Journal of Machine Learning Research 27 (2026) 1-83 Submitted 9/25; Revised 5/26; Published 10/26

Learning with Importance-weighted Variational Inference

Kamélia Daudel DAUDEL@ESSEC.EDU
ESSEC Business School, France
François Roueff ROUEFF@TELECOM-PARIS.FR
LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Editor: Alp Kucukelbir

Abstract

Several variational bounds involving importance weighting ideas generalize the Evidence Lower BOund (ELBO) for marginal likelihood optimization, such as the Importance-weighted Auto-Encoder (IWAE), Variational Rényi (VR) and VR-IWAE bounds. Yet, it remains unclear how the joint choice of bound and gradient estimator impacts the behavior of the resulting variational inference (VI) algorithms. This paper provides a unified theoretical comparison of reparameterized (REP) and doubly-reparameterized (DREP) gradient estimators tied to the IWAE, VR and VR-IWAE bounds. Through asymptotic analyses of the Signal-to-Noise Ratio as the number of Monte Carlo samples N goes to infinity, we identify a bias-variance tradeoff in these gradient estimators and we formally justify the superiority of DREP over REP in importance-weighted VI. An additional asymptotic analysis for challenging regimes, where both N and the Kullback-Leibler divergence between the variational and posterior densities go to infinity, indicates that importance-weighted VI gradient estimators point in a well-founded direction even when the variational approximation deteriorates. Together, these complementary results characterize the optimization trajectory in importance-weighted VI from poor initialization to final convergence. Importantly, our proof techniques establish general theoretical tools for the study of sample means ratios whose scope extends beyond VI and constitute an independent contribution to the field of Monte Carlo methods.

Keywords: Variational Inference, Importance-weighted Auto-encoders, Alpha-divergence, Reparametrization Trick, Signal-to-Noise Ratio, Importance Sampling, Monte Carlo methods

1. Introduction

Variational inference (VI) methods seek to find the best approximation to an unknown target posterior density within a more tractable family of probability densities $\mathcal{Q}$ (Neal and Hinton, 1998; Jordan et al., 1999; Blei et al., 2017). In the standard Bayesian setting, the model is typically assumed fixed and the aim is to identify the member of $\mathcal{Q}$ that minimizes a certain measure of dissimilarity to the posterior. More generally, VI is applicable to models depending on a parameter θ, in which case the primary goal is to optimize the marginal log-likelihood. In both scenarios, the intractability of the marginal likelihood necessitates the maximization of variational bounds, which involve the variational family $\mathcal{Q}$ and are constructed as surrogate objective functions to the marginal log-likelihood that are amenable to optimization.

While the most traditional variational bound is the Evidence Lower BOund (ELBO), an active line of research in VI is interested in deriving alternative bounds that can improve on the outcome of VI. In particular, popular generalizations of the ELBO relying on importance

©2026 Kamélia Daudel and François Roueff.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2087.html.

weighting ideas have been proposed, such as the importance-weighted auto-encoder (IWAE) bound in Burda et al. (2016) that is parametrized by the number of Monte Carlo samples N and the Variational Rényi (VR) bound in Li and Turner (2016) that is parametrized by the Rényi alpha-divergence parameter α. Daudel et al. (2023) also introduced the VR-IWAE bound, a variational bound depending on $(N, \alpha) \in \mathbb{N} \times [0, 1)$ that unifies the ELBO, IWAE and VR bounds. A shared feature of these variational bounds is that they can readily be optimized using reparameterized gradient estimators (Kingma and Welling, 2014; Rezende et al., 2014; Burda et al., 2016; Li and Turner, 2016; Tucker et al., 2019; Daudel et al., 2023).

Yet, the theoretical understanding of how the joint choice of importance-weighted variational bound and gradient estimator governs the optimization trajectory remains limited. The properties of the variational bounds themselves are relatively well-established: Maddison et al. (2017); Nowozin (2018); Domke and Sheldon (2018) study the asymptotic behavior of the IWAE bound as $N \to \infty$, with Daudel et al. (2023) exploring the more general VR-IWAE case. Furthermore, recent works investigate the desirable characteristics of the variational approximation at optimality for the VR bound (Margossian et al., 2024; Margossian and Saul, 2025) and the IWAE bound (Cherief-Abdellatif et al., 2025). Despite these insights into importance-weighted variational bounds and their optima, a comprehensive analysis of their gradient estimators is currently missing from the literature.

More precisely, current analyses of reparameterized (REP) and doubly-reparameterized (DREP) gradient estimators, which are the primary gradient estimators in importance-weighted VI, are fragmented and incomplete. Although Daudel et al. (2023) capture the asymptotics of the REP gradient estimator of the VR-IWAE bound as $N \to \infty$, the role of α is completely overlooked outside of the distinction between the pathological case $\alpha = 0$ (IWAE) identified in Rainforth et al. (2018) and the case $\alpha \in (0, 1)$. As for the DREP gradient estimator of the VR-IWAE bound, existing results are restricted to an informal argument specific to the IWAE setting (Tucker et al., 2019). Beyond these theoretical gaps, there is also a general lack of understanding regarding how REP and DREP gradient estimators perform as the quality of the variational approximation deteriorates.

As a result, comparisons between REP and DREP gradient estimators have been confined to empirical studies suggesting that tuning α improves performance and that the DREP gradient estimator outperforms the REP one (Li and Turner, 2016; Daudel et al., 2023; Geffner and Domke, 2020; Dhaka et al., 2021). Our work addresses these limitations through two main contributions to importance-weighted VI, both of which are grounded in broader theoretical results for the Monte Carlo estimation of sample mean ratios we establish. Namely, our contributions are as follows:

(i) **A unified comparison between REP and DREP gradient estimators in importance-weighted VI.** We establish an asymptotic characterization of the expectation and variance for Monte Carlo estimators taking the form of a ratio of sample means. By applying this theory to importance-weighted VI, we obtain asymptotic analyses as $N \to \infty$ that hold under mild conditions and that allow the systematic comparison of REP and DREP gradient estimators tied to the IWAE, VR and VR-IWAE bounds. Our analyses are based on the Signal-to-Noise Ratio (SNR), a metric commonly used in VI that quantifies the quality of an estimator via the ratio of its expected value to its standard deviation. We show that the hyperparameter α is responsible for a bias-variance tradeoff in the REP and DREP

gradient estimators of the VR-IWAE bound. Moreover, we provide the first theoretical justification for the empirical superiority of DREP over REP in importance-weighted VI.

(ii) **A study of importance-weighted gradient estimators in the challenging regime where the variational approximation deteriorates.** We analyze the case where the quality of the variational density deteriorates in the gradient estimators. This approach is motivated by observing that the asymptotic regimes in (i) may only kick in for a large N in practice if the variational density chosen within $\mathcal{Q}$ does not match the posterior density well enough. Specifically, we derive a second analysis at the SNR level when both N and the Kullback-Leibler divergence between the variational and posterior densities go to infinity. Our insights indicate that importance-weighted VI gradient estimators point in a well-founded direction even in such settings. As with our first contribution, the technical tools used in (ii) are valid outside of importance-weighted VI, which signals that their utility might extend beyond the context considered here.

Crucially, our theoretical analyses in (i) and (ii) are complementary analyses that reflect different stages in importance-weighted VI algorithms. While (ii) captures the behavior in challenging settings such as poor parameter initialization, (i) corresponds to settings where the variational density is an accurate approximation of the posterior density, such as during training. In addition, provided that the variational family is sufficiently expressive, we obtain that optimization under the challenging regime of (ii) drives the trajectory toward the well-behaved regime of (i). We validate our theory over three experiments of increasing complexity, before outlining directions for future research. Deferred results, proofs and numerical experiments can be found in the appendix.

2. Background

Consider a model with joint distribution $p_\theta(x, z)$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^a$, where x denotes an observation and z is a latent variable valued in $\mathbb{R}^d$. In Bayesian inference, one seeks to sample from the posterior density $p_\theta(z|x)$. However, in many important cases sampling directly from the posterior density is impossible. This then hinders usual inference tasks such as marginal likelihood optimization, where the goal is to find the optimal $\theta \in \Theta$ which maximizes the marginal log likelihood

$$\ell(\theta; x) = \log\left(\int p_\theta(x, z)\mathrm{d}z\right). \tag{1}$$

To tackle this challenge, VI methods introduce a probability density $q_\phi(z|x)$ parameterized by $\phi \in \Phi$, whose distribution is easy to sample from and where typically $\Phi \subseteq \mathbb{R}^b$. Specifically, in the context of marginal likelihood maximization, VI methods solve an optimization problem involving a variational bound, that is a lower bound on the marginal log likelihood $\ell(\theta; x)$. While the most common example of variational bound is the ELBO

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log w_{\theta,\phi}(z; x)\ \mathrm{d}z, \quad w_{\theta,\phi}(z; x) = \frac{p_\theta(x, z)}{q_\phi(z|x)}\ ,\ z \in \mathbb{R}^d, \tag{2}$$

several alternatives to the ELBO based on importance-weighting ideas have been introduced. Two notable instances of importance-weighted variational bounds are the IWAE bound

(Burda et al., 2016) and the VR-IWAE bound (Daudel et al., 2023), where the latter is defined for all $N \in \mathbb{N}^\star$ and all $\alpha \in [0, 1)$ by

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \int \prod_{i=1}^N q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}. \tag{3}$$

The VR-IWAE bound serves as a unifying objective that bridges several existing VI frameworks, inheriting appealing theoretical and practical properties from these connections. It recovers the ELBO when $N = 1$ (or as $\alpha \to 1$) and the IWAE bound (Burda et al., 2016) when $\alpha = 0$. Daudel et al. (2023) also show that it is a principled approximation of the VR bound ($N \to \infty$), a flexible variational bound which seeks to improve on the ELBO by employing Rényi's alpha-divergences (Li and Turner, 2016). We expand below on the desirable properties of the VR-IWAE bound originating from its ties to the ELBO, IWAE and VR bounds.

(i) *Tightness.* Increasing the sample size N and/or decreasing α in the VR-IWAE bound yields a tighter lower bound on the marginal log likelihood provided that the support of $p_\theta(\cdot|x)$ is contained within the support of $q_\phi(\cdot|x)$: for all $\alpha, \alpha' \in [0, 1)$ with $\alpha \leq \alpha'$ and all $N \in \mathbb{N}^*$,

$$\mathrm{ELBO}(\theta, \phi; x) \leq \ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_{N+1}^{(\alpha)}(\theta, \phi; x) \leq \mathrm{VR}^{(\alpha)}(\theta, \phi; x) \leq \ell(\theta; x) \tag{4}$$

$$\mathrm{ELBO}(\theta, \phi; x) \leq \ell_N^{(\alpha')}(\theta, \phi; x) \leq \ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_N^{(0)}(\theta, \phi; x) \leq \ell(\theta; x), \tag{5}$$

where $\mathrm{VR}^{(\alpha)}(\theta, \phi; x)$ denotes the VR bound from Li and Turner (2016) and is given by

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x)\ w_{\theta,\phi}(z; x)^{1-\alpha}\ \mathrm{d}z \right), \quad \alpha \in [0, 1) \cup (1, \infty). \tag{6}$$

More precisely, Daudel et al. (2023, Theorem 3) establish that: as $N \to \infty$

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \frac{1}{2N} \gamma^{(\alpha)}(\theta, \phi; x)^2 + o\left(\frac{1}{N}\right) \tag{7}$$

$$\text{where} \quad \gamma^{(\alpha)}(\theta, \phi; x)^2 = \frac{1}{1-\alpha} \mathbb{V}_{Z \sim q_\phi(\cdot|x)} \left(\frac{w_{\theta,\phi}(Z; x)^{1-\alpha}}{\mathbb{E}_{Z' \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z'; x)^{1-\alpha})} \right), \tag{8}$$

meaning that the VR-IWAE bound is a finite sample approximation of the VR bound which gets tighter and converges to the VR bound at a $1/N$ rate. Tightness is a natural criterion for assessing the quality of a variational bound as an approximation of the marginal log-likelihood, although by itself tightness does not guarantee more favorable optimization geometry or stationary points closer to a marginal log likelihood maximizer.

(ii) *Ease of optimization.* The VR-IWAE bound can be straightforwardly optimized via stochastic gradient ascent (SGA; Robbins, 1951; Bottou et al., 2018) paired up with the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014; Burda et al., 2016; Daudel et al., 2023). The reparameterization trick is a common tool used in VI which has led to empirical advantages when estimating gradients with respect to ϕ and which relies on the assumption that $Z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ with $\varepsilon \sim \tilde{q}$ (Kucukelbir et al., 2017; Xu

et al., 2019; Mohamed et al., 2020). Under the assumption that Z can be reparameterized, the gradient of the VR-IWAE bound $\nabla_{\theta,\phi}\ell_N^{(\alpha)}(\theta,\phi;x)$ reads

$$\int \prod_{i=1}^{N} \tilde{q}(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(f(\varepsilon_\ell,\phi;x);x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x) \right) \mathrm{d}\varepsilon_{1:N} \tag{9}$$

and since this gradient can be estimated by the unbiased Monte Carlo estimator

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{w_{\theta,\phi}(f(\varepsilon_{m,j},\phi;x);x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(f(\varepsilon_{m,\ell},\phi;x);x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_{m,j},\phi;x);x) \tag{10}$$

with $\varepsilon_{m,1},\dots,\varepsilon_{m,N}$ being i.i.d. samples generated from $\tilde{q}$ for all $m = 1 \dots M$, the resulting procedure enjoys the typical convergence guarantees of unbiased SGA (Hoffman et al., 2013; Bottou et al., 2018; Garrigos and Gower, 2023). In particular, (10) recovers the unbiased gradient estimator used for ELBO optimization in Kingma and Welling (2014) when $N = 1$ (or $\alpha \to 1$) and the one used for IWAE bound optimization in Burda et al. (2016) when $\alpha = 0$. Moreover, (10) is used in Li and Turner (2016) as an approximation of $\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ to derive a practical SGA scheme based on the VR bound as

$$\nabla_{\theta,\phi}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\int \tilde{q}(\varepsilon)\, w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\, d\varepsilon}{\int \tilde{q}(\varepsilon)\, w_{\theta,\phi}(f(\varepsilon,\phi;x);x)^{1-\alpha} d\varepsilon}. \tag{11}$$

Consequently, on top of encompassing the typical ELBO and IWAE optimization frameworks, the VR-IWAE bound provides theoretical grounding for SGA schemes employing (10) with $\alpha \in (0,1)$: while these schemes are biased when viewed through the angle of the VR bound, they are unbiased when the objective function is the VR-IWAE bound (Daudel et al., 2023).

By unifying the ELBO, IWAE and VR bounds, the VR-IWAE bound allows one to treat the choice of the variational bound as a problem of hyperparameter tuning within a shared framework. This motivates a closer examination of how (N,α) influence the optimization trajectory, which we will achieve in this paper via a thorough characterization of key importance-weighted gradient estimators associated to the VR-IWAE bound.

3. Analyzing Importance-weighted VI Gradient Estimators

We begin by defining the notation and the gradient estimators that form the basis of our study.

3.1 Notation and Gradient Estimators Definitions

In the following, we let x be a fixed observation, $\alpha \in [0,1)$, $\Theta \subseteq \mathbb{R}^a$ and $\Phi \subseteq \mathbb{R}^b$ be two open subsets, and we assume that for all $\theta \in \Theta, \phi \in \Phi$, the following holds:

(A1) $\int p_\theta(x,z)\, \nu(\mathrm{d}z) < \infty$, $\int q_\phi(z|x)\, \nu(\mathrm{d}z) = 1$ and $q_\phi(z|x), p_\theta(z|x) > 0$ for ν-a.e. z.

Here ν is a σ-finite measure on $\mathbb{R}^d$, typically the Lebesgue measure on $\mathbb{R}^d$ as used in (1)–(11), or the Lebesgue measure on a subset of $\mathbb{R}^d$, or any other convenient dominating measure on the Borel sets of $\mathbb{R}^d$. To avoid specifying the dominating measure, we will use the notation

$\mathbb{E}$ throughout the paper. If the probability distribution of the involved random variables has not been specified beforehand, we use a subscript to the symbol $\mathbb{E}$ to indicate their densities with respect to ν. Using this convention, (3) and (6) are then respectively written as

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \mathbb{E}_{Z_j \overset{\text{i.i.d.}}{\sim} q_\phi(\cdot|x)} \left(\log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(Z_j; x)^{1-\alpha} \right) \right) \tag{12}$$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \mathbb{E}_{Z \sim q_\phi(\cdot|x)} \left(w_{\theta,\phi}(Z; x)^{1-\alpha} \right), \tag{13}$$

with (A1) ensuring $w_{\theta,\phi}(\cdot; x)$ is well-defined and positive a.s. under the expectation $\mathbb{E}$.

On top of (A1) and in order to compute gradient estimators, we assume that the mappings $\theta \mapsto p_\theta(x)$, $(z, \theta) \mapsto p_\theta(z|x)$, and $(z, \phi) \mapsto q_\phi(z|x)$ are differentiable on Θ, $\mathbb{R}^d \times \Theta$ and $\mathbb{R}^d \times \Phi$, respectively. We also assume that integration with respect to $q_\phi(\cdot|x)$ can always be reparameterized as an expectation independent of ϕ, and that the involved functions of θ and ϕ inside this expectation are differentiable, meaning (A2) below holds:

(A2) There exist a function f and a density $\tilde{q}$ such that $f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ with $\varepsilon \sim \tilde{q}$. In addition, for $\varepsilon \sim \tilde{q}$, we have a.s. that the mapping $\phi \mapsto f(\varepsilon, \phi; x)$ is differentiable on Φ.

For a differentiable function g on $\mathbb{R}^D$ with $D \in \mathbb{N}^\star$, $\partial_{y_k} g(y)$ denotes the partial derivative of $g(y_1, \ldots, y_D)$ with respect to y_k evaluated at $y = (y_1, \ldots, y_D) \in \mathbb{R}^D$ as in (14) and $\nabla_y g(y)$ denotes the gradient vector $(\partial_{y_k} g(y))_{k=1,\ldots,D}$ as in (9). When differentiating, it can be convenient to evaluate the resulting function at a different variable: we write $[h(y, z)]\big|_{y=y'}$ to indicate that the multivariate function $(y, z) \mapsto h(y, z)$ is evaluated at (y', z). In the following, we let $M, N \in \mathbb{N}^\star$ and $\varepsilon_{m,1}, \ldots, \varepsilon_{m,N}$ be i.i.d. samples generated from $\tilde{q}$ for all $m = 1 \ldots M$, we set $Z_{m,j} = f(\varepsilon_{m,j}, \phi; x)$ for all $j = 1 \ldots N$ and all $m = 1 \ldots M$ and we let ψ denote a component of the $\mathbb{R}^{a+b}$-valued variable $(\theta, \phi) = (\theta_1, \ldots, \theta_a, \phi_1, \ldots, \phi_b)$. We next define below the two gradient estimators of the VR-IWAE bound that are central to our study.

Definition 1 (REP and DREP gradient estimators of the VR-IWAE bound) *The reparameterized (REP) gradient estimator of the VR-IWAE bound w.r.t. ψ is given by*

$$\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}} \, \partial_\psi \log w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi; x); x). \tag{14}$$

Further assuming ψ is a component of the $\mathbb{R}^b$-valued variable $\phi = (\phi_1, \ldots, \phi_b)$, the doubly-reparameterized (DREP) gradient estimator of the VR-IWAE bound w.r.t. ψ is given by

$$\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \left(h_{m,j}^{(\alpha)}(\theta, \phi; x) \, \left[\partial_{\psi'} \log w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi'; x); x) \right] \big|_{\phi'=\phi} \right), \tag{15}$$

where, for all $m = 1 \ldots M$ and all $j = 1 \ldots N$,

$$h_{m,j}^{(\alpha)}(\theta, \phi; x) = \alpha \, \frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}} + (1-\alpha) \, \left(\frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}} \right)^2. \tag{16}$$

The REP and DREP gradient estimators are both unbiased estimators of the gradient of the VR-IWAE bound (Daudel et al., 2023). Establishing such properties rigorously requires assumptions to interchange derivatives and expectations, which are typically taken for granted and omitted in VI (see, for instance, Daudel et al., 2023). For the sake of completeness, we provide precise sufficient conditions ensuring that the interchange is valid in Appendix A: $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ for the REP estimator and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$ for the DREP estimator. We now briefly comment on the specificities of the REP and DREP gradient estimators of the VR-IWAE bound.

As previously mentioned, the REP gradient estimator of the VR-IWAE bound acts as a bridge between the REP gradient estimator of the ELBO ($N = 1$ or $\alpha \to 1$), the REP gradient estimator of the IWAE bound ($\alpha = 0$) and the (biased) REP gradient estimator of the VR bound ($\alpha \in (0, 1)$). The DREP gradient estimator of the VR-IWAE bound (15) was introduced in Daudel et al. (2023) as a generalization of the DREP gradient estimator of the ELBO (Roeder et al., 2017) and of the IWAE bound (Tucker et al., 2019). The main motivation behind DREP gradient estimators comes from the observation made in Roeder et al. (2017) that:

$$\partial_\psi \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) = \left[\partial_{\psi'} \log w_{\theta,\phi}(f(\varepsilon, \phi'; x); x) - \partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)\right]\big|_{\phi'=\phi}.$$

This identity reveals that the REP gradient estimator of the ELBO ($N = 1$ in (14)) has a nonzero variance at optimality, since the left-hand side above has nonzero variance when $q_\phi(\cdot|x)$ matches the exact posterior density $p_\theta(\cdot|x)$ everywhere due to the score function term $\left[\partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)\right]\big|_{\phi'=\phi}$. By contrast, the DREP gradient estimator of the ELBO ($N = 1$ in (15)) enjoys a variance which goes to zero as the variational density approaches the posterior one, and is shown in Roeder et al. (2017) to lead to improved empirical performance as well as studied theoretically in Domke et al. (2023); Kim et al. (2024).

In the spirit of Roeder et al. (2017), the DREP gradient estimator of the IWAE bound (Tucker et al., 2019) and of the VR-IWAE bound (Daudel et al., 2023) maintain the zero-variance property at optimality. However, the behavior of these gradient estimators during the optimization procedure compared to their REP counterparts is much less straightforward. To clarify these dynamics and advance the field of importance-weighted VI, we seek to provide the first informative asymptotic analyses for the REP and DREP gradient estimators of the VR-IWAE bound enabling their comparison. To that end, we will start by proving a new general result capturing the asymptotic behavior in expectation and variance of sample mean ratios as the number of Monte Carlo samples increases.

3.2 Key Technical Result on the Asymptotic Moments of Sample Mean Ratios

We present below a key technical result which captures the asymptotic behavior in expectation and variance for ratios of sample means as the number of Monte Carlo samples increases.

Theorem 1 *Let $(X_1, W_1), (X_2, W_2), \ldots$ be i.i.d. with the same distribution as a generic pair (X, W) valued in $\mathbb{R} \times \mathbb{R}_+$. Denote, for all $N \geq 1$, $\overline{X}_N = N^{-1} \sum_{i=1}^N X_i$ and $\overline{W}_N = N^{-1} \sum_{i=1}^N W_i$. Suppose that there exists $\eta > 0$ such that*

$$\sup_{u>0} \left(u^{-\eta}\, \mathbb{P}\left(W \leq u\right)\right) < \infty. \tag{17}$$

Then the following assertions hold.

(i) If $\mathbb{E}(|X|) < \infty$, *then*

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) = \frac{\mathbb{E}(X)}{(\mathbb{E}(W))^\mu} \,. \tag{18}$$

(ii) If $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(W^2) < \infty$ *and* $\mathbb{E}(|X|\,W) < \infty$, *then*

$$\lim_{N\to\infty} N\left(\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) - \frac{\mathbb{E}(X)}{(\mathbb{E}(W))^\mu}\right) = \frac{(\mu+1)\mu\,\mathbb{E}(X)\,\mathbb{V}(W)}{2\,(\mathbb{E}(W))^{\mu+2}} - \frac{\mu\,\mathbb{C}\mathrm{ov}(X,W)}{(\mathbb{E}(W))^{\mu+1}} \,. \tag{19}$$

Moreover, if instead $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(W) < \infty$ *and* $\mathbb{E}(|\mathbb{E}(W)X - \mathbb{E}(X)W|\,W) < \infty$, *then the convergence in (19) still holds when* $\mu = 1$ *with a limit expressed by*

$$\lim_{N\to\infty} N\left(\mathbb{E}\left(\frac{\overline{X}_N}{\overline{W}_N}\right) - \frac{\mathbb{E}(X)}{\mathbb{E}(W)}\right) = \frac{\mathbb{E}\left(W\left[\mathbb{E}(X)W - \mathbb{E}(W)X\right]\right)}{(\mathbb{E}(W))^3}. \tag{20}$$

(iii) If $\mathbb{E}(|X|^2) < \infty$, *then*

$$\lim_{N\to\infty} \mathbb{V}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) = 0 \,. \tag{21}$$

(iv) If $\mathbb{E}(|X|^2) < \infty$ *and* $\mathbb{E}(W^2) < \infty$, *then*

$$\lim_{N\to\infty} N\,\mathbb{V}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) = \frac{\mathbb{V}\left(\mathbb{E}(W)X - \mu\mathbb{E}(X)W\right)}{(\mathbb{E}(W))^{2\mu+2}} \,. \tag{22}$$

Moreover, if instead $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(W) < \infty$ *and* $\mathbb{E}(|\mathbb{E}(W)X - \mathbb{E}(X)W|^2) < \infty$, *then the convergence in* (22) *still holds when* $\mu = 1$ *thus giving*

$$\lim_{N\to\infty} N\,\mathbb{V}\left(\frac{\overline{X}_N}{\overline{W}_N}\right) = \frac{\mathbb{V}\left(\mathbb{E}(W)X - \mathbb{E}(X)W\right)}{(\mathbb{E}(W))^4} \,. \tag{23}$$

The proof of Theorem 1 is deferred to Appendix B. Observe that (i) and (iii) establish the intuitive limiting behavior for the expectation and variance of $\overline{X}_N/(\overline{W}_N)^\mu$, while (ii) and (iv) capture the behavior of the corresponding first-order terms. Together, (i)-(iv) will provide a general framework to derive the asymptotics for the REP (14) and DREP (15) gradient estimators of the VR-IWAE bound via selected choices of $\overline{X}_N, \overline{W}_N$ and μ. Before proceeding to do so in Section 3.3, we comment on the assumptions we make in Theorem 1 and we discuss the implications of this result for VI and more broadly Monte Carlo methods.

Assumptions. Essentially, the assumptions made in Theorem 1 ensure that the quantities of interest appearing in this theorem are well-defined. This is clear from the specific conditions associated with assertions (i)-(iv) (e.g., the integrability condition $\mathbb{E}(|X|) < \infty$ in (i) that guarantees the existence of $\mathbb{E}(X)$) and we now discuss the condition (17), which is common to all the assertions (i)-(iv).

Assuming that there exists $\eta > 0$ such that (17) holds ensures that the distribution of W does not place excessive mass near 0, since $\mathbb{P}(W \leq u) = O(u^\eta)$ as $u \downarrow 0$. Hence this condition slightly strengthens the right-continuity at zero of the distribution function $u \mapsto \mathbb{P}(W \leq u)$ by imposing a power decay of arbitrarily low positive exponent. Note that without right-continuity, the quantities studied in Theorem 1 would be ill-defined due to the nonzero probability of a vanishing denominator in $\overline{X}_N/(\overline{W}_N)^\mu$. As shown in the lemma below, (17) admits equivalent formulations frequently encountered in the related literature.

Lemma 1 *Let $W_1, \ldots, W_N$ be i.i.d. with the same distribution as a generic random variable W valued in $\mathbb{R}_+$. Then the following three assertions are equivalent.*

(i) There exists $\eta > 0$ such that (17) holds.

(ii) There exists $\mu > 0$ and $N \geq 1$ such that $\mathbb{E}((\overline{W}_N)^{-\mu}) < \infty$.

(iii) There exists $\mu > 0$ such that $\mathbb{E}(W^{-\mu}) < \infty$.

Lemma 1 is a direct consequence of the more precise statement Proposition 5, which can be found in Appendix B.2 alongside its proof. Conditions akin to (ii), although usually stated for a specific value of μ, are typically used when investigating importance-weighted VI bounds or their gradients (see, e.g., Maddison et al., 2017; Domke and Sheldon, 2018; Daudel et al., 2023). As for (iii), it appears in Deligiannidis et al. (2025), which studies the asymptotic bias of self-normalized importance sampling. The assertion (ii) can in fact be equivalently stated with an arbitrary $\mu > 0$ or an arbitrary $N \geq 1$ (see Proposition 5), which yields (iii) as the special case where N is arbitrarily fixed to $N = 1$.

Related work and broader impact. Theorem 1 subsumes several results studying importance-weighted estimators. Specifically, Theorem 1 contributes to the Monte Carlo literature by broadening the applicability of existing importance-weighting results while operating under strictly less restrictive assumptions.

Indeed, the most closely related existing studies have focused on deriving (20) and (23), or specific instantiations of these, under stronger assumptions and different proof techniques compared to ours (Rainforth et al., 2018; Daudel et al., 2023; Deligiannidis et al., 2025); see Table 1 for a quick overview and more details are also available in Appendix B.5. Furthermore, to the best of our knowledge, existing studies are entirely restricted to settings where $\mu = 1$ while Theorem 1 accommodates the general case $\mu > 0$. As we shall see, this extension is particularly relevant in the context of VI and will emerge as a theoretical prerequisite for analyzing the DREP gradient estimator of the VR-IWAE bound. This will permit us to bridge the gap left by Tucker et al. (2019) and Daudel et al. (2023), who proposed the DREP gradients for the IWAE and VR-IWAE bounds, respectively, but did not provide asymptotic guarantees.

Ultimately, our framework provides a theoretical foundation for analyzing importance-weighted estimators, both within and well beyond VI. Applications outside of VI include

any setting where an importance-weighted estimator is employed to approximate a quantity of interest, e.g., the fine-tuning of Large Language Models (Liao et al., 2026).

3.3 Asymptotics of gradient estimators of the VR-IWAE bound as $N \to \infty$

Let us now build a unified framework that enables the comparison between the REP (14) and DREP (15) gradient estimators of the VR-IWAE bound. The structure of the REP and DREP gradient estimators suggests that the asymptotic behavior in expectation and variance of these estimators can be derived by applying Theorem 1 for selected choices of $\overline{X}_N, \overline{W}_N$ and μ. Letting $\alpha \in [0,1)$, $\varepsilon \sim \tilde{q}$, $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. copies of ε and denoting

$$\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x) = w_{\theta,\phi}(f(\varepsilon, \phi'; x); x), \quad \theta \in \Theta\,,\ \phi, \phi' \in \Phi, \tag{24}$$

we notably see that setting $X = (1-\alpha)^{-1}\partial_\psi \left[\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}\right]$, $W = \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}$ and $\mu = 1$ in Theorem 1 permits to capture the asymptotics of the REP gradient estimator in expectation and variance. Starting for instance with (18), we directly get that, under the assumptions of (i) in Theorem 1,

$$\lim_{N\to\infty} \mathbb{E}\left(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)\right) = \frac{1}{1-\alpha} \frac{\mathbb{E}(\partial_\psi \left[\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}\right])}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha})}. \tag{25}$$

While (25) provides an asymptotic limit, its practical implications are not immediately transparent. The same conclusion holds for results obtained by substituting the above choice for (X, W, μ) into Theorem 1. The remainder of this subsection is then devoted to deriving meaningful insights on the behavior of the REP and DREP gradient estimators of the VR-IWAE bound by leveraging Theorem 1 as a theoretical foundation. To that end, we will rely on the following assumption, which corresponds to a rewriting of the condition (17) from Theorem 1 for $W = \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}$ that does not depend on α.

(A3) There exists $\delta > 0$ such that $\sup_{t>0} \left(t^{-\delta}\mathbb{P}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x) \le t)\right) < \infty$.

The theorem below then analyses the behavior in expectation of the REP gradient estimator (14) as $N \to \infty$.

Theorem 2 *Assume* (A3) *and* ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$). *Then, as* $N \to \infty$,

$$\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x) + o\,(1)\,. \tag{26}$$

Further assuming that $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}\right) < \infty$ *yields: as* $N \to \infty$,

$$\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \frac{1}{2N}\partial_\psi[\gamma^{(\alpha)}(\theta, \phi; x)^2] + o\left(\frac{1}{N}\right). \tag{27}$$

Table 1: Closely-related results subsumed by Theorem 1 under less stringent conditions

Rainforth et al. (2018) & Daudel et al. (2023)	(20) and (23) with stronger moment conditions on (X, W) and problem-specific choices of (X, W)
Deligiannidis et al. (2025)	(20) with stronger moment condition on $\lvert\mathbb{E}(W)X - \mathbb{E}(X)W\rvert W$

The proof of Theorem 2 is deferred to Appendix C.1.2. By Theorem 2, $\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ thus converges to $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ at a $1/N$ rate, where $\mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ is the VR bound (6). We now discuss the methodological implications of Theorem 2.

Theorem 2 predicts that the REP gradient estimator of the VR-IWAE bound points on average in the direction which maximizes the VR bound when $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) \neq 0$ and which minimizes $\gamma^{(\alpha)}(\theta,\phi;x)^2$ when $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = 0$. In both cases, the REP gradient estimator of the VR-IWAE bound targets a well-founded direction that is expected to improve on ELBO maximization.

Indeed, the VR bound with $\alpha \in (0,1)$ is known to penalize under-dispersion, driving $q_\phi(\cdot|x)$ to capture regions of posterior probability mass that the standard ELBO typically ignores (Li and Turner, 2016; Margossian et al., 2024; Minka, 2005; Bui et al., 2016; Margossian and Saul, 2025). A similar conclusion applies when $\alpha = 0$, with the particularity that $\mathrm{VR}^{(0)}(\theta,\phi;x) = \ell(\theta;x)$ so that the REP gradient estimator of the VR-IWAE bound points on average in the direction which maximizes the marginal log-likelihood for $\psi \in \{\theta_1,\dots,\theta_a\}$ and minimizes $\gamma^{(0)}(\theta,\phi;x)^2 = \mathbb{V}\left(p_\theta(z|x)/q_\phi(z|x)\right)$ for $\psi \in \{\phi_1,\dots,\phi_b\}$ (Rainforth et al., 2018).

Since $\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \partial_\psi \ell_N^{(\alpha)}(\theta,\phi;x)$ under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, Theorem 2 provides an asymptotic result for the gradient of the VR-IWAE bound which is coherent with the one established in Daudel et al. (2023, Theorem 3) and reviewed in (7). Note also that, assumption-wise, (A3) is inherited from Theorem 1, $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ permits to leverage the reparameterization trick assumption and $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$ ensures that $\gamma^{(\alpha)}(\theta,\phi;x)^2$ is well-defined.

Our next step to further comprehend the behavior of the REP gradient estimator (14) is to capture the behavior of its Signal-to-Noise ratio (SNR) as $N \to \infty$. Recalling that for a random variable X the SNR is given by $\mathrm{SNR}[X] = |\mathbb{E}(X)|/\sqrt{\mathbb{V}(X)}$, an unbiased gradient estimator is expected to be accurate if the SNR is high (so that the target dominates the additive stochastic error) and noisy otherwise. Since the summands of the $\sum_m$ summations in (14) are i.i.d., we have $\mathrm{SNR}[\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{M}\,\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ and we can interpret the $\sum_m$ summations as a way to improve the SNR by the factor $\sqrt{M}$ through averaging, while multiplying the computational cost by M. As such, it suffices to consider the setting $M = 1$ in the following and to focus on the role of (N,α) in $\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$. We then have the theorem below.

Theorem 3 *Assume* (A3) *and* $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$. *Further assume that* $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$ *and that* $\mathbb{V}\left(\partial_\psi\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$. *Then, as* $N \to \infty$,

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{N}\,\frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \frac{1}{2N}\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2] + o\left(\frac{1}{N}\right)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)} + o\left(1\right)}\,, \quad (28)$$

$$\textit{where} \quad \psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \frac{1}{(1-\alpha)^2}\mathbb{V}\left(\partial_\psi\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)\right)\,. \quad (29)$$

The proof of Theorem 3 is deferred to Appendix C.1.3. Besides $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2]$, which we had already identified as key quantities in Theorem 2 to analyze the REP gradient estimator (14), it is now clear that $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ also plays an important role in the success of this estimator. Note that compared to Theorem 2, Theorem 3 further requires $\mathbb{V}\left(\partial_\psi\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$ so that $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ is well-defined.

Asymptotic results related to (28) appear in Rainforth et al. (2018) and in the proof of Daudel et al. (2023, Theorem 1), albeit under more restrictive assumptions (see Appendix B.5). For the specific case $\alpha = 0$, Rainforth et al. (2018) establish that the SNR scales as $\mathcal{O}(\sqrt{N})$ when learning θ and decays at a rate of $\mathcal{O}(1/\sqrt{N})$ when learning ϕ. Furthermore, they highlight the quantities $\partial_\psi \mathrm{VR}^{(0)}(\theta, \phi; x)$, $\partial_\psi[\gamma^{(0)}(\theta, \phi; x)^2]$ and $\psi\text{-}V^{(0,\mathrm{REP})}(\theta, \phi; x)$ in their SNR result. Daudel et al. (2023) show that the SNR decay for ϕ can be mitigated by setting $\alpha \in (0, 1)$, but the quantities $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x)$, $\partial_\psi[\gamma^{(\alpha)}(\theta, \phi; x)^2]$ and $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ do not appear explicitly in their SNR result. Theorem 3 thus generalizes the approach of Rainforth et al. (2018) to the entire range of α on $[0, 1)$, and operates under strictly weaker assumptions than those required by the two aforementioned previous works. This in turn yields important implications regarding the role of α across the entire range $[0, 1)$.

More precisely, when $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x) \neq 0$ (which is a reasonable assumption to make unless $\alpha = 0$ and ψ is among the components of $\phi = (\phi_1, \ldots, \phi_b)$ or we are at a local optimum for (θ, ϕ)), (28) simplifies to

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \sqrt{N}\, \frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x) + o(1)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)} + o(1)},$$

which highlights a bias-variance tradeoff at play in the REP gradient estimator of the VR-IWAE bound via the choice of the hyperparameter α. Here, the gradient of the VR bound represents the bias-controlling component, while $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ is the variance term. As α decreases from 1 to 0, the VR bound interpolates monotonically from the ELBO ($\alpha \to 1$) toward the marginal log-likelihood ($\alpha = 0$) (a monotonicity property recovered, for instance, by taking $N \to \infty$ in (5)). While this suggests a preference for values of α near zero in marginal likelihood optimization, the practical choice of α is in fact constrained by the simultaneous evolution of the variance term $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$. Consequently, Theorem 3 formally explains how the tuning of α has led to improved empirical performance in tasks seeking to optimize the marginal log-likelihood using the REP gradient estimator (14) (see, e.g., Li and Turner, 2016; Daudel et al., 2023).

Crucially, $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ can be further interpreted in light of Roeder et al. (2017); Tucker et al. (2019): the intuition that the score function adds variance to REP gradient estimators of the IWAE bound ($\alpha = 0$) is now verified theoretically for the general REP gradient estimator (14) ($\alpha \in [0, 1)$) by looking at the asymptotics of its SNR as $N \to \infty$. The score function adds variance via $\left[\partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)\right]\big|_{\phi'=\phi}$, which can notably be observed by considering the case $p_\theta(\cdot|x) = q_\phi(\cdot|x)$ in (29) leading to (see (115) of Appendix C.1.3 for details):

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \mathbb{V}\left(\left[\partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)\right]\big|_{\phi'=\phi}\right). \tag{30}$$

Although the DREP gradient estimator (15) was designed to reduce the variance of the REP one (14), a theoretical justification of this property is currently lacking. We therefore next derive the asymptotic SNR of the DREP estimator to formally compare the REP and DREP approaches. Since the DREP gradient estimator (15) only impacts the learning of ϕ, our analysis focuses on its asymptotic behavior w.r.t. ϕ. This leads us to the theorem below.

Theorem 4 *Assume* (A3) *and* ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$). *Further assume that* $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$ *and* $\mathbb{V}\left([\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})]|_{\phi'=\phi}\right) < \infty$, *where* ψ *denotes a component of the* $\mathbb{R}^b$*-valued variable* $\phi = (\phi_1, \ldots, \phi_b)$. *The following assertions hold.*

(i) If $\alpha \in (0,1)$ *and* $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right) < \infty$, *then, as* $N \to \infty$,

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)] = \sqrt{N}\ \frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) + o\left(1\right)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)} + o(1)} \tag{31}$$

$$\text{with } \psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \frac{\alpha^2}{(1-\alpha)^2}\mathbb{V}\left(\left[\partial_{\psi'}\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})}\right)\right]\Big|_{\phi'=\phi}\right). \tag{32}$$

(ii) If $\alpha = 0$, *then, as* $N \to \infty$,

$$\mathrm{SNR}\left[\psi\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)\right] = \sqrt{N}\ \frac{\left|\frac{1}{2}\partial_\psi[\gamma^{(0)}(\theta,\phi;x)^2] + o\left(1\right)\right|}{\sqrt{\psi\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x)} + o(1)}, \tag{33}$$

$$\begin{aligned}\text{with } \psi\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = \mathbb{V}\Bigg(&\frac{\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^2\right)\right]|_{\phi'=\phi}}{2\,\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))^2} \\ &- \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\,\mathbb{E}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^2\right)\right]|_{\phi'=\phi}\right)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))^3}\Bigg).\end{aligned} \tag{34}$$

The proof of Theorem 4 is deferred to Appendix C.1.4. Similarly to Theorem 3 for the REP gradient estimator (14), Theorem 4 highlights a bias-variance tradeoff in the DREP gradient estimator (15) through the choice of the hyperparameter $\alpha \in [0,1)$.

However, a critical distinction arises when comparing the two quantities $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ and $\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)$. Unlike $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$, which remains non-zero even when the variational approximation matches the true posterior ($p_\theta(\cdot|x) = q_\phi(\cdot|x)$), we get that $\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = 0$ in that case. Theorems 3 and 4 thus suggest choosing the DREP gradient estimator over the REP one, aligning with the original intuitions of Tucker et al. (2019); Roeder et al. (2017). From an optimization perspective, the property that $\psi\text{-}V^{(\alpha,\mathrm{DREP})}$ vanishes near an optimum where the variational approximation matches the posterior, hence driving the asymptotic variance of the DREP gradient estimator (15) to zero, is referred to as the interpolation condition. When this condition holds, the convergence rate of SGA is known to improve significantly (Vaswani et al., 2019).

To the best of our knowledge, Theorem 4 is the first asymptotic result in importance-weighted VI which supports the use of DREP gradient estimators over REP gradient estimators. Indeed, the asymptotic $\sqrt{N}$ rate for the SNR of the DREP gradient estimator of the IWAE bound ($\alpha = 0$) was only informally derived in Tucker et al. (2019, Section 8.2). Moreover, the focus of Tucker et al. (2019) was on how the $\sqrt{N}$ rate of the DREP gradient estimator improved on the $\sqrt{1/N}$ rate of the REP gradient estimator when learning ϕ thanks to a variance reduction phenomenon.

In contrast, our approach shows under mild conditions that the SNR rate scales like $\sqrt{N}$ for both the REP and DREP gradient estimators when $\alpha \in [0,1)$ (except when $\alpha = 0$ and

$\psi \in \{\phi_1, \dots, \phi_b\}$ in the REP gradient estimator). Importantly, we go beyond the variance reduction phenomenon that originates from increasing N by taking a closer look at the constants appearing asymptotically in the SNR of the REP and DREP gradient estimators. We obtain that within a same class of gradient estimators (REP or DREP), the tuning of α leads to further variance reduction in parallel to increasing N, at the cost of potentially increasing the bias. Furthermore, ψ-$V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ and ψ-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ are quantities that enable the comparison between the REP and DREP gradient estimators, with theoretical evidence suggesting that the latter is expected to outperform the former. As an aside, observe that all the results we obtained extend naturally to the case $\alpha = 1$ corresponding to the ELBO, albeit in a simpler non-asymptotic way (see Remarks 4 and 6 in the appendix for more details).

We now present two insightful examples where Theorems 2 to 4 apply and in which all the quantities of interest appearing in these results are analytically tractable.

Example 1 *Let $\theta, \phi \in \mathbb{R}^d$. Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix. Consider the reparameterization given by $Z = f(\varepsilon, \phi; x) = \varepsilon + \phi$ where $\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and the goal is to learn ϕ. Then, we can apply Theorems 2 to 4 and all the terms appearing in these results are analytically tractable. In particular, to emphasize the impact of the latent space dimension d, we set $\theta - \phi = \epsilon \cdot \boldsymbol{u}_d$ where $\boldsymbol{u}_d$ denotes the d-dimensional vector whose coordinates are all equal to 1 and $\epsilon > 0$. Then, we have that: for all $\alpha \in [0, 1)$ and $k = 1 \dots d$,*

$$\mathbb{E}(\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \epsilon\alpha + \frac{\epsilon(1-\alpha)\mathrm{e}^{(1-\alpha)^2 d\epsilon^2}}{N} + o\left(\frac{1}{N}\right) \quad \text{as } N \to \infty,$$

$$\phi_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \mathrm{e}^{(1-\alpha)^2 d\epsilon^2}\left(1 + (1-\alpha)^2\epsilon^2\right),$$

$$\phi_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = \begin{cases} 0 & \text{if } \alpha \in (0,1), \\ \epsilon^2 \mathrm{e}^{2d\epsilon^2}\left(\mathrm{e}^{4d\epsilon^2} - 4\mathrm{e}^{2d\epsilon^2} + 4\mathrm{e}^{d\epsilon^2} - 1\right) & \text{if } \alpha = 0. \end{cases}$$

The derivation details for this example are deferred to Appendix C.2.2.

Example 1 showcases how α enables a bias-variance tradeoff for the REP and DREP gradient estimators of the VR-IWAE bound. For the REP gradient estimator, the variance term ϕ_k-$V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ is reduced by increasing α at the cost of increasing the bias in the gradient of the VR bound $\alpha\epsilon$. Viewing this result in terms of SNR, Theorem 3 yields: as $N \to \infty$,

$$\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \begin{cases} \dfrac{\sqrt{N}\epsilon\left(\alpha \exp\left(\frac{-(1-\alpha)^2 d\epsilon^2}{2}\right)\right)}{\sqrt{1 + (1-\alpha)^2\epsilon^2}}(1 + o(1)) & \text{if } \alpha \in (0,1) \\ O\left(1/\sqrt{N}\right) & \text{if } \alpha = 0. \end{cases}$$

Hence setting $\alpha \in (0, 1)$ not only improves the asymptotic SNR rate in N, but increasing α within that range also improves the leading order constant. For the DREP gradient estimator, it is worth highlighting that ϕ_k-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ is 0 for all $\alpha \in (0, 1)$ as opposed to the case $\alpha = 0$, meaning that letting $\alpha \in (0, 1)$ in this estimator improves the SNR rate in N (note that the setting where ψ-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = 0$ for certain values of α seems unlikely enough in practice to warrant a specific theoretical study of the leading order term in the SNR).

Example 2 *Let $\theta \in \mathbb{R}^d$, $\phi = (\tilde{a}, \tilde{b}) \in \mathbb{R}^d \times \mathbb{R}^d$ and $A = \mathrm{diag}(\tilde{a})$. Set $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z; Ax + \tilde{b}, 2/3\ \boldsymbol{I}_d)$ as in Rainforth et al. (2018). Consider the reparameterization given by $Z = \sqrt{\frac{2}{3}}\varepsilon + Ax + \tilde{b}$, where $\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then, the assumptions from Theorems 2 to 4 are met and all the terms appearing in these results are analytically tractable. In particular, denoting $x = (x_1, \ldots, x_d)$, $\theta = (\theta_1, \ldots, \theta_d)$, $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_d)$, letting $Ax + \tilde{b} = (\theta + x)/2 + \epsilon \boldsymbol{u}_d$ with $\epsilon > 0$, we have that: for all $\alpha \in [0, 1)$ and $k = 1 \ldots d$,*

$$\mathbb{E}(\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = -\frac{6\epsilon\alpha}{4-\alpha} - \frac{24\epsilon(1-\alpha)(4-\alpha)^{d-1}\mathrm{e}^{\frac{24(1-\alpha)^2 d\epsilon^2}{(5-2\alpha)(4-\alpha)}}}{N 3^{d/2}(5-2\alpha)^{\frac{d}{2}+1}} + o\left(\frac{1}{N}\right), \quad \text{as } N \to \infty$$

$$\tilde{b}_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \frac{4(4-\alpha)^d}{(15-6\alpha)^{d/2}}\mathrm{e}^{\frac{24(1-\alpha)^2}{(4-\alpha)(5-2\alpha)}d\epsilon^2}\left(\frac{2}{5-2\alpha} + \left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2\right)$$

$$\tilde{b}_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = \frac{\alpha^2}{16}\ \tilde{b}_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x), \quad \alpha \in (0, 1)$$

and $\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x)$ is given by (165) *in Appendix C.2.3. The derivation details for this example are deferred to Appendix C.2.3.*

Example 2 illustrates again the bias-variance tradeoff occurring in the REP and DREP gradient estimators as α varies. Furthermore, the reduction in asymptotic variance obtained by using the DREP gradient estimator over the REP one is remarkably simple to visualize when $\alpha \in (0, 1)$: it is captured by a multiplicative factor $\alpha^2/16$, which itself translates into a direct improvement in terms of asymptotic SNR: as $N \to \infty$,

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)] = 4\alpha^{-1}\ \mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)](1 + o(1)), \quad \alpha \in (0, 1).$$

3.4 Gradient estimators in deteriorating variational approximation regimes

The asymptotic results of Section 3.3 are obtained as $N \to \infty$ for a given model and fixed parameters θ and ϕ. However, how large N needs to be in practice for the provided asymptotic behaviors to kick in depends on how big the mismatch between the target density and the variational density is.

One way to see this is to keep $\alpha \in [0, 1)$ fixed in Examples 1 and 2 and to focus on the impact of d and ϵ in those examples. Increasing d or ϵ can be interpreted as worsening the mismatch between the posterior density $p_\theta(\cdot|x)$ and its variational approximation $q_\phi(\cdot|x)$: as d or ϵ increase, the main terms in the predicted asymptotic behaviors of Examples 1 and 2 vanish, and the variances will blow up. Consequently, as the mismatch between the posterior and variational densities grows, the sample size N required to reach the asymptotic regimes predicted by Theorems 2 to 4 may become prohibitively large. This calls for a specific study of importance-weighted VI gradient estimators in settings where the variational approximation of the posterior density deteriorates.

To that end, we consider in the following an asymptotic regime where $N \to \infty$ in such a way that the mismatch between the target density and the variational one increases in the meantime. Since this mismatch can be the result of various causes, we start by picking a criterion reflecting this mismatch, before making and motivating assumptions that enable us

to carry out the analysis. Specifically, we propose to use the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\left(\theta, \phi; x\right) := \ell(\theta; x) - \mathrm{ELBO}(\theta, \phi; x) \tag{35}$$

as a natural measure of mismatch and we seek to capture the behavior of importance-weighted VI gradient estimators as $N, \mathrm{KL}\left(\theta, \phi; x\right) \to \infty$. Here, it is understood that the considered target and/or variational density depend on N in any possible way, such as through θ, ϕ, or x (for example via their dimensions) but with the constraint that both $N \to \infty$ and $\mathrm{KL}\left(\theta, \phi; x\right) \to \infty$. This regime is highly challenging since the gap between the marginal log-likelihood and the ELBO (that is, the VR-IWAE bound with $N = 1$) goes to infinity and the variational approximation thus severely deteriorates.

We now introduce two assumptions that enable a rigorous treatment of this regime for the REP gradient estimator (14). The first assumption is on the log-weight $\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)$:

(B1) For $\varepsilon \sim \tilde{q}$, the random field $\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a Gaussian random field.

The second assumption, denoted $(\mathsf{B}^{\mathrm{REP}}_{\mathrm{df}})$, guarantees that we can interchange derivative and expectation signs as needed when (B1) also holds. It is postponed to Appendix A.2 to not overload the paper with technical assumptions that can be overlooked in a first read and we concentrate instead on (B1). Assumption (B1) permits us to use the convenient representation

$$\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x) = \mathbb{E}_{\varepsilon\sim\tilde{q}}(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)) - \sqrt{\mathbb{V}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right)} S(\varepsilon, \theta, \phi; x), \tag{36}$$

where $\left(S(\varepsilon, \theta, \phi; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a centered Gaussian random field with unit variance. Intuitively, as importance weights often asymptotically approach a log-normal distribution (see, e.g., Daudel et al., 2023, for an explanation via the Central Limit Theorem), this assumption is meant to capture the limiting behavior where the weights are exactly log-normal. Based on (36), we can also link our measure of mismatch to the variance of the log-weights as

$$\frac{1}{2}\mathbb{V}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right) = \ell(\theta; x) - \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right) = \mathrm{KL}\left(\theta, \phi; x\right), \tag{37}$$

which follows from (A2), (35) and the formula for exponential moments of Gaussian variables

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right) = \mathrm{e}^{\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right) + \frac{1}{2}\mathbb{V}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)}.$$

The representation (36) will turn out to be key in our proofs since the REP gradient estimator involves differentiating $\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)$.

We now investigate the asymptotic behavior of the REP gradient estimator (14) under (B1)-$(\mathsf{B}^{\mathrm{REP}}_{\mathrm{df}})$ as $N, \mathrm{KL}\left(\theta, \phi; x\right) \to \infty$. This leads us to the next proposition.

Proposition 1 *Assume* (B1) *and* $(\mathsf{B}^{\mathrm{REP}}_{\mathrm{df}})$. *Let* $N, \mathrm{KL}\left(\theta, \phi; x\right) \to \infty$ *with*

$$\log N = o\left(\mathrm{KL}\left(\theta, \phi; x\right)\right) \tag{38}$$

$$\limsup \left|\mathrm{C}_\psi(\theta, \phi; x)\right| < 1, \tag{39}$$

where

$$\mathrm{C}_\psi(\theta, \phi; x) := \mathbb{C}\mathrm{orr}_{\varepsilon\sim\tilde{q}}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x), \partial_\psi \log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right). \tag{40}$$

Then, it holds that

$$\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = O\left(\sqrt{\log N}\ |\mathrm{C}_\psi(\theta,\phi;x)|\right), \tag{41}$$

with

$$\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) := \mathrm{SNR}\left[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \mathbb{E}\left(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right)\right]. \tag{42}$$

The proof of Proposition 1 is deferred to Appendix D.4.1. Proposition 1 provides an upper bound on $\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ as $N, \mathrm{KL}\,(\theta,\phi;x) \to \infty$ under two conditions: the first condition (38) ties N to $\mathrm{KL}\,(\theta,\phi;x)$ in such a way that $\log N$ should not grow quicker than $\mathrm{KL}\,(\theta,\phi;x)$, while the second condition (39) is a mild condition to avoid the degenerate case where the log likelihood and its derivative tend to be colinear (by requiring these correlations to stay away from 1 and -1).

Here, $\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ captures whether the REP gradient estimator (14) with $N > 1$ behaves significantly differently from the case $N = 1$. More precisely, it evaluates how the difference $\mathbb{E}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) - \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ compares to the stochastic error in the REP gradient estimator, that is, if the shift obtained by using $N > 1$ instead of $N = 1$ is large enough to dominate the noise of the REP gradient estimator. In the theorem below, we provide more precise results characterizing its asymptotic behavior.

Theorem 5 *Assume* (B1) *and* ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$). *Let* $N, \mathrm{KL}\,(\theta,\phi;x) \to \infty$ *with (38) and*

$$\sqrt{\log N}\,\mathrm{C}_\psi(\theta,\phi;x) = O(1). \tag{43}$$

Then, it holds that

$$\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \sqrt{2\log N}\ |\mathrm{C}_\psi(\theta,\phi;x)|\,(1+o(1)). \tag{44}$$

If moreover,

$$\sqrt{\log N}\,\mathrm{C}_\psi(\theta,\phi;x) = o(1), \tag{45}$$

then we have

$$\mathrm{SNR}\left[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right] = \mathrm{SNR}\left[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right]\,(1+o(1)) + o(1)\,. \tag{46}$$

The proof of Theorem 5 can be found in Appendix D.4.2. Theorem 5 shows that, as $N, \mathrm{KL}\,(\theta,\phi;x) \to \infty$ with (38) and under the condition (43) (which ensures (39) holds), the upper bound given in (41) is tight. In addition, it captures the asymptotic behavior of $\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ under the stronger condition (45). We now further delve into the interpretation of Theorem 5.

Assumptions made in Theorem 5. As previously mentioned, (38) assumes that we are in a regime where $\log N$ does not grow faster than $\mathrm{KL}\,(\theta,\phi;x)$. We next elucidate the conditions (43) and (45). By appealing to the equivalent formulation of $|\mathrm{C}_\psi(\theta,\phi;x)|$ proved in (194) of Appendix D.3, we get that (43) and (45) are assumptions on the behavior of

$$\sqrt{\log N}\,|\mathrm{C}_\psi(\theta,\phi;x)| = \sqrt{\frac{\log N}{2\mathrm{KL}\,(\theta,\phi;x)}}\mathrm{SNR}\left[\partial_\psi \log\left(\frac{q_\phi(f(\varepsilon,\phi;x)|x)}{p_\theta(f(\varepsilon,\phi;x)|x)}\right)\right]$$

Hence, (43) and (45) are conditions on $\partial_\psi \log\left(q_\phi(f(\varepsilon,\phi;x)|x)/p_\theta(f(\varepsilon,\phi;x)|x)\right)$, which is an estimator of $\partial_\psi \mathrm{KL}\left(\theta,\phi;x\right)$. As a result, the assumptions made in Theorem 5 not only tie N to $\mathrm{KL}\left(\theta,\phi;x\right)$, but also to an estimator of $\partial_\psi \mathrm{KL}\left(\theta,\phi;x\right)$: they quantify how the mismatch between the target density and the variational one evolves as a function of N.

Interpreting (44) and (46). The asymptotic result (44) demonstrates that the optimization trajectory does not change significantly by increasing N beyond $N=1$ in the REP gradient estimator (14), since $\mathrm{sSNR}^{(\alpha,\mathrm{REP})}_{\psi,N}(\theta,\phi;x)$ is bounded as N grows (see Remark 8 in the appendix for a finer analysis). Furthermore, under the stronger condition (45), $\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{1,N}(\theta,\phi;x)]$ reverts to $\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{1,1}(\theta,\phi;x)]$ by (46).

Hence, once the variational approximation deteriorates beyond the threshold of (38), Theorem 5 suggests that utilizing $N>1$ in importance-weighted VI methods imposes a computational burden without improving over the $N=1$ ELBO baseline. However, a key takeaway message from Theorem 5 is also that importance-weighted VI methods continue to yield valid gradient estimators for learning the parameters of interest (θ,ϕ) even when the variational approximation deteriorates significantly. Remarkably, the collapse of the REP gradient estimator to the $N=1$ ELBO baseline in fact ensures that the optimization trajectory refines the variational approximation by minimizing $\mathrm{KL}\left(\theta,\phi;x\right)$ w.r.t. ϕ. Consequently, if the variational family is sufficiently expressive, $\mathrm{KL}\left(\theta,\phi;x\right)$ is expected to decrease during training, thus allowing a practical sample budget $N>1$ to become algorithmically viable again. We illustrate Theorem 5 by revisiting Example 1.

Example 3 *Consider the setting of Example 1 where $\theta-\phi=\epsilon\cdot\boldsymbol{u}_d$ with $\epsilon>0$. Then,* $\mathrm{KL}\left(\theta,\phi;x\right)=\frac{1}{2}\epsilon^2 d$, $|\mathrm{C}_\psi(\theta,\phi;x)|=1/\sqrt{d}$ *and,* (38) *and* (45) *respectively become*

$$\log N \ll \epsilon^2 d \quad \textit{and} \quad \log N \ll d\,. \tag{47}$$

Now applying Theorem 5: as $N,\epsilon^2 d\to\infty$, we have under (47) *that for all $\alpha\in[0,1)$ and $\psi\in\{\phi_1,\ldots,\phi_d\}$,*

$$\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{1,N}(\theta,\phi;x)]=\epsilon\,(1+o(1))+o(1). \tag{48}$$

We also have that $\mathbb{V}(\psi\text{-}g^{(\alpha,\mathrm{DREP})}_{1,1}(\theta,\phi;x))=0$. The derivation details for this example are deferred to Appendix D.6.

As $N,\epsilon^2 d\to\infty$, Example 3 states that the condition (47) amounts to assuming that the number of samples N does not grow quicker than exponentially with $\mathrm{e}^{\epsilon^2 d}$, in which case the REP gradient estimator with $N>1$ reverts to the REP gradient estimator with $N=1$ (ELBO). Bypassing (47) imposes a heavy computational budget as the latent space dimension d increases unless ϵ decreases to compensate, that is, the variational approximation closely matches the posterior density.

Note that $\mathbb{V}(\psi\text{-}g^{(\alpha,\mathrm{DREP})}_{1,1}(\theta,\phi;x))=0$ for this example. As it turns out, this property is also true if we consider a general Gaussian model with known positive definite covariance matrix (see Appendix D.5 for details). Such models represent the main models of interest satisfying (B1) and we argue that, although it is possible to study the DREP gradient estimator as $N,\mathrm{KL}\left(\theta,\phi;x\right)\to\infty$ for these models, it is better to skip this study for the sake of conciseness in order to avoid lengthy derivations that are only applied to the corner case where

$\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = 0$. Instead, we wrap up this subsection by putting our results into perspective with the existing literature and by discussing extensions of the assumption (B1).

Related work and broader impact. The degradation of importance weights in high-mismatch regimes is well-documented (Bengtsson et al., 2008). Notably, Agapiou et al. (2017); Chatterjee and Diaconis (2018) analyze the sample size N leading to accurate estimation of self-normalized importance sampling (SNIS) estimators. Under suitable concentration of the log-weight under the target distribution, Chatterjee and Diaconis (2018) demonstrate that having N that is exponential with the KL divergence between the proposal and target distributions is sufficient to control the error of SNIS estimators in probability for test functions that are square-integrable under the target distribution. Although this divergence generally differs from $\mathrm{KL}\,(\theta,\phi;x)$, they coincide under (B1), so that Chatterjee and Diaconis (2018) place the sample-size threshold for accurate SNIS estimation around $\log N = \mathrm{KL}\,(\theta,\phi;x)$.

Yet, existing results do not characterize the consequences of SNIS estimation failure for learning in importance-weighted VI. Specifically, they do not explain whether the resulting gradient estimators retain a useful SNR when the sample size is insufficient for accurate SNIS estimation. Our work shows that the REP gradient estimator (14) does not break down when $\log N = o(\mathrm{KL}\,(\theta,\phi;x))$ but rather collapses to the ELBO case $N = 1$ and minimizes $\mathrm{KL}\,(\theta,\phi;x)$ w.r.t. ϕ. Hence, $\mathrm{KL}\,(\theta,\phi;x)$ is expected to decrease during training, thus reducing the sample budget $N > 1$ needed to escape the weight collapse regime. In doing so, our analysis fills a known theoretical gap in the VI literature (Geffner and Domke, 2020; Vaitl et al., 2022).

Relatedly, Daudel et al. (2023) show that the VR-IWAE bound reverts to the ELBO in the exact Gaussian random field case as well as in an instance of approximately Gaussian random field case. Our work departs from Daudel et al. (2023) in two fundamental ways as (i) it shifts the focus from variational bounds to gradient estimators (ii) it considers the setting in which the variational approximation deteriorates and its deterioration manifests through the KL divergence $\mathrm{KL}\,(\theta,\phi;x)$. In particular, moving from variational bounds to their gradient estimators has strong consequences in terms of proof techniques since we need to differentiate (36) in the proofs of Proposition 1 and Theorem 5. To achieve this, we rely on a novel bound which is key to show the collapse of self-normalized weighted averages (see Proposition 7 in Appendix D.2). Due to the generality of Proposition 7, we anticipate that this bound will be of use beyond our current work.

Beyond the assumption (B1). A logical extension of (B1) is to assume that the first two moments of $\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)$ exist. We can then write without any loss of generality that

$$\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x) = \mathbb{E}_{\varepsilon\sim\tilde{q}}(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)) - \sqrt{\mathbb{V}_{\varepsilon\sim\tilde{q}}(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))}\,\tilde{S}(\varepsilon,\theta,\phi,\phi';x), \quad (49)$$

where $(\tilde{S}(\varepsilon,\theta,\phi,\phi';x))_{(\theta,\phi,\phi')\in\Theta\times\Phi^2}$ is a centered and normalized random field. Under adequate conditions, the centered and normalized log likelihood ratio $\tilde{S}(\varepsilon,\theta,\phi,\phi';x)$ is expected to be well approximated by a Gaussian random field. We conjecture that the gradient estimators $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ and $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)$ will revert to the case $N = 1$ under conditions akin to (and possibly stronger than) the ones from Theorem 5.

Adapting our results to this more general case in order to capture the behavior of both the REP and DREP gradient estimators requires further intricate derivations, which are left

for future work. Indeed, our proofs build on the property that $S(\varepsilon, \theta, \phi; x)$ and its derivative $\partial_\psi S(\varepsilon, \theta, \phi; x)$ are two independent and centered Gaussian processes under (B1) (see ($\mathrm{P}^{\mathrm{REP}}$) in Appendix D.3); the property ($\mathrm{P}^{\mathrm{REP}}$) does not hold anymore when we move away from (B1) to consider settings of the form (49) where $\tilde{S}(\varepsilon, \theta, \phi, \phi'; x)$ is well approximated by a Gaussian random field under adequate conditions.

3.5 Discussion and practical guidance

Our approaches in Sections 3.3 and 3.4 are complementary approaches that characterize different behaviors/phases occurring in importance-weighted VI algorithms.

On the one hand, Section 3.3 captures the setting where the variational approximation matches the posterior density reasonably well, in which case the benefits of importance-weighted VI methods are fully leveraged. For this setting, increasing N leads to a variance reduction phenomenon and SNRs that scale with $\sqrt{N}$ (except for edge cases), with the tuning of α leading to further variance reduction. Furthermore, the gradient estimators $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ and $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ point on average in the direction that is expected to improve on ELBO maximization.

On the other hand, Section 3.4 sheds light on the behavior of importance-weighted VI algorithms in the challenging regime where the variational density is not suitably close to the posterior density in terms of KL divergence. While a weight collapse occurs when $\log N = o(\mathrm{KL}\,(\theta, \phi; x))$ where $\mathrm{KL}\,(\theta, \phi; x)$ can be large, this regime actively seeks to reduce $\mathrm{KL}\,(\theta, \phi; x)$ by improving the quality of the variational approximation.

These behaviors align with the two training phases hypothesized by Vaitl et al. (2022): Section 3.4 describes the initial training phase of importance-weighted VI algorithms and Section 3.3 corresponds to the final training phase near convergence. While Vaitl et al. (2022) treat these phases as decoupled, our analysis ties them together. Provided that the variational family is sufficiently expressive, the initial training phase indeed naturally drives the optimization trajectory toward the final one by lowering the exponential threshold $\mathrm{e}^{\mathrm{KL}(\theta,\phi;x)}$. From there, we can derive the following practical considerations:

Budget allocation. Given a fixed total computational budget of SN (where S is the number of gradient steps), our results lead to a two-phase allocation strategy:

- Initial training phase (high mismatch): Setting $N > 1$ imposes a strict computational penalty without altering the optimization trajectory. Therefore, the practitioner should set $N = 1$ to maximize S, allowing the model to traverse the parameter space as rapidly as possible to reduce $\mathrm{KL}\,(\theta, \phi; x)$.
- Final training phase (low mismatch): When the variational approximation is sufficiently close to the target in terms of KL divergence, the practitioner should increase the budget to $N > 1$.

On-the-fly tuning of N and α. Although $\mathrm{KL}\,(\theta, \phi; x)$ cannot be calculated on-the-fly, practitioners can cheaply monitor the Effective Sample Size (ESS) or the variance of the log-weights as a proxy. For instance, if the monitored ESS collapses toward 1, it signals that the gradient estimators are suffering from a weight collapse. In this case, the practitioner can either automatically drop N to 1 to save computational effort, or dynamically tune

α toward a value that balances the bias-variance tradeoff we identified in this paper. A detailed investigation of these tuning mechanisms is left for future work.

High dimensions. Although high dimensionality typically initializes the variational approximation into the initial training phase, the fundamental quantity in importance-weighted VI is the mismatch between the variational approximation and the posterior density, rather than the latent dimension d itself. Our analysis suggests that, should the variational family be flexible enough, the variational approximation is improved during the optimisation procedure until it enters the first regime even in high dimensions. As we shall see now in our numerical experiments, one can perfectly emulate this high-dimensional behavior even within a moderate latent dimension d, so long as the variational approximation is sufficiently poor.

4. Numerical experiments

In this section, we provide empirical evidence supporting our theoretical claims.

4.1 Gaussian example

We consider the Gaussian example from Example 1, for which Theorem 3 predicts the following asymptotic behavior of $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ for all $\alpha \in [0,1)$: as $N \to \infty$,

$$\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{\sqrt{N}\epsilon\left(\alpha\exp\left(\frac{-(1-\alpha)^2 d\epsilon^2}{2}\right)+\frac{1-\alpha}{N}\exp\left(\frac{(1-\alpha)^2 d\epsilon^2}{2}\right)\right)}{\sqrt{1+(1-\alpha)^2\epsilon^2}}(1+o(1)).$$

To check the validity of this asymptotic result, we first study the case $\alpha \in (0,1)$. Consequently, we investigate whether the asymptotic behavior as $N \to \infty$

$$\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{\sqrt{N}\epsilon\alpha\exp\left(\frac{-(1-\alpha)^2 d\epsilon^2}{2}\right)}{\sqrt{1+(1-\alpha)^2\epsilon^2}}(1+o(1)), \quad \alpha \in (0,1) \tag{50}$$

matches with the observed $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ as N increases (note that only the leading order term remains in (50)). To this end, we let $\epsilon \in \{0.2, 1., 2\}$, $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $d \in \{10, 100\}$, $N \in \{2^j, j = 1 \dots 14\}$ and ϕ_k be a random coordinate in $(\phi_1 \dots \phi_d)$. The results are shown in Figure 1.

Observe first that, in the favourable setting of a low dimension d with a small perturbation ϵ near the optimum (that is $(d,\epsilon) = (10, 0.2)$), the asymptotic behavior predicted by (50) for each α ends up matching the corresponding observed SNR as N increases. In particular, the bias-variance tradeoff behavior highlighted in our theoretical analysis of Example 1 is empirically confirmed for N large enough, with the bias and variance of the REP gradient estimator increasing and decreasing with α, respectively. In addition, the higher the value of α (and the lower the value of ϵ), the smaller N needs to be before the observed SNR and the prediction made by (50) coincide. This follows from the fact that the second-order term in the numerator of the SNR when $\alpha \in (0,1)$ and $\epsilon > 0$ is expected to vanish quicker with N as α increases and ϵ decreases.

As d increases, $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ should revert to $\mathrm{SNR}[\phi_k\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ according to Example 3 unless the number of samples N is very large or ϵ is small, meaning

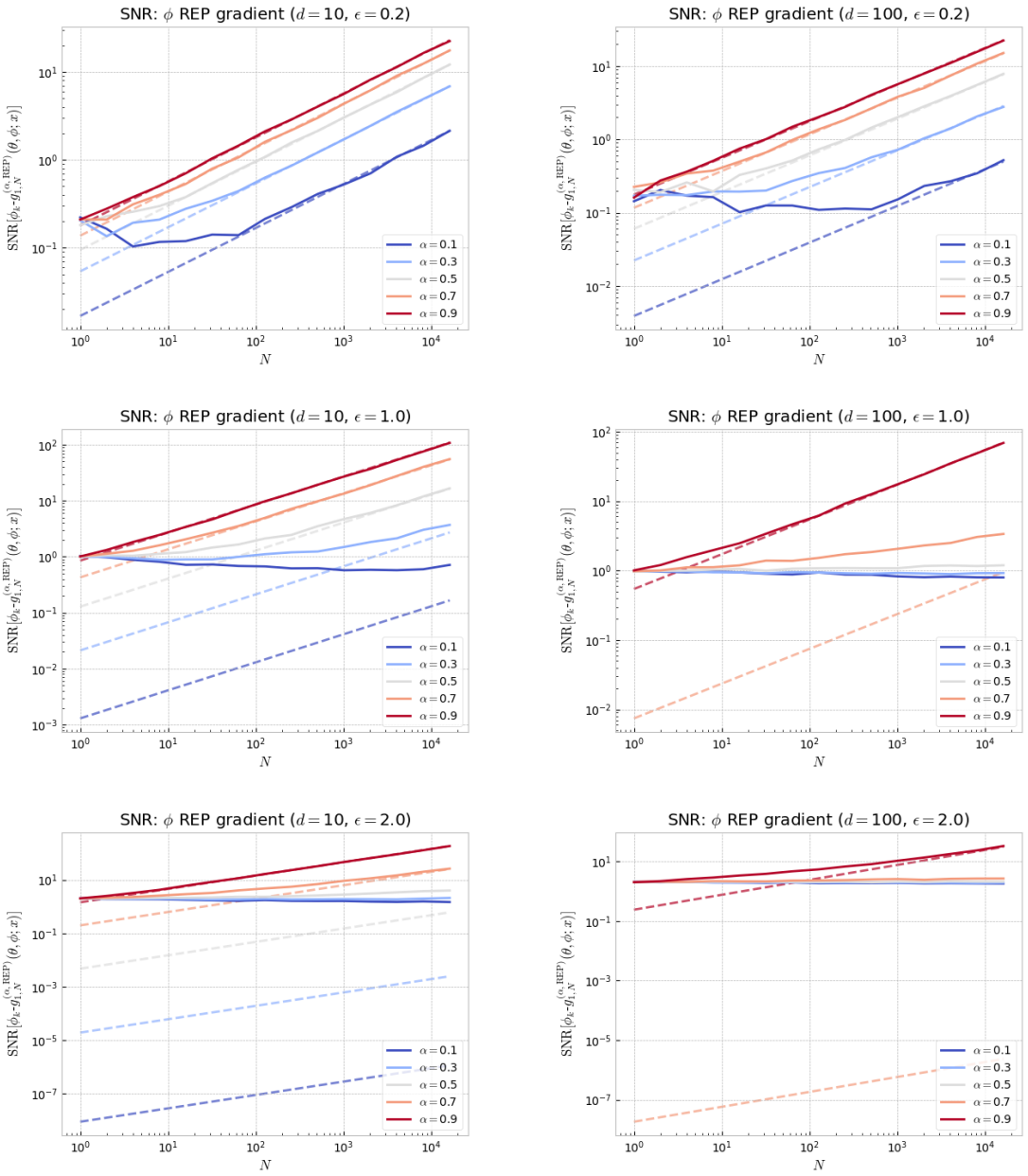


Figure 1: Plotted is $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N, for varying values of (α, d, ϵ) and a random coordinate ϕ_k. The solid lines correspond to $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$, the dashed lines correspond to predictions of the form (50).

that the asymptotic behavior (50) will not reflect the observed SNRs anymore. This is indeed what we observe in Figure 1 ($N = 10^0$ corresponds to $\mathrm{SNR}[\phi_k\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \epsilon$), with the weight collapse occurring quicker the higher the value of ϵ, that is the further the

variational approximation is from the target posterior density. Interpreting the different values of ϵ as different stages of the training procedure, Example 3 sheds light on the role the flexibility of the variational family plays in the success of the REP gradient estimator: the advantages of using the REP gradient estimator with $N > 1$ instead of with $N = 1$ arise after an initial training period ensuring that the quality of the variational approximation increases.

Additional results for the DREP gradient estimator and for the IWAE case ($\alpha = 0$) leading to similar conclusions are provided in Appendix E.1 for the sake of completeness.

4.2 Linear Gaussian example

We are next interested in the linear Gaussian example from Example 2. The data set $\mathcal{D} = \{x_1, ..., x_T\}$ is generated by sampling $T = 1024$ datapoints from $\mathcal{N}(0, 2\boldsymbol{I}_d)$. We let $\theta = \theta^\star + 2\epsilon$, $A = A^\star$ and $\tilde{b} = \tilde{b}^\star + 2\epsilon$ with $\epsilon > 0$, $\theta^\star = T^{-1}\sum_{t=1}^T x_t$, $A^\star = \frac{1}{2}\boldsymbol{I}_d$ and $\tilde{b}^\star = \frac{1}{2}\theta^\star$, so that $Ax + \tilde{b} = \frac{1}{2}(x+\theta) + \epsilon$. For all $\alpha \in [0, 1)$ and all $k = 1 \dots d$, Example 2 thus states that Theorems 3 and 4 predict the asymptotic behavior as $N \to \infty$ of $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)]$ respectively.

Let us now check the validity of the asymptotic results obtained in Example 2. Focusing first on the learning of $\tilde{b}$ when $\alpha \in (0, 1)$, this boils down to checking whether as N increases the observed $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)]$ match with the asymptotic behaviors: as $N \to \infty$,

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \frac{\sqrt{N}\frac{3\epsilon\alpha}{4-\alpha}(1 + o(1))}{\frac{(4-\alpha)^{d/2}}{(15-6\alpha)^{d/4}} \exp\left(\frac{12(1-\alpha)^2}{(4-\alpha)(5-2\alpha)} d\epsilon^2\right) \sqrt{\frac{2}{5-2\alpha} + \left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2}} \quad (51)$$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)] = 4\alpha^{-1}\ \mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)](1 + o(1)) \quad (52)$$

(note that once more only the leading order term remains in (51)). Now let $\epsilon \in \{0.2, 1\}$, $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $d \in \{10, 100\}$, $N \in \{2^j, j = 1 \dots 14\}$ and $\tilde{b}_k$ be a random coordinate in $(\tilde{b}_1, \dots, \tilde{b}_d)$. The results are shown in Figures 2 and 3.

Starting with Figure 2, we see that the conclusions drawn in Section 4.1 for the REP gradient estimator regarding the validity of the asymptotic behaviors predicted by Theorems 3 and 4 when $\alpha \in (0, 1)$ apply here too. Moving on to Figure 3, we have that these conclusions are also valid for the DREP gradient estimator, which notably confirms empirically that the DREP gradient estimator outperforms the REP one by a $4\alpha^{-1}$ factor in terms of SNR for the setting considered here. Additional plots in which we replace $\tilde{b}_k$ by θ_k and/or we set $\alpha = 0$ are available in Appendix E.2. They further validate our conclusions regarding the empirical validity of Theorems 3 and 4.

As d and ϵ increase, $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)]$ reverts to $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)]$ in Figure 2 unless N is increasingly large and we observe a similar behavior for the DREP gradient estimator in Figure 3, which is coherent with our conjecture that an increasing mismatch between the posterior and variational densities takes precedence over the choice of the importance-weighted variational bound. This effect can to some extent be mitigated by increasing α or N, although we argue that improving on the quality of the variational

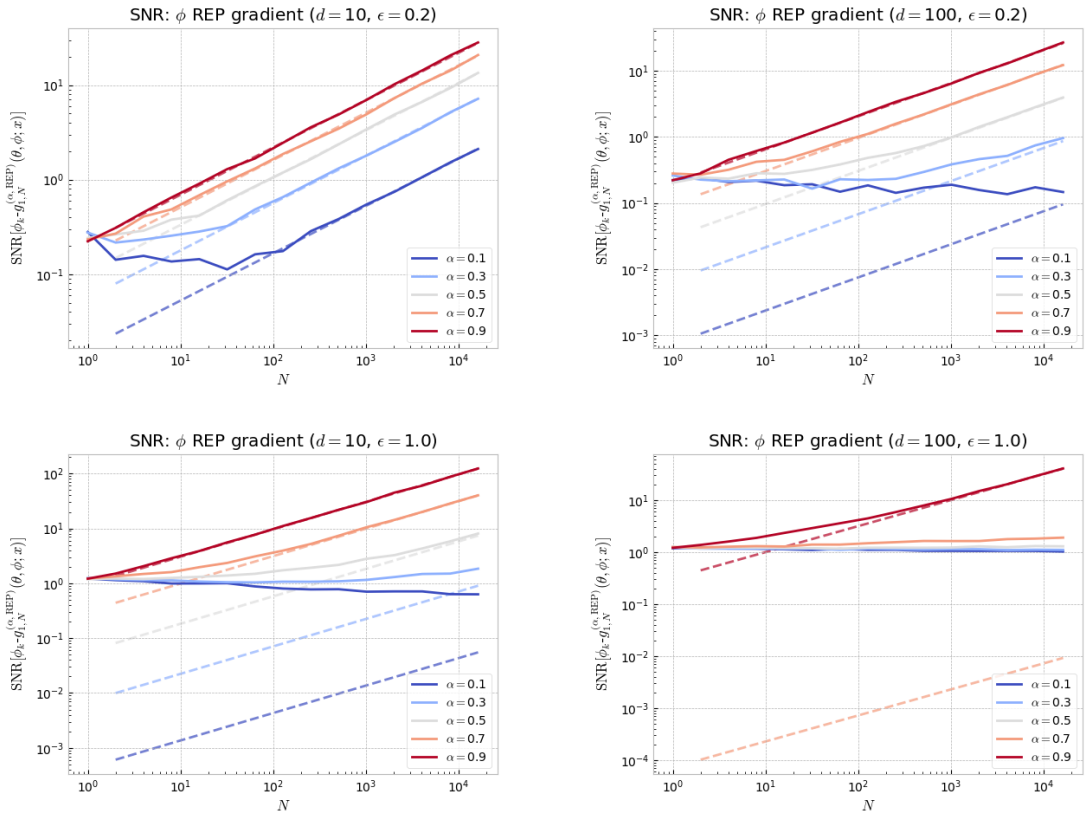


Figure 2: Plotted is $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$, the dashed lines correspond to predictions of the form (51).

approximation through the learning of ϕ is a more direct approach to fully leverage the advantages of importance-weighted VI gradient estimators.

4.3 Variational auto-encoder

We consider the case of a variational auto-encoder (VAE) model (Rezende et al., 2014; Kingma and Welling, 2014) 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.

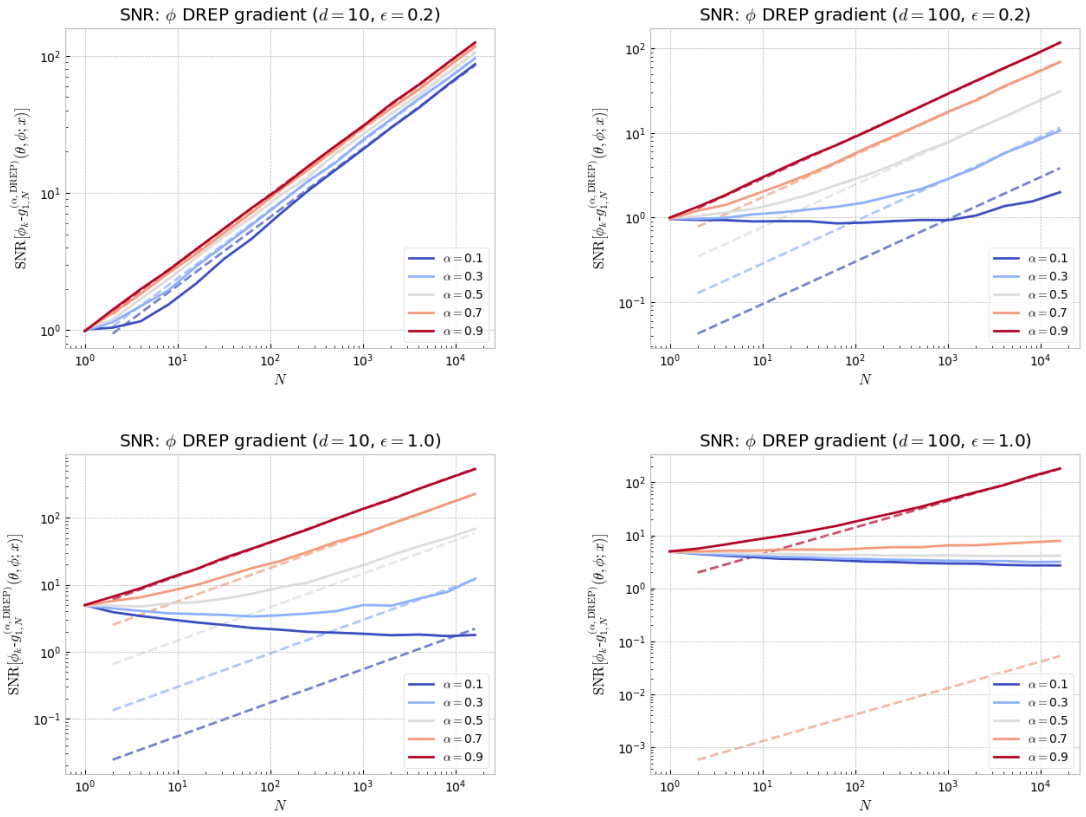


Figure 3: Plotted is $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)]$, the dashed lines correspond to predictions of the form (52).

Let us then check the validity of our asymptotic results for this more involved real-data example. Focusing on the case $\alpha \in (0,1)$, Theorem 3 predicts that: as $N \to \infty$,

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{N}\,\frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) + o(1)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)} + o\left(1\right)}. \tag{53}$$

Unlike the two previous examples, we do not have a closed form expression for $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ here and we will resort to using approximations instead. Letting $N' \in \mathbb{N}^\star$, we approximate $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ by $\psi\text{-}g_{1,N'}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ following Theorem 2 and we approximate $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ by

$$N' \sum_{i=1}^{N'} \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon_{1,i};x)^{1-\alpha}}{\sum_{\ell=1}^{N'} \tilde{w}_{\theta,\phi,\phi}(\varepsilon_{1,\ell};x)^{1-\alpha}} \left[\partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon_{1,i};x)\right) - \psi\text{-}g_{1,N'}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right] \right)^2$$

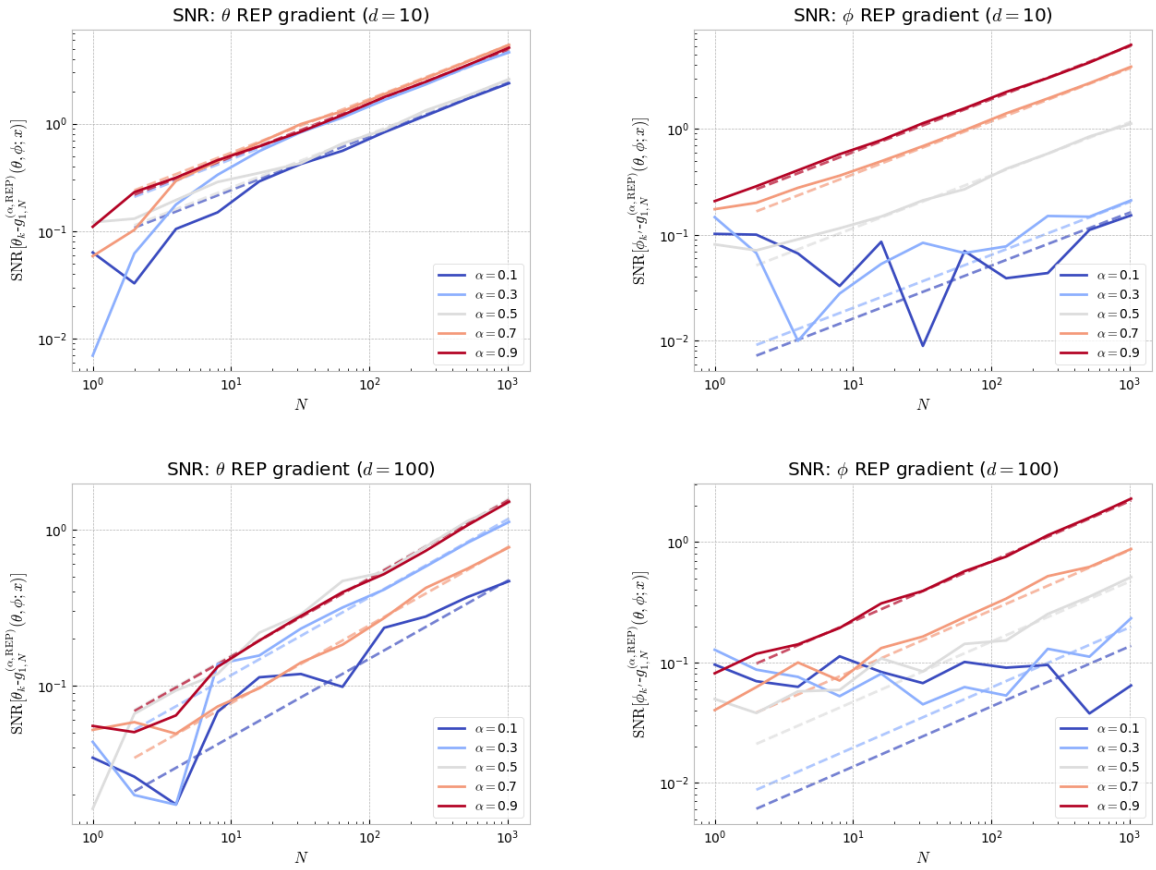


Figure 4: Plotted is $\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the VAE example described in Section 4.3 as a function of N, for varying values of (α, d), $\psi \in \{\theta_k, \phi_{k'}\}$ and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$, the dashed lines correspond to predictions of the form (53) in which $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ have also been approximated.

with $\varepsilon_{1,1},\ldots,\varepsilon_{1,N'}$ being i.i.d. samples generated from $\tilde{q}$. Letting $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $d \in \{10, 100\}$, $N \in \{2^j, j = 1\ldots 10\}$, $N' = 2^{10}$, we then obtain Figure 4 in which $\psi \in \{\theta_k, \phi_{k'}\}$ with k and k' that were randomly selected.

Even though we had to resort to an additional layer of approximation here, our conclusions remain similar to the ones obtained in Sections 4.1 and 4.2, in the sense that the behavior predicted in (53) matches the observed rates for N sufficiently large and we can observe the impact of increasing the mismatch between the variational and posterior densities via d. The theoretical studies we have carried out in this paper thus provide a useful framework to understand importance-weighted VI methods in real-data scenarios.

5. Conclusion

We proposed two complementary SNR analyses which advance the understanding of importance-weighted VI methods by casting a new light on the innerworkings of the ELBO, IWAE, VR, and VR-IWAE bound methodologies at the gradient level. Our proof techniques rely on novel theoretical tools that apply more broadly within importance weighting and are of independent interest. The validity of our analyses is demonstrated over several toy and real-data examples and we now mention three potential directions of research to extend our work.

A first direction of research, also outlined at the end of Section 3.4, is to investigate the extension of our results for the Gaussian random field case to the approximate Gaussian random field one in order to encompass more general settings of interest. Another possibility is to see how our results can be further exploited to tune α and/or to choose the variational family. Finally, our work showcases that there is much to gain theoretically and empirically from investigating the connections between sampling and importance-weighted VI. One may seek to delve deeper into those connections with importance weighting and related Monte Carlo approximations.

Acknowledgments

We are grateful to the Associate Editor and to the reviewers for their insightful comments which helped us improve the manuscript. We also want to thank Pierre E. Jacob and El Mahdi Khribch for stimulating discussions that encouraged us to pursue weaker positive moment conditions for (20) in Theorem 1. These exchanges prompted a re-examination of our proof strategy, which ultimately allowed us to obtain the refined results presented in Theorem 1.

Appendix A. Interchanging Derivative and Expectation Signs

Throughout the paper, we often come across expectations of the form $\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(g(\psi,\varepsilon)\right)$ for some real valued parameter ψ in a parameter space Ψ, which we want to differentiate w.r.t ψ. More specifically, using that $\tilde{q}$ does not depend on ψ, we want compute the derivative by interchanging the derivative and the expectation signs, that is

$$\partial_\psi \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(g(\psi,\varepsilon)\right) = \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\partial_\psi g(\psi,\varepsilon)\right), \quad \psi\in\Psi\,. \tag{54}$$

General conditions to make the identity (54) valid are well known (see, e.g., L'Ecuyer (1995)) and specifying them can be overlooked at first reading as it sometimes burdens the technical content with lengthy assumptions whose sole aim is to make interchanges of derivatives and expectations well justified. Nevertheless, we specify hereafter sufficient conditions for all the interchanges of derivatives and expectations to be valid in our main results.

To this end, we will repeatedly use the notation $\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)$ defined in (24) which, for every ε and x, is differentiable with respect to (θ,ϕ,ϕ') on $\Theta\times\Phi^2$ following the differentiability conditions already assumed on $\theta\mapsto p_\theta(x)$, $(z,\theta)\mapsto p_\theta(z|x)$, $(z,\phi)\mapsto q_\phi(z|x)$ and $\phi\mapsto f(\varepsilon,\phi;x)$ in the beginning of Section 3. Furthermore, ψ will denote a component of the $\mathbb{R}^{a+b}$-valued variable $(\theta,\phi)=(\theta_1,\ldots,\theta_a,\phi_1,\ldots,\phi_b)$ and we will say that $\mathcal{V}$ is a ψ-neighborhood of $(\theta,\phi)\in\mathbb{R}^{a+b}$ if there exists $r>0$ such that for all $\psi'\in(\psi-r,\psi+r)$, the vector obtained by replacing ψ by ψ' in (θ,ϕ) belongs to $\mathcal{V}$.

The remaining of Appendix A is then concerned with providing sufficient conditions to ensure that the interchanges of derivatives and expectations necessary to establish the results of Section 3.3 and Section 3.4 are valid.

A.1 Interchanging Derivative and Expectation Signs in Section 3.3

The first assumption below is concerned with the REP gradient estimator.

$(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ All $(\theta, \phi) \in \Theta \times \Phi$ admit a ψ-neighborhood $\mathcal{V} \subset \Theta \times \Phi$ such that, for $\ell = 0, 1, 2$,

$$\mathbb{E}_{\varepsilon \sim \tilde{q}} \left(\sup_{(\theta', \phi') \in \mathcal{V}} \left| \tilde{w}_{\theta', \phi', \phi'}(\varepsilon; x)^{\ell(1-\alpha)} \, \partial_{\psi'} \left(\log \tilde{w}_{\theta', \phi', \phi'}(\varepsilon; x) \right) \right| \right) < \infty \, . \tag{55}$$

We have the following result which will cover all the interchanges of derivatives and expectations that are necessary for our study of the REP gradient estimator in Section 3.3.

Proposition 2 *Under* $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$*, we have for all* $(\theta, \phi) \in \Theta \times \Phi$*,*

$$\partial_\psi \mathbb{E}_{\varepsilon \sim \tilde{q}} \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha} \right) = \mathbb{E} \left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha} \right) \right) \, , \tag{56}$$

$$\partial_\psi \mathbb{E}_{\varepsilon \sim \tilde{q}} \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{2(1-\alpha)} \right) = \mathbb{E} \left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{2(1-\alpha)} \right) \right) \, . \tag{57}$$

Moreover, for all $N \geq 1$*,*

$$\partial_\psi \mathbb{E}_{\varepsilon_j \overset{i.i.d.}{\sim} \tilde{q}} \left(\log \left(\sum_{j=1}^{N} \tilde{w}_{\theta,\phi,\phi}(\varepsilon_j; x)^{1-\alpha} \right) \right) = \mathbb{E}_{\varepsilon_j \overset{i.i.d.}{\sim} \tilde{q}} \left(\partial_\psi \log \left(\sum_{j=1}^{N} \tilde{w}_{\theta,\phi,\phi}(\varepsilon_j; x)^{1-\alpha} \right) \right) \, . \tag{58}$$

Proof Suppose that $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ holds. Observe that, for $\ell = 1, 2$, by (24), we have

$$\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{\ell(1-\alpha)} \right) = \ell(1-\alpha) \, w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{\ell(1-\alpha)} \, \partial_\psi \left(\log w_{\theta,\phi}(f(\varepsilon, \phi; x); x) \right) \, .$$

Thus, by the usual dominated convergence argument for interchanging the expectation and derivative signs, the condition in (55) for $\ell = 1, 2$ leads to (56) and (57), respectively.

Let now $N \geq 1$. To get (58), by the same dominated convergence argument, it is sufficient to show that all $(\theta, \phi) \in \Theta \times \Phi$ admit a ψ-neighborhood $\mathcal{V} \subset \Theta \times \Phi$ such that

$$\mathbb{E}_{\varepsilon_j \overset{\text{i.i.d.}}{\sim} \tilde{q}} \left(\sup_{(\theta', \phi') \in \mathcal{V}} \left| \partial_{\psi'} \log \left(\sum_{j=1}^{N} \tilde{w}_{\theta', \phi', \phi'}(\varepsilon_j; x)^{1-\alpha} \right) \right| \right) < \infty \, .$$

Differentiating and using that the sup of a sum is bounded by the sum of the sup from above, the latter condition is implied by having that for all $j = 1, \ldots, N$,

$$\mathbb{E}_{\varepsilon_j \overset{\text{i.i.d.}}{\sim} \tilde{q}} \left(\sup_{(\theta', \phi') \in \mathcal{V}} \frac{\left| \partial_{\psi'} \left(\tilde{w}_{\theta', \phi', \phi'}(\varepsilon_j; x)^{1-\alpha} \right) \right|}{\sum_{k=1}^{N} \tilde{w}_{\theta', \phi', \phi'}(\varepsilon_k; x)^{1-\alpha}} \right) < \infty \, .$$

Using that $\tilde{w}_{\theta', \phi', \phi'}(\varepsilon_k; x) \geq 0$ for all $k \neq j$, this ratio is bounded from above by

$$\frac{\left| \partial_\psi \left(\tilde{w}_{\theta', \phi', \phi'}(\varepsilon_j; x)^{1-\alpha} \right) \right|}{\tilde{w}_{\theta', \phi', \phi'}(\varepsilon_j; x)^{1-\alpha}} = (1-\alpha) \, \left| \partial_\psi \left(\log \tilde{w}_{\theta', \phi', \phi'}(\varepsilon_j; x) \right) \right| \, .$$

Hence assuming (55) with $\ell = 0$ ensures that the previous condition holds and we have concluded the proof of (58). ■

The next assumption is concerned with the DREP gradient estimator and is only required for a component ψ of the $\mathbb{R}^b$-valued variable $\phi = (\phi_1, \ldots, \phi_b)$,

($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$) All (θ, ϕ) in $\Theta \times \Phi$ admit a ψ-neighborhood $\mathcal{V} \subset \Theta \times \Phi$ such that:

(i) If $\alpha \in (0, 1)$, we have

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi',\phi'}(\varepsilon;x)^{1-\alpha}\right)\right|\right) < \infty\,. \tag{59}$$

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right|\right) < \infty\,. \tag{60}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{1-\alpha}\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\log q_{\phi'}(Z|x)\right)\,\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right|\right) < \infty\,. \tag{61}$$

(ii) If $\alpha = 0$, we have

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi',\phi'}(\varepsilon;x)^{2}\right)\right|\right) < \infty\,, \tag{62}$$

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2}\right)\right|\right) < \infty\,, \tag{63}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{2}\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\log q_{\phi'}(Z|x)\right)\,\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right|\right) < \infty\,. \tag{64}$$

We have the following result which will cover all the interchanges of derivatives and expectations that are necessary for our study of the DREP gradient estimator in Section 3.3.

Proposition 3 *Let ψ denote a component of the $\mathbb{R}^b$-valued variable $\phi = (\phi_1, \ldots, \phi_b)$. Under ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$), we have the following assertions for all (θ, ϕ) in $\Theta \times \Phi$.*

(i) If $\alpha \in (0, 1)$, we have

$$\left[\partial_{\psi'}\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi} = \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi}\right), \tag{65}$$

$$\partial_{\psi}\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) = \alpha\,\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi}\right). \tag{66}$$

(ii) If $\alpha = 0$, we have

$$\partial_{\psi}\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{2}\right) = -\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2}\right)\right]\Big|_{\phi'=\phi}\right). \tag{67}$$

Proof First observe that (67) in Assertion (ii) corresponds to (66) in Assertion (i) with α replaced by -1. In addition, the case $\alpha = 0$ in Assumption ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$) corresponds to

the case $\alpha \in (0,1)$ with α also replaced by -1. We thus only have to prove that for any $\alpha \in (0,1) \cup \{-1\}$, Conditions (59)–(61) imply the identities (65) and (66).

We will in fact prove that more generally: for all $\alpha \in \mathbb{R}$, Conditions (59)–(61) imply the identities (65) and (66). The case $\alpha = 1$ is immediate so let us pick $\alpha \in \mathbb{R} \setminus \{1\}$ in the following and assume that (59)–(61) hold. Note first that (65) is a simple interchange of the derivative and expectation signs which holds by dominated convergence using (60). To then prove (66), observe that by (59), we can interchange the derivative and expectation signs on the left-hand side of (66), that is: for all $\alpha \in (0,1)$,

$$\partial_\psi \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha} \right) = \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha} \right) \right).$$

Since $\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha} \right) = \left[\partial_{\psi'} \left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha} \right) + \partial_{\psi'} \left(\tilde{w}_{\theta,\phi',\phi}(\varepsilon;x)^{1-\alpha} \right) \right]\big|_{\phi'=\phi}$, to obtain (66) it remains to prove that

$$\frac{1}{\alpha-1} \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\left[\partial_{\psi'} \left(\tilde{w}_{\theta,\phi',\phi}(\varepsilon;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} \right) = \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\left[\partial_{\psi'} \left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} \right) . \quad (68)$$

Observe then that, using (24), it holds that

$$\frac{1}{\alpha-1} \left[\partial_{\psi'} \left(\tilde{w}_{\theta,\phi',\phi}(\varepsilon;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} = \left[w_{\theta,\phi}(z;x)^{1-\alpha} \frac{\partial_\psi q_\phi(z)}{q_\phi(z|x)} \right]\Big|_{z=f(\varepsilon,\phi;x)} .$$

Hence the left-hand side of (68) reads

$$\begin{aligned} \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\left[w_{\theta,\phi}(z;x)^{1-\alpha} \frac{\partial_\psi q_\phi(z|x)}{q_\phi(z|x)} \right]\Big|_{z=f(\varepsilon,\phi;x)} \right) &= \mathbb{E}_{Z\sim q_\phi(\cdot|x)} \left(w_{\theta,\phi}(Z;x)^{1-\alpha} \frac{\partial_\psi q_\phi(Z|x)}{q_\phi(Z|x)} \right) \\ &= \int w_{\theta,\phi}(Z;x)^{1-\alpha} \, \partial_\psi q_\phi(Z|x) \, \nu(\mathrm{d}z) \end{aligned}$$

and the proof of (68) will be concluded if we can prove that

$$\begin{aligned} \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\left[\partial_{\psi'} \left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} \right) &\overset{\text{(a)}}{=} \left[\partial_{\psi'} \mathbb{E}_{\varepsilon\sim\tilde{q}} \left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} \\ &\overset{\text{(b)}}{=} \left[\partial_{\psi'} \mathbb{E}_{Z\sim q_{\phi'}(\cdot|x)} \left(w_{\theta,\phi}(Z;x)^{1-\alpha} \right) \right]\Big|_{\phi'=\phi} \\ &\overset{\text{(c)}}{=} \int w_{\theta,\phi}(Z;x)^{1-\alpha} \, \partial_\psi q_\phi(Z|x) \, \nu(\mathrm{d}z). \end{aligned}$$

The equality (a) follows from interchanging the integral and the derivative signs again thanks to (60) and the equality (b) follows from the reparameterization trick. As for the equality (c), it is obtained by interchanging the integral and the derivative thanks to the fact that for a ψ-neighborhood $\mathcal{V}$ of (θ,ϕ),

$$\int w_{\theta,\phi}(z;x)^{1-\alpha} \sup_{(\theta,\phi')\in\mathcal{V}} \left| \partial_{\psi'} q_{\phi'}(z|x) \right| \, \nu(\mathrm{d}z) < \infty ,$$

which holds using (61) and that

$$\sup_{(\theta,\phi')\in\mathcal{V}} \left| \partial_{\psi'} q_{\phi'}(Z|x) \right| = \sup_{(\theta,\phi')\in\mathcal{V}} \left| \left(\partial_{\psi'} \log q_{\phi'}(Z|x) \right) \frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)} \right| q_\phi(Z|x) .$$

The proof of Proposition 3 is thus concluded. ■

The conclusions from Proposition 3 will be used to express the constants in the asymptotic behavior of the SNR of the DREP estimator, see the proof of Theorem 4.

A.2 Interchanging Derivative and Expectation Signs in Section 3.4

We now examine the specific case where the Gaussian assumption (B1) holds. In this special case, we rely on the following assumption for the REP gradient estimator.

($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$) All (θ, ϕ) in $\Theta \times \Phi$ admit a ψ-neighborhood $\mathcal{V} \subset \Theta \times \Phi$ such that

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}}\left(\partial_{\psi'}\log\tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)^2\right)<\infty\,, \tag{69}$$

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}}\left(\log\tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)^2\right)<\infty\,. \tag{70}$$

We have the following result which will cover all the interchanges of derivatives and expectations that are necessary for our study of the REP gradient estimator in Proposition 1.

Proposition 4 *Let $\varepsilon \sim \tilde{q}$. Under* (B1) *and* ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$)*, we have that: for all $(\theta, \phi) \in \Theta \times \Phi$,*

$$\partial_\psi \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)=\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\partial_\psi\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right) \tag{71}$$

$$\partial_\psi \mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)^2\right)=\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\partial_\psi\left(\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)^2\right)\right) \tag{72}$$

and $\left(\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x), \partial_\psi\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a Gaussian process.

Proof Using the dominated convergence theorem, (69) in ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$) implies (71) but also that $\partial_\psi \log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)$ can be seen as the L^2 limit of a linear combination of the process $\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$.

It follows by (36) that $\left(\left(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x), \partial_\psi\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a Gaussian process. As for (72), it follows from the dominated convergence theorem and

$$\begin{aligned}&\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\left(\log\tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)^2\right)\right|\right)\\ &\leq 2\left(\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}}\left(\partial_{\psi'}\log\tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)^2\right)\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}}\left(\log\tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)^2\right)\right)^{\frac{1}{2}}<\infty\,,\end{aligned}$$

where the first inequality uses the Cauchy-Schwarz inequality and the second is a consequence of (69) and (70). ■

Appendix B. Deferred Proofs of Section 3.2

In this section, we let $(X_1, W_1), \dots, (X_N, W_N)$ be i.i.d. with the same distribution as a generic pair (X, W) valued in $\mathbb{R} \times \mathbb{R}_+$ and we set: for all $v \geq 0$,

$$\phi(v) = \mathbb{E}\left(\mathrm{e}^{-v\,W}\right) \quad \text{and} \quad \phi_1(v) = \mathbb{E}\left(X\,\mathrm{e}^{-v\,W}\right), \tag{73}$$

$$\overline{W}_N := \frac{1}{N}\sum_{i=1}^{N} W_i \quad \text{and} \quad \overline{X}_N := \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{74}$$

$$\overline{X^2}_N := \frac{1}{N}\sum_{i=1}^{N} X_i^2 \quad \text{and} \quad \overline{X}_N^{(i\neq j)} := \frac{2}{N(N-1)} \sum_{1\leq i<j\leq N} X_i X_j, \tag{75}$$

$$\overline{X^c}_N := \frac{1}{N}\sum_{i=1}^{N} X_i^c \quad \text{and} \quad \overline{X^c}_N^{(i\neq j)} := \frac{2}{N(N-1)} \sum_{1\leq i<j\leq N} X_i^c X_j^c, \tag{76}$$

where we set $X_i^c := X_i - \mathbb{E}(X)$.

B.1 Preliminary Lemmas

In the following, we provide two preliminary lemmas that gather useful properties of the functions ϕ and ϕ_1 defined in (73).

Lemma 2 (Properties of the Laplace transform ϕ) *The following assertions hold.*

(i) The function ϕ is continuous and bounded on $\mathbb{R}_+$.

(ii) Assume that $\mathbb{E}(W) = 1$. Then, for all $N \in \mathbb{N}^*$ and all $u \geq 0$, $\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \geq 0$.

(iii) Assume that $\mathbb{E}(W) = 1$ and $\mathbb{E}(W^2) < \infty$. Then, for all $N \in \mathbb{N}^*$ and all $u \in \left[0, \frac{N}{\mathbb{E}(W^2)}\right]$,

$$0 \leq N\left(\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u}\right) \leq \frac{\mathbb{E}(W^2)}{2} u^2 \mathrm{e}^{-u/2} . \tag{77}$$

Furthermore, for all $u \geq 0$,

$$\lim_{N\to\infty} N\left(\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u}\right) = \frac{\mathbb{V}(W)}{2}\, u^2\, \mathrm{e}^{-u} . \tag{78}$$

(iv) Assume that there exists $\eta > 0$ such that

$$\sup_{t>0}\left(t^\eta\, \mathbb{E}\left(\mathrm{e}^{-t\,W}\right)\right) < \infty . \tag{79}$$

Then, for any exponent $m > 0$, there exists $c, p > 0$ such that: for all $q \geq p$ and $v \geq 0$,

$$\left(\phi\left(\frac{v}{q}\right)\right)^q \leq 1 \wedge \left(c\, v^{-m}\right) .$$

Proof The assertion (i) is immediate since ϕ is valued in $[0,1]$ and is continuous by the dominated convergence theorem. In addition, (ii) follows from Jensen's inequality applied to the convex function $W \mapsto \mathrm{e}^{-uW/N}$ and paired up with the fact that $\mathbb{E}(W) = 1$ by assumption. We now move on to the proof of (iii) and (iv).

- **Proof of (iii).** First note that for all $x \geq 0$,

$$-\frac{x^2}{2}\left(1-\mathrm{e}^{-x}\right) \leq \mathrm{e}^{-x} - 1 + x - \frac{x^2}{2} \leq 0 \tag{80}$$

since for all $x \geq 0$: $1 - \mathrm{e}^{-t} \leq 1 - \mathrm{e}^{-x}$ for $0 \leq t \leq x$ and

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

Applying the second inequality in (80) with $x = vW$ and taking expectation yields:

$$\phi(v) \leq 1 - v + \frac{v^2}{2}\mathbb{E}(W^2), \quad v \geq 0.$$

We deduce from the inequality above and (ii) that: for all $u \geq 0$ and $N \in \mathbb{N}^*$,

$$0 \leq \left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \leq \mathrm{e}^{-u}\left(\mathrm{e}^{u+N\ln\left(1-\frac{u}{N}+\frac{u^2}{2N^2}\mathbb{E}(W^2)\right)} - 1\right).$$

Since $\ln(1+x) \leq x$, we then get that for all $u \geq 0$ and $N \in \mathbb{N}^*$,

$$0 \leq \left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \leq \mathrm{e}^{-u}\left(\mathrm{e}^{\frac{u^2}{2N}\mathbb{E}(W^2)} - 1\right) \leq e^{-u}\frac{u^2}{2N}\mathbb{E}\left(W^2\right)\mathrm{e}^{\frac{u^2}{2N}\mathbb{E}(W^2)}$$

where the last inequality follows from $\mathrm{e}^x - 1 \leq x\mathrm{e}^x$ for all $x \in \mathbb{R}$. Multiplying by N and restricting u to the set $\left[0, \frac{N}{\mathbb{E}(W^2)}\right]$ yields (77).

Now let $u \geq 0$ be given. Applying (80) with $x = u\,W/N$ and taking expectation, we get that for all $N \in \mathbb{N}^*$,

$$\left|\phi\left(\frac{u}{N}\right) - 1 + \frac{u}{N} - \frac{\mathbb{E}\left(W^2\right)}{2}\frac{u^2}{N^2}\right| \leq \frac{u^2}{2N^2}\,\mathbb{E}\left(W^2\left(1-\mathrm{e}^{-W\,u/N}\right)\right).$$

By dominated convergence, the expectation in the upper bound converges to 0 as $N \to \infty$. Hence we have that, as $N \to \infty$,

$$\ln\phi\left(\frac{u}{N}\right) = \ln\left(1 - \frac{u}{N} + \frac{\mathbb{E}\left(W^2\right)}{2}\frac{u^2}{N^2} + o\left(\frac{1}{N^2}\right)\right) = -\frac{u}{N} + \frac{\mathbb{V}\left(W\right)}{2}\frac{u^2}{N^2} + o\left(\frac{1}{N^2}\right),$$

where we used that $\ln(1+x) = x - x^2/2 + o(x^2)$ as $x \to 0$ and $\mathbb{E}\left(W^2\right) - 1 = \mathbb{V}\left(W\right)$. It follows that, as $N \to \infty$,

$$N\left(\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u}\right) = N\,\mathrm{e}^{-u}\left(\mathrm{e}^{N\ln\phi\left(\frac{u}{N}\right)+u} - 1\right) = \mathrm{e}^{-u}\,N\left(\mathrm{e}^{\frac{\mathbb{V}(W)}{2}\frac{u^2}{N}+o\left(\frac{1}{N}\right)} - 1\right),$$

which yields the given limit (78) using that $\mathrm{e}^x = 1 + x + o(x)$ as $x \to 0$.

• **Proof of (iv).** First note that for any $v \geq 0$, $p \mapsto \left(\phi\left(\frac{v}{p}\right)\right)^p$ is non-increasing as p increases (since for all $q \geq p$, $\left(\phi\left(\frac{v}{p}\right)\right)^p = \left(\phi\left(\frac{v}{q}\frac{q}{p}\right)\right)^p \geq \left(\phi\left(\frac{v}{q}\right)\right)^q$ by Jensen's inequality). Furthermore, by (79), we have: for any $p > 0$,

$$c_1(p) := \sup_{t>0} \left((t/p)^\eta \, \phi(t/p)\right) < \infty \, .$$

Let $m > 0$ and take $p = m/\eta$. Then we have for all $q \geq p$ and $v \geq 0$,

$$\left(\phi\left(\frac{v}{q}\right)\right)^q \leq \left(\phi\left(\frac{v}{p}\right)\right)^p \leq (v/p)^{-p\,\eta} \, \left(\sup_{t>0} (t/p)^\eta \, \phi\left(\frac{t}{p}\right)\right)^p = c \, v^{-m} \, ,$$

where we set $c = p^{p\eta} \, (c_1(p))^p$, hence the proof of (iv) is completed.

■

Lemma 3 (Properties of ϕ_1) *The following assertions hold.*

(i) Assume that $\mathbb{E}(|X|) < \infty$. Then, ϕ_1 is continuous and bounded on $\mathbb{R}_+$.

(ii) Assume that $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|\, W) < \infty$. Then, for all $u \geq 0$,

$$\sup_{N \geq 1} \left(N \left| \phi_1\left(\frac{u}{N}\right) + \mathbb{E}\left(XW\right) \, \frac{u}{N} \right| \right) \leq \mathbb{E}(|X|W) u \, ,$$
$$\lim_{N \to \infty} N \left| \phi_1\left(\frac{u}{N}\right) + \mathbb{E}\left(XW\right) \frac{u}{N} \right| = 0 \, .$$

(iii) Assume that $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|\, W^\delta) < \infty$ with $0 < \delta < 1$. Then, for all $u \geq 0$,

$$\sup_{N \geq 1} \left(N^\delta \left| \phi_1\left(\frac{u}{N}\right) \right| \right) \leq \mathbb{E}\left(|X|\, W^\delta\right) \, u^\delta \, ,$$
$$\lim_{N \to \infty} N^\delta \left| \phi_1\left(\frac{u}{N}\right) \right| = 0 \, .$$

Proof The assertion (i) is immediate since ϕ_1 is valued in $[-\mathbb{E}(X_-), \mathbb{E}(X_+)]$ and it is continuous by the dominated convergence theorem when $\mathbb{E}(|X|) < \infty$. As for (ii), first note that for all $x \geq 0$,

$$0 \leq \mathrm{e}^{-x} - 1 + x \leq x \left(1 - \mathrm{e}^{-x}\right) , \tag{81}$$

since for all $x \geq 0$, $1 - \mathrm{e}^{-t} \leq 1 - \mathrm{e}^{-x}$ for $0 \leq t \leq x$ and

$$\mathrm{e}^{-x} - 1 + x = \int_0^x \left(1 - \mathrm{e}^{-t}\right) \mathrm{d}t.$$

Using $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|\,W) < \infty$, we have, for all $u \geq 0$ and $N \geq 1$,

$$\begin{aligned}\left|\phi_1\left(\frac{u}{N}\right) + \frac{u\,\mathbb{E}\,(X\,W)}{N}\right| &= \left|\mathbb{E}\left(X\left(\mathrm{e}^{-u\,W/N} - 1 + u\,W/N\right)\right)\right| \\ &\leq \mathbb{E}\left(|X|\left(\mathrm{e}^{-u\,W/N} - 1 + u\,W/N\right)\right) \\ &\leq \frac{u}{N}\,\mathbb{E}\left(\,|X|\,W\,\left(1 - \mathrm{e}^{-u\,W/N}\right)\right)\,,\end{aligned}$$

where in the last inequality we applied (81) with $x = u\,W/N$. The two claims of the assertion (ii) easily follow.

Moving on to (iii), we use that, for all $x \geq 0$, $0 \leq 1 - \mathrm{e}^{-x} \leq x$, which implies

$$0 \leq 1 - \mathrm{e}^{-x} = \left(1 - \mathrm{e}^{-x}\right)^{\delta}\,\left(1 - \mathrm{e}^{-x}\right)^{1-\delta} \leq x^{\delta}\,\left(1 - \mathrm{e}^{-x}\right)^{1-\delta}\,.$$

We get that, since $\mathbb{E}(X) = 0$, for all $u \geq 0$ and $N \geq 1$,

$$\begin{aligned}\left|\phi_1\left(\frac{u}{N}\right)\right| &= \left|\mathbb{E}\left(X\left(\mathrm{e}^{-uW/N} - 1\right)\right)\right| \\ &\leq \mathbb{E}\left(|X|\left(1 - \mathrm{e}^{-uW/N}\right)\right) \\ &\leq \left(\frac{u}{N}\right)^{\delta}\,\mathbb{E}\left(|X|\,W^{\delta}\,\left(1 - \mathrm{e}^{-uW/N}\right)^{1-\delta}\right)\end{aligned}$$

and the desired assertions in (iii) follow under $\mathbb{E}\left(|X|\,W^{\delta}\right) < \infty$ using that $\left(1 - \mathrm{e}^{-uW/N}\right)^{1-\delta}$ is bounded by 1 and vanishes as $N \to \infty$ for $u\,W > 0$. ■

Next, we provide useful equivalences for the condition (79).

B.2 Useful Equivalences

Proposition 5 *Set $\overline{W}_N = N^{-1}\sum_{i=1}^{N} W_i$ for all $N \in \mathbb{N}^{\star}$, where $W_1, \ldots, W_N$ are positive i.i.d. random variables distributed as a generic random variable W. Then the following assertions hold.*

(i) For all $\mu > 0$, we have

$$\limsup_{N\to\infty} \mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) < \infty \Longleftrightarrow \exists N \geq 1\,,\, \mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) < \infty. \tag{82}$$

Furthermore, if the assertions of the equivalence (82) hold for some $\mu > 0$ and the distribution of W does not reduce to a Dirac measure, there exists $N_0 \geq 1$ such that $\mathbb{E}\left(\left(1/\overline{W}_N\right)^{\mu}\right) = \infty$ for $1 \leq N < N_0$ and the sequence $\left(\mathbb{E}\left(\left(1/\overline{W}_N\right)^{\mu}\right)\right)_{N \geq N_0}$ is strictly decreasing in $(0, \infty)$.

(ii) For all $\eta > 0$, the two conditions (17) and (79) are equivalent.

(iii) If condition (79) holds for $\eta > 0$, then $\mathbb{E}\left((\overline{W}_N)^{-\mu}\right) < \infty$ for all $\mu > 0$ and $N \geq 1$ such that $N > \mu/\eta$.

(iv) If condition (79) holds for $\eta > 0$, then $\mathbb{E}(W^{-\mu}) < \infty$ for $0 < \mu < \eta$.

(v) If $\mathbb{E}(W^{-\mu}) < \infty$ for $\mu > 0$, then condition (17) holds for $0 < \eta \leq \mu$.

(vi) If $\mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) < \infty$ for some $\mu > 0$ and $N \geq 1$, then (79) hold with $\eta = \mu/N$.

(vii) Suppose that (17) holds for some $\eta > 0$. Then, for all $\mu > 0$, we have

$$\lim_{N\to\infty} \mathbb{E}\left((1/\overline{W}_N)^{\mu}\right) = \begin{cases} 0 & \text{if } \mathbb{E}(W) = \infty, \\ (\mathbb{E}(W))^{-\mu} & \text{otherwise.} \end{cases} \tag{83}$$

Proof We prove Assertions (i)–(vii) successively.

(i) **Proof of Assertion (i).** Let Γ denote the Euler Gamma function, so that $\Gamma(\mu) = \int_0^\infty t^{\mu-1}\mathrm{e}^{-t}\,\mathrm{d}t$ for all $\mu > 0$. Following an idea of Cressie et al. (1981), we have: for all $N \geq 1$,

$$\begin{aligned} \mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) &= (\Gamma(\mu))^{-1} \int_0^\infty t^{\mu-1}\mathbb{E}\left(\mathrm{e}^{-\overline{W}_N\, t}\right)\,\mathrm{d}t \\ &= (\Gamma(\mu))^{-1} \int_0^\infty t^{\mu-1}\left(\mathbb{E}\left(\mathrm{e}^{-W\, t/N}\right)\right)^N\,\mathrm{d}t, \end{aligned} \tag{84}$$

where we have used that, for all $x > 0$, $x^{-\mu} = (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\mathrm{e}^{-x\,t}\,\mathrm{d}t$ and Tonelli's theorem. For all $t \geq 0$ and $N \geq 1$, by Jensen's inequality and strict convexity of $x \mapsto x^{(N+1)/N}$ on $(0,\infty)$, we have, for all $t > 0$,

$$\left(\mathbb{E}\left(\mathrm{e}^{-W\,t/N}\right)\right)^N = \left(\mathbb{E}\left(\left(\mathrm{e}^{-W\,t/(N+1)}\right)^{(N+1)/N}\right)\right)^N \geq \left(\mathbb{E}\left(\mathrm{e}^{-W\,t/(N+1)}\right)\right)^{N+1},$$

with equality if and only if both sides of the inequality are infinite or $\mathrm{e}^{-W\,t/(N+1)}$ is equal to its mean a.s. and we can thus deduce Assertion (i).

(ii) **Proof of Assertion (ii).** Let $\eta > 0$. We are interested in proving the equivalence between the two conditions

$$\sup_{t>0}\left(t^\eta\,\mathbb{E}\left(\mathrm{e}^{-t\,W}\right)\right) < \infty \tag{85}$$

and

$$\sup_{u>0}\left(u^{-\eta}\,\mathbb{P}\left(W \leq u\right)\right) < \infty. \tag{86}$$

Let us first assume that (85) holds and show (86). We have, for all $u > 0$, by Markov's inequality,

$$\mathbb{P}\left(W \leq u\right) = \mathbb{P}\left(\mathrm{e}^{\eta(1-W/u)} \geq 1\right) \leq \mathrm{e}^\eta\,\mathbb{E}\left(\mathrm{e}^{-\eta\,W/u}\right) \leq \mathrm{e}^\eta\,C\,\eta^{-\eta}\,u^\eta,$$

where C is the sup in the left-hand side of (85). Hence (86) follows.

Let us now assume (86) and show that (85) holds. We have, for all $t > 0$,

$$\mathbb{E}\left(\mathrm{e}^{-t\,W}\right) = t\int_0^\infty \mathrm{e}^{-t\,u}\,\mathbb{P}\left(W \leq u\right)\,\mathrm{d}u = \int_0^\infty \mathrm{e}^{-v}\,\mathbb{P}\left(W \leq v/t\right)\,\mathrm{d}v\,.$$

Let C now denote the (finite) sup in the left-hand side of (86), so that $\mathbb{P}(W \leq u) \leq C\, u^\eta$ for all $u > 0$. Using this bound in the previous integral, we get that, for all $t > 0$,

$$\mathbb{E}\left(\mathrm{e}^{-t\,W}\right) \leq C\, t^{-\eta} \int_0^\infty \mathrm{e}^{-v}\, v^\eta \,\mathrm{d}v\ .$$

Since the latter integral is a finite constant, this concludes the proof of (ii).

(iii) **Proof of Assertion (iii).** Suppose that (79) holds for some $\eta > 0$. Then there exists $c > 0$ such that, for all $t \geq 0$, $\mathbb{E}\left(\mathrm{e}^{-W\,t}\right) \leq 1 \wedge (c\, t^{-\eta})$. Hence for $\mu > 0$ and $N \geq 1$ such that $N > \mu/\eta$, the right-hand side of (84) is finite and $\mathbb{E}\left((1/\overline{W}_N)^\mu\right)$ as well.

(iv) **Proof of Assertion (iv).** This is a special case of (iii) imposing $N = 1$.

(v) **Proof of Assertion (v).** Suppose that $\mathbb{E}(W^{-\mu}) < \infty$ for some $\mu > 0$. Then $\mathbb{P}(W = 0) = 0$ and, for all $u > 0$,

$$\mathbb{P}(W \leq u) = \mathbb{P}(W^{-\mu} u^\mu \geq 1) \leq u^\mu \mathbb{E}(W^{-\mu})\ ,$$

where we used the Markov inequality.

(vi) **Proof of Assertion (vi).** Applying Assertions (v) and (ii) successively with W replaced by $\overline{W}_N$, the given assumption implies

$$\sup_{t>0}\left(t^\mu \mathbb{E}\left(\mathrm{e}^{-t\overline{W}_N}\right)\right) < \infty\ .$$

Since $\mathbb{E}\left(\mathrm{e}^{-t\overline{W}_N}\right) = \left(\mathbb{E}\left(\mathrm{e}^{-t\,W/N}\right)\right)^N$, we get, setting $u = t/N$,

$$\sup_{u>0}\left(u^\mu \left(\mathbb{E}\left(\mathrm{e}^{-u\,W}\right)\right)^N\right) < \infty\ .$$

The claimed implication follows.

(vii) **Proof of Assertion (vii).** By the strong law of large numbers we have $\overline{W}_N \to \mathbb{E}(W)$ as $N \to \infty$ a.s., including in the case where $\mathbb{E}(W) = \infty$. Hence the sequence $\left((1/\overline{W}_N)^\mu\right)_{N\geq 1}$ converges to the right-hand side of (83) a.s. To conclude the proof of Assertion (vii), it suffices to show that, for all $\mu > 0$, the sequence $\left((1/\overline{W}_N)^\mu\right)_{N\geq N_0}$ is uniformly integrable for some $N_0 \geq 1$ (see for example Theorem 3.5 of Billingsley, 1971). By the Markov inequality, the uniform integrability then follows from the fact that: for some $\epsilon > 0$,

$$\limsup_{N\to\infty} \mathbb{E}\left(\left((1/\overline{W}_N)^\mu\right)^{1+\epsilon}\right) < \infty.$$

In other words, the uniform integrability will follow from

$$\limsup_{N\to\infty} \mathbb{E}\left((1/\overline{W}_N)^{\mu'}\right) < \infty,$$

with $\mu' = \mu(1+\epsilon) > \mu$. By Assertion (iii), the assertions of the equivalence (82) hold for all $\mu > 0$. The proof is concluded by taking $\mu' = \mu(1+\epsilon)$ in lieu of μ in the assertions of the equivalence (82).

■

Remark 1 *Assertion (ii) in Proposition 5 is related to the famous Abelian and Tauberian theorems for the Laplace transform of a probability measure on $\mathbb{R}_+$, see (Feller, 1991, Section XIII.5).*

Remark 2 *Assertion (i) in Proposition 5 can be deduced from the proof of Daudel et al. (2023, Lemma 4). Here we provided a simpler proof of this assertion and, more importantly, we obtained the novel assertions (ii)–(vii).*

B.3 Asymptotic Behavior of Sample Means Ratios: Key Special Cases

In this section, we investigate the asymptotic behavior of sample means ratios in some special cases that are pivotal to prove Theorem 1. We start with a preliminary lemma.

Lemma 4 *Suppose that (79) holds for some $\eta > 0$ and let $\mu > 0$. Then the following assertions hold.*

(i) If X is valued in $\mathbb{R}_+$, then the two following identities hold for $N \geq 1$ and $N \geq 2$ respectively:

$$\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) = \int_0^\infty \frac{t^{\mu-1}}{\Gamma(\mu)}\,\phi_1\left(\frac{t}{N}\right)\,\left(\phi\left(\frac{t}{N}\right)\right)^{N-1}\mathrm{d}t, \tag{87}$$

$$\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) = \int_0^\infty \frac{t^{\mu-1}}{\Gamma(\mu)}\left(\phi_1\left(\frac{t}{N}\right)\right)^2\left(\phi\left(\frac{t}{N}\right)\right)^{N-2}\mathrm{d}t. \tag{88}$$

(ii) If $\mathbb{E}(|X|) < \infty$, then there exists $N_0 \geq 3$ such that for all $N \geq N_0$, the identities (87) and (88) hold with well defined and finite integrals. Moreover the following identity holds for all $N \geq N_0$:

$$\mathbb{E}\left(\frac{\overline{X^c}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) = \mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) - 2\mathbb{E}\left(\frac{\mathbb{E}(X)\,\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) + \mathbb{E}\left(\frac{(\mathbb{E}(X))^2}{\left(\overline{W}_N\right)^\mu}\right). \tag{89}$$

Proof We prove (i) and (ii) sequentially.

(i) **Proof of (i).** Proceeding as in (84) and using that $x^{-\mu} = (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\mathrm{e}^{-x\,t}\,\mathrm{d}t$ for all $x > 0$, we have by Tonelli's theorem (since $\overline{X}_N$ and $\overline{W}_N$ are nonnegative): for all $N \geq 1$,

$$\begin{aligned}
\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) &= (\Gamma(\mu))^{-1}\,\mathbb{E}\left(\overline{X}_N\int_0^\infty t^{\mu-1}\,\mathrm{e}^{-\overline{W}_N\,t}\mathrm{d}t\right)\\
&= (\Gamma(\mu))^{-1}\,\mathbb{E}\left(X_1\int_0^\infty t^{\mu-1}\,\mathrm{e}^{-\overline{W}_N\,t}\mathrm{d}t\right)\\
&= (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\left(\mathbb{E}\left(X_1\mathrm{e}^{-\overline{W}_N\,t}\right)\right)\mathrm{d}t\\
&= (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\,\phi_1(t/N)\,\left(\phi(t/N)\right)^{N-1}\mathrm{d}t.
\end{aligned}$$

As for (88), it is obtained similarly by using that for all $t > 0$ and $N \geq 2$,

$$\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \mathbb{E}\left(X_i X_j \mathrm{e}^{-\overline{W}_N\, t}\right) = \left(\phi_1\left(\frac{t}{N}\right)\right)^2 \left(\phi\left(\frac{t}{N}\right)\right)^{N-2} ,$$

so that

$$\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^{\mu}}\right) = (\Gamma(\mu))^{-1} \int_0^{\infty} t^{\mu-1}\, \phi_1(t/N)^2\, (\phi(t/N))^{N-2}\, \mathrm{d}t.$$

(ii) **Proof of (ii).** Setting $u = (N-1)t/N$ for $N \geq 2$ and $u = (N-2)t/N$ for $N \geq 3$ respectively yield that (87) and (88) can be equivalently written as

$$\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^{\mu}}\right) = \left(\frac{N}{N-1}\right)^{\mu} \int_0^{\infty} \frac{u^{\mu-1}}{\Gamma(\mu)}\, \phi_1\left(\frac{u}{N-1}\right)\, \left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1} \mathrm{d}u, \tag{90}$$

$$\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^{\mu}}\right) = \left(\frac{N}{N-2}\right)^{\mu} \int_0^{\infty} \frac{u^{\mu-1}}{\Gamma(\mu)} \left(\phi_1\left(\frac{u}{N-2}\right)\right)^2 \left(\phi\left(\frac{u}{N-2}\right)\right)^{N-2} \mathrm{d}u. \tag{91}$$

Based on the proof (i), (90) will be valid when X is not valued in $\mathbb{R}_+$ provided that Fubini's theorem applies. In other words, (90) will hold if $\mathbb{E}\left(\overline{|X|}_N / \left(\overline{W}_N\right)^{\mu}\right) < \infty$ where $\overline{|X|}_N = N^{-1} \sum_{i=1}^N |X_i|$. Denoting $\tilde{\phi}_1(v) = \mathbb{E}\left(|X| \mathrm{e}^{-v\, W}\right)$, we have by applying (90) that

$$\mathbb{E}\left(\frac{\overline{|X|}_N}{\left(\overline{W}_N\right)^{\mu}}\right) = \left(\frac{N}{N-1}\right)^{\mu} \frac{1}{\Gamma(\mu)} \int_0^{\infty} u^{\mu-1}\, \tilde{\phi}_1\left(\frac{u}{N-1}\right)\, \left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1} \mathrm{d}u\, .$$

Since $\mathbb{E}(|X|) < \infty$ by assumption, the function $\tilde{\phi}_1$ is bounded on $\mathbb{R}_+$. Furthermore, (iv) of Lemma 2 with $m = \mu + 1$ yields that: there exist $c > 0$ and $N_0 \geq 2$ such that, for all $N \geq N_0$,

$$\left| u^{\mu-1}\, \tilde{\phi}_1\left(\frac{u}{N-1}\right) \left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1} \right| \leq \mathbb{E}(|X|)\; u^{\mu-1} \left[1 \wedge \left(c\, u^{-\mu-1}\right)\right], \quad u > 0. \tag{92}$$

Consequently $\mathbb{E}\left(\overline{|X|}_N / \left(\overline{W}_N\right)^{\mu}\right) < \infty$ for all $N \geq N_0$. Hence, for all $N \geq N_0$ Fubini's theorem applies and (90) holds, and so does its equivalent form (87). The proof is concluded by applying the same line of reasoning to (91) and using that there exist $c > 0$ and $N_0 \geq 3$ such that, for all $N \geq N_0$,

$$\left| u^{\mu-1}\, \tilde{\phi}_1\left(\frac{u}{N-2}\right)^2 \left(\phi\left(\frac{u}{N-2}\right)\right)^{N-2} \right| \leq (\mathbb{E}(|X|))^2\; u^{\mu-1} \left[1 \wedge \left(c\, u^{-\mu-1}\right)\right], \quad u > 0. \tag{93}$$

Hence we get (88). We conclude by observing that the identity (89) is obtained by applying (87) and (88) to X and X_i replaced by $X^c = X - \mathbb{E}(X)$ and X_i^c by $X_i - \mathbb{E}(X)$ (with corresponding similar upper bounds in (92) and (93)) and noting that $\phi_1^c(u) = \mathbb{E}\left(X^c \mathrm{e}^{-u\, W}\right)$ satisfies $\phi_1^c = \phi_1 - \mathbb{E}(X)\, \phi$.

■

Proposition 6 *Suppose that (79) holds for some $\eta > 0$ and let $\mu > 0$. Then the following assertions hold.*

(i) If $\mathbb{E}(W) = 1$ and $\mathbb{E}(W^2) < \infty$, then

$$\lim_{N\to\infty} N \left(\mathbb{E}\left(\left(\overline{W}_N \right)^{-\mu} \right) - 1 \right) = \frac{(\mu+1)\mu}{2} \mathbb{V}(W) \ .$$

(ii) If $\mathbb{E}(|X|) < \infty$, then

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N \right)^{\mu}} \right) = \frac{\mathbb{E}(X)}{(\mathbb{E}(W))^{\mu}} \quad \text{and} \quad \lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N \right)^{\mu}} \right) = \frac{(\mathbb{E}(X))^2}{(\mathbb{E}(W))^{\mu}} \ .$$

(iii) If $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|\, W) < \infty$, then

$$\lim_{N\to\infty} N\, \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N \right)^{\mu}} \right) = -\frac{\mu\, \mathbb{C}\mathrm{ov}\,(X, W)}{(\mathbb{E}(W))^{\mu+1}} \ .$$

(iv) If $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|^2) < \infty$, then

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\left(\overline{X}_N \right)^2}{\left(\overline{W}_N \right)^{\mu}} \right) = 0 \ .$$

(v) If $\mathbb{E}(X) = 0$ and $\mathbb{E}\left(|X|\, W^{\delta} \right)) < \infty$ with $0 < \delta < 1$, then

$$\lim_{N\to\infty} N^{\delta}\, \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N \right)^{\mu}} \right) = 0.$$

Further assuming that $\delta = 1/2$ and $\mathbb{E}(|X|^2) < \infty$, it holds that

$$\lim_{N\to\infty} N\, \mathbb{E}\left(\frac{\left(\overline{X}_N \right)^2}{\left(\overline{W}_N \right)^{\mu}} \right) = \frac{\mathbb{V}(X)}{(\mathbb{E}(W))^{\mu}} \ . \tag{94}$$

Proof We prove Assertions (i)–(v) successively.

(i) **Proof of (i).** Using (84) and the definition of the Euler Gamma function we obtain that, for all $N \in \mathbb{N}^*$,

$$\begin{aligned} N \left(\mathbb{E}\left(\left(\overline{W}_N \right)^{-\mu} \right) - 1 \right) &= (\Gamma(\mu))^{-1} \int_0^{\infty} u^{\mu-1} N \left(\left(\phi\left(\frac{u}{N} \right) \right)^N - \mathrm{e}^{-u} \right) \mathrm{d}u \\ &=: (\Gamma(\mu))^{-1} \left(A_N + B_N \right) \ , \end{aligned}$$

where A_N and B_N are defined by separating the integral in two parts and setting

$$A_N = \int_0^\infty u^{\mu-1}\, N \left(\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \right) \mathbb{1}_{u \in \left[0, \frac{N}{\mathbb{E}(W^2)}\right]}\, \mathrm{d}u$$

$$B_N = \int_0^\infty u^{\mu-1}\, N \left(\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \right) \mathbb{1}_{u \in \left(\frac{N}{\mathbb{E}(W^2)}, \infty\right)}\, \mathrm{d}u.$$

Using next (77) and (78) of Lemma 2, we obtain that the dominated convergence theorem applies and leads to

$$\lim_{N\to\infty} A_N = \frac{\mathbb{V}\,(W)}{2} \int_0^\infty u^{\mu+1}\, \mathrm{e}^{-u}\, \mathrm{d}u = \frac{\mathbb{V}\,(W)\ \Gamma(\mu+2)}{2}.$$

As for B_N, (iv) of Lemma 2 with $m = \mu + 2$ yields that: there exist $c > 0$ and $N_0 \geq 1$ such that, for all $N \geq N_0$,

$$\left(\phi\left(\frac{u}{N}\right)\right)^N - \mathrm{e}^{-u} \leq \left(\phi\left(\frac{u}{N}\right)\right)^N \leq c\, u^{-\mu-2}\ .$$

Combining this with (ii) of Lemma 2, gives us that, for all $N \geq N_0$,

$$0 \leq B_N \leq c\, N \int_{\frac{N}{\mathbb{E}(W^2)}}^\infty u^{-3}\, \mathrm{d}u \leq \frac{c}{2}\ \left(\mathbb{E}\left(W^2\right)\right)^2\ N^{-1}\ .$$

Thus B_N converges to zero and the given limit is proved using the well known identity $\Gamma(\mu+2) = (\mu+1)\,\mu\,\Gamma(\mu)$.

(ii) **Proof of (ii).** Let us first consider the case where $\mathbb{E}(X) = 0$, that is, we show that

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu} \right) = \lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu} \right) = 0\ . \tag{95}$$

By (ii) of Lemma 4, we can use the identities (87) and (88) for N large enough, or their equivalent versions (90) and (91). The desired result (95) will follow if we can apply the dominated convergence theorem in the right-hand sides of (90) and (91). Using (iv) of Lemma 2 and since $\mathbb{E}(|X|) < \infty$ ($\mathbb{E}(X) = 0$ by assumption), we have that: for all $m > \mu$, there exist $c > 0$ and $N_0 \geq 3$ such that, for all $N \geq N_0$, (92) and (93) hold. Consequently we can apply the dominated convergence theorem: using (i) of Lemma 3 with $\phi_1(0) = \mathbb{E}(X) = 0$ and the fact that ϕ is bounded by 1 yields (95). Now considering the case where $\mathbb{E}(X) \neq 0$, we get from the first limit in (95) that

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X}_N - \mathbb{E}(X)}{\left(\overline{W}_N\right)^\mu} \right) = 0$$

and the first desired result follows using (vii) of Proposition 5. As for the second one, applying the second limit in (95) with X and X_i replaced by $X^c = X - \mathbb{E}(X)$ and $X_i^c = X_i - \mathbb{E}(X)$ for all $i = 1 \ldots N$ yields

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{\overline{X^c}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu} \right) = 0.$$

Under the present assumptions, we can apply (ii) of Lemma 4 and, in particular Identity (89), which, inserted in the previous limit, gives

$$\lim_{N\to\infty}\left[\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^{\mu}}\right)-2\mathbb{E}(X)\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^{\mu}}\right)+\mathbb{E}(X)^2\mathbb{E}\left(\frac{1}{\left(\overline{W}_N\right)^{\mu}}\right)\right]=0$$

The limits of $\mathbb{E}\left(\overline{X}_N/\left(\overline{W}_N\right)^{\mu}\right)$ and $\mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right)$ as $N\to\infty$ have been established previously to be $\mathbb{E}(X)/(\mathbb{E}(W))^{\mu}$ and $(\mathbb{E}(W))^{-\mu}$, and this leads to the second desired limit.

(iii) **Proof of (iii).** Let us define

$$I_N := N\,\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^{\mu}}\right)+\mathbb{E}\left(X\,W\right)\left(\frac{N}{N-1}\right)^{\mu+1}\frac{\Gamma(\mu+1)}{\Gamma(\mu)}\mathbb{E}\left(\frac{1}{\left(\overline{W}_{N-1}\right)^{\mu+1}}\right)\,. \tag{96}$$

By (ii) with $X=1$, the second term in the right-hand side of (96) converges to

$$\mathbb{E}\left(X\,W\right)\frac{\Gamma(\mu+1)}{\Gamma(\mu)}\frac{1}{(\mathbb{E}(W))^{\mu+1}}=\frac{\mu\,\mathbb{E}\left(X\,W\right)}{(\mathbb{E}(W))^{\mu+1}}\,,$$

where we used the well known identity of the Euler Gamma function. Now using (84) with N replaced by $N-1$ and μ replaced by $\mu+1$, we have: for all $N\geq 2$,

$$\mathbb{E}\left(\frac{1}{\left(\overline{W}_{N-1}\right)^{\mu+1}}\right)=\frac{1}{\Gamma(\mu+1)}\int_0^{\infty}u^{\mu}\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\,\mathrm{d}u$$

and by pairing it up with (ii) of Lemma 4 ($\mathbb{E}(|X|)<\infty$ since $\mathbb{E}(X)=0$), we get that for N large enough

$$I_N=N\,\left(\frac{N}{N-1}\right)^{\mu}\int_0^{\infty}\frac{u^{\mu-1}}{\Gamma(\mu)}\left(\phi_1\left(\frac{u}{N-1}\right)+\frac{u\,\mathbb{E}\left(X\,W\right)}{N-1}\right)\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\,\mathrm{d}u\,.$$

Therefore, to conclude the proof of (iii), it remains to show that I_N converges to zero as $N\to\infty$, which is implied by

$$\lim_{N\to\infty}\int_0^{\infty}\frac{u^{\mu-1}}{\Gamma(\mu)}\left\{N\,\left(\phi_1\left(\frac{u}{N-1}\right)+\frac{u\,\mathbb{E}\left(X\,W\right)}{N-1}\right)\right\}\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\,\mathrm{d}u=0\,. \tag{97}$$

Using (iv) of Lemma 2 and the first claim in (ii) of Lemma 3, we get that the absolute value of the integrand in (97) is bounded by an integrable function for N large enough, therefore the dominated convergence theorem applies. Furthermore, using (i) of Lemma 2 and the second claim in (ii) of Lemma 3, the point-wise limit of the integrand is zero. Thus (97) follows from the dominated convergence theorem and the proof of (iii) is concluded.

(iv) Proof of (iv). Using the definitions introduced in (75), we have

$$\mathbb{E}\left(\frac{\left(\overline{X}_N\right)^2}{\left(\overline{W}_N\right)^\mu}\right) = \frac{1}{N}\mathbb{E}\left(\frac{\left(\overline{X^2}_N\right)}{\left(\overline{W}_N\right)^\mu}\right) + \frac{N(N-1)}{N^2}\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) .$$

Under the assumption $\mathbb{E}(|X|^2) < \infty$, (ii) applies to both expectations in the right-hand side (with X and X_i replaced by their squares for the first one) and yields the result.

(v) **Proof of (v).** Let $0 < \delta < 1$. Again we can apply the identities of (ii) in Lemma 4 for N large enough. Here we pick the equivalent form (91) to write, for N large enough,

$$N^\delta \mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu}\right) = \left(\frac{N}{N-1}\right)^{\mu+\delta}\int_0^\infty \frac{u^{\mu-1}}{\Gamma(\mu)}(N-1)^\delta\,\phi_1\left(\frac{u}{N-1}\right)\,\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\mathrm{d}u.$$

The first claim then follows from the dominated convergence theorem since (iv) of Lemma 2 paired up with (iii) of Lemma 3 yield: there exist $c > 0$ and $N_0 \geq 3$ such that, for all $N \geq N_0$,

$$\left|u^{\mu-1}(N-1)^\delta\,\phi_1\left(\frac{u}{N-1}\right)\,\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\right| \leq u^{\mu+\delta-1}\mathbb{E}(|X|W^\delta)\left[1\wedge\left(c\,u^{-\mu-\delta-1}\right)\right]$$

$$\text{and}\quad \lim_{N\to\infty}\left|u^{\mu-1}(N-1)^\delta\,\phi_1\left(\frac{u}{N-1}\right)\,\left(\phi\left(\frac{u}{N-1}\right)\right)^{N-1}\right| = 0,$$

where we have also used the fact that ϕ is bounded by 1.

Moving on to the second claim where $\delta = 1/2$ and $\mathbb{E}(|X|^2) < \infty$, we observe that

$$N\,\mathbb{E}\left(\frac{\left(\overline{X}_N\right)^2}{\left(\overline{W}_N\right)^\mu}\right) = \mathbb{E}\left(\frac{\overline{X^2}_N}{\left(\overline{W}_N\right)^\mu}\right) + (N-1)\,\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) ,$$

where $\overline{X^2}_N$ and $\overline{X}_N^{(i\neq j)}$ are defined in (75). By applying (ii) with X and X_i replaced by their squares the first expectation in the right-hand side converges to the right-hand side of (94). Thus the proof of (94) will be complete if we can show that

$$\lim_{N\to\infty} N\,\mathbb{E}\left(\frac{\overline{X}_N^{(i\neq j)}}{\left(\overline{W}_N\right)^\mu}\right) = 0 .$$

Under our assumptions, we can apply (ii) of Lemma 4, notably giving that, the previous display is equivalent to

$$\lim_{N\to\infty}\left(\frac{N}{N-2}\right)^{\mu+1}\int_0^\infty \frac{u^{\mu-1}}{\Gamma(\mu)}\left((N-2)^{1/2}\phi_1\left(\frac{u}{N-2}\right)\right)^2\left(\phi\left(\frac{u}{N-2}\right)\right)^{N-2}\mathrm{d}u = 0 .$$

Again (iv) of Lemma 2 paired up with (iii) of Lemma 3 (with $\delta = 1/2$) allow us to apply the dominated convergence theorem, which yields the claimed limit.

■

B.4 Proof of Theorem 1

By (ii) of Proposition 5, the condition (17) is equivalent to the condition (79). Consequently, we will use the facts established in Proposition 6 (which relies on (79)) to prove Theorem 1 (which assumes (17)).

(i) **Proof of (i).** This is exactly (ii) of Proposition 6.

(ii) **Proof of (ii).** We first start with the general case $\mu > 0$. Multiplying both sides of (19) by $(\mathbb{E}(W))^\mu$ and replacing $W/\mathbb{E}(W)$, $W_i/\mathbb{E}(W)$ by W and W_i respectively, we see that we can assume $\mathbb{E}(W) = 1$ without loss of generality, in which case (19) reads

$$\lim_{N\to\infty} N \left(\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu} \right) - \mathbb{E}(X) \right) = \frac{(\mu+1)\mu\,\mathbb{E}(X)}{2}\,\mathbb{V}\left(W\right) - \mu\,\mathbb{C}\mathrm{ov}\left(X, W\right) \ . \tag{98}$$

Under our assumptions and since $\mathbb{E}(W) = 1$, we can apply (i) and (iii) of Proposition 6, so that

$$\lim_{N\to\infty} N \left(\mathbb{E}(X)\mathbb{E}\left(\frac{1}{\left(\overline{W}_N\right)^\mu} \right) - \mathbb{E}(X) \right) = \frac{(\mu+1)\mu\,\mathbb{E}(X)}{2}\,\mathbb{V}\left(W\right) \ , \tag{99}$$

$$\lim_{N\to\infty} N\,\mathbb{E}\left(\frac{\overline{X^c}_N}{\left(\overline{W}_N\right)^\mu} \right) = -\mu\,\mathbb{C}\mathrm{ov}\left(X, W\right) \ , \tag{100}$$

where $\overline{X^c}_N$ is defined in (76). Observing that

$$\frac{\overline{X^c}_N}{\left(\overline{W}_N\right)^\mu} = \frac{\overline{X}_N}{\left(\overline{W}_N\right)^\mu} - \frac{\mathbb{E}(X)}{\left(\overline{W}_N\right)^\mu} \ ,$$

we see that (98) is obtained by summing (99) and (100).

Moving on to the case $\mu = 1$, we can assume once again that $\mathbb{E}(W) = 1$ without loss of generality, in which case (20) becomes

$$\lim_{N\to\infty} N \left(\mathbb{E}\left(\frac{\overline{X}_N}{\overline{W}_N} \right) - \mathbb{E}(X) \right) = \mathbb{E}\left(W\left[\mathbb{E}(X)\,W - X\right]\right) \ . \tag{101}$$

Now replacing X and X_i by $(X - \mathbb{E}(X)W)$ and $(X_i - \mathbb{E}(X)W_i)$, respectively, we can further assume that $\mathbb{E}(X) = 0$ and (101) now reads

$$\lim_{N\to\infty} N\,\mathbb{E}\left(\frac{\overline{X}_N}{\overline{W}_N} \right) = -\mathbb{E}\left(X\,W\right) \ ,$$

under the assumptions $\mathbb{E}(W) = 1$, $\mathbb{E}(X) = 0$ and $\mathbb{E}(|X|\,W) < \infty$, which follows directly from an application of (iii) in Proposition 6 with $\mu = 1$.

(iii) **Proof of (iii).** By Minkowski's inequality, it holds that

$$\left(\mathbb{V}\left(\frac{\overline{X}_N}{(\overline{W}_N)^\mu}\right)\right)^{\frac{1}{2}} \leq \left(\mathbb{V}\left(\frac{\overline{X}_N - \mathbb{E}(X)}{(\overline{W}_N)^\mu}\right)\right)^{\frac{1}{2}} + |\mathbb{E}(X)| \left(\mathbb{V}\left(\frac{1}{(\overline{W}_N)^\mu}\right)\right)^{\frac{1}{2}}$$
$$\leq \left(\mathbb{E}\left(\frac{\left(\overline{X^c}_N\right)^2}{(\overline{W}_N)^{2\mu}}\right)\right)^{\frac{1}{2}} + |\mathbb{E}(X)| \left(\mathbb{V}\left(\frac{1}{(\overline{W}_N)^\mu}\right)\right)^{\frac{1}{2}} .$$

The variance in the second term of the sum in the last right-hand side converges to 0 by applying (vii) of Proposition 5 twice. As for the first term, it converges to zero by applying (iv) of Proposition 6 (with μ, X_i and X replaced by 2μ, X^c and X_i^c), which concludes the proof of (iii).

(iv) **Proof of (iv).** We start with the general case $\mu > 0$. First observe that we can assume that $\mathbb{E}(W) = 1$ again without loss of generality. For all $N \in \mathbb{N}^*$,

$$\mathbb{V}\left(\frac{\overline{X}_N}{(\overline{W}_N)^\mu}\right) = V_N^{(1)} + (\mathbb{E}(X))^2 \, V_N^{(2)} + 2\,\mathbb{E}(X)\, C_N \,, \tag{102}$$

where

$$V_N^{(1)} := \mathbb{V}\left(\frac{\overline{X^c}_N}{(\overline{W}_N)^\mu}\right) , \quad V_N^{(2)} := \mathbb{V}\left(\left(\overline{W}_N\right)^{-\mu}\right) \quad \text{and} \quad C_N := \mathbb{C}\mathrm{ov}\left(\frac{\overline{X^c}_N}{(\overline{W}_N)^\mu}, \left(\overline{W}_N\right)^{-\mu}\right) .$$

We treat each term separately.

- First variance term: $V_N^{(1)}$. Note that $\mathbb{E}(W^2) < \infty$ and $\mathbb{E}(|X|^2) < \infty$ imply that $\mathbb{E}(|X|\,W^{1/2})$ and $\mathbb{E}(|X|\,W)$ are finite. We can thus apply (94) and (iii) of Proposition 6 (with μ, X_i and X replaced by 2μ, X_i^c and X^c), which gives us the asymptotic behavior of the two expectations in the expression

$$V_N^{(1)} = \mathbb{E}\left(\frac{\left(\overline{X^c}_N\right)^2}{(\overline{W}_N)^{2\mu}}\right) - \left(\mathbb{E}\left(\frac{\overline{X^c}_N}{(\overline{W}_N)^\mu}\right)\right)^2 .$$

  We obtain

$$\lim_{N\to\infty} N\, V_N^{(1)} = \mathbb{V}\,(X) \ . \tag{103}$$

- Second variance term: $V_N^{(2)}$. For all $N \in \mathbb{N}^*$,

$$\mathbb{V}\left(\left(\overline{W}_N\right)^{-\mu}\right) = \left\{\mathbb{E}\left(\left(\overline{W}_N\right)^{-2\mu}\right) - 1\right\} - \left\{\left(1 + \left[\mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) - 1\right]\right)^2 - 1\right\} ,$$

  Applying (i) of Proposition 6 to get the asymptotic behavior of the expectations in the right-hand side, we get that

$$\lim_{N\to\infty} N\,\mathbb{V}\left(\left(\overline{W}_N\right)^{-\mu}\right) = (2\mu+1)\,\mu\,\mathbb{V}\,(W) - (\mu+1)\,\mu\,\mathbb{V}\,(W)$$
$$= \mu^2\,\mathbb{V}\,(W) \ . \tag{104}$$

- Covariance term: C_N.

$$C_N = \mathbb{E}\left(\frac{\overline{X^c}_N}{(\overline{W}_N)^{2\mu}}\right) - \mathbb{E}\left(\frac{\overline{X^c}_N}{(\overline{W}_N)^{\mu}}\right)\mathbb{E}\left(\left(\overline{W}_N\right)^{-\mu}\right) \ .$$

  Applying (iii) of Proposition 6 twice, once with μ replaced by 2μ and once again without changing μ, along with (vii) of Proposition 5, we get that

$$\lim_{N\to\infty} N\, C_N = -2\mu\, \mathbb{C}\mathrm{ov}\,(X,W) + \mu\, \mathbb{C}\mathrm{ov}\,(X,W) = -\mu\, \mathbb{C}\mathrm{ov}\,(X,W) \ .$$

Combining (102) with (103), (104) and the above limit, we obtain that

$$\lim_{N\to\infty} N\, \mathbb{V}\left(\frac{\overline{X}_N}{(\overline{W}_N)^{\mu}}\right) = \mathbb{V}\,(X) + (\mathbb{E}(X))^2\ \mu^2\, \mathbb{V}\,(W) - 2\,\mathbb{E}(X)\,\mu\, \mathbb{C}\mathrm{ov}\,(X,W) \ .$$

This is exactly the desired result (iv) in the case $\mathbb{E}(W) = 1$, and the proof is completed for a general $\mu > 0$.

Moving on to the case $\mu = 1$, we can assume once more that $\mathbb{E}(W) = 1$ without loss of generality, in which case (23) reads

$$\lim_{N\to\infty} N\, \mathbb{V}\left(\frac{\overline{X}_N}{\overline{W}_N}\right) = \mathbb{V}\,(X - \mathbb{E}(X)W) \ . \tag{105}$$

Replacing X and X_i by $(X - \mathbb{E}(X)W)$ and $(X_i - \mathbb{E}(X)W_i)$, respectively, we can further assume that $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) < \infty$ (transposing the original $\mathbb{E}(|\mathbb{E}(W)X - \mathbb{E}(X)W|^2) < \infty$ to the new definition of X when $\mathbb{E}(W) = 1$) and (105) reads

$$\lim_{N\to\infty} N\, \mathbb{V}\left(\frac{\overline{X}_N}{\overline{W}_N}\right) = \mathbb{V}\,(X) \ . \tag{106}$$

Now using that $\mathbb{E}(|X|^2) < \infty$ and $\mathbb{E}(W) = 1$, which implies $\mathbb{E}(|X|\sqrt{W}) < \infty$, we can apply the two assertions in (v) of Proposition 6 with $\mu = 1, 2$ and $\delta = 1/2$ so that

$$\lim_{N\to\infty} N\left(\mathbb{E}\left(\frac{\overline{X}_N}{\left(\overline{W}_N\right)}\right)\right)^2 = 0 \quad \text{and} \quad \lim_{N\to\infty} N\,\mathbb{E}\left(\left(\frac{\overline{X}_N}{\overline{W}_N}\right)^2\right) = \mathbb{V}\,(X).$$

This yields (106) and the proof is completed.

B.5 Related works for Theorem 1

(i) Deligiannidis et al. (2025) concurrently established a result that is comparable to (20): their analysis in (Deligiannidis et al., 2025, Theorem 2.1) focused on the self-normalized importance sampling estimator, which corresponds to taking $(X, W) = (f(Z)\omega(Z), \omega(Z))$ with $Z \sim q$, where $\omega = \pi/q$ is the importance weight, π is the target distribution, q the proposal distribution and f is a function of interest. Letting $Z_1, \ldots, Z_N$ be i.i.d. random variables with the same distribution as Z, (20) then becomes

$$\lim_{N\to\infty} N\left(\mathbb{E}\left(\frac{\sum_{i=1}^N w(Z_i)f(Z_i)}{\sum_{i=1}^N w(Z_i)}\right) - \mathbb{E}(w(Z)f(Z))\right) = -\mathbb{E}\left(w(Z)^2\left[f(Z) - \mathbb{E}(w(Z)f(Z))\right]\right),$$

which is exactly the result obtained in Deligiannidis et al. (2025). However, their derivation relies on a distinct proof strategy that necessitates a stronger positive moment condition on $|\mathbb{E}(W)X - \mathbb{E}(X)W|W$.

(ii) Rainforth et al. (2018) and Daudel et al. (2023) directed their analyses exclusively towards the REP gradient estimator of the IWAE and VR-IWAE bound respectively. They showed that (20) and (23) hold for specific choices of (X, W) under much stringent conditions than that of Theorem 1. More precisely, letting $\varepsilon \sim \tilde{q}$, they take

$$X = w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha} \partial_\psi \log w_{\theta,\phi}(f(\varepsilon, \phi; x); x)$$
$$W = w_{\theta,\phi}(f(\varepsilon, \phi; x); x)^{1-\alpha},$$

with Rainforth et al. (2018) considering the case $\alpha = 0$ and assuming that X and W have moments of order 6, while Daudel et al. (2023) considers the case $\alpha \in (0, 1)$ and assumes moments of order 8 for X and W. The derivations in Rainforth et al. (2018) and Daudel et al. (2023) rely on distinct proof strategies that necessitates stronger conditions compared to those used in Theorem 1.

Appendix C. Deferred proofs and results for Section 3.3

C.1 Proofs of Theorems 2 to 4

C.1.1 Notation

Let us first introduce some notation and resulting identities that will be used for the proofs of Theorems 2 to 4. Letting $\varepsilon \sim \tilde{q}$ and with $\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)$ as in (24), we define

$$X = (1-\alpha)^{-1} \partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha} \right) \quad \text{and} \quad W = \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha} ,$$

Furthermore, let $\varepsilon_1, \ldots, \varepsilon_N$ be i.i.d. copies of ε, and define X_i and W_i for all $i = 1, 2, \ldots$ as X and W but with ε_i replacing ε. Further denote $\overline{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$ and $\overline{W}_N = \frac{1}{N} \sum_{i=1}^N W_i$ for all $N \in \mathbb{N}^*$. Then, using the above definitions in (14), we have

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) \right) = \mathbb{E}\left(\frac{\overline{X}_N}{\overline{W}_N} \right) , \tag{107}$$

$$\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \mathbb{V}\left(\frac{\overline{X}_N}{\overline{W}_N} \right) , \tag{108}$$

Remark 3 *We note that with these definitions,* (A3) *means that (17) holds with* $\eta = \delta/(1-\alpha)$, $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}\right) < \infty$ *that* $\mathbb{E}\left(W^2\right) < \infty$, $\mathbb{V}\left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}\right)\right) < \infty$ *that* $\mathbb{E}\left(|X|^2\right) < \infty$ *and* $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ *implies* $\mathbb{E}\left(|X|\, W^{\ell-1}\right) < \infty$ *for* $\ell = 0, 1, 2$.

Remark 4 *The above quantities are defined for* $0 \leq \alpha < 1$. *However, the definition of* W *boils down for* $\alpha = 1$ *to* $W = 1$, *hence* $\overline{W}_N = 1$ *for all* $N \geq 1$. *As for* X, *since it holds for*

all $\alpha \in [0,1)$ *that* $X = W\partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)$*, it becomes* $X = \partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)$ *for* $\alpha = 1$*. In this case, explicit formulas hold for (107) and (108), namely,*

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \mathbb{E}\left(\partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)\right) = \partial_\psi \mathrm{ELBO}(\theta,\phi;x) \ ,$$
$$\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \frac{1}{N}\mathbb{V}\left(\partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)\right) \ .$$

where the second equality in the first displayed line follows from $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$*, which for* $\alpha = 1$ *takes the form*

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \left|\partial_{\psi'}\left(\log \tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right)\right|\right) < \infty \ .$$

Thus the results of Theorems 2 and 3 hold for $\alpha = 1$ *in a non-asymptotic way, namely, with all o-terms replaced by zeros,* $\mathrm{VR}^{(1)}(\theta,\phi;x) = \mathrm{ELBO}(\theta,\phi;x)$*,* $\psi\text{-}V^{(1,\mathrm{REP})}(\theta,\phi;x) = \mathbb{V}\left(\partial_\psi \left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right)\right)$ *and* $\partial_\psi[\gamma^{(1)}(\theta,\phi;x)^2] = 0$*.*

C.1.2 Proof of Theorem 2

Under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, it holds that

$$\partial_\psi \ell_N^{(\alpha)}(\theta,\phi;x) = \mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) \tag{109}$$
$$\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{\mathbb{E}(X)}{\mathbb{E}(W)} \tag{110}$$
$$\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2] = 2\,\frac{\mathbb{E}\left(W\left[\mathbb{E}(W)\,X - \mathbb{E}(X)\,W\right]\right)}{\left(\mathbb{E}(W)\right)^3}, \tag{111}$$

which follows from Proposition 2, the definitions of Appendix C.1.1, (12), (13), (8) and (56)–(58), as they imply

$$\begin{aligned}
&\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\,\mathbb{E}\left(\log \overline{W}_N\right) \ ,\\
&\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\,\log \mathbb{E}\left(W\right) \ ,\\
&\gamma^{(\alpha)}(\theta,\phi;x)^2 = \frac{1}{1-\alpha}\,\left(\frac{\mathbb{E}(W^2)}{\mathbb{E}(W)^2} - 1\right) \ ,\\
&\partial_\psi \mathbb{E}(W) = (1-\alpha)\,\mathbb{E}(X) \ ,\\
&\partial_\psi \mathbb{E}\left(W^2\right) = 2\,(1-\alpha)\,\mathbb{E}\left(X\,W\right) \ ,\\
&\partial_\psi \mathbb{E}\left(\log \overline{W}_N\right) = (1-\alpha)\,\mathbb{E}\left(\frac{\overline{X}_N}{\overline{W}_N}\right) \ .
\end{aligned} \tag{112}$$

We can conclude by using Remark 3 and the interchanges of the derivative and expectation signs obtained above under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ to rewrite the asymptotic results (i) and the second part of (ii) from Theorem 1 into the convenient formulations (26) and (27) respectively.

C.1.3 Proof of Theorem 3

Note that the conditions of Theorem 3 are stronger than those required in Theorem 2. We thus already have that (27) holds as $N \to \infty$, which gives that $|\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))|$ behaves as the numerator of the r.h.s of (28). To conclude the proof, it remains to study the asymptotic behavior of $\mathbb{V}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ as $N \to \infty$. As in the proof of Theorem 2, we first note that, following Remark 3, the assumptions of Theorem 3 imply

(W1) There exists $\eta > 0$ such that (17) holds and $\mathbb{E}(|X|)$, $\mathbb{E}(W)$ and $\mathbb{E}\left(\left|\mathbb{E}(W)X - \mathbb{E}(X)W\right|^2\right)$ are all finite.

Then the claimed asymptotic behavior of the variance is obtained in two steps: (i) we show a first asymptotic behavior for $\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ as $N \to \infty$ which holds under (W1) and (ii) we show how the conclusions of Proposition 2 allow us to express this asymptotic result under the convenient formulation (29).

First step. Under (W1), we can apply the second part of Assertion (iv) of Theorem 1. Using (23) and (108) we have that, as $N \to \infty$,

$$\begin{aligned}\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) &= \frac{1}{N}\frac{\mathbb{V}\left(\mathbb{E}(W)X - \mathbb{E}(X)W\right)}{\left(\mathbb{E}(W)\right)^4} + o\left(\frac{1}{N}\right) \\ &= \frac{1}{N}\mathbb{V}\left(\frac{X}{\mathbb{E}(W)} - W\frac{\mathbb{E}(X)}{\mathbb{E}(W)^2}\right) + o\left(\frac{1}{N}\right).\end{aligned}$$

Second step. We prove that, under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, the constant defined in (29) reads

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{V}\left(\frac{X}{\mathbb{E}(W)} - W\frac{\mathbb{E}(X)}{\mathbb{E}(W)^2}\right). \tag{113}$$

By using the definitions of X and W in Appendix C.1.1 and (56) of Proposition 2, we have

$$\partial_\psi\left(\frac{W}{\mathbb{E}(W)}\right) = \frac{\partial_\psi W}{\mathbb{E}(W)} - W\frac{\partial_\psi\mathbb{E}(W)}{(\mathbb{E}(W))^2} = (1-\alpha)\left(\frac{X}{\mathbb{E}(W)} - W\frac{\mathbb{E}(X)}{\mathbb{E}(W)^2}\right)$$

meaning that the two variables in the variances of the right-hand sides of (29) and (113) coincide up to the $(1-\alpha)$ multiplicative term, which is squared when put in front of the variance. The identity (113) follows.

Remark 5 *As a byproduct of the proof of Appendix C.1.3, we have established* (113)*, that is using the definition of X and W in Appendix C.1.1,*

$$\begin{aligned}&\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) \\ &= \frac{1}{(1-\alpha)^2}\mathbb{V}\left(\frac{\partial_\psi(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})} - \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\mathbb{E}(\partial_\psi(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}))}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})^2}\right) \\ &= \mathbb{V}\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\left(\partial_\psi(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)) - \mathbb{E}\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\partial_\psi(\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))\right)\right)\right).\end{aligned}$$

Now considering the case $p_\theta(\cdot|x) = q_\phi(\cdot|x)$ in the above and using

$$\partial_\psi \log w_{\theta,\phi}(f(\varepsilon,\phi;x);x) = \left[\partial_{\psi'} \log w_{\theta,\phi}(f(\varepsilon,\phi';x);x) - \partial_{\psi'} \log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\big|_{\phi'=\phi}, \tag{114}$$

it follows that

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{V}\left(\left[\partial_{\psi'} \log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\big|_{\phi'=\phi}\right) \quad \textit{when } p_\theta(\cdot|x) = q_\phi(\cdot|x). \tag{115}$$

C.1.4 Proof of Theorem 4

We first introduce some additional notation for the DREP estimators. Namely, we set

$$X^{(1)} = \frac{1}{1-\alpha}\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi} = \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\left[\partial_\psi \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right]\big|_{\phi'=\phi}$$
$$X^{(2)} = \frac{1}{2(1-\alpha)}\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)}\right)\right]\Big|_{\phi'=\phi} = \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha} X^{(1)} = W\, X^{(1)}\,.$$

Then, $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$ means that $\mathbb{E}\left(|W|^2\right) < \infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right) < \infty$ that $\mathbb{E}\left(|X^{(1)}|^2\right) < \infty$ and $\mathbb{V}\left([\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})]|_{\phi'=\phi}\right) < \infty$ that $\mathbb{E}\left(|X^{(2)}|^2\right) < \infty$. With Remark 3, this leads to the following two lists of conditions valid respectively in Part (i) and (ii) of Theorem 4:

(W2) There exists $\eta > 0$ such that (17) holds and $\mathbb{E}\left(\left|X^{(1)}\right|\right)$, $\mathbb{E}\left(|X^{(2)}|^2\right)$, $\mathbb{E}(W)$ and $\mathbb{E}\left(|\mathbb{E}(W)X^{(1)} - \mathbb{E}\left(X^{(1)}\right)W|^2\right)$ are all finite.

(W3) There exists $\eta > 0$ such that (17) holds and $\mathbb{E}\left(\left|X^{(2)}\right|^2\right)$ and $\mathbb{E}(W^2)$ are finite.

As in Appendix C.1.1 for X_i and W_i, the random variables $X_i^{(1)}$ and $X_i^{(2)}$ are defined for all $i \geq 1$ as $X^{(1)}$ and $X^{(2)}$ respectively but with ε_i replacing ε, and we set $\overline{X}_N^{(\ell)} = \frac{1}{N}\sum_{i=1}^N X_i^{(\ell)}$ for $\ell = 1, 2$. Then, using the above definitions in (15), we have

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = \mathbb{E}\left(\alpha\,\frac{\overline{X}_N^{(1)}}{\overline{W}_N} + \frac{1-\alpha}{N}\,\frac{\overline{X}_N^{(2)}}{(\overline{W}_N)^2}\right) \tag{116}$$

$$\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = \mathbb{V}\left(\alpha\,\frac{\overline{X}_N^{(1)}}{\overline{W}_N} + \frac{1-\alpha}{N}\,\frac{\overline{X}_N^{(2)}}{(\overline{W}_N)^2}\right) \tag{117}$$

and we further define

$$\psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \alpha\,\frac{\mathbb{E}\left(X^{(1)}\right)}{\mathbb{E}(W)}\,, \tag{118}$$

$$\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) = \frac{\mathbb{E}\left(X^{(2)}\right)}{\left(\mathbb{E}(W)\right)^2} \tag{119}$$

as well as

$$\psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \alpha^2\,\mathbb{V}\left(\frac{X^{(1)}}{\mathbb{E}(W)} - W\frac{\mathbb{E}(X^{(1)})}{\mathbb{E}(W)^2}\right) \tag{120}$$

$$\text{and}\quad \psi\text{-}V_2^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = (1-\alpha)^2\,\mathbb{V}\left(\frac{X^{(2)}}{\mathbb{E}(W)^2} - 2W\frac{\mathbb{E}(X^{(2)})}{\mathbb{E}(W)^3}\right)\,. \tag{121}$$

As for the proofs of Theorems 2 and 3, the proof of Theorem 4 is made of two steps. In the first step, we establish that under (W2) with $\alpha \in (0,1)$: as $N \to \infty$,

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)] = \sqrt{N}\,\frac{\left|\psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right| + o(1)}{\sqrt{\psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)} + o(1)}\,, \tag{122}$$

while under (W3) with $\alpha = 0$: as $N \to \infty$,

$$\mathrm{SNR}\left[\psi\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)\right] = \sqrt{N}\,\frac{\left|\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x)\right| + o\left(1\right)}{\sqrt{\psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x)} + o(1)}\,. \tag{123}$$

In the second step we use the conclusions of Proposition 3 to rewrite (122) and (123) under the convenient formulations (31) and (33).

First step. Applying (i) and (iii)–(iv) from Theorem 1 with $X^{(1)}$ or $X^{(2)}$ replacing X, we obtain the following implications:

$$\mathbb{E}\left(\left|X^{(1)}\right|\right) < \infty \Longrightarrow \lim_{N\to\infty}\mathbb{E}\left(\frac{\overline{X}_N^{(1)}}{\overline{W}_N}\right) = \frac{\mathbb{E}\left(X^{(1)}\right)}{\mathbb{E}(W)}\,, \tag{124}$$

$$\mathbb{E}\left(\left|X^{(2)}\right|\right) < \infty \Longrightarrow \lim_{N\to\infty}\mathbb{E}\left(\frac{\overline{X}_N^{(2)}}{\left(\overline{W}_N\right)^2}\right) = \frac{\mathbb{E}\left(X^{(2)}\right)}{\left(\mathbb{E}(W)\right)^2}\,, \tag{125}$$

$$\mathbb{E}\left(|X^{(2)}|^2\right) < \infty \Longrightarrow \lim_{N\to\infty}\mathbb{V}\left(\frac{\overline{X}_N^{(2)}}{\left(\overline{W}_N\right)^2}\right) = 0\,, \tag{126}$$

$$\left.\begin{array}{c}\mathbb{E}\left(\left|X^{(1)}\right|\right)<\infty,\,\mathbb{E}(W)<\infty,\\ \mathbb{V}\left(\mathbb{E}(W)X^{(1)}-\mathbb{E}(X^{(1)})W\right)<\infty\end{array}\right\} \Longrightarrow \lim_{N\to\infty} N\,\mathbb{V}\left(\frac{\overline{X}_N^{(1)}}{\overline{W}_N}\right) = \frac{\mathbb{V}\left(\mathbb{E}(W)X^{(1)} - W\mathbb{E}(X^{(1)})\right)}{\left(\mathbb{E}(W)\right)^4}\,, \tag{127}$$

$$\left.\begin{array}{c}\mathbb{E}\left(|X^{(2)}|^2\right)<\infty\\ \mathbb{E}(W^2)<\infty\end{array}\right\} \Longrightarrow \lim_{N\to\infty} N\,\mathbb{V}\left(\frac{\overline{X}_N^{(2)}}{\left(\overline{W}_N\right)^2}\right) = \frac{\mathbb{V}\left(\mathbb{E}(W)\,X^{(2)} - 2\,W\,\mathbb{E}(X^{(2)})\right)}{\left(\mathbb{E}(W)\right)^6}\,. \tag{128}$$

From (116), (124) and (125), using the quantities defined in (118) and (119) we get that if $\alpha \in (0,1)$ and $\mathbb{E}\left(\left|X^{(i)}\right|\right) < \infty$ for $i = 1, 2$: as $N \to \infty$,

$$\mathbb{E}\left(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) + o(1)\,. \tag{129}$$

Pairing (126) and (127) up, we get that if $\alpha \in (0,1)$ and (W2) holds: as $N \to \infty$,

$$\mathbb{V}\left(\alpha \frac{\overline{X}_N^{(1)}}{\overline{W}_N} + \frac{1-\alpha}{N}\frac{\overline{X}_N^{(2)}}{\overline{W}_N^2}\right) = \frac{\alpha^2}{N}\frac{\mathbb{V}\left(X^{(1)} - W\frac{\mathbb{E}(X^{(1)})}{\mathbb{E}(W)}\right)}{(\mathbb{E}(W))^2} + o\left(\frac{1}{N}\right) .$$

Combining this with (117) and (129) yields (122).

From (116) and (125), we get that if $\alpha = 0$ and $\mathbb{E}\left(\left|X^{(2)}\right|\right) < \infty$: as $N \to \infty$,

$$\mathbb{E}\left(\psi\text{-}g_{M,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)\right) = \frac{1}{N}\,\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) + o\left(\frac{1}{N}\right) . \tag{130}$$

Now, finally, (117), (121) and (128) in the case $\alpha = 0$ imply that, under (W3), as $N \to \infty$,

$$\mathbb{V}\left(\psi\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)\right) = \frac{\psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x) + o(1)}{N^3},$$

which, combined with (129) yields (123).

Second step. Under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$, the conclusions of Proposition 3 hold and we can show that:

(i) If $\alpha \in (0,1)$, (122) is the same as (31), that is, we have:

$$\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) , \tag{131}$$

$$\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) . \tag{132}$$

Indeed by definitions (13) and (118), (131) reads

$$\frac{1}{1-\alpha}\frac{\partial_\psi \mathbb{E}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)}{\mathbb{E}(W)} = \frac{\alpha}{1-\alpha}\frac{\mathbb{E}\left(\left[\partial_{\psi'}\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right]\big|_{\phi'=\phi}\right)}{\mathbb{E}(W)} ,$$

which holds by (66). As for (132), by (32) and (120), it easily follows by differentiating inside the variance in (32) and using (65).

(ii) If $\alpha = 0$, (123) is the same as (33), that is, we have:

$$\partial_\psi\left(\gamma^{(0)}(\theta,\phi;x)^2\right) = -2\,\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) , \tag{133}$$

$$\psi\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = \psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x) . \tag{134}$$

Indeed, by (8), with $\alpha = 0$, we have $\gamma^{(0)}(\theta,\phi;x)^2 = \left(\frac{\mathbb{E}(W^2)}{(\mathbb{E}(W))^2} - 1\right)$. Observe moreover that since $\alpha = 0$, we have $\mathbb{E}(W) = \int p_\theta(x,z)\,\nu(\mathrm{d}z)$ and thus $\mathbb{E}(W)$ does not depend on ϕ and therefore not on ψ. We thus get that $\partial_\psi\left(\gamma^{(0)}(\theta,\phi;x)^2\right) = \partial_\psi\mathbb{E}(W^2)/\left(\mathbb{E}(W)\right)^2$. By definition of $X^{(2)}$ and (119), (133) then follows from (67). As for (134), it simply follows by identifying (34) with (121) using $\alpha = 0$ and the definitions of W and $X^{(2)}$.

Remark 6 *As in Remark 4 for the REP estimator, let us briefly examine the case $\alpha = 1$. In this case we get $\psi\text{-}g_{1,N}^{(1,\mathrm{DREP})}(\theta,\phi;x) = \overline{X}_N^{(1)}$ with $X^{(1)} = \left[\partial_\psi \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right]\big|_{\phi'=\phi}$. When*

the DREP estimator is unbiased (under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$ *adapted to* $\alpha = 1$*), we further have that* $\mathbb{E}\left(X^{(1)}\right) = \partial_\psi \mathrm{ELBO}(\theta, \phi; x)$*, so that, for all* $N \geq 1$*,*

$$\mathrm{SNR}[\psi\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)] = \sqrt{N}\ \frac{|\partial_\psi \mathrm{ELBO}(\theta, \phi; x)|}{\sqrt{\psi\text{-}V^{(1,\mathrm{DREP})}(\theta, \phi; x)}}\ ,$$

with $\psi\text{-}V^{(1,\mathrm{DREP})}(\theta, \phi; x) = \mathbb{V}\left(\left[\partial_\psi \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\big|_{\phi'=\phi}\right)$*. This extends the asymptotic formula (31) obtained for* $\alpha \in (0, 1)$ *to the case* $\alpha = 1$*.*

C.2 Proof of examples

Before providing the proofs for Examples 1 and 2, we first recall below the two settings considered in these two examples alongside with some known results taken from Daudel et al. (2023).

Example 4 (Known results from Daudel et al. (2023) associated to Example 1) *Let* $\theta, \phi \in \mathbb{R}^d$*. Set* $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ *and* $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$*, where* $\boldsymbol{I}_d$ *is the* d*-dimensional identity matrix. Then, denoting the Euclidean norm of a finite dimensional vector* x *with real entries by* $\|x\|$ *and its associated inner product by* $\langle \cdot, \cdot \rangle$*,*

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x) - \frac{\alpha\|\theta - \phi\|^2}{2} \quad \text{and} \quad \gamma^{(\alpha)}(\theta, \phi; x)^2 = \frac{\exp\left((1-\alpha)^2\|\theta - \phi\|^2\right) - 1}{1 - \alpha}.$$

In addition, it holds that

$$\log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) = -\frac{\|\theta - \phi\|^2}{2} - \|\theta - \phi\| S, \quad S = \frac{\langle z - \phi, \phi - \theta\rangle}{\|\theta - \phi\|}, \quad z \in \mathbb{R}^d.$$

For $\phi = \theta + \boldsymbol{u}_d$*, we have*

$$\log\left(\frac{p_\theta(Z|x)}{q_\phi(Z|x)}\right) = -\frac{d}{2} - \sqrt{d} S, \quad S \sim \mathcal{N}(0, 1).$$

Example 5 (Known results from Daudel et al. (2023) associated to Example 2) *Let* $\theta \in \mathbb{R}^d$*,* $\phi = (\tilde{a}, \tilde{b}) \in \mathbb{R}^d \times \mathbb{R}^d$ *and* $A = \mathrm{diag}(\tilde{a})$*. Set* $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$*,* $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ *and* $q_\phi(z|x) = \mathcal{N}(z; Ax + \tilde{b}, 2/3\ \boldsymbol{I}_d)$ *(as in Rainforth et al., 2018). Then,*

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x) + \frac{d}{2}\left(\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right) - \frac{3\alpha}{4-\alpha}\left\|Ax + \tilde{b} - \frac{\theta + x}{2}\right\|^2$$

$$\gamma^{(\alpha)}(\theta, \phi; x)^2 = \frac{1}{1-\alpha}\left((4-\alpha)^d(15 - 6\alpha)^{-\frac{d}{2}} \exp\left(\frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)}\left\|Ax + \tilde{b} - \frac{\theta + x}{2}\right\|^2\right) - 1\right).$$

In addition, it holds that

$$\log\left(\frac{p_\theta(Z|x)}{q_\phi(Z|x)}\right) = -\log \mathbb{E}(\exp(-\sigma\sqrt{d} S)) - \sigma\sqrt{d} S, \quad S = \sum_{j=1}^{d} \frac{\xi_j}{\sigma\sqrt{d}}$$

with $\sigma^2 = \mathbb{V}(\xi_1) = 1/18 + 8\lambda^2/3 < \infty$ *and* $\xi_j = y_j^2/4 - 2\lambda y_j - 1/6$ *for all* $j = 1 \dots d$*, where* $\lambda = \left\|\frac{\theta+x}{2} - Ax - \tilde{b}\right\|/\sqrt{d}$*,* $y = (y_1, \dots, y_d) \sim \mathcal{N}(0, 2/3\boldsymbol{I}_d)$ *and* $\xi_1, \dots, \xi_d$ *are centered i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure.*

We next review the assumptions that need to be checked to obtain Examples 1 and 2.

C.2.1 Checking assumptions

First note that (i) the assumption (A1) holds with ν as the Lebesgue measure on $\mathbb{R}^d$ for both Examples 4 and 5 and (ii) $\theta \mapsto p_\theta(x)$, $(z,\theta) \mapsto p_\theta(z|x)$, and $(z,\phi) \mapsto q_\phi(z|x)$ are differentiable on Θ, $\mathbb{R}^d \times \Theta$ and $\mathbb{R}^d \times \Phi$, with, $\Theta = \Phi = \mathbb{R}^d$ for Example 4 ($p_\theta(x)$ is arbitrary in Example 4 since only $p_\theta(z|x)$ is given) and $\Theta = \mathbb{R}^d$ and $\Phi = \mathbb{R}^{2d}$ for Example 5 (here $(\phi_1, \dots, \phi_{2d}) = \left(\tilde{a}_1, \dots, \tilde{a}_d, \tilde{b}_1, \dots, \tilde{b}_d\right)$ since A is diagonal). Examples 1 and 2 build on the settings described in Examples 4 and 5 respectively, hence these assumptions also hold for Examples 1 and 2. In addition, (A2) holds with the reparametrization proposed in Example 1 (that is, $q = \mathcal{N}(0, \boldsymbol{I}_d)$ and $f(\varepsilon, \phi; x) = \varepsilon + \phi$) and the one proposed in Example 2 (that is, $q = \mathcal{N}(0, \boldsymbol{I}_d)$ and $f(\varepsilon, \phi; x) = \sqrt{\frac{2}{3}}\varepsilon + Ax + \tilde{b}$). The assumptions that remain to be checked are then (A3), $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$, $\mathbb{V}\left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$, $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, $\mathbb{V}\left(\left[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})\right]|_{\phi'=\phi}\right) < \infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right) < \infty$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$.

- **Checking (A3).** By (v) in Proposition 5, it is sufficient to prove that

$$\text{there exists } \eta > 0 \text{ such that } \quad \mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{-\eta}\right) < \infty. \tag{135}$$

- **Checking $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$, $\mathbb{V}\left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$.** Using the Hölder inequality, $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$, $\mathbb{V}\left(\partial_\psi \left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ hold if we can show that, for all $(\theta,\phi) \in \Theta \times \Phi$ and all real value $p > 0$, there exists a ψ-neighborhood $\mathcal{V}$ of (θ,ϕ) such that

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)^p\right) < \infty \tag{136}$$

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \left|\partial_{\psi'} \log \tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)\right|^p\right) < \infty, \tag{137}$$

where ψ is a given component of (θ,ϕ).

- **Checking $\mathbb{V}\left(\left[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})\right]|_{\phi'=\phi}\right) < \infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right) < \infty$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$.** Using the Hölder inequality once more, $\mathbb{V}\left(\left[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})\right]|_{\phi'=\phi}\right) < \infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right) < \infty$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$ hold if we can show that, for all $(\theta,\phi) \in \Theta \times \Phi$ and all real value $p > 0$, there exists a ψ-neighborhood $\mathcal{V}$ of (θ,ϕ) such

that (136) and (137) hold and, in addition,

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^p\right)<\infty \tag{138}$$

$$\mathbb{E}_{\varepsilon\sim\tilde{q}}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\log\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right|^p\right)<\infty \tag{139}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}}w_{\theta,\phi'}(Z;x)^p\right)<\infty \tag{140}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left(\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right)^p\right)<\infty \tag{141}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi}\log q_{\phi'}(Z|x)\right|^p\right)<\infty, \tag{142}$$

where ψ is a given component of ϕ.

When proving Examples 1 and 2, we will then notably show that the conditions (135)-(142) are satisfied, so that all the theorems of Section 3.3 apply to the settings described in Examples 4 and 5 with the reparameterizations of Examples 1 and 2.

C.2.2 Proof of Example 1

Recall that in this setting ψ is a component of ϕ, say ϕ_k for all $k=1\dots d$, $p_\theta(z|x)=\mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z|x)=\mathcal{N}(z;\phi,\boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix.

- **Checking (A3), $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)<\infty$, $\mathbb{V}\left(\partial_\psi\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right)<\infty$ ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$), $\mathbb{V}\left(\left[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})\right]|_{\phi'=\phi}\right)<\infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\big|_{\phi'=\phi}\right)<\infty$ and ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$).** Let us prove (135)-(142). To this end, observe that

$$\log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right)=-\frac{\|\theta-\phi\|^2}{2}-\langle z-\phi,\phi-\theta\rangle,\quad z\in\mathbb{R}^d. \tag{143}$$

Since $\langle Z-\phi,\phi-\theta\rangle$ is a Gaussian variable if $Z\sim q_\phi(\cdot|x)$, it admits exponential moments of any exponent and we have that (135) holds. We further get from (143) that $\log\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)=c_0(\theta,\phi,\phi')-\langle\varepsilon,\phi-\theta\rangle$ for some continuous function c_0 defined on $\Theta\times\Phi^2$. Conditions (136)–(142) follow using that ε is a Gaussian variable.

- **Closed-form expressions for the quantities appearing in Theorems 2 to 4.** First recall from Example 4 that

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x)=\ell(\theta;x)-\frac{\alpha\|\theta-\phi\|^2}{2},\quad \gamma^{(\alpha)}(\theta,\phi;x)^2=\frac{\exp\left((1-\alpha)^2\|\theta-\phi\|^2\right)-1}{1-\alpha}\,. \tag{144}$$

Computing $\partial_\psi\mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2]$ follows directly from differentiating in (144) w.r.t. ϕ_k:

$$\partial_{\phi_k}\mathrm{VR}^{(\alpha)}(\theta,\phi;x)=-\alpha(\phi_k-\theta_k) \tag{145}$$

$$\partial_{\phi_k}[\gamma^{(\alpha)}(\theta,\phi;x)^2]=2(1-\alpha)(\phi_k-\theta_k)\exp\left((1-\alpha)^2\|\phi-\theta\|^2\right). \tag{146}$$

Furthermore, using (143): for all $y > 0$ and all $\phi' \in \mathbb{R}^d$,

$$\log\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^y\right) = -\frac{y\|\theta-\phi\|^2}{2} - y\langle\varepsilon+\phi'-\phi,\phi-\theta\rangle$$
$$= yg(\theta,\phi,\phi') - y\langle\varepsilon,\phi-\theta\rangle, \tag{147}$$

where $g(\theta,\phi,\phi') = -\frac{\|\theta-\phi\|^2}{2} - \langle\phi'-\phi,\phi-\theta\rangle$. Setting $y = 1-\alpha$ in (147), we deduce that:

$$\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})} = \frac{\exp\left(-(1-\alpha)\langle\varepsilon,\phi-\theta\rangle\right)}{\mathbb{E}(\exp\left(-(1-\alpha)\langle\varepsilon,\phi-\theta\rangle\right))} = \frac{\exp\left(-(1-\alpha)\|\theta-\phi\|\tilde{S}\right)}{\exp\left(\frac{1}{2}(1-\alpha)^2\|\theta-\phi\|^2\right)} \tag{148}$$

with $\tilde{S} = \frac{\langle\varepsilon,\phi-\theta\rangle}{\|\theta-\phi\|} \sim \mathcal{N}(0,1)$. Hence,

$$\partial_{\phi_k}\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right) = -(1-\alpha)\left(\varepsilon^{(k)} + (1-\alpha)(\phi_k-\theta_k)\right)\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}$$
$$\left[\partial_{\phi'_k}\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})}\right)\right]\Bigg|_{\phi'=\phi} = 0,$$

where $\varepsilon^{(k)}$ denotes the k-th element of $\varepsilon \in \mathbb{R}^d$. Consequently,

$$\phi_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{E}\left(\left(\varepsilon^{(k)} + (1-\alpha)(\phi_k-\theta_k)\right)^2\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^2\right)$$
$$\phi_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = 0, \quad \alpha \in (0,1). \tag{149}$$

Thus, using (148), we have that: for all $\psi \in \{\theta_1,\ldots,\theta_d,\phi_1,\ldots,\phi_d\}$,

$$\phi_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathrm{e}^{-(1-\alpha)^2\|\theta-\phi\|^2}\mathbb{E}\left((\varepsilon^{(k)} + (1-\alpha)(\phi_k-\theta_k))^2\mathrm{e}^{-2(1-\alpha)\langle\varepsilon,\phi-\theta\rangle}\right)$$
$$= \mathrm{e}^{-(1-\alpha)^2\|\theta-\phi\|^2}\mathrm{e}^{2(1-\alpha)^2\|\theta-\phi\|^2}\left(1+(1-\alpha)^2(\phi_k-\theta_k)^2\right)$$
$$= \mathrm{e}^{(1-\alpha)^2\|\theta-\phi\|^2}\left(1+(1-\alpha)^2(\phi_k-\theta_k)^2\right). \tag{150}$$

Now considering the special case $\alpha = 0$, it holds that using (114):

$$\phi_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = \mathbb{V}\left(A_{\theta,\phi}(\varepsilon;x) - 2\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\mathbb{E}\left(A_{\theta,\phi}(\varepsilon;x)\right)\right)$$
$$= \mathbb{E}\left(\left(A_{\theta,\phi}(\varepsilon;x) - 2\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\mathbb{E}\left(A_{\theta,\phi}(\varepsilon;x)\right)\right)^2\right) - \mathbb{E}(A_{\theta,\phi}(\varepsilon;x))^2 \tag{151}$$

with

$$A_{\theta,\phi}(\varepsilon;x) = \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^2}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))^2}\left(\partial_{\phi_k}\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x) + \left[\partial_{\phi'_k}\log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\Big|_{\phi'=\phi}\right).$$

Since $\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)) = \exp(\ell(\theta;x))$ and thus does not depend on ϕ_k, we deduce that

$$A_{\theta,\phi}(\varepsilon;x) = \frac{1}{2}\partial_{\phi_k}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^2\right) + \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^2}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))^2}\left[\partial_{\phi'_k}\log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\Big|_{\phi'=\phi}. \tag{152}$$

Using (148) with $\alpha = 0$, taking the square and differentiating, we have that

$$\frac{1}{2}\partial_{\phi_k}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^2\right) = -\left(\varepsilon^{(k)} + \phi_k - \theta_k\right)\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^2.$$

Furthermore, $\left[\partial_{\phi'_k}\log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\Big|_{\phi'=\phi} = \left[\varepsilon^{(k)} + \phi_k - \phi'_k\right]\Big|_{\phi'=\phi} = \varepsilon^{(k)}$ and thus

$$A_{\theta,\phi}(\varepsilon;x) = (\theta_k - \phi_k)\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^2.$$

Consequently, since for all $\ell \in \mathbb{N}^\star$,

$$\mathbb{E}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^\ell\right) = \mathrm{e}^{\frac{\ell(\ell-1)}{2}\|\theta-\phi\|^2}, \tag{153}$$

we obtain that

$$\phi_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = (\theta_k - \phi_k)^2\mathrm{e}^{2\|\theta-\phi\|^2}\left(\mathrm{e}^{4\|\theta-\phi\|^2} - 4\mathrm{e}^{2\|\theta-\phi\|^2} + 4\mathrm{e}^{\|\theta-\phi\|^2} - 1\right). \tag{154}$$

Lastly, we obtain the desired expressions in Example 1 by taking $\theta - \phi = \epsilon \cdot \boldsymbol{u}_d$ with $\epsilon > 0$ in (145), (146), (149), (150) and applying Theorems 2 to 4.

C.2.3 Proof of Example 2

Recall that in this setting ψ is a component of $(\theta, \tilde{b})$, say θ_k or $\tilde{b}_k$ for some $k = 1 \dots d$, $p_\theta(z) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x;z,\boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z;Ax+\tilde{b},2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\tilde{a} = (\tilde{a}_1,\dots,\tilde{a}_d)$. We thus have that

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

$$p_\theta(x) = \int p_\theta(x,z)\mathrm{d}z = \mathcal{N}(x;\theta,2\boldsymbol{I}_d), \tag{155}$$

$$p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2,1/2\ \boldsymbol{I}_d)\ . \tag{156}$$

- **Checking (A3), $\mathbb{V}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) < \infty$, $\mathbb{V}\left(\partial_\psi\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)\right) < \infty$ ($\mathsf{A}^{\mathrm{REP}}_{\mathrm{df}}$), $\mathbb{V}\left([\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2(1-\alpha)})]|_{\phi'=\phi}\right) < \infty$, $\mathbb{V}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]|_{\phi'=\phi}\right) < \infty$ and ($\mathsf{A}^{\mathrm{DREP}}_{\mathrm{df}}$).** Let us prove (135)-(142). To this end, using the densities and conditional densities above, we obtain that

$$\log w_{\theta,\phi}(z;x) = C_1(\theta,\phi;x) - \frac{1}{4}\|z\|^2 + \langle z, \theta + x - \frac{3}{2}\left(Ax+\tilde{b}\right)\rangle\ ,$$

where $C_1(\cdot;x)$ is a $\mathcal{C}^\infty$ function from $\Theta\times\Phi$ to $\mathbb{R}$. For $Z\sim q_\phi(\cdot|x)$, it is a Gaussian vector with scalar covariance matrix equal to $2/3$ on its diagonal, and we get that (135) holds for any $\eta<3$. Next, using that $f(\varepsilon,\phi';x)=\sqrt{\frac{2}{3}}\varepsilon+A'x+\tilde{b}'$. we have

$$\log\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)=C_2(\theta,\phi,\phi';x)-\frac{1}{6}\|\varepsilon\|^2-\frac{1}{\sqrt{6}}\left\langle\varepsilon,A'x+\tilde{b}'-2\left(\theta+x\right)+3\left(Ax+\tilde{b}\right)\right\rangle\ ,$$

where $C_2(\cdot;x)$ is a $\mathcal{C}^\infty$ function from $\Theta\times\Phi^2$ to $\mathbb{R}$. Since $\tilde{q}=\mathcal{N}(0,\boldsymbol{I}_d)$, we easily get (136)–(139). Conditions (140)–(142) are obtained similarly.

- **Closed-form expressions for the quantities appearing in Theorems 2 to 4.** First recall from Example 5 that

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x)=\ell(\theta;x)+\frac{d}{2}\left(\log\left(\frac{4}{3}\right)+\frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right)-\frac{3\alpha}{4-\alpha}\left\|Ax+\tilde{b}-\frac{\theta+x}{2}\right\|^2$$

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

Computing $\partial_\psi\mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2]$ follows directly from differentiating the equations above w.r.t. θ/ϕ paired up with (155). so that

$$\partial_{\theta_k}\mathrm{VR}^{(\alpha)}(\theta,\phi;x)=\frac{x_k-\theta_k}{2}+\frac{3\alpha}{4-\alpha}\left(\tilde{a}_kx_k+\tilde{b}_k-\frac{\theta_k+x_k}{2}\right)$$

$$\partial_{\tilde{b}_k}\mathrm{VR}^{(\alpha)}(\theta,\phi;x)=-2\cdot\frac{3\alpha}{4-\alpha}\left(\tilde{a}_kx_k+\tilde{b}_k-\frac{\theta_k+x_k}{2}\right)$$

and

$$\partial_{\tilde{b}_k}[\gamma^{(\alpha)}(\theta,\phi;x)^2]=2\cdot\frac{24(1-\alpha)(4-\alpha)^{d-1}}{3^{\frac{d}{2}}(5-2\alpha)^{\frac{d}{2}+1}}\exp\left(\frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)}\left\|Ax+\tilde{b}-\frac{\theta+x}{2}\right\|^2\right)\times\left(\tilde{a}_kx_k+\tilde{b}_k-\frac{\theta_k+x_k}{2}\right).$$

As for computing $\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)$ above, we have that, for all $\phi,\phi'\in\mathbb{R}^{2d}$,

$$\frac{p_\theta(f(\varepsilon,\phi';x)|x)}{q_\phi(f(\varepsilon,\phi';x)|x)}=C\exp\left(\frac{3}{4}\|A'x+\tilde{b}'-Ax-\tilde{b}\|^2\right)\times\exp\left(-\frac{1}{2\cdot3}\left(\|\varepsilon\|^2+2\left\langle\varepsilon,\sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right)\ ,\quad(157)$$

where C does not depend on A nor $\tilde{b}$. Hence,

$$\mathbb{E}\left(\left(\frac{p_\theta(f(\varepsilon,\phi';x)|x)}{q_\phi(f(\varepsilon,\phi';x)|x)}\right)^{1-\alpha}\right)=C^{1-\alpha}\exp\left(\frac{3(1-\alpha)}{4}\|A'x+\tilde{b}'-Ax-\tilde{b}\|^2\right)\times\mathbb{E}\left(\exp\left(-\frac{1-\alpha}{2\cdot3}\left(\|\varepsilon\|^2+2\left\langle\varepsilon,\sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right)\right)\ ,$$

with

$$\mathbb{E}\left(\exp\left(-\frac{1-\alpha}{2\cdot 3}\left(\|\varepsilon\|^2+2\left\langle\varepsilon,\sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right)\right)$$
$$=\left(\frac{3}{4-\alpha}\right)^{\frac{d}{2}}\exp\left(\frac{(1-\alpha)^2}{4(4-\alpha)}\left\|A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right\|^2\right).$$

Consequently: for all $\phi,\phi'\in\mathbb{R}^{2d}$,

$$\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})}=\exp\left(-\frac{1-\alpha}{2\cdot 3}\left(\|\varepsilon\|^2+2\left\langle\varepsilon,\sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right)$$
$$\times\left(\frac{4-\alpha}{3}\right)^{d/2}\exp\left(-\frac{(1-\alpha)^2}{4(4-\alpha)}\left\|A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right\|^2\right). \quad (158)$$

Hence, for all $k=1,\ldots,d$,

$$\partial_{\theta_k}\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)$$
$$=(1-\alpha)\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)}+\frac{1-\alpha}{4-\alpha}[4(\tilde{a}_k x_k+\tilde{b}_k)-2(\theta_k+x_k)]\right)\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})} \quad (159)$$

and we also have

$$\partial_{\tilde{b}_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}=-2\partial_{\theta_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}, \quad (160)$$

where $\varepsilon^{(k)}$ denotes the k-th element of $\varepsilon\in\mathbb{R}^d$. Furthermore,

$$\left[\partial_{\tilde{b}'_k}\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})}\right]\Bigg|_{\phi'=\phi}=-\frac{1}{2}\partial_{\theta_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}. \quad (161)$$

In addition,

$$\theta_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$$
$$=\mathbb{E}\left(\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)}+\frac{1-\alpha}{4-\alpha}[4(\tilde{a}_k x_k+\tilde{b}_k)-2(\theta_k+x_k)]\right)^2\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^2\right) \quad (162)$$

and now using (158), we have that: for all $\ell\in\mathbb{N}^\star$,

$$\tilde{q}(\varepsilon)\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^\ell=\left(\frac{4-\alpha}{3}\right)^{d\ell/2}\exp\left(-\frac{(1-\alpha)^2\ell}{4(4-\alpha)}\|\mu\|^2\right)$$
$$(2\pi)^{-d/2}\exp\left(-\frac{1}{2}\|\varepsilon\|^2\right)\exp\left(-\frac{(1-\alpha)\ell}{2\cdot 3}\left(\|\varepsilon\|^2+2\left\langle\varepsilon,\sqrt{\frac{3}{2}}\mu\right\rangle\right)\right)$$

where for convenience we have denoted $\mu_k = 4(\tilde{a}_k x_k + \tilde{b}_k) - 2(\theta_k + x_k)$ for all $k = 1 \dots d$ and $\mu = (\mu_1, \dots, \mu_d)$. Thus,

$$\tilde{q}(\varepsilon) \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)^{1-\alpha})} \right)^{\ell} = h(\ell) \left(\frac{2\pi \cdot 3}{3 + (1-\alpha)\ell} \right)^{-d/2} \exp\left(-\frac{3 + (1-\alpha)\ell}{2 \cdot 3} \left\| \varepsilon + \frac{(1-\alpha)\ell}{3 + (1-\alpha)\ell} \sqrt{\frac{3}{2}} \mu \right\|^2 \right) \quad (163)$$

with

$$\begin{aligned} h(\ell) &= \left(\frac{4-\alpha}{3} \right)^{d\ell/2} \exp\left(-\frac{(1-\alpha)^2 \ell}{4(4-\alpha)} \|\mu\|^2 \right) \left(\frac{3}{3 + (1-\alpha)\ell} \right)^{d/2} \exp\left(\frac{(1-\alpha)^2 \ell^2}{4(3 + (1-\alpha)\ell)} \|\mu\|^2 \right) \\ &= \left(\frac{4-\alpha}{3} \right)^{d\ell/2} \exp\left(\frac{3(1-\alpha)^2 \ell(\ell-1)}{4(4-\alpha)(3 + (1-\alpha)\ell)} \|\mu\|^2 \right) \left(\frac{3}{3 + (1-\alpha)\ell} \right)^{d/2}. \end{aligned}$$

Consequently, taking $\ell = 2$ in (163) and plugging this in (162), we get that

$$\theta_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \frac{(4-\alpha)^d}{(15 - 6\alpha)^{d/2}} \mathrm{e}^{\frac{(1-\alpha)^2}{(4-\alpha)(5-2\alpha)} \frac{3}{2} \|\mu\|^2} \left(\frac{2}{5 - 2\alpha} + \frac{(1-\alpha)^2 3^2}{(5-2\alpha)^2 (4-\alpha)^2} \mu_k^2 \right),$$

from which we can also deduce $\tilde{b}_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ using (160). Similarly, using (161), we can deduce from the above $\tilde{b}_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ for all $\alpha \in (0, 1)$. Now considering the case $\alpha = 0$, we deduce using (159) and (161) with $\alpha = 0$ that:

$$\frac{1}{2} \partial_{\tilde{b}_k} \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right)^2 = -2 \left(\sqrt{\frac{2}{3}} \varepsilon^{(k)} + \frac{1}{4} \mu_k \right) \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right)^2.$$

Hence, $A_{\theta,\phi}(\varepsilon; x)$ as defined in (152) with $\phi_k = \tilde{b}_k$ becomes

$$A_{\theta,\phi}(\varepsilon; x) = -\frac{1}{2} \left(\sqrt{\frac{2}{3}} \varepsilon^{(k)} + \mu_k \right) \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right)^2,$$

where we have used that $\left[\partial_{\tilde{b}'_k} \log q_{\phi'}(f(\varepsilon, \phi; x)|x) \right]\Big|_{\phi'=\phi} = \sqrt{3/2}\ \varepsilon^{(k)}$. Now using (163) with $\alpha = 0$ and $\ell = 2$, we obtain that

$$\mathbb{E}(A_{\theta,\phi}(\varepsilon; x)) = -\frac{3}{2 \times 5} \frac{4^d}{15^{d/2}} \mathrm{e}^{\frac{3}{8 \times 5} \|\mu\|^2} \mu_k.$$

Next, using the above as well as (163) with $\alpha = 0$ and $\ell = 2$ and (151) with $\phi_k = \tilde{b}_k$, we can compute all the terms appearing in $\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x)$ since

$$\begin{aligned} \tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x) = \mathbb{E}(A_{\theta,\phi}(\varepsilon; x)^2) &- 4\mathbb{E}\left(A_{\theta,\phi}(\varepsilon; x) \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right) \mathbb{E}\left(A_{\theta,\phi}(\varepsilon; x) \right) \\ &+ \left(4\mathbb{E}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right)^2 \right) - 1 \right) \mathbb{E}(A_{\theta,\phi}(\varepsilon; x))^2 \end{aligned}$$

with

$$\mathbb{E}(A_{\theta,\phi}(\varepsilon;x)^2) = \frac{1}{4}\mathbb{E}\left(\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)} + \mu_k\right)^2 \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^4\right)$$
$$= \frac{1}{4}\left(\frac{4}{3}\right)^{2d}\left(\frac{3}{7}\right)^{d/2} \mathrm{e}^{\frac{9}{4\cdot 7}\|\mu\|^2}\left(\frac{2}{7} + \frac{3^2}{7^2}\mu_k^2\right)$$

and

$$\mathbb{E}\left(A_{\theta,\phi}(\varepsilon;x)\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right) = -\frac{1}{2}\mathbb{E}\left(\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)} + \mu_k\right)\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^3\right) = -\frac{1}{4}h(3)\mu_k.$$

Thus, we finally get

$$\begin{aligned}\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = {} & \frac{1}{4}\left(\frac{4}{3}\right)^{2d}\left(\frac{3}{7}\right)^{d/2}\mathrm{e}^{\frac{9}{4\cdot 7}\|\mu\|^2}\left(\frac{2}{7}+\frac{3^2}{7^2}\mu_k^2\right) \\ & - \left(\frac{4}{3}\right)^{3d/2}\left(\frac{1}{2}\right)^{d/2} e^{\frac{3}{4\cdot 4}\|\mu\|^2}\frac{3}{2\cdot 5}\frac{4^d}{15^{d/2}}\mathrm{e}^{\frac{3}{8\times 5}\|\mu\|^2}\mu_k^2 \\ & + \left(4\frac{4^d}{15^{d/2}}\mathrm{e}^{\frac{3}{8\cdot 5}\|\mu\|^2} - 1\right)\frac{3^2}{2^2\times 5^2}\frac{4^{2d}}{15^d}\mathrm{e}^{\frac{3}{4\times 5}\|\mu\|^2}\mu_k^2.\end{aligned} \tag{164}$$

Lastly, we obtain the desired expressions in Example 2 by taking $Ax + \tilde{b} = (\theta + x)/2 + \epsilon \boldsymbol{u}_d$ with $\epsilon > 0$ so that $\mu_k = 4\epsilon$ and $\|\mu\|^2 = 16d\epsilon^2$. Notice in particular that (164) becomes:

$$\begin{aligned}\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = {} & \frac{1}{4}\left(\frac{4}{3}\right)^{2d}\left(\frac{3}{7}\right)^{d/2}\mathrm{e}^{\frac{36}{7}d\epsilon^2}\left(\frac{2}{7}+\left(\frac{12\epsilon}{7}\right)^2\right) \\ & - \left(\frac{4}{3}\right)^{3d/2}\left(\frac{1}{2}\right)^{d/2} e^{3d\epsilon^2}\frac{3}{2\cdot 5}\frac{4^d}{15^{d/2}}\mathrm{e}^{\frac{6}{5}d\epsilon^2}\mu_k^2 + \left(4\frac{4^d}{15^{d/2}}\mathrm{e}^{\frac{6}{5}d\epsilon^2} - 1\right)\frac{3^2}{2^2\times 5^2}\frac{4^{2d}}{15^d}\mathrm{e}^{\frac{12}{5}d\epsilon^2}\mu_k^2.\end{aligned} \tag{165}$$

Appendix D. Deferred proofs and results for Section 3.4

D.1 L^p norms

For a real valued random variable X we denote by $\|X\|_p$ its L^p norm, that is, $\|X\|_\infty$ denotes the smallest a.s. upper bound of $|X|$ and, for all $1 \le p < \infty$,

$$\|X\|_p = \left(\mathbb{E}\left(|X|^p\right)\right)^{\frac{1}{p}} .$$

Recall that the L^p space of random variables with finite L^p-norms is a Banach space when endowed with this L^p-norm and that, by Jensen's inequality we have $\|X\|_p \le \|X\|_{p'}$ for all $p \le p'$. We will also repeatedly use the Hölder inequality

$$\|XY\|_1 \le \|X\|_p \|Y\|_{p'} ,$$

which holds for all $X \in L^p$ and $Y \in L^{p'}$ with $p, p' \in [1,\infty]$ such that $1/p + 1/p' = 1$.

D.2 Preliminary results for the proofs of Proposition 1

In the following we will use the asymptotic equivalence

$$1 - \Phi(u) = u^{-1}\varphi(u)(1 + o(1)) \quad \text{as } u \to \infty, \tag{166}$$

where φ denotes the standard Normal density and Φ denotes the cumulative distribution function of the standard Normal distribution. In fact it is easy to show that the o-term in this asymptotic equivalence is non-positive as soon as $u > 0$, since, in this case,

$$1 - \Phi(u) = \int_u^\infty \varphi(y)\, \mathrm{d}y \leq u^{-1} \int_u^\infty y\, \varphi(y)\, \mathrm{d}y = u^{-1}\varphi(u). \tag{167}$$

We then have the following two lemmas.

Lemma 5 *Let $Y_1, \ldots, Y_N$ be i.i.d. standard normal random variables and set $M_N = \max_{1 \leq i \leq N} Y_i$. Then for all $c > 0$, we have, as $N \to \infty$,*

$$\mathbb{P}\left(M_N > \sqrt{2\,(1+c)\,\log N}\right) = O\left((\log N)^{-1/2}\, N^{-c}\right), \tag{168}$$

$$\mathbb{P}\left(M_N < \sqrt{2\,(1+c)^{-1}\,\log N}\right) = O\left(\mathrm{e}^{-N^{c/(1+c)}/\sqrt{4\pi \log N}}\right). \tag{169}$$

Proof For any $u \in \mathbb{R}$ and $N \geq 1$, we have

$$\mathbb{P}(M_N > u) = 1 - \exp\left(N \log\left(1 - (1 - \Phi(u))\right)\right).$$

Then using (166) with $u = \sqrt{2x \log N}$ and $x > 0$, we have, as $N \to \infty$,

$$1 - \Phi(u) = \frac{N^{-x}}{\sqrt{4\pi\, x \log N}}\,(1 + o(1)).$$

We successively get the following assertions depending on where x lies w.r.t. 1

$$\text{if } x > 1, \text{ as } N \to \infty, \quad \mathbb{P}\left(M_N > \sqrt{2x \log N}\right) = \frac{N^{1-x}}{\sqrt{4\pi\, x\, \log N}}\,(1 + o(1)),$$

$$\text{if } x \in (0,1), \text{ as } N \to \infty, \quad \mathbb{P}\left(M_N \leq \sqrt{2x \log N}\right) = O\left(\exp\left(-N^{1-x}/\sqrt{4\pi \log N}\right)\right),$$

The two bounds (168) and (169) easily follow. ∎

The next lemma extends Daudel et al. (2023, Lemma 1), which follows from the case $m = 1$.

Lemma 6 *Let $Y_1, \ldots, Y_N$ be i.i.d. standard normal random variables and set $M_N = \max_{1 \leq i \leq N} Y_i$. Let $m \in [1, \infty)$. Then, as $N \to \infty$,*

$$\mathbb{E}\left(|M_N|^m\right) = (2 \log N)^{m/2}\left(1 + O\left(\frac{\log\log N}{\log N}\right)\right) \tag{170}$$

$$\mathbb{E}\left((M_N)_-^m\right) = O\left(\rho^N\right) \quad \textit{for any } \rho \in (1/2, 1), \tag{171}$$

where $(x)_- = \max(-x, 0)$ denotes the negative part of x.

Proof The Gaussian distribution belongs to the maximum domain of attraction of the Gumbel distribution. More precisely, we have (see de Haan and Ferreira, 2007)

$$\lim_{N\to\infty} \mathbb{P}\left(a_N^{-1}\left(M_N - b_N\right) \leq x\right) = \exp(-\mathrm{e}^{-x}),$$

with $a_N = 1/\sqrt{2\log N}$, $b_N = \sqrt{2\log N} - \frac{1}{2}(\log\log N + \log 4\pi)/(\sqrt{2\log N})$. Since the Gaussian distribution has finite moments of all orders and so does the Gumbel distribution, (Pickands III, 1968, Theorem 2.1) yields

$$\lim_{N\to\infty} \mathbb{E}\left(|L_N|^m\right) = \int_{-\infty}^{\infty} |x|^m \, \exp(-x - \mathrm{e}^{-x}) \, \mathrm{d}x,$$

where we have set $L_N = a_N^{-1}(M_N - b_N)$. Now writing $b_N^{-1} M_N = 1 + \frac{a_N}{b_N} L_N$ and since $m \in [1,\infty)$ by assumption, the Minkowski inequality yields

$$\left| \left(\mathbb{E}\left(\left|b_N^{-1} M_N\right|^m\right)\right)^{1/m} - 1 \right| \leq \left(\mathbb{E}\left(\left|b_N^{-1} M_N - 1\right|^m\right)\right)^{1/m} = \frac{a_N}{b_N} \, \left(\mathbb{E}\left(|L_N|^m\right)\right)^{1/m}.$$

Hence $\mathbb{E}\left(|M_N|^m\right) = b_N^m \left(1 + O\left(\frac{a_N}{b_N}\right)\right)^m$, which leads to (170). To get (171), we use that $\mathbb{E}\left((M_N)_-^m\right) = \mathbb{E}\left(|M_N|^m \, \mathbb{1}_{\{M_N<0\}}\right)$, and, for any $p > 1$, the Hölder inequality leads to

$$\mathbb{E}\left((M_N)_-^m\right) \leq \left(\mathbb{E}\left(|M_N|^{m\,p}\right)\right)^{1/p} \left(\mathbb{P}\left(M_N < 0\right)\right)^{(p-1)/p} = \left(\mathbb{E}\left(|M_N|^{m\,p}\right)\right)^{1/p} \left(\frac{1}{2}\right)^{N\,(p-1)/p}.$$

Since (170) also holds with m replaced by $m\,p$, we get the claimed bound by choosing p large enough to have $\left(\frac{1}{2}\right)^{(p-1)/p} \in (\frac{1}{2}, \rho)$ with $\rho \in (\frac{1}{2}, 1)$. ■

We next present a key proposition, which provides a general bound that will be used to show the collapse of self-normalized weighted averages (in the sense that one weight dominates over all others in the self-normalized weighted averages).

Proposition 7 *Let $\left(\left(W_i, \tilde{Y}_i\right)\right)_{i\geq 1}$ be a sequence of i.i.d. $\mathbb{R}_+^* \times \mathbb{R}$-valued random variables. Let F denote the distribution function of $\log W_1$, and its generalized inverse by $F^{\leftarrow}$. In addition, let $Z \sim \mathcal{N}(0,1)$ and let us define, for all $s \in \mathbb{R}$,*

$$\zeta(s) = \frac{\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z) - F^{\leftarrow}\circ\Phi(s)} \,\middle|\, Z \leq s\right]}{1 - \Phi\left(s\right)}. \tag{172}$$

Let us further define, for all $N \geq 1$, $\tilde{Y}_N^* = \max\left(\left|\tilde{Y}_1\right|, \ldots, \left|\tilde{Y}_N\right|\right)$ and for all $u \in \mathbb{R}_+$, $s \in \mathbb{R}$, $m \in [1,\infty)$ and $p \in (1,\infty)$,

$$\tilde{\zeta}_N^{(m,p)}(u,s) = N\left(\left\|\tilde{Y}_1\right\|_{p\,m} \, s^{(p-1)/(pm)} + \left\|\tilde{Y}_1 \, \mathbb{1}_{\left\{\tilde{Y}_N^* > u\right\}}\right\|_m\right). \tag{173}$$

Then, there exists a non-decreasing random sequence $(J_N)_{N\geq 1}$ such that, for all $N \geq 1$, J_N is valued in $\{1,\ldots,N\}$ with $W_{J_N} = \max\left(W_1,\ldots,W_N\right)$ and, for all $m, \delta \in [1,\infty)$,

$0 < \underline{b} < \sqrt{2} < \overline{b}$, $p \in (1, \infty)$, $u > 0$ and $N \in \mathbb{N}^*$, we have

$$\left\| \sum_{i=1}^{N} \overline{W}_{i,N}^{\delta} \tilde{Y}_i - \tilde{Y}_{J_N} \right\|_m \leq C \left(u \sup_{\substack{b \in [\underline{b}, \overline{b}] \\ q=1,m}} \left(\zeta(b\sqrt{\log N}) \right)^{\frac{q}{m}} + \tilde{\zeta}_N^{(m,p)} \left(u, (\log N)^{-\frac{1}{2}} \; N^{1-\frac{\underline{b}^2}{2}} \right) \right) , \tag{174}$$

where $C > 0$ is a constant only depending on p, $\underline{b}$, $\overline{b}$, δ and m, and $\overline{W}_{i,N}$ denotes the self-normalized weight defined by

$$\overline{W}_{i,N} := \left(\sum_{1 \leq k \leq N} W_k \right)^{-1} W_i \, , \qquad 1 \leq i \leq N.$$

Proof Let us denote by Q the probability kernel of a regular version of the conditional distribution of $\tilde{Y}_1$ given W_1 (which exists since the real line is a Polish space, see (Douc et al., 2018, Theorem B.3.11)), that is $\mathbb{P}\left[\tilde{Y}_1 \leq \tilde{y} \,\middle|\, W_1 \right] = Q(W_1, (-\infty, \tilde{y}])$ $\mathbb{P}$-a.s. for all $\tilde{y} \in \mathbb{R}$. Further denote the associated conditional generalized inverse by $Q^{\leftarrow}$, that is

$$\forall w \in \mathbb{R}_+^*, \, \forall u \in (0,1), \quad Q^{\leftarrow}(w,u) = \inf \left\{ \tilde{y} \in \mathbb{R} \, : \, Q(w, (-\infty, \tilde{y}]) \geq u \right\}.$$

Recall that $Z \sim \mathcal{N}(0,1)$ and let U be independent of Z with uniform distribution on $[0,1]$. By definition of F and Q, $\left(W_1, \tilde{Y}_1\right)$ have the same distribution as $\left(\mathrm{e}^{F^{\leftarrow} \circ \Phi(Z)}, Q^{\leftarrow}(\mathrm{e}^{F^{\leftarrow} \circ \Phi(Z)}, U) \right)$. Therefore, from now on, we let $(Z_i)_{i \geq 1}$ and $(U_i)_{i \geq 1}$ be two independent i.i.d. sequence distributed as Z and U, respectively and set, without loss of generality,

$$\forall i \geq 1, \quad W_i = \mathrm{e}^{F^{\leftarrow} \circ \Phi(Z_i)} \quad \text{and} \quad \tilde{Y}_i = Q^{\leftarrow}(W_i, U_i) \, . \tag{175}$$

Now denoting by I_N the (random) index in $\{1, \ldots, N\}$ such that $Z_{I_N} = \max(Z_1, \ldots, Z_N)$, we have that I_N is a.s. uniquely defined since $Z_1, \ldots, Z_N$ are independent and Φ is continuous. Furthermore, as required in the proposition, $(I_N)_{N \geq 1}$ is a non-decreasing random sequence such that: for all $N \geq 1$, I_N is valued in $\{1, \ldots, N\}$ with $W_{I_N} = \max(W_1, \ldots, W_N)$.

Next let $m, \delta \in [1, \infty)$, $0 < \underline{b} < \sqrt{2} < \overline{b}$ and $p \in (1, \infty)$. It remains to show that there exists a constant $C > 0$ only depending on these constants such that the bound (174) holds for all $u > 0$ and $N \in \mathbb{N}^*$. To this end, we first write

$$\sum_{i=1}^{N} \left(\overline{W}_{i,N} \right)^{\delta} \tilde{Y}_i = \tilde{Y}_{I_N} - A_N + B_N \, ,$$

where, denoting $J_N := \{1, \ldots, N\} \setminus \{I_N\}$, we set

$$A_N := \left(1 - \left(\overline{W}_{I_N,N} \right)^{\delta} \right) \tilde{Y}_{I_N} \quad \text{and} \quad B_N := \sum_{i \in J_N} \left(\overline{W}_{i,N} \right)^{\delta} \tilde{Y}_i \, .$$

To evaluate the left-hand side of (174), that is $\left\| \sum_{i=1}^{N} \overline{W}_{i,N}^{\delta} \tilde{Y}_i - \tilde{Y}_{I_N} \right\|_m$, the idea is to separate the integration domain in $\left\{ \underline{b} \leq \frac{Z_{I_N}}{\sqrt{\log N}} \leq \overline{b} \right\} \cap \left\{ \tilde{Y}_N^* \leq u \right\}$ and its complementary set. More

specifically, to prove (174) it suffices to show that

$$\mathbb{E}\left(|A_N|^m \, \mathbb{1}_{\left\{\underline{b}\leq \frac{z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}\cap\left\{\tilde{Y}_N^*\leq u\right\}}\right) \leq \delta^m \, C_2 \, u^m \sup_{b\in\left[\underline{b},\bar{b}\right],q=1,m} \left(\zeta(b\sqrt{\log N})\right)^q \tag{176}$$

$$\mathbb{E}\left(|B_N|^m \, \mathbb{1}_{\left\{\underline{b}\leq \frac{z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}\cap\left\{\tilde{Y}_N^*\leq u\right\}}\right) \leq C_2 \, u^m \sup_{b\in\left[\underline{b},\bar{b}\right],q=1,m} \left(\zeta(b\sqrt{\log N})\right)^q \tag{177}$$

$$\mathbb{E}\left(\left(\sum_{i=1}^N \left|\tilde{Y}_i\right|\right)^m \mathbb{1}_{\left\{\frac{z_{I_N}}{\sqrt{\log N}}\notin\left[\underline{b},\bar{b}\right]\right\}\cup\left\{\tilde{Y}_N^*>u\right\}}\right) \leq C_1 \, \left(\tilde{\zeta}_N^{(m,p)}(u,(\log N)^{-\frac{1}{2}} \, N^{1-\bar{b}^2/2})\right)^m, \tag{178}$$

where C_1 only depends on p, $\underline{b}$ and $\bar{b}$ and C_2 only depends on m. The desired result will then follow by combining (176), (177), the Minkowski inequality, (178) and the fact that since $|A_N| \leq \left|\tilde{Y}_{I_N}\right|$ and $|B_N| \leq \sum_{i\in J_N}\left|\tilde{Y}_i\right|$, we have $|A_N|+|B_N| \leq \sum_{i=1}^N \left|\tilde{Y}_i\right|$. Let us thus prove separately that (i) the inequalities (176) and (177) hold and (ii) the inequality (178) holds.

(i) *Proof of (176) and (177).* For all $N \geq 2$, set

$$D_N = \sum_{i\in J_N} \frac{W_i}{W_{I_N}}. \tag{179}$$

Then, for all $N \geq 2$,

$$|A_N| = \left(1-(1+D_N)^{-\delta}\right)\left|\tilde{Y}_{I_N}\right| \leq \delta \, D_N \, \tilde{Y}_N^*,$$

where in the last inequality we used the definition of $\tilde{Y}_N^*$ and the fact that $1-(1+x)^{-\delta} \leq \delta x$ for all $\delta \geq 1$ and $x \geq 0$. Furthermore, for all $N \geq 2$,

$$|B_N| \leq (1+D_N)^{-\delta} \sum_{i\in J_N}\left(\frac{W_i}{W_{I_N}}\right)^\delta \left|\tilde{Y}_i\right| \leq \sum_{i\in J_N}\left(\frac{W_i}{W_{I_N}}\right)^\delta \left|\tilde{Y}_i\right| \leq D_N \, \tilde{Y}_N^* \, ,$$

where we used that $\delta \geq 1$, and $W_i \leq W_{I_N}$ for all $i \in J_N$. To get (176) and (177), notice that

$$\mathbb{E}\left(D_N^m \, (\tilde{Y}_N^*)^m \, \mathbb{1}_{\left\{\underline{b}\leq \frac{z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}\cap\left\{\tilde{Y}_N^*\leq u\right\}}\right) \leq u^m \mathbb{E}\left(D_N^m \, \mathbb{1}_{\left\{\underline{b}\leq \frac{z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}}\right),$$

thus it only remains to show that there exists $C_2 > 0$ which only depends on m such that

$$\mathbb{E}\left(D_N^m \, \mathbb{1}_{\left\{\underline{b}\leq \frac{z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}}\right) \leq C_2 \sup_{b\in\left[\underline{b},\bar{b}\right],q=1,m} \left(\zeta(b\sqrt{\log N})\right)^q. \tag{180}$$

By definition of D_N in (179) and using (175), we have that

$$D_N = \sum_{k\in J_N} \mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)-F^{\leftarrow}\circ\Phi(Z_{I_N})}$$

hence conditioning on Z_{I_N}, we get

$$\mathbb{E}\left(D_N^m \mathbb{1}_{\left\{\underline{b}\le \frac{Z_{I_N}}{\sqrt{\log N}}\le \bar{b}\right\}}\right)$$
$$= \mathbb{E}\left(\mathbb{1}_{\left\{\underline{b}\le \frac{Z_{I_N}}{\sqrt{\log N}}\le \bar{b}\right\}} \mathbb{E}\left[\left(\sum_{i=1}^{N-1} \mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)-F^{\leftarrow}\circ\Phi(s)}\right)^m \,\middle|\, Z_i \le Z_{I_N}\,,\, 1\le i\le N-1\right]\right). \quad (181)$$

Since the $Z_1,\dots,Z_N$ are i.i.d. with the same distribution as Z, by the Rosenthal inequality (see Theorem 2.12 Petrov, 1995), we have, for a constant $C_4>0$ only depending on m,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{N-1} \mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)-F^{\leftarrow}\circ\Phi(s)}\right)^m \,\middle|\, Z_i \le Z_{I_N}\,,\, 1\le i\le N-1\right] \\
&\qquad \le C_4\,\Big[(N-1)\,\mathbb{E}\left[\mathrm{e}^{m\,(F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s))}\,\middle|\, Z\le Z_{I_N}\right] \\
&\qquad\quad + \Big((N-1)\,\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s)}\,\middle|\, Z\le Z_{I_N}\right]\Big)^m\Big] \\
&\qquad \le C_4 \sum_{q=1,m} \Big((N-1)\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s)}\,\middle|\, Z\le Z_{I_N}\right]\Big)^q,
\end{aligned}$$

where, in the last inequality, we used that $m\ge 1$ and that the exponent of the exponential in the first line is non-positive when conditioning on $\{Z\le Z_{I_N}\}$. The expectation appearing in the last line is equal to $(1-\Phi(Z_{I_N}))\,\zeta(Z_{I_N})$. Using this with (181), we get that

$$\begin{aligned}
\mathbb{E}\left(D_N^m \mathbb{1}_{\left\{\underline{b}\frac{Z_{I_N}}{\sqrt{\log N}}\le \bar{b}\right\}}\right) &\le C_4 \sum_{q=1,m} \mathbb{E}\left(\mathbb{1}_{\left\{\underline{b}\le \frac{Z_{I_N}}{\sqrt{\log N}}\le \bar{b}\right\}}\,\{(N-1)(1-\Phi(Z_{I_N}))\,\zeta(Z_{I_N})\}^q\right) \\
&\le 2\,C_4 \max_{q=1,m}\left((N-1)^q\,\mathbb{E}\left((1-\Phi(Z_{I_N}))^q\right) \sup_{b\in[\underline{b},\bar{b}]}\left(\zeta(b\sqrt{\log N})\right)^q\right).
\end{aligned}$$

Since Φ is continuous, we further have that, for any $q\ge 1$,

$$\mathbb{E}\left((1-\Phi(Z_{I_N}))^q\right) = \mathbb{E}\left((1-U_{(N,N)})^q\right) = N\int_0^1 (1-x)^q\,x^{N-1}\,\mathrm{d}x = \frac{\Gamma(q+1)\,N!}{\Gamma(N+q+1)} \le \frac{\Gamma(q+1)}{(N+1)^q}\,,$$

where $U_{(N,N)}$ is the maximum of N i.i.d. random variables with uniform distribution in $[0,1]$. Hence (180) holds with $C_2 = 2\,C_4\,\Gamma(m+1)$, which concludes the proof of (176) and (177).

(ii) *Proof of (178).* Since $m\ge 1$, $x\mapsto x^m$ is convex and we have $\left(\frac{1}{N}\sum_{i=1}^N \left|\tilde{Y}_i\right|\right)^m \le \frac{1}{N}\sum_{i=1}^N \left|\tilde{Y}_i\right|^m$, leading to

$$\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^N \left|\tilde{Y}_i\right|\right)^m \mathbb{1}_{\mathbf{D}_N}\right) &\le N^{m-1}\sum_{i=1}^N \mathbb{E}\left(\left|\tilde{Y}_i\right|^m \mathbb{1}_{\mathbf{D}_N}\right) \\
&= N^m\,\mathbb{E}\left(\left|\tilde{Y}_1\right|^m \mathbb{1}_{\mathbf{D}_N}\right),
\end{aligned}$$

where we used that the domain $\mathbf{D}_N = \left\{ \frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b}, \overline{b}\right] \right\} \cup \left\{ \tilde{Y}_N^* > u \right\}$ is stable by permutation of the indices i within $\{1, \dots, N\}$ so that $\left|\tilde{Y}_i\right|^m \mathbb{1}_{\mathbf{D}_N}$ has the same distribution as $\left|\tilde{Y}_1\right|^m \mathbb{1}_{\mathbf{D}_N}$ for all $i = 1, \dots, N$. Now we note that $\mathbb{1}_{\mathbf{D}_N} \leq \mathbb{1}_{\left\{ \frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b}, \overline{b}\right] \right\}} + \mathbb{1}_{\left\{ \tilde{Y}_N^* > u \right\}}$ and, by definition of $\tilde{\zeta}_N^{(m,p)}$ in (173), to get (178) with $C_1 = \max\left(1, C_3^{(p-1)/p}\right)$, it only remains to show that

$$\mathbb{E}\left(\left|\tilde{Y}_i\right|^m \mathbb{1}_{\left\{ \frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b}, \overline{b}\right] \right\}} \right) \leq \left(\mathbb{E}\left(\left|\tilde{Y}_1\right|^{p\, m} \right) \right)^{1/p} \left(C_3 \, (\log N)^{-1/2} \, N^{1 - \overline{b}^2/2} \right)^{(p-1)/p}, \quad (182)$$

for a certain $C_3 > 0$ only depending on $\underline{b}$ and $\overline{b}$. Using Lemma 5 with $c = \overline{b}^2/2 - 1$ and $c = 2/\underline{b}^2 - 1$ successively in (168) and (169), we indeed find for such a constant C_3,

$$\mathbb{P}\left(\frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b}, \overline{b}\right] \right) \leq C_3 \left((\log N)^{-1/2} \, N^{1 - \overline{b}^2/2} \right).$$

Then we get (182) by applying the Hölder inequality and the proof of (178) is concluded.

■

Remark 7 *Proposition 7 investigates the behavior of the weighted average of* $\left(\tilde{Y}_i\right)_{i=1,\dots,N}$ *when using self-normalized positive weights* $\left(\overline{W}_{i,N}\right)_{i=1,\dots,N}$ *to some power* $\delta \geq 1$. *More precisely, (174) provides a bound of the* L^m*-norm of the error when approximating this average by a single* $\tilde{Y}_{I_N}$*, with* I_N *corresponding to an index with maximal weight.*

Under standard moment conditions, and using the law of large numbers, such an average should be well approximated as $N \to \infty$ *by* $N^{1-\delta} \mathbb{E}\left(W_1^\delta \tilde{Y}_1\right) \left(\mathbb{E}\left(W_1\right)\right)^{-\delta}$ *rather than by the (random)* $\tilde{Y}_{I_N}$ *as in (174). However, we will apply* (174) *with* F *depending on* N *in such a way that the maximal weight* W_{I_N} *tends to dominate over all other weights.*

In particular, in our applications of Proposition 7, we will take advantage of the fact that the constant C *does not depend on* F *in the bound (174). Furthermore, since (174) holds for any* $u > 0$*, we will choose* u *for a given* N *so that the right-hand side of (174) is as small as possible. This will be achieved by compromising between the* u *in the first term (which increases as* u *increases) and the second expectation in (173) defining* $\tilde{\zeta}_N^{(m,p)}(u, s)$ *which decreases as* u *increases. Finally, the first term between the parentheses of (173) will be made small by taking* $\overline{b}$ *large enough to compensate the* N *in front of the parentheses.*

D.3 Additional notation, useful first properties and derivations

In the remainder of Appendix D, we let $\varepsilon \sim \tilde{q}$ and $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. copies of ε and we assume that (B1) holds and so does ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$). We now introduce the following helpful notation:

$$\Lambda_{\theta,\phi}(\cdot; x) = \log \tilde{w}_{\theta,\phi,\phi}(\cdot; x) \tag{183}$$

$$B(\theta, \phi; x)^2 = \mathbb{V}_{\varepsilon\sim\tilde{q}}\left(\Lambda_{\theta,\phi}(\varepsilon; x)\right) = 2\mathrm{KL}\left(\theta, \phi; x\right) \tag{184}$$

$$\overline{\mathcal{W}}_{j,N}^{(\alpha)} = \left(\sum_{k=1}^{N} \mathrm{e}^{(1-\alpha)\,\Lambda_{\theta,\phi}(\varepsilon_k; x)}\right)^{-1} \mathrm{e}^{(1-\alpha)\,\Lambda_{\theta,\phi}(\varepsilon_j; x)}\,, \qquad 1 \le j \le N \tag{185}$$

$$\widetilde{\Lambda}_{\theta,\phi}(\cdot; x) = \frac{\Lambda_{\theta,\phi}(\cdot; x) - \mathbb{E}\left(\Lambda_{\theta,\phi}(\varepsilon; x)\right)}{B(\theta, \phi; x)} \tag{186}$$

$$\widetilde{B}^{(\mathrm{REP})}(\theta, \phi; x) = \sqrt{\mathbb{V}\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right)}\,. \tag{187}$$

Under (B1), the expectation and variance of $\Lambda_{\theta,\phi}(\varepsilon; x)$ are well defined. Furthermore, $\widetilde{\Lambda}_{\theta,\phi}$ is well defined in (186) if $B(\theta, \phi; x) > 0$ and $\widetilde{B}^{(\mathrm{REP})}(\theta, \phi; x)^2$ is well defined if $(\theta', \phi') \mapsto \widetilde{\Lambda}_{\theta',\phi'}(\varepsilon; x)$ is well defined and differentiable in a neighborhood of (θ, ϕ). Using Proposition 4, all these conditions hold providing that $B(\theta, \phi; x) > 0$ (or equivalently $\mathrm{KL}\left(\theta, \phi; x\right) > 0$), which we assume in the following since we are interested in the setting where $N, \mathrm{KL}\left(\theta, \phi; x\right) \to \infty$. More precisely, we get the following assertions:

(i) We have that $\partial_\psi B(\theta, \phi; x)$ is well defined;

(ii) Since $B(\theta, \phi; x) > 0$, $(\theta', \phi') \mapsto \widetilde{\Lambda}_{\theta',\phi'}(\varepsilon; x)$ is well defined and differentiable in a neighborhood of (θ, ϕ);

(iii) We have that $\widetilde{B}^{(\mathrm{REP})}(\theta, \phi; x)$ is well defined;

(iv) We have that $\left(\left(\widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x), \partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a Gaussian process;

(v) We can interchange the derivative ∂_ψ with the expectation sign for $\widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)$ as well as its square.

Using Assertion (v) and since $\widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)$ is centered with variance 1, we then get that

$$\mathbb{E}\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right) = \mathbb{C}\mathrm{ov}\left(\widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x), \partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right) = 0\,.$$

Combining this with Assertion (iv), we thus have the following useful property:

($\mathrm{P}^{\mathrm{REP}}$) $\left(\widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ and $\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ are two independent and centered Gaussian processes.

Plugging in the notation we introduced in (14) and using straightforward algebra, the REP gradient estimator reads

$$\begin{aligned}\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi}(\varepsilon; x)) + \partial_\psi B(\theta, \phi; x) \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j; x) \\ + B(\theta, \phi; x) \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j; x)\,.\end{aligned} \tag{188}$$

Furthermore, differentiating $B^2(\theta,\phi;x) = \mathbb{V}\left(\Lambda_{\theta,\phi}(\varepsilon;x)\right)$, we get that

$$B(\theta,\phi;x)\,\partial_\psi B(\theta,\phi;x) = \frac{1}{2}\partial_\psi B^2(\theta,\phi;x) = \mathbb{C}\mathrm{ov}\left(\Lambda_{\theta,\phi}(\varepsilon;x), \partial_\psi \Lambda_{\theta,\phi}(\varepsilon;x)\right) \ .$$

Next differentiating the product $B(\theta,\phi;x)\widetilde{\Lambda}_{\theta,\phi}(\varepsilon;x)$, which equals $\Lambda_{\theta,\phi}(\varepsilon;x)$ up to an added non-random constant, we obtain the sum of two uncorrelated variables and taking the variance leads to

$$\mathbb{V}\left(\partial_\psi \Lambda_{\theta,\phi}(\varepsilon;x)\right) = \left(\partial_\psi B(\theta,\phi;x)\right)^2 + \left(B(\theta,\phi;x)\,\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2 \ . \tag{189}$$

Since we can rewrite the correlation from (40) as

$$\mathrm{C}_\psi(\theta,\phi;x) = \frac{\mathbb{C}\mathrm{ov}\left(\Lambda_{\theta,\phi}(\varepsilon;x), \partial_\psi \Lambda_{\theta,\phi}(\varepsilon;x)\right)}{\sqrt{B(\theta,\phi;x)^2 \mathbb{V}\left(\partial_\psi \Lambda_{\theta,\phi}(\varepsilon;x)\right)}}$$

we also get using the expressions above that

$$\mathrm{C}_\psi(\theta,\phi;x) = \frac{\partial_\psi B(\theta,\phi;x)}{\sqrt{\left(\partial_\psi B(\theta,\phi;x)\right)^2 + \left(B(\theta,\phi;x)\,\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2}} \ . \tag{190}$$

To conclude this section we provide expression for $\mathrm{SNR}[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ as well as $\mathrm{SNR}[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \partial_\psi \ell(\theta;x)]$, where the latter is linked to $\mathrm{C}_\psi(\theta,\phi;x)$. Using (188) with $N=1$, we obtain that

$$\begin{aligned}\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi}(\varepsilon;x)) \\ + \partial_\psi B(\theta,\phi;x)\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_1;x) + B(\theta,\phi;x)\,\partial_\psi \widetilde{\Lambda}_{\theta,\phi}(\varepsilon_1;x) \ .\end{aligned} \tag{191}$$

It follows from (191) and Property ($\mathrm{P}^{\mathrm{REP}}$) that

$$\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \left(\partial_\psi B(\theta,\phi;x)\right)^2 + \left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2 \tag{192}$$

$$\mathrm{SNR}[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{\left|\partial_\psi \mathbb{E}(\Lambda_{\theta,\phi}(\varepsilon;x))\right|}{\sqrt{\left(\partial_\psi B(\theta,\phi;x)\right)^2 + \left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2}} \tag{193}$$

and, using (37) and (190), which leads to

$$\mathrm{SNR}[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \partial_\psi \ell(\theta;x)] = B(\theta,\phi;x)\ \left|\mathrm{C}_\psi(\theta,\phi;x)\right| \ . \tag{194}$$

Notice in particular that we also have

$$\begin{aligned}\frac{\partial_\psi \ell(\theta;x) - \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))}{\sqrt{\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))}} &= \frac{B(\theta,\phi;x)\partial_\psi B(\theta,\phi;x)}{\sqrt{\left(\partial_\psi B(\theta,\phi;x)\right)^2 + \left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2}} \\ &= B(\theta,\phi;x)\,\mathrm{C}_\psi(\theta,\phi;x)\end{aligned} \tag{195}$$

D.4 Proof of Proposition 1 and Theorem 5

We start with two lemmas.

Lemma 7 *Let $Z \sim \mathcal{N}(0,1)$. Then, for all $s \in \mathbb{R}$ and $\sigma > 0$,*

$$\mathbb{E}\left(\mathrm{e}^{\sigma(Z-s)} \,|\, Z \le s\right) = \frac{\Phi(s-\sigma)}{\Phi(s)} \, \frac{\varphi(s)}{\varphi(\sigma - s)} \,. \tag{196}$$

It follows that, defining ζ by (172) with $F^{\leftarrow} = \mu + \sigma\Phi^{-1}$ where $\mu \in \mathbb{R}$ and $\sigma > 0$, there exists a universal constant $C > 0$ such that, for all $\sigma \ge 2s \ge 1$,

$$\zeta(s) \le C \, \frac{s}{\sigma} \,. \tag{197}$$

Proof We have, for all $s \in \mathbb{R}$ and $\sigma > 0$,

$$\begin{aligned}
\mathbb{E}\left(\mathrm{e}^{\sigma(Z-s)} \,|\, Z \le s\right) &= \frac{1}{\Phi(s)} \, \mathbb{E}\left(\mathrm{e}^{\sigma(Z-s)} \, \mathbb{1}_{\{Z \le s\}}\right) \\
&= \frac{1}{\Phi(s)} \, \mathbb{E}\left(\frac{\varphi(s)\,\varphi(Z-\sigma)}{\varphi(\sigma-s)\,\varphi(Z)} \mathbb{1}_{\{Z \le s\}}\right) \\
&= \frac{1}{\Phi(s)} \, \frac{\varphi(s)}{\varphi(\sigma-s)} \, \mathbb{E}\left(\mathbb{1}_{\{Z+\sigma \le s\}}\right),
\end{aligned}$$

where, in the last line, we used that $Z + \sigma$ has density $u \mapsto \varphi(u - \sigma)$ hence the likelihood ratio $\varphi(Z-\sigma)/\varphi(Z)$ amounts to change Z into $Z + \sigma$. We thus get (196). Now, defining ζ by (172) with $F^{\leftarrow} = \mu + \sigma\Phi^{-1}$, we get, for all $s > 0$ and $\sigma > 0$,

$$\zeta(s) = \frac{1}{\Phi(s)} \frac{m(\sigma - s)}{m(s)},$$

where the function m corresponds to the Mills ratio $m(u) = \frac{1-\Phi(u)}{\varphi(u)}$ with $u > 0$. Since

$$\forall u > 0, \quad \frac{u}{u^2+1} < m(u) < \frac{1}{u},$$

(see Gordon, 1941) we finally deduce that: for all $\sigma > 2s \ge 1$,

$$\zeta(s) \le \frac{1}{\Phi(1)} \frac{s^2+1}{s} \frac{1}{\sigma - s} \le C \, \frac{s}{\sigma},$$

where $C = 6\Phi(1)^{-1}$ and where we have used that $s \ge 1$ and $s \le \sigma/2$. ∎

Lemma 8 *Let $\xi_1, \xi_2, \ldots$ be i.i.d. standard normal random variables and set, for all $N = 1, 2, \ldots,$ and $\beta \in \mathbb{R}$,*

$$\overline{W}_{j,N}(\beta) := \left(\sum_{k=1}^{N} \mathrm{e}^{\beta \xi_k}\right)^{-1} \mathrm{e}^{\beta \xi_j} \,, \qquad 1 \le j \le N \,. \tag{198}$$

Let $\delta \geq 1$, $\lambda \in [0,1]$ and $q \geq 1$. Then we have, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$,

$$\mathbb{E}\left(\sum_{j=1}^{N}(\overline{W}_{j,N}(\beta))^{\delta}\,\xi_j\right) = (2\,\log N)^{1/2}\,(1+o(1))\,, \tag{199}$$

$$\mathbb{V}\left(\sum_{j=1}^{N}[\lambda\overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\,\xi_j\right) = o\,(\log N)\,, \tag{200}$$

$$\mathbb{E}\left(\sum_{j=1}^{N}\left(\overline{W}_{j,N}(\beta)\right)^{\delta}\right) = 1 + o(1). \tag{201}$$

Proof We will first show that, for any $m \in [1,\infty)$, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$,

$$\left\|\sum_{i=1}^{N}(\overline{W}_{i,N}(\beta))^{\,\delta}\,\xi_i - \max(\xi_1,\ldots,\xi_N)\right\|_m = o\left(\sqrt{\log N}\right)\,, \tag{202}$$

$$\left\|\sum_{i=1}^{N}(\overline{W}_{i,N}(\beta))^{\,\delta} - 1\right\|_m = O\left(\left(\frac{\sqrt{\log N}}{\beta}\right)^{\frac{1}{m}} + N^{-\frac{q}{m}}\right)\,. \tag{203}$$

Then we will show that (202) implies (199)-(200), and that (203) implies (201).

(i) *Proof of (202) and (203).* We use Proposition 7 successively in the two settings

Setting 1) for all $j = 1\ldots N$, $W_j = \mathrm{e}^{\beta\,\xi_j}$ and $\tilde{Y}_j = \xi_j$;

Setting 2) for all $j = 1\ldots N$, $W_j = \mathrm{e}^{\beta\,\xi_j}$ and $\tilde{Y}_j = 1$.

In both settings, we have $F^{\leftarrow}(u) = \beta\,\Phi^{-1}(u)$. Then Proposition 7 gives us that there exists a non-decreasing random sequence $(I_N)_{N\geq 1}$ such that, for all $N \geq 1$, I_N is valued in $\{1,\ldots,N\}$ with $W_{I_N} = \max(W_1,\ldots,W_N)$ and, for all $m,\delta \in [1,\infty)$, $0 < \underline{b} < \sqrt{2} < \overline{b}$, $p \in (1,\infty)$, $u > 0$ and $N \in \mathbb{N}^*$, we have in Setting 1) and Setting 2),

$$\left\|\sum_{i=1}^{N}(\overline{W}_{i,N}(\beta))^{\,\delta}\,\tilde{Y}_i - \tilde{Y}_{I_N}\right\|_m \leq C\left(u\sup_{\substack{b\in\left[\underline{b},\overline{b}\right]\\ q=1,m}}\left(\zeta(b\,\sqrt{\log N})\right)^{\frac{q}{m}} + \tilde{\zeta}_N^{(m,p)}\left(u, (\log N)^{-\frac{1}{2}}\;N^{1-\frac{\overline{b}^2}{2}}\right)\right)\,,$$

where ζ is defined in (172), $\tilde{\zeta}_N^{(m,p)}$ is defined in (173) and $C > 0$ is a constant only depending on p, $\underline{b}$, $\overline{b}$, δ and m. By Lemma 7, using $F^{\leftarrow}$ as above, we have that there exists a universal constant $C_0 > 0$ such that, for all $b > 0$,

$$\beta \geq 2b\sqrt{\log N} \geq 1 \Longrightarrow \zeta(b\,\sqrt{\log N}) \leq C_0\;\frac{b\,\sqrt{\log N}}{\beta}\,.$$

Note that if $\underline{b} \leq b \leq \overline{b}$, $\beta \geq \overline{b}\,(2 \vee C_0)\sqrt{\log N}$ and $\underline{b} \geq 1$, then the condition on the left-hand side is satisfied and the upper bound in the right-hand side is at most 1. Using this in the

previous display for the special case $p = 2$ and $\underline{b} = 1$, we get that for all $m, \delta \in [1, \infty)$ and $\overline{b} > \sqrt{2}$, there exists $C, C_0 > 0$ such that, for all $u > 0$, if $\beta \geq \overline{b}\,(2 \vee C_0)\sqrt{\log N} \geq 1$,

$$\left\| \sum_{i=1}^{N} (\overline{W}_{i,N}(\beta))^{\,\delta}\, \tilde{Y}_i - \tilde{Y}_{I_N} \right\|_m \leq C \left(u \left(\frac{\overline{b}\, C_0 \sqrt{\log N}}{\beta} \right)^{\frac{1}{m}} + \tilde{\zeta}_N^{(m,2)} \left(u, (\log N)^{-\frac{1}{2}}\ N^{1-\frac{\overline{b}^2}{2}} \right) \right) . \tag{204}$$

To prove (202) and (203) we now consider Setting 1) and Setting 2) separately by taking $\tilde{Y}_j = \xi_j$ and $\tilde{Y}_j = 1$ respectively in $\tilde{\zeta}_N^{(m,2)} \left(u, (\log N)^{-\frac{1}{2}}\ N^{1-\frac{\overline{b}^2}{2}} \right)$ and choosing adequate values of $\overline{b}$ and u in the upper bound above. Namely, we show that for $\overline{b} = \sqrt{4m + 4q + 2}$ and $q \geq 1$

- If $\tilde{Y}_j = \xi_j$ for all $j \geq 1$, then: as $N \to \infty$,

$$\tilde{\zeta}_N^{(m,2)} \left(\overline{b}\sqrt{\log N}, (\log N)^{-\frac{1}{2}}\ N^{1-\frac{\overline{b}^2}{2}} \right) = O\left((\log N)^{-1/(4m)} N^{-q/m} \right) . \tag{205}$$

- If $\tilde{Y}_j = 1$ for all $j \geq 1$, then: for all $N \geq 2$,

$$\tilde{\zeta}_N^{(m,2)} \left(2, (\log N)^{-1/2}\, N^{1-\frac{\overline{b}^2}{2}} \right) = (\log N)^{-1/(4m)}\, N^{-q/m} . \tag{206}$$

First note that (206) follows directly from plugging in $(u, s) = (2, (\log N)^{-1/2}\, N^{1-\frac{\overline{b}^2}{2}})$ in (173). As for (205), setting $\tilde{Y}_j = \xi_j$ for all $j \geq 1$ and plugging in

$$(u, s) = \left(\overline{b}\sqrt{\log N}, (\log N)^{-1/2} N^{1-\overline{b}^2/2} \right)$$

in (173) yields that, for all $N \geq 2$,

$$\tilde{\zeta}_N^{(m,2)} \left(\overline{b}\sqrt{\log N}, (\log N)^{-1/2} N^{1-\overline{b}^2/2} \right) = \|\xi_1\|_{2m}\, (\log N)^{-1/(4m)} N^{-q/m} + N \left\| \xi_1\, \mathbb{1}_{\left\{ \xi_N^* > \overline{b}\sqrt{\log N} \right\}} \right\|_m ,$$

where $\xi_N^* = \max(|\xi_1|, \ldots, |\xi_N|)$ and we used that $\overline{b}^{\,2} = 4m + 4q + 2$. Since

$$\left\| \xi_1\, \mathbb{1}_{\left\{ \xi_N^* > \overline{b}\sqrt{\log N} \right\}} \right\|_m \leq \|\xi_1\|_{2m} \left\| \mathbb{1}_{\left\{ \xi_N^* > \overline{b}\sqrt{\log N} \right\}} \right\|_{2m} ,$$

the desired result (205) will follow if we can show that: as $N \to \infty$,

$$\mathbb{P}\left(\xi_N^* > \overline{b}\sqrt{\log N} \right) = O\left((\log N)^{-1/2} N^{-2m-2q} \right) .$$

Writing that $\xi_N^* = \max(M_N, M_N')$ where we have set $M_N = \max(\xi_1, \ldots, \xi_N)$ and $M_N' = \max(-\xi_1, \ldots, -\xi_N)$, it holds that $\mathbb{P}(\xi_N^* > u) \leq \mathbb{P}(M_N > u) + \mathbb{P}(M_N' > u) = 2\mathbb{P}(M_N > u)$ for any $u > 0$, hence the last display follows from (168) in Lemma 5 with $c = \overline{b}^{\,2}/2 - 1$.

Now, for the choice $\overline{b} = \sqrt{4m+4q+2}$, if $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$, we eventually have $\beta \geq \overline{b}(2 \vee C_0)\sqrt{\log N} \geq 1$, so that (204) eventually applies. If $\tilde{Y}_j = \xi_j$ for all j, we have $\tilde{Y}_{I_N} = \max(\xi_1, \ldots, \xi_N)$, and the bound (204) with $u = \overline{b}\sqrt{\log N}$, together with (205), gives us that, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$, (202) hold. If $\tilde{Y}_j = 1$ for all j, we have $\tilde{Y}_{I_N} = 1$, and the bound (204) with $u = 2$ and the same choice of $\overline{b}$, together with (206), gives (203).

(ii) *Proofs of (199) and (200).* Since the expectation is 1-Lipschitz for the L^1-norm, the asymptotic behavior (202) with $m = 1$ implies that, for any $\delta \geq 1$,

$$\mathbb{E}\left(\sum_{i=1}^{N} (\overline{W}_{i,N}(\beta))^{\,\delta}\, \xi_i\right) = \mathbb{E}\left(\max(\xi_1, \ldots, \xi_N)\right) + o\left(\sqrt{\log N}\right) .$$

Applying Lemma 6 with $m = 1$ the first term in the right-hand side is asymptotically equivalent to $\sqrt{2\log N}$ and we get (199). The cases $\delta = 1, 2$ in particular together give that, for any $\lambda \in [0,1]$,

$$\mathbb{E}\left(\sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j\right) = (2\,\log N)^{1/2}\,(1 + o(1)) \ . \tag{207}$$

By the Minkowski inequality for the L^2-norm, we obtain that

$$\begin{aligned}
&\left| \left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j \right\|_2 - \|\max(\xi_1, \ldots, \xi_N)\|_2 \right| \\
&\leq \left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j - \max(\xi_1, \ldots, \xi_N) \right\|_2 \\
&\leq \lambda \left\| \sum_{j=1}^{N} \overline{W}_{j,N}(\beta)\, \xi_j - \max(\xi_1, \ldots, \xi_N) \right\|_2 + (1-\lambda) \left\| \sum_{j=1}^{N} (\overline{W}_{j,N}(\beta))^2\, \xi_j - \max(\xi_1, \ldots, \xi_N) \right\|_2 .
\end{aligned}$$

Hence, using (202) again, this time with $m = 2$, we deduce

$$\left| \left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j \right\|_2 - \|\max(\xi_1, \ldots, \xi_N)\|_2 \right| = o\left(\sqrt{\log N}\right),$$

which we can rewrite equivalently as

$$\left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j \right\|_2 = \|\max(\xi_1, \ldots, \xi_N)\|_2 + o\left(\sqrt{\log N}\right).$$

Using Lemma 6 with $m = 2$ to evaluate the first term in the right-hand side and taking the square, we get that

$$\mathbb{E}\left(\left(\sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\, \xi_j\right)^2\right) = 2\,\log N\,(1 + o(1)) .$$

This asymptotic behavior with the square of (207) yields (200).

(iii) *Proof of (201).* The proof of (201) is obtained similarly to the proofs of (199) and (200), but starting from (203) instead of (202).

■

We can now prove Proposition 1 and Theorem 5.

D.4.1 Proof of Proposition 1

The proof is made of two steps. Firstly we show that: as $N, B(\theta,\phi;x) \to \infty$ with (38) holding,

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi}(\varepsilon;x)) + \sqrt{2\log N}\;\partial_\psi B(\theta,\phi;x)\;(1+o(1))\,, \tag{208}$$

$$\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) \geq \left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2 (1+o(1))\,. \tag{209}$$

Secondly, we deduce the claim of the proposition.

(i) **Proof of (208) and (209).** Recall that $\varepsilon_1, \varepsilon_2, \ldots$ are i.i.d. copies of ε. Since $\overline{\mathcal{W}}_{1,N}^{(\alpha)}, \ldots, \overline{\mathcal{W}}_{N,N}^{(\alpha)}$ as defined in (185) can be expressed as functions of $\Lambda_{\theta,\phi}(\varepsilon_k;x)$ for all $k = 1 \ldots N$, by Property ($\mathrm{P}^{\mathrm{REP}}$) they are independent of the centered random variables $\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_1;x), \ldots, \partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_N;x)$. Furthermore, Property ($\mathrm{P}^{\mathrm{REP}}$) also gives us that $\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_1;x), \ldots, \widetilde{\Lambda}_{\theta,\phi}(\varepsilon_N;x)$ are i.i.d. $\mathcal{N}(0,1)$. Next noticing that $\overline{\mathcal{W}}_{1,N}^{(\alpha)}, \ldots, \overline{\mathcal{W}}_{N,N}^{(\alpha)}$ can be rewritten using $\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_k;x)$ for all $k = 1 \ldots N$, we can apply Lemma 8 with $\xi_k = \widetilde{\Lambda}_{\theta,\phi}(\varepsilon_k;x)$, and $\beta = (1-\alpha)B(\theta,\phi;x)$ to prove (208) and (214). More precisely, by taking the expectation in (188) we get that

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi}(\varepsilon;x)) + \partial_\psi B(\theta,\phi;x)\;\mathbb{E}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\;\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right)$$

and, as $N, B(\theta,\phi;x) \to \infty$ with (38) holding, (199) with $\delta = 1$ gives us that (208) holds. Similarly, by taking the variance in (188) we have that

$$\begin{aligned}\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = (\partial_\psi B(\theta,\phi;x))^2\;\mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\;\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right)\\ + (B(\theta,\phi;x))^2\;\mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\;\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right).\end{aligned} \tag{210}$$

Thus,

$$\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) \geq (B(\theta,\phi;x))^2\;\mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\;\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right)$$

and to get (209) it only remains to show that the second variance in the right-hand side above is $\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)^2\,(1+o(1))$. Using Property ($\mathrm{P}^{\mathrm{REP}}$) again,

$$\begin{aligned}\mathbb{V}\left(\sum_{j=1}^{N}\overline{W}_{j,N}^{(\alpha)}\;\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right) &= \mathbb{E}\left(\left(\sum_{j=1}^{N}\overline{W}_{j,N}^{(\alpha)}\;\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right)^2\right)\\ &= \mathbb{E}\left(\sum_{j=1}^{N}\left(\overline{W}_{j,N}^{(\alpha)}\right)^2\;\mathbb{E}\left(\left(\partial_\psi\widetilde{\Lambda}_{\theta,\phi}(\varepsilon_j;x)\right)^2\right)\right)\\ &= \widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)^2\,\mathbb{E}\left(\sum_{j=1}^{N}\left(\overline{W}_{j,N}^{(\alpha)}\right)^2\right)\\ &= \widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)^2\,(1+o(1)) \qquad (211)\end{aligned}$$

where in the last line we have used (201) with $\delta=2$.

(ii) We can now prove the claim of the proposition. By Property ($\mathrm{P}^{\mathrm{REP}}$), the stochastic terms in the second line of (191) have mean zero and by (42) and (208), we have

$$\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)=\frac{\sqrt{2\log N}|\;\partial_\psi B(\theta,\phi;x)|\;\;(1+o(1))}{\sqrt{\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right)}}\,. \qquad (212)$$

Applying (209), we get that

$$\mathrm{sSNR}_{\psi,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\leq\frac{\sqrt{2\log N}\;\;|\partial_\psi B(\theta,\phi;x)|\;\;(1+o(1))}{B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\,(1+o(1))}\,.$$

Now, by (190), we have

$$\frac{|\partial_\psi B(\theta,\phi;x)|}{B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)}=\left((|\mathrm{C}_\psi(\theta,\phi;x)|)^{-2}-1\right)^{-1/2} \qquad (213)$$

which is $O\left(|\mathrm{C}_\psi(\theta,\phi;x)|\right)$ under (39). Hence we get (41).

D.4.2 Proof of Theorem 5

Since (39) is implied by (43), we can build on the equations established in the proof of Proposition 1 such as (210). We then have that the first variance in the right-hand side of (210) is $o\left(\log N\right)$ as $N, B(\theta,\phi;x)\to\infty$ with (38) holding thanks to (200) with $\lambda=1$ in Lemma 8. Pairing this up with (211) leads to

$$\begin{aligned}\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) &= \left(\partial_\psi B(\theta,\phi;x)\right)^2\;\;o\left(\log N\right) \qquad (214)\\ &\quad+\left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2(1+o(1))\,.\end{aligned}$$

In addition, the condition (43) paired up with (213) yields

$$\frac{|\partial_\psi B(\theta,\phi;x)|}{B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)}=|\mathrm{C}_\psi(\theta,\phi;x)|\;\;(1+o(1))=O\left((\log N)^{-1/2}\right)\,.$$

so that (214) becomes

$$\mathbb{V}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \left(B(\theta,\phi;x)\widetilde{B}^{(\mathrm{REP})}(\theta,\phi;x)\right)^2(1+o(1)) \\ = \mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))\,(1+o(1)) \quad (215)$$

Combining this with (192) and (212) leads to (44). Furthermore, (189), (190) and (208) yield

$$\mathbb{E}\left(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) \\ + \sqrt{\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))}\sqrt{2\log N}\;\mathrm{C}_\psi(\theta,\phi;x)\,(1+o(1))\,. \quad (216)$$

If we further assume the stronger condition (45), (46) immediately follows from the two above displays paired up with (192) and (208).

Remark 8 *As* $N, \mathrm{KL}\,(\theta,\phi;x) \to \infty$ *under* (38) *and* (43)*, no improvement in variance is obtained by increasing the sample size* N *beyond the case* $N=1$ *in* $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ *since* (215) *shows that* $\mathbb{V}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ *reverts to* $\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$*. The case of* (216) *is more subtle: using* (195) *in Appendix D.3,* (216) *reads equivalently*

$$\mathbb{E}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) \\ + \sqrt{\frac{\log N}{\mathrm{KL}\,(\theta,\phi;x)}}\;\left[\partial_\psi \ell(\theta;x) - \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))\right](1+o(1))\,. \quad (217)$$

Then, $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ *points in the direction of* $\mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ *on average when* $\psi \in \{\phi_1,\ldots,\phi_b\}$*, since* $\partial_\psi \ell(\theta;x) = 0$ *and*

$$\mathbb{E}(\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))\left(1-\sqrt{\frac{\log N}{\mathrm{KL}\,(\theta,\phi;x)}}+o(1)\right).$$

There is thus no improvement in terms of gradient direction for learning ϕ *that follows from taking* $N>1$ *instead of* $N=1$*. When* $\psi \in \{\theta_1,\ldots,\theta_a\}$*, the answer depends on which term dominates in* (217)*. Indeed,* $\psi\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ *points on average (i) in the direction of* $\mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ *should the first one dominate and (ii) in the direction of a convex combination of* $\mathbb{E}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ *and* $\partial_\psi \ell(\theta;x)$ *should both terms be of the same order. Note that the case where the second term dominates amounts to assuming that* $\mathrm{SNR}[\psi\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = o(1)$ *under* (43) *so we leave it out.*

D.5 General Gaussian model

Consider a general Gaussian model with known positive definite covariance matrix $\Sigma(x)$ obtained by setting $p_\theta(z|x) = \mathcal{N}(z;\mu(\theta;x),\Sigma(x))$ $q_\phi(z|x) = \mathcal{N}(z;\tilde{\mu}(\phi;x),\Sigma(x))$, where μ and $\tilde{\mu}$ are the mean vectors parameterized by the true and variational parameters respectively

and where we use the reparametrization $f(\varepsilon, \phi; x) = \Sigma(x)^{1/2}\,\varepsilon + \tilde{\mu}(\phi; x)$ with $\varepsilon \sim \mathcal{N}(0, \mathbf{I}_d)$. Then, (B1) holds and we have

$$\begin{aligned}\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x) = \ell(\theta; x) &+ \frac{1}{2}\langle \left(\tilde{\mu}(\phi; x) - \tilde{\mu}(\phi'; x)\right)\,,\, \Sigma(x)^{-1/2}\left(\tilde{\mu}(\phi; x) - \tilde{\mu}(\phi'; x)\right)\rangle \\ &- \frac{1}{2}\langle \left(\tilde{\mu}(\phi'; x) - \mu(\theta; x)\right)\,,\, \Sigma(x)^{-1}\left(\tilde{\mu}(\phi'; x) - \mu(\theta; x)\right)\rangle \\ &+ \langle \varepsilon\,,\, \Sigma(x)^{-1/2}\left(\tilde{\mu}(\phi; x) - \mu(\theta; x)\right)\rangle.\end{aligned}$$

Observe then that the stochastic part above does not depend on ϕ' hence $\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) = 0$ when $\psi \in \{\phi_1, \ldots, \phi_b\}$. We also obtain that

$$\mathrm{KL}\,(\theta, \phi; x) = \frac{1}{2}\,\|\Delta(\theta, \phi; x)\|^2 \quad \text{and} \quad \mathrm{C}_\psi(\theta, \phi; x) = \frac{\langle \partial_\psi \Delta(\theta, \phi; x)\,,\, \Delta(\theta, \phi; x)\rangle}{\|\Delta(\theta, \phi; x)\|\;\|\partial_\psi \Delta(\theta, \phi; x)\|}$$

with $\Delta(\theta, \phi; x) := \Sigma(x)^{-1/2}\left(\tilde{\mu}(\phi; x) - \mu(\theta; x)\right)$.

D.6 Proof of Example 3

This example is a specific case of the general one in Appendix D.5 with $\Sigma(x) = \mathbf{I}_d$, $\mu(\theta; x) = \theta$ and $\tilde{\mu}(\phi; x) = \phi$. Thus we get from Appendix D.5 that (B1) holds, $\Delta(\theta, \phi; x) = \phi - \theta$, which implies $\|\partial_\psi(\phi - \theta)\| = 1$ for any component ψ among $\{\theta_1, \ldots, \theta_a, \phi_1, \ldots, \phi_b\}$, and:

$$\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x) = \ell(\theta; x) - \frac{\|\theta - \phi\|^2}{2} - \langle \varepsilon + \phi' - \phi, \phi - \theta\rangle\,, \tag{218}$$

$$\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x) = \ell(\theta; x) - \frac{\|\theta - \phi\|^2}{2} - \langle \varepsilon, \phi - \theta\rangle\,, \tag{219}$$

$$\mathrm{KL}\,(\theta, \phi; x) = \frac{1}{2}\,\|\theta - \phi\|^2. \tag{220}$$

We easily get from (219) that ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$) also holds. Furthermore in the setting $\theta - \phi = \epsilon \cdot \boldsymbol{u}_d$, we get that

$$\mathrm{KL}\,(\theta, \phi; x) = \frac{1}{2}\epsilon^2 d \quad \text{and} \quad |\mathrm{C}_\psi(\theta, \phi; x)| = d^{-1/2}\,.$$

Therefore, in this specific case, (38) is equivalent to (47), (39) to eventually have $d \geq 2$ and (43) and (45) to $\log N = O(d)$ and $\log N \ll d$, respectively.

Applying Theorem 5, (47) then implies (46) which takes the form (48) since (219) straightforwardly leads to

$$\mathrm{SNR}[\phi_k\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \epsilon \quad \text{for } k = 1, \ldots, a.$$

Furthermore, as far as the DREP estimator is concerned, we get from (218) that

$$\left[\partial_{\phi'_k} \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\Big|_{\phi'=\phi} - 0 = 0 - \phi_k + \theta_k\,,$$

thus $\mathbb{V}(\psi\text{-}g_{1,1}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) = 0$.

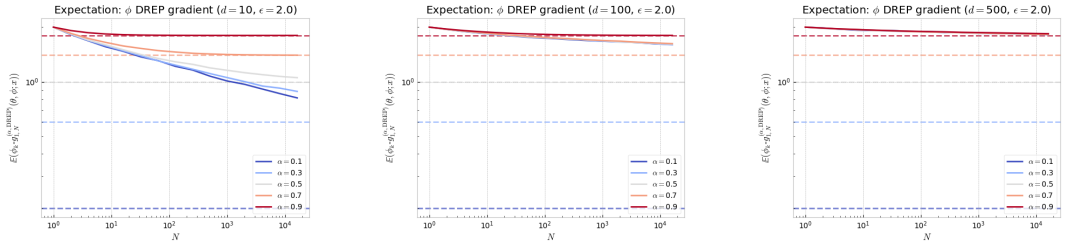


Figure 5: Plotted is $\mathbb{E}(\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x))$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N, for varying values of (α, d), $\epsilon = 2$ and a random coordinate ϕ_k. The solid lines correspond to $\mathbb{E}(\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x))$, the dashed lines correspond to predictions of the form $y = \epsilon\alpha$.

Appendix E. Additional numerical experiments

E.1 Gaussian experiment from Section 4.1

Example 3 predicts that the REP gradient estimator with $N > 1$ reverts to the case $N = 1$ as d increases, and notably that $\mathbb{E}(\phi_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ reverts to $\mathbb{E}(\phi_k\text{-}g_{1,1}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$. Similarly to the REP case detailed in Section 4.1, this behavior can be observed in the DREP case by setting a high value for ϵ, see Figure 5.

The IWAE case unfolds in the same manner as d increases. By Example 1: as $N \to \infty$,

$$\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(0,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{\frac{1}{N}}\frac{\epsilon\exp\left(\frac{d\epsilon^2}{2}\right)}{\sqrt{1+\epsilon^2}}(1+o(1)) \tag{221}$$

$$\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)] = \frac{\sqrt{N}(1+o(1))}{\left(\exp\left(4d\epsilon^2\right) - 4\exp\left(2d\epsilon^2\right) + 4\exp\left(d\epsilon^2\right) - 1\right)^{\frac{1}{2}}} \tag{222}$$

We let $d \in \{10, 100\}$, $\epsilon = 0.2$, $N \in \{2^j, j = 1 \dots 15\}$ and our results are plotted on Figure 6. Similarly to the REP case with $\alpha \in (0, 1)$ detailed in Section 4.1, in the favourable setting of a low dimension $d = 10$, the behavior of the REP and DREP gradient estimators predicted by (221) and (222) respectively match as N increases. As expected, this is no longer true as d increases.

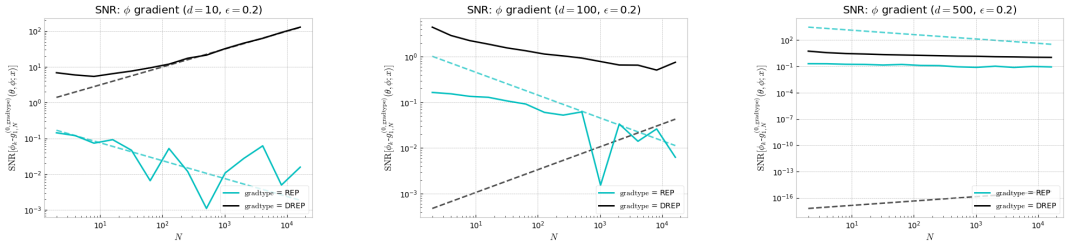


Figure 6: Plotted are $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(0,\mathrm{REP})}(\theta,\phi;x)]$ and $\mathrm{SNR}[\phi_k\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N and with $\epsilon = 0.2$. The solid lines correspond to the SNRs, while the dashed lines correspond to predictions of the form (221) and (222).

E.2 Linear Gaussian experiment from Section 4.2

We provide here additional experiments for the cases $\psi = \theta_k$ with $\alpha \in [0,1)$ and $\psi = \tilde{b}_k$ with $\alpha = 0$. By Example 2: as $N \to \infty$, for all $\alpha \in [0,1)$,

$$\mathrm{SNR}[\theta_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{N}\,\frac{\left|\frac{x_k-\theta_k}{2}+\frac{3\epsilon\alpha}{4-\alpha}\right|(1+o(1))}{\frac{(4-\alpha)^{d/2}}{(15-6\alpha)^{d/4}}\exp\left(\frac{12(1-\alpha)^2}{(4-\alpha)(5-2\alpha)}d\epsilon^2\right)\sqrt{\frac{2}{5-2\alpha}+\left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2}} \tag{223}$$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)] = \sqrt{N}\frac{\frac{24\epsilon 4^{d-1}\exp\left(\frac{24d\epsilon^2}{4\cdot 5}\right)}{3^{d/2}5^{\frac{d}{2}+1}}}{\sqrt{\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x)}}(1+o(1)) \tag{224}$$

We let $d \in \{10,100,500\}$, $\epsilon \in \{0.2,1\}$, $N \in \{2^j, j = 1\ldots 15\}$ and our results are plotted on Figures 7 and 8. Similarly to the cases detailed in Section 4.1, in the favourable setting of a low dimension $d = 10$, the behavior of the REP and DREP gradient estimators predicted by (223) and (224) respectively match as N increases. This is no longer true as d increases. Interestingly, we see that when it comes to the learning of θ, the SNR is not monotonic in α, which is not surprising given the numerator term $\left|\frac{x_k-\theta_k}{2}+\frac{3\epsilon\alpha}{4-\alpha}\right|$ appearing in (223).

References

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431, 2017.

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.

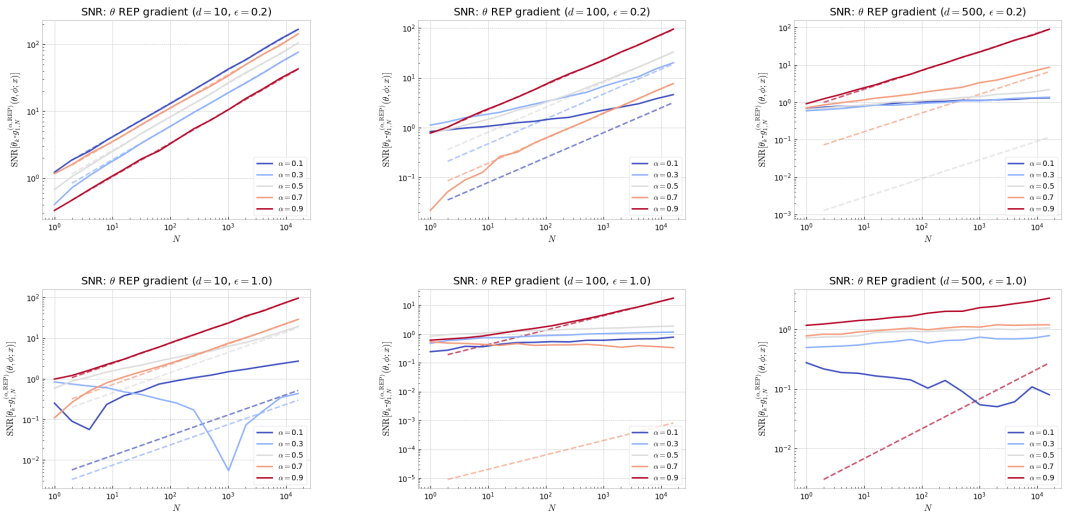


Figure 7: Plotted is $\mathrm{SNR}[\theta_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\theta_k\text{-}g_{1,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]$, the dashed lines correspond to predictions of the form (223).

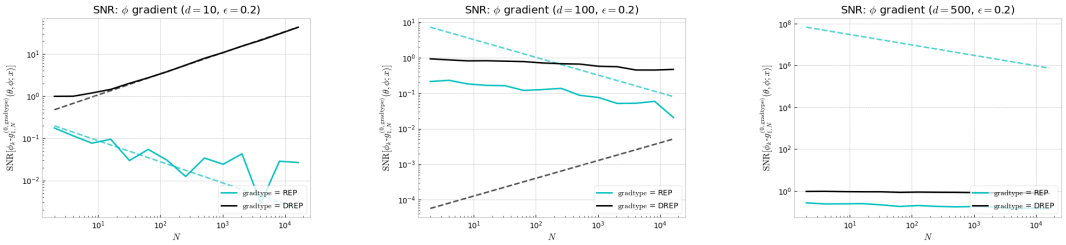


Figure 8: Plotted are $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(0,\mathrm{REP})}(\theta,\phi;x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g_{1,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, with $\epsilon = 0.2$ and for a randomly selected datapoint x. The solid lines correspond to the SNRs, the dashed lines correspond to predictions of the form (223) and (224).

Patrick Billingsley. *Weak Convergence of Measures*. Society for Industrial and Applied Mathematics, 1971.

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.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

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.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Badr-Eddine Cherief-Abdellatif, Randal Douc, Arnaud Doucet, and Hugo Marival. On the asymptotics of importance weighted variational inference. *arXiv preprint arXiv:2501.08477*, 2025.

Noel Cressie, Anne S. Davis, J. Leroy Folks, and George E. Policello, II. The moment-generating function and negative integer moments. *Amer. Statist.*, 35(3):148–150, 1981.

Kamélia Daudel, Joe Benton, Yuyang Shi, and Arnaud Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83, 2023.

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.

George Deligiannidis, Pierre E. Jacob, El Mahdi Khribch, and Guanyang Wang. On importance sampling and independent metropolis-hastings with an unbounded weight function, 2025. URL `https://arxiv.org/abs/2411.09514`.

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, 2021.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems 31*, pages 4470–4479. 2018.

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

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 2. John Wiley & Sons, 1991.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

Robert D. Gordon. Values of Mills' Ratio of Area to Bounding Ordinate and of the Normal Probability Integral for Large Values of the Argument. *The Annals of Mathematical Statistics*, 12(3):364 – 366, 1941.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of machine learning research*, 2013.

Michael Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Kyurae Kim, Yian Ma, and Jacob Gardner. Linear convergence of black-box variational inference: Should we stick the landing? In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 235–243. PMLR, 02–04 May 2024. URL `https://proceedings.mlr.press/v238/kim24a.html`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

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

Pierre L'Ecuyer. On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41(4):738–748, 1995.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Wenjie Liao, Like Wu, Liangjie Zhao, Shihui Xu, and Shigeru Fujimura. Iris: Interpolative r\'enyi iterative self-play for large language model fine-tuning. *arXiv preprint arXiv:2604.20933*, 2026.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems 30*, pages 6573–6583. 2017.

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

Charles C. Margossian, Loucas Pillaud-Vivien, and Lawrence K. Saul. Variational inference for uncertainty quantification: an analysis of trade-offs. *arXiv:2403.13748*, 2024.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Radford M. Neal and Geoffrey E. Hinton. *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_12. URL https://doi.org/10.1007/978-94-011-5014-9_12.

Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyZoi-WRb.

Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

James Pickands III. Moment Convergence of Sample Extremes. *The Annals of Mathematical Statistics*, 39(3):881 – 889, 1968.

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 *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4277–4285, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rezende14.html.

Herbert E. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951. URL https://api.semanticscholar.org/CorpusID:16945044.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

Lorenz Vaitl, Kim A Nicoli, Shinichi Nakajima, and Pan Kessel. Gradients should stay on path: better estimators of the reverse-and forward kl divergence for normalizing flows. *Machine Learning: Science and Technology*, 3(4):045006, 2022.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Ming Xu, Matias Quiroz, Robert Kohn, and Scott A Sisson. Variance reduction properties of the reparameterization trick. In *The 22nd international conference on artificial intelligence and statistics*, pages 2711–2720. PMLR, 2019.