Journal of Machine Learning Research 24 (2023) 1-83 Submitted 10/22; Revised 6/23; Published 8/23

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics

Kamélia Daudel KAMELIA.DAUDEL@STATS.OX.AC.UK

Joe Benton* BENTON@STATS.OX.AC.UK

Yuyang Shi* YSHI@STATS.OX.AC.UK

Arnaud Doucet DOUCET@STATS.OX.AC.UK

Department of Statistics, University of Oxford
Oxford OX1 3TG, United Kingdom

Editor: Pierre Alquier

Abstract

Several algorithms involving the Variational Rényi (VR) bound have been proposed to minimize an alpha-divergence between a target posterior distribution and a variational distribution. Despite promising empirical results, those algorithms resort to biased stochastic gradient descent procedures and thus lack theoretical guarantees. In this paper, we formalize and study the VR-IWAE bound, a generalization of the importance weighted auto-encoder (IWAE) bound. We show that the VR-IWAE bound enjoys several desirable properties and notably leads to the same stochastic gradient descent procedure as the VR bound in the reparameterized case, but this time by relying on unbiased gradient estimators. We then provide two complementary theoretical analyses of the VR-IWAE bound and thus of the standard IWAE bound. Those analyses shed light on the benefits or lack thereof of these bounds. Lastly, we illustrate our theoretical claims over toy and real-data examples.

Keywords: variational inference, alpha-divergence, importance weighted auto-encoder, weight collapse, high dimension

1. Introduction

Variational inference methods aim at finding the best approximation to a target posterior density within a so-called variational family of probability densities. This best approximation is traditionally obtained by minimizing the exclusive Kullback–Leibler divergence (Wainwright and Jordan, 2008; Blei et al., 2017), however this divergence is known to have some drawbacks (for instance variance underestimation, see Minka, 2005).

As a result, alternative divergences have been explored (Minka, 2005; Li and Turner, 2016; Bui et al., 2016; Dieng et al., 2017; Li and Gal, 2017; Wang et al., 2018; Daudel et al., 2021, 2023; Daudel and Douc, 2021; Rodríguez-Santana and Hernández-Lobato, 2022), in particular the class of alpha-divergences. This family of divergences is indexed by a scalar α. It provides additional flexibility that can in theory be used to overcome the obstacles associated to the exclusive Kullback–Leibler divergence (which is recovered by letting $\alpha \to 1$).

[*]. Equal contribution

Among those methods, techniques involving the Variational Rényi (VR) bound introduced in Li and Turner (2016) have led to promising empirical results and have been linked to key algorithms such as the importance weighted auto-encoder (IWAE) algorithm (Burda et al., 2016) in the special case $\alpha = 0$ and the black-box alpha (BB-α) algorithm (Hernandez-Lobato et al., 2016).

Yet methods based on the VR bound are seen as lacking theoretical guarantees. This comes from the fact that they are classified as biased in the community: by selecting the VR bound as the objective function, those methods indeed resort to biased gradient estimators (Li and Turner, 2016; Hernandez-Lobato et al., 2016; Bui et al., 2016; Li and Gal, 2017; Geffner and Domke, 2020, 2021; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022).

Geffner and Domke (2020) have recently provided insights from an empirical perspective regarding the magnitude of the bias and its impact on the outcome of the optimization procedure when the (biased) reparameterized gradient estimator of the VR bound is used. They observe that the resulting algorithm appears to require an impractically large amount of computations to actually optimise the VR bound as the dimension increases (and otherwise seems to simply return minimizers of the exclusive Kullback–Leibler divergence). They postulate that this effect might be due to a weight degeneracy behavior (Bengtsson et al., 2008), but this behavior is not quantified precisely from a theoretical point of view.

In this paper, our goal is to (i) develop theoretical guarantees for VR-based variational inference methods and (ii) construct a theoretical framework elucidating the weight degeneracy behavior that has been empirically observed for those techniques. The rest of this paper is organized as follows:

- In Section 2, we provide some background notation and we review the main concepts behind the VR bound.

- In Section 3, we introduce the VR-IWAE bound. We show in Proposition 1 that this bound, previously defined by Li and Turner (2016) as the expectation of the biased Monte Carlo approximation of the VR bound, can be actually interpreted as a variational bound which depends on an hyperparameter α with $\alpha \in [0, 1)$. In addition, we obtain that the VR-IWAE bound leads to the same stochastic gradient descent procedure as the VR bound in the reparameterized case. Unlike the VR bound, the VR-IWAE bound relies on *unbiased* gradient estimators and coincides with the IWAE bound for $\alpha = 0$, fully bridging the gap between both methodologies.

  We then generalize the approach of Rainforth et al. (2018)—which characterizes the signal-to-noise ratio (SNR) of the reparameterized gradient estimators of the IWAE—to the VR-IWAE bound and establish that the VR-IWAE bound with $\alpha \in (0, 1)$ enjoys better theoretical properties than the IWAE bound (Theorem 1). To further tackle potential SNR difficulties, we also extend the doubly-reparameterized gradient estimator of the IWAE (Tucker et al., 2019) to the VR-IWAE bound (Theorem 2).

- In Section 4, we provide a thorough theoretical study of the VR-IWAE bound. Following Domke and Sheldon (2018), we start by investigating the case where the dimension of the latent space d is fixed and the number of Monte Carlo samples N in the VR-IWAE bound goes to infinity (Theorem 3). Our analysis shows that the hyperparameter α

allows us to balance between an error term depending on both the encoder and the decoder parameters (θ, ϕ) and a term going to zero at a $1/N$ rate. This suggests that tuning α can be beneficial to obtain the best empirical performances.

However, the relevance of such analysis can be limited for a high-dimensional latent space d (Examples 1 and 2). We then propose a novel analysis where N does not grow as fast as exponentially with d (Theorems 4 and 5) or sub-exponentially with $d^{1/3}$ (Theorem 6), which we use to revisit Examples 1 and 2 in Examples 3 and 4 respectively. This analysis suggests that in these regimes the VR-IWAE bound, and hence in particular the IWAE bound, are of limited interest.

- In Section 5, we detail how our work relates to the existing litterature.
- Lastly, Section 6 provides empirical evidence illustrating our theoretical claims for both toy and real-data examples.

2. Background

Given a model with joint distribution $p_\theta(x, z)$ parameterized by θ, where x denotes an observation and z is a latent variable valued in $\mathbb{R}^d$, one is interested in finding the parameter θ which best describes the observations $\mathcal{D} = \{x_1, \ldots, x_T\}$. This will be our running example. The corresponding posterior density satisfies:

$$p_\theta(\boldsymbol{z}|\mathcal{D}) \propto \prod_{i=1}^{T} p_\theta(x_i, z_i) \tag{1}$$

with $\boldsymbol{z} = (z_1, \ldots, z_T)$, so that the marginal log likelihood reads

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^{T} \ell(\theta; x_i) \quad \text{with} \quad \ell(\theta; x) := \log p_\theta(x) = \log \left(\int p_\theta(x, z) \mathrm{d}z \right). \tag{2}$$

Unfortunately as this marginal log likelihood is typically intractable, finding θ maximizing it is difficult. Variational bounds are then designed to act as surrogate objective functions more amenable to optimization.

Let $q_\phi(z|x)$ be a variational encoder parameterized by ϕ, common variational bounds are the Evidence Lower BOund (ELBO) and the IWAE bound (Burda et al., 2016):

$$\begin{aligned} \mathrm{ELBO}(\theta, \phi; x) &= \int q_\phi(z|x) \log w_{\theta,\phi}(z; x) \ \mathrm{d}z, \\ \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) &= \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x) \right) \mathrm{d}z_{1:N}, \quad N \in \mathbb{N}^\star \end{aligned}$$

where for all $z \in \mathbb{R}^d$,

$$w_{\theta,\phi}(z; x) = \frac{p_\theta(x, z)}{q_\phi(z|x)}.$$

The IWAE bound generalizes the ELBO (which is recovered for $N = 1$) and acts as a lower bound on $\ell(\theta; x)$ that can be estimated in an unbiased manner. Instead of maximizing $\ell(\theta; \mathcal{D})$ defined in (2), one then considers the surrogate objective

$$\sum_{i=1}^{T} \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x_i)$$

which is optimized by performing stochastic gradient descent steps w.r.t. (θ, ϕ) on it combined to mini-batching. Optimizing this objective w.r.t. ϕ is difficult due to high-variance gradients with low signal-to-noise ratio (Rainforth et al., 2018). To mitigate this problem, reparameterized (Kingma and Welling, 2014; Burda et al., 2016) and doubly-reparameterized gradient estimators (Tucker et al., 2019) have been proposed.

Crucially, stochastic gradient schemes on the IWAE bound (and hence on the ELBO) only resort to unbiased estimators in both the reparameterized (Kingma and Welling, 2014; Burda et al., 2016) and the doubly-reparameterized (Tucker et al., 2019) cases, providing theoretical justifications behind those approaches. In particular, under the assumption that z can be reparameterized (that is $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$) and under common differentiability assumptions, the reparameterized gradient w.r.t. ϕ of the IWAE bound is given by

$$\frac{\partial}{\partial \phi} \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) = \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}$$

and the doubly-reparameterized one by

$$\begin{aligned} &\frac{\partial}{\partial \phi} \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) \\ &\quad = \int \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \left(\frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \right)^2 \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x); x)|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}. \quad (3) \end{aligned}$$

Unbiased Monte Carlo estimators of both gradients are hence respectively given by

$$\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \tag{4}$$

and

$$\sum_{j=1}^{N} \left(\frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \right)^2 \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x); x)|_{\phi'=\phi},$$

with $\varepsilon_1, \ldots, \varepsilon_N$ being i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \ldots N$. Maddison et al. (2017) and Domke and Sheldon (2018) in particular established that the variational gap - that is the difference between the IWAE bound and the marginal log-likelihood - goes to zero at a fast $1/N$ rate when the dimension of the latent space d is fixed and the number of samples N goes to infinity.

Another example of variational bound is the Variational Rényi (VR) bound introduced by Li and Turner (2016): it is defined for all $\alpha \in \mathbb{R} \setminus \{1\}$ by

$$\mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) \; w_{\theta,\phi}(z;x)^{1-\alpha} \, \mathrm{d}z \right) \tag{5}$$

and it generalizes the ELBO (which corresponds to the extension by continuity of the VR bound to the case $\alpha = 1$, see Li and Turner, 2016, Theorem 1). It is also a lower (resp. upper) bound on the marginal log-likelihood $\ell(\theta; x)$ for all $\alpha > 0$ (resp. $\alpha < 0$).

In the spirit of the IWAE bound optimisation framework, the VR bound is used for variational inference purposes in (Li and Turner, 2016, Section 4.1, 4.2 and 5.2) to optimise the marginal log-likelihood $\ell(\theta, \mathcal{D})$ defined in (2) by considering the global objective function

$$\sum_{i=1}^{T} \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$$

and by performing stochastic gradient descent steps w.r.t. (θ, ϕ) on it paired up with mini-batching and reparameterization. This VR bound methodology has provided positive empirical results compared to the usual case $\alpha = 1$ and has been widely adopted in the literature (Li and Turner, 2016; Bui et al., 2016; Hernandez-Lobato et al., 2016; Li and Gal, 2017; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022). As discussed in the remark below, this methodology is obviously not limited to the choice of posterior density defined in (1) and is more broadly applicable.

Remark 1 (Black-box alpha energy function) *Let $p_0(z)$ be a prior on a latent variable z valued in $\mathbb{R}^d$ and by $p(x|z)$ the likelihood of the observation x given z, we might consider the posterior density*

$$p(z|\mathcal{D}) \propto p_0(z) \prod_{i=1}^{T} p(x_i|z), \tag{6}$$

leading to the marginal log-likelihood

$$\tilde{\ell}(\mathcal{D}) = \log \left(\int p(\mathcal{D}, z) \mathrm{d}z \right) = \log \left(p_0(z) \prod_{i=1}^{T} p(x_i|z) \mathrm{d}z \right).$$

Here, the latent variable z valued in $\mathbb{R}^d$ is shared across all the observations. Now further assume that the prior density $p_0(z) = \exp(s(z)^T \phi_0 - \log Z(\phi_0))$ has an exponential form, with ϕ_0 and s being the natural parameters and the sufficient statistics respectively and $Z(\phi_0)$ being the normalizing constant ensuring that p_0 is a probability density function.

In order to find the best approximation to the posterior density (6), Hernandez-Lobato et al. (2016) offers to minimize the black-box alpha (BB-α) energy function, which is defined by: for all $\alpha \in \mathbb{R} \setminus \{1\}$,

$$\mathcal{E}(\phi) = \log Z(\phi_0) - \log Z(\tilde{\phi}) - \frac{1}{1-\alpha} \sum_{i=1}^{T} \log \left(\int q_\phi(z) \left(\frac{p(x_i|z)}{f_\phi(z)} \right)^{1-\alpha} \mathrm{d}z \right)$$

where $f_\phi(z) = \exp(s(z)^T \phi)$ is within the same exponential family as the prior and $q_\phi(z) = \exp(s(z)^T \tilde{\phi} - \log Z(\tilde{\phi}))$ with $\tilde{\phi} = T\phi + \phi_0$ denoting the natural parameters of q_ϕ and $Z(\tilde{\phi})$ its normalizing constant. Here, the minimisation is carried out via stochastic gradient descent w.r.t. ϕ combined with mini-batching and reparameterization.

As observed in Li and Gal (2017), minimizing $\mathcal{E}(\phi)$ w.r.t. ϕ is equivalent to maximizing the sum of VR bounds

$$\sum_{i=1}^{T} \frac{T}{1-\alpha} \log \left(\int q_\phi(z) w_{\theta,\phi}(z;x)^{\frac{1-\alpha}{T}} \mathrm{d}z \right)$$

w.r.t. ϕ, where this time $w_{\theta,\phi}(z;x) = p(x_i|z)^T p_0(z)/q_\phi(z)$.

However, the stochastic gradient descent scheme originating from having selected the VR bound as the objective function suffers from one important shortcoming: it relies on biased gradient estimators for all $\alpha \notin \{0,1\}$, meaning that there exists no convergence guarantees for the whole scheme. Indeed, Li and Turner (2016) show that the gradient of the VR bound w.r.t. ϕ satisfies

$$\frac{\partial}{\partial \phi} \mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon},$$

with $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$. The gradient above being intractable, they approximate it using

$$\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k;x)^{1-\alpha}} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x), \tag{7}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \ldots N$. The cases $\alpha = 0$ and $\alpha = 1$ recover the stochastic reparameterized gradients of the IWAE bound (4) and of the ELBO (consider (4) with $N = 1$). As a result, we can trace them back to unbiased stochastic gradient descent schemes for IWAE bound and ELBO optimisation respectively. Yet, this is no longer the case when $\alpha \notin \{0,1\}$, hence impeding the theoretical guarantees of the scheme.

In addition, due to the log function, the VR bound itself can only be approximated using biased Monte Carlo estimators, with (Li and Turner, 2016, Section 4.1) using

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(Z_j;x)^{1-\alpha} \right) \tag{8}$$

where $Z_1, \ldots, Z_N$ are i.i.d. samples generated from q_ϕ. Furthermore, while the VR bound and the IWAE bound approaches are linked via the gradient estimator (7), the VR bound does not recover the IWAE bound when $\alpha = 0$.

The next section aims at overcoming the theoretical difficulties regarding the VR bound mentioned above.

3. The VR-IWAE Bound

For all $\alpha \in \mathbb{R} \setminus \{1\}$, let us introduce the quantity

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}, \tag{9}$$

which we will refer to as the *VR-IWAE bound.* Note that for the VR-IWAE bound to be well-defined we will assume that the following assumption holds in the rest of the paper.

(A1) It holds that $0 < p_\theta(x) < \infty$ and the support of $q_\phi(\cdot|x)$ and of $p_\theta(\cdot|x)$ are equal.

We may omit the dependency on x in $z \mapsto q_\phi(z|x)$ and $z \mapsto w_{\theta,\phi}(z; x)$ for notational convenience and we now make two remarks regarding the VR-IWAE bound defined in (9).

- Contrary to the VR bound, VR-IWAE bound (i) can be approximated using an unbiased Monte Carlo estimator and (ii) recovers the IWAE bound by setting $\alpha = 0$. Under common differentiability assumptions, we also have that

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x)$$

  (see Appendix A.1 for details), meaning that the VR-IWAE bound interpolates between the IWAE bound and the ELBO.

- Li and Turner (2016) interpreted the quantity defined in (9) as the expectation of the biased Monte Carlo approximation of the VR bound (8). They established in (Li and Turner, 2016, Theorem 2) some properties on this quantity. In particular, they showed that (i) for all $\alpha \leq 1$ and all $N \in \mathbb{N}^\star$,

$$\ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_{N+1}^{(\alpha)}(\theta, \phi; x) \leq \mathcal{L}^{(\alpha)}(\theta, \phi; x)$$

  and (ii) for all $\alpha \in \mathbb{R}$, $\ell_N^{(\alpha)}(\theta, \phi; x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta, \phi; x)$ as N goes to infinity if the function $z \mapsto w_{\theta,\phi}(z)$ is assumed to be bounded.

Based on the two previous remarks, $\ell_N^{(\alpha)}(\theta, \phi; x)$ seems to be an interesting candidate as a variational bound which generalizes the IWAE bound. We take here another perspective on the quantity $\ell_N^{(\alpha)}(\theta, \phi; x)$ by wanting to frame it as a variational bound with ties to the Rényi's α-divergence variational inference methodology of Li and Turner (2016) and to the IWAE bound, hence the name VR-IWAE bound. We now need to check that the VR-IWAE bound can indeed be used as a variational bound for marginal log-likelihood optimisation in the context of our running example.

3.1 The VR-IWAE Bound as a Variational Bound

As underlined in the following proposition, the VR-IWAE bound is a variational bound for all $\alpha \in [0, 1)$ which enjoys properties akin to those obtained for the IWAE bound (Burda et al., 2016) and which becomes looser as α increases.

Proposition 1 (Properties of the VR-IWAE bound) *The following properties hold for the VR-IWAE bound.*

1. *For all $\alpha \in [0,1)$ and all $N \in \mathbb{N}^\star$,*

$$\mathrm{ELBO}(\theta,\phi;x) \leq \ell_N^{(\alpha)}(\theta,\phi;x) \leq \ell_{N+1}^{(\alpha)}(\theta,\phi;x) \leq \mathcal{L}^{(\alpha)}(\theta,\phi;x) \leq \ell(\theta;x). \quad (10)$$

Moreover, if the function $z \mapsto w_{\theta,\phi}(z)$ is bounded, then $\ell_N^{(\alpha)}(\theta,\phi;x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta,\phi;x)$ as N goes to infinity.

2. *For all $\alpha_1, \alpha_2 \in (0,1)$ such that $\alpha_1 > \alpha_2$ and all $N \in \mathbb{N}^\star$,*

$$\ell_N^{(\alpha_1)}(\theta,\phi;x) \leq \ell_N^{(\alpha_2)}(\theta,\phi;x) \leq \ell_N^{(\mathrm{IWAE})}(\theta,\phi;x), \quad (11)$$

where the case of equality is reached if and only if $z \mapsto w_{\theta,\phi}(z)$ is constant for ν-almost all $z \in \mathbb{R}^d$ (with ν denoting the Lebesgue measure).

3. *Further assuming that z can be reparameterized, that is $z = f(\varepsilon,\phi) \sim q_\phi$ where $\varepsilon \sim q$, we have under common differentiability assumptions that*

$$\begin{aligned}&\frac{\partial}{\partial\phi}\ell_N^{(\alpha)}(\theta,\phi;x) \\ &\qquad = \int\int \prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi))\right)\mathrm{d}\varepsilon_{1:N} \quad (12)\end{aligned}$$

and an unbiased estimator of $\partial\ell_N^{(\alpha)}(\theta,\phi;x)/\partial\phi$ is given by

$$\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi)) \quad (13)$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j,\phi)$ for all $j = 1 \ldots N$.

The proof of Proposition 1 is deferred to Appendix A.2 and we now comment on Proposition 1. Observe that both the VR and VR-IWAE bounds share the same estimated reparameterized gradient w.r.t. ϕ, that is (7) is exactly (13), hence they lead to the same stochastic gradient descent algorithm. However, when $\alpha \in (0,1)$, a key difference is that in the VR bound case this estimator is biased while it is unbiased for VR-IWAE bound.

This motivates the VR-IWAE bound as a generalization of the IWAE bound that overcomes the theoretical difficulties of the VR bound, as unbiased gradient estimates provide the convergence of the stochastic gradient descent procedure (under proper conditions on the learning rate). In fact, and as we shall see next, the estimated reparameterized gradient w.r.t. ϕ written in (7) and (13) - that we have now properly justified using the VR-IWAE bound - also enjoys an advantageous signal-to-noise ratio behavior when $\alpha \in (0,1)$.

3.2 Signal-to-noise Ratio (SNR) Analysis

Rainforth et al. (2018) identified some issues associated to using reparameterized gradient estimators of the IWAE bound. They did so by looking at the signal-to-noise ratio (SNR) of those estimates: their main theorem (Rainforth et al., 2018, Theorem 1) shows that while increasing N leads to a tighter IWAE bound and improves the SNR for learning θ, it actually worsens the SNR for learning ϕ.

Let us now investigate if and how the conclusions of (Rainforth et al., 2018, Theorem 1) extend to the VR-IWAE bound. To this end, we first recall the definition of the SNR used in Rainforth et al. (2018). Given a random vector $X = (X_1, \dots, X_L)$ of dimension $L \in \mathbb{N}^\star$, the SNR may be defined as follows:

$$\mathrm{SNR}[X] = \left(\frac{|\mathbb{E}(X_1)|}{\sqrt{\mathbb{V}(X_1)}}, \dots, \frac{|\mathbb{E}(X_L)|}{\sqrt{\mathbb{V}(X_L)}} \right).$$

Writing $\theta = (\theta_1, \dots, \theta_L)$ and $\phi = (\phi_1, \dots, \phi_{L'})$ and with $L, L' \in \mathbb{N}^\star$, we now consider for all $\ell = 1 \dots L$ and all $\ell' = 1 \dots L'$ the unbiased estimates of the reparameterized gradient of the VR-IWAE bound w.r.t. θ_ℓ and w.r.t. $\phi_{\ell'}$ given by: for all $M, N \in \mathbb{N}^\star$ and all $\alpha \in [0, 1)$,

$$\delta^{(\alpha)}_{M,N}(\theta_\ell) = \frac{1}{(1-\alpha)M} \sum_{m=1}^{M} \frac{\partial}{\partial \theta_\ell} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi))^{1-\alpha} \right), \tag{14}$$

$$\delta^{(\alpha)}_{M,N}(\phi_{\ell'}) = \frac{1}{(1-\alpha)M} \sum_{m=1}^{M} \frac{\partial}{\partial \phi_{\ell'}} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi))^{1-\alpha} \right), \tag{15}$$

where $(\varepsilon_{m,j})_{1 \leq m \leq M, 1 \leq j \leq N}$ are i.i.d. samples generated from q and $z_{m,j} = f(\varepsilon_{m,j}, \phi)$ for all $m = 1 \dots M$ and all $n = 1 \dots N$. Note that the link with the reparameterized gradient estimator (13) from Proposition 1 can be made by considering the case $M = 1$ in (15). We then have the following theorem.

Theorem 1 (SNR analysis) *Let $\alpha \in [0, 1)$ and for all $N \in \mathbb{N}^\star$ and all $j = 1 \dots N$, define $\tilde{w}_{1,j} = w_{\theta,\phi}(f(\varepsilon_{1,j}, \phi))$ and $\hat{Z}_{1,N,\alpha} = N^{-1} \sum_{j=1}^{N} \tilde{w}_{1,j}^{1-\alpha}$. Assume that the eighth moments of $\tilde{w}_{1,1}^{1-\alpha}$, $\partial \tilde{w}_{1,1}^{1-\alpha}/\partial\theta_\ell$ and $\partial \tilde{w}_{1,1}^{1-\alpha}/\partial\phi_{\ell'}$ are finite, where ℓ is an integer between 1 and L and ℓ' is an integer between 1 and L'. Furthermore, assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{1,N,\alpha})^4) < \infty$. Lastly, assume that $\partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha})/\partial\theta_\ell \neq 0$ and that*

$$\begin{aligned} &\partial \mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})/\partial\phi_{\ell'} > 0, \quad \textit{if } \alpha = 0 \\ &\partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha})/\partial\phi_{\ell'} \neq 0, \quad \textit{if } \alpha \in (0, 1). \end{aligned} \tag{16}$$

Then, under common differentiability assumptions, the SNR of the VR-IWAE bound reparameterized gradient estimates w.r.t θ_ℓ and w.r.t $\phi_{\ell'}$ defined in (14) and (15) respectively satisfy

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] = \Theta(\sqrt{MN}) \tag{17}$$

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\phi_{\ell'})] = \begin{cases} \Theta(\sqrt{M/N}) & \textit{if } \alpha = 0, \\ \Theta(\sqrt{MN}) & \textit{if } \alpha \in (0, 1). \end{cases} \tag{18}$$

The proof of Theorem 1 can be found in Appendix A.3. Theorem 1 states that for $\alpha \in (0,1)$, the SNR for learning the generative network (θ) and for learning the inference network (ϕ) both improve as N increases, unlike the IWAE bound case $\alpha = 0$ where the second SNR worsens as N increases. This provides theoretical support suggesting that taking $\alpha > 0$ in the VR-IWAE bound may help to ensure a good training signal, thus leading to improved empirical performances compared to the IWAE bound.

In the following, we investigate another way to provide gradient estimators of the VR-IWAE bound with an advantageous SNR behavior in practice.

3.3 Doubly-reparameterized Gradient for the VR-IWAE Bound

To remedy the SNR issue identified in Rainforth et al. (2018), Tucker et al. (2019) proposed a new estimator of the gradient of the IWAE bound (3) under the name doubly-reparameterized gradient estimator. As written in the theorem below, the doubly-reparameterized gradient estimator of the IWAE bound (3) in fact generalizes to the case $\alpha \in (0,1)$.

Theorem 2 (Generalized doubly-reparameterized gradient) *Under common differentiability assumptions and assuming that z can be reparameterized, that is $z = f(\varepsilon, \phi) \sim q_\phi$ where $\varepsilon \sim q$, we have that: for all $\alpha \in [0,1]$,*

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}, \tag{19}$$

with $z_j = f(\varepsilon_j, \phi)$ for all $j = 1 \ldots N$ and

$$h_j(\alpha) = \alpha \ \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} + (1-\alpha) \ \left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \right)^2.$$

An unbiased estimator of $\partial \ell_N^{(\alpha)}(\theta, \phi; x)/\partial \phi$ is then given by

$$\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi} \tag{20}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi)$ for all $j = 1 \ldots N$.

The proof of Theorem 2 is deferred to Appendix A.4. One can then check that we recover the usual doubly-reparameterized gradient estimator of the IWAE bound (resp. the ELBO) when $\alpha = 0$ (resp. $\alpha = 1$). Like the reparameterized gradient estimator (13), which we have studied in Section 3.2 as it corresponds to the special case $M = 1$ in Theorem 1, this second (doubly-reparameterized) gradient estimator too may lead to improved empirical performances. From there, large-scale learning occurs by using that Proposition 1 implies

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^T \ell(\theta; x_i) \geq \sum_{i=1}^T \mathcal{L}^{(\alpha)}(\theta, \phi; x_i) \geq \sum_{i=1}^T \ell_N^{(\alpha)}(\theta, \phi; x_i)$$

and by following the training procedure for the IWAE bound. Indeed, we have access to an unbiased estimator of the lower bound of the full data set $\sum_{i=1}^T \ell_N^{(\alpha)}(\theta, \phi; x_i)$ (as well as an unbiased estimator of its reparameterized/doubly-reparameterized gradient) using mini-batching. Seeking to maximize the objective function $\sum_{i=1}^T \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$ by optimising $\sum_{i=1}^T \ell_N^{(\alpha)}(\theta, \phi; x_i)$ in fact amounts to seeking to minimize a specific Rényi's α-divergence with a mean-field assumption on the variational approximation (see Remark 2 for detail).

Remark 2 *Define $\boldsymbol{z} = (z_1, \ldots, z_T)$ and $q_\phi(\boldsymbol{z}) = \prod_{i=1}^T q_\phi(z_i)$. Then, for all $\alpha \in (0, 1)$:*

$$\begin{aligned}\sum_{i=1}^T \mathcal{L}^{(\alpha)}(\theta, \phi; x_i) &= \sum_{i=1}^T \frac{1}{1-\alpha} \log \left(\int q_\phi(z) w_{\theta,\phi}(z; x_i)^{1-\alpha} \mathrm{d}z \right) \\ &= \sum_{i=1}^T \frac{1}{1-\alpha} \log \left(\int q_\phi(z_i) w_{\theta,\phi}(z_i; x_i)^{1-\alpha} \mathrm{d}z_i \right) \\ &= \frac{1}{1-\alpha} \log \left(\int \int \prod_{i=1}^T q_\phi(z_i) \prod_{j=1}^T w_{\theta,\phi}(z_j; x_j)^{1-\alpha} \mathrm{d}z_{1:T} \right) \\ &= \frac{1}{1-\alpha} \log \left(\int q_\phi(\boldsymbol{z}) w_{\theta,\phi}(\boldsymbol{z}; \mathcal{D})^{1-\alpha} \mathrm{d}\boldsymbol{z} \right)\end{aligned}$$

where $p_\theta(\mathcal{D}, \boldsymbol{z}) = \prod_{i=1}^T p(x_i, z_i)$ and $w_{\theta,\phi}(\boldsymbol{z}; \mathcal{D}) = p_\theta(\mathcal{D}, \boldsymbol{z})/q_\phi(\boldsymbol{z})$. Observe that the last equality is a VR bound, meaning that maximizing the global objective function $\sum_{i=1}^T \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$ is equivalent to minimizing the Rényi's α-divergence between the two probability distributions with associated probability densities $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z}|\mathcal{D})$ respectively w.r.t. the Lebesgue measure. Hence, this approach belongs to alpha-divergence variational inference methods with the particularity that it makes a mean-field assumption on the variational approximation $q_\phi(\boldsymbol{z})$.

At this stage, we have formalized and motivated the VR-IWAE bound. We now want to get an understanding of its theoretical properties.

4. Theoretical Study of the VR-IWAE Bound

The starting point of our approach is to exploit the fact that prior theoretical works study the particular case $\alpha = 0$ (corresponding to the IWAE bound) when the dimension of the latent space $\dim(z) = d$ is fixed and the number of samples N goes to infinity.

4.1 Behavior of the VR-IWAE Bound when d is Fixed and N goes to Infinity

A quantity that has been of interest to assess the quality of the IWAE bound is the *variational gap*, which is defined as the difference between the IWAE bound and the marginal log-likelihood:

$$\Delta_N(\theta, \phi; x) := \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) - \ell(\theta; x) = \int \int \prod_{i=1}^N q_\phi(z_i) \log \left(\frac{1}{N} \sum_{j=1}^N \overline{w}_{\theta,\phi}(z_j) \right) \mathrm{d}z_{1:N} \quad (21)$$

where for all $z \in \mathbb{R}^d$

$$\overline{w}_{\theta,\phi}(z) := \frac{w_{\theta,\phi}(z)}{\mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z))} = \frac{w_{\theta,\phi}(z)}{p_\theta(x)},$$

so that $\overline{w}_{\theta,\phi}(z_1), \dots, \overline{w}_{\theta,\phi}(z_N)$ correspond to the relative weights. The analysis of the variational gap (21), first performed in Maddison et al. (2017) and then refined in Domke and Sheldon (2018), investigated the case where $\dim(z) = d$ is fixed and N goes to infinity. Informally, they obtained in their Theorem 3 that the variational gap behaves as follows

$$\Delta_N(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

with γ_0 denoting the variance of the relative weights, that is

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}(Z)).$$

This result suggests that using N is very beneficial to reduce the variational gap, as it goes to zero at a fast $1/N$ rate. It motivates a study - in a regime where d is fixed and N goes to infinity - of the more general variational gap defined for all $\alpha \in [0, 1)$ by

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x).$$

The following result generalizes (Domke and Sheldon, 2018, Theorem 3) to the VR-IWAE bound.

Theorem 3 *Let $\alpha \in [0, 1)$. Then, it holds that*

$$0 < \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha}) < \infty. \tag{22}$$

Further assume that there exists $\beta > 0$ such that

$$\mathbb{E}_{Z \sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta}) < \infty, \tag{23}$$

where we have defined $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$. Lastly, assume that the following condition holds

$$\limsup_{N \to \infty} \mathbb{E}(1/R_{\alpha,N}) < \infty, \tag{24}$$

where, for all $N \in \mathbb{N}^\star$, $R_{\alpha,N} = N^{-1} \sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha}$ and $Z_1, \dots, Z_N$ are i.i.d. samples generated according to q_ϕ. Then, denoting $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$, we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right). \tag{25}$$

The proof of this result is deferred to Appendix B.1 and we now aim at interpreting Theorem 3, starting with the conditions (23) and (24).

4.1.1 Conditions (23) and (24)

A first remark is that the conditions (23) and (24) stated in Theorem 3 exactly generalize the ones from (Domke and Sheldon, 2018, Theorem 3), which are recovered by setting $\alpha = 0$. This then prompt us to investigate in the following proposition how restrictive the conditions (23) and (24) are as a function of α.

Proposition 2 *Let $\alpha_1, \alpha_2 \in [0,1)$ with $\alpha_1 > \alpha_2$. Then, the two following assertions hold.*

1. *If* (23) *holds with $\alpha = \alpha_2$, then* (23) *holds with $\alpha = \alpha_1$.*
2. *If* (24) *holds with $\alpha = \alpha_2$, then* (24) *holds with $\alpha = \alpha_1$.*

The proof of this result is deferred to Appendix B.2. It notably relies the fact that the condition (24) is equivalent to the statement that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}(1/R_{\alpha,N}) < \infty$, which follows from Lemma 4 in Appendix A.3 with $k = 1$. Notice that this provides an interesting equivalent condition to (24) that might be easier to check empirically.

Proposition 2 then states that the conditions (23) and (24) with $\alpha = \alpha_1$ are at worse as restrictive as the case $\alpha = \alpha_2$, where $\alpha_1 > \alpha_2$. Putting this into perspective with Domke and Sheldon (2018), the conditions (23) and (24) when $\alpha > 0$ are hence not more restrictice than the conditions presented in (Domke and Sheldon, 2018, Theorem 3) for the more usual IWAE bound case $\alpha = 0$. In fact, one would even be inclined to think that those conditions become easier to satisfy as α increases, motivating once again the use of $\alpha \in (0,1)$ in practice to be in the conditions of application of Theorem 3.

4.1.2 Interpreting (25)

Under the assumptions of Theorem 3, (25) states: for all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

The variational gap $\Delta_N^{(\alpha)}(\theta,\phi;x)$ is hence composed of two main terms:

- A term going to zero at a $1/N$ rate that depends on γ_α^2. Here γ_α^2 is controlled thanks to (23), as (23) implies that $\mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) < \infty$ or equivalently that $\gamma_\alpha^2 < \infty$.
- An error term $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$. This term decreases away from zero as α increases due to the fact that $\mathcal{L}^{(\alpha)}(\theta,\phi;x)$ decreases away from its upper bound $\ell(\theta;x)$ as α increases (see for example Li and Turner, 2016, Theorem 1). It is equal to zero when $\alpha = 0$ or when the posterior and the encoder distributions are equal to one another.

  Unless $\alpha = 0$ or the posterior and encoder distributions are matching, the error term $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ hence maintains a dependency in (θ,ϕ) in the variational gap even as N goes to infinity. This is coherent with Theorem 1, in the sense that the case $\alpha \in (0,1)$ might ensure a better learning of both θ and ϕ in practice compared to the case $\alpha = 0$ (as the latter does not keep a dependency in ϕ as N goes to infinity).

Since the error term $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$ is decreasing away from zero as α increases and the term going to zero at a $1/N$ rate depends on the behavior of γ_α^2 (with γ_α^2 going to 0 as α goes to 1, see Lemma 5 of Appendix B.3), there might then be a tradeoff to achieve when choosing α in order to obtain the best empirical performances.

To the best of our knowledge, and by appealing to the link between the VR-bound and the VR-IWAE bound methodologies established in Section 3.1, Theorem 3 is the first result shedding light via (25) on how the quantity γ_α^2 alongside with the error term

$\mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x)$ may play a role to guarantee the success of gradient-based methods involving the VR-bound. While the result obtained in Theorem 3 is encouraging and might further motivate the use of $\alpha \in (0,1)$ in practice, one may seek to identify potential limitations of Theorem 3.

4.1.3 Limitations of Theorem 3

To investigate the limitations of Theorem 3, let us provide below two insightful examples in which all the terms appearing in (25) are tractable.

Example 1 *Let $\sigma > 0$, $S_1,\dots,S_N$ be i.i.d. normal random variables and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1),\dots,\ \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form*

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad i = 1\dots N, \tag{26}$$

where the relationship between mean and variance ensures that the relative weights have expectation 1. Then, we can apply Theorem 3: for all $\alpha \in [0,1)$,

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right]-1}{1-\alpha}.$$

In particular, we can write the weights under the form (26) with $\sigma = 1$ by setting $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$, $q_\phi(z|x) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$, $\theta = 0\cdot\boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix and $\boldsymbol{u}_d$ the d-dimensional vector whose coordinates are all equal to 1.

The proof of Example 1 is deferred to Appendix B.4 and we now comment on Example 1. A first comment is that as α increases, the error term $\mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x)$ worsens linearly with α while γ_α^2 decreases with α, which supports our claim that there might exist an optimal α that balances between the two terms appearing in the variational gap as a rule of thumb.

Furthermore, the variance of the relative weights and more generally γ_α^2 is exponential with d. This means that the analysis of Domke and Sheldon (2018)—that we extended to $\alpha \in [0,1)$ in Theorem 3—may not capture what is happening in some high-dimensional scenarios as we may never use N large enough in high-dimensional settings for the asymptotic regime of Theorem 3 to kick in. We now present our second example.

Example 2 *We consider the linear Gaussian example from Rainforth et al. (2018), that is $p_\theta(z) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x;z,\boldsymbol{I}_d)$ with $\theta \in \mathbb{R}^d$, and $q_\phi(z|x) = \mathcal{N}(z;Ax+b, 2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\phi = (\tilde{a},b) \in \mathbb{R}^d \times \mathbb{R}^d$. Here, the optimal parameter values $(\theta^\star,\phi^\star)$ are given by $\theta^\star = T^{-1}\sum_{t=1}^T x_t$ and $\phi^\star = (a^\star,b^\star)$ with $a^\star = 1/2\boldsymbol{u}_d$ and $b^\star = \theta^\star/2$ (see Rainforth et al., 2018, Appendix B). Furthermore, the true marginal likelihood and true posterior density are given by $p_\theta(x) = \mathcal{N}(x;\theta,2\boldsymbol{I}_d)$ and $p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2, 1/2\ \boldsymbol{I}_d)$ respectively. Then, we can apply Theorem 3: for all $\alpha \in [0,1)$,*

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right),$$

with

$$\mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x)=\frac{d}{2}\left[\log\left(\frac{4}{3}\right)+\frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right]-\frac{3\alpha}{4-\alpha}\left\|Ax+b-\frac{\theta+x}{2}\right\|^2$$
$$\gamma_\alpha^2=\frac{1}{1-\alpha}\left[(4-\alpha)^d(15-6\alpha)^{-\frac{d}{2}}\exp\left(\frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)}\left\|Ax+b-\frac{\theta+x}{2}\right\|^2\right)-1\right].$$

The proof of Example 2 is deferred to Appendix B.5. To interpret Example 2, observe that the case of optimality is particularly telling in this example, since when $(\theta,\phi)=(\theta^\star,\phi^\star)$ it holds that $\gamma_\alpha^2=(1-\alpha)^{-1}[(4-\alpha)^d(15-6\alpha)^{-d/2}-1]$ and γ_α^2 is thus exponential in d despite the parameters (θ,ϕ) being optimal for the setting considered.

Hence, and in line with our conclusions for Example 1, the relevance of Theorem 3 can be limited for a high-dimensional latent space d. This calls for an in-depth study of the variational gap as both d and N go to infinity.

4.2 Behavior of The VR-IWAE Bound when both d and N go to Infinity

To better capture what is happening to the VR-IWAE bound in high-dimensional scenarios, we now let $d,N\to\infty$ in the variational gap

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x):=\ell_{N,d}^{(\alpha)}(\theta,\phi;x)-\ell_d(\theta;x),$$

where we have emphasized notationally the dependence on d in the VR-IWAE bound (9), the log-likelihood (2) and in the variational gap. We will consider the two cases:

$$\text{(i)}\quad d,N\to\infty\quad\text{with}\quad\frac{\log N}{d}\to 0,$$
$$\text{(ii)}\quad d,N\to\infty\quad\text{with}\quad\frac{\log N}{d^{1/3}}\to 0,$$

that is, N grows slower than exponentially with d as in (i) or slower than sub-exponentially with $d^{1/3}$ as in (ii). As we shall see, those two cases will rely on a different set of assumptions each in order to carry out the analysis. In both scenarios, we will prove that a single importance weight dominates all the others, which strongly impacts the variational gap. To this end, let us rewrite the variational gap $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$ under a more convenient form. Writing $\overline{w}_i=\overline{w}_{\theta,\phi}(z_i)$ for all $i=1\ldots N$, we first re-order the weights $\overline{w}_1,\ldots,\overline{w}_N$ as

$$\overline{w}^{(1)}<\overline{w}^{(2)}<\cdots<\overline{w}^{(N-1)}<\overline{w}^{(N)},$$

where we have made the assumption that the weights have no tie almost surely. Now denoting by $q_\phi^{(N)}$ the density of $\overline{w}^{(N)}$ and defining for all $\alpha\in[0,1)$

$$T_{N,d}^{(\alpha)}=\sum_{j=1}^{N-1}\left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}}\right)^{1-\alpha}\tag{27}$$

(we have dropped the dependency in x appearing in $T_{N,d}^{(\alpha)}$ for notational ease here), we then have the following proposition.

Proposition 3 *For all $\alpha \in [0,1)$, the variational gap $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$ can be rewritten as*

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) = \Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x) + R_{N,d}^{(\alpha)}(\theta,\phi;x) \tag{28}$$

where

$$\Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x) = \int q_\phi^{(N)}(\overline{w}^{(N)}) \log\left(\overline{w}^{(N)}\right) \mathrm{d}\overline{w}^{(N)} + \frac{\log N}{\alpha - 1} \tag{29}$$

$$0 \leq R_{N,d}^{(\alpha)}(\theta,\phi;x) \leq \frac{1}{1-\alpha}\mathbb{E}(T_{N,d}^{(\alpha)}). \tag{30}$$

The proof of Proposition 3 can be found in Appendix B.6. To continue the analysis, the key intuition will be that the log weights typically satisfy a central limit theorem (CLT), hence the weights are approximately log-normal as the dimension d increases. One such case for instance arises when the posterior and variational distributions are such that the log weights satisfy

$$\log \overline{w}_i = \sum_{j=1}^{d} X_{i,j}, \quad i = 1 \ldots N, \tag{31}$$

where, for all $i = 1 \ldots N$, $X_{i,1}, \ldots X_{i,d}$ are i.i.d. random variables and $\mathbb{E}(\exp(\sum_{j=1}^d X_{i,j})) = 1$ (since the relative weights satisfy $\mathbb{E}(\overline{w}_i) = 1$). Indeed, denoting $\xi_{i,j} = -(X_{i,j} - \mathbb{E}(X_{1,1}))$, $\sigma^2 = \mathbb{V}(\xi_{1,1})$ and $S_i = \sum_{j=1}^d \xi_{i,j}/(\sigma\sqrt{d})$, (31) can equivalently be rewritten as

$$\log \overline{w}_i = -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) - \sigma\sqrt{d}S_i, \quad i = 1 \ldots N, \tag{32}$$

where under the assumption that $\sigma^2 < \infty$, S_i converges in distribution to the standard normal distribution by the CLT for all $i = 1 \ldots N$. Consequently, the distribution of the weights originating from (32) can be approximated in high-dimensional settings by the log-normal distribution from Example 1, that is

$$\log \overline{w}_i = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \ldots N.$$

For this reason, we first show in the following how the rest of the analysis unfolds when the distribution of the weights is assumed to be exactly log-normal. We will then use this analysis as a stepping stone to treat the more general case where the distribution of the weights is approximately log-normal of the form (32).

4.2.1 Log-normal Distribution Assumption for the Weights

Let $S_1, \ldots, S_N$ be i.i.d. random variables and let the weights $\overline{w}_1, \ldots, \overline{w}_N$ be of the form

$$\log \overline{w}_i = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \ldots N, \tag{33}$$

that is we consider the case where the distribution of the weights is log-normal. Let $S^{(1)} \leq \ldots \leq S^{(N)}$ denote the ordered sequence of $S_1, \ldots, S_N$ and recall that $\overline{w}^{(1)} \leq \ldots \leq \overline{w}^{(N)}$ denotes the ordered sequence of $\overline{w}_1, \ldots, \overline{w}_N$. We then have the following lemma, which provides asymptotic results on the expectation of $S^{(1)}$ as $N \to \infty$.

Lemma 1 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Then,*

$$\mathbb{E}(S^{(1)}) = -\sqrt{2 \log N} + O\left(\frac{\log \log N}{\sqrt{\log N}}\right). \tag{34}$$

The proof of this lemma can be found in Appendix B.7.1. Intuitively, Lemma 1 will serve as the basis to study the two terms appearing in Equation (28) of Proposition 3, as both $\Delta_{N,d}^{(\alpha,MAX)}(\theta, \phi; x)$ and $\mathbb{E}(T_{N,d}^{(\alpha)})$ depend on $S^{(1)}$ through the relation

$$\log \overline{w}^{(N)} = -\frac{\sigma^2 d}{2} - \sigma \sqrt{d} S^{(1)}.$$

From there, we can derive the two propositions below.

Proposition 4 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0, 1)$,*

$$\lim_{N,d \to \infty} \Delta_{N,d}^{(\alpha,MAX)}(\theta, \phi; x) + \frac{d\sigma^2}{2} \left(1 - 2\sqrt{\frac{2 \log N}{d\sigma^2}} + \frac{1}{1-\alpha} \frac{2 \log N}{d\sigma^2} + O\left(\frac{\log \log N}{\sqrt{d \log N}}\right)\right) = 0.$$

Proposition 5 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0, 1)$, we have*

$$\lim_{\substack{N,d \to \infty \\ \log N/d \to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0. \tag{35}$$

The proof of these two propositions are deferred to Appendix B.7.2 and Appendix B.7.3 respectively. Importantly, Proposition 5 implies that the largest weight $\overline{w}^{(N)}$ converges to 1 in probability, meaning that there is a weight collapse when $N, d \to \infty$ with $\log N/d \to 0$ (following the definition of weight collapse given in Bengtsson et al., 2008). By using (35) with $\alpha = 0$, this weight collapse indeed follows from Markov's inequality (in order to get that $T_{N,d}^{(0)}$ converges to 0 in probability) combined with the fact that $\overline{w}^{(N)} = (1 + T_{N,d}^{(0)})^{-1}$.

Building on Proposition 3, Proposition 4 and Proposition 5, we now deduce the following theorem, which describes the asymptotic behavior of the variational gap as $N, d \to \infty$ in the log-normal distribution case for values of α in $[0, 1)$.

Theorem 4 (i.i.d. normal random variables) *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0, 1)$, we have*

$$\lim_{\substack{N,d \to \infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{d\sigma^2}{2} \left(1 - 2\sqrt{\frac{2 \log N}{d\sigma^2}} + \frac{1}{1-\alpha} \frac{2 \log N}{d\sigma^2} + O\left(\frac{\log \log N}{\sqrt{d \log N}}\right)\right) = 0.$$

While Theorem 4 states that increasing N decreases the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$ for N large enough, it does so by a factor which is negligible compared to the term $-d\sigma^2/2$. This is in sharp contrast to Theorem 3 and more specifically to Example 1, which predicts that for log-normal weights the variational gap decreases in $1/N$ in the fixed d, large N regime.

Contrary to Example 1, the term $-d\sigma^2/2$ does not depend on α here. In fact, by taking the expectation in (33), $\mathrm{ELBO}_d(\theta,\phi;x) - \ell_d(\theta;x) = -d\sigma^2/2$, meaning that the following approximation of the variational gap in the context of Theorem 4 holds: for all $\alpha \in [0,1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) \approx \mathrm{ELBO}_d(\theta,\phi;x) - \ell_d(\theta;x), \quad \text{as } N,d\to\infty \text{ with } \tfrac{\log N}{d}\to 0.$$

Hence, Theorem 4 shows that in high-dimensional scenarios and under the log-normal distribution assumption (33), we cannot expect to gain much from the VR-IWAE bound unless N grows exponentially with d, in the sense that the improvement is negligible compared to the ELBO. This result holds *for all values of α in* $[0,1)$, thus it holds for the IWAE bound ($\alpha = 0$) as well.

We obtain the following slightly more general result by building on the proof of Theorem 4.

Theorem 5 (General i.i.d. normal random variables) *Let $S_1,\dots,S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1,\dots,\overline{w}_N$ satisfy*

$$\log \overline{w}_i = -\frac{B_d^2}{2} - B_d S_i, \quad i = 1\dots N, \tag{36}$$

and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_-\sqrt{d}$. Then, for all $\alpha \in [0,1)$, we have

$$\lim_{\substack{N,d\to\infty \\ \log N/d\to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{B_d^2}{2}\left\{1 - 2\,\frac{\sqrt{2\log N}}{B_d} + \frac{1}{1-\alpha}\,\frac{2\log N}{B_d^2} + O\left(\frac{\log\log N}{B_d\sqrt{\log N}}\right)\right\} = 0.$$

The proof of Theorem 5 can be found in Appendix B.7.4. We now revisit the Gaussian example given in Example 1 in the context of Theorem 5.

Example 3 *Set $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$, with $\theta,\phi \in \mathbb{R}^d$. Denoting $B_d = \|\theta-\phi\|$, we can write the weights $\overline{w}_1,\dots,\overline{w}_N$ under the form* (36) *(see* (81) *of Appendix B.4). Hence, Theorem 5 applies if there exists $\sigma_- > 0$ such that $B_d \geq \sigma_-\sqrt{d}$. This is for example the case if $\theta = 0\cdot\boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$ with $\sigma_- = 1$.*

As we shall see next, our conclusion regarding the behavior of the VR-IWAE bound in high-dimensional settings extends to cases where the log-normal assumption does not necessarily hold exactly, that is if we assume instead that (32) holds, where $S_1,\dots S_N$ are i.i.d. random variables whose distribution is close to a normal as $N,d\to\infty$.

4.2.2 Beyond the Log-normal Distribution Assumption

Following (32), let us set

$$\log \overline{w}_i = -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) - \sigma\sqrt{d}S_i, \quad i = 1\dots N, \tag{37}$$

where the i.i.d. random variables $S_1,\dots,S_N$ are defined as follows:

$$S_i = \frac{1}{\sigma\sqrt{d}}\sum_{j=1}^{d}\xi_{i,j}, \quad i = 1\dots N. \tag{38}$$

The assumption (A2) below ensures that $S_1,\dots S_N$ have a distribution that is close to a normal as $N,d\to\infty$, so that (33) is recovered in the limit.

(A2) For all $i = 1 \dots N$,

(a) $\xi_{i,1}, \dots, \xi_{i,d}$ are i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure and satisfy $\mathbb{E}(\xi_{i,1}) = 0$ and $\mathbb{V}(\xi_{i,1}) = \sigma^2 < \infty$.

(b) There exists $K > 0$ such that:

$$|\mathbb{E}(\xi_{i,1}^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Here, the condition (A2)b corresponds to the well-known Bernstein condition. Paired up with (A2)a, this condition permits us to appeal to classical limit theorems for large deviations in order to enlarge the so-called zone of normal convergence beyond the CLT (Petrov, 1995; Saulis and Statulevičius, 2000). This enables us to establish preliminary results which are used to prove the results of Section 4.2.2 we will now present (we refer to Appendix B.8.2 for the statement of those preliminary results). We first provide the equivalent of Lemma 1 in the more general context of (38) and under (A2).

Lemma 2 *Assume* (A2). *Let $S_1, \dots, S_N$ be as in* (38). *Then, as $N, d \to \infty$, with $\frac{\log N}{d^{1/3}} \to 0$,* (34) *holds.*

The proof of this result is deferred to Appendix B.8.3. Notice that we are now assuming that N grows slower than sub-exponentially with $d^{1/3}$ in Lemma 2. The following two propositions give results akin to those obtained in Proposition 4 and Proposition 5.

Proposition 6 *Assume* (A2). *Let $S_1, \dots, S_N$ be as in* (38). *Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (37). *Then, setting*

$$a := \log \mathbb{E}(\exp(-\xi_{1,1})), \tag{39}$$

we have that $a > 0$ and that for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \Delta_{N,d}^{(\alpha, MAX)}(\theta, \phi; x) + da \left\{ 1 - \frac{\sigma}{a} \sqrt{\frac{\log N}{d}} + O\left(\frac{\log \log N}{\sqrt{d \log N}} \right) \right\} = 0.$$

Proposition 7 *Assume* (A2). *Let $S_1, \dots, S_N$ be as in* (38). *Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (37). *Then, for all $\alpha \in [0, 1)$,*

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0.$$

The proof of Proposition 6 and Proposition 7 can be found in Appendix B.8.4 and Appendix B.8.5 respectively.

Remark 3 *The log-normal case corresponds to setting $a = \sigma^2/2$ in Proposition 6 (this can be checked using the definition of a in* (39) *combined with* (94) *from the proof of Proposition 6 in Appendix B.8.4). Contrary to Proposition 4, the $(1-\alpha)^{-1}(d\sigma^2)^{-1} 2 \log N$ term is now subsumed by the final $O(\log \log N / \sqrt{d \log N})$ term in Proposition 6, which comes from the fact that Proposition 6 makes the additional assumption $\log N / d^{1/3} \to 0$ as $N, d \to \infty$. Hence, Proposition 4 and Proposition 6 agree with each other in the log-normal case.*

Proposition 3, Proposition 6 and Proposition 7 lead to the theorem below, which characterizes the asymptotics of the variational gap as $N, d \to \infty$ for $\alpha \in [0, 1)$ in the more general case where the distribution of the weights is approximately log-normal according to (37).

Theorem 6 (i.i.d. random variables) *Assume* (A2)*. Let $S_1, \ldots, S_N$ be as in* (38)*. Further assume that the weights $\overline{w}_1, \ldots, \overline{w}_N$ satisfy* (37) *and let $a > 0$ be defined as in* (39)*. Then, for all $\alpha \in [0, 1)$,*

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da \left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d \log N}}\right)\right) = 0.$$

We have thus obtained that, under the assumptions of Theorem 6, the VR-IWAE bound is of limited interest for all values of $\alpha \in [0, 1)$ unless N grows at least sub-exponentially with $d^{1/3}$. In fact, by taking the expectation in the expression of the log-weights, we have that $\mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x) = -da$ (using for example (95) from the proof of Proposition 6 in Appendix B.8.4). Hence, the following approximation of the variational gap holds in the context of Theorem 6: for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) \approx \mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x), \quad \text{as } N, d \to \infty \text{ with } \tfrac{\log N}{d^{1/3}} \to 0.$$

Since the weights are assumed to be *approximately* log-normal this time as opposed to Section 4.2.1, the condition that N should grow at least exponentially with d to avoid a weight collapse effect has now been replaced by the less restrictive yet still stringent condition that N should grow at least sub-exponentially with $d^{1/3}$.

As described below, the assumptions on the distribution of the weights—that is, on the ratio between the posterior and the variational distributions—appearing in Theorem 6 are met for the linear Gaussian setting from Example 2.

Example 4 *We consider the linear Gaussian setting from Rainforth et al. (2018) that we recalled in Example 2. Denoting $\lambda = \|\frac{\theta+x}{2} - Ax - b\|/\sqrt{d}$, the weights can be written in the form of* (37) *with $\sigma^2 = 1/18 + 8/3\lambda^2$ and we also have $a = \lambda^2 + 1/6 + 1/2\log(3/4)$. As a result, we can apply Theorem 6 if* (A2) *holds. This is for example the case at optimality when $(\theta, \phi) = (\theta^\star, \phi^\star)$ (and the derivation details for this example can be found in Appendix B.8.6).*

Example 4 states that we are in the conditions of application of Theorem 6 when the parameters (θ, ϕ) are optimal, with corresponding optimal posterior density $p_{\theta^\star}(z|x) = \mathcal{N}(z; (\theta^\star + x)/2, 1/2\boldsymbol{I}_d)$ and optimal variational density $q_{\phi^\star}(z|x) = \mathcal{N}(z; (\theta^\star + x)/2, 2/3\boldsymbol{I}_d)$.

This example showcases how, in some instances where the variational family is not large enough to contain the target density, there can be a weight collapse phenomenon as d increases that severely impacts the VR-IWAE bound, even when the parameters (θ, ϕ) are set to be the optimal ones for the problem considered. This concludes our theoretical study of the VR-IWAE bound, which sheds lights on the conditions behind the success or failure of this bound. In the next section, we describe how our theoretical results relate to the existing literature.

5. Related Work

Alpha-divergence variational inference. Our work provides the theoretical grounding behind VR-bound gradient-based schemes (Hernandez-Lobato et al., 2016; Bui et al., 2016; Li and Turner, 2016; Dieng et al., 2017; Li and Gal, 2017; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022). It also unifies the VR and IWAE bound methodologies (Burda et al., 2016; Rainforth et al., 2018; Tucker et al., 2019; Domke and Sheldon, 2018; Maddison et al., 2017) and serves as a foundation for improving on both methodologies.

Proof techniques. Several of our theoretical results generalize known findings from the literature in order to build the VR-IWAE bound methodology and to characterize its asymptotics. Some of our proofs are straightforwardly derived from existing ones, such as the proofs of Theorems 2 and 3 (which are established by directly adapting the proofs written in Tucker et al. (2019) and in Domke and Sheldon (2018) respectively). However, a number of our proof techniques differs significantly from/alter parts of known proofs (see Appendix C for details). Lastly, the derivations made in Section 4.2 for the asymptotics of the VR-IWAE bound when $N, d \to \infty$ are, to the best of our knowledge, the first of their kind.

Importance sampling. Common variational bounds and their gradients can often be expressed in terms of the importance weights $w_{\theta,\phi}$ (with our novel VR-IWAE bound being no exception to that rule). As such, the success of gradient-based variational inference has been known to depend on the behavior of the importance weights and there has been a growing interest in understanding this behavior through the use of insights and tools from the importance sampling (IS) literature (Maddison et al., 2017; Domke and Sheldon, 2018; Dhaka et al., 2021; Geffner and Domke, 2021). In particular, it is well-known that IS can perform poorly in high dimensions unless the target and reference/proposal distributions are close.

Picklands III (1975) for instance showed that, under commonly satisfied assumptions, the right tail of the importance weights distribution approximates a generalized Pareto distribution, that is, a heavy-tailed distribution with three parameters (u, σ, k) and moments of order up to $\lfloor 1/k \rfloor$. This behavior is typical in high dimensions and it makes IS fail, as the IS estimators are dominated by the few largest terms. Leveraging this result, Dhaka et al. (2021) considered the case of black-box variational inference and viewed the importance weights as approximately drawn from a generalized Pareto distribution with tail index k. The importance weights taken to an exponent $1 - \alpha$ are then approximately distributed according to a generalized Pareto distribution with tail index $(1 - \alpha)k$ and they deduced that the estimates should be more stable as α increases towards 1 due to lighter tails.

The analysis from Dhaka et al. (2021) goes hand in hand with our findings, as (i) Theorems 1 and 3 predict improvements in terms of SNR and variance as α increases, at the cost of an increasing bias and (ii) our results from Section 4.2 show that, as d increases, the VR-IWAE bound fails regardless of the value of $\alpha \in [0, 1)$ and provides negligible improvements compared to the ELBO ($\alpha = 1$). However, one main specificity of our work is our precise characterization of how the distribution of the importance weights impacts the tightness of the VR-IWAE bound. Specifically, Theorem 3 generalizes Domke and Sheldon (2018) to the VR-IWAE bound, while the results from Section 4.2 provide the first theoretical justification behind the empirical findings from Geffner and Domke (2021) regarding the impact of weight collapse on the tightness of variational bounds.

The next section is devoted to illustrating the theoretical claims we have made thus far over toy and real-data experiments.

6. Numerical Experiments

In this section, our goal is to verify the validity of the theoretical results we established over several numerical experiments, starting with a Gaussian example in which the distribution of the weights is exactly log-normal.

6.1 Gaussian Example

We consider the Gaussian example described in Example 3, for which the weights $\overline{w}_1, \dots, \overline{w}_N$ can be written under the form (36) with $B_d = \|\theta - \phi\|$, meaning that the distribution of the weights is log-normal. On the one hand, Theorem 3 predicts that for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = -\frac{\alpha B_d^2}{2} - \frac{\exp\left[(1-\alpha)^2 B_d^2\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right) \tag{40}$$

(this follows from a straightforward adaptation of Example 1). On the other hand, Theorem 5 tells us that if there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$, then: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d \to \infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{B_d^2}{2}\left\{1 - 2\,\frac{\sqrt{2\log N}}{B_d} + \frac{1}{1-\alpha}\,\frac{2\log N}{B_d^2} + O\left(\frac{\log\log N}{B_d\sqrt{\log N}}\right)\right\} = 0. \tag{41}$$

We now want to check the validity of the two asymptotic results above. To do so, we need to be able to approximate the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$, which can be done using the unbiased Monte Carlo (MC) estimator given for all $N \in \mathbb{N}^\star$ by

$$\frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} \overline{w}_{\theta,\phi}(Z_j)^{1-\alpha}\right),$$

with $Z_1, \dots, Z_N$ being i.i.d. samples generated according to q_ϕ. As for the approximation returned by Theorem 3, we will represent it according to (40) through functions of the form

$$c_1 \mapsto -\frac{\alpha B_d^2}{2} - \frac{\exp\left[(1-\alpha)^2 B_d^2\right] - 1}{2(1-\alpha)N} + \frac{c_1}{N} \tag{42}$$

and for the approximation returned by Theorem 5, we will represent it according to (41) through functions of the form

$$c_2 \mapsto -\frac{B_d^2}{2} + B_d\sqrt{2\log N} + \frac{\log N}{\alpha - 1} + \frac{c_2 B_d \log\log N}{\sqrt{\log N}}. \tag{43}$$

We first consider the case where $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$. In that setting, $B_d = \sqrt{d}$ and we have that (i) the $1/N$ term from (40) is exponential in $(1-\alpha)^2 d$ and (ii) we are in the conditions of application of Theorem 5 by setting $\sigma_- = 1$.

Consequently, for this choice of (θ, ϕ) and *regardless of the value of* $\alpha \in [0, 1)$, we are expecting Theorem 5 to capture the behavior of the variational gap as d and N increase in such a way that $\log N/d$ decreases. This is indeed what we observe in Figure 1, in which we let $d \in \{10, 100, 1000\}$, $\alpha \in \{0., 0.2, 0.5\}$, $N \in \{2^j \ : \ j = 1 \dots 9\}$ and we compare the behavior of the variational gap to the behavior predicted by Theorem 5 through curves of the form (43).

Unsurprisingly, although valid in low dimensions for a proper choice of α, the analysis of Theorem 3 requires an unpractical amount of samples N to properly capture the behavior of the variational gap as d increases (additional plots providing the comparison with Theorem 3 are made available in Appendix D.1 for the sake of completeness).

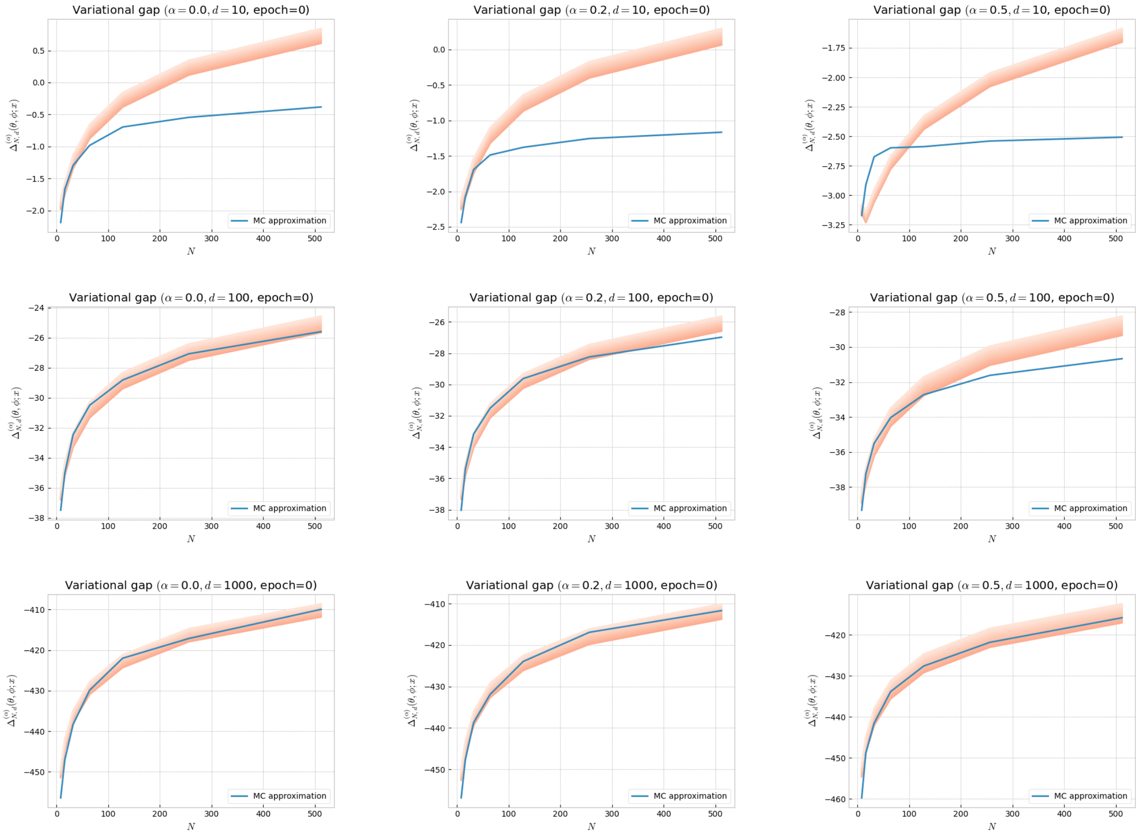

Figure 1: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the toy example described in Section 6.1 as a function of N, for varying values of (α, d) and with $(\theta, \phi) = (0 \cdot \boldsymbol{u_d}, \boldsymbol{u}_d)$ so that $B_d = \sqrt{d}$. Plotted in orange are curves of the form (43) with tailored values of c_2.

We next train the parameter ϕ in order to measure the impact of the training procedure on the validity of our asymptotic results. Here, this impact is reflected in the quantity B_d

through the simple relation $B_d = \|\theta - \phi\|$. In case the training is successful, $B_d/\sqrt{d}$ is then anticipated to decrease from 1 to 0 (having set $\theta = 0 \cdot \boldsymbol{u_d}$ and initialized with $\phi = \boldsymbol{u_d}$).

Hence, as the training progresses, we will be less and less able to find $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$, which will contradict the assumption we make in Theorem 5. At the same time, the $1/N$ term from (40) will decrease thanks to its dependency in B_d, meaning that (40) may become a better approximation than (41) during the training procedure.

This behavior is empirically confirmed in Figure 2 (and we also check in Figure 14 of Appendix D.1 that $B_d/\sqrt{d}$ indeed goes from 1 to 0 during the training procedure). In those plots, the parameter ϕ was optimised via stochastic gradient descent using the reparameterized gradient estimator (13) with $N = 100$ and we set $\alpha = 0.2$ and $d = 1000$ (and a similar trend can be observed for other values of α and d).

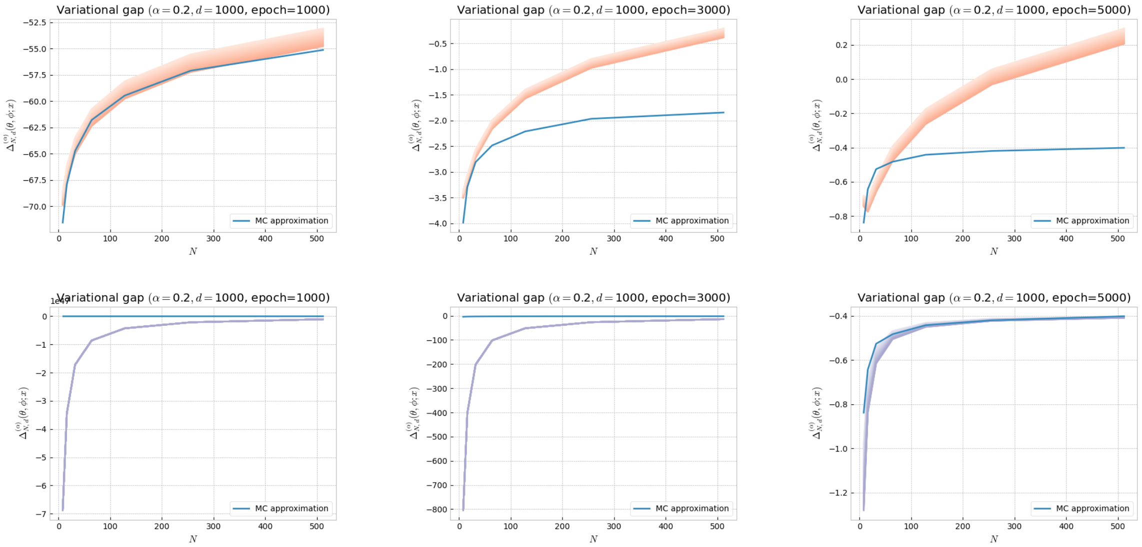

Figure 2: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) at epochs $\{1000, 3000, 5000\}$ for the toy example described in Section 6.1 as a function of N, for $\alpha = 0.2$ and $d = 1000$. Plotted in orange (resp. in purple) are curves of the form (43) with tailored values of c_2 (resp. of the form (42) with tailored values of c_1).

The main insight we get from our first numerical experiment is then that: as the dimension d increases and N does not grow faster than exponentially with d, we should not expect much empirically from the VR-IWAE bound as a lower bound to the marginal log-likelihood when the distribution of the weights is log-normal. This is true unless the encoder and decoder distributions become very close to one another, in which case Theorem 3 does apply instead of Theorem 5.

Thus, while this limitation of the VR-IWAE bound holds for all $\alpha \in [0, 1)$, it may be mitigated by (i) proposing successful training procedures (further shedding light on the importance of finding gradient estimators with good SNR properties) and (ii) selecting suitable variational families which can capture the complexity within the target posterior density. Furthermore, the analysis provided by Theorem 3 may also apply in lower dimensional settings, under the condition that the variance term appearing in Theorem 3 is well-behaved and that the value of α is properly tuned. We next present a second numerical experiment, where this time the weights are not exactly log-normal.

6.2 Linear Gaussian Example

We are interested in the linear Gaussian example from Rainforth et al. (2018), which we already highlighted in Examples 2 and 4. The data set $\mathcal{D} = \{x_1, \dots, x_T\}$ is generated by sampling $T = 1024$ datapoints from $\mathcal{N}(0, 2\boldsymbol{I}_d)$ and we will consider three initializations for the parameters (θ, ϕ) involving a Gaussian perturbation of standard deviation σ_{perturb} of the ground truth values $(\theta^\star, \phi^\star)$:

(i) $\sigma_{\text{perturb}} = 0.5$: the parameters are initialized far from $(\theta^\star, \phi^\star)$,
(ii) $\sigma_{\text{perturb}} = 0.01$: the parameters are initialized close to $(\theta^\star, \phi^\star)$,
(iii) $\sigma_{\text{perturb}} = 0.$: the parameters are equal to $(\theta^\star, \phi^\star)$.

The first two initializations follow from Rainforth et al. (2018) and should notably permit us to approximately characterize the behavior of the linear model before and after training.

Our first step is to check that, as written in Example 4, the distribution of the weights is approximately log-normal as d increases for the initializations above. To do so, we randomly select a datapoint x, draw $N = 1000000$ weight samples in dimension $d = \{20, 100, 1000\}$ for $\sigma_{\text{perturb}} \in \{0.5, 0.01, 0.\}$, before plotting for each d a histogram of the resulting log-weight distribution as well as a Q-Q plot to test the normality assumption of those log-weights.

The results are shown on Figure 3 and we see that while the log-normality phenomenon happens in dimension $d = 100$ when a large perturbation is being considered, even a small perturbation to no perturbation at all can induce some log-normality of the weights as d further increases, which is in line with the theory (and similar plots can be observed for other randomly selected datapoints).

We next want to test the validity of our asymptotic results. On the one hand, Theorem 3 predicts that: for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d(\theta; x) - \frac{\gamma_{\alpha,d}^2}{2N} + o\left(\frac{1}{N}\right), \tag{44}$$

where $\mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d$ and $\gamma_{\alpha,d}^2$ can be analytically computed using Example 2 (and we have emphasized the dependency in d in each of those terms). On the other hand, Theorem 6 predicts under (A2) that: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N / d^{1/3} \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d \log N}}\right)\right) = 0,$$

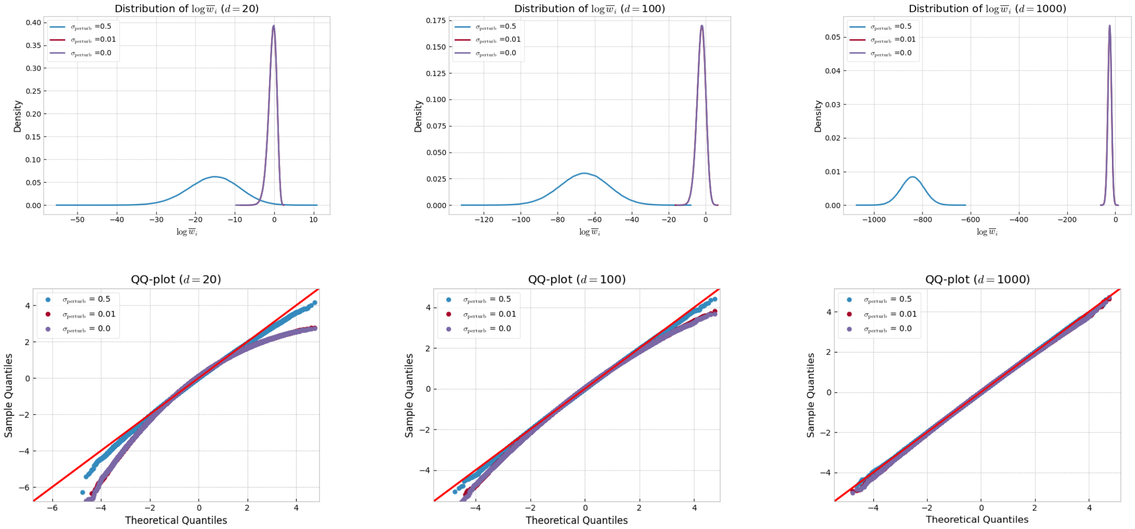

Figure 3: Plotted is the distribution of $\log \overline{w}_i$ and the corresponding QQ-plot for the linear Gaussian example described in Section 6.2 for a randomly selected datapoint x, varying values of d and three different initializations of the parameters (θ, ϕ).

where σ^2 and a can be computed analytically according to Example 4. Hence, to check whether these results apply, we want to look at functions of the form

$$\text{(Theorem 3)} \quad c_1 \mapsto \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d(\theta; x) - \frac{\gamma_{\alpha,d}^2}{2N} + \frac{c_1}{N} \tag{45}$$

$$\text{(Theorem 6)} \quad c_2 \mapsto -da + \sqrt{d}\sigma\sqrt{2\log N} + \frac{c_2\sqrt{d}\log\log N}{\sqrt{\log N}} \tag{46}$$

and see how well they approximate the behavior of the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$. Based on Example 4, we are expecting the regime predicted by Theorem 6 to apply as d increases if N does not grow faster than $d^{1/3}$ and this is indeed what we observe in Figure 4.

While this process is noticeably quicker for an initialization that is far from the optimum ($\sigma_{\text{perturb}} = 0.5$), all three initializations considered here eventually exhibit the behavior predicted by Theorem 6 as d further increases. As already mentioned in Section 4.2.2, this sends the important message that the VR-IWAE bound can strongly deteriorate as d increases due to a mismatch between the targeted density and its variational approximation, even though the parameters themselves are optimal.

As for Theorem 3, we obtain that this theorem applies in low to medium dimensions when the value of α is well-chosen and/or the parameters are close to being optimal, but fails as d increases unless we use an unpractical amount of samples N (see Appendix D.2.1).

We now want to get insights regarding the training of the VR-IWAE bound in practice. We follow the methodology used in Rainforth et al. (2018), which looked into the convergence

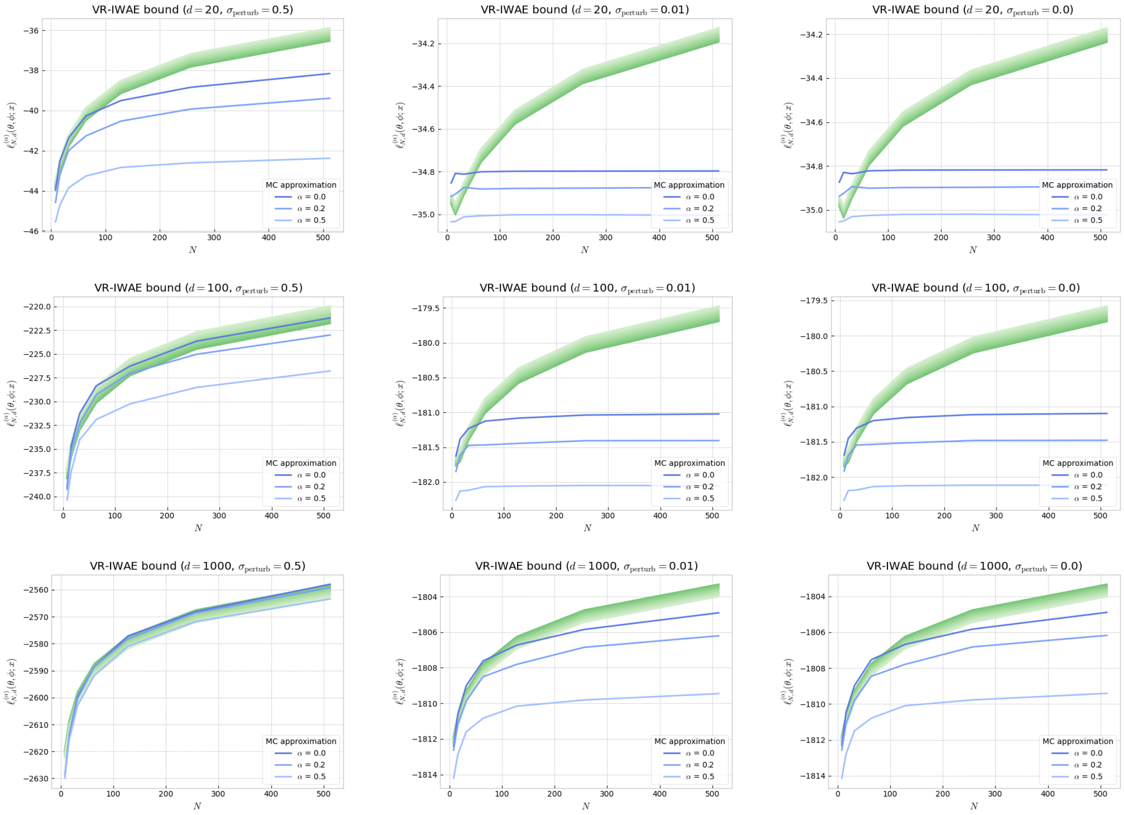

Figure 4: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d) and three different initializations of (θ, ϕ). Plotted in green are curves of the form (46) with tailored values of c_2.

of the SNR for the numerical example considered here in the specific case of the IWAE bound ($\alpha = 0$). Our goal is thus to check whether we can observe the SNR advantages when $\alpha > 0$ predicted by Theorem 1 in the reparameterized case.

Let us decompose θ as $(\theta_\ell)_{1 \leq \ell \leq d}$ and ϕ as $(\phi_{\ell'})_{1 \leq \ell' \leq d+1}$. We then look at the reparameterized estimated gradients of the VR-IWAE bound $(\delta^{(\alpha)}_{1,N}(\theta_\ell))_{1 \leq \ell \leq d}$ and $(\delta^{(\alpha)}_{1,N}(\phi_{\ell'}))_{1 \leq \ell' \leq d+1}$ defined in (14) and (15) respectively as a function of N, for varying values of α, varying values of d and for the two initializations $\sigma_{\text{perturb}} = 0.01$ and $\sigma_{\text{perturb}} = 0.5$. The results are shown in Figure 5 (resp. Figure 6) and they have been obtained by randomly selecting 10 indexes ℓ ranging between 1 and d and averaging over the resulting $\text{SNR}(\delta^{(\alpha)}_{1,N}(\theta_\ell))$ values (resp. by randomly selecting 10 indexes ℓ' ranging between 1 and $d+1$ and averaging over the resulting $\text{SNR}(\delta^{(\alpha)}_{1,N}(\phi_{\ell'}))$ values). Theoretical lines have also been added to Figures 5 and 6 in order to reflect the asymptotic regimes predicted by Theorem 1.

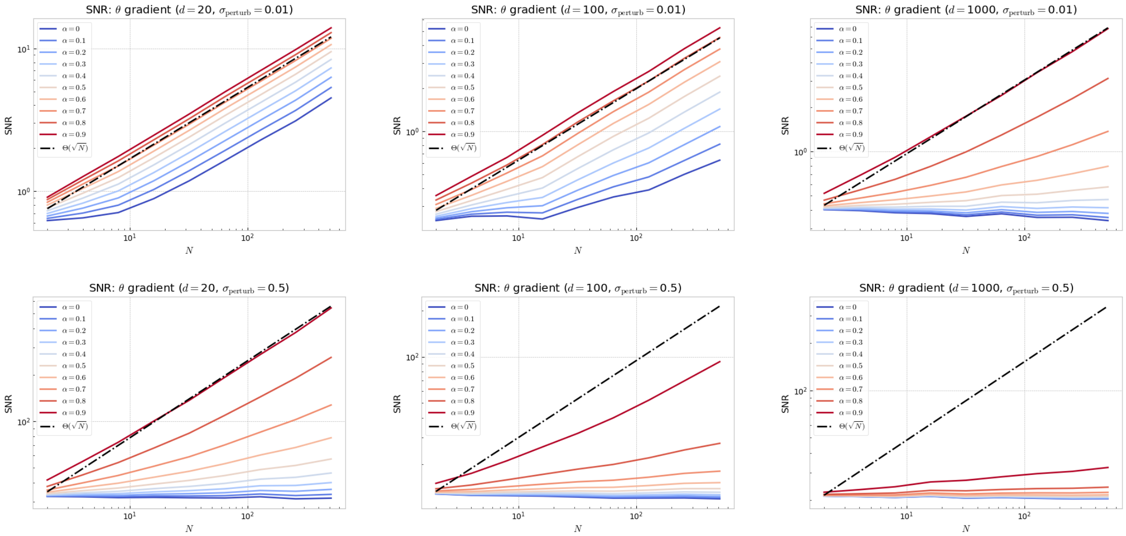

Figure 5: Plotted is the SNR of the generative network (θ) gradients in the reparameterized case (computed over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

Observe that, in the favourable setting of low to medium dimensions with a small perturbation near the optimum (that is $d \in \{20, 100\}$ with $\sigma_{\text{perturb}} = 0.01$), the asymptotic rates predicted by Theorem 1 for the SNR match the observed rates. In particular, the SNR of the inference network gradients vanishes for $\alpha = 0$ while it does not for $\alpha > 0$, which showcases the potential benefits of using the VR-IWAE bound with $\alpha > 0$ instead of the IWAE bound. More generally, increasing α increases the SNR of both the generative and the inference networks, with what seems to be a monotonic increase with α.

However, the improvement in SNR for both the generative and inference networks becomes less pronounced as we get further away from the optimum ($\sigma_{\text{perturb}} = 0.5$) and/or increase d ($d = 1000$). We relate this behavior to the weight collapse effect established in Theorem 6 and anticipate that observing the asymptotic rates predicted by Theorem 1 requires an unpractical amount of samples N as d increases, regardless of the value of $\alpha \in [0, 1)$. Note that the use of doubly-reparameterized gradient estimators for ϕ mitigates the decay in SNR (see Figure 16 of Appendix D.2.2).

Lastly, the behavior of the VR-IWAE bound as well as the SNR behavior of its gradient estimators are not the only way to measure the success of gradient-based methods involving the VR-IWAE bound. For example, we observe that while increasing α does not lower the Mean Squared Error (MSE) for log-likelihood estimation, it can be useful in lowering the MSE of the θ gradient estimates (see Figures 17 and 18 of Appendix D.2.2). We now move on to our third and final numerical experiment, in which we examine a real-data scenario.

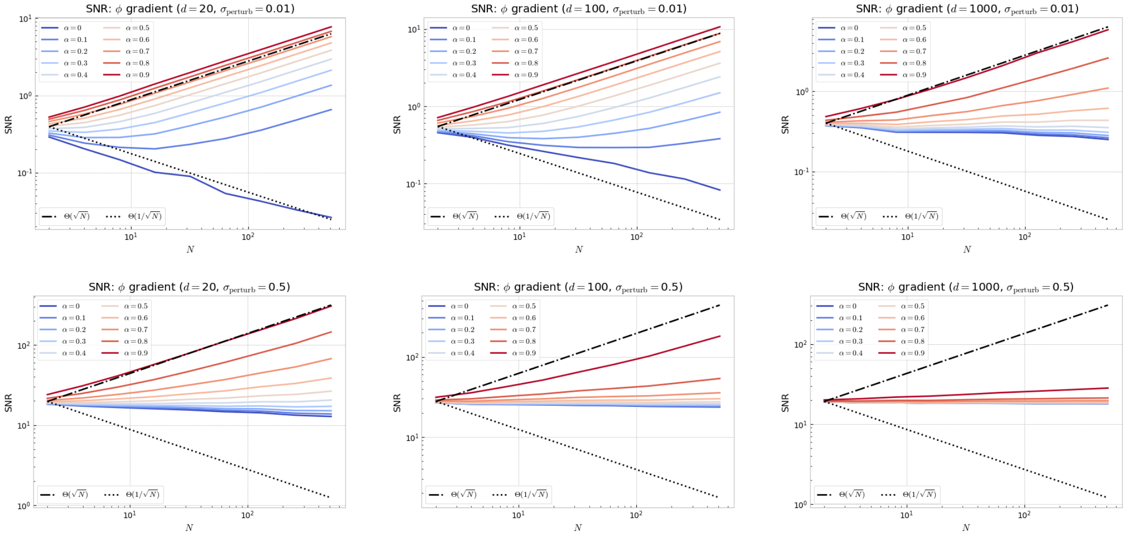

Figure 6: Plotted is the SNR of the inference network (ϕ) gradients in the reparameterized case (computed over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

6.3 Variational Auto-encoder

We consider the case of a variational auto-encoder (VAE) model designed to generate MNIST digits with a d-dimensional latent space, where $p_\theta(z)$ is a fixed standard Gaussian distribution, $p_\theta(x|z)$ is a product over the output dimensions of independent Bernoulli random variables with logits $\pi_\theta(z)$, $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma_\phi(x))$ and the functions $\pi_\theta(z)$ and $(\mu_\phi(x), \sigma_\phi(x))$ are parameterized by neural networks. More precisely, both the encoding and decoding networks are MLPs with two hidden layers of size 200 and tanh nonlinearities.

We first want to investigate whether the distribution of the weights appears to become log-normal in this setting as the dimension of the latent space d increases.

To verify this claim empirically, we randomly select a datapoint x in the test set and for $d \in \{5, 10, 50, 100, 1000, 5000\}$ we randomly generate some model parameters (θ, ϕ), before drawing $N = 1000000$ (unnormalized) weight samples. For each d, we then normalize the weights and plot a histogram of the resulting log-weight distribution, alongside with a QQ-plot to test the normality assumption of those log-weights. The results are shown in Figure 7 and they illustrate the fact that the weights tend to become log-normal as d increases (and similar plots can be obtained for other randomly selected datapoints and other initializations of the parameters (θ, ϕ)).

From there, we want to check the validity of our asymptotic results. To do so, a first comment is that, regardless of the distribution of the weights, Theorem 3 predicts the

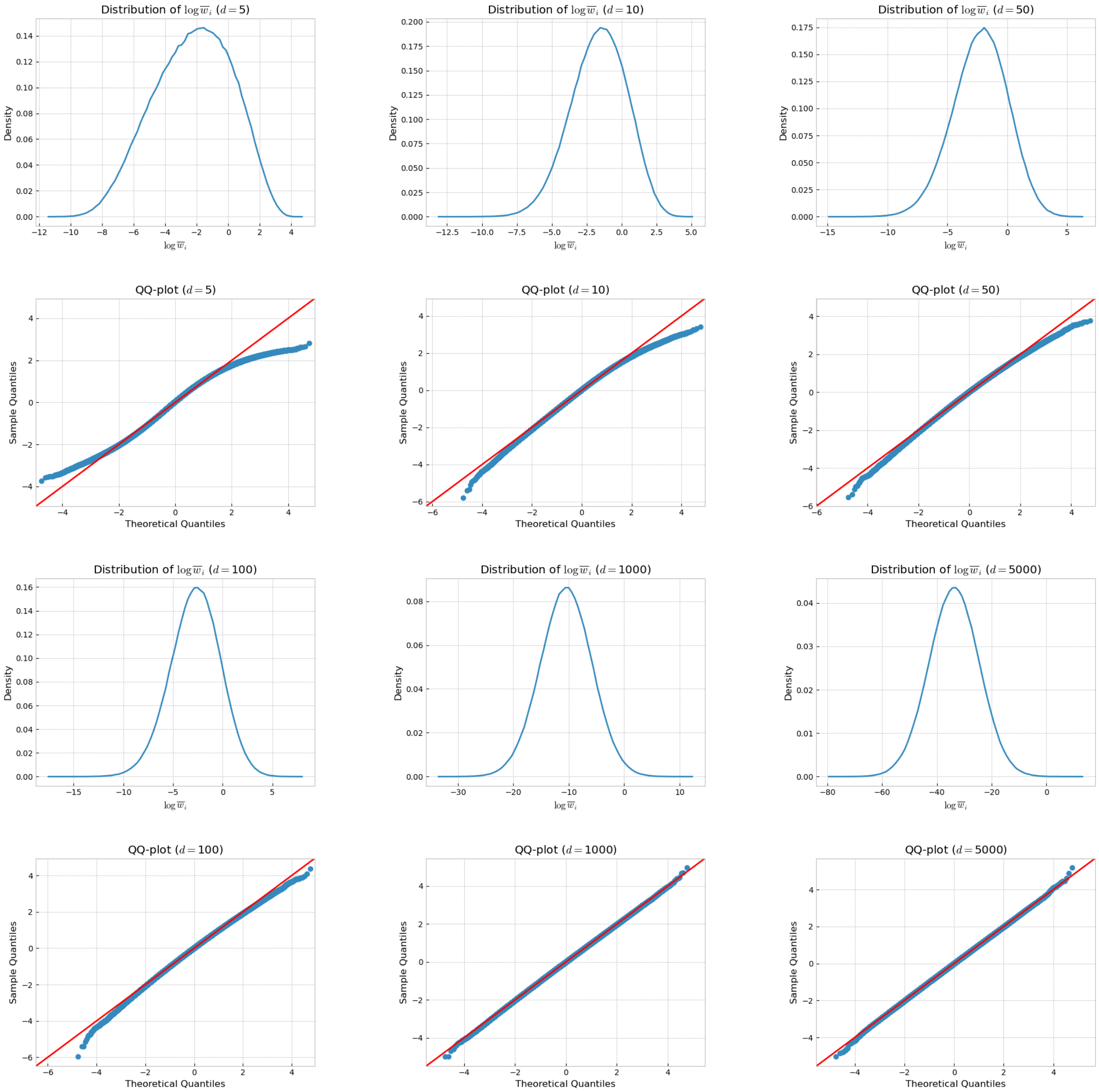

Figure 7: Plotted is the distribution of $\log \overline{w}_i$ and the corresponding QQ-plot for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of d.

following: for all $\alpha \in [0, 1)$,

$$\ell_{N,d}^{(\alpha)}(\theta, \phi; x) = \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \frac{\gamma_{\alpha,d}^2}{2N} + o\left(\frac{1}{N}\right), \tag{47}$$

where, $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ denotes the VR-IWAE bound, $\mathcal{L}_d^{(\alpha)}(\theta, \phi; x)$ the VR-bound and $\gamma_{\alpha,d}^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$ (and we have emphasized the dependency in d in each of those terms). If we further make the assumption that the weights are of the form (37) (which appears to approximately be the case as the dimension d increases as per Figure 7), then Theorem 6 predicts under (A2) that: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \ell^{(\alpha)}_{N,d}(\theta,\phi;x) - \mathrm{ELBO}_d(\theta,\phi;x) - \sqrt{d}\sigma\sqrt{2\log N} + O\left(\frac{\sqrt{d}\log\log N}{\sqrt{\log N}}\right) = 0. \quad (48)$$

Here we have emphasized the dependency in d in $\mathrm{ELBO}_d(\theta,\phi;x)$ and we have also used the fact that $\mathrm{ELBO}_d(\theta,\phi;x) - \ell_d(\theta;x) = -da$ (as previously stated, this follows from taking the expectation in (95) from the proof of Proposition 6 in Appendix B.8.4). Hence, to check whether these results apply, we want to look at functions of the form

$$\text{(Theorem 3)} \quad c_1 \mapsto \mathcal{L}^{(\alpha)}_d(\theta,\phi;x) - \frac{\gamma^2_{\alpha,d}}{2N} + \frac{c_1}{N} \quad (49)$$

$$\text{(Theorem 6)} \quad c_2 \mapsto \mathrm{ELBO}_d(\theta,\phi;x) + \sqrt{d}\sigma\sqrt{2\log N} + \frac{c_2\sqrt{d}\log\log N}{\sqrt{\log N}} \quad (50)$$

and see how well they approximate the behavior of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta,\phi;x)$. Although the functions above contain unknown terms, those terms can all be estimated: the VR bound can be estimated using MC sampling as in (8) and so can the ELBO. As for σ, it can be estimated from the sample standard deviation of the log-weights (and $\gamma^2_{\alpha,d}$ can be estimated in a similar fashion).

Note as a side remark that we are considering the VR-IWAE bound as the quantity of interest whose behavior shall be mimicked by (47) or (48) (through (49) or (50)). Indeed, while we were working with the variational gap in our previous numerical experiments, computing this quantity requires us to estimate both the VR-IWAE bound and the log-likelihood here, which would have incurred an additional source of randomness that we have been able to avoid in both (47) and (48).

Based on Figure 7, we expect two situations to arise at this stage: (i) the asymptotic regime suggested by Theorem 3 captures the behavior of the VR-IWAE bound in low to medium dimensions and (ii) the asymptotic regime predicted by Theorem 6 is accurate as d increases and N does not grow faster than $d^{1/3}$.

This is exactly what we observe in Figures 8 and 9, in which σ is estimated with the 1000000 (unnormalized) weight samples used to build Figure 7 and so are the VR bound and the ELBO (additional plots are also available in Figures 19 and 20 of Appendix D.3). In particular, we see in Figure 8 that the asymptotic regime of Theorem 3 mimics the behavior of the VR-IWAE bound in low to medium dimensions as long as $\gamma^2_{\alpha,d}$ does not grow too quickly with d (and we already observe a mismatch between the two for $\alpha = 0$ and $d = 100$).

Nevertheless, the VR-IWAE bound ends up straying away from the behavior predicted by Theorem 3 as d increases unless N becomes impractically large (with the particularity that this process happens slower as α increases). We then see on Figure 9 that, as d increases to reach high-dimensional settings so that the ratio $\log N/d^{1/3}$ becomes small for the values of N considered here, the behavior predicted by Theorem 6 starts to emerge.

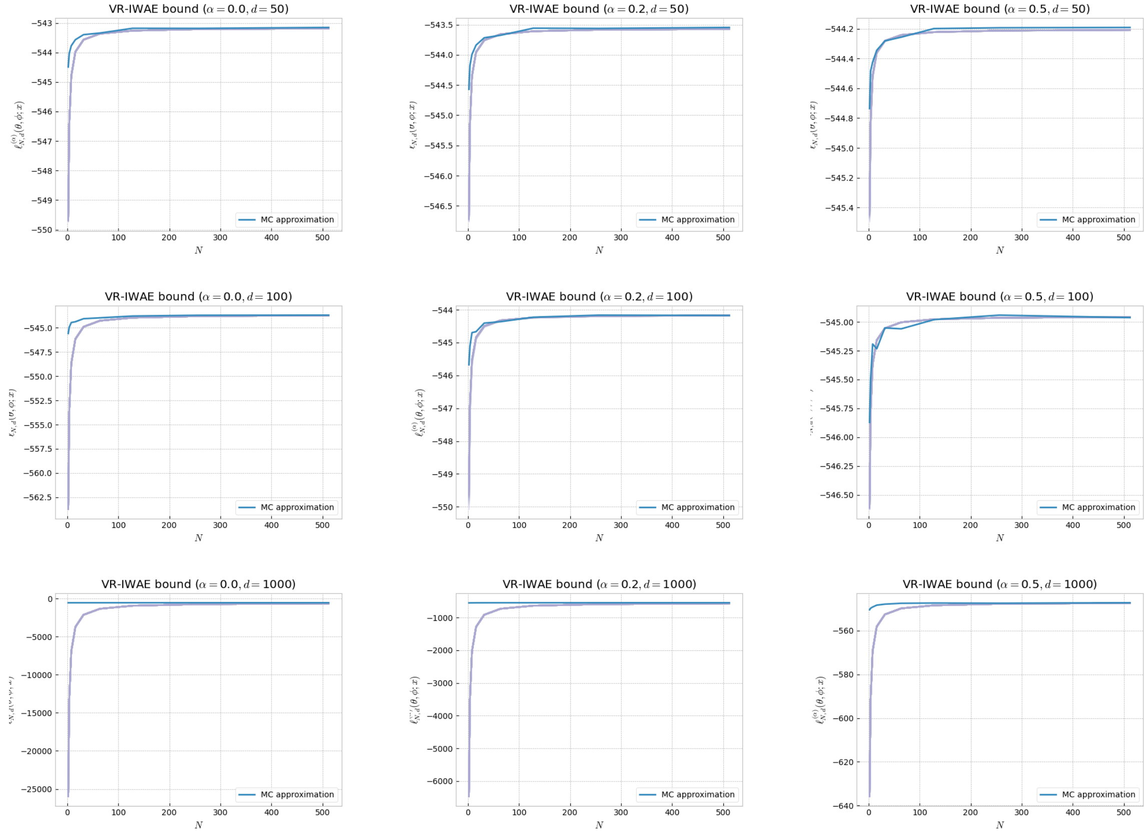

Figure 8: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in purple are curves of the form (49) with tailored values of c_1.

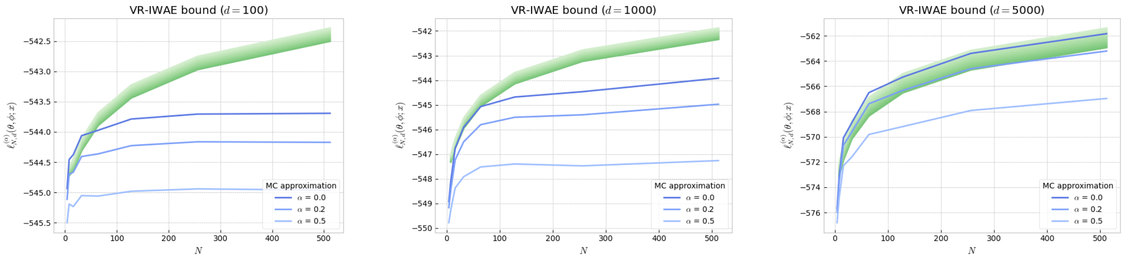

Figure 9: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in green are curves of the form (50) with tailored values of c_2.

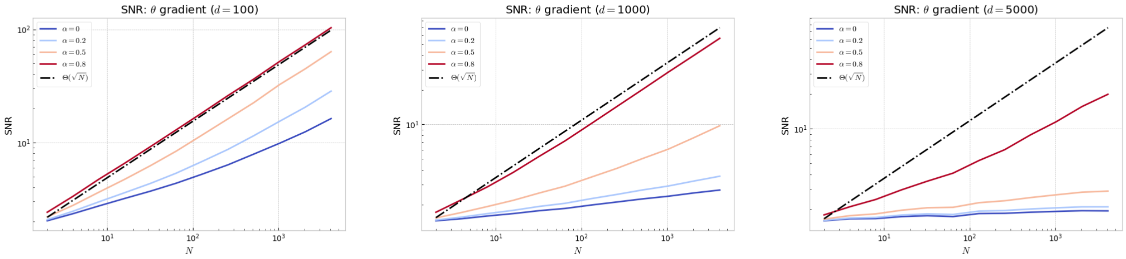

Figure 10: Plotted is the SNR of the generative network (θ) gradients in the reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

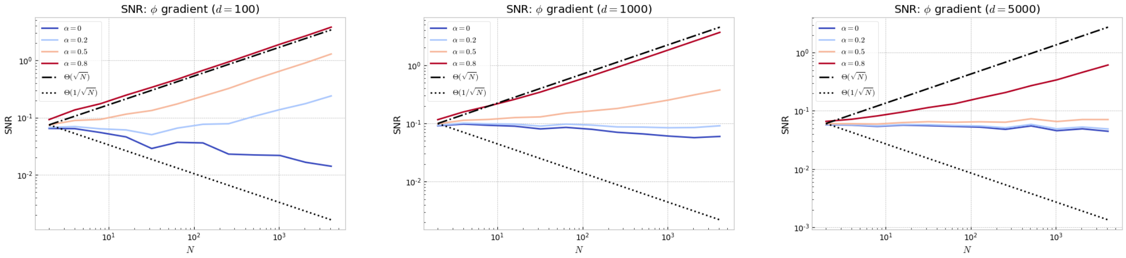

Figure 11: Plotted is the SNR of the inference network (ϕ) gradients in the reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

We now look into the training of the VR-IWAE bound and more specifically into the SNR in medium to high dimensions at initialization, since this scenario corresponds to situations where the VR-IWAE bound seems to resemble more and more the behavior predicted by Theorem 6 (as observed in Figure 9). Following the methodology from the previous subsection, the results are presented in Figures 10 and 11, in which we have plotted the SNR for the generative network and for the inference network respectively in the reparameterized case alongside theoretical lines that reflect the asymptotic regimes predicted by Theorem 1.

As already observed in Section 6.2, the SNR benefits from setting $\alpha > 0$ and the asymptotic rates predicted by Theorem 1 do not capture the SNR behavior as d increases (unless N is unpractically large and/or we appeal to higher values of α). Furthermore, and as we can see in Figure 12, resorting to doubly-reparameterized estimators improves the SNR.

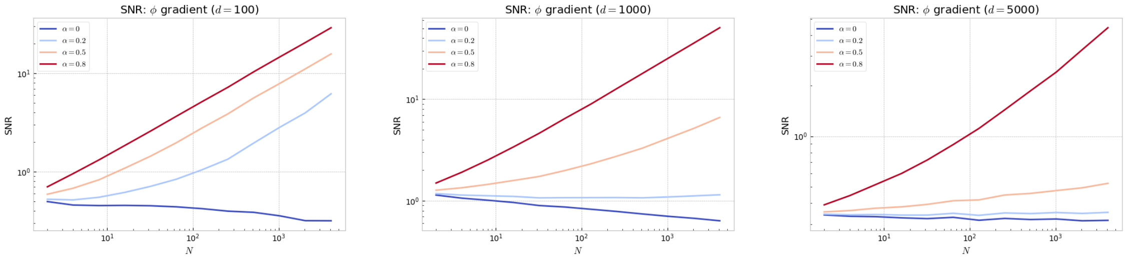

Figure 12: Plotted is the SNR of the inference network (ϕ) gradients in the doubly-reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

We thus confirmed that approximately log-normal weights can arise in real data scenarios as d increases and that our theoretical study provides a useful framework to capture the impact of the weights on the VR-IWAE bound as a function of N, d and α. In line with our empirical findings for the SNR, we also obtained that the asymptotic rates predicted by Theorem 1 match the observed rates in low to medium dimensions and we postulated that the weight collapse occuring in the VR-IWAE bound as d increases may deteriorate the SNR too.

One aspect that remains unexplored empirically is the role of M in the VR-IWAE bound methodology, and in particular the interplay between M and N in the learning outcome. Indeed, the total number of samples needed per iteration in the gradient descent procedure is $N \times M$, with M being responsible for the usual $1/M$ variance reduction in gradient estimators such as (14). Intuitively, and following a similar line of reasoning as for α, we expect to see a bias-variance tradeoff between increasing M or N while keeping $M \times N$ fixed (we refer to Appendix D.3 for details).

Furthermore, if the weight collapse appearing in the VR-IWAE bound as d increases ends up badly impacting the associated gradient descent, our results indicate that practitioners should either (i) set $N = 1$ and allocate the maximum computational budget to M, which in fact corresponds to setting $\alpha = 1$ in the VR-IWAE bound, (ii) find more suitable variational approximations that can capture the complexity within the posterior density or (iii) resort to/construct better gradient estimates (e.g. doubly-reparameterized gradient estimators).

7. Conclusion

In this paper, we formalized the VR-IWAE bound, a variational bound depending on an hyperparameter $\alpha \in [0, 1)$ which generalizes the standard IWAE bound ($\alpha = 0$). We showed that the VR-IWAE bound provides theoretical guarantees behind various VR bound-based schemes proposed in the alpha-divergence variational inference literature and identified other additional desirable properties of this bound.

We then provided two complementary analyses of the variational gap, that is of the difference between the VR-IWAE bound and the marginal log-likelihood. The first analysis shed light on how α may play a role in reducing the variational gap. We then proposed a second analysis to better capture the behavior of the variational gap in high-dimensional scenarios, establishing that the variational gap suffers in this case from a damaging weight collapse phenomenon for all $\alpha \in [0, 1)$. Lastly, we illustrated our theoretical results over several toy and real-data examples.

Overall, our work provides foundations for improving the IWAE and VR methodologies and we now state potential directions of research to extend it. Firstly, one may investigate whether the weight collapse behavior applies beyond the cases we highlighted. Looking into how this weight collapse affects the gradient descent procedures associated to the VR-IWAE bound could be a second direction of research. Thirdly, and in order to improve on the VR-IWAE bound methodology beyond the weight collapse phenomenon, one may seek to further build on the fact that the VR-IWAE bound is the theoretically-sound extension of the IWAE bound that originates from the alpha-divergence variational inference methodology.

Acknowledgments

We would like to thank the action editor and the reviewers for helpful comments and suggestions on the paper. Kamélia Daudel and Arnaud Doucet acknowledge support of the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. Joe Benton was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). Arnaud Doucet also acknowledges support from the EPSRC grant EP/R034710/1.

Appendix A. Deferred Proofs of Section 3

A.1 Extension of $\ell_N^{(\alpha)}$ to the Case $\alpha = 1$

We prove that under common differentiability assumptions, the following limit holds:

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x).$$

Proof Setting $f(\alpha) = \int \int \prod_{i=1}^N q_\phi(z_i) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j)^{1-\alpha} \right) \mathrm{d}z_{1:N}$, the bound $\ell_N^{(\alpha)}(\theta, \phi; x)$ can be rewritten as

$$\ell_N^{(\alpha)}(\theta, \phi; x) = -\frac{f(\alpha) - f(1)}{\alpha - 1}$$

and hence, $\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = -f'(1)$. We then get the desired result by observing that

$$f'(\alpha) = \int \int \prod_{i=1}^N q_\phi(z_i) \left(\frac{\frac{1}{N} \sum_{j=1}^N -\log(\bar{w}_{\phi,\theta}(z_j)) \bar{w}_{\phi,\theta}(z_j)^{1-\alpha}}{\frac{1}{N} \sum_{j=1}^N \bar{w}_{\phi,\theta}(z_j)^{1-\alpha}} \right) \mathrm{d}z_{1:N}$$

and letting $\alpha \to 1$ in the quantity above. ■

A.2 Proof of Proposition 1

Proof of Proposition 1 The results for the case $\alpha = 0$ follow from Burda et al. (2016) and we focus on the case $\alpha \in (0,1)$ in the proof below.

1. One the one hand, (Li and Turner, 2016, Theorem 1) implies that: for all $\alpha \in (0,1)$,

$$\mathcal{L}^{(\alpha)}(\theta,\phi;x) \leq \ell(\theta;x). \tag{51}$$

On the other hand, we obtain from (Li and Turner, 2016, Theorem 2) that:

- For all $N \in \mathbb{N}^\star$ and all $\alpha < 1$

$$\ell_N^{(\alpha)}(\theta,\phi;x) \leq \ell_N^{(N+1)}(\theta,\phi;x) \leq \mathcal{L}^{(\alpha)}(\theta,\phi;x)$$

which gives (10) when paired with (51).

- If the function $z \mapsto w_{\theta,\phi}(z)$ is bounded, then $\ell_N^{(\alpha)}(\theta,\phi;x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta,\phi;x)$ as N goes to infinity.

2. Let $1 > \alpha_1 > \alpha_2 > 0$. Then, the functions $u \mapsto u^{\frac{1-\alpha_1}{1-\alpha_2}}$ and $u \mapsto u^{1-\alpha_2}$ are concave for all $u > 0$ and hence Jensen's inequality implies

$$\begin{aligned}\ell_N^{(\alpha_1)}(\theta,\phi;x) &= \frac{1}{1-\alpha_1}\int\int\prod_{i=1}^N q_\phi(z_i|x)\log\left(\frac{1}{N}\sum_{j=1}^N\left[w_{\theta,\phi}(z_i)^{1-\alpha_2}\right]^{\frac{1-\alpha_1}{1-\alpha_2}}\right)\mathrm{d}z_{1:N}\\ &\leq \ell_N^{(\alpha_2)}(\theta,\phi;x)\\ &\leq \ell_N^{(0)}(\theta,\phi;x)\end{aligned}$$

The desired result (11) follows by using that $\ell_N^{(0)}(\theta,\phi;x) = \ell_N^{(\mathrm{IWAE})}(\theta,\phi;x)$. As for the case of equality, it is obtained as the case of equality of Jensen's inequality.

3. Under the reparameterization trick,

$$\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\int\int\prod_{i=1}^N q(\varepsilon_i)\log\left(\frac{1}{N}\sum_{j=1}^N w_{\theta,\phi}(f(\varepsilon_j,\phi))^{1-\alpha}\right)\mathrm{d}\varepsilon_{1:N}$$

leading, under common differentiability assumptions, to

$$\frac{\partial}{\partial\phi}\ell_N^{(\alpha)}(\theta,\phi;x) = \int\int\prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N\frac{w_{\theta,\phi}(f(\varepsilon_j,\phi))^{-\alpha}\frac{\partial}{\partial\phi}w_{\theta,\phi}(f(\varepsilon_j,\phi))}{\sum_{k=1}^N w_{\theta,\phi}(f(\varepsilon_k,\phi))^{1-\alpha}}\right)\mathrm{d}\varepsilon_{1:N}.$$

The desired result (12) is then obtained using the REINFORCE trick

$$\frac{\partial}{\partial\phi}w_{\theta,\phi}(f(\varepsilon_j,\phi)) = w_{\theta,\phi}(f(\varepsilon_j,\phi))\frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi))$$

and the unbiased estimator (13) follows immediately.

■

A.3 Proof of Theorem 1

The proof of Theorem 1 is based on the proof of the corresponding result in (Rainforth et al., 2018, Theorem 1) (arxiv version of 5 Mar 2019). First, we prove the following useful lemma, which is an extension of (Rainforth et al., 2018, Lemma 1).

Lemma 3 *Suppose we have random variables $X_{i,j}$ for all $i = 1 \dots r$ and $j = 1 \dots N$ satisfying*

(i) $\mathbb{E}\left(X_{i,j}\right) = 0$ *for all* $i = 1 \dots r$ *and* $j = 1 \dots N$*;*

(ii) $\mathbb{E}\left(|X_{i,j}|^r\right) < \infty$ *for all* $i = 1 \dots r$ *and* $j = 1 \dots N$*;*

(iii) for each $i = 1 \dots r$*, the random variables* $X_{i,1}, \dots, X_{i,N}$ *are i.i.d.;*

(iv) for each $i = 1 \dots r$ *and* $j = 1 \dots N$*, the random variables* $X_{i,j}$ *and* $\{X_{i',j'}\}_{i'=1\dots r;\, j' \neq j}$ *are independent.*

Then

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right) \dots \left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right)\right] = \begin{cases} O(N^{-r/2}) & \textit{if } r \textit{ is even,} \\ O(N^{-(r+1)/2}) & \textit{if } r \textit{ is odd.} \end{cases}$$

Proof We have

$$\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right) \dots \left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right) = \frac{1}{N^r} \sum_{\{A_1,\dots,A_t\}} \sum_{(j_1,\dots,j_t)} \left(\prod_{i \in A_1} X_{i,j_1}\right) \dots \left(\prod_{i \in A_t} X_{i,j_t}\right) \quad (52)$$

where the first sum is over all partitions $\{A_1, \dots, A_t\}$ of the set $\{1, \dots, r\}$ (so that t is an integer between 1 and r for each partition) and the second sum is over all tuples $(j_1, \dots, j_t)$ with each element being a distinct integer between 1 and N (e.g. for the partition $\{A_1\}$ with $A_1 = \{1, \dots, r\}$ and $t = 1$, the second sum reduces to the sum over $j_1 = 1 \dots N$ and for the partition $\{A_1, \dots, A_r\}$ with $A_p = \{p\}$ for all $p = 1 \dots r$ and $t = r$, the second sum corresponds to the sum over all permutations over the subsets of length r of $(1, \dots, N)$).

Now consider the case where $|A_p| = 1$ for some integer p between 1 and t in a certain partition $\{A_1, \dots, A_t\}$. Without any loss of generality, we let $|A_1| = 1$. Then, by the independence of condition (iv) followed by condition (i), we have

$$\begin{aligned} \mathbb{E}\left[\left(\prod_{i \in A_1} X_{i,j_1}\right) \dots \left(\prod_{i \in A_t} X_{i,j_t}\right)\right] &= \mathbb{E}\left[X_{i^*,j_1}\right] \mathbb{E}\left[\left(\prod_{i \in A_2} X_{i,j_2}\right) \dots \left(\prod_{i \in A_t} X_{i,j_t}\right)\right] \\ &= 0 \end{aligned}$$

where i^* is the single element of A_1. Hence, we can restrict the sum over $\{A_1, \dots, A_t\}$ to only consider partitions where every partition has at least two elements. Furthermore, by the generalized Hölder's inequality (i.e. given the r random variables $X_1, \dots, X_r$, it holds that $\mathbb{E}(|\prod_{p=1}^{r} X_p|) \leq \prod_{p=1}^{r} \mathbb{E}(|X_p|^r)^{1/r}$) and conditions (ii) and (iii), we also have

$$\mathbb{E}\left[\left|\left(\prod_{i \in A_1} X_{i,j_1}\right) \dots \left(\prod_{i \in A_t} X_{i,j_t}\right)\right|\right] \leq \prod_{i=1}^{r} \mathbb{E}\left(|X_{i,1}|^r\right)^{1/r} < \infty,$$

where we have used that the product on the l.h.s. of the equation above contains exactly r terms since $\{A_1, \ldots, A_t\}$ is a partition of $\{1, \ldots, r\}$. Putting this together with (52) yields:

$$\begin{aligned}\left|\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right)\cdots\left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right)\right]\right| &\leq \frac{1}{N^r}\sum_{\substack{\{A_1,\ldots,A_t\}\\ \text{all } |A_p|\geq 2}}\ \sum_{(j_1,\ldots,j_t)}\left(\prod_{i=1}^{r}\mathbb{E}\left(|X_{i,1}|^r\right)^{1/r}\right)\\ &\leq \frac{1}{N^r}\sum_{\substack{\{A_1,\ldots,A_t\}\\ \text{all } |A_p|\geq 2}} N^t\left(\prod_{i=1}^{r}\mathbb{E}\left(|X_{i,1}|^r\right)^{1/r}\right).\end{aligned}$$

Finally, note that (i) any partition $\{A_1, \ldots, A_t\}$ of $\{1, \ldots, r\}$ where each part has size at least 2 can have at most $\lfloor r/2 \rfloor$ parts, so $t \leq \lfloor r/2 \rfloor$ and (ii) we can crudely bound the number of partitions by r^r. Hence

$$\begin{aligned}\left|\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right)\cdots\left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right)\right]\right| &\leq \frac{r^r}{N^{r-\lfloor r/2\rfloor}}\left(\prod_{i=1}^{r}\mathbb{E}\left(|X_{i,1}|^r\right)^{1/r}\right)\\ &= \begin{cases} O(N^{-r/2}) & \text{if } r \text{ is even}\\ O(N^{-(r+1)/2}) & \text{if } r \text{ is odd.}\end{cases}\end{aligned}$$

■

We next prove a second lemma.

Lemma 4 *Let k be a positive integer. Set $R_{\alpha,N} = N^{-1}\sum_{i=1}^{N} w_{\theta,\phi}(Z_i)^{1-\alpha}$, where $Z_1, \ldots, Z_N$ are i.i.d. samples generated according to q_ϕ. Then, the condition*

$$\limsup_{N\to\infty}\mathbb{E}\left((1/R_{\alpha,N})^k\right) < \infty \tag{53}$$

is equivalent to the statement that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/R_{\alpha,N})^k) < \infty$.

Proof of Lemma 4 Fix a positive integer $N \geq 2$. For all $x \in [0, 1)$, we have by convexity of the function $x \mapsto (1-x)^{-k}$ that

$$\left(\frac{1}{1-x}\right)^k \geq \left(\frac{N}{N-1}\right)^k + k\left(\frac{N}{N-1}\right)^{k+1}\left(x - \frac{1}{N}\right).$$

It follows that if $x_1, \ldots, x_N \in (0, 1)$ are such that $\sum_{i=1}^{N} x_i = 1$, then

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{1-x_i}\right)^k \geq \left(\frac{N}{N-1}\right)^k.$$

Given $\alpha \in [0, 1)$ and N positive reals $w_1, \ldots, w_N$, we may set $x_i = w_i^{1-\alpha}/\left(\sum_{i=1}^{n} w_i^{1-\alpha}\right)$ in the above to get

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{N-1}{\sum_{\substack{j=1\\ j\neq i}}^{N} w_j^{1-\alpha}}\right)^k \geq \left(\frac{N}{\sum_{j=1}^{N} w_j^{1-\alpha}}\right)^k.$$

Now consider setting $w_i = w_{\theta,\phi}(Z_i)$ where the Z_i are i.i.d. samples generated according to q_ϕ. We see that the r.h.s. of the above expression is distributed as $(1/R_{\alpha,N})^k$, while each term in the sum on the l.h.s. is distributed as $(1/R_{\alpha,N-1})^k$. We conclude that

$$\mathbb{E}\left((1/R_{\alpha,N-1})^k\right) \geq \mathbb{E}\left((1/R_{\alpha,N})^k\right)$$

for all $N \geq 2$. Hence $\mathbb{E}\left((1/R_{\alpha,N})^k\right)$ is decreasing in N, so $\limsup_{N\to\infty} \mathbb{E}\left((1/R_{\alpha,N})^k\right) < \infty$ if and only if there exists some N such that $\mathbb{E}\left((1/R_{\alpha,N})^k\right) < \infty$. ■

We now move on to the proof of Theorem 1.

Proof of Theorem 1 We use the following shorthand notation

$$\begin{aligned}&\tilde{w}_{m,j} = w_{\theta,\phi}(f(\varepsilon_{m,j},\phi)), \quad m = 1\ldots M,\ j = 1\ldots N\\ &Z_\alpha = \mathbb{E}_{\varepsilon\sim q}(w_{\theta,\phi}(f(\varepsilon,\phi))^{1-\alpha})\end{aligned}$$

and we also recall the notation

$$\hat{Z}_{1,N,\alpha} = \frac{1}{N}\sum_{j=1}^N \tilde{w}_{1,j}^{1-\alpha}.$$

We will first prove that

$$\mathrm{SNR}[\delta_{M,N}^{(\alpha)}(\theta_\ell)] = \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial\theta_\ell} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial\theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right) + O\left(\frac{1}{N}\right)}} \tag{54}$$

$$\mathrm{SNR}[\delta_{M,N}^{(\alpha)}(\phi_{\ell'})] = \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial\phi_{\ell'}} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial\phi_{\ell'}}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\phi_{\ell'}} - \frac{\partial\log Z_\alpha}{\partial\phi_{\ell'}}\right]^2\right) + O\left(\frac{1}{N}\right)}}. \tag{55}$$

As the two expressions above follow the same form, it is in fact enough to only prove (54). We will do so by studying the asymptotic variance and expected value of $\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell) := (1-\alpha)\delta_{M,N}^{(\alpha)}(\theta_\ell)$ separately, before combining them to deduce (54).

- **Study of $\mathbb{V}(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell))$.**

We start from the identity

$$\frac{\partial\log\hat{Z}_{1,N,\alpha}}{\partial\theta_\ell} = \frac{\partial\log Z_\alpha}{\partial\theta_\ell} + \frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right),$$

which can for example be verified by using the following identity (which is a version of the Taylor expansion to first order with an explicit form for the remainder)

$$\log(1+x) = x - \int_0^x \frac{t}{1+t}dt,$$

substituting $x = (\hat{Z}_{1,N,\alpha} - Z_\alpha)/Z_\alpha$, differentiating with respect to θ_ℓ and using the chain rule where necessary. Hence,

$$\begin{aligned}
M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right) &= \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{1,N}(\theta_\ell)\right) = \mathbb{V}\left(\frac{\partial \log(\hat{Z}_{1,N,\alpha})}{\partial \theta_\ell}\right) \\
&= \mathbb{V}\left(\frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&= \mathbb{V}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&= \mathbb{V}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&\quad + 2\,\mathrm{Cov}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right), \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&\quad + \mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right)
\end{aligned}$$

Furthermore, observe that

$$\begin{aligned}
\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) &= \frac{1}{N}\sum_{j=1}^N \frac{\partial}{\partial \theta_\ell}\left(\frac{\tilde{w}_{1,j}^{1-\alpha} - Z_\alpha}{Z_\alpha}\right) \\
&= \frac{1}{N}\sum_{j=1}^N \frac{Z_\alpha \frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial \theta_\ell} - \tilde{w}_{1,j}^{1-\alpha} \frac{\partial Z_\alpha}{\partial \theta_\ell}}{Z_\alpha^2}. \qquad (56)
\end{aligned}$$

As a result,

$$\mathbb{E}\left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right] = \frac{\mathbb{E}[\frac{\partial(\tilde{w}_{1,1}^{1-\alpha})}{\partial \theta_\ell}] - \frac{\partial Z_\alpha}{\partial \theta_\ell}}{Z_\alpha} = 0, \qquad (57)$$

where we have used that under common differentiability assumptions $\mathbb{E}[\partial(\tilde{w}_{1,1}^{1-\alpha})/\partial\theta_\ell] = \partial Z_\alpha/\partial\theta_\ell$, that is we can interchange the order of integration and differentiation. Consequently, we can simplify the expression of $M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right)$ to obtain that

$$\begin{aligned}
M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right) &= \mathbb{E}\left(\left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right]^2\right) \\
&\quad + 2\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right]^2\right) \\
&\quad + \mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \qquad (58)
\end{aligned}$$

We now control the three terms in the r.h.s. of (58) separately.

1. *First term in the r.h.s. of* (58)*.* To control the first term in the r.h.s. of (58), we use (56) to get that

$$\begin{aligned}
\mathbb{E}\left(\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right) &= \frac{1}{N^2}\mathbb{E}\left(\sum_{j=1}^N\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\
&= \frac{1}{N^2}\mathbb{E}\left(\sum_{j=1}^N\left[\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}\right]^2\right)\\
&= \frac{1}{NZ_\alpha^4}\mathbb{E}\left(\left[Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right]^2\right)\\
&= \frac{1}{NZ_\alpha^4}\mathbb{E}\left(\left[\tilde{w}_{1,1}^{-\alpha}\left\{(1-\alpha)Z_\alpha\frac{\partial\tilde{w}_{1,1}}{\partial\theta_\ell}-\tilde{w}_{1,1}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right\}\right]^2\right),
\end{aligned} \tag{59}$$

where the cross-terms disappeared due to the independence of the $(\varepsilon_{1,j})_{1\leq j\leq N}$ paired up with (57).

2. *Second term in the r.h.s of* (58)*.* We deal with the cross term in (58) by splitting it into two parts

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\
&\quad=\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\
&\quad+\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\right).
\end{aligned} \tag{60}$$

Using the expression of $\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)$ given in (56) and the fact that

$$\hat{Z}_{1,N,\alpha}-Z_\alpha=\frac{1}{N}\sum_{j=1}^N\left(\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha\right), \tag{61}$$

we set: for all $j=1\ldots J$,

$$\begin{aligned}
X_{1,j}&=\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha,\\
X_{2,j}=X_{3,j}&=\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}
\end{aligned}$$

and we can then apply Lemma 3 with $r=3$ (by noting in particular that the required moments are finite under our assumptions): we thus obtain that

$$\mathbb{E}\left(\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)=O\left(\frac{1}{N^2}\right)$$

which controls the first term in the r.h.s. of (60). The second term in the r.h.s. of (60) can be bounded as follows

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\right)\\
&\leq \mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)^2\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right)^{1/2}\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2}. \qquad (62)
\end{aligned}$$

By taking this time: for all $j = 1 \dots N$,

$$X_{1,j} = X_{2,j} = \tilde{w}_{1,j}^{1-\alpha} - Z_\alpha,$$

$$X_{3,j} = X_{4,j} = X_{5,j} = X_{6,j} = \frac{Z_\alpha \frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell} - \tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},$$

in Lemma 3 with $r = 6$ (and noting once again that the required moments are finite under our assumptions), we see that the first term in the r.h.s. of (62) is $O(N^{-3/2})$. As for the second term of the r.h.s. of (62), Cauchy-Schwarz implies that

$$\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right) \leq \mathbb{E}\left(\frac{1}{\hat{Z}_{1,N,\alpha}^4}\right)^{1/2}\mathbb{E}\left(\left(Z_\alpha-\hat{Z}_{1,N,\alpha}\right)^4\right)^{1/2}.$$

Now note that under our assumptions, Lemma 4 can be applied with $k = 4$ so that (53) with $k = 4$ holds and controls the first term in the r.h.s. above. Furthermore, applying Lemma 3 with $r = 4$ and for all $j = 1 \dots N$,

$$X_{1,j} = X_{2,j} = X_{3,j} = X_{4,j} = \tilde{w}_{1,j}^{1-\alpha} - Z_\alpha$$

yields

$$\mathbb{E}\left(\left(Z_\alpha-\hat{Z}_{1,N,\alpha}\right)^4\right) = O\left(\frac{1}{N^2}\right).$$

We can then conclude that

$$\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2} = O\left(\frac{1}{N^{1/2}}\right), \qquad (63)$$

and so the r.h.s. of (62) is bounded above by $O(N^{-2})$. It follows that the second term in the r.h.s. of (60) is $O(N^{-2})$ too and we can conclude that

$$\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right) = O\left(\frac{1}{N^2}\right) \qquad (64)$$

that is the second term of the r.h.s. of (58) is $O(N^{-2})$.

3. *Third term of the r.h.s. in* (58). For the third term of the r.h.s. in (58), note that

$$\begin{aligned}
&\mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)\\
&\leq \mathbb{E}\left(\left[\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\
&\leq \mathbb{E}\left(\left[\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right)^{1/2}\mathbb{E}\left(\frac{1}{\hat{Z}_{1,N,\alpha}^4}\right)^{1/2}
\end{aligned}$$

where the final line follows from Cauchy–Schwarz. As a result, using (56) and (61), taking for all $j = 1 \dots N$

$$\begin{aligned}
X_{1,j} = X_{2,j} = X_{3,j} = X_{4,j} &= \tilde{w}_{1,j}^{1-\alpha} - Z_\alpha\\
X_{5,j} = X_{6,j} = X_{7,j} = X_{8,j} &= \frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell} - \tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},
\end{aligned}$$

and since the required moments are finite under our assumptions, Lemma 3 with $r = 8$ implies that

$$\mathbb{E}\left(\left[\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right) = O\left(\frac{1}{N^4}\right).$$

Combined with (53) with $k = 4$ (which holds under our assumptions by Lemma 4 with $k = 4$), this implies that

$$\mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right) = O\left(\frac{1}{N^2}\right). \tag{65}$$

Putting (58), (59), (64) and (65) together, we see that

$$\begin{aligned}
\mathbb{V}\left(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell)\right) &= \frac{1}{MNZ_\alpha^4}\mathbb{E}\left(\left[\tilde{w}_{1,1}^{-\alpha}\left\{(1-\alpha)Z_\alpha\frac{\partial\tilde{w}_{1,1}}{\partial\theta_\ell} - \tilde{w}_{1,1}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right\}\right]^2\right) + O\left(\frac{1}{MN^2}\right)\\
&= \frac{1}{MNZ_\alpha^2}\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right) + O\left(\frac{1}{MN^2}\right)
\end{aligned}$$

and it follows that

$$\begin{aligned}
\sqrt{\mathbb{V}(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell))} &= \frac{1}{\sqrt{MN}Z_\alpha}\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right) + O\left(\frac{1}{N}\right)}\\
&= \frac{1}{\sqrt{MN}Z_\alpha}\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)}\sqrt{1 + O\left(\frac{1}{N}\right)}\\
&= \frac{1}{\sqrt{MN}Z_\alpha}\left(\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)\right).
\end{aligned} \tag{66}$$

- **Study of** $\mathbb{E}(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell))$**.**

We start from the identity

$$\frac{\partial \log \hat{Z}_{1,N,\alpha}}{\partial \theta_\ell} = \frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \frac{1}{2}\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right) \\ + \left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right),$$

which can for example be proved using the following identity (which is a version of the Taylor expansion to second order with an explicit form for the remainder)

$$\log(1+x) = x - \frac{x^2}{2} + \int_0^x \frac{t^2}{1+t}\mathrm{d}t,$$

substituting $x = (\hat{Z}_{1,N,\alpha} - Z_\alpha)/Z_\alpha$, differentiating with respect to θ_ℓ and using the chain rule where necessary. It follows that

$$\begin{aligned}\mathbb{E}\left(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell)\right) &= \mathbb{E}\left(\tilde{\delta}_{1,N}^{(\alpha)}(\theta_\ell)\right) = \mathbb{E}\left(\frac{\partial \log \hat{Z}_{1,N,\alpha}}{\partial \theta_\ell}\right) \\ &= \mathbb{E}\Bigg(\frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \frac{1}{2}\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right) \\ &\qquad + \left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\Bigg) \\ &= \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2}\mathbb{E}\left(\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right)\right) + R_2(\hat{Z}_{1,N,\alpha}) \\ &= \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + R_2(\hat{Z}_{1,N,\alpha}) \end{aligned} \tag{67}$$

where we denote

$$R_2(\hat{Z}_{1,N,\alpha}) = \mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right)$$

and where we have used (56), (57), (61) and the fact that under common differentiability assumptions, we have that

$$\begin{aligned}\mathbb{E}\left(\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]^2\right)\right) &= 2\mathbb{E}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]\right)\right)\\ &= \frac{2}{N^2}\mathbb{E}\left(\sum_{j=1}^N \frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\cdot\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}\right)\\ &= \frac{1}{N}\mathbb{E}\left(\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\right]^2\right)\right)\\ &= \frac{1}{N}\frac{\partial}{\partial\theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right]\end{aligned}$$

(here the cross-terms disappear due to the independence of the $(\varepsilon_{1,j})_{1\leq j\leq N}$ paired up with (57)). Notice then that we can split up $R_2(\hat{Z}_{1,N,\alpha})$ as

$$\begin{aligned}R_2(\hat{Z}_{1,N,\alpha}) = \mathbb{E}&\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)\\ &+\mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)\end{aligned}\tag{68}$$

The first term in (68) can be bounded by applying Lemma 3 with $r=3$ and for all $j=1\dots N$,

$$\begin{aligned}X_{1,j} &= X_{2,j} = \tilde{w}_{1,j}^{1-\alpha}-Z_\alpha,\\ X_{3,j} &= \frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},\end{aligned}$$

noting the required moments are finite under our assumptions, so that

$$\mathbb{E}\left(\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)^2\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right) = O\left(\frac{1}{N^2}\right).$$

The second term in (68) can be bounded using Cauchy-Schwarz as follows

$$\begin{aligned}&\mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)\\ &\leq\mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)^2\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)^{1/2}\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2}\end{aligned}$$

The second term is $O(N^{-1/2})$ by (63), while the first can be bounded by $O(N^{-3/2})$ using Lemma 3 with $r=6$ and for all $j=1\dots N$:

$$\begin{aligned}X_{1,j} &= X_{2,j} = X_{3,j} = X_{4,j} = \tilde{w}_{1,j}^{1-\alpha}-Z_\alpha,\\ X_{5,j} &= X_{6,j} = \frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}.\end{aligned}$$

Hence by combining with (67), we have

$$\mathbb{E}\left(\tilde{\delta}^{(\alpha)}_{M,N}\right) = \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^2}\right) \tag{69}$$

- **Deducing** $\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)]$**.**

Finally, putting (66) and (69) together, we get

$$\begin{aligned}\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] &= \frac{\left|\frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^2}\right)\right|}{\frac{1}{\sqrt{MN}Z_\alpha}\left(\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial \log \tilde{w}_{1,1}}{\partial \theta_\ell} - \frac{\partial \log Z_\alpha}{\partial \theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)\right)} \\ &= \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial \theta_\ell} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial \log \tilde{w}_{1,1}}{\partial \theta_\ell} - \frac{\partial \log Z_\alpha}{\partial \theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)}\end{aligned}$$

which is exactly (54). Since we have assumed that $\frac{\partial Z_\alpha}{\partial \theta_\ell}$ is non-zero and since this term corresponds to the leading order term, we then deduce that

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] = \Theta(\sqrt{MN})$$

and we thus recover (17).

Similarly, (55) holds for $\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\phi_{\ell'})]$ and we obtain the desired result (18) by splitting the cases $\alpha \in (0,1)$ and $\alpha = 0$. In the former, we have $\partial Z_\alpha / \partial \phi_{\ell'} = \partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha})/\partial \phi_{\ell'} \neq 0$ by (16), so the leading order term is $\Theta(\sqrt{MN})$, while in the latter case we have $\partial Z_\alpha / \partial \phi_{\ell'} = 0$ while $\partial \mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})/\partial \phi_{\ell'} > 0$ and so the leading order term is $\Theta(\sqrt{M/N})$ ■

A.4 Proof of Theorem 2

Proof of Theorem 2 Recall from Proposition 1 that

$$\frac{\partial}{\partial \phi}\ell_N^{(\alpha)}(\theta, \phi; x) = \int\int \prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial \phi}\log w_{\theta,\phi}(f(\varepsilon_j, \phi))\right)\mathrm{d}\varepsilon_{1:N}.$$

We will now follow the reasoning of Tucker et al. (2019). To do so, we expand the total derivative of $\ell_N^{(\alpha)}$ with respect to ϕ by using that

$$\frac{\partial}{\partial \phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi)) = -\frac{\partial}{\partial \phi}\log q_\phi(f(\varepsilon_j,\phi'))|_{\phi'=\phi} + \frac{\partial}{\partial \phi}f(\varepsilon_j,\phi)\;\frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)$$

which gives

$$\begin{aligned}\frac{\partial}{\partial\phi}\ell_N^{(\alpha)}(\theta,\phi;x) &= -\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(f(\varepsilon_j,\phi'))|_{\phi'=\phi}\right)\mathrm{d}\varepsilon_{1:N} \\ &\quad+\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\ \frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)\right)\mathrm{d}\varepsilon_{1:N} \\ &:= -A+B. \end{aligned} \tag{70}$$

Notice now that

$$\begin{aligned} A &= \sum_{j=1}^N\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(f(\varepsilon_j,\phi'))|_{\phi'=\phi}\right)\mathrm{d}\varepsilon_{1:N} \\ &= \sum_{j=1}^N\int\int\prod_{i=1}^N q_\phi(z_i)\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j'=z_j}\right)\mathrm{d}z_{1:N}, \end{aligned} \tag{71}$$

where we have set $z_j' = f(\varepsilon_j,\phi')$. Observe in addition that for all $j = 1\ldots N$, the reparameterization trick implies:

$$\begin{aligned}&\int q_\phi(z_j)\ \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j'=z_j}\ \mathrm{d}z_j \\ &\qquad\qquad = \int q(\varepsilon_j)\ \frac{\partial}{\partial z_j}\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\right)\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\ \mathrm{d}\varepsilon_j\end{aligned}$$

and hence

$$\begin{aligned}&\int q_\phi(z_j)\ \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j=z_j'}\ \mathrm{d}z_j \\ &= (1-\alpha)\int q(\varepsilon_j)\left[\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} - \left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\right)^2\right]\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)\ \mathrm{d}\varepsilon_j.\end{aligned}$$

The desired equality (19) is then obtained by combining the last equality above with (70) and (71). ■

Appendix B. Deferred Proofs and Results of Section 4

B.1 Proof of Theorem 3

Proof of Theorem 3 For convenience in the proof, let us first introduce the notation

$$R_\alpha = w_{\theta,\phi}(Z)^{1-\alpha}$$

with $Z \sim q_\phi$ and let us observe that under (A1) we have that $\mathbb{E}(R_\alpha) > 0$. Furthermore, (A1) and Jensen's inequality applied to the concave function $u \mapsto u^{1-\alpha}$ yield

$$\mathbb{E}(R_\alpha) \leq p_\theta(x)^{1-\alpha} < \infty, \tag{72}$$

meaning that (22) holds. Now decompose the variational gap into the two following terms:

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \left[\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x)\right] + \left[\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)\right].$$

To get the desired result (25), we only need to study the behavior of the term inside the first bracket. This will be done via an adaptation of the proof of (Domke and Sheldon, 2018, Theorem 3) to our more general framework, which is provided here for the sake of completeness. We write

$$\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\mathbb{E}\left(\log\left(1+\delta_{\alpha,N}\right)\right)$$

where for all $z \in \mathbb{R}^d$,

$$\delta_{\alpha,N} = \frac{R_{\alpha,N}}{\mathbb{E}(R_\alpha)} - 1 \in (-1,\infty).$$

The second-order Taylor expansion of $\log\left(1+\delta_{\alpha,N}\right)$ gives

$$\log\left(1+\delta_{\alpha,N}\right) = \delta_{\alpha,N} - \frac{1}{2}\delta_{\alpha,N}^2 + \int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x.$$

Now using that $\mathbb{E}(\delta_{\alpha,N}) = 0$ and that $\mathbb{E}(\delta_{\alpha,N}^2) = \mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))/N$, we deduce

$$\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x) = -\frac{\gamma_\alpha^2}{2N} + \frac{1}{1-\alpha}\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right).$$

All that is left to prove is then that

$$\lim_{N\to\infty} N\left|\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right)\right| = 0. \tag{73}$$

By (Domke and Sheldon, 2018, Lemma 7), we have that for all $\varepsilon > 0$ and all $\beta \in (0,1]$ there exist positive constants C_ε and D_β such that

$$\left|\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right| \leq C_\varepsilon \left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}} + D_\beta |\delta_{\alpha,N}|^{2+\beta}$$

and as a result

$$N\left|\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right)\right| \leq C_\varepsilon N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) + D_\beta N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right). \tag{74}$$

Recall that under our assumptions, there exists $\beta > 0$ such that (23) holds. Without loss of generality, one can assume that $\beta \in (0,1]$. [Indeed, assuming that $\beta > 1$, we can find $0 < \beta' \leq 1 < \beta$ so that

$$\mathbb{E}_{Z\sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta'}) < \infty,$$

which follows from Jensen's inequality applied to the concave function $u \mapsto u^{(2+\beta')/(2+\beta)}$ and from (23).] Let us now show that the two terms in (74) go to 0 as $N \to \infty$ for a suitable choice of ε ($\varepsilon = \beta/3$).

- First term of (74). Observe first that Hölder's inequality with $p = (1+\varepsilon)/\varepsilon$ and $q = 1+\varepsilon$ implies the following:

$$\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) \leq \mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|\right)^{\frac{\varepsilon}{1+\varepsilon}} \mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}.$$

From there, we deduce that

$$\limsup_{N\to\infty} N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right)$$
$$\leq \limsup_{N\to\infty} \mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|\right)^{\frac{\varepsilon}{1+\varepsilon}} \limsup_{N\to\infty}\left[N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}\right],$$

having used that for any two sequences of non-negative real numbers $(a_N)_{N\in\mathbb{N}^*}$ and $(b_N)_{N\in\mathbb{N}^*}$, $\limsup_{N\to\infty}(a_N b_N) \leq \limsup_{N\to\infty} a_N \cdot \limsup_{N\to\infty} b_N$.

We then obtain that the first limit is bounded by a constant by appealing to (24). [Indeed, (24) means that for sufficiently large N, $\mathbb{E}(1/R_{\alpha,N})$ is bounded by a constant, and hence so is $\mathbb{E}(|1/(1+\delta_{\alpha,N})|)$ by combining the boundedness of $\mathbb{E}(1/R_{\alpha,N})$ with (72)].

As for the second limit, (Domke and Sheldon, 2018, Lemma 5) with $s = 2+3\varepsilon \geq 2$ and $U_i = \overline{w}_{\theta,\phi}^{(\alpha)}(Z_i) - 1$ implies that there exists a constant $B_\varepsilon > 0$ such that

$$\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right) \leq B_\varepsilon N^{-(2+3\varepsilon)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1\right|^{2+3\varepsilon}\right).$$

Setting $\varepsilon = \beta/3$, we can rewrite the term on the r.h.s. as

$$B_{\beta/3} N^{-(2+\beta)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1\right|^{2+\beta}\right),$$

leading in particular to the inequality

$$\mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right) \leq B_{\beta/3} N^{-(2+\beta)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1\right|^{2+\beta}\right). \tag{75}$$

Hence, by (23) and since $N^{-(2+\beta)/2} = o(N^{-1})$, we obtain

$$\limsup_{N\to\infty}\left[N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}\right] = 0 \quad \text{when } \varepsilon = \beta/3.$$

As a consequence

$$\limsup_{N\to\infty} N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) = 0 \quad \text{when } \varepsilon = \beta/3. \tag{76}$$

- Second term of (74). Using (75) combined with (23) and since $N^{-(2+\beta)/2} = o(N^{-1})$, we deduce:

$$\lim_{N\to\infty} D_\beta N \mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right) = 0. \tag{77}$$

Combining (74) with (76) and (77) yields (73) and the proof is concluded. ■

B.2 Proof of Proposition 2

Proof of Proposition 2 We prove the two assertions separately.

1. Assume that (23) holds with $\alpha = \alpha_2$, that is, there exists $\beta > 0$ such that

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}^{(\alpha_2)}_{\theta,\phi}(Z) - 1\right|^{2+\beta}\right) < \infty$$

or equivalently using (22) with $\alpha = \alpha_2$ and setting $a_2 := \mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_2})$ so that $a_2 \in (0,\infty)$,

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^{2+\beta}\right) < \infty. \tag{78}$$

We now want to prove that (78) implies (23) with $\alpha = \alpha_1$. Using that $|u^\eta - 1| \le |u-1|$ for all $u \ge 0$ and all $\eta \in (0,1)$, we have: for all $z \in \mathbb{R}^d$,

$$\left|\overline{w}^{(\alpha_1)}_{\theta,\phi}(z) - 1\right| \le \left|\frac{w_{\theta,\phi}(z)^{1-\alpha_2}}{\mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_1})^{\frac{1-\alpha_2}{1-\alpha_1}}} - 1\right|$$

where we have set $\eta = (1-\alpha_1)/(1-\alpha_2)$. Hence,

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}^{(\alpha_1)}_{\theta,\phi}(Z) - 1\right|^{2+\beta}\right) \le \tilde{a}_1^{-1}\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^{2+\beta}\right), \tag{79}$$

where $\tilde{a}_1 = a_1^{(1-\alpha_2)/(1-\alpha_1)}$ with $a_1 := \mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_1})$. Note in particular that a_1 belongs to $(0,\infty)$ as a consequence of (22) with $\alpha = \alpha_1$ and thus so does $\tilde{a}_1$. Now observe that, setting $p = 2+\beta > 1$, Minkowski's inequality implies that

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^p\right)^{\frac{1}{p}} \le \mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^p\right)^{\frac{1}{p}} + \mathbb{E}_{Z\sim q_\phi}\left(|a_2 - \tilde{a}_1|^p\right)^{\frac{1}{p}}$$

that is

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^{2+\beta}\right)^{\frac{1}{2+\beta}} \le \mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^{2+\beta}\right)^{\frac{1}{2+\beta}} + |a_2 - \tilde{a}_1|$$

We then deduce that (23) holds with $\alpha = \alpha_1$ by combining (79) with the inequality above and the fact that (i) $\tilde{a}_1^{-1} < \infty$, (ii) $\mathbb{E}_{Z\sim q_\phi}(|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2|^{2+\beta})^{1/(2+\beta)} < \infty$ by (78) and (iii) $|a_2 - \tilde{a}_1| < \infty$.

2. Assume (24) holds for $\alpha = \alpha_2$. Then by Lemma 4 with $k = 1$ we may pick N such that $\mathbb{E}(1/R_{\alpha_2,N}) < \infty$. Observe now that

$$\begin{aligned}\mathbb{E}\left(1/R_{\alpha_1,N}\right) &= \mathbb{E}\left(\frac{N}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1}}\right)\\ &\leq \mathbb{E}\left(\frac{N}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1}}\;\middle|\; w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1, \dots, N\right)\end{aligned}$$

where we have used that $N/(\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1})$ is a decreasing function of each $w_{\theta,\phi}(Z_i)$, with $w_{\theta,\phi}(Z_i)$ being independent random variables. Since $\alpha_1 > \alpha_2$, it follows that

$$\begin{aligned}\mathbb{E}\left(1/R_{\alpha_1,N}\right) &\leq \mathbb{E}\left(\frac{N p_\theta(x)^{\alpha_1-\alpha_2}}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_2}}\;\middle|\; w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1 \dots N\right)\\ &\leq p_\theta(x)^{\alpha_1-\alpha_2}\mathbb{E}\left(\frac{1}{R_{\alpha_2,N}}\;\middle|\; w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1 \dots N\right)\\ &\leq p_\theta(x)^{\alpha_1-\alpha_2}\frac{\mathbb{E}(1/R_{\alpha_2,N})}{\mathbb{P}\left(w_{\theta,\phi}(Z) \leq p_\theta(x)\right)^N}\\ &< \infty\end{aligned}$$

since $\mathbb{E}(w_{\theta,\phi}(Z)) = p_\theta(x)$ which implies that $\mathbb{P}\left(w_{\theta,\phi}(Z) \leq p_\theta(x)\right) > 0$. We see that there exists a choice of N for which $\mathbb{E}\left(1/R_{\alpha_1,N}\right) < \infty$, from which (24) follows by Lemma 4 with $k = 1$.

■

B.3 Behavior of γ_α^2

Lemma 5 *Let $\alpha \in [0, 1)$. Then, under common integrability and differentiability assumptions*

$$\lim_{\alpha \to 1} \gamma_\alpha^2 = 0.$$

Proof By definition of γ_α^2, we have that

$$\gamma_\alpha^2 = \frac{1}{\mathbb{E}(w_{\theta,\phi}^{1-\alpha})^2} \cdot \frac{1}{1-\alpha}\mathbb{E}\left(\left[w_{\theta,\phi}^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right]^2\right).$$

On the one hand, we have that $\mathbb{E}(w_{\theta,\phi}^{1-\alpha}) \to 1$ as $\alpha \to 1$ under convenient integrability assumptions. On the other hand, for all $z \in \mathbb{R}^d$,

$$\begin{aligned}&\lim_{\alpha \to 1} \frac{1}{1-\alpha}\left[w_{\theta,\phi}(z)^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right]^2\\ &\quad= \lim_{\alpha \to 1}\left\{2\left[w_{\theta,\phi}(z)^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right] \cdot \left[-w_{\theta,\phi}(z)^{1-\alpha}\log w_{\theta,\phi}(z) + \mathbb{E}\left(w_{\theta,\phi}^{1-\alpha}\log w_{\theta,\phi}\right)\right]\right\}\end{aligned}$$

so that under convenient differentiability assumptions,

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} \mathbb{E}\left(\left[w_{\theta,\phi}^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha}) \right]^2 \right) = 0$$

thus implying that $\lim_{\alpha \to 1} \gamma_\alpha^2 = 0$. ■

B.4 Proof of Example 1

Proof of Example 1 Using (26), we first deduce that: for all $m \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{Z \sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^m\right) &= \mathbb{E}_{S \sim \mathcal{N}(0,1)}\left[\exp\left(-\frac{m\sigma^2 d}{2} - m\sigma\sqrt{d}S\right)\right] \\
&= \exp\left(-\frac{m\sigma^2 d}{2}\right)\mathbb{E}_{S \sim \mathcal{N}(0,1)}\left[\exp\left(-m\sigma\sqrt{d}S\right)\right] \\
&= \exp\left(-\frac{m\sigma^2 d}{2}\right)\exp\left(\frac{m^2\sigma^2 d}{2}\right) \\
&= \exp\left(\frac{m(m-1)\sigma^2 d}{2}\right).
\end{aligned}$$

Therefore, plugging in $m = 1-\alpha$,

$$\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x) = \frac{1}{1-\alpha}\log \mathbb{E}_{Z \sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha}\right) = -\frac{\alpha\sigma^2 d}{2},$$

which gives the desired result for $\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)$. In addition: for all $m \in \mathbb{R}$,

$$\mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^m) = \exp\left(\frac{m(m-1)\sigma^2 d}{2}\right) p_\theta(x)^m. \tag{80}$$

Now note that (26) can be rewritten as: for all $i = 1 \dots N$,

$$\log w_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i + \log p_\theta(x), \quad S_i \sim \mathcal{N}(0,1).$$

Hence, we get that: for all $i = 1 \dots N$,

$$\begin{aligned}
\log \overline{w}_{\theta,\phi}^{(\alpha)}(z_i) &= (1-\alpha)\log w_{\theta,\phi}(z_i) - \log \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha}) \\
&= -(1-\alpha)\sigma\sqrt{d}S_i - \frac{(1-\alpha)^2\sigma^2 d}{2}
\end{aligned}$$

where we have used (80) with $m = 1-\alpha$. As a result,

$$\begin{aligned}
\mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) &= \mathbb{V}_{S \sim \mathcal{N}(0,1)}\left(\exp\left\{-(1-\alpha)\sigma\sqrt{d}S\right\}\exp\left\{-\frac{(1-\alpha)^2\sigma^2 d}{2}\right\}\right) \\
&= \exp\left((1-\alpha)^2\sigma^2 d\right)\left(\exp\left((1-\alpha)^2\sigma^2 d\right) - 1\right)\exp\left(-(1-\alpha)^2\sigma^2 d\right) \\
&= \exp\left((1-\alpha)^2\sigma^2 d\right) - 1,
\end{aligned}$$

which yields the desired result for γ_α^2. In addition, we show that (23) and (24) both hold, meaning that we can apply Theorem 3. To see this, note that $\mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^m)$ and $\mathbb{E}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^m)$ are well-defined and finite for all $\alpha, m \in \mathbb{R}$. Furthermore, observe that by convexity of the function $u \mapsto |u|^{2+\beta}$ with $\beta > 0$, it holds that

$$\mathbb{E}_{Z\sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta}) \leq 2^{1+\beta}\left(\mathbb{E}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^{2+\beta}) + 1\right)$$

thus (23) holds for all $\beta > 0$. Lastly, $\mathbb{E}(1/R_{\alpha,N}) \leq \mathbb{E}(N^{-1}\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{\alpha-1})$ by the HM-AM inequality and the r.h.s. is finite for all $\alpha \in \mathbb{R}$ thus (24) also holds.

In the particular case $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$: for all $i = 1\ldots N$,

$$\begin{aligned}
\log \overline{w}_{\theta,\phi}(z_i) &= -\frac{1}{2}\left(\|z_i-\theta\|^2 - \|z_i-\phi\|^2\right)\\
&= -\frac{1}{2}\left(\|\theta\|^2 - \|\phi\|^2 + 2\langle z_i, \phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(\langle\theta-\phi,\theta+\phi\rangle + 2\langle z_i,\phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(\langle\theta-\phi,\theta-\phi+2\phi\rangle + 2\langle z_i,\phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(B_d^2 + 2\langle z_i-\phi,\phi-\theta\rangle\right)\\
&= -\frac{B_d^2}{2} - B_d S_i \quad \text{with } S_i = \tfrac{1}{B_d}\langle z_i-\phi,\phi-\theta\rangle, \qquad (81)
\end{aligned}$$

where we have set $B_d = \|\phi-\theta\|$. Since $z_i - \phi \sim \mathcal{N}(0,\boldsymbol{I}_d)$, it follows that $S_i \sim \mathcal{N}(0,1)$ as required. Lastly, when $\theta = 0\cdot\boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, we have that $B_d = \sqrt{d}$. ■

B.5 Proof of Example 2

Proof Let us first prove that $p_\theta(x) = \mathcal{N}(x;\theta,2\boldsymbol{I}_d)$ and $p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2, 1/2\ \boldsymbol{I}_d)$. To see this, note that

$$p_\theta(x,z) = \left(\frac{1}{(2\pi)^{d/2}}\right)^2 \exp\left(-\frac{1}{2}\left\{\|z-\theta\|^2 + \|x-z\|^2\right\}\right).$$

As a result, only considering the dependency in z, we have that

$$p_\theta(x,z) \propto \exp\left(-\frac{1}{2}\cdot 2\left\|z - \frac{\theta+x}{2}\right\|^2\right),$$

which implies that $p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2, 1/2\ \boldsymbol{I}_d)$. Furthermore,

$$\begin{aligned}
p_\theta(x) &= \int p_\theta(x,z)\mathrm{d}z\\
&= \frac{1}{(2\pi\cdot 2)^{d/2}}\int \frac{1}{(2\pi\cdot 1/2)^{d/2}}\exp\left(-\frac{1}{2}\cdot 2\left\|z-\frac{\theta+x}{2}\right\|^2\right)\mathrm{d}z\cdot\exp\left(-\frac{1}{2}\cdot\frac{1}{2}\|\theta-x\|^2\right)\\
&= \mathcal{N}(x;\theta,2\boldsymbol{I}_d).
\end{aligned}$$

Hence, for all $z \in \mathbb{R}^d$

$$\log \overline{w}_{\theta,\phi}(z) = \log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) = \log\left(\frac{(2\pi \cdot 2/3)^{d/2}}{(2\pi \cdot 1/2)^{d/2}} \exp\left[-\left\|z - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z - Ax - b\|^2\right]\right)$$

and from there, we can straightforwardly deduce that: for all $i = 1 \dots N$,

$$\log \overline{w}_{\theta,\phi}(z_i) = \frac{d}{2}\log\left(\frac{4}{3}\right) - \left\|z_i - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z_i - Ax - b\|^2. \tag{82}$$

Using (82) we can write that

$$\begin{aligned}\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\overline{w}_{\theta,\phi}(z)^{1-\alpha}\mathrm{d}z\right) \\ &= \frac{d}{2}\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log(I)\end{aligned}$$

where

$$\begin{aligned}I &= \int \frac{1}{(4\pi/3)^{d/2}}\exp\left[-\frac{3}{4}\|z - Ax - b\|^2 + (1-\alpha)\left(-\left\|z - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z - Ax - b\|^2\right)\right]\mathrm{d}z \\ &= \int \frac{1}{(4\pi/3)^{d/2}}\exp\left[-\frac{3\alpha}{4}\|z - Ax - b\|^2 - (1-\alpha)\left\|z - \frac{\theta + x}{2}\right\|^2\right]\mathrm{d}z.\end{aligned}$$

I is well-defined and finite for all $\alpha < 4$. Completing the square leads to

$$\begin{aligned}I = \left(\frac{3}{4-\alpha}\right)^{d/2}\exp\Bigg(\left(\frac{4}{4-\alpha}\right)\Bigg\|\frac{3\alpha}{4}(Ax + b) \\ +(1-\alpha)\frac{\theta + x}{2}\Bigg\|^2 - \frac{3\alpha}{4}\|Ax + b\|^2 - (1-\alpha)\left\|\frac{\theta + x}{2}\right\|^2\Bigg).\end{aligned}$$

As a result,

$$\begin{aligned}\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) &= \frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] \\ &+ \frac{4}{(4-\alpha)(1-\alpha)}\left\|\frac{3\alpha}{4}(Ax + b) + (1-\alpha)\frac{\theta + x}{2}\right\|^2 - \frac{3\alpha}{4(1-\alpha)}\|Ax + b\|^2 - \left\|\frac{\theta + x}{2}\right\|^2.\end{aligned}$$

It can then be checked that the second line simplifies to $-\frac{3\alpha}{4-\alpha}\left\|Ax + b - \frac{\theta + x}{2}\right\|^2$, from which we deduce the desired result for $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$. [Notice in particular that $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ is well-defined for all $\alpha \neq 1, \alpha < 4$ with continuous extension at $\alpha = 1$.] On the other hand, we have that

$$\gamma_\alpha^2 = \frac{1}{1-\alpha}\mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) = \frac{1}{1-\alpha}\left(\frac{\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{2-2\alpha}\right)}{\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha}\right)^2} - 1\right).$$

Furthermore, for all $\alpha' \in \mathbb{R} \setminus \{1\}$, it holds that

$$\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha'}\right) = \exp\left((1-\alpha')\left[\mathcal{L}^{(\alpha')}(\theta, \phi; x) - \ell(\theta; x)\right]\right). \tag{83}$$

Now using (83) with $\alpha' = \alpha$ and $\alpha' = 2\alpha - 1$ and combining with the expression of $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$, we obtain

$$\gamma_\alpha^2 = \frac{1}{1-\alpha} \left(\exp(A) - 1 \right)$$

with

$$\begin{aligned} A &= 2(1-\alpha) \left\{ \frac{d}{2} \left[\log\left(\frac{4}{3}\right) + \frac{1}{2(1-\alpha)} \log\left(\frac{3}{5-2\alpha}\right) \right] - \frac{3(2\alpha-1)}{5-2\alpha} \left\| Ax + b - \frac{\theta + x}{2} \right\|^2 \right\} \\ &\quad - 2(1-\alpha) \left\{ \frac{d}{2} \left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha} \log\left(\frac{3}{4-\alpha}\right) \right] - \frac{3\alpha}{4-\alpha} \left\| Ax + b - \frac{\theta + x}{2} \right\|^2 \right\} \\ &= \frac{d}{2} \log\left(\frac{(4-\alpha)^2}{5-2\alpha} \right) + \frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)} \left\| Ax + b - \frac{\theta + x}{2} \right\|^2 \end{aligned}$$

from which we deduce the desired result for γ_α^2 [notice in particular that γ_α^2 is well-defined for all $\alpha < 5/2$].

In addition, the assumptions made in Theorem 3 are satisfied for all $\alpha \in [0, 1)$. To see this, set $m = 1 - \alpha'$ in (83) and use that $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ is well-defined for all $\alpha \neq 1, \alpha < 4$ with continuous extension at $\alpha = 1$, so that $\mathbb{E}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}(Z)^m)$ and thus $\mathbb{E}_{Z \sim q_\phi}\left(w_{\theta,\phi}(Z)^m\right)$ and $\mathbb{E}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^m)$ are well-defined and finite for all $m > -3$. Furthermore, the convexity of the function $u \mapsto |u|^{2+\beta}$ with $\beta > 0$ implies that

$$\mathbb{E}_{Z \sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta}) \leq 2^{1+\beta} \left(\mathbb{E}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^{2+\beta}) + 1 \right)$$

thus (23) holds for all $\beta > 0$. Lastly, $\mathbb{E}(1/R_{\alpha,N}) \leq \mathbb{E}(N^{-1} \sum_{i=1}^N w_{\theta,\phi}(Z_i)^{\alpha-1})$ by the HM-AM inequality and the r.h.s. is finite for all $\alpha > -2$ thus (24) also holds. ■

B.6 Proof of Proposition 3

Proof of Proposition 3 For all $\alpha \in [0, 1)$, we can rewrite the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi)$ as

$$\begin{aligned} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) &= \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i) \log\left(\frac{1}{N} \sum_{j=1}^N \overline{w}_j^{1-\alpha} \right) \mathrm{d}\overline{w}_{1:N} \\ &= \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q_\phi(z_i) \log\left(\frac{1}{N} \sum_{j=1}^N (\overline{w}^{(j)})^{1-\alpha} \right) \mathrm{d}\overline{w}_{1:N} \\ &= \frac{1}{1-\alpha} \left[\int \int \prod_{i=1}^N q_\phi(z_i) \log\left(\frac{1}{N} (\overline{w}^{(N)})^{1-\alpha} \right) \mathrm{d}\overline{w}_{1:N} \right. \\ &\qquad \left. + \int \int \prod_{i=1}^N q_\phi(z_i) \log\left(1 + \sum_{j=1}^{N-1} \left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}} \right)^{1-\alpha} \right) \mathrm{d}\overline{w}_{1:N} \right] \\ &= \Delta_{N,d}^{(\alpha,MAX)}(\theta, \phi; x) + R_{N,d}^{(\alpha)}(\theta, \phi; x) \end{aligned}$$

where we have used (29) and where we have set

$$R_{N,d}^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i)\log\left(1+\sum_{j=1}^{N-1}\left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}}\right)^{1-\alpha}\right)\mathrm{d}\overline{w}_{1:N}.$$

All that is left to do is now to prove (30). Observe that by definition of $T_{N,d}^{(\alpha)}$ in (27) and since $\alpha \in [0,1)$, we can write

$$\begin{aligned}0 \le R_{N,d}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\int\int\prod_{i=1}^{N} q_\phi(z_i)\log\left(1+T_{N,d}^{(\alpha)}\right)\mathrm{d}\overline{w}_{1:N}\\ &\le \frac{1}{1-\alpha}\int\prod_{i=1}^{N} q_\phi(z_i)\ T_{N,d}^{(\alpha)}\ \mathrm{d}\overline{w}_{1:N}\\ &= \frac{1}{1-\alpha}\mathbb{E}(T_{N,d}^{(\alpha)}),\end{aligned}$$

which concludes the proof. ■

B.7 Deferred Proofs of Section 4.2.1

B.7.1 Proof of Lemma 1

Proof of Lemma 1 First, note that since $S_1,\dots,S_N$ are i.i.d. normal random variables, so are $-S_1,\dots,-S_N$. Setting $M_N = \max_{1\le i\le N} -S_i$, we also have $S^{(1)} = -M_N$. A standard result (obtained, for example, by combining Theorem 1.1.2 and Example 1.1.7 in de Haan and Ferreira, 2007) is that for all $x \in \mathbb{R}$,

$$\lim_{N\to\infty} P\left(a_N^{-1}\left(M_N - b_N\right) \le x\right) = \exp(-e^{-x}) \tag{84}$$

with $a_N = 1/\sqrt{2\log N}$ and $b_N = \sqrt{2\log N} - \frac{1}{2}(\log\log N + \log 4\pi)/(\sqrt{2\log N})$. Since $\mathbb{E}(|M_N|) \le \mathbb{E}(M_N^2)^{1/2} \le \mathbb{E}(\sum_{i=1}^N S_i^2)^{1/2} \le N^{1/2} < \infty$ for all N, it follows by (Pickands III, 1968, Theorem 2.1) that

$$\lim_{N\to\infty} a_N^{-1}\left(\mathbb{E}\left(M_N\right) - b_N\right) = \mathbb{E}(U),$$

where U is a Gumbel random variable and $\mathbb{E}(U)$ is given by the Euler–Masceroni constant. Using that $S^{(1)} = -M_N$, we deduce

$$\lim_{N\to\infty} -a_N^{-1}\left(\mathbb{E}(S^{(1)}) + b_N\right) = \mathbb{E}(U).$$

Finally, plugging in the definition of a_N and b_N, we obtain

$$\begin{aligned}\mathbb{E}(S^{(1)}) &= -\sqrt{2\log N} + \frac{\log\log N + \log 4\pi}{2\sqrt{2\log N}} - \frac{\mathbb{E}(U)}{\sqrt{2\log N}} + o\left(\frac{1}{\sqrt{2\log N}}\right)\\ &= -\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\end{aligned}$$

and we have thus recovered (34). ■

B.7.2 Proof of Proposition 4

Proof of Proposition 4 First note that

$$\log \overline{w}^{(N)} = -\frac{d\sigma^2}{2} - \sqrt{d}\sigma S^{(1)}.$$

Combining this result with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi; x)$ in (29) yields

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{d\sigma^2}{2} - \sqrt{d}\sigma \mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}.$$

Now using (34), we deduce

$$\begin{aligned}\Delta_{N,d}^{MAX}(\theta, \phi) &= -\frac{d\sigma^2}{2} + \sqrt{d}\sigma\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha - 1} \\ &= -\frac{d\sigma^2}{2}\left\{1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\},\end{aligned}$$

which concludes the proof. ■

B.7.3 Proof of Proposition 5

We first prove a useful intermediate lemma regarding the concentration of $S^{(1)}$.

Lemma 6 *Let $S_1, \ldots, S_N$ be i.i.d. normal random variables, set $S^{(1)} = \min_{1\leq i\leq N} S_i$, and define $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$. Then as $N \to \infty$, we have*

$$\mathbb{P}(S^{(1)} \notin I_N) = O\left(\frac{1}{N^4}\right).$$

Proof We control the probability of the events $\{S^{(1)} > -\sqrt{\log N}\}$ and $\{S^{(1)} < -4\sqrt{\log N}\}$ separately. First, note that

$$\begin{aligned}\log \mathbb{P}(S^{(1)} > -\sqrt{\log N}) &= N\log(\overline{\Phi}(-\sqrt{\log N})) \\ &= N\log\left(1 - \overline{\Phi}(\sqrt{\log N})\right) \\ &= -N\overline{\Phi}(\sqrt{\log N})(1 + o(1)) \\ &= -N\frac{\phi(\sqrt{\log N})}{\sqrt{\log N}}(1 + o(1)) \qquad (85)\end{aligned}$$

where in the final line we have used the standard approximation

$$\overline{\Phi}(x) = \frac{\phi(x)}{x}(1 + o(1)) \qquad (86)$$

as $x \to \infty$. We deduce that

$$\mathbb{P}(S^{(1)} > -\sqrt{\log N}) = \exp\left\{-N\frac{N^{-1/2}}{\sqrt{2\pi}\sqrt{\log N}}(1 + o(1))\right\} = O\left(\frac{1}{N^4}\right) \qquad (87)$$

as $N \to \infty$. Second, as $N \to \infty$ we have, by a union bound, that

$$\begin{aligned}\mathbb{P}(S^{(1)} < -4\sqrt{\log N}) &\leq N\Phi(-4\sqrt{\log N}) \\ &= N\frac{\phi(4\sqrt{\log N})}{4\sqrt{\log N}}(1+o(1)) \\ &= \frac{N^{-7}}{4\sqrt{2\pi}\sqrt{\log N}}(1+o(1)) \\ &= O\left(\frac{1}{N^4}\right) \qquad (88)\end{aligned}$$

and so the result follows. ■

We now prove Proposition 5 by building on the proof from Snyder et al. (2008) and on Lemma 6.

Proof of Proposition 5 Denote $\sigma_\alpha = (1-\alpha)\sigma$ for all $\alpha \in [0,1)$. A first remark is that, conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N-1$ i.i.d. random variables

$$\mathbb{E}(T^{(\alpha)}_{N,d}|S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S - S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{\phi(z)}{\overline{\Phi}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with $\phi(z)$ denoting the standard normal density and $\overline{\Phi}(x) = \int_x^\infty \phi(z)\mathrm{d}z$ denoting the normalizing constant. Then,

$$\mathbb{E}(T^{(\alpha)}_{N,d}|S^{(1)}) = \frac{(N-1)\int_{S^{(1)}}^\infty \exp\left(-\sigma_\alpha\sqrt{d}\left(z - S^{(1)}\right)\right)\phi(z)\mathrm{d}z}{\overline{\Phi}(S^{(1)})}.$$

We can then calculate explicitly

$$\begin{aligned}&\int_{S^{(1)}}^\infty \exp\left(-\sigma_\alpha\sqrt{d}\left(z - S^{(1)}\right)\right)\phi(z)\mathrm{d}z \\ &= \exp(\sigma_\alpha\sqrt{d}S^{(1)} + \sigma_\alpha^2 d/2)\int_{S^{(1)}}^\infty (\sqrt{2\pi})^{-1}\exp\left(-\frac{1}{2}(z+\sigma_\alpha\sqrt{d})^2\right)\mathrm{d}z \\ &= \exp(\sigma_\alpha\sqrt{d}S^{(1)} + \sigma_\alpha^2 d/2)\overline{\Phi}(\sigma_\alpha\sqrt{d} + S^{(1)}).\end{aligned}$$

Denoting $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$, on the event $\{S^{(1)} \in I_N\}$, as $N, d \to \infty$ with $\log N/d \to 0$, we have

$$\sigma_\alpha\sqrt{d} + S^{(1)} = \sigma_\alpha\sqrt{d}(1 + o_{N,d}(1)),$$

where we use the notation $o_{N,d}(1)$ to denote that the implicit constant, which goes to zero as $N, d \to \infty$ with $\log N/d \to 0$, does not depend on $S^{(1)}$. Using the approximation (86) for $\Phi(x)$ as $x \to \infty$,

$$\overline{\Phi}(\sigma_\alpha\sqrt{d} + S^{(1)}) = \frac{\phi(\sigma_\alpha\sqrt{d} + S^{(1)})}{\sigma_\alpha\sqrt{d} + S^{(1)}}(1 + o_{N,d}(1)).$$

Hence, observing that $\exp(\sigma_\alpha\sqrt{d}S^{(1)}+\sigma_\alpha^2 d/2)\phi(\sigma_\alpha\sqrt{d}+S^{(1)})=\phi(S^{(1)})$, it follows that

$$\int_{S^{(1)}}^{\infty}\exp\left(-\sigma_\alpha\sqrt{d}\left(z-S^{(1)}\right)\right)\phi(z)\mathrm{d}z=\frac{\phi(S^{(1)})}{\sigma_\alpha\sqrt{d}}(1+o_{N,d}(1))$$

on the event $\{S^{(1)}\in I_N\}$. Using (86), we can also write

$$\Phi(S^{(1)})=\overline{\Phi}(-S^{(1)})=\frac{\phi(-S^{(1)})}{-S^{(1)}}(1+o_{N,d}(1)).$$

Combined with $\phi(-S^{(1)})=\phi(S^{(1)})$, this allows us to deduce that

$$\begin{aligned}\int_{S^{(1)}}^{\infty}\exp\left(-\sigma_\alpha\sqrt{d}\left(z-S^{(1)}\right)\right)\phi(z)\mathrm{d}z&=\frac{(-S^{(1)})\Phi(S^{(1)})}{\sigma_\alpha\sqrt{d}}(1+o_{N,d}(1))\\&\leq\Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o_{N,d}(1)),\end{aligned}\tag{89}$$

all on the event $\{S^{(1)}\in I_N\}$. Finally, since $S^{(1)}<-\sqrt{\log N}$ implies

$$\overline{\Phi}(S^{(1)})=1+o_{N,d}(1),$$

we have

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)})\leq(N-1)\Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o_{N,d}(1)).$$

We conclude by using the tower law of expectation and splitting according to whether $S^{(1)}\in I_N$, giving

$$\begin{aligned}\mathbb{E}(T_{N,d}^{(\alpha)})&=\mathbb{E}\left(\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)})\right)\\&\leq\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}(N-1)\Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o_{N,d}(1))\right)+\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\notin I_N\}}\right)\\&\leq(N-1)\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}\mathbb{E}\left(\Phi(S^{(1)})\right)(1+o(1))+\mathbb{P}(S^{(1)}\notin I_N)\\&\leq\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o(1))+O\left(\frac{1}{N^4}\right)\to 0\end{aligned}$$

where in the final line we have used Lemma 6 and that $\Phi(S^{(1)})$ is distributed as the minimum of N independent uniform random variables on $[0,1]$ and so $\mathbb{E}(\Phi(S^{(1)}))=\frac{1}{N+1}$.

■

B.7.4 Proof of Theorem 5

First note that Lemma 1 and Lemma 6 are not affected by the change in the distribution of the weights we have made in (36). As for Proposition 4 and Proposition 5, they are modified according to Proposition 8 and Proposition 9 below.

Proposition 8 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (36) *and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. Then, for all $\alpha \in [0, 1)$,*

$$\lim_{N,d\to\infty} \Delta_{N,d}^{(\alpha,MAX)}(\theta, \phi; x) + \frac{B_d^2}{2}\left\{1 - 2\ \frac{\sqrt{2\log N}}{B_d} + \frac{1}{1-\alpha}\ \frac{2\log N}{B_d^2} + O\left(\frac{\log\log N}{B_d\sqrt{\log N}}\right)\right\} = 0.$$

Proof First note that

$$\log \overline{w}^{(N)} = -\frac{B_d^2}{2} - B_d S^{(1)}.$$

Combining this result with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi; x)$ in (29) yields

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{B_d^2}{2} - B_d \mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}.$$

Now using (34) of Lemma 1, we deduce

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{B_d^2}{2} + B_d\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha - 1},$$

which concludes the proof. ■

Proposition 9 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (36) *and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. Then, for all $\alpha \in [0, 1)$, we have*

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0.$$

Proof Conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N - 1$ i.i.d. random variables

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-(1-\alpha)B_d(S - S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{\phi(z)}{\overline{\Phi}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with $\phi(z)$ denoting the standard normal density and $\overline{\Phi}(x) = \int_x^\infty \phi(z)\mathrm{d}z$ denoting the normalizing constant. Now denoting $\sigma_\alpha = (1-\alpha)\sigma_-$ for all $\alpha \in [0, 1)$ and using that

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) \leq (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S - S^{(1)})\right)\right)$$

we obtain by following the proof of Proposition 5 that the term on the r.h.s. above goes to 0 as $\log N/d \to 0$ with $N, d \to \infty$. We deduce the desired result by combining this with the fact that $\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) \geq 0$. ■

The proof of Theorem 5 then follows immediately from Proposition 8 and Proposition 9.

B.8 Deferred Proofs and Results of Section 4.2.2

We start by recalling some useful results from Saulis and Statulevičius (2000) regarding large deviations for sums of independent random variables.

B.8.1 Large Deviations for Sums of Independent Random Variables

The random variable ξ is said to satisfy the assumption (A-ξ) if the following holds.

(A-ξ) There exists $\Delta > 0$ such that $|\Gamma_k(\xi)| \leq \frac{k!}{\Delta^{k-2}}$ for all integer $k \geq 3$, where $\Gamma_k(\xi)$ denotes the k-th cumulant of ξ.

We now state without proof (Saulis and Statulevičius, 2000, Lemma 2.3) and (Saulis and Statulevičius, 2000, Theorem 3.1) in the particular case $\gamma = 0$.

Lemma 7 ((Saulis and Statulevičius, 2000, Lemma 2.3) with $\gamma = 0$) *Let ξ be a random variable with $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\xi^2) = 1$. Denote by $G(\cdot)$ the cdf of ξ. Assume that* (A-ξ) *holds and set*

$$\Delta_0 = \frac{\sqrt{2}}{36}\Delta.$$

Then, in the interval $0 \leq x < \Delta_0$, the relations of large deviations

$$1 - G(x) = (1 - \Phi(x)) \exp(P(x)) \left(1 + \theta_1 f(x) \frac{x+1}{\Delta_0}\right)$$
$$G(-x) = \Phi(-x) \exp(P(-x)) \left(1 + \theta_2 f(x) \frac{x+1}{\Delta_0}\right)$$

are valid, with Φ denoting the standard normal distribution. Here, P and f are defined by

$$P(x) = \sum_{k=3}^{\infty} \lambda_k x^k + \theta \left(x/\Delta_0\right)^3$$
$$f(x) = \frac{60(1 + 10\Delta_0^2 \exp\left\{-\left(1 - x/\Delta_0\right)\sqrt{\Delta_0}\right\})}{1 - x/\Delta_0},$$

where $\theta, \theta_1, \theta_2$ are some variables not exceeding 1 in absolute value and where for all $k \geq 3$

$$|\lambda_k| \leq \frac{2}{k}\left(16/\Delta\right)^{k-2}$$

so that

$$P(x) \leq \frac{x^3}{2(x + 8\Delta_0)} \quad \text{and} \quad P(-x) \geq -\frac{x^3}{3\Delta_0}.$$

Theorem 1 ((Saulis and Statulevičius, 2000, Theorem 3.1) with $\gamma = 0$) *Let $\xi_1, \ldots, \xi_d$ be independent random variables with $\mathbb{E}(\xi_j) = 0$ and $\sigma_j^2 = \mathbb{V}(\xi_j) < \infty$. Set*

$$\boldsymbol{S}_d = \frac{1}{B_d}\left(\xi_1 + \ldots + \xi_d\right),$$

where $B_d^2 = \sum_{j=1}^d \sigma_j^2$*. Assume that there exists* $K > 0$ *such that: for all* $j = 1 \dots d$*,*

$$|\mathbb{E}(\xi_j^k)| \leq k! K^{k-2} \sigma_j^2, \quad k \geq 3. \tag{90}$$

Then,

$$|\Gamma_k(\boldsymbol{S}_d)| \leq \frac{k!}{\Delta_d^{k-2}}, \quad k \geq 3$$

with

$$\Delta_d = \frac{B_d}{K_d}, \quad \textit{where} \quad K_d = 2 \max \left\{ K, \max_{1 \leq j \leq d} \sigma_j \right\},$$

that is, (A-ξ) *holds with* $\xi = \boldsymbol{S}_d$ *and* $\Delta = \Delta_d$*.*

B.8.2 Preliminary Results

Building on Lemma 7 and Theorem 1, we can now state some preliminary results that will come in handy when proving the results from Section 4.2.2.

Lemma 8 *Let* $\xi_1, \dots, \xi_d$ *be i.i.d. random variables with* $\mathbb{E}(\xi_1) = 0$ *and* $\sigma^2 = \mathbb{V}(\xi_1) < \infty$*. Set*

$$\boldsymbol{S}_d = \frac{1}{B_d} (\xi_1 + \ldots + \xi_d),$$

where $B_d = \sigma\sqrt{d}$*. Assume that there exists* $K > 0$ *such that:*

$$|\mathbb{E}(\xi_1^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Set $\Delta_d = B_d / K_d$ *where* $K_d = 2 \max\{K, \sigma\}$*. Then, as* $d \to \infty$*, there exists an analytic function* P_d *such that the cdf of* $\boldsymbol{S}_d$*, denoted* $G_d(\cdot)$*, satisfies*

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x)) \exp(P_d(x))(1 + o(1)) \\ G_d(-x) &= \Phi(-x) \exp(P_d(-x))(1 + o(1)) \end{aligned}$$

uniformly for all $x \geq 0$ *and* $x = o(\sqrt{d})$*. Here,* Φ *denotes the standard normal distribution,* P_d *is such that*

$$P_d(x) = \sum_{k=3}^{\infty} \lambda_{k,d} x^k$$

with

$$|\lambda_{k,d}| \leq A (c/\sqrt{d})^{k-2}, \quad k \geq 3$$

for some constants $A, c > 0$*.*

Proof Observe first that $\boldsymbol{S}_d$ satisfies $\mathbb{E}(\boldsymbol{S}_d) = 0$ and $\mathbb{E}(\boldsymbol{S}_d^2) = 1$ with $B_d = \sigma\sqrt{d}$ and $K_d = 2\max(K, \sigma)$. Furthermore, (A-ξ) holds with $\xi = \boldsymbol{S}_d$ and $\Delta = \Delta_d$ by Theorem 1. Then, we can apply Lemma 7 with $\xi = \boldsymbol{S}_d$ and $\Delta = \Delta_d$ to obtain that in the interval $0 \leq x < \Delta_{0,d}$, the relations of large deviations

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x)) \exp(P(x)) \left(1 + \theta_1 f(x) \frac{x+1}{\Delta_{0,d}}\right), \\ G_d(-x) &= \Phi(-x) \exp(P(-x)) \left(1 + \theta_2 f(x) \frac{x+1}{\Delta_{0,d}}\right) \end{aligned}$$

are valid. Here, $\Delta_{0,d} = \frac{\sqrt{2}}{36} \Delta_d$ and P, f are defined by

$$\begin{aligned} P(x) &= P_d(x) + \theta \left(x/\Delta_{0,d}\right)^3 \\ f(x) &= \frac{60(1 + 10\Delta_{0,d}^2 \exp\left\{-\left(1 - x/\Delta_{0,d}\right)\sqrt{\Delta_{0,d}}\right\})}{1 - x/\Delta_{0,d}} \\ P_d(x) &= \sum_{k=3}^{\infty} \lambda_{k,d} x^k, \end{aligned}$$

where $\theta, \theta_1, \theta_2$ are some variables not exceeding 1 in absolute value and

$$|\lambda_{k,d}| \leq \frac{2}{k}(16/\Delta_d)^{k-2} \leq A(c/\sqrt{d})^{k-2}, \quad k \geq 3$$

for some constants $A, c > 0$. Under the assumption $x = o(\sqrt{d})$, $P(x) = P_d(x) + o(1)$, $f(x)\frac{x+1}{\Delta_{0,d}} = o(1)$ and we can thus deduce that as $d \to \infty$ the relations of large deviations become

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x)) \exp(P_d(x)) \left(1 + o(1)\right) \\ G_d(-x) &= \Phi(-x) \exp(P_d(-x)) \left(1 + o(1)\right) \end{aligned}$$

uniformly for all $x \geq 0$ and $x = o(\sqrt{d})$. ■

The corollary below then follows from Lemma 8.

Corollary 1 *Under the assumptions of Lemma 8, as $d \to \infty$,*

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x))(1 + o(1)) \\ G_d(-x) &= \Phi(-x)(1 + o(1)) \end{aligned}$$

uniformly for all $x \geq 0$ and $x = o(d^{1/6})$.

Proof Since we consider the case $x \geq 0$ and $x = o(d^{1/6})$, we can apply Lemma 8 to get: as $d \to \infty$,

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x)) \exp(P_d(x))(1 + o(1)) \\ G_d(-x) &= \Phi(-x) \exp(P_d(-x))(1 + o(1)) \end{aligned}$$

where P_d is defined in Lemma 8. In addition, using successively that (i) $|\lambda_{k,d}| \leq A(c/\sqrt{d})^{k-2}$ by Lemma 8 (ii) $x = o(\sqrt{d})$ and (iii) $x^3 = o(\sqrt{d})$, we have that:

$$\begin{aligned}
|P_d(x)| &\leq \sum_{k=3}^{\infty} |\lambda_{k,d}| x^k \\
&\leq Acx^3 d^{-1/2} \sum_{k=3}^{\infty} \left(cxd^{-1/2}\right)^{k-3} \\
&\leq Acx^3 d^{-1/2}(1 + o(1)) \\
&= o(1).
\end{aligned}$$

Similarly, $|P_d(-x)| = o(1)$ and consequently,

$$\begin{aligned}
1 - G_d(x) &= (1 - \Phi(x))(1 + o(1)) \\
G_d(-x) &= \Phi(-x)(1 + o(1))
\end{aligned}$$

uniformly for all $x \geq 0$ and $x = o(d^{1/6})$. ■

We also prove the following concentration result, which parallels the corresponding result Lemma 6 from the exact log-normal case and which will be useful in subsequent proofs.

Lemma 9 *Let $S_1, \ldots, S_N$ be i.i.d. distributed according to* (38)*, set $S^{(1)} = \min_{1 \leq i \leq N} S_i$ and define $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$. Then as $N, d \to \infty$ with $\log N / d^{1/3} \to 0$, we have*

$$\mathbb{P}(S^{(1)} \notin I_N) = O\left(\frac{1}{N^4}\right).$$

Proof The proof follows the same structure as the proof of Lemma 6, using Corollary 1 to relate the approximately log-normal case to the exact case.

We control the probability of the events $\{S^{(1)} > -\sqrt{\log N}\}$ and $\{S^{(1)} < -4\sqrt{\log N}\}$ separately. First, since $\sqrt{\log N} = o(d^{1/6})$ as $N, d \to \infty$ with $\log N / d^{1/3} \to 0$, by Corollary 1 we have

$$\begin{aligned}
\log \mathbb{P}(S^{(1)} > -\sqrt{\log N}) &= N \log\left(1 - G_d(-\sqrt{\log N})\right) \\
&= N \log\left(1 - (1 + o(1))\overline{\Phi}(\sqrt{\log N})\right) \\
&= -N \frac{\phi(\sqrt{\log N})}{\sqrt{\log N}}(1 + o(1))
\end{aligned}$$

using the same method as in (85). Following (87), we deduce that

$$\mathbb{P}(S^{(1)} > -\sqrt{\log N}) = O\left(\frac{1}{N^4}\right)$$

Second, we can write

$$\begin{aligned}
\mathbb{P}(S^{(1)} < -4\sqrt{\log N}) &\leq N G_d(-4\sqrt{\log N}) \\
&= N\overline{\Phi}(4\sqrt{\log N})(1 + o(1)) \\
&= O\left(\frac{1}{N^4}\right)
\end{aligned}$$

using the same method as in (88), from which the result follows. ■

B.8.3 Proof of Lemma 2

Proof of Lemma 2 The idea of the proof will be to relate it to the case where $S_1, \dots, S_N$ are exactly normally distributed, which was proved in Section B.7.1. Recall that we have $S_1, \dots, S_N$ with cdf G_d and $S^{(1)} = \min_{1 \leq i \leq N} S_i$. Also, let $\tilde{S}_1, \dots, \tilde{S}_N$ be auxiliary i.i.d. standard Gaussian random variables, and set $\tilde{S}^{(1)} = \min_{1 \leq i \leq N} S_i$.

By the assumption that the $\xi_{i,j}$ are absolutely continuous with respect to the Lebesgue measure, G_d is continuous and hence we can construct $S_1, \dots, S_N$ and $\tilde{S}_1, \dots, \tilde{S}_N$ on a common probability space by drawing N uniform random variables $U_1, \dots, U_N \sim U[0,1]$ and setting $S_i = G_d^{-1}(U_i)$, $\tilde{S}_i = \Phi^{-1}(U_i)$. We then have that $S^{(1)} = G_d^{-1}(U^{(1)})$ and $\tilde{S}^{(1)} = \Phi^{-1}(U^{(1)})$.

From Lemma 1 we know that

$$\mathbb{E}(\tilde{S}^{(1)}) = -\sqrt{2 \log N} + O\left(\frac{\log \log N}{\sqrt{\log N}}\right)$$

so it suffices to prove that

$$\mathbb{E}(|\tilde{S}^{(1)} - S^{(1)}|) = O\left(\frac{\log \log N}{\sqrt{\log N}}\right). \tag{91}$$

Letting $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$, we will split the above expectation according to whether $\tilde{S}^{(1)} \in I_N$.

- Assuming first that $\tilde{S}^{(1)} \in I_N$, so that $\tilde{S}^{(1)} = o(d^{1/6})$, if we let $h \in \mathbb{R}$ be an arbitrary real satisfying $h = o_{N,d}(1)$, then using Corollary 1 we can write

$$\begin{aligned}
G_d(\tilde{S}^{(1)} + h) &= \Phi(\tilde{S}^{(1)} + h)(1 + o_{N,d}(1)) \\
&= -\frac{\phi(\tilde{S}^{(1)} + h)}{\tilde{S}^{(1)} + h}(1 + o_{N,d}(1)) \\
&= \Phi(\tilde{S}^{(1)}) \frac{\phi(\tilde{S}^{(1)} + h)}{\phi(\tilde{S}^{(1)})}(1 + o_{N,d}(1)) \\
&= U^{(1)} \exp\left\{-h\tilde{S}^{(1)} - h^2/2\right\}(1 + o_{N,d}(1))
\end{aligned}$$

  and so it follows by the continuity of G_d that there is a choice of h, satisfying $h = O_{N,d}(1/\sqrt{\log N})$, such that $G_d(\tilde{S}^{(1)} + h) = U^{(1)}$. We conclude that

$$|\tilde{S}^{(1)} - S^{(1)}| \leq O_{N,d}\left(\frac{1}{\sqrt{\log N}}\right)$$

  and so

$$\mathbb{E}\left(|\tilde{S}^{(1)} - S^{(1)}| \mathbb{1}_{\{\tilde{S}^{(1)} \in I_N\}}\right) \leq O\left(\frac{1}{\sqrt{\log N}}\right). \tag{92}$$

- On the other hand, we may also write

$$\begin{aligned}
\mathbb{E}\left(|S_1| \mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) &\leq \mathbb{E}\left(|S_1| \mathbb{1}_{\{|S_1| \geq N^2\}}\right) + \mathbb{E}\left(|S_1| \mathbb{1}_{\{|S_1| < N^2\} \cap \{\tilde{S}^{(1)} \notin I_N\}}\right) \\
&\leq \frac{1}{N^2}\mathbb{E}(|S_1|^2) + N^2 \mathbb{P}\left(\tilde{S}^{(1)} \notin I_N\right) \\
&\leq O\left(\frac{1}{N^2}\right)
\end{aligned}$$

where we have used $\mathbb{E}(|S_1|^2) = 1$ and Lemma 6 to bound the second term.

The same result also holds with $\tilde{S}_1$ in place of S_1 (e.g. by considering taking the ξ_i to be i.i.d. Gaussians), and so we see that

$$\mathbb{E}\left(|\tilde{S}^{(1)} - S^{(1)}|\mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) \leq \sum_{i=1}^{N} \mathbb{E}\left(|\tilde{S}_i - S_i|\mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) = O\left(\frac{1}{N}\right) \tag{93}$$

Combining (92) and (93) yields (91) and the proof is concluded. ■

B.8.4 Proof of Proposition 6

Proof of Proposition 6 First, note that since the weights satisfy (37), we may write

$$\log \overline{w}^{(N)} = -\log\left(\mathbb{E}(\exp(-\sigma\sqrt{d}S_1))\right) - \sigma\sqrt{d}S^{(1)}.$$

In addition, using the definition of S_1 written in (38), that is

$$S_1 = \frac{1}{\sigma\sqrt{d}}\sum_{j=1}^{d}\xi_{1,j},$$

where the $\xi_{1,1}, \ldots, \xi_{1,d}$ are i.i.d. random variables, we have that

$$\begin{aligned}\mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) &= \prod_{j=1}^{d}\mathbb{E}(\exp(-\xi_{1,j})) \\ &= \left(\mathbb{E}(\exp(-\xi_{1,1}))\right)^{d}.\end{aligned}$$

Thus,

$$-\log\left(\mathbb{E}(\exp(-\sigma\sqrt{d}S_1))\right) = -d\log\mathbb{E}(\exp(-\xi_{1,1})) = -da \tag{94}$$

By Jensen's inequality applied to the strictly convex function $u \mapsto -\log(u)$, we have that

$$a < \mathbb{E}(\xi_{1,1}) = 0.$$

Hence,

$$\log \overline{w}^{(N)} = -da - \sigma\sqrt{d}S^{(1)} \tag{95}$$

with $a > 0$. Following the proof of Proposition 4 in Appendix B.7.2, we can then conclude by combining (95) with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi)$ in (29). Indeed,

$$\Delta_{N,d}^{MAX}(\theta, \phi; x) = -da - \sigma\sqrt{d}\mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}$$

and using (34), we deduce:

$$\begin{aligned}\Delta_{N,d}^{MAX}(\theta,\phi;x) &= -da + \sigma\sqrt{d}\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha - 1} \\ &= -da\left\{1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + \frac{1}{1-\alpha}\frac{\log N}{da} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\},\end{aligned}$$

Hence, using now that $\frac{1}{1-\alpha}\frac{\log N}{da} = O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)$ under the assumption $\log N/d^{1/3} \to 0$, we can deduce

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi) + da\left\{1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\} = 0,$$

which yields the desired result. ■

B.8.5 Proof of Proposition 7

Proof of Proposition 7 The proof will build on the proof of Proposition 5. As in that proof, we denote $\sigma_\alpha = (1-\alpha)\sigma$ for all $\alpha \in [0,1)$ and observe that, conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N-1$ i.i.d. random variables

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S - S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{g_d(z)}{\overline{G_d}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with g_d denoting the pdf of S_1 and $\overline{G_d}(x) = \int_x^\infty g_d(z)\mathrm{d}z$ for all $x \in \mathbb{R}$, that is

$$\mathbb{E}(T_{N,d}|S^{(1)}) = (N-1)\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}.$$

We are thus required to show that

$$(N-1)\mathbb{E}\left[\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right] \to 0 \tag{96}$$

as $N, d \to 0$ with $\log N/d^{1/3} \to 0$. First, we show that contributions due to extreme values of $S^{(1)}$ are negligible. To see this, note that

$$\left|\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right| \leq 1,$$

so that Lemma 9 implies

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\notin I_N\}}\frac{\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right]\leq(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\notin I_N\}}\right)\to 0.$$

Hence it suffices to show that

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\frac{\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right]\to 0.$$

Note that by Corollary 1, we have $\overline{G}_d(S^{(1)})\geq\overline{G}_d(0)=1-\Phi(0)(1+o(1))$ on the event $\{S^{(1)}\in I_N\}$ as $N,d\to\infty$ with $\log N/d^{1/3}\to 0$, so $\overline{G}_d(S^{(1)})$ is uniformly bounded below. It thus suffices to prove

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z\right]\to 0.$$

We will in fact show that

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right]\to 0 \tag{97}$$

and

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\left|\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\left(g_d(z)-\phi(z)\right)\mathrm{d}z\right|\right]\to 0. \tag{98}$$

- **Proof of** (97)**.** Following the proof of Proposition 5, we see that (89) holds whenever $S^{(1)}\in I_N$. Restricting to the event $\{S^{(1)}\in I_N\}$ and taking expectations over $S^{(1)}$, we get

  $$\begin{aligned}&(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right]\\ &\leq(N-1)\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\Phi(S^{(1)})\right](1+o(1)).\end{aligned}$$

  Since $S^{(1)}=o(d^{1/6})$, Corollary 1 implies that

  $$\Phi(S^{(1)})=G_d(S^{(1)})(1+o_{N,d}(1)),$$

  where the uniformity over x in the statement of the theorem implies that the implicit constant is independent of $S^{(1)}$. Restricting to $\{S^{(1)}\in I_N\}$ and taking expectations again, noting that $G_d(S^{(1)})$ is distributed as the minimum of N uniform random variables on $[0,1]$, we see

  $$\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\Phi(S^{(1)})\right]=\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}G_d(S^{(1)})\right](1+o(1))\leq\frac{1}{N+1}(1+o(1)). \tag{99}$$

  We conclude that

  $$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right]\leq\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o(1))\to 0,$$

  proving (97).

- **Proof of** (98)**.** Two applications of integration by parts give

$$\int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}g_d(z)\mathrm{d}z = -G_d(S^{(1)}) \\ + \int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}G_d(z)\mathrm{d}z$$

and

$$\int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}\phi(z)\mathrm{d}z = -\Phi(S^{(1)}) \\ + \int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}\Phi(z)\mathrm{d}z.$$

It follows that

$$\left|\int_{S^{(1)}}^{\infty} \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))\,(g_d(z) - \phi(z))\,\mathrm{d}z\right| \\ \leq \left|\Phi(S^{(1)}) - G_d(S^{(1)})\right| + \left|\int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}(G_d(z) - \Phi(z))\mathrm{d}z\right|.$$

We now deal with each of these terms separately. On the event $\{S^{(1)} \in I_N\}$, we know $\Phi(S^{(1)}) = G_d(S^{(1)})(1 + o_{N,d}(1))$ so we have

$$\left|\Phi(S^{(1)}) - G_d(S^{(1)})\right| \leq o_{N,d}(1)G_d(S^{(1)})$$

and so by restricting to $\{S^{(1)} \in I_N\}$ and taking expectations

$$(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}\left|\Phi(S^{(1)}) - G_d(S^{(1)})\right|\right) \leq (N-1)\cdot o(1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}G_d(S^{(1)})\right) \to 0.$$

along the same lines as (99).

We split the second term as an integral from $S^{(1)}$ to 0 and an integral from 0 to ∞ and we write:

$$\left|\int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}(G_d(z) - \Phi(z))\mathrm{d}z\right| \leq A_1 + A_2$$

with

$$A_1 = \int_{S^{(1)}}^{0} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}\left|G_d(z) - \Phi(z)\right|\mathrm{d}z \\ A_2 = \int_{0}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z - S^{(1)})\}\mathrm{d}z.$$

Again, we bound each term individually. For A_2, assuming that $S^{(1)} \in I_N$ we have the bound

$$A_2 = \exp\{\sigma_\alpha\sqrt{d}S^{(1)}\} \leq \exp\{-\sigma_\alpha\sqrt{d\log N}\} \leq \exp\{-\sigma_\alpha(\log N)^2\}$$

for sufficiently large N, d with $\log N/d^{1/3} \to \infty$ and so $(N-1)\mathbb{E}(\mathbb{1}_{\{S^{(1)} \in I_N\}} A_2) \to 0$. To bound A_1, note that by Corollary 1

$$|G_d(z) - \Phi(z)| \leq o_{N,d}(1)\Phi(z)$$

for all $z \in [S^{(1)}, 0]$ so long as $S^{(1)} \in I_N$, and hence under this assumption we can write

$$A_1 \leq o_{N,d}(1) \int_{S^{(1)}}^{0} \sigma_\alpha \sqrt{d} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\Phi(z)\mathrm{d}z.$$

Changing the upper limit from 0 to ∞, which can only weaken the bound, and then integrating by parts gives

$$\begin{aligned} &\leq o_{N,d}(1) \int_{S^{(1)}}^{\infty} \sigma_\alpha \sqrt{d} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\Phi(z)\mathrm{d}z \\ &\leq o_{N,d}(1) \left\{ \Phi(S^{(1)}) + \int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\phi(z)\mathrm{d}z \right\}. \end{aligned}$$

We conclude that

$$\begin{aligned} (N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} A_1\right) \leq o(1)\Big\{ &(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} \Phi(S^{(1)})\right) \\ &+ (N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} \int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\phi(z)\mathrm{d}z\right) \Big\} \end{aligned}$$

The former term tends to zero by (99) and the latter tends to zero by (97). We thus see that (98) holds, completing the proof.

■

B.8.6 Proof of Example 4

Proof of Example 4 Recall that by (82): for all $i = 1 \dots N$,

$$\log \overline{w}_i = \frac{d}{2} \log\left(\frac{4}{3}\right) - \left\| z_i - \frac{\theta + x}{2} \right\|^2 + \frac{3}{4}\|z_i - Ax - b\|^2.$$

We want to show that if $z_i \sim q_\phi(\cdot|x) = \mathcal{N}(Ax + b, 2/3\ \boldsymbol{I}_d)$, then $\log \overline{w}_i$ can be written in the form of (37). For this purpose, denote $\mathbf{1} = (1, \dots, 1)^T$ and observe that there exists an orthogonal matrix U such that $U\left(\frac{\theta+x}{2} - Ax - b\right) = \lambda \mathbf{1}$. We can then sample $z_i \sim \mathcal{N}(Ax + b, 2/3\ \boldsymbol{I}_d)$ by setting $z_i = U^{-1}y_i + Ax + b$ where $y_i \sim \mathcal{N}(0, 2/3\ \boldsymbol{I}_d)$. With this parameterization, (82) becomes

$$\begin{aligned} \log \overline{w}_i &= -\left\| U^{-1}y_i + Ax + b - \frac{\theta + x}{2} \right\|^2 + \frac{3}{4}\|U^{-1}y_i\|^2 + const. \\ &= -\|y_i - \lambda \mathbf{1}\|^2 + \frac{3}{4}\|y_i\|^2 + const. \\ &= -\sum_{j=1}^{d} \left\{ (y_{ij} - \lambda)^2 - \frac{3}{4} y_{ij}^2 \right\} + const. \end{aligned}$$

where *const.* denotes a fixed constant which depends only on d, θ, A, b and x.

Let us now set $\zeta_{ij} = (y_{ij} - \lambda)^2 - 3y_{ij}^2/4$ and $\xi_{i,j} = \zeta_{ij} - \mathbb{E}(\zeta_{ij})$. Since $y_i \sim \mathcal{N}(0, 2/3\boldsymbol{I}_d)$ it follows that $\xi_{i,1}, \ldots, \xi_{i,d}$ are i.i.d. random variables with $\mathbb{E}(\xi_{i,j}) = 0$. Now defining $\sigma^2 = \mathbb{V}(\xi_{1,1}) < \infty$, we have that σ^2 can be computed analytically by observing that

$$\mathbb{E}(\zeta_{ij}) = \mathbb{E}\left((y_{ij} - \lambda)^2\right) - \mathbb{E}\left(\frac{3}{4}y_{ij}^2\right) = \frac{1}{4}\mathbb{E}(y_{ij}^2) + \lambda^2 = \frac{1}{6} + \lambda^2$$

from which we can deduce that

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left([\zeta_{ij} - \mathbb{E}(\zeta_{ij})]^2\right) = \mathbb{E}\left(\left[\frac{1}{4}y_{ij}^2 - 2\lambda y_{ij} - \frac{1}{6}\right]^2\right)\\ &= \frac{1}{16}\mathbb{E}(y_{ij}^4) + \left(4\lambda^2 - \frac{1}{12}\right)\mathbb{E}(y_{ij}^2) + \frac{1}{36}\\ &= \frac{1}{18} + \frac{8}{3}\lambda^2 < \infty.\end{aligned}$$

Hence, (37) holds by defining S_i as in (38) (noting that the constant terms must match since $\overline{w}_i$ is normalised and so has expected value 1). In addition, we can also analytically compute the quantity a defined in (39) by noting that

$$\begin{aligned}\mathbb{E}\left(\exp\left(-\xi_{1,1}\right)\right) &= \int_{-\infty}^{\infty} \exp\left(-\left[\frac{1}{4}u^2 - 2\lambda u - \frac{1}{6}\right]\right) \cdot \frac{1}{\sqrt{2\pi \cdot \frac{2}{3}}} e^{-\frac{3}{4}u^2}\, \mathrm{d}u\\ &= \sqrt{\frac{3}{4}} \exp\left(\lambda^2 + \frac{1}{6}\right)\end{aligned}$$

so that

$$a = \lambda^2 + \frac{1}{6} + \frac{1}{2}\log\left(\frac{3}{4}\right).$$

Finally, we check that (A2) holds in the case where $(\theta, \phi) = (\theta^\star, \phi^\star)$. For this choice of parameters, $\lambda = 0$, hence σ is independent of d. Furthermore, $\xi_{i,j}$ is clearly absolutely continuous with respect to the Lebesgue measure, and the distribution of $\xi_{i,j}$ is independent of d. It follows that (A2)a holds using our previous observations.

To now check (A2)b, we let $k \geq 3$. In that case, $u \mapsto |u|^k$ is convex and for all real-valued u_1, u_2 and u_3, we have that:

$$\begin{aligned}\left|\frac{1}{3}u_1 + \frac{1}{3}u_2 + \frac{1}{3}u_3\right|^k &\leq \left|\frac{1}{3}|u_1| + \frac{1}{3}|u_2| + \frac{1}{3}|u_3|\right|^k\\ &\leq \frac{1}{3}\left(|u_1|^k + |u_2|^k + |u_3|^k\right)\end{aligned}$$

so that, setting $u_1 = (y_{ij} - \lambda)^2$, $u_2 = -\frac{3}{4}y_{ij}^2$ and $u_3 = -\mathbb{E}(\zeta_{ij})$, it holds that

$$\left|(y_{ij} - \lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}(\zeta_{ij})\right|^k \leq 3^{k-1}\left((y_{ij} - \lambda)^{2k} + \left(\frac{3}{4}y_{ij}\right)^{2k} + |\mathbb{E}(\zeta_{ij})|^k\right).$$

Using a similar argument applied to $(y_{ij} - \lambda)^{2k}$, we then deduce that

$$\left|(y_{ij} - \lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}(\zeta_{ij})\right|^k \leq 3^{k-1}\left(2^{2k-1}\left(y_{ij}^{2k} + \lambda^{2k}\right) + \left(\frac{3}{4}y_{ij}\right)^{2k} + |\mathbb{E}(\zeta_{ij})|^k\right).$$

Hence,

$$\begin{aligned}
\mathbb{E}\left(|\xi_{i,1}|^k\right) &= \mathbb{E}\left(\left|(y_{ij}-\lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left(2^{2k-1}\left(\mathbb{E}(y_{ij}^{2k}) + \lambda^{2k}\right) + \frac{3^{2k}}{4^{2k}}\mathbb{E}(y_{ij}^{2k}) + \left|\mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left((2^{2k-1} + \frac{3^{2k}}{4^{2k}}) \cdot (2k-1)!! \cdot \frac{2^k}{3^k} + 2^{2k-1}\lambda^{2k} + \left|\mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left((2^{2k-1} + \frac{3^{2k}}{4^{2k}}) \cdot \frac{2^{2k}}{3^k} \cdot k! + 2^{2k-1}\lambda^{2k} + \left(\frac{1}{6} + \lambda^2\right)^k\right)
\end{aligned}$$

where we have used that the $(2k)$-th moment of a standard Gaussian random variable is $(2k-1)!!$. Finally, since $\lambda = 0$ in our case, we obtain that

$$\mathbb{E}\left(|\xi_{i,1}|^k\right) \leq k! K^{k-2}\sigma^2$$

for some sufficiently large choice of K which is independent of d, and (A2) thus holds. ■

Remark 4 *We have obtained that* (A2) *holds when (θ, ϕ) are equal to the optimal parameters. If we now assume that x, θ and ϕ are initially drawn from Gaussian distributions with bounded covariance matrices (like it is the case in Rainforth et al., 2018), we anticipate that* (A2) *should approximately hold even for values of the parameters other than the optimal choice. Notice indeed that in that case we inuitively expect $\|\frac{\theta+x}{2} - Ax - b\| = O(\sqrt{d})$ for most values of x, θ and ϕ. It follows that we should expect $\lambda = O(1)$ and $\sigma = \Theta(1)$ in practice as $d \to \infty$, and so* (A2) *should approximately hold.*

Appendix C. Futher Details Regarding Related Proof Techniques

As mentioned in Section 5, a number of our proof techniques differs significantly from/alter parts of known proofs, which in some cases impacts the corresponding theoretical results. Namely,

- **Theorem 1**. The proof of this result is based on the proof for the case $\alpha = 0$ written in the arxiv version of 5 Mar 2019 of (Rainforth et al., 2018, Theorem 1), which was the latest version available to us. Nevertheless, and contrary to Rainforth et al. (2018), we (i) use an explicit form for the remainder term in Taylor's theorem rather than the mean value form of the remainder, allowing us to get more precise control on the magnitude of the remainder and its gradients, and (ii) we consequently rely on Lemma 3, which is a non-immediate extension of (Rainforth et al., 2018, Lemma 1).

  This significantly impacts the proof technique and as a result, the main difference in terms of assumptions compared to (Rainforth et al., 2018, Theorem 1) is that, for a given $\alpha \in [0, 1)$, we are requiring the eighth moments of $\tilde{w}_{1,1}^{1-\alpha}$, $\partial\tilde{w}_{1,1}^{1-\alpha}/\partial\theta_\ell$ and $\partial\tilde{w}_{1,1}^{1-\alpha}/\partial\phi_{\ell'}$ to be finite in Theorem 1, where (Rainforth et al., 2018, Theorem 1) asked for the fourth moments to be finite with $\alpha = 0$. In addition, we need to further assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{1,N,\alpha})^4) < \infty$.

- **Proposition 5**. The proof of this result mostly mirrors the proof written in (Snyder et al., 2008, Section 4.a) which considers the case $\alpha = 0$ and is used in the context of particle filtering. The main difference is that we require an additional lemma (Lemma 6 of Appendix B.7.3) to provide us with a more precise concentration result for $S^{(1)}$. This lemma will also have other uses in the rest of the paper. This does not result in any change of assumptions compared to (Snyder et al., 2008, Section 4.a).

- **Proposition 7**. The proof of this result is arguably the one that required the most alterations out of the three discussed here. It borrows some ideas from the proofs written in the context of particle filtering in Li et al. (2005); Bengtsson et al. (2008); Li et al. (2005), which all aimed at establishing that $\mathbb{E}(T_{N,d}^{(0)}) \to 0$ in the approximate log-normal case under some conditions on N and d. However, we are significantly more thorough in our control of the error terms, which we bound precisely with the aid of two results from Saulis and Statulevičius (2000) and Lemma 9, rather than simply working under convergence in probability.

  In terms of assumptions, it is closest to Li et al. (2005), except for the fact that our result relies on a Bernstein condition while Li et al. (2005) uses Cramer's conditions. Both are in fact equivalent in the i.i.d. setting, and we choose to use Bernstein condition as we believe it might make it easier to generalize our result beyond the i.i.d. case using the results from Saulis and Statulevičius (2000) recalled in Appendix B.8.1.

Appendix D. Additional Numerical Experiments and Derivation Details

D.1 Gaussian Example from Section 6.1

Figure 13 empirically confirms that the asymptotic regime predicted by Theorem 3 does not reflect what is happening in reality in the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$ when the dimension d increases, N is small and the distribution of the weight is log-normal.

Note that we only plotted the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$ for the cases $d = \{10, 100\}$ in the figure above. This is due to the fact that when $d = 1000$, computing γ_α^2 the $1/N$ term returns an overflow, further illustrating the limitations of the approach from Theorem 3 in the specific setting considered here. Note also that since the variance term is exponential in $(1-\alpha)^2 d$, increasing α does play a role in decreasing γ_α^2 so that the asymptotic regime predicted by Theorem 3 applies in lower dimensions (e.g. $d = 10$ with $\alpha = 0.5$).

D.2 Linear Gaussian Example from Section 6.2

D.2.1 Empirical Experiments for Theorem 3 in the Context of Section 6.2

Figure 15 empirically confirms that we need an unpractical amount of samples N for the asymptotic regime predicted by Theorem 3 to capture the behavior of the variational gap as d increases when $\sigma_{\text{perturb}} = 0$. Note that similar plots and conclusions can be obtained for $\sigma_{\text{perturb}} \in \{0.01, 0.5\}$. Those are not given here for the sake of conciseness.

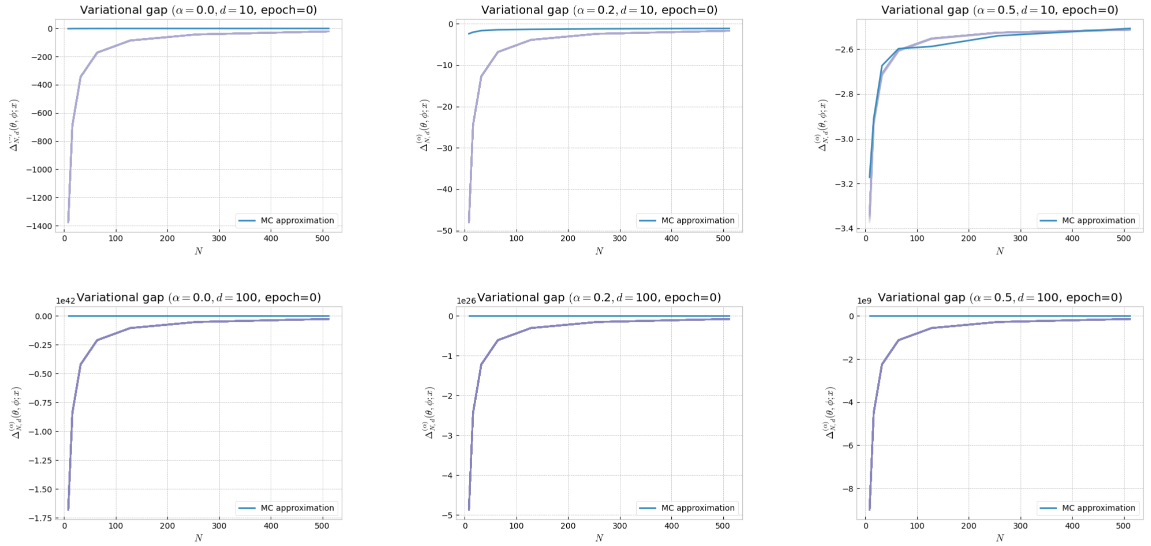

Figure 13: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the toy example described in Section 6.1 as a function of N, for varying values of (α, d) and with $(\theta, \phi) = (0 \cdot \boldsymbol{u_d}, \boldsymbol{u}_d)$ so that $B_d = \sqrt{d}$. Plotted in purple are curves of the form (42) with tailored values of c_1.

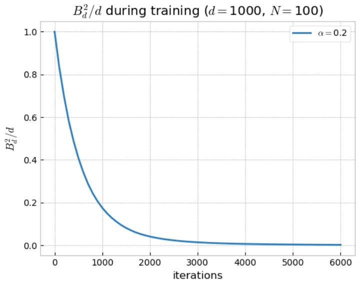

Figure 14: Evolution of B_d^2/d during the training of the ϕ parameter for the toy example described in Section 6.1.

D.2.2 Additional Experimental Results for Section 6.2

We provide some additional results in the context of Section 6.2 regarding the signal-to-noise ratio (SNR) in the doubly-reparameterized case and the Mean Squared Error (MSE) for the VR-IWAE bound and its θ, ϕ gradients.

- **SNR in the doubly-reparameterized case.** In line with Tucker et al. (2019), we observe in Figure 16 that using the doubly-reparameterized gradient estimator for ϕ increases the SNR when $\alpha = 0$. We in fact see that the SNR is increased for all values

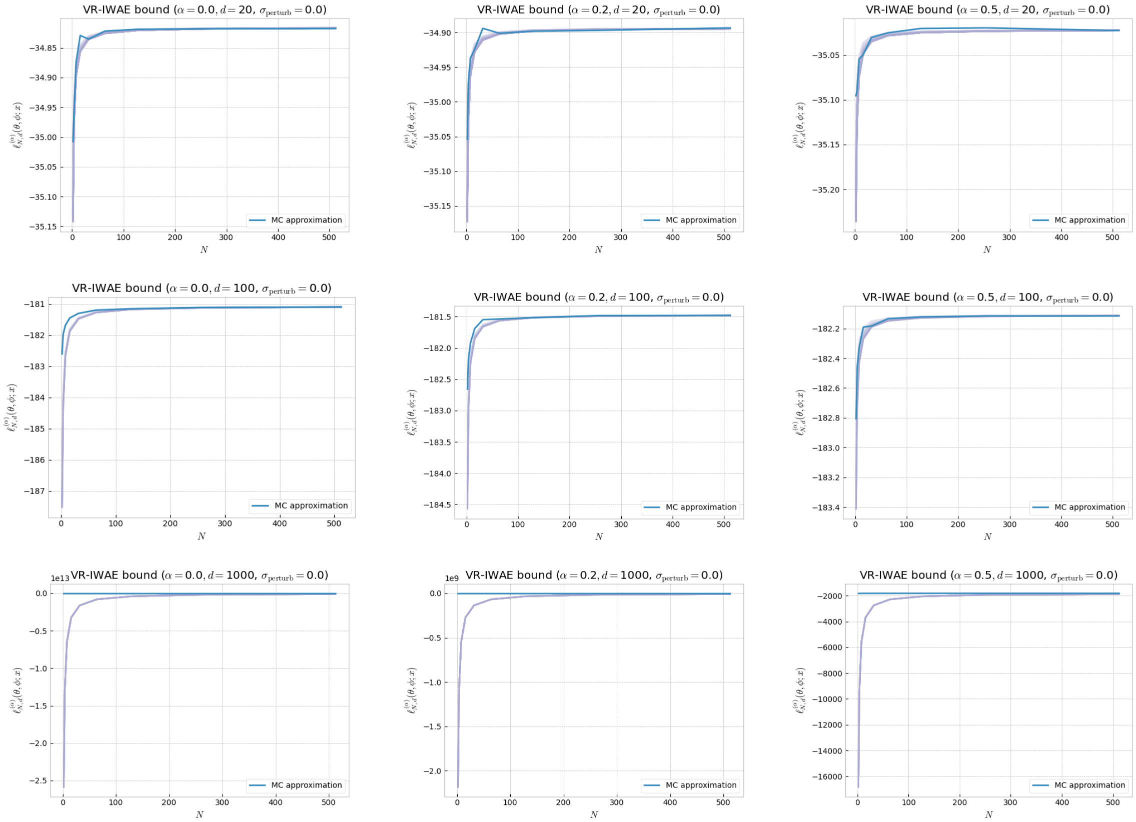

Figure 15: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d). Plotted in purple are curves of the form (45) with tailored values of c_1.

of α, extending the conclusions from Tucker et al. (2019) to $\alpha \in [0, 1)$ in the example considered here.

However, as we get further away from the optimum ($\sigma_{\text{perturb}} = 0.5$) and/or increase the dimension ($d = 1000$), we observe that it still remains challenging to obtain an increasing SNR for the ϕ gradients for small values of α, even when using doubly-reparameterized gradient estimators.

- **MSE for the VR-IWAE bound and its θ, ϕ gradients.** We observe on Figures 17 and 18 that while increasing α does not lower the MSE of the VR-IWAE estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(Z_j; x)^{1-\alpha} \right)$$

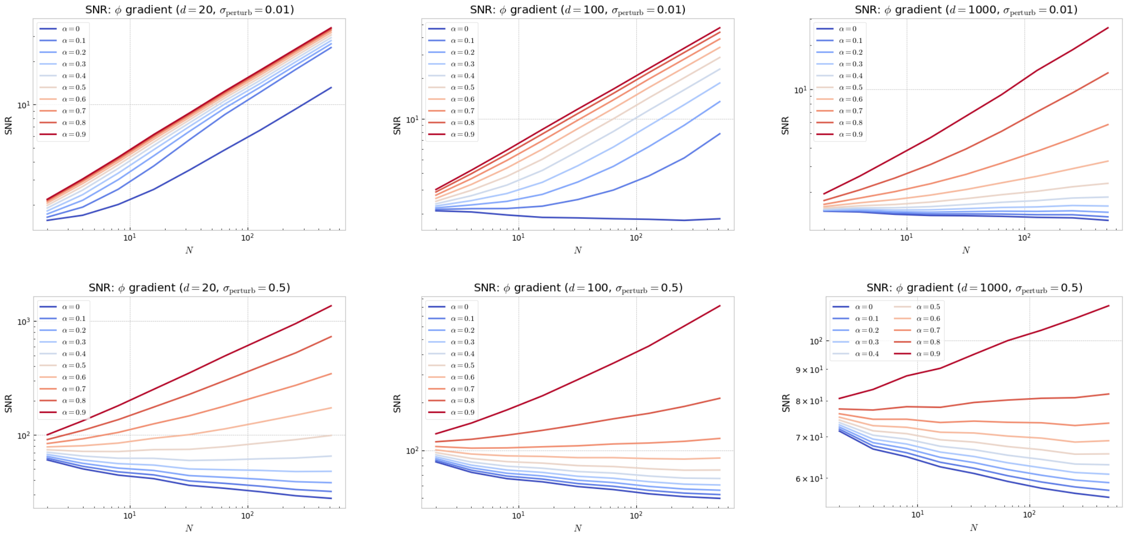

Figure 16: Plotted is the SNR of the inference network (ϕ) gradients in the doubly-reparameterized case (computed over 1000 MC samples) for the linear Gaussian example in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

for log-likelihood estimation, it can be useful in lowering the MSE of its θ gradients

$$\frac{1}{1-\alpha}\nabla_\theta \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(Z_j;x)^{1-\alpha}\right)$$

compared to the θ gradients of the true log-likelihood $\nabla_\theta \ell_d(\theta;x)$.

In the low perturbation regime ($\sigma_{\text{perturb}} = 0.01$) and in medium to high dimensions ($d = 100, 1000$), we indeed see in Figure 17 that every tested value of $\alpha > 0$ achieves lower θ gradient MSE than $\alpha = 0$ for low values of N. As we increase to $N = 2^9$, the value of α achieving the lowest MSE is $\alpha = 0.3$ for $d = 100$, and $\alpha = 0.8$ for $d = 1000$. This sheds light on a bias-variance tradeoff between low bias at $\alpha = 0$ and low variance at $\alpha = 1$, and is in line with the findings of Theorem 3.

In the high perturbation regime ($\sigma_{\text{perturb}} = 0.5$), we see in Figure 17 that the choice of α appears to make less of a difference, especially when the dimension d is high. This suggests that bias reduction may be more important when the inference distribution $q_\phi(z|x)$ is far from the optimum.

D.3 Variational Auto-encoder from Section 6.3

We present additional results for the VAE example discussed in Section 6.3.

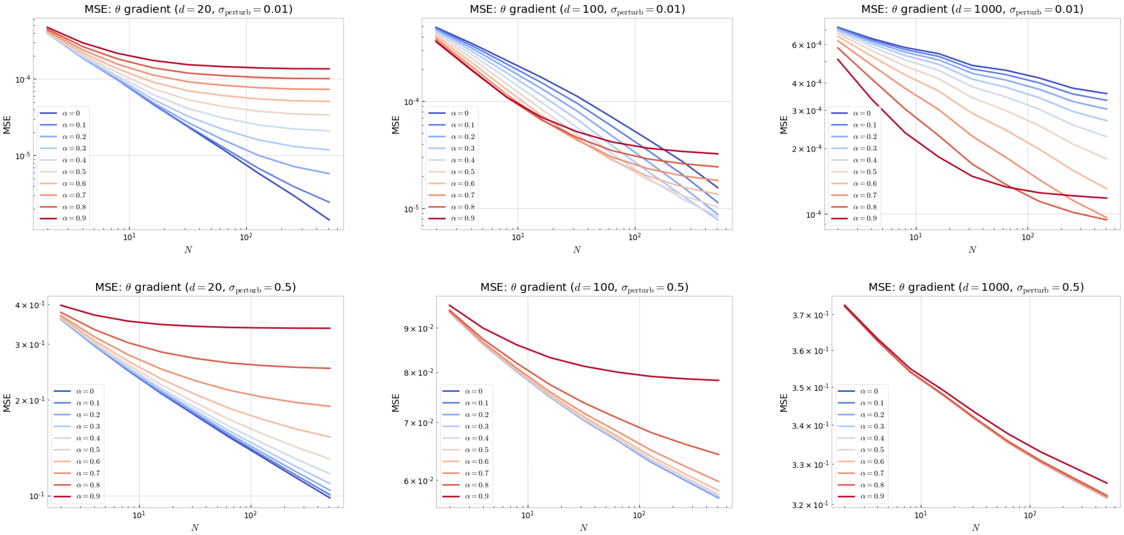

Figure 17: Plotted is the MSE of the generative network (θ) gradients (computed over 1000 MC samples) compared to the log-likelihood gradients for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

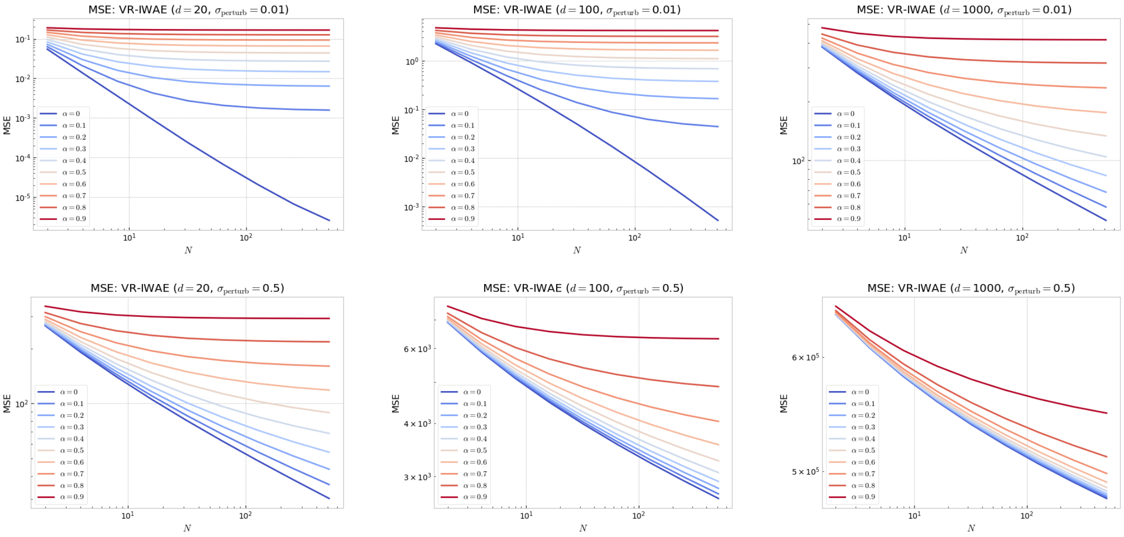

Figure 18: Plotted is the MSE of the VR-IWAE estimate (computed over 1000 MC samples) compared to the log-likelihood gradients for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

D.3.1 Complementary Plots for the VR-IWAE Bound

Figures 19 and 20 provide additional plots to Figures 8 and 9 reinforcing the conclusions drawn in Section 6.3.

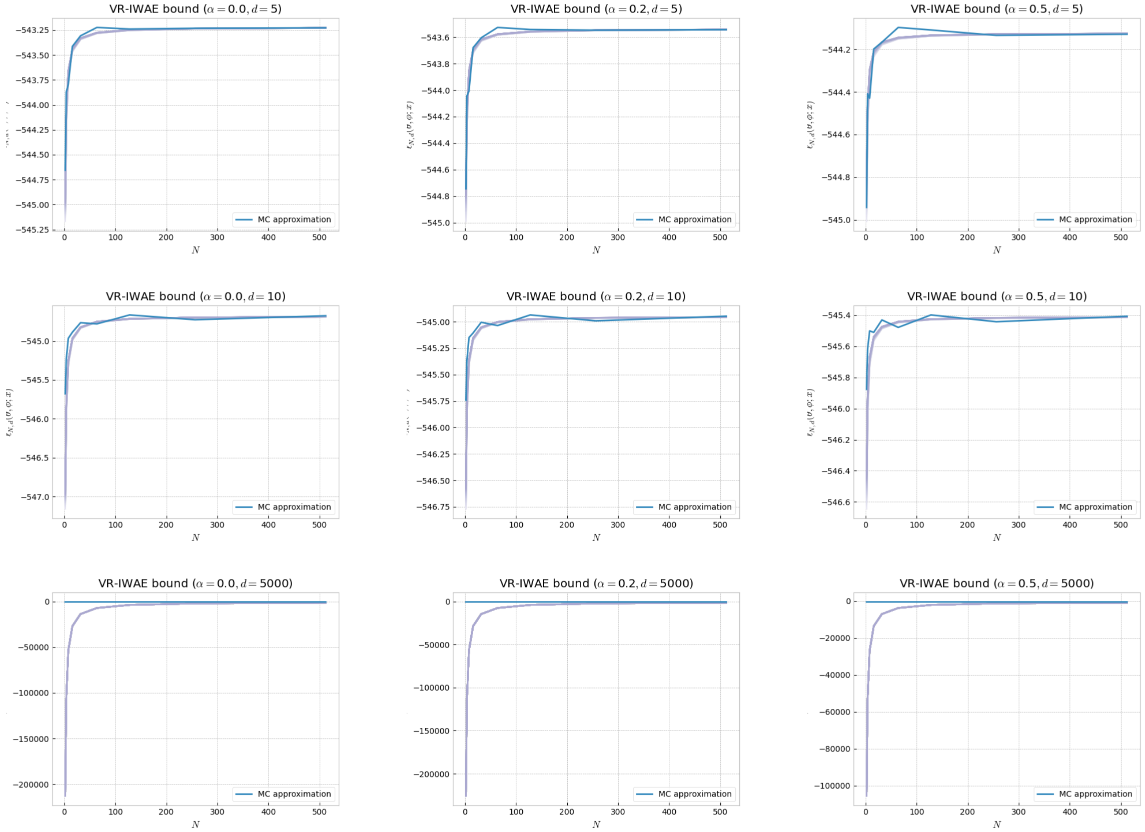

Figure 19: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the testing set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in purple are curves of the form (49) with tailored values of c_1.

D.3.2 Impact of α and of M on Empirical Performances

We discuss here the impact of α and of M on the empirical performances of the VR-IWAE bound methodology in the reparameterized and doubly-reparameterized cases.

- **Impact of α on the empirical performances.** We investigate how the choice of α impacts the Negative Log Likelihood (NLL) after training the VAE with the VR-IWAE

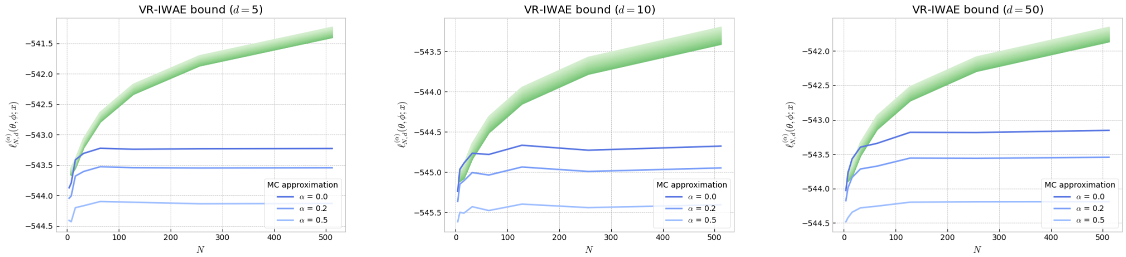

Figure 20: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the testing set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in green are curves of the form (50) with tailored values of c_2.

bound. The NLL can indeed be used to evaluate the empirical performances of VAEs (since a lower NLL corresponds to a higher likelihood of the data under the VAE model, which indicates better training of the generative network θ). Furthermore, although the NLL is intractable, following Burda et al. (2016) it can be approximated using the negative IWAE bound with $N = 5000$.

We plot in Figure 21 the NLL estimate on the MNIST test set as a function of α after training VAEs on the MNIST training set using either the reparameterized ("rep") or the doubly-reparameterized ("drep") gradient estimators of the VR-IWAE objective with $N = 10, 100$ and $d = 50$. Here, all the models are trained for 1000 epochs using the Adam optimizer with learning rate $1e-3$ and batch size 100.

We observe that the doubly-reparameterized gradient estimator generally achieves better NLL results than the reparameterized one when α is fixed. In addition, for both cases the value of α achieving the best NLL performance lies in the middle of $(0, 1)$, around $\alpha = 0.5$. In line with Theorem 3, this suggests that there is a bias-variance tradeoff to consider when choosing α, and that the best setting can lie between the standard IWAE ($\alpha = 0$, low bias) and ELBO ($\alpha = 1$, low variance) objectives, with the optimal choice of α being dependent on the data set, model architecture, as well as the stochastic gradient descent procedure used for training.

- **Impact of M on the empirical performances.** We investigate how the choice of M and N affects the training of VAE when $M \times N$ is fixed in the VR-IWAE bound methodology. We plot in Figure 22 the NLL on the MNIST test set after training the VAE on the MNIST training set for 1000 epochs, with $M \times N = 100$, $d = 50$ and $\alpha \in \{0, 0.2\}$.

  We observe a bias-variance tradeoff that is similar to the analysis above for the impact of α. Indeed, the cases $M = 100$ and $M = 1$ have a particular meaning in the plots of Figure 22: $M = 100$ corresponds to the ELBO with maximum computational

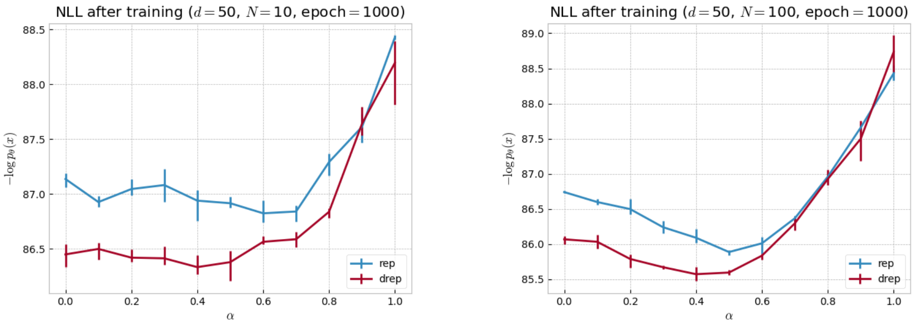

Figure 21: Plotted is the Negative Log Likelihood (NLL) estimate on the test set of the MNIST data set as described in Section 6.3 as a function of α, after training on the train set for 1000 epochs with $N \in \{10, 100\}$. The error bars are computed over 3 trials with different network initialisations and seeds during training.

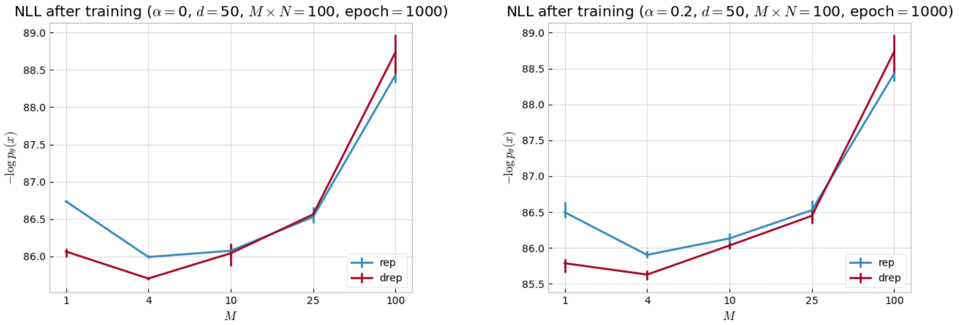

Figure 22: Plotted is the Negative Log Likelihood (NLL) estimate on the test set of the MNIST data set as described in Section 6.3 as a function of M while fixing $M \times N = 100$, after training on the train set for 1000 epochs with $\alpha = 0, 0.2$. The error bars are computed over 3 trials with different network initialisations and seeds during training.

budget for M (i.e. $\alpha = 1$, low variance), while $M = 1$ corresponds to the VR-IWAE bound with maximum computational budget for N (i.e. low bias, with the lowest bias being achieved for $\alpha = 0$). Here, the best value of test NLL is obtained for $\alpha = 0.2, M = 4, N = 25$ among our tested combinations. Note that one potential advantage of tuning α instead of M and N is that α resides on a one-dimensional continuous interval, whereas M and N are integer values and so their choice is more limited (that is, they can be more difficult to tune for a given computational budget).

References

Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, 2008.

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

Thang D. Bui, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Yingzhen Li. Black-box α-divergence for deep generative models. In *NIPS Workshop on Approximate inference*, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*, 2016.

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.

Kamélia Daudel, Randal Douc, and François Roueff. Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.

Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250 – 2270, 2021. doi: 10.1214/20-AOS2035.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007. ISBN 9780387344713.

Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael Riis Andersen, Jonathan H Huggins, and Aki Vehtari. Challenges and opportunities in high-dimensional variational inference. volume 34, pages 7787–7798. Neural information processing systems foundation, 3 2021.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ-upper bound minimization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4470–4479. Curran Associates, Inc., 2018.

Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In Marina Meila and Tong Zhang, editors, *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.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pages 1511–1520. PMLR, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Bo Li, Thomas Bengtsson, and Peter Bickel. Curse-of-dimensionality revisited: collapse of importance sampling in very large scale systems. Technical Report 696, Department of Statistics, University of California at Berkeley, 2005.

Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2052–2061. PMLR, 06–11 Aug 2017.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6573–6583. Curran Associates, Inc., 2017.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

Valentin V. Petrov. *Limit Theorems of Probability Theory.* 06 1995. ISBN 9780198534990.

James Pickands III. Moment Convergence of Sample Extremes. *The Annals of Mathematical Statistics*, 39(3):881 – 889, 1968. doi: 10.1214/aoms/1177698320.

James Picklands III. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1 1975. doi: 10.1214/AOS/1176343003.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference* *on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4277–4285. PMLR, 10–15 Jul 2018.

Simón Rodríguez-Santana and Daniel Hernández-Lobato. Adversarial α-divergence minimization for bayesian approximate inference. *Neurocomputing*, 471:260–274, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.09.076.

Leonas Saulis and Vytautas Statulevičius. *Limit Theorems on Large Deviations*, pages 185–266. 01 2000. ISBN 978-3-642-08170-5. doi: 10.1007/978-3-662-04172-7_5.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629 – 4640, 2008. doi: https://doi.org/10.1175/2008MWR2529.1.

George Tucker, Dieterich Lawson, Shixiang Shane Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

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. ISSN 1935-8237. doi: 10.1561/2200000001.

Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ruqi Zhang, Yingzhen Li, Christopher De Sa, Sam Devlin, and Cheng Zhang. Meta-learning divergences for variational inference. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4024–4032. PMLR, 13–15 Apr 2021.