9 \cdot 10^{-4}$ | $1.9 \cdot 10^{-4} \pm 7.5 \cdot 10^{-5}$* | $1.5 \cdot 10^{-3} \pm 1.6 \cdot 10^{-3}$* | $5.9 \cdot 10^{-1} \pm 5.3 \cdot 10^{-1}$ | $7.9 \cdot 10^{-1} \pm 7.1 \cdot 10^{-1}$ | $9.1 \cdot 10^{-4} \pm 5.7 \cdot 10^{-4}$ | $3.2 \cdot 10^{-3} \pm 1.9 \cdot 10^{-3}$ |
| | Laplace | $9.7 \cdot 10^{-1} \pm 5.7 \cdot 10^{-5}$ | $1.3 \cdot 10^{0} \pm 7.1 \cdot 10^{-4}$ | $2.1 \cdot 10^{-3} \pm 1.6 \cdot 10^{-3}$ | $2.7 \cdot 10^{-2} \pm 3.7 \cdot 10^{-4}$ | $9.7 \cdot 10^{-1} \pm 1.1 \cdot 10^{-4}$ | $1.3 \cdot 10^{0} \pm 8.9 \cdot 10^{-3}$ | $1.0 \cdot 10^{-3} \pm 9.1 \cdot 10^{-4}$ | $1.6 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$ |
| | Student-t | $9.7 \cdot 10^{-1} \pm 3.0 \cdot 10^{-5}$ | $1.3 \cdot 10^{0} \pm 8.1 \cdot 10^{-4}$ | $2.7 \cdot 10^{-4} \pm 1.3 \cdot 10^{-4}$ | $1.6 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$ | $9.7 \cdot 10^{-1} \pm 2.7 \cdot 10^{-4}$ | $1.3 \cdot 10^{0} \pm 7.3 \cdot 10^{-3}$ | $1.2 \cdot 10^{-3} \pm 1.1 \cdot 10^{-3}$ | $1.4 \cdot 10^{-2} \pm 2.1 \cdot 10^{-4}$ |
| Ellipse | Gaussian | $5.2 \cdot 10^{-4} \pm 3.5 \cdot 10^{-4}$ | $3.5 \cdot 10^{-3} \pm 4.4 \cdot 10^{-3}$ | $5.5 \cdot 10^{-4} \pm 2.9 \cdot 10^{-4}$* | $2.6 \cdot 10^{-4} \pm 1.3 \cdot 10^{-4}$* | $1.2 \cdot 10^{-3} \pm 2.4 \cdot 10^{-4}$ | $2.5 \cdot 10^{-4} \pm 1.9 \cdot 10^{-4}$ | $4.5 \cdot 10^{-4} \pm 2.5 \cdot 10^{-4}$ | $2.7 \cdot 10^{-4} \pm 2.8 \cdot 10^{-4}$ |
| | Laplace | $4.8 \cdot 10^{-1} \pm 6.7 \cdot 10^{-3}$ | $2.3 \cdot 10^{-1} \pm 2.8 \cdot 10^{-3}$ | $1.1 \cdot 10^{-3} \pm 6.1 \cdot 10^{-4}$ | $2.2 \cdot 10^{-4} \pm 2.5 \cdot 10^{-4}$ | $1.3 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$ | $4.6 \cdot 10^{-3} \pm 4.7 \cdot 10^{-3}$ | $1.7 \cdot 10^{-3} \pm 2.1 \cdot 10^{-3}$ | $2.0 \cdot 10^{-3} \pm 2.6 \cdot 10^{-3}$ |
| | Student-t | $4.9 \cdot 10^{-1} \pm 1.5 \cdot 10^{-3}$ | $2.7 \cdot 10^{-1} \pm 2.2 \cdot 10^{-3}$ | $6.7 \cdot 10^{-4} \pm 4.0 \cdot 10^{-4}$ | $2.1 \cdot 10^{-4} \pm 1.2 \cdot 10^{-4}$ | $1.4 \cdot 10^{-3} \pm 8.8 \cdot 10^{-4}$ | $6.9 \cdot 10^{-4} \pm 5.7 \cdot 10^{-4}$ | $5.8 \cdot 10^{-4} \pm 2.6 \cdot 10^{-4}$ | $2.9 \cdot 10^{-3} \pm 4.8 \cdot 10^{-3}$ |
| A-GMM | Gaussian | $3.9 \cdot 10^{-1} \pm 4.9 \cdot 10^{-1}$ | $1.7 \cdot 10^{-1} \pm 3.1 \cdot 10^{-1}$ | $2.6 \cdot 10^{-4} \pm 1.6 \cdot 10^{-4}$* | $2.6 \cdot 10^{-4} \pm 2.8 \cdot 10^{-4}$* | $1.6 \cdot 10^{-3} \pm 9.3 \cdot 10^{-4}$ | $9.9 \cdot 10^{-3} \pm 3.5 \cdot 10^{-4}$ | $4.9 \cdot 10^{-2} \pm 2.0 \cdot 10^{-3}$ | $1.5 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$ |
| | Laplace | $9.7 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$ | $8.5 \cdot 10^{-1} \pm 2.7 \cdot 10^{-1}$ | $5.8 \cdot 10^{-1} \pm 8.0 \cdot 10^{-3}$ | $2.5 \cdot 10^{-2} \pm 5.8 \cdot 10^{-4}$ | $9.7 \cdot 10^{-1} \pm 2.7 \cdot 10^{-4}$ | $1.0 \cdot 10^{0} \pm 2.8 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1} \pm 3.2 \cdot 10^{-3}$ | $1.6 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$ |
| | Student-t | $9.7 \cdot 10^{-1} \pm 5.7 \cdot 10^{-3}$ | $7.2 \cdot 10^{-1} \pm 7.7 \cdot 10^{-4}$ | $9.6 \cdot 10^{-2} \pm 8.0 \cdot 10^{-4}$ | $1.9 \cdot 10^{-2} \pm 1.1 \cdot 10^{-3}$ | $3.4 \cdot 10^{-1} \pm 3.5 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1} \pm 3.0 \cdot 10^{-1}$ | $8.5 \cdot 10^{-2} \pm 2.1 \cdot 10^{-3}$ | $9.9 \cdot 10^{-3} \pm 2.1 \cdot 10^{-3}$ |
| GMM (4D) | Gaussian | $2.0 \cdot 10^{-3} \pm 1.7 \cdot 10^{-3}$ | $5.7 \cdot 10^{-2} \pm 9.8 \cdot 10^{-4}$ | $1.3 \cdot 10^{-3} \pm 4.4 \cdot 10^{-4}$* | $9.3 \cdot 10^{-4} \pm 2.1 \cdot 10^{-4}$* | $1.7 \cdot 10^{-3} \pm 6.9 \cdot 10^{-4}$ | $2.0 \cdot 10^{-2} \pm 1.2 \cdot 10^{-3}$ | $8.8 \cdot 10^{-4} \pm 3.5 \cdot 10^{-4}$ | $2.2 \cdot 10^{-2} \pm 7.8 \cdot 10^{-4}$ |
| | Student-t | $3.1 \cdot 10^{-3} \pm 2.3 \cdot 10^{-3}$ | $9.3 \cdot 10^{-2} \pm 1.4 \cdot 10^{-3}$ | $1.4 \cdot 10^{-3} \pm 4.7 \cdot 10^{-4}$ | $4.1 \cdot 10^{-2} \pm 1.7 \cdot 10^{-4}$ | $2.6 \cdot 10^{-3} \pm 1.1 \cdot 10^{-3}$ | $1.3 \cdot 10^{-2} \pm 2.0 \cdot 10^{-3}$ | $1.2 \cdot 10^{-3} \pm 1.3 \cdot 10^{-4}$ | $6.0 \cdot 10^{-2} \pm 5.4 \cdot 10^{-4}$ |
| GMM (8D) | Gaussian | $1.4 \cdot 10^{-1} \pm 3.1 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1} \pm 1.8 \cdot 10^{-1}$ | $9.1 \cdot 10^{-4} \pm 1.6 \cdot 10^{-4}$* | $9.6 \cdot 10^{-4} \pm 1.5 \cdot 10^{-4}$* | $6.6 \cdot 10^{-3} \pm 1.1 \cdot 10^{-2}$ | $2.1 \cdot 10^{-2} \pm 7.2 \cdot 10^{-4}$ | $\mathbf{2.4 \cdot 10^{-1} \pm 5.3 \cdot 10^{-1}}$ | $6.4 \cdot 10^{-2} \pm 1.1 \cdot 10^{-1}$ |
| | Student-t | $6.9 \cdot 10^{-1} \pm 5.6 \cdot 10^{-4}$ | $4.7 \cdot 10^{-1} \pm 6.8 \cdot 10^{-4}$ | $9.6 \cdot 10^{-4} \pm 2.4 \cdot 10^{-4}$ | $3.3 \cdot 10^{-2} \pm 4.0 \cdot 10^{-4}$ | $2.2 \cdot 10^{-3} \pm 1.6 \cdot 10^{-3}$ | $1.5 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$ | $1.6 \cdot 10^{-3} \pm 3.8 \cdot 10^{-4}$ | $3.8 \cdot 10^{-2} \pm 6.8 \cdot 10^{-4}$ |
| GMM (16D) | Gaussian | $2.1 \cdot 10^{-3} \pm 1.5 \cdot 10^{-3}$ | $9.2 \cdot 10^{-2} \pm 7.3 \cdot 10^{-4}$ | $1.2 \cdot 10^{-3} \pm 2.8 \cdot 10^{-4}$* | $9.7 \cdot 10^{-4} \pm 1.3 \cdot 10^{-4}$* | $2.5 \cdot 10^{-3} \pm 9.9 \cdot 10^{-4}$ | $1.4 \cdot 10^{-2} \pm 1.4 \cdot 10^{-3}$ | $1.8 \cdot 10^{-3} \pm 4.7 \cdot 10^{-4}$ | $6.9 \cdot 10^{-3} \pm 6.1 \cdot 10^{-4}$ |
| | Student-t | $7.0 \cdot 10^{-1} \pm 3.0 \cdot 10^{-4}$ | $4.9 \cdot 10^{-1} \pm 1.1 \cdot 10^{-4}$ | $1.2 \cdot 10^{-3} \pm 2.9 \cdot 10^{-4}$ | $2.3 \cdot 10^{-2} \pm 2.0 \cdot 10^{-4}$ | $2.6 \cdot 10^{-3} \pm 7.3 \cdot 10^{-4}$ | $1.6 \cdot 10^{-2} \pm 6.5 \cdot 10^{-4}$ | $2.0 \cdot 10^{-3} \pm 3.5 \cdot 10^{-4}$ | $2.1 \cdot 10^{-2} \pm 5.5 \cdot 10^{-4}$ |
| Ellipse (4D) | Gaussian | $1.8 \cdot 10^{-3} \pm 4.8 \cdot 10^{-4}$ | $1.6 \cdot 10^{-3} \pm 6.9 \cdot 10^{-4}$ | $2.1 \cdot 10^{-3} \pm 9.1 \cdot 10^{-4}$* | $2.7 \cdot 10^{-3} \pm 7.9 \cdot 10^{-4}$* | $4.3 \cdot 10^{-3} \pm 1.7 \cdot 10^{-3}$ | $3.3 \cdot 10^{-3} \pm 1.4 \cdot 10^{-3}$ | $3.7 \cdot 10^{-3} \pm 1.4 \cdot 10^{-3}$ | $3.1 \cdot 10^{-3} \pm 4.0 \cdot 10^{-4}$ |
| | Student-t | $2.7 \cdot 10^{-3} \pm 7.6 \cdot 10^{-4}$ | $3.3 \cdot 10^{-3} \pm 1.5 \cdot 10^{-3}$ | $2.7 \cdot 10^{-3} \pm 2.0 \cdot 10^{-3}$ | $3.1 \cdot 10^{-3} \pm 4.5 \cdot 10^{-4}$ | $5.9 \cdot 10^{-3} \pm 3.0 \cdot 10^{-3}$ | $6.1 \cdot 10^{-3} \pm 2.8 \cdot 10^{-3}$ | $2.9 \cdot 10^{-3} \pm 6.8 \cdot 10^{-4}$ | $3.9 \cdot 10^{-3} \pm 1.4 \cdot 10^{-3}$ |
| Ellipse (8D) | Gaussian | $4.9 \cdot 10^{-4} \pm 1.8 \cdot 10^{-4}$ | $1.1 \cdot 10^{-3} \pm 1.3 \cdot 10^{-3}$ | $9.3 \cdot 10^{-4} \pm 3.2 \cdot 10^{-4}$* | $1.0 \cdot 10^{-3} \pm 8.4 \cdot 10^{-5}$* | $1.3 \cdot 10^{-3} \pm 5.4 \cdot 10^{-4}$ | $1.3 \cdot 10^{-3} \pm 2.7 \cdot 10^{-4}$ | $9.6 \cdot 10^{-4} \pm 2.3 \cdot 10^{-4}$ | $8.5 \cdot 10^{-4} \pm 2.1 \cdot 10^{-4}$ |
| | Student-t | $1.8 \cdot 10^{-3} \pm 5.8 \cdot 10^{-4}$ | $2.0 \cdot 10^{-3} \pm 5.4 \cdot 10^{-4}$ | $1.3 \cdot 10^{-3} \pm 1.3 \cdot 10^{-4}$ | $8.5 \cdot 10^{-4} \pm 1.7 \cdot 10^{-4}$ | $1.9 \cdot 10^{-3} \pm 7.0 \cdot 10^{-4}$ | $1.6 \cdot 10^{-3} \pm 1.4 \cdot 10^{-4}$ | $1.1 \cdot 10^{-3} \pm 1.7 \cdot 10^{-4}$ | $1.2 \cdot 10^{-3} \pm 1.0 \cdot 10^{-4}$ |
| Ellipse (16D) | Gaussian | $8.3 \cdot 10^{-4} \pm 5.7 \cdot 10^{-5}$ | $5.8 \cdot 10^{-4} \pm 6.2 \cdot 10^{-5}$ | $1.1 \cdot 10^{-3} \pm 2.4 \cdot 10^{-4}$* | $9.0 \cdot 10^{-4} \pm 4.3 \cdot 10^{-5}$* | $1.9 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$ | $1.4 \cdot 10^{-3} \pm 1.7 \cdot 10^{-4}$ | $\mathbf{1.3 \cdot 10^{-2} \pm 1.6 \cdot 10^{-2}}$ | $4.7 \cdot 10^{-3} \pm 2.3 \cdot 10^{-3}$ |
| | Student-t | $1.5 \cdot 10^{-3} \pm 2.8 \cdot 10^{-4}$ | $2.2 \cdot 10^{-3} \pm 1.4 \cdot 10^{-4}$ | $1.1 \cdot 10^{-3} \pm 1.1 \cdot 10^{-4}$ | $9.7 \cdot 10^{-4} \pm 6.5 \cdot 10^{-5}$ | $1.9 \cdot 10^{-3} \pm 5.0 \cdot 10^{-4}$ | $1.5 \cdot 10^{-3} \pm 1.7 \cdot 10^{-4}$ | $1.5 \cdot 10^{-3} \pm 2.5 \cdot 10^{-4}$ | $1.3 \cdot 10^{-3} \pm 1.5 \cdot 10^{-4}$ |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction promise theoretical results for exact mean and correlation recovery with the FKL and $\alpha$-DIV, as well as experiments showcasing their impact in practice, which we provide in Sec. 3 and Sec. 4 respectively.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide an explicit discussion of limitations at the end of Sec. 5

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide full proofs for all of our theorems in App. B and cleanly state the assumptions made (see (A1), (A2)).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our main experimental setup in Sec. 4 (declaring the optimizer, hyperparameters, number of update steps and model selection) and further expand on it in App. C, where we provide detailed information on the used targets (App. C.1), initializations (Tab. 3), and implementation of the variational families, as well as the used hardware.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code for our experiments in a public Github repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we declare our training details, model selection, and evaluation metrics in Sec. 4 and App. C.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tab. 4 reports results accompanied by error bars, which are computed from 5 initializations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the used hardware specification in App. C.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work does not involve human subjects nor publicly available or sensitive data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work focuses on the theory of variational inference. We do not believe any particular societal impacts must be highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not use data or models that have a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We do not use publicly available data, but cite packages and code bases that we build our implementation on in App. C.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We used LLM assistance for coding and occasional brainstorming. In particular, LLM assistance was useful for gathering references for the convexity arguments used in the paper. All proofs and experiments were designed and executed by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.