Learning with Importance Weighted Variational Inference: Asymptotics for Gradient Estimators of the VR-IWAE Bound

Kamélia Daudel, François Roueff

Learning with Importance Weighted Variational Inference: Asymptotics for Gradient Estimators of the VR-IWAE Bound

Kamélia Daudel
ESSEC Business School
daudel@essec.edu

François Roueff
Télécom Paris, Institut Polytechnique de Paris
francois.roueff@telecom-paris.fr

ABSTRACT

Several popular variational bounds involving importance weighting ideas have been proposed to generalize and improve on the Evidence Lower BOund (ELBO) in the context of maximum likelihood optimization, such as the Importance Weighted Auto-Encoder (IWAE) and the Variational Rényi (VR) bounds. The methodology to learn the parameters of interest using these bounds typically amounts to running gradient-based variational inference algorithms that incorporate the reparameterization trick. However, the way the choice of the variational bound impacts the outcome of variational inference algorithms can be unclear. Recently, the VR-IWAE bound was introduced as a variational bound that unifies the ELBO, IWAE and VR bounds methodologies. In this paper, we provide two analyses for the reparameterized and doubly-reparameterized gradient estimators of the VR-IWAE bound, which reveal the advantages and limitations of these gradient estimators while enabling us to compare of the ELBO, IWAE and VR bounds methodologies. Our work advances the understanding of importance weighted variational inference methods and we illustrate our theoretical findings empirically.

***K*eywords** Variational Inference · Alpha-Divergence · Importance Weighted Auto-encoder · High dimension · Signal-To-Noise Ratio

Contents

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 $\mathcal{Q}$ (Jordan et al., 1999; Blei et al., 2017). A common setting where VI is applied is when one is given a model that depends on a parameter θ and the goal is to optimize the associated marginal log likelihood, with the posterior density being intractable. Since direct optimization of the marginal log likelihood cannot be carried out, variational bounds involving the variational family $\mathcal{Q}$ are constructed as surrogate objective functions to the marginal log likelihood that are more amenable to optimization.

While the most traditional variational bound is the Evidence Lower BOund (ELBO), popular alternatives to the ELBO that rely on importance weighting ideas have been proposed to improve on VI, such as the Importance Weighted Auto-Encoder (IWAE) bound in Burda et al., 2016 and the Variational Rényi (VR) bound in Li and Turner, 2016. An active line of research in VI is then concerned with how the choice of the variational bound affects the outcome of VI algorithms (Dieng et al., 2017; Li and Gal, 2017; Maddison et al., 2017; Rainforth et al., 2018; Wang et al., 2018; Tucker et al., 2019; Geffner and Domke, 2020; Daudel and Douc, 2021; Daudel et al., 2021; Dhaka et al., 2021; Geffner and Domke, 2021; Rudner et al., 2021; Guilmeau et al., 2022; Knoblauch et al., 2022; Rodríguez-Santana and Hernández-Lobato, 2022; Daudel et al., 2023a,b; Guilmeau et al., 2024; Margossian et al., 2024).

Among these works, Daudel et al., 2023a introduce and study the VR-IWAE bound, a variational bound that depends on two hyperparameters $(N, \alpha) \in \mathbb{N}^\star \times [0, 1)$ and that unifies the ELBO, IWAE and VR methodologies when the reparametrization trick (Kingma and Welling, 2014) is available. Daudel et al., 2023a notably provide two analyses of the VR-IWAE bound that elucidate the role N and α play in this bound. Yet, solely focusing on the behavior of a variational bound is insufficient to assess the effectiveness of algorithms based on this bound at learning the parameters of interest (Rainforth et al., 2018).

In this paper, we study the role of N and α in two gradient estimators of the VR-IWAE bound that are at the center of the methodology built in Daudel et al., 2023a, namely the reparameterized (REP) and doubly-reparameterized (DREP) gradient estimators of the VR-IWAE bound. In doing so, our aim is to provide insights that apply to widely-used gradient estimators of the IWAE and VR bounds and more broadly to further advance the understanding of importance weighted VI methods. The paper is then organized as follows.

In Section 2, we review the main concepts behind the VR-IWAE bound methodology and the existing asymptotic studies for this methodology. In Section 3, we provide two asymptotic studies for the REP and DREP gradient estimators of the VR-IWAE bound.

(i) The first analysis, in which the number of Monte Carlo samples N goes to infinity, shows that α enables a bias-variance tradeoff in these gradient estimators and highlights how the DREP gradient estimator can significantly outperform the REP one (Theorems 1 to 3). This analysis is illustrated in Examples 3 and 4. Examples 3 and 4 also reveal the impact the dimension of the latent space d may have on the relevance of Theorems 1 to 3 as d increases.

(ii) To account for the effect of a large dimension d in the REP and DREP gradient estimators of the VR-IWAE bound, we derive a second analysis in Theorems 4 and 5 as $N, d \to \infty$ that is illustrated in Example 5. This analysis suggests that in certain regimes the REP and DREP gradients of the VR-IWAE bound suffer from a weight collapse phenomenon for all $\alpha \in [0, 1)$ which impedes the VR-IWAE bound methodology.

Section 4 provides empirical evidence supporting our claims. Directions for future research are outlined in Section 5. Proofs and additional experiments are provided in the appendix.

2 Background on the VR-IWAE bound methodology

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 bayesian inference tasks such as maximization 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 maximum likelihood optimization, 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

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

several generalizations of the ELBO exist, such as the IWAE bound (Burda et al., 2016)

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

and the VR-IWAE bound (Daudel et al., 2023a): for all $N \in \mathbb{N}^\star$ and all $\alpha \in [0, 1)$,

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

Both variational bounds recover the ELBO when $N = 1$ (or when $\alpha \to 1$ for the VR-IWAE bound) with the VR-IWAE bound generalizing the IWAE bound ($\alpha = 0$). Furthermore, a straightforward way to optimize these variational bounds is to use stochastic gradient ascent (SGA) paired up with the reparameterization trick (Kingma and Welling, 2014; Burda et al., 2016; Daudel et al., 2023a). 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 $\varepsilon \sim q$. 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

$$\int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x)^{1-\alpha}}{\sum_{j'=1}^N w_{\theta,\phi}(f(\varepsilon_{j'}, \phi; x); x)^{1-\alpha}} \nabla_{\theta,\phi} \log w_{\theta,\phi}(f(\varepsilon_j; \phi; x); x)\right) \mathrm{d}\varepsilon_{1:N} \tag{4}$$

and since this gradient can be estimated with an unbiased Monte Carlo estimator, the resulting procedure enjoys the typical theoretical guarantees of unbiased SGA.

One key consequence of this result is that the VR-IWAE bound provides the theoretical grounding for VI methods which seek to improve on the ELBO methodology by relying instead on the VR bound, a flexible variational bound defined in Li and Turner, 2016 by

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x)\, w_{\theta,\phi}(z;x)^{1-\alpha}\, \mathrm{d}z \right), \quad \alpha \in (0,1) \cup (1, \infty). \tag{5}$$

Indeed, SGA steps on the VR bound lead to the same SGA scheme as the VR-IWAE bound one, but the SGA is biased for the VR bound (Li and Turner, 2016; Daudel et al., 2023a).

To provide insights as to how the variational bound should be selected in VI, Daudel et al., 2023a then provide two asymptotic analyses of the VR-IWAE bound. Those analyses will lay the foundations for our work in this paper and we will thus briefly recall them next. Before that, note that from now on 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)–(5), 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 (5) are then respectively written as

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \mathbb{E}_{Z_j \overset{\text{i.i.d.}}{\sim} q_\phi(\cdot|x)} \left(\log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(Z_j; x)^{1-\alpha} \right) \right) \tag{6}$$

$$\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \mathbb{E}_{Z \sim q_\phi(z|x)} \left(w_{\theta,\phi}(Z; x)^{1-\alpha} \right), \tag{7}$$

with (A1) ensuring $w_{\theta,\phi}(\cdot; x)$ is well-defined and positive a.s. under the expectation $\mathbb{E}$. We now review the first asymptotic analysis of the VR-IWAE bound from Daudel et al., 2023a.

2.1 Asymptotics of the VR-IWAE bound as $N \to \infty$

Under precise conditions stated in Daudel et al., 2023a, Theorem 3, it holds that:

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{VR}^{(\alpha)}(\theta, \phi; x) - \frac{\gamma^{(\alpha)}(\theta, \phi; x)^2}{2N} + o\left(\frac{1}{N}\right), \tag{8}$$

where $\mathrm{VR}^{(\alpha)}(\theta, \phi; x)$ is the VR bound defined in (7) and

$$\gamma^{(\alpha)}(\theta, \phi; x)^2 = \frac{1}{1-\alpha} \mathbb{V}_{Z \sim q_\phi(\cdot|x)} \left(\frac{w_{\theta,\phi}(Z; x)^{1-\alpha}}{\mathbb{E}_{Z' \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z'; x)^{1-\alpha})} \right). \tag{9}$$

The asymptotic result (8) shows that the VR-IWAE bound converges at a fast $1/N$ rate towards the VR bound and hence encourages increasing N (as the VR bound is meant to improve on the ELBO and it

is exactly the true marginal log likelihood when $\alpha = 0$, see Li and Turner, 2016). It also highlights a bias-variance tradeoff: the VR bound is a bias term with $\mathrm{VR}^{(\alpha)}(\theta, \phi; x) = \ell(\theta; x)$ for $\alpha = 0$ and $\gamma^{(\alpha)}(\theta, \phi; x)^2$ is a variance term with $\gamma^{(\alpha)}(\theta, \phi; x)^2 \to 0$ as $\alpha \to 1$. Daudel et al., 2023a, Theorem 3 thus suggests using the VR-IWAE bound with $N > 1$ over the ELBO ($N = 1$) and, depending on the bias-variance tradeoff, using the VR-IWAE bound with $\alpha \in (0, 1)$ over the the IWAE bound ($\alpha = 0$).

While the asymptotic regime predicted by (8) is promising, Daudel et al., 2023a noticed that it may not kick in for high-dimensional latent spaces d unless N is unpractically large. Therefore, (8) may not reflect the behavior of the VR-IWAE bound in practice. As a result, Daudel et al., 2023a developed a second asymptotic approach when $N, d \to \infty$.

2.2 Asymptotics of the VR-IWAE bound as $N, d \to \infty$

Daudel et al., 2023a argue in their Section 4.2 that the relative weight

$$\frac{w_{\theta,\phi}(Z; x)}{\mathbb{E}_{Z' \sim q_\phi(\cdot|x)}(w_{\theta,\phi}(Z'; x))} = \frac{p_\theta(Z|x)}{q_\phi(Z|x)}, \quad Z \sim q_\phi(\cdot|x)$$

is likely to become approximately log-normal as the dimension of the latent space d increases. This intuition is based on a central limit theorem argument, so that the distribution of the relative weight can be approximated in high dimensions by a log-normal distribution of the form

$$\log\left(\frac{p_\theta(Z|x)}{q_\phi(Z|x)}\right) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S, \quad S \sim \mathcal{N}(0, 1),\ \sigma^2 < \infty. \tag{10}$$

Under (10), Daudel et al., 2023a then show that the VR-IWAE bound collapses to the ELBO as $N, d \to \infty$ when N does not grow as fast as exponentially with d, before extending their study to approximately log-normal cases of the form

$$\log\left(\frac{p_\theta(Z|x)}{q_\phi(Z|x)}\right) = -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S)) - \sigma\sqrt{d}S, \quad S = \sum_{j=1}^{d} \frac{\xi_j}{\sigma\sqrt{d}},\ \sigma^2 = \mathbb{V}(\xi_1) < \infty \tag{11}$$

with $\xi_1, \ldots, \xi_d$ being centered i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure. Contrary to (8), this second study warns about the potential computational overhead when $N > 1$ and the weights are approximately log-normal.

2.3 Two key examples

Daudel et al., 2023a showcase the validity of their results over two examples recalled below.

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 $\langle \cdot, \cdot \rangle$,*

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

In addition, (10) holds with $\sigma^2 = 1$ for $\phi = \theta + \boldsymbol{u}_d$, where $\boldsymbol{u}_d$ denotes the d-dimensional vector whose coordinates are all equal to 1 and more generally we have for this example

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

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, (11) holds with $\sigma^2 = 1/18 + 8\lambda^2/3$ and $\xi_j = y_j^2/4 - 2\lambda y_j - 1/6$ for all $j = 1 \dots d$, where $\lambda = \left\|\frac{\theta+x}{2} - Ax - \tilde{b}\right\|/\sqrt{d}$ and $y = (y_1, \dots, y_d) \sim \mathcal{N}(0, 2/3\boldsymbol{I}_d)$.

2.4 Limitations of existing results

The asymptotic analyses from Daudel et al., 2023a recalled in Subsections 2.1 and 2.2 focus on the role of N, α and d in the VR-IWAE bound. However, assessing the effectiveness of the VR-IWAE bound methodology requires understanding the behavior of the gradient estimators of the bound rather than solely focusing on the bound itself. For example, while the IWAE bound can be shown to converge to the target marginal likelihood at a $1/N$ rate for a fixed d (see Domke and Sheldon, 2018, or set $\alpha = 0$ in (8)), increasing N may be harmful to the learning of ϕ in the reparameterized case (Rainforth et al., 2018).

Daudel et al., 2023a, Theorem 1 provides an asymptotic analysis for the Signal-to-Noise Ratio (SNR) of the reparameterized (REP) gradient estimator of the VR-IWAE bound. Indeed, let $M \in \mathbb{N}^\star$, $\varepsilon_{m,1}, \dots, \varepsilon_{m,N}$ be i.i.d. samples generated from q for all $m = 1 \dots M$ and set $Z_{m,j} = f(\varepsilon_{m,j}, \phi; x)$ for all $m = 1 \dots M$ and all $j = 1 \dots N$. Let ψ denote a component of the $\mathbb{R}^{a+b}$-valued variable $(\theta, \phi) = (\theta_1, \dots, \theta_a, \phi_1, \dots, \phi_b)$ and consider the unbiased REP gradient estimator of the VR-IWAE bound w.r.t. to the component ψ given by

$$\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{N}\frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}}\,\partial_\psi \log w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi; x); x). \tag{12}$$

Then, Daudel et al., 2023a, Theorem 1 yields that:

$$\begin{aligned}
&\text{For all } k = 1 \dots a, \quad \mathrm{SNR}[\theta_k\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] \asymp (\sqrt{MN}).\\
&\text{For all } k' = 1 \dots b, \quad \mathrm{SNR}[\phi_{k'}\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] \asymp \begin{cases} \sqrt{M/N} & \text{if } \alpha = 0,\\ \sqrt{MN} & \text{if } \alpha \in (0, 1), \end{cases}
\end{aligned} \tag{13}$$

which motivates the VR-IWAE bound with $\alpha \in (0, 1)$ and $N > 1$ over the ELBO ($N = 1$) and IWAE bound ($\alpha = 0$). Daudel et al., 2023a also derive a doubly-reparameterized (DREP) gradient estimator in the spirit of Tucker et al., 2019, which leads to better performances empirically compared to the REP gradient estimator.

Yet, the SNR results for the REP gradient estimator from Daudel et al., 2023a only distinguish between the cases $\alpha = 0$ and $\alpha \in (0, 1)$, with no further information regarding the role of α across the interval $(0, 1)$. Furthermore, the DREP gradient estimator is not analyzed and the impact of a high-dimensional latent space in the REP and DREP gradient estimators of the VR-IWAE bound is not elucidated either.

In the next section, we will first show that we can obtain more refined results regarding the role of α compared to those presented in Daudel et al., 2023a, Theorem 1 that also require less restrictive assumptions. We will then use our more precise results as a stepping stone towards deriving relevant asymptotic analyses at the gradient level for the VR-IWAE bound that shed light on the role of N, α and d in the VR-IWAE bound methodology.

3 Analyzing the VR-IWAE bound at the gradient level

From here on, 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 q such that $f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ with $\varepsilon \sim q$. In addition, for $\varepsilon \sim q$, 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 $g(y_1, \dots, y_D)$ with respect to y_k evaluated at $y = (y_1, \dots, y_D) \in \mathbb{R}^D$ as in (12) and $\nabla_y g(y)$ denotes the gradient vector $(\partial_{y_k} g(y))_{k=1,\dots,D}$ as in (4). When differentiating, it can be convenient to evaluate the resulting function at a different variable: we write $[h(y, z)]|_{y=y'}$ 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 $\varepsilon_{m,1}, \dots, \varepsilon_{m,N}$ be i.i.d. samples generated from q for all $m = 1 \dots M$, we set $Z_{m,j} = f(\varepsilon_{m,j}, \phi; x)$ for all $j = 1 \dots N$ and all $m = 1 \dots M$ and we let ψ denote a component of the $\mathbb{R}^{a+b}$-valued variable $(\phi, \theta) = (\theta_1, \dots, \theta_a, \phi_1, \dots, \phi_b)$. We consider two estimators of the gradient of the VR-IWAE bound.

• **Reparameterized gradient estimator.** The reparameterized (REP) gradient estimator of the VR-IWAE bound w.r.t. ψ is given by (12), that is

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

• **Doubly-reparameterized gradient estimator.** If ψ is a component of the $\mathbb{R}^b$-valued variable $\phi = (\phi_1, \dots, \phi_b)$, we can use the doubly-reparameterized (DREP) gradient estimator of the VR-IWAE bound w.r.t. ψ given by

$$\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{N} \left(h_{m,j}^{(\alpha)}(\theta, \phi; x) \left[\partial_{\psi'} \log w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi'; x); x)\right]|_{\phi'=\phi} \right), \tag{14}$$

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

$$h_{m,j}^{(\alpha)}(\theta, \phi; x) = \alpha \frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}} + (1-\alpha) \left(\frac{w_{\theta,\phi}(Z_{m,j}; x)^{1-\alpha}}{\sum_{\ell=1}^{N} w_{\theta,\phi}(Z_{m,\ell}; x)^{1-\alpha}} \right)^2. \tag{15}$$

The gradient estimator (14), which was introduced in Daudel et al., 2023a, generalizes the DREP gradient estimator 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:

$$\partial_\psi \log w_{\theta,\phi}(f(\varepsilon,\phi;x);x) = [\partial_{\psi'} \log w_{\theta,\phi}(f(\varepsilon,\phi';x);x) - \partial_{\psi'} \log q_{\phi'}(f(\varepsilon,\phi;x)|x)]|_{\phi'=\phi} \tag{16}$$

meaning that the left-hand side will have nonzero variance even when $q_\phi(\cdot|x)$ matches the exact posterior density $p_\theta(\cdot|x)$ everywhere. As informally shown in Tucker et al., 2019, the DREP gradient estimator of the IWAE bound can then improve on the REP gradient estimator. More generally, (14) can perform better than (12) in practice (Daudel et al., 2023a).

The REP and DREP gradient estimators are both unbiased estimators of the gradient of the VR-IWAE bound (Daudel et al., 2023a). Establishing such properties requires assumptions to interchange derivatives and expectations. We provide in Appendix A.2 precise sufficient conditions to ensure that this can be done rigorously (without overloading the paper with technical assumptions that can be skipped on a first reading) : $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ will be used for the REP estimator and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$ for the DREP estimator.

Let us now derive relevant asymptotic analyses for these gradient estimators.

3.1 Asymptotics of gradient estimators as $N \to \infty$

Let $\alpha \in [0,1)$. Let $\varepsilon \sim q$, $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. copies of ε. Let us denote

$$\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x) = w_{\theta,\phi}(f(\varepsilon,\phi';x);x), \quad \theta \in \Theta\,,\ \phi,\phi' \in \Phi. \tag{17}$$

We will rely on the following assumptions.

(A3) There exist $\mu > 1$ and $N \in \mathbb{N}^\star$ such that $\mathbb{E}(|N^{-1}\sum_{i=1}^N \tilde{w}_{\theta,\phi,\phi}(\varepsilon_j;x)^{1-\alpha}|^{-\mu}) < \infty$.

(A_h) We have that $\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{(1-\alpha)h}) < \infty$.

$(\mathsf{A}_{h'}^{\mathrm{REP}})$ We have that $\mathbb{E}\left(|\partial_\psi\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right)|^{h'}\right) < \infty$.

In (A_h) and $(\mathsf{A}_{h'}^{\mathrm{REP}})$, h and h' are positive exponents to be precised. For ease of exposition, an in-depth discussion of (A3), (A_h) and $(\mathsf{A}_{h'}^{\mathrm{REP}})$ is postponed to Section 3.1.1. We present our first result, which analyses the gradient of the VR-IWAE bound as $N \to \infty$.

Theorem 1 *Assume* (A3), (A_h) *with* $h > 2$, $(\mathsf{A}_{h'}^{\mathrm{REP}})$ *with* $h' > 1$ *and* $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$. *If* $h' < 2$ *assume moreover that* $2/h + 1/h' < 1$. *Then, as* $N \to \infty$,

$$\partial_\psi \ell_N^{(\alpha)}(\theta,\phi;x) = \partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \frac{1}{2N}\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2] + o\left(\frac{1}{N}\right). \tag{18}$$

The proof of Theorem 1 is deferred to Appendix B.2.2. Theorem 1 provides an asymptotic result for the gradient of the VR-IWAE bound which is coherent with the one established in Daudel et al., 2023a, Theorem 3 and reviewed in Section 2.1.

Indeed, the gradient of the VR-IWAE bound is made of two main terms that involve the gradient of quantities identified in (8): the gradient of the VR bound and a fast $1/N$ term incorporating the gradient of $\gamma^{(\alpha)}(\theta,\phi;x)^2$. Theorem 1 thus reinforces the idea that the core quantities to analyze the VR-IWAE bound methodology are the VR bound and $\gamma^{(\alpha)}(\theta,\phi;x)^2$, and that the benefits of using the VR-IWAE bound depend on how those quantities and especially their gradients behave. Note that it is reasonable to assume that $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) \neq 0$ unless $\alpha = 0$ and ψ is among the components of $\phi = (\phi_1,\dots,\phi_b)$ (so that $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = 0$) or we are at a local optimum for (θ,ϕ). In this case the first order term in (18) is the leading term.

As it turns out, the proof of Theorem 1 consists of two steps. Firstly, under (A3), (A_h) and ($\mathsf{A}^{\mathrm{REP}}_{h'}$) with h, h' as in Theorem 1, we show a precise asymptotic result for the expectation of the REP gradient estimator of the VR-IWAE bound as $N \to \infty$. Secondly, we appeal to ($\mathsf{A}^{\mathrm{REP}}_{\mathrm{df}}$) and we rewrite this asymptotic result under the convenient formulation (18), which notably involves using that $\mathbb{E}(\psi\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x)) = \partial_\psi \ell^{(\alpha)}_N(\theta,\phi;x)$ under ($\mathsf{A}^{\mathrm{REP}}_{\mathrm{df}}$). As a byproduct of Theorem 1, we thus have access to an asymptotic result for $\mathbb{E}(\psi\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x))$ as $N \to \infty$ and our next step to further comprehend the behavior of the REP gradient estimator is to capture the behavior of its SNR as $N \to \infty$. Recalling that for a random variable X the SNR is given by $\mathrm{SNR}[X] = |\mathbb{E}(X)|/\sqrt{\mathbb{V}(X)}$, we then have the following theorem.

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

$$\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x)] = \sqrt{MN}\,\frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) - \frac{1}{2N}\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2] + o\left(\frac{1}{N}\right)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)} + o\left(1\right)}\,, \tag{19}$$

$$\textit{where}\quad \psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \frac{1}{(1-\alpha)^2}\mathbb{V}_{\varepsilon\sim q}\left(\partial_\psi\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)\right)\,. \tag{20}$$

The proof of Theorem 2 is deferred to Appendix B.2.3. Once more, the proof is made of two steps: we first establish an asymptotic result for $\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x)]$ as $N \to \infty$ which holds under (A3) and (A_h)–($\mathsf{A}^{\mathrm{REP}}_{h'}$) with $h, h' > 2$, before rewriting this asymptotic result under the convenient formulation (19) thanks to ($\mathsf{A}^{\mathrm{REP}}_{\mathrm{df}}$). Let us now comment on Theorem 2.

• When $\alpha = 0$ in (19) and (20), we recognize the asymptotic SNR results for the REP gradient estimator of the IWAE bound from Rainforth et al., 2018 and the SNR scales like $\sqrt{MN}$ when learning the components of θ, like $\sqrt{M/N}$ when learning the components of ϕ.

• Theorem 2 is a refinement of Daudel et al., 2023a, Theorem 1. Indeed, asymptotic results close to (19) already appear in equations (54) and (55) from the proof of Daudel et al., 2023a, Theorem 1. Those results encompass the case $\alpha = 0$ studied in Rainforth et al., 2018 and encourage setting $\alpha \in (0,1)$ as well as increasing N. However, these results hold under stronger assumptions than the ones we make here and the role that $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$, $\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2]$ and $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ play in the SNR is not uncovered in Daudel et al., 2023a. As already mentioned earlier, the full discussion regarding our assumptions is postponed to Section 3.1.1.

• 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 to study in order to analyze the VR-IWAE bound methodology, it is now clear that $\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ also plays an important role in the success the VR-IWAE bound methodology based on the REP gradient

estimator. Crucially, this quantity can be interpreted in light of Roeder et al., 2017; Tucker et al., 2019: the intuition of Roeder et al., 2017 that the score function adds variance to the REP gradient estimator of the IWAE bound (which led Tucker et al., 2019 to introduce the DREP gradient estimator of the IWAE bound) is confirmed and justified for the REP gradient estimator of the more general VR-IWAE bound by looking at the asymptotics as $N \to \infty$. In this more general setting, the score function adds variance via $[\partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)]|_{\phi'=\phi}$. This can notably be observed by considering the case $p_\theta(\cdot|x) = q_\phi(\cdot|x)$ for which we get (see (107) of Appendix B.2.3 for details) that:

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \mathbb{V}\left([\partial_{\psi'} \log q_{\phi'}(f(\varepsilon, \phi; x)|x)]|_{\phi'=\phi}\right). \tag{21}$$

With the last point in mind, we turn to the DREP gradient estimator of the VR-IWAE bound. Since the DREP gradient estimator only impacts the learning of ϕ, it remains to understand the asymptotic behavior for the variance of the DREP gradient estimator w.r.t. ϕ. We introduce the following assumptions, in which h' and $\tilde{h}$ are positive real numbers to be precised.

$(\mathsf{A}^{\mathrm{DREP}}_{h'})$ We have that $\mathbb{E}\left(\left|[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^{1-\alpha})]|_{\phi'=\phi}\right|^{h'}\right) < \infty.$

$(\widetilde{\mathsf{A}}^{\mathrm{DREP}}_{\tilde{h}})$ We have that $\mathbb{E}\left(\left|\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^{2(1-\alpha)}\right)\right]\right|_{\phi'=\phi}\right|^{\tilde{h}}\right) < \infty.$

This leads us to the theorem below.

Theorem 3 *Assume* (A3), (A_h), ($\widetilde{\mathsf{A}}^{\mathrm{DREP}}_{\tilde{h}}$) *with* $h, \tilde{h} > 2$ *and* ($\mathsf{A}^{\mathrm{DREP}}_{\mathrm{df}}$). *Let* ψ *denote 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* ($\mathsf{A}^{\mathrm{DREP}}_{h'}$) *holds with* $h' > 2$, *then, as* $N \to \infty$,

$$\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N}(\theta, \phi; x)] = \sqrt{MN}\, \frac{\left|\partial_\psi \mathrm{VR}^{(\alpha)}(\theta, \phi; x) + o(1)\right|}{\sqrt{\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)} + o(1)}, \tag{22}$$

$$\textit{where} \quad \psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = \frac{\alpha^2}{(1-\alpha)^2}\mathbb{V}\left(\left[\partial_{\psi'}\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^{1-\alpha})}\right)\right]\Big|_{\phi'=\phi}\right). \tag{23}$$

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

$$\mathrm{SNR}\left[\psi\text{-}g^{(0,\mathrm{DREP})}_{M,N}(\theta, \phi; x)\right] = \sqrt{MN}\, \frac{\frac{1}{2}\partial_\psi[\gamma^{(0)}(\theta, \phi; x)^2] + o(1)}{\sqrt{\psi\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x)} + o(1)}, \tag{24}$$

$$\textit{where} \quad \psi\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x) = \mathbb{V}\left(\frac{[\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^2)]|_{\phi'=\phi}}{2\,\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))^2} - \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\,\mathbb{E}\left([\partial_{\psi'}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)^2)]|_{\phi'=\phi}\right)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))^3}\right). \tag{25}$$

The proof of Theorem 3 is deferred to Appendix B.2.4 and we next comment on Theorem 3.

• The case $\alpha = 0$ in Theorem 3 recovers the asymptotic $\sqrt{MN}$ rate for the SNR of the DREP gradient estimator of the IWAE bound informally derived in Tucker et al., 2019, Section 8.2 and which was used to motivate this estimator over the REP one when $\alpha = 0$. Theorem 3 thus enables us to (i) state conditions for the asymptotic analysis written in Tucker et al., 2019, Section 8.2 to hold and (ii) extend the asymptotic analysis for $\alpha = 0$ to the more general case of the VR-IWAE bound which was not covered in Daudel et al., 2023a.

• Theorem 3 indicates how increasing N in the DREP gradient estimator of the VR-IWAE bound will be beneficial in practice for all $\alpha \in [0, 1)$. Furthermore, and contrary to ψ-$V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ which is non-zero even when $p_\theta(\cdot|x) = q_\phi(\cdot|x)$, we have ψ-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = 0$ when $p_\theta(\cdot|x) = q_\phi(\cdot|x)$. This encourages considering the DREP gradient estimator as an alternative to the REP one. Interestingly, the discontinuity between the cases $\alpha = 0$ and $\alpha \in (0, 1)$ (expressed through the very different expression of ψ-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ in each of these cases) originates from the fact that when $\alpha \in (0, 1)$, the part that dominates in the variance of the DREP gradient estimator as N increases is the variance of the first term in the r.h.s. of (15).

We have obtained that ψ-$V^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ and ψ-$V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ are additional quantities that should be monitored with $\partial_\psi \ell_N^{(\alpha)}(\theta, \phi; x)$ and $\partial_\psi[\gamma^{(\alpha)}(\theta, \phi; x)^2]$ to assess the success of the VR-IWAE bound methodology. Before providing examples for which these quantities are tractable and our results apply, we discuss the assumptions made in Theorems 1 to 3.

3.1.1 Assumptions made in Theorems 1 to 3

We start with Assumption (A3), which is common to Theorems 1 to 3.

Assumption (A3). A condition akin to (A3) already appears in Daudel et al., 2023a, Theorem 1, with the particularity that in Daudel et al., 2023a this condition is assumed to hold for a specific choice of μ, namely $\mu = 4$. In fact, and as implied by Lemma 1 below, the condition assumed in Daudel et al., 2023a, Theorem 1 is equivalent to (A3) and we can also simplify the verification of (A3) by providing a characterization that is equivalent.

Lemma 1 *Set* $\overline{W}_N = N^{-1} \sum_{i=1}^N W_i$ *for all*

Assumptions (A_h) and ($\mathsf{A}_{h'}^{\mathrm{REP}}$). These assumptions are moment conditions which are used together in Theorems 1 and 2 to control the numerator and denominator terms appearing in the REP gradient estimator of the VR-IWAE bound. Similarly to (A3), these conditions are present in Daudel et al., 2023b, Theorem 1 for a specific choice of (h, h'), that is $h = h' = 8$. They also appear in Rainforth et al., 2018 (which studies the case $\alpha = 0$) with $h = h' = 4$. By constrast, our conditions in Theorem 2 only require $h, h' > 2$. Hence, the SNR results of Theorem 2 hold under much less restrictive moment assumptions compared to existing SNR results, rendering them applicable to a wider range of scenarios.

Assumptions ($\widetilde{\mathsf{A}}_{\tilde{h}}^{\mathrm{DREP}}$) and ($\mathsf{A}_{h'}^{\mathrm{DREP}}$). These assumptions are moment conditions used in Theorem 3 to deal with the numerator terms appearing in the DREP gradient estimator of the VR-IWAE bound. Paired up with (A3) and (A_h), they are to the best of our knowledge the first set of conditions permitting to capture the behavior as $N \to \infty$ of the DREP gradient estimator of the VR-IWAE bound. Notice that ($\mathsf{A}_{h'}^{\mathrm{DREP}}$) is only utilized to handle the case $\alpha \in (0, 1)$ in Theorem 3, since only the second term remains in (15) when $\alpha = 0$.

Assumptions ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$) and ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$). As mentioned earlier, these assumptions ensure that interchanging the derivation and expectation signs is valid when needed in our proofs, see Appendix A.2 for a statement of these assumptions and how they are used in our proofs.

3.1.2 Illustrating Theorems 1 to 3

We revisit Examples 1 and 2 by studying this time the gradient of the VR-IWAE bound.

Example 3 (Example 1 revisited) *Consider the setting of Example 1 with the reparameterization given by $Z = f(\varepsilon, \phi; x) = \varepsilon + \phi$ where $\varepsilon \sim \mathcal{N}(0 \cdot \boldsymbol{u}_d, \boldsymbol{I}_d)$ and the goal is to learn ϕ. Then, we can apply Theorems 1 to 3 and all the terms appearing in these results are analytically tractable. In particular, letting $k = 1, \dots d$, $\theta = \epsilon \cdot \boldsymbol{u}_d$ and $\phi = 0 \cdot \boldsymbol{u_d}$ with $\epsilon > 0$, we have that: for all $\alpha \in [0, 1)$, as $N \to \infty$,*

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

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

$$\mathrm{SNR}[\phi_k\text{-}g_{M,N}^{(0,\mathrm{DREP})}(\theta, \phi; x)] = \frac{\sqrt{MN}(1 + o(1))}{\left(\exp\left(4d\epsilon^2\right) - 4\exp\left(2d\epsilon^2\right) + 4\exp\left(d\epsilon^2\right) - 1\right)^{\frac{1}{2}}},$$

$$\mathrm{SNR}[\phi_k\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)] \gg \sqrt{MN} \qquad \text{if } \alpha \in (0, 1).$$

The derivation details for this example are deferred to Appendix B.3.2.

Example 3 supports the claim that α enables a bias-variance tradeoff for the REP and DREP gradient estimators of the VR-IWAE bound. Indeed, the SNR in the REP case is improved by taking $\alpha \in (0, 1)$ and more specifically by increasing α at the cost of increasing the bias through the quantity $\alpha\epsilon$ in the numerator (corresponding to the gradient of the VR bound). It also turns out that $\phi_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = 0$ for $\alpha \in (0, 1)$. Although the setting where $\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta, \phi; x) = 0$ for certain values of α seems unlikely enough in practice to warrant a specific theoretical study of the leading order term in the SNR, it is noteworthy that in Example 3 the case $\alpha \in (0, 1)$ for the DREP gradient estimator improves the rate in N.

If we now keep $\alpha \in [0, 1)$ and $\epsilon > 0$ fixed in Example 3, we see that this example hints at the fact that the asymptotic regimes predicted by Theorems 1 to 3 may not kick in as the dimension d increases unless N is unpractically large. In other words, Theorems 1 to 3 may not capture the behavior of the REP gradient estimator of the VR-IWAE bound as d increases, much like Daudel et al., 2023b noted that (8) may not capture the behavior of the VR-IWAE bound as d increases. We now move on to a second example.

Example 4 (Example 2 revisited) *Consider the setting of Example 2 with the reparameterization given by* $Z = \sqrt{\frac{2}{3}}\varepsilon + Ax + \tilde{b}$, *where* $\varepsilon \sim \mathcal{N}(0 \cdot \boldsymbol{u}_d, \boldsymbol{I}_d)$. *Then, the assumptions from Theorems 1 to 3 are met and all the terms appearing in these results are analytically tractable. In particular, denoting* $x = (x_1, \dots, x_d)$, $\theta = (\theta_1, \dots, \theta_d)$, $\tilde{b} = (\tilde{b}_1, \dots, \tilde{b}_d)$, *letting* $Ax + \tilde{b} = (\theta + x)/2 + \epsilon \boldsymbol{u}_d$ *with* $\epsilon > 0$ *and* $k = 1 \dots d$, *we have that: for all* $\alpha \in [0, 1)$, *as* $N \to \infty$,

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

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \sqrt{MN}\, \frac{\left(\frac{3\epsilon\alpha}{4-\alpha} + \frac{12\epsilon(1-\alpha)(4-\alpha)^{d-1} \exp\left(\frac{24(1-\alpha)^2 d\epsilon^2}{(5-2\alpha)(4-\alpha)}\right)}{N 3^{d/2} (5-2\alpha)^{\frac{d}{2}+1}}\right) (1 + o(1))}{\frac{(4-\alpha)^{d/2}}{(15-6\alpha)^{d/4}} \exp\left(\frac{12(1-\alpha)^2}{(4-\alpha)(5-2\alpha)} d\epsilon^2\right) \sqrt{\frac{2}{5-2\alpha} + \left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2}}$$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)] = \begin{cases} \sqrt{MN} \frac{\frac{24\epsilon 4^{d-1} \exp\left(\frac{24d\epsilon^2}{4\cdot 5}\right)}{3^{d/2} 5^{\frac{d}{2}+1}}}{\sqrt{\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x)}} (1 + o(1)) & \text{if } \alpha = 0, \\ 4\alpha^{-1}\, \mathrm{SNR}[\tilde{b}_k\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] & \text{if } \alpha \in (0, 1), \end{cases}$$

where $\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta, \phi; x)$ *is given by* (156) *in Appendix B.3.3. The derivation details for this example are deferred to Appendix B.3.3.*

Example 4 illustrates again the bias-variance tradeoff occurring in the REP and DREP gradient estimators of the VR-IWAE bound as α varies. It also showcases how the DREP gradient estimator can significantly outperform the REP one when $\alpha \in (0, 1)$ via a direct improvement in terms of SNR that is expressed through the multiplicative factor $4\alpha^{-1}$. Lastly, the conclusions drawn for Example 3 regarding the impact of d on the SNRs transfer to Example 4, in the sense that Theorems 1 to 3 may not capture the asymptotic behavior of the REP and DREP gradient estimators as d increases. This then calls for a study of the REP and DREP gradient estimators of the VR-IWAE bound as both d and N go to infinity.

3.2 Asymptotics of gradient estimators as $N, d \to \infty$

We now investigate the behavior of the REP and DREP gradient estimators of the VR-IWAE bound in high dimensions. From here on, we denote these two gradient estimators by $\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)$ and $\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)$ respectively to highlight their dependency with respect to d. In principle, θ, ϕ and x may also depend on d, but we do not make this dependency explicit in the notation for the sake of conciseness. Now observe that, letting $Z \sim q_\phi(\cdot|x)$, the difference $\log w_{\theta,\phi}(Z; x) - \mathbb{E}_{Z \sim q_\phi(\cdot|x)}(\log w_{\theta,\phi}(Z; x))$ can be seen as a centered log likelihood ratio over the latent space $\mathbb{R}^d$ and, provided that the first two moments of

$\log w_{\theta,\phi}(Z;x)$ exist, we can write without any loss of generality that

$$\log w_{\theta,\phi}(Z;x) = \mathbb{E}_{Z\sim q_\phi(\cdot|x)}(\log w_{\theta,\phi}(Z;x)) - B_d S_d, \tag{27}$$

where $B_d^2 = \mathbb{V}(\log w_{\theta,\phi}(Z;x))$ and S_d is a centered and normalized random variable. As $d \to \infty$, the log likelihood ratio $\log w_{\theta,\phi}(Z;x)$ will typically behave after centering and normalizing as a standardized Normal variable, which corresponds to taking $S_d \sim \mathcal{N}(0,1)$ in (27). Consequently, we offer in our study to analyze the Gaussian case in order to derive insights regarding the behavior of the REP and DREP gradient estimators in high dimensional settings.

Remark 1 *The setting studied in Daudel et al., 2023a can easily be recovered as a special case of* (27)*. Indeed,* (27) *can be rewritten as*

$$\log\left(\frac{p_\theta(Z|x)}{q_\phi(Z|x)}\right) = -\log\mathbb{E}\left(\exp(-B_d S_d)\right) - B_d S_d,$$

since $\mathbb{E}_{Z\sim q_\phi(\cdot|x)}(p_\theta(Z|x)/q_\phi(Z|x)) = 1$*. The link with the cases studied in Daudel et al., 2023a and recalled in Section 2.2 is then immediate by setting* $B_d^2 = \sigma^2 d$ *with* $\sigma^2 > 0$ *and considering the choices of* S_d *described in* (10) *and* (11) *respectively.*

Since the REP and DREP gradient estimators of the VR-IWAE bound involve differentiating $\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)$, we will in fact need to account for the dependency in (θ,ϕ,ϕ') of the log-weights. This amounts to considering instead of (27)

$$\begin{aligned}&\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x) = \mathbb{E}_{\varepsilon\sim q}(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)) - B_d(\theta,\phi,\phi';x) S_d(\varepsilon,\theta,\phi,\phi';x)\\ &\text{with}\quad B_d(\theta,\phi,\phi';x) = \mathbb{V}_{\varepsilon\sim q}\left(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right).\end{aligned} \tag{28}$$

As we focus on the Gaussian case in our study, we also work under the following assumption.

(B1) For $\varepsilon \sim q$, the random field $(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))_{(\theta,\phi,\phi')\in\Theta\times\Phi^2}$ is a Gaussian random field.

To guaranty that we can interchange derivative and expectation signs when needed for our study of the REP and DREP gradient estimators under (B1), we further rely on additional assumptions, denoted by $(\mathrm{B}_{\mathrm{df}}^{\mathrm{REP}})$ and $(\mathrm{B}_{\mathrm{df}}^{\mathrm{DREP}})$ respectively. These assumptions, which do not depend on $\alpha \in [0,1)$, are postponed to Appendix A.2.2 to not overload the paper with technical assumptions that can be overlooked in a first read (as we did for $(\mathrm{A}_{\mathrm{df}}^{\mathrm{REP}})$ and $(\mathrm{A}_{\mathrm{df}}^{\mathrm{DREP}})$).

Starting with the REP gradient estimator and denoting the marginal log likelihood by $\ell_d(\theta;x)$ to indicate that this quantity depends on d, we then have the following theorem.

Theorem 4 *Assume* (B1) *and* $(\mathrm{B}_{\mathrm{df}}^{\mathrm{REP}})$*. If, as* $N, d \to \infty$*,*

$$\log N = o\left(B_d^2(\theta,\phi,\phi;x)\right) \tag{29}$$

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \partial_\psi \ell_d(\theta;x)]\,\frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)} = o(1), \tag{30}$$

then: as $N, d \to \infty$*,*

$$\mathrm{SNR}[\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)]\,(1+o(1)) + o(1). \tag{31}$$

The proof of Theorem 4 is deferred to Appendix C.3.1 and we now comment on this result. Theorem 4 provides conditions tying N to d under which the REP gradient estimator of the VR-IWAE bound suffers from a weight collapse phenomenon as $N, d \to \infty$ in the Gaussian random field case for all $\alpha \in [0, 1)$. Specifically: as $N, d \to \infty$, if (i) N does not grow exponentially with $B_d(\theta, \phi; x)^2$ and (ii) $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) - \partial_\psi \ell_d(\theta; x)]$ does not grow faster than $\sqrt{\log N}/B_d(\theta, \phi, \phi; x)$, then the improvement at the SNR level obtained by using the VR-IWAE bound with $N \gg 1$ is negligible compared to having used $N = 1$ (ELBO) instead.

Indeed, by distinguishing between the two settings $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = o(1)$ and $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] \gg 1$ in (31), we see that in both settings the case $N \gg 1$ does not yield significant gains compared to the case $N = 1$ in terms of SNR. In fact, we get an even more precise result when $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] \gg 1$ since (31) then reads $\mathrm{SNR}[\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] = \mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)](1 + o(1))$ so that it becomes equivalent to use $N \gg 1$ or $N = 1$ at the SNR level as $N, d \to \infty$ under the conditions stated in Theorem 4.

Consequently, Theorem 4 reveals that the behavior as a function of d of $B_d(\theta, \phi; x)^2$ and $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) - \partial_\psi \ell_d(\theta; x)]$ is crucial, as these quantities have a direct impact on the computational budget required to benefit from the VR-IWAE bound methodology based on the REP gradient estimator. As we shall see in our forthcoming Example 5, we can find instances where $B_d(\theta, \phi, \phi)^2$ is of order d so that the condition (29) calls for N to not grow faster than exponentially in d. In this scenario, circumventing the condition (29) for a high dimension d would thus require using an unpractical amount of samples N in practice. As for (30), focusing first on the case $\psi \in \{\phi_1, \ldots, \phi_b\}$ it holds that:

$$\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) - \partial_\psi \ell_d(\theta; x) = \psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) = \partial_\psi \log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)$$

so that (30) also reads

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)] \, \frac{\sqrt{\log N}}{B_d(\theta, \phi, \phi; x)} = o(1), \tag{32}$$

Thus, (30) directly involves the weight collapse term $\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)]$ appearing in (31) and we can interpret (30) as follows: this condition requires the REP gradient estimator of the VR-IWAE bound with $N = 1$ to have a high enough SNR as d increases if we wish to benefit from the VR-IWAE bound methodology. More generally, since

$$\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x) - \partial_\psi \ell_d(\theta; x) = \partial_\psi \log \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))} \right) = \partial_\psi \log \left(\frac{p_\theta(f(\varepsilon, \phi; x)|x)}{q_\phi(f(\varepsilon, \phi; x)|x)} \right)$$

the condition (30) emphasizes how important the behavior of the log of the normalized weight $\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)/\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x))$ is to make the most of the VR-IWAE bound methodology: if the SNR of its gradient is not high enough, it may take an unpractical amount of samples N as d increases to obtain an improvement in the SNR using the REP gradient.

Crucially, we see thanks to the expression above that (30) does not depend on α. Hence, none of the assumptions made in Theorem 4 depend on α and if the weight collapse phenomenon occurs for one value of α in the context of Theorem 4, it will thus occur for all $\alpha \in [0, 1)$. Furthermore, (30) highlights how the choice of the variational family $\mathcal{Q}$ strongly impacts the VR-IWAE bound methodology: the more

expressive this family is, the closer $q_\phi(f(\cdot,\phi;x)|x)$ can get to the target $p_\theta(f(\cdot,\phi;x)|x)$ and the more we can avoid a weight collapse in the REP gradient estimator of the VR-IWAE bound.

Note that an underlying assumption when stating (30) is to have $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \mathbb{V}(\partial_\psi \log \tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)) > 0$. Following earlier comments we made in relation to Roeder et al., 2017; Tucker et al., 2019, we expect this quantity to be positive in practice. We next turn to the study of the DREP gradient estimator and we have the following theorem.

Theorem 5 *Assume* (B1) *and* ($\mathsf{B}_{\mathrm{df}}^{\mathrm{DREP}}$). *Let* ψ *denote a component of the* $\mathbb{R}^b$*-valued variable* $\phi = (\phi_1,\ldots,\phi_b)$. *The following assertions hold.*

(i) If, as $N,d\to\infty$, *there exists* $r>0$ *such that* $\sqrt{\log N} \ll B_d(\theta,\phi,\phi;x) \ll N^r$ *and*

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) - [\partial_{\psi'}\log \mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi}] \frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)} = o(1), \tag{33}$$

$$\mathrm{SNR}\left[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right] \left(\frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)}\right)^{1/2} = o(1), \tag{34}$$

then: as $N,d\to\infty$,

$$\mathrm{SNR}[\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)] = \mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)]\ (1+o(1)) + o(1). \tag{35}$$

(ii) If, as $N,d\to\infty$, *there exists* $r>0$ *such that* $\sqrt{\log N} \ll B_d(\theta,\phi,\phi;x) \ll N^r$ *and* $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = 0$ *for* d *large enough, then: as* $N,d\to\infty$,

$$\mathbb{E}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = \mathbb{E}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x))(1+o(1))\ , \tag{36}$$

$$\mathbb{V}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = \left(\mathbb{E}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x))\right)^2 O\left(\frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)}\right)\ . \tag{37}$$

The proof of Theorem 5 is deferred to Appendix C.3.2. In the same vein as Theorem 4, Theorem 5 provides conditions on N and d leading to a weight collapse in the Gaussian case for the DREP gradient estimator of the VR-IWAE bound with $\alpha\in[0,1)$ and $N,d\to\infty$.

More specifically, assertion (i) is the analogue of Theorem 4 when considering the DREP gradient estimator instead of the REP one. In particular, (33) and (34) are the counterpart of (30) (recall indeed that (30) can be rewritten as (32) when $\psi\in\{\phi_1,\ldots,\phi_b\}$). Since

$$\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) - [\partial_{\psi'}\log \mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi} = \left[\partial_{\psi'}\log\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))}\right)\right]\Bigg|_{\phi'=\phi},$$

(33) showcases once more how the success of the VR-IWAE bound methodology depends on the behavior of the log of the normalized weights, while (34) highlights the need to have a high enough SNR when $N=1$ if we wish to benefit from the VR-IWAE bound methodology based on the DREP gradient estimator as d increases.

In the spirit of the DREP gradient methodology (Roeder et al., 2017; Tucker et al., 2019) discussed earlier, we include in assertion (ii) of our analysis the setting where $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) =$

$\mathbb{V}(\partial_{\psi'}\,[\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)]|_{\phi'=\phi}) = 0$ for d large enough. Interestingly, the main take-away message does not change between assertions (i) and (ii), as the weight collapse occurs in both settings, with the condition (29) that already appeared in Theorem 4 being shared between both assertions (in addition to the mild condition that there exists $r > 0$ such that, as $N, d \to \infty$, $B_d(\theta, \phi, \phi; x) \ll N^r$).

We now revisit the setting from Examples 1 and 3, this time for high dimensions.

Example 5 (Examples 1 and 3 revisited for high dimensions) *Consider the setting of Examples 1 and 3, in which we let $\theta = \epsilon \cdot \boldsymbol{u}_d$ and $\phi = 0 \cdot \boldsymbol{u_d}$ with $\epsilon > 0$. Then, (29) becomes*

$$\frac{\log N}{d} = o(1) \tag{38}$$

and we can apply Theorems 4 and 5. Specifically, for all $\alpha \in [0, 1)$ and all $\psi \in \{\phi_1, \ldots, \phi_d\}$: as $N, d \to \infty$, we have under (38) that (31) holds, and if moreover $d \ll N^r$ for some $r > 0$, (36) and (37) hold. In particular, we have that

$$\begin{aligned}
&\mathrm{SNR}[\psi\text{-}g^{(\alpha,\mathrm{REP})}_{1,N,d}(\theta,\phi;x)] = \epsilon + o(1)\ , \\
&\mathbb{E}(\psi\text{-}g^{(\alpha,\mathrm{DREP})}_{1,N,d}(\theta,\phi;x)) = -\epsilon(1+o(1))\ , \\
&\mathbb{V}(\psi\text{-}g^{(\alpha,\mathrm{DREP})}_{1,N,d}(\theta,\phi;x)) = O\left(\sqrt{\frac{\log N}{d}}\right)\ .
\end{aligned}$$

The derivation details for this example are deferred to Appendix C.4.

As $N, d \to \infty$, Example 5 states that the condition (38), that is the number of samples N does not grow quicker than exponentially with the dimension of the latent space d, leads to a weight collapse in the REP and DREP gradient estimators of the VR-IWAE bound for all $\alpha \in [0, 1)$ (in the DREP case we also require $d \ll N^r$ for some $r > 0$, which is a very mild condition in practice). Bypassing (38) imposes a heavy computational budget as d increases. Example 5 thus illustrates how the VR-IWAE bound methodology with $N > 1$ and $\alpha \in [0, 1)$, which includes the IWAE bound methodology ($\alpha = 0$), may not lead to improvements over the ELBO one ($N = 1$) in high dimensions due to a collapse in its gradient estimators. Let us now put Example 5 in perspective with the findings of Daudel et al., 2023a.

Daudel et al., 2023a, Example 3 shows that for the setting described in Examples 1 and 3: as $N, d \to \infty$ and under (38), a weight collapse phenomenon occurs in the VR-IWAE bound for all $\alpha \in [0, 1)$. Consequently, Example 5 lends credence to the idea that the behavior of the VR-IWAE bound and of its REP and DREP gradient estimators are intertwined through conditions of the form (38) (and more generally of the form (29)). While Daudel et al., 2023a focused only on the behavior of the VR-IWAE bound, we uncover in Theorems 4 and 5 additional conditions on top of conditions of the form (29) which pertain to the SNR of the REP and DREP gradient estimators when $N = 1$ and permit us to capture the behavior of these estimators for the Gaussian case when $N > 1$ in high-dimensions.

We conjecture that, under adequate conditions making the centered and normalized log likelihood ratio $S_d(\varepsilon, \theta, \phi, \phi'; x)$ in (28) well approximated by a Gaussian random field as d increases, there will be a weight collapse in the REP and DREP gradient estimators of the VR-IWAE bound under conditions akin to (possibly stronger than) (29), (30), (33) and (34). Indeed, we already know from Daudel et al., 2023a that in instances where (11) holds (the setting described in Example 2 is one such instance), the condition for collapse in the VR-IWAE bound is to have N that does not grow quicker than exponentially with $d^{1/3}$.

Establishing associated weight collapse results for the REP and DREP gradient estimators of the VR-IWAE bound is left for future work.

We next present several numerical experiments.

4 Numerical experiments

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

4.1 Gaussian example

We consider the Gaussian example described in Example 1 with the reparameterization given by $Z = f(\varepsilon, \phi; x) = \varepsilon + \phi$ where $\varepsilon \sim \mathcal{N}(0 \cdot \boldsymbol{u}_d, \boldsymbol{I}_d)$. Letting $k = 1, \dots d$, $\theta = \epsilon \cdot \boldsymbol{u}_d$ and $\phi = 0 \cdot \boldsymbol{u_d}$ with $\epsilon > 0$, we know from Example 3 that Theorem 2 predicts the following asymptotic behavior of $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta, \phi; x)]$ for all $\alpha \in [0, 1)$: as $N \to \infty$,

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

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 $N \to \infty$

$$\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta, \phi; x)] = \frac{\sqrt{MN}\epsilon\alpha \exp\left(\frac{-(1-\alpha)^2 d\epsilon^2}{2}\right)}{\sqrt{1 + (1-\alpha)^2\epsilon^2}}(1 + o(1)), \quad \alpha \in (0, 1) \tag{39}$$

matches with the observed $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta, \phi; x)]$ as N increases (note that only the leading order term remains in (39)). To this end, we let $\epsilon \in \{0.2, 1., 2\}$, $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $d \in \{10, 100, 500\}$, $N \in \{2^j, j = 1 \dots 15\}$, $M = 1$ 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 (39) 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 3 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 highest the value of α (and the lowest the value of ϵ), the lowest N needs to be before the observed SNR and the prediction made by (39) coincide. This can be explained by 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, we expect $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta, \phi; x)]$ to collapse to $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,1,d}(\theta, \phi; x)]$ unless we use an unpractical amount of samples N (Example 5), meaning that the asymptotic behavior (39) will not reflect the observed SNRs anymore. This is indeed what we observe in Figure 1 ($N = 10^0$ corresponds to $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{1,1,d}(\theta, \phi; x)] = \epsilon$), with the particularity that the weight collapse occurs quicker the higher the value of ϵ, that is the further the variational approximation is to the target posterior density. Interpreting the different values of ϵ as different stages of the training procedure, Example 5 sheds light on the role the flexibility of the variational family plays in the success of the VR-IWAE bound framework.

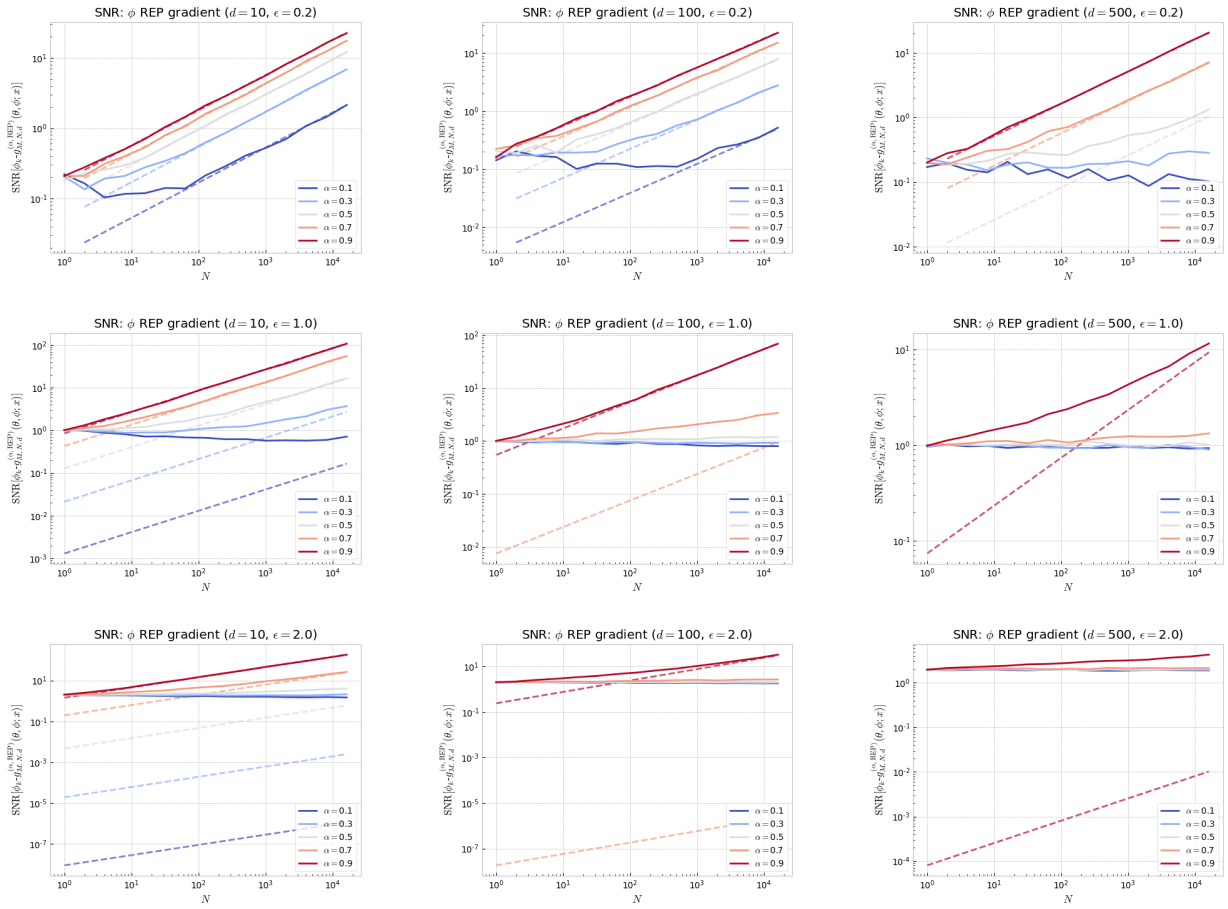

Figure 1: Plotted is $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N, for varying values of (α, d, ϵ) and a random coordinate ϕ_k. The solid lines correspond to $\mathrm{SNR}[\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$, while the dashed lines correspond to predictions of the form (39).

Example 5 also predicts the collapse of the DREP gradient estimator as d increases, such as the collapse $\mathbb{E}(\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x))$ to $\mathbb{E}(\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,1,d}(\theta,\phi;x))$. Similarly to the REP case, this collapse is well-observed by setting a high value for ϵ, see Figure 2 in which $\epsilon = 2$. The IWAE case ($\alpha = 0$) unfolds in the same manner as d increases and it is provided in Appendix D.1 for the sake of completeness. We now move on to a second numerical experiment.

4.2 Linear Gaussian example

We are next interested in the linear Gaussian example from Rainforth et al., 2018, which we reviewed in Example 2. The data set $\mathcal{D} = \{x_1, ..., x_T\}$ is generated by sampling $T = 1024$ datapoints from $\mathcal{N}(0, 2\boldsymbol{I}_d)$ and we consider the reparameterization given by $Z = \sqrt{\frac{2}{3}}\varepsilon + Ax + \tilde{b}$ where $\varepsilon \sim \mathcal{N}(0 \cdot \boldsymbol{u}_d, \boldsymbol{I}_d)$. Letting $\theta = \theta^\star + 2\epsilon$, $A = A^\star$ 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 $\tilde{b}^\star = \frac{1}{2}\theta^\star$, we have that $Ax + \tilde{b} = \frac{1}{2}(x + \theta) + \epsilon$. For all $\alpha \in [0, 1)$, we thus know from Example 4 that Theorems 2 and 3 predict the

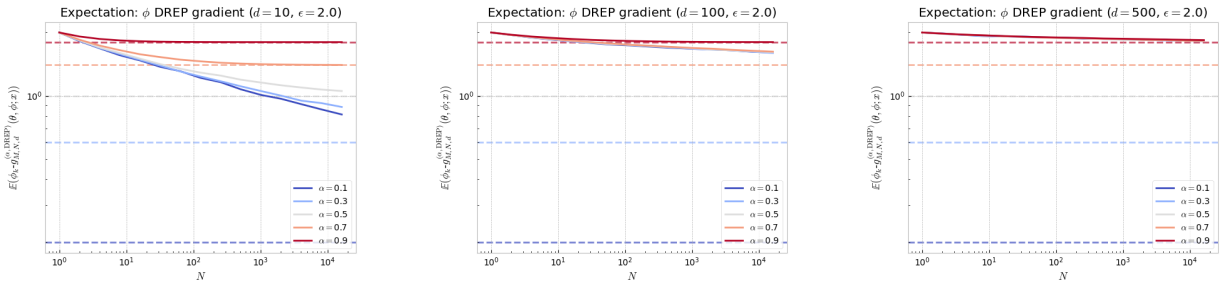

Figure 2: Plotted is $\mathbb{E}(\phi_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x))$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N, for varying values of (α, d) and a random coordinate ϕ_k. The solid lines correspond to $\mathbb{E}(\phi_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x))$, while the dashed lines correspond to predictions of the form $y=\epsilon\alpha$.

asymptotic behavior as $N\to\infty$ of $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$ respectively. Let us now check the validity of the asymptotic results obtained in Example 4. Letting $k=1\dots d$ and focusing first on the learning of $\tilde{b}$ when $\alpha\in(0,1)$, this boils down to checking whether as N increases the observed $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$ match with the asymptotic behaviors: as $N\to\infty$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)] = \frac{\sqrt{MN}\frac{3\epsilon\alpha}{4-\alpha}(1+o(1))}{\frac{(4-\alpha)^{d/2}}{(15-6\alpha)^{d/4}}\exp\left(\frac{12(1-\alpha)^2}{(4-\alpha)(5-2\alpha)}d\epsilon^2\right)\sqrt{\frac{2}{5-2\alpha}+\left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2}} \tag{40}$$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)] = 4\alpha^{-1}\,\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x)] \tag{41}$$

(note that once more only the leading order term remains in (40)). Now let $\epsilon\in\{0.2,1\}$, $\alpha\in\{0.1,0.3,0.5,0.7;0.9\}$, $d\in\{10,100,500\}$, $N\in\{2^j, j=1\dots15\}$, $M=1$ and $\tilde{b}_k$ be a random coordinate in $(\tilde{b}_1,\dots,\tilde{b}_d)$. The results are shown in Figures 3 and 4.

Starting with Figure 3, 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 2 and 3 when $\alpha\in(0,1)$ apply here too. Moving on to Figure 4, 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 D.2. They further validate our conclusions regarding the empirical validity of Theorems 2 and 3.

As d increases, a weight collapse phenomenon can be observed in Figures 3 and 4 for the REP and DREP gradient estimators of the VR-IWAE bound. This is coherent with the work of Daudel et al., 2023a, in which they show (and empirically check) that in the context of Example 4 there is weight collapse phenomenon occuring in the VR-IWAE bound as d increases unless N is unpractically large.

5 Conclusion

We proposed two complementary SNR analyses for the REP and DREP gradient estimators of the VR-IWAE bound. Those analyses improve on and enrich the existing studies by casting a new light on

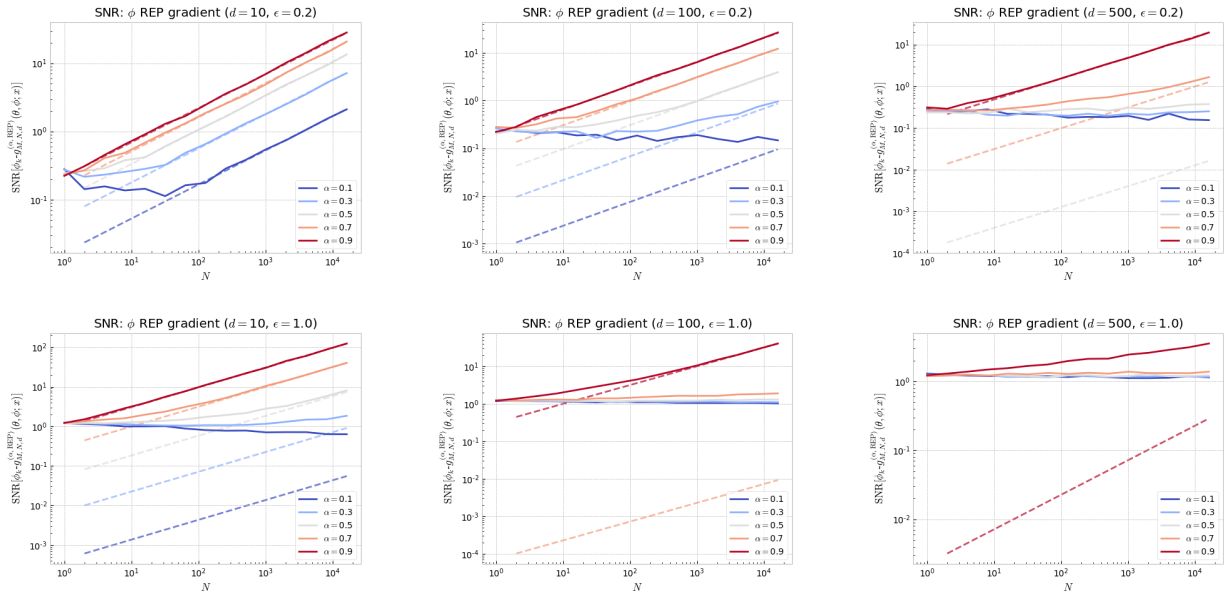

Figure 3: Plotted is $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$, while the dashed lines correspond to predictions of the form (40).

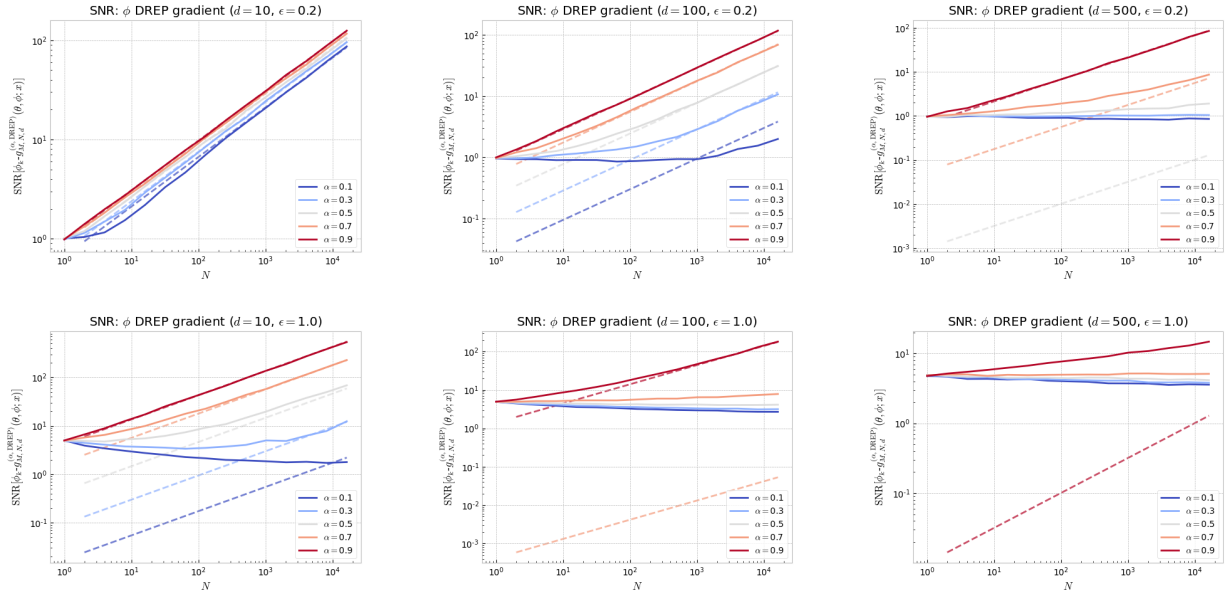

Figure 4: Plotted is $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(\alpha,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$, while the dashed lines correspond to predictions of the form (40).

the innerworkings of the VR-IWAE bound methodology as well as enabling a comparison of the ELBO, IWAE and VR bounds methodologies at the gradient level.

Our work advances the understanding of importance weighted variational inference methods and we now briefly mention potential direction of research to extend it. One may for instance investigate how our theoretical results for the Gaussian random field case can be extended to the approximate Gaussian random field case of Section 3.2. Another possibility is to see if there are instances where the weight collapse can be avoided, for example via the tuning of α.

A Preliminaries

A.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 $1 \leq p < \infty$,

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

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 $\|X\|_p \leq \|X\|_{p'}$ for all $p \leq p'$. We will also repeatedly use the Hölder inequality

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

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

A.2 Interchanging derivative and expectation signs

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

$$\partial_\psi \mathbb{E}_{\varepsilon\sim q}\left(g(\psi, \varepsilon)\right) = \mathbb{E}_{\varepsilon\sim q}\left(\partial_\psi g(\psi, \varepsilon)\right), \quad \psi \in \Psi . \tag{42}$$

General conditions to make the identity (42) 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 (17) 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.2 is then concerned with providing sufficient conditions to ensure that the interchanges of derivatives and expectations necessary to establish the results of Section 3.1 and Section 3.2 are valid.

A.2.1 Interchanging derivative and expectation signs in Section 3.1

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 $\ell = 0, 1, 2$,

$$\mathbb{E}_{\varepsilon \sim q} \left(\sup_{(\theta', \phi') \in \mathcal{V}} \left| \tilde{w}_{\theta', \phi', \phi'}(\varepsilon; x)^{\ell(1-\alpha)} \, \partial_{\psi'} \left(\log \tilde{w}_{\theta', \phi', \phi'}(\varepsilon; x) \right) \right| \right) < \infty \, . \tag{43}$$

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 3.1.

Proposition 1 *Under* ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$)*, we have for all* $(\theta, \phi) \in \Theta \times \Phi$*,*

$$\partial_\psi \mathbb{E}_{\varepsilon \sim q} \left(\tilde{w}_{\theta, \phi, \phi}(\varepsilon; x)^{1-\alpha} \right) = \mathbb{E} \left(\partial_\psi \left(\tilde{w}_{\theta, \phi, \phi}(\varepsilon; x)^{1-\alpha} \right) \right) \, , \tag{44}$$

$$\partial_\psi \mathbb{E}_{\varepsilon \sim q} \left(\tilde{w}_{\theta, \phi, \phi}(\varepsilon; x)^{2(1-\alpha)} \right) = \mathbb{E} \left(\partial_\psi \left(\tilde{w}_{\theta, \phi, \phi}(\varepsilon; x)^{2(1-\alpha)} \right) \right) \, . \tag{45}$$

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

$$\partial_\psi \mathbb{E}_{\varepsilon_j \overset{\text{i.i.d.}}{\sim} q} \left(\log \left(\sum_{j=1}^{N} \tilde{w}_{\theta, \phi, \phi}(\varepsilon_j; x)^{1-\alpha} \right) \right) = \mathbb{E}_{\varepsilon_j \overset{\text{i.i.d.}}{\sim} q} \left(\partial_\psi \log \left(\sum_{j=1}^{N} \tilde{w}_{\theta, \phi, \phi}(\varepsilon_j; x)^{1-\alpha} \right) \right) \, . \tag{46}$$

Proof. Suppose that ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$) holds. Observe that, for $\ell = 1, 2$, by (17), 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 (43) for $\ell = 1, 2$ leads to (44) and (45), respectively.

Let now $N \geq 1$. To get (46), 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

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

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 $j = 1, \ldots, N$,

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

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 (43) with $\ell = 0$ ensures that the previous condition holds and we have concluded the proof of (46). □

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

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi',\phi'}(\varepsilon;x)^{1-\alpha}\right)\right|\right) < \infty\,. \tag{47}$$

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right|\right) < \infty\,. \tag{48}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{1-\alpha}\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\log q_{\phi'}(Z|x)\right)\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right|\right) < \infty\,. \tag{49}$$

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

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi',\phi'}(\varepsilon;x)^{2}\right)\right|\right) < \infty\,, \tag{50}$$

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2}\right)\right|\right) < \infty\,, \tag{51}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{2}\sup_{(\theta,\phi')\in\mathcal{V}}\left|\partial_{\psi'}\left(\log q_{\phi'}(Z|x)\right)\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right|\right) < \infty\,. \tag{52}$$

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 3.1.

Proposition 2 *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

$$\left[\partial_{\psi'}\mathbb{E}_{\varepsilon\sim q}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi} = \mathbb{E}_{\varepsilon\sim q}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi}\right), \tag{53}$$

$$\partial_{\psi}\mathbb{E}_{\varepsilon\sim q}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}\right) = \alpha\,\mathbb{E}_{\varepsilon\sim q}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}\right)\right]\Big|_{\phi'=\phi}\right). \tag{54}$$

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

$$\partial_{\psi}\mathbb{E}_{\varepsilon\sim q}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{2}\right) = -\mathbb{E}_{\varepsilon\sim q}\left(\left[\partial_{\psi'}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{2}\right)\right]\Big|_{\phi'=\phi}\right). \tag{55}$$

Proof. First observe that (55) in Assertion (ii) corresponds to (54) 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 (47)–(49) imply the identities (53) and (54).

We will in fact prove that more generally: for all $\alpha \in \mathbb{R}$, Conditions (47)–(49) imply the identities (53) and (54). The case $\alpha = 1$ is immediate so let us pick $\alpha \in \mathbb{R} \setminus \{1\}$ in the following and assume that (47)–(49) hold. Note first that (53) is a simple interchange of the derivative and expectation signs which holds by dominated convergence using (48). To then prove (54), observe that by (47), we can interchange the derivative and expectation signs on the left-hand side of (54), that is: for all $\alpha \in (0, 1)$,

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

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 (54) it remains to prove that

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

Observe then that, using (17), it holds that

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

Hence the left-hand side of (56) reads

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

and the proof of (56) will be concluded if we can prove that

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

The equality (a) follows from interchanging the integral and the derivative signs again thanks to (48) 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 (49) 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 2 is thus concluded. □

The conclusions from Proposition 2 will be used to express the constants in the asymptotic behavior of the SNR of the DREP estimator, see the proof of Theorem 3.

A.2.2 Interchanging derivative and expectation signs in Section 3.2

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

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \left(\partial_{\psi'} \log \tilde{w}_{\theta',\phi',\phi'}(\varepsilon; x)\right)^2\right) < \infty \,, \tag{57}$$

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \left(\log \tilde{w}_{\theta',\phi',\phi'}(\varepsilon; x)\right)^2\right) < \infty \,. \tag{58}$$

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 Theorem 4.

Proposition 3 *Let $\varepsilon \sim q$. Under (B1) and ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$), we have that: for all $(\theta, \phi) \in \Theta \times \Phi$,*

$$\partial_\psi \mathbb{E}_{\varepsilon\sim q}\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right) = \mathbb{E}_{\varepsilon\sim q}\left(\partial_\psi \log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right) \tag{59}$$

$$\partial_\psi \mathbb{E}_{\varepsilon\sim q}\left(\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right)^2\right) = \mathbb{E}_{\varepsilon\sim q}\left(\partial_\psi \left(\left(\log \tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)\right)^2\right)\right) \tag{60}$$

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, (57) in ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$) implies (59) 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 (28) 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 (60), it follows from the dominated convergence theorem and

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

where the first inequality uses the Cauchy-Schwarz inequality and the second is a consequence of (57) and (58). □

We now consider the assumption used for the DREP gradient estimator.

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

$$\mathbb{E}\left(\sup_{(\theta,\phi')\in\mathcal{V}} \left(\partial_{\psi'} \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right)^2\right) < \infty \,, \tag{61}$$

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

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 Theorem 5.

Proposition 4 *Let ψ denote a component of the $\mathbb{R}^b$-valued variable $\phi = (\phi_1, \ldots, \phi_b)$ and $\varepsilon \sim q$. Under (B1) and ($\mathsf{B}_{\mathrm{df}}^{\mathrm{DREP}}$), we have that: for all $(\theta, \phi) \in \Theta \times \Phi$,*

$$[\partial_{\psi'} \mathbb{E}\left(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right)]\big|_{\phi'=\phi} = \mathbb{E}\left([\partial_{\psi'} \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)]\big|_{\phi'=\phi} \right) \tag{63}$$

$$\left[\partial_{\psi'} \mathbb{E}\left(\left(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right)^2\right)\right]\big|_{\phi'=\phi} = \mathbb{E}\left(\left[\partial_{\psi'} \left(\left(\log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon; x)\right)^2\right)\right]\big|_{\phi'=\phi} \right) \tag{64}$$

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

Proof. The proof is similar to that of Proposition 3 and is therefore omitted. □

B Deferred proofs and results for Section 3.1

B.1 Preliminary results for the proofs of Theorems 1 to 3

We first provide three lemmas.

Lemma 2 *Let $\eta_0 < \mu_0$ be two real exponents. Then, for all $a, b \geq 0$, we have*

$$|a-b|^{\mu_0} \leq |a-b|^{\eta_0} \left(a^{\mu_0-\eta_0} + b^{\mu_0-\eta_0}\right) .$$

Proof. Since both sides of the inequality are unchanged by permuting a and b, it is sufficient to show that it holds for all $0 \leq a \leq b$. We then indeed have

$$|a-b|^{\mu_0} = (b-a)^{\eta_0}\ (b-a)^{\mu_0-\eta_0} \leq (b-a)^{\eta_0}\ b^{\mu_0-\eta_0} \leq |a-b|^{\eta_0} \left(a^{\mu_0-\eta_0} + b^{\mu_0-\eta_0}\right) ,$$

where in the first inequality we used that $x \mapsto x^{\mu_0-\eta_0}$ is non-decreasing on $\mathbb{R}_+$ and in the second that $a^{\mu_0-\eta_0} \geq 0$. □

Lemma 3 *Let Z be a real valued random variable, let $Z_1, \ldots, Z_N$ be i.i.d. copies of Z and denote $\overline{Z}_N = N^{-1}\sum_{i=1}^N Z_i$ for all $N \in \mathbb{N}^\star$. Then, for all real $p \geq 1$,*

$$\mathbb{E}(|Z|^{2p}) < \infty \Longrightarrow \sup_{N\in\mathbb{N}^\star} \left(N^p\, \mathbb{E}\left(\left|\overline{Z}_N - \mathbb{E}\left(Z\right)\right|^{2p} \right) \right) < \infty , \tag{65}$$

$$\mathbb{E}(|Z|^{p}) < \infty \Longrightarrow \sup_{N\in\mathbb{N}^\star} \mathbb{E}\left(\left|\overline{Z}_N\right|^{p} \right) < \infty , \tag{66}$$

$$\mathbb{E}(|Z|^{p}) < \infty \Longrightarrow \lim_{N\to\infty} \mathbb{E}\left(\left|\overline{Z}_N - \mathbb{E}\left(Z\right)\right|^{p} \right) = 0 . \tag{67}$$

Proof of Lemma 3. Since $2p \geq 2$, (65) is a straightforward consequence of Petrov, 1995, Theorems 2.10 with $S_N = N(\overline{Z}_N - \mathbb{E}(Z))$. To prove (66), we apply the Minkowski inequality which yields the bound

$$\left\|\overline{Z}_N\right\|_p \leq \frac{1}{N}\sum_{i=1}^N \left\|Z_i\right\|_p = \left\|Z\right\|_p . \tag{68}$$

Let us now prove (67). For any $M > 0$ we set

$$V^{(\leq M)} = Z\,\mathbb{1}_{\{|Z|\leq M\}} - \mathbb{E}\left(Z\,\mathbb{1}_{\{|Z|\leq M\}}\right) \quad \text{and} \quad V^{(>M)} = Z\,\mathbb{1}_{\{|Z|>M\}} - \mathbb{E}\left(Z\,\mathbb{1}_{\{|Z|>M\}}\right)\,,$$

and we define similarly $V_i^{(\leq M)}$ and $V_i^{(>M)}$ for all $i \geq 1$, while $\overline{V_N^{(\leq M)}}$ and $\overline{V_N^{(>M)}}$ denote their N-sample means. Then we observe that, for any $M > 0$,

$$\left\|\overline{Z}_N - \mathbb{E}(Z)\right\|_p \leq \left\|\overline{V_N^{(\leq M)}}\right\|_p + \left\|\overline{V_N^{(>M)}}\right\|_p \leq \left\|\overline{V_N^{(\leq M)}}\right\|_{2\vee p} + \left\|\overline{V_N^{(>M)}}\right\|_p\,.$$

Applying (66) with $V^{(\leq M)}$ (which admits finite moments of all orders) replacing Z and $2 \vee p$ replacing $2p$ we get that the first term in this upper bound satisfies $\left\|\overline{V_N^{(\leq M)}}\right\|_{2\vee p} = O\left(N^{-1/2}\right) = o(1)$ as $N \to \infty$. Hence, to obtain (67), it only remains to show

$$\mathbb{E}(|Z|^p) < \infty \Longrightarrow \lim_{M\to\infty} \limsup_{N\to\infty} \left\|\overline{V_N^{(>M)}}\right\|_p = 0\,. \tag{69}$$

Using the same bound as in (68) but with $V_N^{(>M)}$ replacing Z, we have

$$\left\|\overline{V_N^{(>M)}}\right\|_p \leq \left\|Z\,\mathbb{1}_{\{|Z|>M\}} - \mathbb{E}\left(Z\,\mathbb{1}_{\{|Z|>M\}}\right)\right\|_p \leq \left\|Z\,\mathbb{1}_{\{|Z|>M\}}\right\|_p + \mathbb{E}\left(|Z|\,\mathbb{1}_{\{|Z|>M\}}\right)$$

This upper bound does not depend on N and converges to zero as $M \to \infty$ by dominated convergence if $\mathbb{E}(|Z|^p) < \infty$. Hence, we get (69). □

Lemma 4 *Let X and Y be two random variables. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be two sequences of i.i.d. random variables with the same distribution as X and Y respectively. Assume that $\mathbb{V}(X) < \infty$, $\mathbb{E}(Y) \neq 0$ and $\mathbb{V}(Y) < \infty$. For all $N \in \mathbb{N}^\star$, denote $\overline{X}_N = N^{-1}\sum_{i=1}^N X_i$ and $\overline{Y}_N = N^{-1}\sum_{i=1}^N Y_i$. Then, for all $\ell \in \mathbb{N}^\star$, as $N \to \infty$,*

$$\sqrt{N}\left(\frac{\overline{X}_N}{\left(\overline{Y}_N\right)^\ell} - \frac{\mathbb{E}(X)}{\mathbb{E}(Y)^\ell}\right) \xrightarrow{\text{dist.}} \mathcal{N}\left(0, \frac{1}{\mathbb{E}(Y)^{2\ell}}\mathbb{V}\left(X - \ell Y\frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\right)\right)\,. \tag{70}$$

Proof. Using the Central Limit Theorem, we have

$$\sqrt{N}\left(\begin{bmatrix}\overline{X}_N\\ \overline{Y}_N\end{bmatrix} - \begin{bmatrix}\mathbb{E}(X)\\ \mathbb{E}(Y)\end{bmatrix}\right) \xrightarrow{\text{dist.}} \mathcal{N}(0, \boldsymbol{\Sigma}) \quad \text{where} \quad \boldsymbol{\Sigma} = \begin{pmatrix}\mathbb{V}(X) & \mathbb{C}\text{ov}(X,Y)\\ \mathbb{C}\text{ov}(X,Y) & \mathbb{V}(Y)\end{pmatrix}\,.$$

Paired up with the Delta method, we obtain that

$$\sqrt{N}\left(\frac{\overline{X}_N}{\left(\overline{Y}_N\right)^\ell} - \frac{\mathbb{E}(X)}{\mathbb{E}(Y)^\ell}\right) \xrightarrow{\text{dist.}} \mathcal{N}\left(0, a\boldsymbol{\Sigma}a^T\right) \quad \text{where} \quad a^T = \begin{bmatrix}\frac{1}{\mathbb{E}(Y)^\ell} & \frac{-\ell\,\mathbb{E}(X)}{\mathbb{E}(Y)^{\ell+1}}\end{bmatrix}^T\,.$$

Eq. (70) then follows by rewriting the asymptotic variance as follows:

$$\begin{aligned} a\boldsymbol{\Sigma}a^T &= \frac{\mathbb{V}(X)}{\mathbb{E}(Y)^{2\ell}} - 2\ell\frac{\mathbb{E}(X)}{\mathbb{E}(Y)^{2\ell+1}}\mathbb{C}\text{ov}(X,Y) + \ell^2\frac{\mathbb{E}(X)^2}{\mathbb{E}(Y)^{2(\ell+1)}}\mathbb{V}(Y)\\ &= \frac{1}{\mathbb{E}(Y)^{2\ell}}\left(\mathbb{V}(X) - 2\ell\frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\mathbb{C}\text{ov}(X,Y) + \ell^2\frac{\mathbb{E}(X)^2}{\mathbb{E}(Y)^2}\mathbb{V}(Y)\right)\\ &= \frac{1}{\mathbb{E}(Y)^{2\ell}}\mathbb{V}\left(X - \ell Y\frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\right)\,, \end{aligned}$$

from which we deduce the desired result. □

We next present a proposition, which will be central to prove Theorems 1 to 3.

Proposition 5 *Set $\overline{Z}_N = N^{-1}\sum_{i=1}^N Z_i$ for all $N \in \mathbb{N}^\star$, where $Z_1, \ldots, Z_N$ are positive i.i.d. random variables. Then the following assertions hold.*

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

$$\limsup_{N\to\infty} \mathbb{E}\left((1/\overline{Z}_N)^\mu\right) < \infty \Longleftrightarrow \exists N \geq 1\ ,\ \mathbb{E}((1/\overline{Z}_N)^\mu) < \infty. \tag{71}$$

Furthermore, if the assertions of the equivalence (71) hold for some $\mu > 0$ and the distribution of Z_1 does not reduce to a Dirac measure, there exists $N_0 \geq 1$ such that $\mathbb{E}\left(\left(1/\overline{Z}_N\right)^\mu\right) = \infty$ for $1 \leq N < N_0$ and the sequence $\left(\mathbb{E}\left(\left(1/\overline{Z}_N\right)^\mu\right)\right)_{N\geq N_0}$ is strictly decreasing in $(0,\infty)$.

(ii) For all $\eta > 0$, we have

$$\sup_{t>0}\left(t^\eta\ \mathbb{E}\left(\mathrm{e}^{-t\,Z_1}\right)\right) < \infty \Longleftrightarrow \sup_{u>0}\left(u^{-\eta}\ \mathbb{P}\left(Z_1 \leq u\right)\right) < \infty. \tag{72}$$

(iii) Suppose that the assertions of the equivalence (72) hold for some $\eta > 0$. Then the assertions of the equivalence (71) hold for all $\mu > 0$.

(iv) Suppose that the assertions of the equivalence (71) hold for some $\mu > 1$. Then the assertions of the equivalence (72) hold for some $\eta > 0$.

(v) Suppose that the assertions of the equivalence (72) hold for some $\eta > 0$. Then, for all $\mu > 0$, we have

$$\lim_{N\to\infty} \mathbb{E}\left((1/\overline{Z}_N)^\mu\right) = \begin{cases} 0 & \text{if } \mathbb{E}(Z_1) = \infty, \\ (\mathbb{E}(Z_1))^{-\mu} & \text{otherwise.} \end{cases} \tag{73}$$

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

(i) **Proof of Assertion (i).** Let Γ denote the 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 $N \geq 1$,

$$\begin{aligned} \mathbb{E}\left((1/\overline{Z}_N)^\mu\right) &= (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\mathbb{E}\left(\mathrm{e}^{-\overline{Z}_N\,t}\right)\,\mathrm{d}t \\ &= (\Gamma(\mu))^{-1}\int_0^\infty t^{\mu-1}\left(\mathbb{E}\left(\mathrm{e}^{-Z_1\,t/N}\right)\right)^N\,\mathrm{d}t, \end{aligned} \tag{74}$$

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 $t > 0$,

$$\left(\mathbb{E}\left(\mathrm{e}^{-Z_1\,t/N}\right)\right)^N = \left(\mathbb{E}\left(\left(\mathrm{e}^{-Z_1\,t/(N+1)}\right)^{(N+1)/N}\right)\right)^N \geq \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\,t/(N+1)}\right)\right)^{N+1},$$

with equality if and only if both sides of the inequality are infinite or $\mathrm{e}^{-Z_1\,t/(N+1)}$ is equal to its mean a.s. and we can thus deduce Assertion (i).

(ii) **Proof of Assertion (ii).** Let $\eta > 0$. We first assume that

$$\sup_{t>0} \left(t^\eta \, \mathbb{E}\left(\mathrm{e}^{-t\, Z_1} \right) \right) < \infty \tag{75}$$

holds and show

$$\sup_{u>0} \left(u^{-\eta} \, \mathbb{P}\left(Z_1 \le u \right) \right) < \infty. \tag{76}$$

We have, for all $u > 0$, by Markov's inequality,

$$\mathbb{P}\left(Z_1 \le u \right) = \mathbb{P}\left(\mathrm{e}^{\eta(1 - Z_1/u)} \ge 1 \right) \le \mathrm{e}^\eta \, \mathbb{E}\left(\mathrm{e}^{-\eta\, Z_1/u} \right) \le \mathrm{e}^\eta \, C \, \eta^{-\eta} \, u^\eta,$$

where C is the sup in the left-hand side of (75). Hence (76) follows. Let us now assume (76) and show that (75) holds. We have, for all $t > 0$,

$$\mathbb{E}\left(\mathrm{e}^{-t\, Z_1} \right) = t \int_0^\infty \mathrm{e}^{-t\, u} \, \mathbb{P}\left(Z_1 \le u \right) \, \mathrm{d}u = \int_0^\infty \mathrm{e}^{-v} \, \mathbb{P}\left(Z_1 \le v/t \right) \, \mathrm{d}v \; .$$

Let C now denote the (finite) sup in the left-hand side of (76), so that $\mathbb{P}\left(Z_1 \le u \right) \le C \, u^\eta$ for all $u > 0$. Using this bound in the previous integral, we get that, for all $t > 0$,

$$\mathbb{E}\left(\mathrm{e}^{-t\, Z_1} \right) \le C \, t^{-\eta} \int_0^\infty \mathrm{e}^{-v} \, v^\eta \, \mathrm{d}v \; .$$

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

(iii) **Proof of Assertion (iii).** Let $\mu > 0$ and suppose that (75) holds. Then we have, for any $N \ge 1$, as $t \to \infty$,

$$t^{\mu-1} \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, t/N} \right) \right)^N = O\left(t^{\mu-1} (t/N)^{-\eta\, N} \right) = O\left(t^{\mu-1-\eta\, N} \right) \; .$$

Hence for N large enough, the right-hand side of (74) is finite and $\mathbb{E}\left((1/\overline{Z}_N)^\mu \right)$ as well. We obtain the right-hand side of the equivalence (71).

(iv) **Proof of Assertion (iv).** Let $\mu > 1$. Note that, since Z_1 is positive, for any $N \ge 1$, $t \mapsto \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, t/N} \right) \right)^N$ is decreasing as $t > 0$ is increasing. Hence, for all $k \ge 0$, the integral $\int_k^{k+1} t^{\mu-1} \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, t/N} \right) \right)^N \, \mathrm{d}t$ is bounded from below by $u_k = \int_k^{k+1} t^{\mu-1} \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, (k+1)/N} \right) \right)^N \, \mathrm{d}t$. Since $\int_k^{k+1} t^{\mu-1} \, \mathrm{d}t = (k+1)^\mu - k^\mu = k^\mu \left((1 + k^{-1})^\mu - 1 \right)$, we have, as $k \to \infty$,

$$u_k = \mu \, w_{k+1,N} \, k^{\mu-1} \left(1 + o(1) \right) \quad \text{where} \quad w_{k,N} = \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, k/N} \right) \right)^N \; .$$

As the equivalent assertions in (71) hold with $\mu > 1$, we have by (74) that: for some $N \ge 1$,

$$\sum_{k=0}^\infty u_k \le \int_0^\infty t^{\mu-1} \left(\mathbb{E}\left(\mathrm{e}^{-Z_1\, t/N} \right) \right)^N \, \mathrm{d}t < \infty \; .$$

Thus, as $k \to \infty$, $u_k \to 0$, which implies $w_{k+1,N} = o\left(k^{1-\mu} \right)$, hence $\mathbb{E}\left(\mathrm{e}^{-Z_1\, (k+1)/N} \right) = o\left(k^{(1-\mu)/N} \right)$. Now for all $t > 1$, taking $k = \lfloor Nt \rfloor - 1$, we get, as $t \to \infty$, $\mathbb{E}\left(\mathrm{e}^{-Z_1\, t} \right) \le \mathbb{E}\left(\mathrm{e}^{-Z_1\, (k+1)/N} \right) = o\left(\left([Nt] - 1 \right)^{(1-\mu)/N} \right) = o(t^{(1-\mu)/N})$ as $t \to \infty$. This implies (75).

(v) **Proof of Assertions (v).** By the strong law of large numbers we have $\overline{Z}_N \to \mathbb{E}(Z_1)$ as $N \to \infty$ a.s., including in the case where $\mathbb{E}(Z_1) = \infty$. Hence the sequence $\left((1/\overline{Z}_N)^\mu\right)_{N\geq 1}$ converges to the right-hand side of (73) a.s. To conclude the proof of Assertion (v), it suffices to show that, for all $\mu > 0$, the sequence $\left((1/\overline{Z}_N)^\mu\right)_{N\geq N_0}$ 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 $\epsilon > 0$,

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

In other words, the uniform integrability will follow from

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

with $\mu' = \mu(1+\epsilon) > \mu$. By Assertion (iii), the assertions of the equivalence (71) 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 (71).

□

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

Proposition 6 *Let (X, Y) be an $\mathbb{R} \times \mathbb{R}_+^*$-valued random vector. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a sequence of i.i.d. random vectors with the same distributions as (X, Y). Set $\overline{X}_N = N^{-1}\sum_{i=1}^N X_i$ and $\overline{Y}_N = N^{-1}\sum_{i=1}^N Y_i$ for all $N \in \mathbb{N}^\star$. Assume that there exist $N \geq 1$ and $\mu > 0$ such that $\mathbb{E}((\overline{Y}_N)^{-\mu}) < \infty$ and that $\mathbb{E}(Y^h)$ and $\mathbb{E}(|X|^{h'})$ are finite for some positive real exponents h and h'. Then the following assertions hold.*

(i) If $h \geq 1$ and $h' > 1$, we have, as $N \to \infty$,

$$\mathbb{E}\left(\frac{\overline{X}_N}{(\overline{Y}_N)^\ell}\right) = \frac{\mathbb{E}(X)}{(\mathbb{E}(Y))^\ell} + o(1), \quad \ell \in \{1, 2\}. \tag{77}$$

(ii) If $h > 2$, we have, as $N \to \infty$,

$$\mathbb{E}\left(\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^3}{\overline{Y}_N}\right) = o\left(\frac{1}{N}\right). \tag{78}$$

(iii) If $h > 2$ and $h' > 1$ with $2/h + 1/h' < 1$, we have, as $N \to \infty$,

$$\mathbb{E}\left(\frac{(\overline{Y}_N - \mathbb{E}(Y))^2\ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\overline{Y}_N}\right) = o\left(\frac{1}{N}\right). \tag{79}$$

Furthermore, if $h, h' \geq 2$ with $1/h + 1/h' < 1$, then (79) holds again as $N \to \infty$.

(iv) Suppose that $h > 2$ and $h' > 1$. If $h' < 2$, suppose moreover that $2/h + 1/h' < 1$. Then, we have, as $N \to \infty$,

$$\mathbb{E}\left(\frac{\overline{X}_N}{\overline{Y}_N}\right) = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} + \frac{1}{N}\left(\frac{\mathbb{E}(X)\,\mathbb{V}(Y)}{(\mathbb{E}(Y))^3} - \frac{\mathbb{C}\mathrm{ov}(X,Y)}{(\mathbb{E}(Y))^2}\right) + o\left(\frac{1}{N}\right) . \tag{80}$$

(v) If $h' > 2$, we have, as $N \to \infty$,

$$\mathbb{V}\left(\frac{\overline{X}_N}{(\overline{Y}_N)^2}\right) = O(1) . \tag{81}$$

(vi) If $h, h' > 2$, we have, as $N \to \infty$,

$$\mathbb{V}\left(\frac{\overline{X}_N}{(\overline{Y}_N)^\ell}\right) = \frac{1}{N}\frac{\mathbb{V}\left(X - \ell Y \frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\right)}{(\mathbb{E}(Y))^{2\ell}} + o\left(\frac{1}{N}\right), \quad \ell \in \{1,2\}. \tag{82}$$

Proof. In the following, we always suppose that $h, h' \geq 1$ so that $\mathbb{E}(Y)$ and $\mathbb{E}(X)$ are well defined. We now state some preliminary facts that will be useful to refer to when deriving the claimed assertions. Applying Lemma 3 yields

$$1 \leq h' \Longrightarrow \left\|\overline{X}_N - \mathbb{E}(X)\right\|_{h'} = o(1) \quad \text{as } N \to \infty , \tag{83}$$

$$2 \leq h' \Longrightarrow \left\|\overline{X}_N - \mathbb{E}(X)\right\|_{h'}^2 = O\left(\frac{1}{N}\right) \quad \text{as } N \to \infty , \tag{84}$$

$$1 \leq h \Longrightarrow \left\|\overline{Y}_N - \mathbb{E}(Y)\right\|_{h} = o(1) \quad \text{as } N \to \infty , \tag{85}$$

$$2 \leq h \Longrightarrow \left\|\overline{Y}_N - \mathbb{E}(Y)\right\|_{h}^2 = O\left(\frac{1}{N}\right) \quad \text{as } N \to \infty. \tag{86}$$

By (iv) and (v) in Proposition 5, for all $\mu > 0$ we have, as $N \to \infty$, $\mathbb{E}((\overline{Y}_N)^{-\mu}) = O(1)$. Therefore, using the Hölder inequality $\mathbb{E}\left((\overline{Y}_N)^{-\mu'}\,|V|^{h''}\right) \leq \left\|(\overline{Y}_N)^{-\mu'}\right\|_r \|V\|_{r'h''}^{h''}$ with $\mu' > 0$, $V = \overline{X}_N - \mathbb{E}(X)$ or $V = \overline{Y}_N - \mathbb{E}(Y)$ and $r, r' > 1$ such that $1/r + 1/r' = 1$ and $r'h'' = h$ or h', (83)–(86) yield, for any $\mu' > 0$,

$$1 \leq h' \Longrightarrow \forall h'' \in (0, h'),\ \mathbb{E}\left(\frac{\left|\overline{X}_N - \mathbb{E}(X)\right|^{h''}}{\left(\overline{Y}_N\right)^{\mu'}}\right) = o(1) \quad \text{as } N \to \infty , \tag{87}$$

$$2 \leq h' \Longrightarrow \forall h'' \in (0, h'),\ \mathbb{E}\left(\frac{\left|\overline{X}_N - \mathbb{E}(X)\right|^{h''}}{\left(\overline{Y}_N\right)^{\mu'}}\right) = O\left(\frac{1}{N^{h''/2}}\right) \quad \text{as } N \to \infty , \tag{88}$$

$$1 \leq h \Longrightarrow \forall h'' \in (0, h),\ \mathbb{E}\left(\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{\left(\overline{Y}_N\right)^{\mu'}}\right) = o(1) \quad \text{as } N \to \infty , \tag{89}$$

$$2 \leq h \Longrightarrow \forall h'' \in (0, h),\ \mathbb{E}\left(\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{\left(\overline{Y}_N\right)^{\mu'}}\right) = O\left(\frac{1}{N^{h''/2}}\right) \quad \text{as } N \to \infty. \tag{90}$$

(i) **Proof of Assertion (i).** We have, for all $N \geq 1$ and $\ell \in \{1,2\}$,

$$\frac{\overline{X}_N}{(\overline{Y}_N)^\ell} = \frac{\mathbb{E}(X)}{(\mathbb{E}(Y))^\ell} + \frac{\overline{X}_N - \mathbb{E}(X)}{(\overline{Y}_N)^\ell} - \frac{\mathbb{E}(X)}{(\mathbb{E}(Y))^\ell} \left(\frac{(\overline{Y}_N)^\ell - (\mathbb{E}(Y))^\ell}{(\overline{Y}_N)^\ell} \right) . \tag{91}$$

Observe then that the equality above can be rewritten as follows when $\ell \in \{1,2\}$:

$$\begin{aligned} &\frac{\overline{X}_N}{(\overline{Y}_N)^\ell} - \frac{\mathbb{E}(X)}{(\mathbb{E}(Y))^\ell} \\ &\qquad = \frac{\overline{X}_N - \mathbb{E}(X)}{(\overline{Y}_N)^\ell} - \frac{\mathbb{E}(X)}{(\mathbb{E}(Y))^\ell} \frac{\overline{Y}_N - \mathbb{E}(Y)}{\overline{Y}_N} - (\ell - 1) \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} \frac{\overline{Y}_N - \mathbb{E}(Y)}{(\overline{Y}_N)^\ell} . \end{aligned} \tag{92}$$

Therefore, in order to get (77), it suffices to have that, as $N \to \infty$,

$$\mathbb{E}\left(\frac{\overline{X}_N - \mathbb{E}(X)}{(\overline{Y}_N)^\ell} \right) = o(1) \quad \text{and} \quad \mathbb{E}\left(\frac{\overline{Y}_N - \mathbb{E}(Y)}{(\overline{Y}_N)^{\mu''}} \right) = o(1) \quad \text{with } \mu'' \in \{1, \ell\} . \tag{93}$$

Suppose now that $h \geq 1$ and $h' > 1$. Then the first assertion in (93) follows by applying (87) with $h'' = 1 \in (0, h')$ and $\mu' = \ell$. If $h > 1$ the second assertion follows by applying (89) with $h'' = 1 \in (0, h)$ and $\mu' = \mu''$. If $h = 1$, we pick any $h'' \in (0, h)$ and use Lemma 2 with $\mu_0 = 1$, $\eta_0 = h''$, $a = \overline{Y}_N$ and $b = \mathbb{E}(Y)$ which yields: for $\mu'' \in \{1, \ell\}$ and all $N \geq 1$,

$$\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|}{(\overline{Y}_N)^{\mu''}} \leq \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{(\overline{Y}_N)^{\mu''+h''-1}} + (\mathbb{E}(Y))^{1-h''} \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{(\overline{Y}_N)^{\mu''}}$$

and the second assertion in (93) follows from (89) with $\mu' = \mu'' + \eta_0 - 1$ or $\mu' = \mu''$.

(ii) **Proof of Assertion (ii).** If $h > 3$, then $h'' = 3$ and $\mu' = 1$ in (90) directly yields (78). Suppose from now on that $2 < h \leq 3$. Take any $2 < h'' < h$ to be chosen later. Applying Lemma 2 with $\mu_0 = 3$, $\eta_0 = h''$, $a = \overline{Y}_N$ and $b = \mathbb{E}(Y)$, we have: for all $N \geq 1$,

$$\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^3}{\overline{Y}_N} \leq \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{\overline{Y}_N^{h''-2}} + (\mathbb{E}(Y))^{3-h''} \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h''}}{\overline{Y}_N}$$

and (78) follows from applying (90) with $\mu' = h'' - 2$ and $\mu' = 1$.

(iii) **Proof of Assertion (iii).** Suppose that $h > 2$ and $h' > 1$ with $2/h + 1/h' < 1$. Let $r' > 1$ be such that $1/h' + 1/r' = 1$. By the Hölder inequality, we have

$$\mathbb{E}\left(\frac{(\overline{Y}_N - \mathbb{E}(Y))^2 \left|\overline{X}_N - \mathbb{E}(X)\right|}{\overline{Y}_N} \right) \leq \left\| \frac{(\overline{Y}_N - \mathbb{E}(Y))^2}{\overline{Y}_N} \right\|_{r'} \left\| \overline{X}_N - \mathbb{E}(X) \right\|_{h'}$$

and since the condition $2/h + 1/h' < 1$ implies $2r' < h$, we then get (79) by applying (90) with $h'' = 2r' < h$ and $\mu' = 1$ and (83).

Suppose now that $h, h' \geq 2$ with $1/h + 1/h' < 1$. Take $1 < h_1 < 2$ close enough to 1 to have

$$h_1/h + 1/h' < 1 . \tag{94}$$

Applying Lemma 2 with $\mu_0 = 2$, $\eta_0 = h_1$, $a = \overline{Y}_N$ and $b = \mathbb{E}(Y)$, we have: for all $N \geq 1$,

$$\frac{(\overline{Y}_N - \mathbb{E}(Y))^2 \ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\overline{Y}_N} \leq \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h_1} \ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\left(\overline{Y}_N\right)^{h_1-1}} + (\mathbb{E}(Y))^{2-h_1} \ \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h_1} \ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\overline{Y}_N} .$$

Now, let $r' > 1$ be once more such that $1/h' + 1/r' = 1$, which with (94) implies $r'h_1 < h$, and, by the Hölder inequality, we have, for any $\mu' > 0$,

$$\mathbb{E}\left(\frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h_1} \ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\left(\overline{Y}_N\right)^{\mu'}}\right) \leq \left\| \frac{\left|\overline{Y}_N - \mathbb{E}(Y)\right|^{h_1}}{\left(\overline{Y}_N\right)^{\mu'}} \right\|_{r'} \left\|\overline{X}_N - \mathbb{E}(X)\right\|_{h'} .$$

Since $h, h' \geq 2$, we can apply (90) with $h'' = r'h_1 < h$ and $\mu' = h_1 - 1$ or $\mu' = 1$. Combining this with (84) and the two previous bounds, gives that: as $N \to \infty$,

$$\mathbb{E}\left(\frac{\left(\overline{Y}_N - \mathbb{E}(Y)\right)^2 \ \left|\overline{X}_N - \mathbb{E}(X)\right|}{\overline{Y}_N}\right) = O\left(\frac{1}{N^{h_1/2+1/2}}\right) .$$

Since $h_1 > 1$, we get (79).

(iv) **Proof of Assertion (iv).** This assertion is a consequence of Assertion (ii) paired up with the following identity, valid for all $N \geq 1$,

$$\mathbb{E}\left(\frac{\overline{X}_N}{\overline{Y}_N}\right) = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} + \frac{1}{N}\left(\frac{\mathbb{E}(X)\,\mathbb{V}(Y)}{(\mathbb{E}(Y))^3} - \frac{\mathbb{C}\mathrm{ov}(X,Y)}{(\mathbb{E}(Y))^2}\right) + \mathbb{E}\left(\frac{\left(\overline{X}_N - \mathbb{E}(X)\right)\left(\overline{Y}_N - \mathbb{E}(Y)\right)^2}{\overline{Y}_N\,(\mathbb{E}(Y))^2} - \frac{\mathbb{E}(X)}{\mathbb{E}(Y)}\frac{\left(\overline{Y}_N - \mathbb{E}(Y)\right)^3}{\overline{Y}_N\,(\mathbb{E}(Y))^2}\right) . \quad (95)$$

Indeed, under the assumptions of Assertion (iv), we can apply Assertions (ii) and (iii), showing that the expectation in the second line of (95) is $o(1/N)$ as $N \to \infty$ (notice in particular that the condition $h > 2$ with $h' \geq 2$ implies $1/h + 1/h' < 1$). To conclude we thus only need to show (95). This follows from observing that

$$\frac{1}{\overline{Y}_N} = \frac{1}{\mathbb{E}(Y)} - \frac{\overline{Y}_N - \mathbb{E}(Y)}{(\mathbb{E}(Y))^2} + \frac{\left(\overline{Y}_N - \mathbb{E}(Y)\right)^2}{\overline{Y}_N\,(\mathbb{E}(Y))^2},$$

plugging this in the r.h.s. of (91) when $\ell = 1$ and then taking the expectation, with the N^{-1} term in the first line of (95) coming from the fact that, for all $N \geq 1$,

$$\mathbb{E}\left(\left(\overline{Y}_N - \mathbb{E}(Y)\right)^2\right) = \frac{1}{N}\mathbb{V}(Y) \quad \text{and} \quad \mathbb{E}\left(\left(\overline{X}_N - \mathbb{E}(X)\right)\left(\overline{Y}_N - \mathbb{E}(Y)\right)\right) = \frac{1}{N}\mathbb{C}\mathrm{ov}(X,Y) .$$

(v) **Proof of Assertion (v).** We have, for all $N \geq 1$,

$$\frac{\overline{X}_N}{\left(\overline{Y}_N\right)^2} = \frac{\overline{X}_N - \mathbb{E}(X)}{\left(\overline{Y}_N\right)^2} + \frac{\mathbb{E}(X)}{\left(\overline{Y}_N\right)^2} .$$

If $h' > 2$, we can take $h'' = 2$ and $\mu' = 4$ in (88) and get that the first term in the right-hand side has its second moment, hence its variance, bounded by $O(N^{-1})$. The second term has a bounded variance as we already saw that by Lemma 1, for all $\mu > 0$ we have: as $N \to \infty$, $\mathbb{E}\left((\overline{Y}_N)^{-\mu}\right) = O(1)$. This yields (81).

(vi) **Proof of Assertion (vi).** Suppose that $h, h' > 2$. Applying Lemma 4, we have that for $\ell = 1, 2$, as $N \to \infty$, $\sqrt{N}\left((\overline{Y}_N)^{-\ell}\overline{X}_N - (\mathbb{E}(Y))^{-\ell}\,\mathbb{E}(X)\right)$ converges in distribution to a random variable whose variance is the constant appearing in the leading asymptotic term of (82). Hence to obtain (82), it suffices to show that we do not only have the convergence in distribution but also the convergence of the variance to the variance of the limit. This will be true if we have the convergence of the moments of order 1 and 2 to those of the limit. By Theorem 3.5 of Billingsley, 1971, we thus only need to show that the sequence $\left(N\left((\overline{Y}_N)^{-\ell}\overline{X}_N - (\mathbb{E}(Y))^{-\ell}\,\mathbb{E}(X)\right)^2\right)_{N\geq 1}$ is uniformly integrable (as it will also imply uniform integrability without the square). By the Markov inequality, this follows if we find $h'' > 2$ such that, as $N \to \infty$,

$$\left\| \frac{\overline{X}_N}{\left(\overline{Y}_N\right)^{\ell}} - \frac{\mathbb{E}(X)}{\left(\mathbb{E}(Y)\right)^{\ell}} \right\|_{h''} = O\left(\frac{1}{N^{1/2}}\right) . \tag{96}$$

Using the identity (92), this is implied by

$$\left\| \frac{\overline{X}_N - \mathbb{E}(X)}{(\overline{Y}_N)^{\ell}} \right\|_{h''} = O\left(\frac{1}{N^{1/2}}\right) \quad \text{and} \quad \left\| \frac{\overline{Y}_N - \mathbb{E}(Y)}{(\overline{Y}_N)^{\mu''}} \right\|_{h''} = O\left(\frac{1}{N^{1/2}}\right) \quad \text{with } \mu'' \in \{1, \ell\} .$$

Since $h, h' > 2$, we can (i) apply (88) with $h'' > 2$ and $\mu' = h''\ell$ and (ii) apply (90) with $h'' > 2$ and $\mu' = h''\mu''$. We then obtain the two previous asymptotic bounds, which concludes the proof of Assertion (vi).

□

B.2 Proofs of Theorems 1 to 3

B.2.1 Notation

Let us first introduce some notation and resulting identities that will be used for the proofs of Theorems 1 to 3. Letting $\varepsilon \sim q$ and with $\tilde{w}_{\theta,\phi,\phi}(\varepsilon; x)$ as in (17), we define

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

Furthermore, let $\varepsilon_1, \ldots, \varepsilon_N$ be i.i.d. copies of ε, and define X_i and Y_i for all $i = 1, 2, \ldots$ as X and Y but with ε_i replacing ε. Further denote $\overline{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\overline{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$ for all $N \in \mathbb{N}^*$. Then, using the above definitions in (12), we have

$$\mathbb{E}\left(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)\right) = \mathbb{E}\left(\frac{\overline{X}_N}{\overline{Y}_N}\right) , \tag{97}$$

$$\mathbb{V}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \frac{1}{M}\mathbb{V}\left(\frac{\overline{X}_N}{\overline{Y}_N}\right) . \tag{98}$$

Also note that with these definitions, (A3) means that there exists $\mu > 1$ and $N \geq 1$ such that $\mathbb{E}\left((\overline{Y}_N)^{-\mu}\right) < \infty$ and (A_h) that $\mathbb{E}\left(|Y|^h\right)$ and $\mathbb{E}\left(|X|^{h'}\right)$ are finite.

B.2.2 Proof of Theorem 1

The proof consists of two steps. In the first step, we show an asymptotic result for the expectation of the REP gradient estimator of the VR-IWAE bound as $N \to \infty$ which holds under (A3), (A_h) and $(\mathsf{A}_{h'}^{\mathrm{REP}})$ with h, h' as in Theorem 1. In the second step, we use the interchanges of the derivative and expectation signs obtained in Proposition 1 under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ to rewrite the asymptotic result obtained in the first step into the convenient formulation (18).

First step. Let us show that: as $N \to \infty$,

$$\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) = \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \frac{1}{2N}\psi\text{-}\mathrm{G}_2^{(\alpha,\mathrm{REP})}(\theta,\phi;x) + o\left(\frac{1}{N}\right) , \tag{99}$$

where

$$\psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} , \tag{100}$$

$$\psi\text{-}\mathrm{G}_2^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = 2\left(\frac{\mathbb{C}\mathrm{ov}\,(X,Y)}{(\mathbb{E}(Y))^2} - \frac{\mathbb{E}(X)\,\mathbb{V}\,(Y)}{(\mathbb{E}(Y))^3}\right) . \tag{101}$$

By (A3), (A_h) and $(\mathsf{A}_{h'}^{\mathrm{REP}})$ with h, h' as in Theorem 1, we can apply Assertion (iv) in Proposition 6. It thus only remains to identify (80) with (99). This follows from (97) and the definitions (100) and (101).

Second step. Under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, the conclusions of Proposition 1 imply that

$$\partial_\psi \ell_N^{(\alpha)}(\theta,\phi;x) = \mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)) \tag{102}$$

$$\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{REP})}(\theta,\phi;x) \tag{103}$$

$$\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2] = \psi\text{-}\mathrm{G}_2^{(\alpha,\mathrm{REP})}(\theta,\phi;x) . \tag{104}$$

More precisely, using the definitions of Appendix B.2.1, (6), (7), (9) and (44)–(46), we get

$$\ell_N^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\,\mathbb{E}\left(\log \overline{Y}_N\right) ,$$
$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\,\log \mathbb{E}\,(Y) ,$$
$$\gamma^{(\alpha)}(\theta,\phi;x)^2 = \frac{1}{1-\alpha}\,\frac{\mathbb{V}(Y)}{\mathbb{E}(Y)^2} ,$$

as well as

$$\partial_\psi \mathbb{E}(Y) = (1-\alpha)\,\mathbb{E}(X) , \tag{105}$$
$$\partial_\psi \mathbb{E}\left(Y^2\right) = 2\,(1-\alpha)\,\mathbb{E}\,(X\,Y) ,$$
$$\partial_\psi \mathbb{E}\left(\log \overline{Y}_n\right) = (1-\alpha)\,\mathbb{E}\left(\frac{\overline{X}_n}{\overline{Y}_n}\right) .$$

and pairing this with (97), (100) sand (101), we deduce (102)–(104).

B.2.3 Proof of Theorem 2

Note that the conditions of Theorem 2 on h, h' are stronger than those required in Theorem 1. We thus already have that (99) holds as $N \to \infty$, which as seen in the proof of Theorem 2 gives under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ that $|\mathbb{E}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))|$ behaves as the numerator of the r.h.s of (19). 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 1, this is done in two steps: (i) we show a first asymptotic behavior for $\mathbb{V}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{REP})}(\theta,\phi;x))$ as $N \to \infty$ which holds under (A3) and (A_h)–$(\mathsf{A}_{h'}^{\mathrm{REP}})$ with $h, h' > 2$ and (ii) we show how the conclusions of Proposition 1 allow us to express this asymptotic result under the convenient formulation (20).

First step. By (A3), (A_h) and $(\mathsf{A}_{h'}^{\mathrm{REP}})$ with $h, h' > 2$, we can apply Assertion (vi) of Proposition 6 with $\ell = 1$. Using (82) and (98) we have that, as $N \to \infty$,

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

Second step. We prove that, under $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$, the constant defined in (20) can be written as

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{V}\left(\frac{X}{\mathbb{E}(Y)} - Y\frac{\mathbb{E}(X)}{\mathbb{E}(Y)^2}\right). \tag{106}$$

By using the definitions of X and Y in Appendix B.2.1 and (44) of Proposition 1, we have

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

meaning that the two variables in the variances of the right-hand sides of (20) and (106) coincide up to the $(1-\alpha)$ multiplicative term, which is squared when put in front of the variance. The identity (106) follows.

Remark 3 *As a byproduct of the proof of Appendix B.2.3, we have established* (106)*, that is using the definition of X and Y in Appendix B.2.1,*

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

Now considering the case $p_\theta(\cdot|x) = q_\phi(\cdot|x)$ in the above and using (16)*, it follows that*

$$\psi\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{V}\left([\partial_{\psi'}\log q_{\psi'}(f(\varepsilon,\phi;x)|x)]|_{\psi'=\phi}\right) \quad \textit{when } p_\theta(\cdot|x) = q_\phi(\cdot|x). \tag{107}$$

B.2.4 Proof of Theorem 3

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

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

As for X_i and Y_i in Appendix B.2.1, 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 (14), we have

$$\mathbb{E}\left(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = \mathbb{E}\left(\alpha\,\frac{\overline{X}_N^{(1)}}{\overline{Y}_N} + \frac{1-\alpha}{N}\,\frac{\overline{X}_N^{(2)}}{(\overline{Y}_N)^2}\right) \tag{108}$$

$$\mathbb{V}(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = \frac{1}{M}\mathbb{V}\left(\alpha\,\frac{\overline{X}_N^{(1)}}{\overline{Y}_N} + \frac{1-\alpha}{N}\,\frac{\overline{X}_N^{(2)}}{(\overline{Y}_N)^2}\right) \tag{109}$$

and we further define

$$\psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \alpha\,\frac{\mathbb{E}\left(X^{(1)}\right)}{\mathbb{E}(Y)}\,, \tag{110}$$

$$\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) = \frac{\mathbb{E}\left(X^{(2)}\right)}{\left(\mathbb{E}(Y)\right)^2} \tag{111}$$

as well as

$$\psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \alpha^2\,\mathbb{V}\left(\frac{X^{(1)}}{\mathbb{E}(Y)} - Y\frac{\mathbb{E}(X^{(1)})}{\mathbb{E}(Y)^2}\right) \tag{112}$$

$$\text{and}\quad \psi\text{-}V_2^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = (1-\alpha)^2\,\mathbb{V}\left(\frac{X^{(2)}}{\mathbb{E}(Y)^2} - 2Y\frac{\mathbb{E}(X^{(2)})}{\mathbb{E}(Y)^3}\right)\,. \tag{113}$$

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

$$\mathrm{SNR}[\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)] = \sqrt{MN}\,\frac{\left|\psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right| + o(1)}{\sqrt{\psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)} + o(1)}\,, \tag{114}$$

while under the conditions of (ii) in Theorem 3: as $N \to \infty$,

$$\mathrm{SNR}\left[\psi\text{-}g_{M,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)\right] = \sqrt{MN}\,\frac{\left|\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x)\right| + o\left(1\right)}{\sqrt{\psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x)} + o(1)}\,. \tag{115}$$

In the second step we use the conclusions of Proposition 2 to rewrite (114) and (115) under the convenient formulations (22) and (24).

First step. Using the definitions above for $X^{(1)}$ and $X^{(2)}$ paired up with the notation from Appendix B.2.1, (A_h) means that $\mathbb{E}\left(|Y|^h\right) < \infty$, ($\mathsf{A}^{\mathrm{DREP}}_{h'}$) that $\mathbb{E}\left(|X^{(1)}|^{h'}\right) < \infty$ and ($\widetilde{\mathsf{A}}^{\mathrm{DREP}}_{\tilde{h}}$) that $\mathbb{E}\left(|X^{(2)}|^{\tilde{h}}\right) < \infty$. Furthermore, Assertions (i) and (v)–(vi) from Proposition 6 with $X^{(1)}$ or $X^{(2)}$ replacing X and $\ell = 1$ or $\ell = 2$ read:

$$h \geq 1\,,\ h' > 1 \Longrightarrow \mathbb{E}\left(\frac{\overline{X}_N^{(1)}}{\overline{Y}_N}\right) = \frac{\mathbb{E}\left(X^{(1)}\right)}{\mathbb{E}(Y)} + o\left(1\right) \quad \text{as } N \to \infty, \tag{116}$$

$$h \geq 1\,,\ \tilde{h} > 1 \Longrightarrow \mathbb{E}\left(\frac{\overline{X}_N^{(2)}}{\left(\overline{Y}_N\right)^2}\right) = \frac{\mathbb{E}\left(X^{(2)}\right)}{\left(\mathbb{E}(Y)\right)^2} + o\left(1\right) \quad \text{as } N \to \infty, \tag{117}$$

$$\tilde{h} > 2 \Longrightarrow \mathbb{V}\left(\frac{\overline{X}_N^{(2)}}{\overline{Y}_N}\right) = O(1) \quad \text{as } N \to \infty, \tag{118}$$

$$h, h' > 2 \Longrightarrow \mathbb{V}\left(\frac{\overline{X}_N^{(1)}}{\overline{Y}_N}\right) = \frac{1}{N}\frac{\mathbb{V}\left(X^{(1)} - Y\frac{\mathbb{E}(X^{(1)})}{\mathbb{E}(Y)}\right)}{\left(\mathbb{E}(Y)\right)^2} + o\left(\frac{1}{N}\right) \quad \text{as } N \to \infty, \tag{119}$$

$$h, \tilde{h} > 2 \Longrightarrow \mathbb{V}\left(\frac{\overline{X}_N^{(2)}}{\left(\overline{Y}_N\right)^2}\right) = \frac{1}{N}\frac{\mathbb{V}\left(X^{(2)} - 2\,Y\frac{\mathbb{E}(X^{(2)})}{\mathbb{E}(Y)}\right)}{\left(\mathbb{E}(Y)\right)^4} + o\left(\frac{1}{N}\right) \quad \text{as } N \to \infty, \tag{120}$$

From (108), (116) and (117), using the quantities defined in (110) and (111) we get that if $h \geq 1$ and $h', \tilde{h} > 1$: as $N \to \infty$,

$$\mathbb{E}\left(\psi\text{-}g_{M,N}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = \begin{cases} \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) + o(1) & \text{if } 0 < \alpha < 1, \\ \frac{1}{N}\,\psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) + o\left(\frac{1}{N}\right) & \text{if } \alpha = 0. \end{cases} \tag{121}$$

If $\alpha \in (0,1)$ and $h, h', \tilde{h} > 2$, pairing (118) and (119) up, we get that: as $N \to \infty$,

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

Combining this with (109) and (121) yields (114). Finally, if $\alpha = 0$ and $h, \tilde{h} > 2$, (109), (113) and (120) in the case $\alpha = 0$ imply that

$$\mathbb{V}\left(\psi\text{-}g_0^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = \frac{\psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x) + o(1)}{M\,N^3},$$

which, combined with (121) yields (115).

Second step. Under ($\mathsf{A}^{\mathrm{DREP}}_{\mathrm{df}}$), the conclusions of Proposition 2 hold and we can show that:

(i) If $\alpha \in (0,1)$, (114) is the same as (22), that is, we have:

$$\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \psi\text{-}\mathrm{G}_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\,, \tag{122}$$

$$\psi\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = \psi\text{-}V_1^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\,. \tag{123}$$

Indeed by definitions (7) and (110), (122) reads

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

which holds by (54). As for (123), by (23) and (112), it easily follows by differentiating inside the variance in (23) and using (53).

(ii) If $\alpha = 0$, (115) is the same as (24), that is, we have:

$$\partial_\psi\left(\gamma^{(0)}(\theta,\phi;x)^2\right) = -2\ \psi\text{-}\mathrm{G}_2^{(0,\mathrm{DREP})}(\theta,\phi;x) , \tag{124}$$

$$\psi\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = \psi\text{-}V_2^{(0,\mathrm{DREP})}(\theta,\phi;x) . \tag{125}$$

Indeed, by (9), with $\alpha = 0$, we have $\gamma^{(0)}(\theta,\phi;x)^2 = \left(\frac{\mathbb{E}(Y^2)}{(\mathbb{E}(Y))^2} - 1\right)$. Observe moreover that since $\alpha = 0$, we have $\mathbb{E}(Y) = \int p_\theta(x,z)\ \nu(\mathrm{d}z)$ and thus $\mathbb{E}(Y)$ does not depend on ϕ and therefore not on ψ. We thus get that $\partial_\psi\left(\gamma^{(0)}(\theta,\phi;x)^2\right) = \partial_\psi\mathbb{E}(Y^2)/\left(\mathbb{E}(Y)\right)^2$. By definition of $X^{(2)}$ and (111), (124) then follows from (55). As for (125), it simply follows by identifying (25) with (113) using $\alpha = 0$ and the definitions of Y and $X^{(2)}$.

B.3 Proof of examples

Before providing the proofs for Examples 3 and 4, we review the assumptions that need to be checked in order to obtain these examples.

B.3.1 Checking assumptions

First note that (i) the assumption (A1) holds with ν as the Lebesgue measure on $\mathbb{R}^d$ for both Examples 1 and 2 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 1 ($p_\theta(x)$ is arbitrary in Example 1 since only $p_\theta(z|x)$ is given) and $\Theta = \mathbb{R}^d$ and $\Phi = \mathbb{R}^{2d}$ for Example 2 (here $(\phi_1,\ldots,\phi_{2d}) = \left(\tilde{a}_1,\ldots,\tilde{a}_d,\tilde{b}_1,\ldots,\tilde{b}_d\right)$ since A is diagonal). Examples 3 and 4 build on the settings described in Examples 1 and 2 respectively, hence these assumptions also hold for Examples 3 and 4. In addition, (A2) holds with the reparametrization proposed in Example 3 (that is, $q = \mathcal{N}(0, \boldsymbol{I}_d)$ and $f(\varepsilon,\phi;x) = \varepsilon + \phi$) and the one proposed in Example 4 (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), (A_h), ($\mathsf{A}_{h'}^{\mathrm{REP}}$) ($\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}}$), ($\mathsf{A}_{h'}^{\mathrm{DREP}}$), ($\widetilde{\mathsf{A}}_{\tilde{h}}^{\mathrm{DREP}}$) and ($\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}}$).

• **Checking (A3).** By Lemma 1, it is equivalent to check (26), which by setting $v = -\log(u)$ is equivalent to finding $\eta > 0$ such that

$$\mathbb{P}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x) \le e^{-v}\right) = O(\mathrm{e}^{-\eta v}) \quad \text{as } v \to \infty.$$

Since by the Markov inequality, we have: for any $\eta > 0$,

$$\begin{aligned}\mathbb{P}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x) \le e^{-v}\right) &= \mathbb{P}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{-\eta}\mathrm{e}^{-\eta v} \ge 1\right)\\ &\le \mathrm{e}^{-\eta v}\ \mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{-\eta}\right)\end{aligned}$$

to check (A3) it is sufficient to prove that

$$\text{there exists } \eta > 0 \text{ such that } \quad \mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)^{-\eta}\right) < \infty. \tag{126}$$

• **Checking (A_h), $(\mathsf{A}_{h'}^{\mathrm{REP}})$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$.** Using the Hölder inequality, (A_h), $(\mathsf{A}_{h'}^{\mathrm{REP}})$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{REP}})$ hold for all $h, h' > 0$ 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

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta',\phi')\in\mathcal{V}} \tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)^p\right) < \infty \tag{127}$$

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta',\phi')\in\mathcal{V}} |\partial_{\psi'} \log \tilde{w}_{\theta',\phi',\phi'}(\varepsilon;x)|^p\right) < \infty, \tag{128}$$

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

• **Checking $(\mathsf{A}_{h'}^{\mathrm{DREP}})$, $(\widetilde{\mathsf{A}}_{\tilde{h}}^{\mathrm{DREP}})$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$.** Using the Hölder inequality once more, $(\mathsf{A}_{h'}^{\mathrm{DREP}})$, $(\widetilde{\mathsf{A}}_{\tilde{h}}^{\mathrm{DREP}})$ and $(\mathsf{A}_{\mathrm{df}}^{\mathrm{DREP}})$ hold for all $h', \tilde{h} > 0$ 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 (127) and (128) hold and, in addition,

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}} \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^p\right) < \infty \tag{129}$$

$$\mathbb{E}_{\varepsilon\sim q}\left(\sup_{(\theta,\phi')\in\mathcal{V}} |\partial_{\psi'} \log \tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)|^p\right) < \infty \tag{130}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}} w_{\theta,\phi'}(Z;x)^p\right) < \infty \tag{131}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}} \left(\frac{q_{\phi'}(Z|x)}{q_\phi(Z|x)}\right)^p\right) < \infty \tag{132}$$

$$\mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(\sup_{(\theta,\phi')\in\mathcal{V}} |\partial_\psi \log q_{\phi'}(Z|x)|^p\right) < \infty, \tag{133}$$

where ψ is a given component of ϕ.

When proving Examples 3 and 4, we will then notably show that the conditions (126)-(133) are satisfied, so that all the theorems of Section 3.1 apply to the settings described in Examples 1 and 2 with the reparameterizations of Examples 3 and 4.

B.3.2 Proof of Example 3

Recall that in this setting ψ is a component of ϕ, say ϕ_k for some 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), (A_h), ($A^{\mathrm{REP}}_{h'}$) ($A^{\mathrm{REP}}_{\mathrm{df}}$), ($A^{\mathrm{DREP}}_{h'}$), ($\widetilde{A}^{\mathrm{DREP}}_{\tilde{h}}$) and ($A^{\mathrm{DREP}}_{\mathrm{df}}$).** Let us prove (126)-(133). To this end, observe that

$$\log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) = -\frac{\|\theta-\phi\|^2}{2} - \langle z-\phi, \phi-\theta\rangle, \quad z\in\mathbb{R}^d. \tag{134}$$

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 (126) holds. We further get from (134) 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 (127)–(133) follow using that ε is a Gaussian variable.

• **Closed-form expressions for the quantities appearing in Theorems 1 to 3.** First recall that

$$\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = \ell(\theta;x) - \frac{\alpha\|\theta-\phi\|^2}{2}, \quad \gamma^{(\alpha)}(\theta,\phi;x)^2 = \frac{\exp\left((1-\alpha)^2\|\theta-\phi\|^2\right)-1}{1-\alpha}. \tag{135}$$

Computing $\partial_\psi \mathrm{VR}^{(\alpha)}(\theta,\phi;x)$ and $\partial_\psi[\gamma^{(\alpha)}(\theta,\phi;x)^2]$ follows directly from differentiating in (135) w.r.t. ϕ_k:

$$\partial_{\phi_k}\mathrm{VR}^{(\alpha)}(\theta,\phi;x) = -\alpha(\phi_k-\theta_k) \tag{136}$$

$$\partial_{\phi_k}[\gamma^{(\alpha)}(\theta,\phi;x)^2] = 2(1-\alpha)(\phi_k-\theta_k)\exp\left((1-\alpha)^2\|\phi-\theta\|^2\right). \tag{137}$$

Furthermore, using (134): for all $y>0$ and all $\phi'\in\mathbb{R}^d$,

$$\begin{aligned}\log\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^y\right) &= -\frac{y\|\theta-\phi\|^2}{2} - y\langle\varepsilon+\phi'-\phi,\phi-\theta\rangle \\ &= yg(\theta,\phi,\phi') - y\langle\varepsilon,\phi-\theta\rangle,\end{aligned} \tag{138}$$

where $g(\theta,\phi,\phi') = -\frac{\|\theta-\phi\|^2}{2} - y\langle\phi'-\phi,\phi-\theta\rangle$. Setting $y=1-\alpha$ in (138), we deduce that:

$$\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})} = \frac{\exp\left(-(1-\alpha)\langle\varepsilon,\phi-\theta\rangle\right)}{\mathbb{E}(\exp\left(-(1-\alpha)\langle\varepsilon,\phi-\theta\rangle\right))} = \frac{\exp\left(-(1-\alpha)\|\theta-\phi\|\tilde{S}\right)}{\exp\left(\frac{1}{2}(1-\alpha)^2\|\theta-\phi\|^2\right)} \tag{139}$$

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

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

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

$$\begin{aligned}\phi_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) &= \mathbb{E}\left(\left(\varepsilon^{(k)} + (1-\alpha)(\phi_k-\theta_k)\right)^2\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^2\right) \\ \phi_k\text{-}V^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) &= 0, \quad \alpha\in(0,1).\end{aligned} \tag{140}$$

Thus, using (139), we have that: for all $\psi \in \{\theta_1, \dots, \theta_d, \phi_1, \dots, \phi_d\}$,

$$\begin{aligned}\phi_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) &= e^{-(1-\alpha)^2\|\theta-\phi\|^2}\mathbb{E}\left((\varepsilon^{(k)}+(1-\alpha)(\phi_k-\theta_k))^2 e^{-2(1-\alpha)\langle\varepsilon,\phi-\theta\rangle}\right)\\ &= e^{-(1-\alpha)^2\|\theta-\phi\|^2} e^{2(1-\alpha)^2\|\theta-\phi'\|^2}\left(1+(1-\alpha)^2(\phi_k-\theta_k)^2\right)\\ &= e^{(1-\alpha)^2\|\theta-\phi\|^2}\left(1+(1-\alpha)^2(\phi_k-\theta_k)^2\right). \end{aligned} \tag{141}$$

Now considering the special case $\alpha = 0$, it holds that using (16):

$$\begin{aligned}\phi_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) &= \mathbb{V}\left(A_{\theta,\phi}(\varepsilon;x) - 2\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\mathbb{E}\left(A_{\theta,\phi}(\varepsilon;x)\right)\right)\\ &= \mathbb{E}\left(\left(A_{\theta,\phi}(\varepsilon;x) - 2\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\mathbb{E}\left(A_{\theta,\phi}(\varepsilon;x)\right)\right)^2\right) - \mathbb{E}(A_{\theta,\phi}(\varepsilon;x))^2 \end{aligned} \tag{142}$$

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

$$A_{\theta,\phi}(\varepsilon;x) = \frac{1}{2}\partial_{\phi_k}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^2\right) + \frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^2}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))^2}\left[\partial_{\phi'_k}\log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\big|_{\phi'=\phi}. \tag{143}$$

Using (139) 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, $\left[\partial_{\phi'_k}\log q_{\phi'}(f(\varepsilon,\phi;x)|x)\right]\big|_{\phi'=\phi} = \left[\varepsilon^{(k)}+\phi-\phi'\right]\big|_{\phi'=\phi} = \varepsilon^{(k)}$ 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 $\ell \in \mathbb{N}^\star$,

$$\mathbb{E}\left(\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x))}\right)^\ell\right) = e^{\frac{\ell(\ell-1)}{2}\|\theta-\phi\|^2}, \tag{144}$$

we obtain that

$$\phi_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = (\theta_k-\phi_k)^2 e^{2\|\theta-\phi\|^2}\left(e^{4\|\theta-\phi\|^2} - 4e^{2\|\theta-\phi\|^2} + 4e^{\|\theta-\phi\|^2} - 1\right). \tag{145}$$

Lastly, we obtain the desired expressions in Example 3 by taking $\theta - \phi = \epsilon \cdot \boldsymbol{u}_d$ with $\epsilon > 0$ in (136), (137), (140), (141) and applying Theorems 1 to 3.

B.3.3 Proof of Example 4

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

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

$$p_\theta(x) = \int p_\theta(x, z)\mathrm{d}z = \mathcal{N}(x; \theta, 2\boldsymbol{I}_d), \tag{146}$$

$$p_\theta(z|x) = \mathcal{N}(z; (\theta + x)/2, 1/2\ \boldsymbol{I}_d)\ . \tag{147}$$

• **Checking (A3), (A_h), ($\mathsf{A}^{\mathrm{REP}}_{h'}$) ($\mathsf{A}^{\mathrm{REP}}_{\mathrm{df}}$), ($\mathsf{A}^{\mathrm{DREP}}_{h'}$), ($\widetilde{\mathsf{A}}^{\mathrm{DREP}}_{\tilde{h}}$) and ($\mathsf{A}^{\mathrm{DREP}}_{\mathrm{df}}$).** Let us prove (126)-(133). 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 (126) 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 $q = \mathcal{N}(0, \boldsymbol{I}_d)$, we easily get (127)–(130). Conditions (131)–(133) are obtained similarly.

• **Closed-form expressions for the quantities appearing in Theorems 1 to 3.** First recall 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 (146). so that

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

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

and

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

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

$$\frac{p_\theta(f(\varepsilon,\phi';x)|x)}{q_\phi(f(\varepsilon,\phi';x)|x)} = C\exp\left(\frac{3}{4}\|A'x+\tilde{b}'-Ax-\tilde{b}\|^2\right) \times \exp\left(-\frac{1}{2\cdot 3}\left(\|\varepsilon\|^2+2\left\langle \varepsilon, \sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right), \quad (148)$$

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

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

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 $\phi,\phi' \in \mathbb{R}^{2d}$,

$$\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})} = \exp\left(-\frac{1-\alpha}{2\cdot 3}\left(\|\varepsilon\|^2+2\left\langle \varepsilon, \sqrt{\frac{3}{2}}\left(A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right)\right\rangle\right)\right) \times \left(\frac{4-\alpha}{3}\right)^{d/2}\exp\left(-\frac{(1-\alpha)^2}{4(4-\alpha)}\left\|A'x+\tilde{b}'+3(Ax+\tilde{b})-2(\theta+x)\right\|^2\right). \quad (149)$$

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

$$\partial_{\theta_k}\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right) = (1-\alpha)\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)}+\frac{1-\alpha}{4-\alpha}[4(\tilde{a}_k x_k+\tilde{b}_k)-2(\theta_k+x_k)]\right)\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})} \quad (150)$$

and we also have

$$\partial_{\tilde{b}_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})} = -2\partial_{\theta_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}, \quad (151)$$

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

$$\left[\partial_{\tilde{b}'_k}\frac{\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)^{1-\alpha})}\right]\Bigg|_{\phi'=\phi} = -\frac{1}{2}\partial_{\theta_k}\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}. \quad (152)$$

In addition,

$$\theta_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \mathbb{E}\left(\left(\sqrt{\frac{2}{3}}\varepsilon^{(k)} + \frac{1-\alpha}{4-\alpha}[4(\tilde{a}_k x_k + \tilde{b}_k) - 2(\theta_k + x_k)]\right)^2 \left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^2\right) \quad (153)$$

and now using (149), we have that: for all $\ell \in \mathbb{N}^\star$,

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

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,

$$q(\varepsilon)\left(\frac{\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha}}{\mathbb{E}(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)^{1-\alpha})}\right)^\ell = h(\ell)\left(\frac{2\pi\cdot 3}{3+(1-\alpha)\ell}\right)^{-d/2}\exp\left(-\frac{3+(1-\alpha)\ell}{2\cdot 3}\left\|\varepsilon + \frac{(1-\alpha)\ell}{3+(1-\alpha)\ell}\sqrt{\frac{3}{2}}\mu\right\|^2\right) \quad (154)$$

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 (154) and plugging this in (153), we get that

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

from which we can also deduce $\tilde{b}_k\text{-}V^{(\alpha,\mathrm{REP})}(\theta,\phi;x)$ using (151). Similarly, using (152), 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 (150) and (152) 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 (143) 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 (154) with $\alpha = 0$ and $\ell = 2$, we obtain that

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

Next, using the above as well as (154) with $\alpha = 0$ and $\ell = 2$ and (142) with $\phi_k = \tilde{b}_k$, we can compute all the terms appearing in $\tilde{b}_k$-$V^{(0,\mathrm{DREP})}(\theta,\phi;x)$ 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

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

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

$$\begin{aligned}\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = &\frac{1}{4}\left(\frac{4}{3}\right)^{2d}\left(\frac{3}{7}\right)^{d/2} e^{\frac{9}{4\cdot 7}\|\mu\|^2}\left(\frac{2}{7} + \frac{3^2}{7^2}\mu_k^2\right)\\ &- \left(\frac{4}{3}\right)^{3d/2}\left(\frac{1}{2}\right)^{d/2} e^{\frac{3}{4\cdot 4}\|\mu\|^2}\frac{3}{2\cdot 5}\frac{4^d}{15^{d/2}}e^{\frac{3}{8\times 5}\|\mu\|^2}\mu_k^2\\ &+ \left(4\frac{4^d}{15^{d/2}}e^{\frac{3}{8\cdot 5}\|\mu\|^2} - 1\right)\frac{3^2}{2^2\times 5^2}\frac{4^{2d}}{15^d}e^{\frac{3}{4\times 5}\|\mu\|^2}\mu_k^2.\end{aligned} \tag{155}$$

Lastly, we obtain the desired expressions in Example 4 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 (155) becomes:

$$\begin{aligned}\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x) = &\frac{1}{4}\left(\frac{4}{3}\right)^{2d}\left(\frac{3}{7}\right)^{d/2} e^{\frac{36}{7}d\epsilon^2}\left(\frac{2}{7} + \left(\frac{12\epsilon}{7}\right)^2\right)\\ &- \left(\frac{4}{3}\right)^{3d/2}\left(\frac{1}{2}\right)^{d/2} e^{3d\epsilon^2}\frac{3}{2\cdot 5}\frac{4^d}{15^{d/2}}e^{\frac{6}{5}d\epsilon^2}\mu_k^2 + \left(4\frac{4^d}{15^{d/2}}e^{\frac{6}{5}d\epsilon^2} - 1\right)\frac{3^2}{2^2\times 5^2}\frac{4^{2d}}{15^d}e^{\frac{12}{5}d\epsilon^2}\mu_k^2.\end{aligned} \tag{156}$$

C Deferred proofs and results for Section 3.2

C.1 Preliminary results for the proofs of Theorems 4 and 5

In the following we will use the asymptotic equivalence

$$1 - \Phi(u) = u^{-1}\varphi(u)(1 + o(1)) \quad \text{as } u \to \infty, \tag{157}$$

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-negative as soon as $u > 0$, since, in this case,

$$1 - \Phi(u) = \int_u^\infty \varphi(y)\, \mathrm{d}y \leq u^{-1} \int_u^\infty y\, \varphi(y)\, \mathrm{d}y = u^{-1}\varphi(u). \tag{158}$$

We then have the following two lemmas.

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

$$\mathbb{P}\left(M_N > \sqrt{2\,(1+c)\,\log N}\right) = O\left((\log N)^{-1/2}\, N^{-c}\right), \tag{159}$$

$$\mathbb{P}\left(M_N < \sqrt{2\,(1+c)^{-1}\,\log N}\right) = O\left(\mathrm{e}^{-N^{c/(1+c)}/\sqrt{4\pi \log N}}\right). \tag{160}$$

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 (157) 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 (159) and (160) easily follow. □

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

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

$$\mathbb{E}\left(|M_N|^m\right) = (2 \log N)^{m/2}\left(1 + O\left(\frac{\log \log N}{\log N}\right)\right) \tag{161}$$

$$\mathbb{E}\left((M_N)_-^m\right) = O\left(\rho^N\right) \quad \textit{for any } \rho \in (1/2, 1), \tag{162}$$

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 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 (161). To get (162), 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 (161) 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 $s \in \mathbb{R}$,*

$$\zeta(s) = \frac{\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z) - F^{\leftarrow}\circ\Phi(s)} \,\middle|\, Z \leq s\right]}{1 - \Phi\left(s\right)}. \tag{163}$$

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 $p \in (1,\infty)$,

$$\tilde{\zeta}_N^{(m,p)}(u,s) = N \, \left(\left\| \tilde{Y}_1 \right\|_{p\,m}^{1/m} \, s^{(p-1)/p} + \left\| \tilde{Y}_1 \, \mathbb{1}_{\left\{\tilde{Y}_N^* > u\right\}} \right\|_m \right). \tag{164}$$

Then, 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 $W_{I_N} = \max\left(W_1,\ldots,W_N\right)$ and, for all $m, \delta \in [1,\infty)$, $\underline{b} < 2 < \overline{b}$, $p \in (1,\infty)$, $u > 0$ and $N \in \mathbb{N}^*$, we have

$$\left\| \sum_{i=1}^{N} \overline{W}_{i,N}^{\delta} \, \tilde{Y}_i - \tilde{Y}_{I_N} \right\|_m \leq C \left(u \sup_{\substack{b\in[\underline{b},\overline{b}] \\ q=1,m}} \left(\zeta(b\sqrt{\log N})\right)^{\frac{q}{m}} + \tilde{\zeta}_N^{(m,p)}\left(u, (\log N)^{-\frac{1}{2}} \, N^{1-\frac{\overline{b}}{2}}\right) \right), \tag{165}$$

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

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

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, Thorem B.3.11), that is $\mathbb{P}\left[\tilde{Y}_1 \le \tilde{y} \,\middle|\, W_1\right] = Q(W_1, (-\infty, \tilde{y}])$ $\mathbb{P}$-a.s. for all $\tilde{y} \in \mathbb{R}$. Further denote the associated conditional generalized inverse by $Q^{\leftarrow}$, that is

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

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 $(Z_i)_{i\ge1}$ and $(U_i)_{i\ge1}$ be two independent i.i.d. sequence distributed as Z and U, respectively and set, without loss of generality,

$$\forall i \ge 1, \quad W_i = \mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)} \quad \text{and} \quad \tilde{Y}_i = Q^{\leftarrow}(W_i, U_i)\,. \tag{166}$$

Now denoting by I_N the (random) index in $\{1,\dots,N\}$ such that $Z_{I_N} = \max(Z_1,\dots,Z_N)$, we have that I_N is a.s. uniquely defined since $Z_1,\dots,Z_N$ are independent and Φ is continuous. Furthermore, as required in the proposition, $(I_N)_{N\ge1}$ is a non-decreasing random sequence such that: for all $N \ge 1$, I_N is valued in $\{1,\dots,N\}$ with $W_{I_N} = \max(W_1,\dots,W_N)$.

Next let $m, \delta \in [1,\infty)$, $\underline{b} < 1 < \overline{b}$ 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 (165) 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 $J_N := \{1,\dots,N\} \setminus \{I_N\}$, 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 (165), 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 $\left\{\underline{b} \le \frac{Z_{I_N}}{\sqrt{\log N}} \le \overline{b}\right\} \cap \left\{\tilde{Y}_N^* \le u\right\}$ and its complementary set. More specifically, to prove (165) it suffices to show that

$$\mathbb{E}\left(|A_N|^m \, \mathbb{1}_{\left\{\underline{b} \le \frac{Z_{I_N}}{\sqrt{\log N}} \le \overline{b}\right\} \cap \left\{\tilde{Y}_N^* \le u\right\}} \right) \le \delta\, C_2\, u^m \sup_{b \in \left[\underline{b},\overline{b}\right], q=1,m} \left(\zeta(b\sqrt{\log N})\right)^q \tag{167}$$

$$\mathbb{E}\left(|B_N|^m \, \mathbb{1}_{\left\{\underline{b} \le \frac{Z_{I_N}}{\sqrt{\log N}} \le \overline{b}\right\} \cap \left\{\tilde{Y}_N^* \le u\right\}} \right) \le C_2\, u^m \sup_{b \in \left[\underline{b},\overline{b}\right], q=1,m} \left(\zeta(b\sqrt{\log N})\right)^q \tag{168}$$

$$\mathbb{E}\left(\left(\sum_{i=1}^N \left|\tilde{Y}_i\right| \right)^m \mathbb{1}_{\left\{\frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b},\overline{b}\right]\right\} \cup \left\{\tilde{Y}_N^* > u\right\}} \right) \le C_1 \left(\tilde{\zeta}_N^{(m,p)}(u, (\log N)^{-\frac{1}{2}}\, N^{1-\overline{b}/2})\right)^m, \tag{169}$$

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 (167), (168), the Minkowski inequality, (169) 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 (167) and (168) hold and (ii) the inequality (169) holds.

(i) *Proof of (167) and (168).* For all $N \geq 2$, set

$$D_N = \sum_{i\in J_N} \frac{W_i}{W_{I_N}}. \tag{170}$$

Then, for all $N \geq 2$,

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

where in the last inequality we used the definition of $\tilde{Y}_N^*$ and the fact that $(1+x)^{\delta} \leq 1+\delta x$ 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 (167) and (168), 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

$$\mathbb{E}\left(D_N^m\,\mathbb{1}_{\left\{\underline{b}\leq \frac{Z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}}\right) \leq C_2 \sup_{b\in\left[\underline{b},\bar{b}\right],q=1,m}\left(\zeta(b\sqrt{\log N})\right)^q. \tag{171}$$

By definition of D_N in (170) and using (166), 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

$$\begin{aligned}
&\mathbb{E}\left(D_N^m\,\mathbb{1}_{\left\{\underline{b}\leq \frac{Z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}}\right) \\
&\qquad = \mathbb{E}\left(\mathbb{1}_{\left\{\underline{b}\leq \frac{Z_{I_N}}{\sqrt{\log N}}\leq \bar{b}\right\}}\,\mathbb{E}\left[\left(\sum_{i=1}^{N-1}\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)-F^{\leftarrow}\circ\Phi(s)}\right)^m \,\middle|\, Z_i \leq Z_{I_N}\,,\,1\leq i\leq N-1\right]\right).
\end{aligned} \tag{172}$$

Since the $Z_1, \dots, Z_N$ are i.i.d. with the same distribution as Z, by the Rosenthal inequality (see Therorem 2.12 Petrov, 1995), we have, for a constant $C_4 > 0$ only depending on m,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{N-1} \mathrm{e}^{F^{\leftarrow}\circ\Phi(Z_i)-F^{\leftarrow}\circ\Phi(s)}\right)^m \,\middle|\, Z_i \le Z_{I_N}\,,\ 1 \le i \le N-1\right] \\
&\qquad \le C_4\ \Big[(N-1)\,\mathbb{E}\left[\mathrm{e}^{m\,(F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s))} \,\middle|\, Z \le Z_{I_N}\right] \\
&\qquad\quad + \left((N-1)\,\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s)} \,\middle|\, Z \le Z_{I_N}\right]\right)^m\Big] \\
&\qquad \le C_4 \sum_{q=1,m} \left((N-1)\mathbb{E}\left[\mathrm{e}^{F^{\leftarrow}\circ\Phi(Z)-F^{\leftarrow}\circ\Phi(s)} \,\middle|\, Z \le Z_{I_N}\right]\right)^q,
\end{aligned}$$

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 (172), we get that

$$\begin{aligned}
\mathbb{E}\left(D_N^m\, \mathbb{1}_{\left\{\underline{b} \le \frac{Z_{I_N}}{\sqrt{\log N}} \le \bar{b}\right\}}\right) &\le C_4 \sum_{q=1,m} \mathbb{E}\left(\mathbb{1}_{\left\{\underline{b} \le \frac{Z_{I_N}}{\sqrt{\log N}} \le \bar{b}\right\}}\ \left\{(N-1)(1-\Phi(Z_{I_N}))\,\zeta(Z_{I_N})\right\}^q\right) \\
&\le 2\,C_4 \max_{q=1,m}\left((N-1)^q\,\mathbb{E}\left((1-\Phi(Z_{I_N}))^q\right) \sup_{b\in[\underline{b},\bar{b}]}\left(\zeta(b\,\sqrt{\log N})\right)^q\right)\,.
\end{aligned}$$

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 (171) holds with $C_2 = 2\,C_4\,\Gamma(m+1)$, which concludes the proof of (167) and (168).

(ii) *Proof of (169).* 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 $\mathbf{D}_N = \left\{\frac{Z_{I_N}}{\sqrt{\log N}} \notin [\underline{b},\bar{b}]\right\} \cup \left\{\tilde{Y}_N^* > u\right\}$ 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 $\mathbb{1}_{\mathbf{D}_N} \le \mathbb{1}_{\left\{\frac{Z_{I_N}}{\sqrt{\log N}} \notin [\underline{b},\bar{b}]\right\}} + \mathbb{1}_{\left\{\tilde{Y}_N^* > u\right\}}$ and, by definition of $\tilde{\zeta}_N^{(m,p)}$ in (164), to get (169) with $C_1 = \max\left(1, C_3^{(p-1)/p}\right)$, it only remains to show that

$$\mathbb{E}\left(\left|\tilde{Y}_i\right|^m\,\mathbb{1}_{\left\{\frac{Z_{I_N}}{\sqrt{\log N}} \notin [\underline{b},\bar{b}]\right\}}\right) \le \left(\mathbb{E}\left(\left|\tilde{Y}_1\right|^{p\,m}\right)\right)^{1/p}\left(C_3\,(\log N)^{-1/2}\,N^{1-\bar{b}/2}\right)^{(p-1)/p}\,, \tag{173}$$

for a certain $C_3 > 0$ only depending on $\underline{b}$ and $\overline{b}$. Using Lemma 5 with $c = \overline{b}/2 - 1$ and $c = 2/\underline{b} - 1$ successively in (159) and (160), we indeed find for such a constant C_3,

$$\mathbb{P}\left(\frac{Z_{I_N}}{\sqrt{\log N}} \notin \left[\underline{b}, \overline{b}\right]\right) \leq C_3 \left((\log N)^{-1/2} \; N^{1-\overline{b}/2}\right) .$$

Then we get (173) by applying the Hölder inequality and the proof of (169) is concluded.

□

Remark 4 *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, (165) 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 (165). However, we will apply (165 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 (165). Furthermore, since (165) holds for any* $u > 0$, *we will choose* u *for a given* N *so that the right-hand side of (165) 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 (164) defining* $\tilde{\zeta}_N^{(m,p)}(u,s)$ *which decreases as* u *increases. Finally, the first term between the parentheses of (164) will be made small by taking* $\overline{b}$ *large enough to compensate the* N *in front of the parentheses.*

C.2 Additional notation, useful first properties and derivations

In the remaining of Appendix C, we let $\varepsilon \sim q$ and $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. copies of ε and we assume that (B1) holds and so does ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$) (when dealing with the REP gradient estimator) and ($\mathsf{B}_{\mathrm{df}}^{\mathrm{DREP}}$) (when dealing with the REP gradient estimator).

We now introduce the following helpful notation:

$$\Lambda_{\theta,\phi,\phi'}(\cdot;x) = \log \tilde{w}_{\theta,\phi,\phi'}(\cdot;x) \tag{174}$$

$$\overline{\mathcal{W}}_{j,N}^{(\alpha)} = \left(\sum_{k=1}^{N} \mathrm{e}^{(1-\alpha)\,\Lambda_{\theta,\phi,\phi}(\varepsilon_k;x)}\right)^{-1} \mathrm{e}^{(1-\alpha)\,\Lambda_{\theta,\phi,\phi}(\varepsilon_j;x)} , \qquad 1 \leq j \leq N \tag{175}$$

$$\widetilde{\Lambda}_{\theta,\phi,\phi'}(\cdot;x) = \frac{\Lambda_{\theta,\phi,\phi'}(\cdot;x) - \mathbb{E}\left(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)\right)}{B_d(\theta,\phi,\phi';x)} \tag{176}$$

$$\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)^2 = \mathbb{V}\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon;x)\right) \tag{177}$$

$$\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2 = \mathbb{V}\left(\left[\partial_{\psi'} \widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon;x)\right]\Big|_{\phi'=\phi}\right) . \tag{178}$$

Under (B1), the expectation and variance of $\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)$ are well defined. Recalling from Section 3.2 that $B_d(\theta,\phi,\phi';x)$ denotes the variance of $\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)$, we have that $\widetilde{\Lambda}_{\theta,\phi,\phi'}$ in well defined in (176) if

$B_d(\theta, \phi, \phi'; x) > 0$. Since we consider the setting where $N, d \to \infty$ under Condition (29) in Theorems 4 and 5, this implies that, eventually, $B_d(\theta, \phi, \phi'; x) > 0$. Consquently, we will always assume that $B_d(\theta, \phi, \phi'; x) > 0$ in the following without loss of generality. Furthermore, using Propositions 3 and 4, we get the following assertions:

(i) We have that $\partial_\psi B_d(\theta, \phi, \phi; x)$ and $[\partial_{\psi'} B_d(\theta, \phi, \phi'; x)]|_{\phi'=\phi}$ are well defined;

(ii) We have that $\widetilde{B}_d^{(\mathrm{REP})}(\theta, \phi; x)^2$ and $\widetilde{B}_d^{(\mathrm{DREP})}(\theta, \phi; x)^2$ are also well defined;

(iii) We have that $\left(\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x), \partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ is a Gaussian process, and so is $\left(\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x), \left[\partial_{\psi'} \widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\Big|_{\phi'=\phi}\right)\right)_{(\theta,\phi)\in\Theta\times\Phi}$.

(iv) We can interchange the derivatives ∂_ψ and $[\partial_{\psi'}(\cdot)]|_{\phi'=\phi}$ with the expectation signs for $\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)$ and $\widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)$, respectively, as well as their squares.

Using Assertion (iv) above and since $\widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)$ is centered with variance 1, we then get that $\mathbb{E}\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right) = \mathbb{C}\mathrm{ov}\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x), \partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right) = 0$ and $\mathbb{E}\left(\left[\partial_{\psi'} \widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\Big|_{\phi'=\phi}\right) = \mathbb{C}\mathrm{ov}\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x), \left[\partial_{\psi'} \widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\Big|_{\phi'=\phi}\right) = 0$. Combining this with Assertion (iii), we thus have the following first properties:

($\mathrm{P}^{\mathrm{REP}}$) $\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ and $\left(\partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ are two independent and centered Gaussian processes with variances equal to 1 and $\left(\widetilde{B}_d^{(\mathrm{REP})}(\theta, \phi; x)^2\right)_{(\theta,\phi)\in\Theta\times\Phi}$, respectively.

($\mathrm{P}^{\mathrm{DREP}}$) $\left(\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon; x)\right)_{(\theta,\phi)\in\Theta\times\Phi}$ and $\left(\left[\partial_{\psi'} \widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon; x)\right]\Big|_{\phi'=\phi}\right)_{(\theta,\phi)\in\Theta\times\Phi}$ are two independent and centered Gaussian processes with variances equal to 1 and $\left(\widetilde{B}_d^{(\mathrm{REP})}(\theta, \phi; x)^2\right)_{(\theta,\phi)\in\Theta\times\Phi}$, respectively.

Finally, let us rewrite the REP and DREP gradient estimators (defined in (12) and (14) respectively) using the above notation. Since their expectation and variance can be deduced from the case $M = 1$:

$$
\begin{aligned}
&\mathbb{E}(\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = \mathbb{E}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) \ , \\
&\mathbb{V}(\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) = M^{-1}\mathbb{V}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta, \phi; x)) \ , \\
&\mathbb{E}(\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) = \mathbb{E}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) \ , \\
&\mathbb{V}(\psi\text{-}g_{M,N,d}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) = M^{-1}\mathbb{V}(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta, \phi; x)) \ ,
\end{aligned}
$$

we in fact only consider the case $M = 1$ for simplicity and we obtain that

$$\begin{aligned}
\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) &= \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\ \partial_\psi \Lambda_{\theta,\phi,\phi}(\varepsilon_j;x) \\
&= \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\ \partial_\psi \left(\Lambda_{\theta,\phi,\phi}(\varepsilon_j;x) - \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x))\right) \\
&= \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + \partial_\psi B_d(\theta,\phi,\phi;x)\ \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\ \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x) \\
&\quad + B_d(\theta,\phi,\phi;x) \sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)}\ \partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\ , \qquad (179)
\end{aligned}$$

as well as

$$\begin{aligned}
\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) &= \sum_{j=1}^{N} \left(\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)} + (1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right) \left[\partial_{\psi'}\Lambda_{\theta,\phi,\phi'}(\varepsilon_j;x)\right]\big|_{\phi'=\phi} \\
&= \left(\alpha + (1-\alpha)\sum_{j=1}^{N}(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right) \left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi} \\
&\quad + \left[\partial_{\psi'}B_d(\theta,\phi,\phi';x)\right]\big|_{\phi'=\phi} \sum_{j=1}^{N} \left(\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)} + (1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right) \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x) \\
&\quad + B_d(\theta,\phi,\phi;x) \sum_{j=1}^{N} \left(\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)} + (1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right) \left[\partial_{\psi'}\left(\widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon_j;x)\right)\right]\Big|_{\phi'=\phi}\ . \qquad (180)
\end{aligned}$$

C.3 Proofs of Theorems 4 and 5

We start by presenting two lemmas.

Lemma 7 *Let $Z \sim \mathcal{N}(0,1)$. Then, for all $s \in \mathbb{R}$ and $\sigma > 0$,*

$$\mathbb{E}\left(\mathrm{e}^{\sigma(Z-s)} \,|\, Z \le s\right) = \frac{\Phi(s-\sigma)}{\Phi(s)}\ \frac{\varphi(s)}{\varphi(\sigma-s)}\ . \qquad (181)$$

It follows that, defining ζ by (163) 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 $\sigma \ge 2s \ge 1$,

$$\zeta(s) \le C\ \frac{s}{\sigma}\ . \qquad (182)$$

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 (181). Now, defining ζ by (163) 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 $\beta \in \mathbb{R}$,*

$$\overline{W}_{j,N}(\beta) := \left(\sum_{k=1}^{N} \mathrm{e}^{\beta\,\xi_k}\right)^{-1} \mathrm{e}^{\beta\,\xi_j}\,, \qquad 1 \le j \le N\,. \tag{183}$$

Let $\delta \ge 1$, $\lambda \in [0,1]$ and $q \ge 1$. Then we have, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$,

$$\mathbb{E}\left(\sum_{j=1}^{N} (\overline{W}_{j,N}(\beta))^{\delta}\,\xi_j\right) = (2\,\log N)^{1/2}\,(1+o(1))\,, \tag{184}$$

$$\mathbb{V}\left(\sum_{j=1}^{N} [\lambda\overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\,\xi_j\right) = o\left(\log N\right), \tag{185}$$

$$\mathbb{E}\left(\sum_{j=1}^{N} \left(\overline{W}_{j,N}(\beta)\right)^{\delta}\right) = 1 + o(1), \tag{186}$$

$$\mathbb{V}\left(\sum_{j=1}^{N} \left(\overline{W}_{j,N}(\beta)\right)^{\delta}\right) = O\left(\frac{\sqrt{\log N}}{\beta} \vee N^{-q}\right). \tag{187}$$

Proof. We will first show that, for any $m \in [1, \infty)$, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$,

$$\left\| \sum_{i=1}^{N} (\overline{W}_{i,N}(\beta))^{\delta} \xi_i - \max(\xi_1, \dots, \xi_N) \right\|_m = o\left(\sqrt{\log N}\right) , \tag{188}$$

$$\left\| \sum_{i=1}^{N} (\overline{W}_{i,N}(\beta))^{\delta} - 1 \right\|_m = O\left(\left(\frac{\sqrt{\log N}}{\beta} \right)^{\frac{1}{m}} + N^{-\frac{q}{m}} \right) . \tag{189}$$

Then we will show that (188) implies (184)-(185), and that (189) implies (186)-(187).

(i) *Proof of (188) and (189).* We use Proposition 7 successively in the two settings

Setting 1) for all $j = 1 \dots N$, $W_j = \mathrm{e}^{\beta \xi_j}$ and $\tilde{Y}_j = \xi_j$;
Setting 2) for all $j = 1 \dots N$, $W_j = \mathrm{e}^{\beta \xi_j}$ 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 $\{1, \dots, N\}$ with $W_{I_N} = \max(W_1, \dots, W_N)$ and, for all $m, \delta \in [1, \infty)$, $\underline{b} < 2 < \overline{b}$, $p \in (1, \infty)$, $u > 0$ and $N \in \mathbb{N}^*$, we have in Setting 1) and Setting 2),

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

where ζ is defined in (163), $\tilde{\zeta}_N^{(m,p)}$ is defined in (164) 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 $b \leq \overline{b}$ and $\beta \geq 2 \, (\overline{b} \vee C_0) \sqrt{\log N} \geq 1$, then the condition on the left-hand side is satified and the upper bound in the right-hand side is at most 1. Using this in the previous display for the special case $p = 2$, we get that for all $m, \delta \in [1, \infty)$ and $\overline{b} > 2$, there exists $C, C_0 > 0$ such that, for all $u > 0$, if $\beta \geq 2 \, (\overline{b} \vee C_0) \sqrt{\log N} \geq 1$,

$$\left\| \sum_{i=1}^{N} (\overline{W}_{i,N}(\beta))^{\delta} \tilde{Y}_i - \tilde{Y}_{I_N} \right\|_m \leq C \left(u \left(\frac{\overline{b} \, C_0 \sqrt{\log N}}{\beta} \right)^{\frac{1}{m}} + \tilde{\zeta}_N^{(m,2)} \left(u, (\log N)^{-\frac{1}{2}} \; N^{1 - \frac{\overline{b}}{2}} \right) \right) . \tag{190}$$

To prove (188) and (189) we now consider Setting 1) and Setting 2) separately by taking $\tilde{Y}_j = \xi_j$ and $\tilde{Y}_j = 1$ respectively in $\tilde{\zeta}_N^{(m,2)} \left(u, (\log N)^{-\frac{1}{2}} \; N^{-1 - \frac{\overline{b}}{2}} \right)$ and choosing adequate values of $\overline{b}$ and u in the upper bound above. Namely, we will show that

- If $\tilde{Y}_j = \xi_j$ for all $j \geq 1$, then: as $N \to \infty$,

$$\tilde{\zeta}_N^{(m,2)} \left(\sqrt{10 \log N}, (\log N)^{-\frac{1}{2}} \; N^{-2} \right) = o\left(\log N \right) . \tag{191}$$

- If $\tilde{Y}_j = 1$ for all $j \geq 1$, then: for all $N \geq 2$,

$$\tilde{\zeta}_N^{(m,2)}\left(2, (\log N)^{1-\frac{\bar{b}}{2}}\right) = (\log N)^{-\frac{1}{4}}\, N^{\frac{3}{2}-\frac{\bar{b}}{4}}\ . \tag{192}$$

Observe then that (192) follows directly from plugging in $(u,s) = \left(2, (\log N)^{1-\frac{\bar{b}}{2}}\right)$ in (164). As for (191): setting $\tilde{Y}_j = \xi_j$ for all $j \geq 1$ and plugging in $(u,s) = (\sqrt{10\log N}, (\log N)^{-\frac{1}{2}}\ N^{-2})$ in (164) yields that for all $N \geq 2$,

$$\tilde{\zeta}_N^{(m,2)}\left(\sqrt{10\log N}, (\log N)^{-\frac{1}{2}}\ N^{-2}\right) = \|\xi_1\|_{pm}^{1/m}\,(\log N)^{-\frac{1}{4}} + N\left\|\xi_1\,\mathbb{1}_{\left\{\xi_N^* > \sqrt{10\log N}\right\}}\right\|_m\ ,$$

where $\xi_N^* = \max(|\xi_1|, \ldots, |\xi_N|)$. Since $\left\|\xi_1\,\mathbb{1}_{\left\{\xi_N^* > \sqrt{10\log N}\right\}}\right\|_m \leq \|\xi_1\|_{2m}\left\|\mathbb{1}_{\left\{\xi_N^* > \sqrt{2(1+c)\log N}\right\}}\right\|_{2m}$, the desired result (191) will follow if we can show that: as $N \to \infty$,

$$\mathbb{P}\left(\xi_N^* > \sqrt{10\log N}\right) = o\left(N^{m/2}\right)\ .$$

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 (159) in Lemma 5 with $c = 4$.

Now, for any given $\bar{b} > 2$, if $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$, we eventually have $\beta \geq 2\,(\bar{b} \vee C_0)\sqrt{\log N} \geq 1$ so that (190) 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 (190) with $\bar{b} = 6$ among with (191), gives us that, as $N, \beta \to \infty$ with $\sqrt{\log N} = o(\beta)$, (188) hold. If now $\tilde{Y}_j = 1$ for all j, we have $\tilde{Y}_{I_N} = 1$ and the bound (190) with $\bar{b}$ large enough among with (192), gives us that (189) holds.

(ii) *Proofs of (184) and (185).* Since the expectation is 1-Lipschitz for the L^1-norm, the asymptotic behavior (188) with $m = 1$ implies that, for any $\delta \geq 1$,

$$\mathbb{E}\left(\sum_{i=1}^N (\overline{W}_{i,N}(\beta))^{\,\delta}\,\xi_i\right) = \mathbb{E}\left(\max(\xi_1, \ldots, \xi_N)\right)_m + o\left(\sqrt{\log N}\right)\ .$$

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 (184). The cases $\delta = 1, 2$ in particular together give that, for any $\lambda \in [0,1]$,

$$\mathbb{E}\left(\sum_{j=1}^N [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\,\xi_j\right) = (2\,\log N)^{1/2}\,(1 + o(1))\ . \tag{193}$$

By the Minkowski inequality for the L^2-norm, we obtain that

$$\begin{aligned}
&\left|\left\|\sum_{j=1}^N [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\,\xi_j\right\|_2 - \|\max(\xi_1, \ldots, \xi_N)\|_2\right| \\
&\quad\leq \left\|\sum_{j=1}^N [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2]\,\xi_j - \max(\xi_1, \ldots, \xi_N)\right\|_2 \\
&\quad\leq \lambda\left\|\sum_{j=1}^N \overline{W}_{j,N}(\beta)\,\xi_j - \max(\xi_1, \ldots, \xi_N)\right\|_2 + (1-\lambda)\left\|\sum_{j=1}^N (\overline{W}_{j,N}(\beta))^2\,\xi_j - \max(\xi_1, \ldots, \xi_N)\right\|_2\ .
\end{aligned}$$

Hence, using (188) again, this time with $m = 2$, we deduce

$$\left| \left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2] \, \xi_j \right\|_2 - \|\max(\xi_1, \dots, \xi_N)\|_2 \right| = o\left(\sqrt{\log N}\right),$$

which we can rewrite equivalently as

$$\left\| \sum_{j=1}^{N} [\lambda \overline{W}_{j,N}(\beta) + (1-\lambda)(\overline{W}_{j,N}(\beta))^2] \, \xi_j \right\|_2 = \|\max(\xi_1, \dots, \xi_N)\|_2 + o\left(\sqrt{\log N}\right).$$

Using Lemma 6 with $m = 2$ to evaluate the first term in the right-hand side and taking the sqaure, 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 (193) yields (185).

(iii) *Proofs of (186) and (187).* The proofs of (186) and (187) are obtained similarly to the proofs of (184) and (185), but starting from (189) instead of (188).

□

We can now prove Theorems 4 and 5.

C.3.1 Proof of Theorem 4

The proof is made of two steps. Firstly we show that: as $N, d \to \infty$ with (29) holding,

$$\mathbb{E}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + (2\log N)^{1/2} \; \partial_\psi B_d(\theta,\phi,\phi;x) \; (1+o(1)) \tag{194}$$

$$\begin{aligned} \mathbb{V}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) &= (\partial_\psi B_d(\theta,\phi,\phi;x))^2 \; o(\log N) \\ &\quad + B_d(\theta,\phi,\phi;x)^2 \widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)^2 \, (1+o(1)) \; . \end{aligned} \tag{195}$$

Secondly, further assuming (30), we show that (31) holds as $N, d \to \infty$.

(i) **Proof of (194) and (195).** Recall that $\varepsilon_1, \varepsilon_2, \dots$ are i.i.d. copies of ε. Since $\overline{\mathcal{W}}_{1,N}^{(\alpha)}, \dots, \overline{\mathcal{W}}_{N,N}^{(\alpha)}$ as defined in (175) can be expressed as functions of $\Lambda_{\theta,\phi,\phi}(\varepsilon_k;x)$ for all $k = 1 \dots N$, by Property ($\mathrm{P}^{\mathrm{REP}}$) they are independent of the centered random variables $\partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_1;x), \dots, \partial_\psi \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_N;x)$. Furthermore, Property ($\mathrm{P}^{\mathrm{REP}}$) also gives us that $\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_1;x), \dots, \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_N;x)$ are i.i.d. $\mathcal{N}(0,1)$. Next noticing that $\overline{\mathcal{W}}_{1,N}^{(\alpha)}, \dots, \overline{\mathcal{W}}_{N,N}^{(\alpha)}$ can be rewritten using $\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_k;x)$ for all $k = 1 \dots N$, we can apply Lemma 8 with $\xi_k = \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_k;x)$, and $\beta = (1-\alpha) B_d(\theta,\phi,\phi;x)$ to prove (194) and (195). More precisely, by taking the expectation in (179) we get that

$$\mathbb{E}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + \partial_\psi B_d(\theta,\phi,\phi;x) \; \mathbb{E}\left(\sum_{j=1}^{N} \overline{\mathcal{W}}_{j,N}^{(\alpha)} \; \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x) \right)$$

and, as $N, d \to \infty$ with (29) holding, (184) with $\delta = 1$ gives us that (194) holds. Similarly, by taking the variance in (179) we have that

$$\mathbb{V}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)\right) = \left(\partial_\psi B_d(\theta,\phi,\phi;x)\right)^2 \mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right) + \left(B_d(\theta,\phi,\phi;x)\right)^2 \mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \partial_\psi\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)$$

and (185) with $\lambda = 1$ in Lemma 8, gives us that the first variance in the right-hand side is $o(\log N)$ as $N, d \to \infty$ with (29) holding. Now observe that, using Property ($\mathrm{P}^{\mathrm{REP}}$) again,

$$\begin{aligned}
\mathbb{V}\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \partial_\psi\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right) &= \mathbb{E}\left(\left(\sum_{j=1}^N \overline{\mathcal{W}}_{j,N}^{(\alpha)}\, \partial_\psi\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)^2\right) \\
&= \mathbb{E}\left(\sum_{j=1}^N \left(\overline{\mathcal{W}}_{j,N}^{(\alpha)}\right)^2 \mathbb{E}\left(\left(\partial_\psi\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)^2\right)\right) \\
&= \widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)^2\, \mathbb{E}\left(\sum_{j=1}^N \left(\overline{\mathcal{W}}_{j,N}^{(\alpha)}\right)^2\right)
\end{aligned}$$

and we can conclude by using (186) with $\delta = 2$.

(ii) **Proof of (31).** The marginal log-likelihood satisfies

$$\ell_d(\theta;x) = \log \mathbb{E}_{Z\sim q_\phi(\cdot|x)}\left(w_{\theta,\phi}(Z;x)\right) = \log \mathbb{E}\left(\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x)\right) = \log \mathbb{E}\left(\mathrm{e}^{\Lambda_{\theta,\phi,\phi}(\varepsilon;x)}\right) .$$

Since $\Lambda_{\theta,\phi,\phi}(\varepsilon;x)$ is a Gaussian variable with variance $B_d(\theta,\phi,\phi;x)^2$, we get that

$$\ell_d(\theta;x) = \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + \frac{1}{2} B_d(\theta,\phi,\phi;x)^2 . \tag{196}$$

Now, using (179) in the case $N = 1$, we obtain that

$$\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \partial_\psi\mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x)) + \partial_\psi B_d(\theta,\phi,\phi;x)\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_1;x) + B_d(\theta,\phi,\phi;x)\,\partial_\psi\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_1;x) .$$

It follows from Property ($\mathrm{P}^{\mathrm{REP}}$) that

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{|\partial_\psi\mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x))|}{\sqrt{\left(\partial_\psi B_d(\theta,\phi,\phi;x)\right)^2 + B_d(\theta,\phi,\phi;x)^2\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)^2}} \tag{197}$$

and, using (196), that

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) - \partial_\psi\ell_d(\theta;x)] = \frac{|B_d(\theta,\phi,\phi;x)\partial_\psi B_d(\theta,\phi,\phi;x)|}{\sqrt{\left(\partial_\psi B_d(\theta,\phi,\phi;x)\right)^2 + B_d(\theta,\phi,\phi;x)^2\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)^2}} .$$

Using the latter equation, we get that (30) is equivalent to have

$$\sqrt{\log N}|\partial_\psi B_d(\theta,\phi,\phi;x)| = o\left(B_d(\theta,\phi,\phi;x)\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)\right) \tag{198}$$

and it also implies that

$$|\partial_\psi B_d(\theta,\phi,\phi;x)| = o\left(B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)\right)\ . \tag{199}$$

Plugging (198) into (194) and (195) thus yields

$$\mathrm{SNR}[\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{|\partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x))| + o\left(B_d(\theta,\phi,\phi;x)\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)\right)}{B_d(\theta,\phi,\phi;x)\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)\ (1+o(1))}\ .$$

On the other hand, plugging (199) in (197) yields

$$\mathrm{SNR}[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = \frac{|\partial_\psi \mathbb{E}(\Lambda_{\theta,\phi,\phi}(\varepsilon;x))|}{B_d(\theta,\phi,\phi;x)\widetilde{B}_d^{(\mathrm{REP})}(\theta,\phi;x)\ (1+o(1))}\ .$$

We deduce (31) from the last two displays.

C.3.2 Proof of Theorem 5

The proof is made of three steps. Firstly, we show that: as $N, d \to \infty$ with $\sqrt{\log N} \ll B_d(\theta,\phi,\phi;x) \ll N^r$, it holds that

$$\begin{aligned}
&\mathbb{E}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right)\\
&= \left([\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi} + (2\log N)^{1/2}\,[\partial_{\psi'}B_d(\theta,\phi,\phi';x)]|_{\phi'=\phi}\right)(1+o(1)) \qquad (200)\\
&\mathbb{V}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = B_d(\theta,\phi,\phi;x)^2\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2(1+o(1))\\
&\quad + [\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi}^2\ O\left(\frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)}\right)\\
&\quad + \left([\partial_{\psi'}B_d(\theta,\phi,\phi';x)]|_{\phi'=\phi}\right)^2\ o\left(\log N\right)\\
&\quad + [\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi}\ [\partial_{\psi'}B_d(\theta,\phi,\phi';x)]|_{\phi'=\phi}\ O\left(\frac{(\log N)^{\frac{3}{4}}}{\sqrt{B_d(\theta,\phi,\phi;x)}}\right). \qquad (201)
\end{aligned}$$

Secondly, we show that: further assuming (33) and (34) as in assertion (i) of Theorem 5, we obtain (35). Thirdly, we show that: further assuming that $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = 0$ for d large enough as in assertion (ii) of Theorem 5, we obtain (36) and (37).

(i) **Proof of (200) and (201).** We proceed as in the first step of the proof of Theorem 4, but this time using Property ($\mathrm{P}^{\mathrm{DREP}}$) and (180). This leads to

$$\begin{aligned}
\mathbb{E}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) &= \left(\alpha + (1-\alpha)\mathbb{E}\left(\sum_{j=1}^N (\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)\right)[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi}\\
&\quad + [\partial_{\psi'}B_d(\theta,\phi,\phi';x)]|_{\phi'=\phi}\,\mathbb{E}\left(\sum_{j=1}^N [\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)} + (1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)\ ,
\end{aligned}$$

and

$$\mathbb{V}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = (1-\alpha)^2\,\left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]|^2_{\phi'=\phi}\,\mathbb{V}\left(\sum_{j=1}^N(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)$$
$$+\left(\left[\partial_{\psi'}B_d(\theta,\phi,\phi';x)\right]|_{\phi'=\phi}\right)^2\,\mathbb{V}\left(\sum_{j=1}^N[\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)$$
$$+B_d(\theta,\phi,\phi;x)^2\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2\,\mathbb{E}\left(\sum_{j=1}^N\left(\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)^2\right)$$
$$+(1-\alpha)\,\left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]|_{\phi'=\phi}\,\left[\partial_{\psi'}B_d(\theta,\phi,\phi';x)\right]|_{\phi'=\phi}$$
$$\times\,\mathbb{C}\mathrm{ov}\left(\sum_{j=1}^N(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2,\sum_{j=1}^N[\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)\,.$$

Now applying Lemma 8 with $\xi_k=\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_k;x)$ and $\beta=(1-\alpha)B_d(\theta,\phi,\phi;x)$, we get that (186) with $\delta=2$, (184) successively with $\delta=1,2$, (187) with $\delta=2$ and $q=r$, (185) with $\lambda=\alpha$, (186) successively with $\delta=2,3,4$ yield: as $N,d\to\infty$ with $\sqrt{\log N}\ll B_d(\theta,\phi,\phi;x)\ll N^r$,

$$\mathbb{E}\left(\sum_{j=1}^N(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)=1+o(1)\,,$$
$$\mathbb{E}\left(\sum_{j=1}^N[\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)=(2\log N)^{1/2}\,(1+o(1))\,,$$
$$\mathbb{V}\left(\sum_{j=1}^N(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)=O\left(\frac{\sqrt{\log N}}{B_d(\theta,\phi,\phi;x)}\right)\,,$$
$$\mathbb{V}\left(\sum_{j=1}^N[\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)=o\left(\log N\right)\,,$$
$$\mathbb{E}\left(\sum_{j=1}^N\left(\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2\right)^2\right)=1+o(1)\,.$$

Pairing up the last two equalities with the Cauchy-Schwarz inequality further imply

$$\mathbb{C}\mathrm{ov}\left(\sum_{j=1}^N(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2,\sum_{j=1}^N[\alpha\overline{\mathcal{W}}_{j,N}^{(\alpha)}+(1-\alpha)(\overline{\mathcal{W}}_{j,N}^{(\alpha)})^2]\,\widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_j;x)\right)=O\left(\frac{(\log N)^{\frac{3}{4}}}{\sqrt{B_d(\theta,\phi,\phi;x)}}\right)\,.$$

Plugging these asymptotic behaviors into the previous expressions of the expectation and variance of the DREP gradient estimator, we then obtain (200) and (201).

(ii) **Proof of (35).** The marginal log-likelihood satisfies

$$\log\mathbb{E}_{Z\sim q_{\phi'}(\cdot|x)}\left(w_{\theta,\phi}(Z;x)\right)=\log\mathbb{E}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right)=\log\mathbb{E}\left(\mathrm{e}^{\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)}\right)\,.$$

Since $\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)$ is a Gaussian variable with variance $B_d(\theta,\phi,\phi';x)^2$, we get that

$$\log \mathbb{E}\left(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right) = \mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x)) + \frac{1}{2} B_d(\theta,\phi,\phi';x)^2 \,. \tag{202}$$

Furthermore, (180) with $N=1$ yields

$$\begin{aligned}\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) = {} & \left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi} + \left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|_{\phi'=\phi} \widetilde{\Lambda}_{\theta,\phi,\phi}(\varepsilon_1;x) \\ & + B_d(\theta,\phi,\phi;x)\left[\partial_{\psi'}\left(\widetilde{\Lambda}_{\theta,\phi,\phi'}(\varepsilon_1;x)\right)\right]\Big|_{\phi'=\phi} \,. \end{aligned} \tag{203}$$

Using Property ($\mathrm{P}^{\mathrm{DREP}}$), we thus have that

$$\mathrm{SNR}\left[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right] = \frac{\left|\left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi}\right|}{\sqrt{\left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|^2_{\phi'=\phi} + B_d(\theta,\phi,\phi;x)^2\, \widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2}} \tag{204}$$

and, using moreover (202), that

$$\begin{aligned}&\mathrm{SNR}\left[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x) - \left[\partial_{\psi'}\log\mathbb{E}(\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi}\right] \\ &\qquad = \frac{\left|B_d(\theta,\phi,\phi';x)\;\left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|_{\phi'=\phi}\right|}{\sqrt{\left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|^2_{\phi'=\phi} + B_d(\theta,\phi,\phi;x)^2\, \widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2}} \,.\end{aligned}$$

From the last display, we can express (33) equivalently as

$$\sqrt{\log N} = o\left(\frac{B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)}{\left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|_{\phi'=\phi}}\right) \,. \tag{205}$$

Observing that (205) also implies that

$$\left[\partial_{\psi'} B_d(\theta,\phi,\phi';x)\right]\big|_{\phi'=\phi} = o\left(B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)\right) \,,$$

and using this condition in (204) gives us that

$$\mathrm{SNR}\left[\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right] = \frac{\left|\left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi}\right|}{B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)}\,(1+o(1)) \,, \tag{206}$$

Using (206), we can express (34) as

$$\left|\left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi}\right| \frac{(\log N)^{\frac{1}{4}}}{\sqrt{B_d(\theta,\phi,\phi;x)}} = o\left(B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)\right) \,.$$

Plugging (205) and the latter condition into (200) and (201), we obtain

$$\begin{aligned}\mathbb{E}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) &= \left[\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))\right]\big|_{\phi'=\phi}\,(1+o(1)) + o\left(B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)\right), \\ \mathbb{V}\left(\psi\text{-}g_{1,N,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) &= B_d(\theta,\phi,\phi;x)^2 \widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x)^2(1+o(1)),\end{aligned}$$

which, together with (206), imply (35).

(iii) **Proof of (36) and (37).** Suppose now that $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = 0$ for d large enough. Using (203) with Property ($\mathrm{P}^{\mathrm{DREP}}$), this variance being zero is equivalent to having

$$[\partial_{\psi'} B_d(\theta,\phi,\phi';x)]|_{\phi'=\phi} = B_d(\theta,\phi,\phi;x)\,\widetilde{B}_d^{(\mathrm{DREP})}(\theta,\phi;x) = 0\,.$$

Plugging this into (200) and (201) and since (203) implies that $\mathbb{E}\left(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)\right) = [\partial_{\psi'}\mathbb{E}(\Lambda_{\theta,\phi,\phi'}(\varepsilon;x))]|_{\phi'=\phi}$, we get (36) and (37).

C.4 Proof of Example 5

We get from Example 1 that:

$$\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x) - \ell_d(\theta;x) = -\frac{\|\theta-\phi\|^2}{2} - \|\theta-\phi\|S, \quad S = \frac{\langle\varepsilon,\phi-\theta\rangle}{\|\theta-\phi\|}. \tag{207}$$

Consequently, (B1) holds with $B_d^2(\theta,\phi,\phi;x) = \|\theta-\phi\|^2$ and differentiating (207) with respect to ϕ_k, we also obtain that

$$\phi_k\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x) = \partial_{\phi_k}\log\tilde{w}_{\theta,\phi,\phi}(\varepsilon;x) = \theta_k - \phi_k - \varepsilon^{(k)}$$

so that $\mathrm{SNR}[\phi_k\text{-}g_{1,1,d}^{(\alpha,\mathrm{REP})}(\theta,\phi;x)] = |\theta_k-\phi_k|$. Now using that $\theta = \epsilon\cdot\boldsymbol{u}_d$ and $\phi = 0\cdot\boldsymbol{u_d}$, we get that (29) is exactly (38), that (30) holds and so does ($\mathsf{B}_{\mathrm{df}}^{\mathrm{REP}}$). Applying Theorem 4, (38) then implies

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

Furthermore, since we also have from Example 1 that:

$$\log\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x) - \ell_d(\theta;x) = -\frac{\|\theta-\phi\|^2}{2} - \|\theta-\phi\|S, \quad S = \frac{\langle\varepsilon+\phi'-\phi,\phi-\theta\rangle}{\|\theta-\phi\|},$$

differentiating w.r.t ϕ_k' and taking the result at $\phi'=\phi$, we get that

$$\left[\partial_{\phi_k'}\log\tilde{w}_{\theta,\phi,\phi'}(\varepsilon;x)\right]\Big|_{\phi'=\phi} - 0 = 0 + \phi_k - \theta_k\,,$$

thus $\mathbb{V}(\psi\text{-}g_{1,1,d}^{(\alpha,\mathrm{DREP})}(\theta,\phi;x)) = 0$ and ($\mathsf{B}_{\mathrm{df}}^{\mathrm{DREP}}$) holds. As a result, (ii) from Theorem 5 applies.

D Additional numerical experiments

D.1 Gaussian experiment from Section 4.1

We provide here additional experiments for the case $\alpha = 0$. By Example 3: as $N\to\infty$,

$$\mathrm{SNR}[\phi_k\text{-}g_{M,N}^{(0,\mathrm{REP})}(\theta,\phi;x)] = \sqrt{\frac{M}{N}}\frac{\exp\left(\frac{d\epsilon^2}{2}\right)}{\sqrt{1+\epsilon^2}}(1+o(1)) \tag{208}$$

$$\mathrm{SNR}[\phi_k\text{-}g_{M,N}^{(0,\mathrm{DREP})}(\theta,\phi;x)] = \frac{\sqrt{MN}(1+o(1))}{\left(\exp\left(4d\epsilon^2\right) - 4\exp\left(2d\epsilon^2\right) + 4\exp\left(d\epsilon^2\right) - 1\right)^{\frac{1}{2}}} \tag{209}$$

We let $d\in\{10,100,500\}$, $\epsilon = 0.2$, $N\in\{2^j, j=1\dots15\}$, $M=1$ and our results are plotted on Figure 5. 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 (208) and (209) respectively match as N increases. As expected, this is no longer true as d increases.

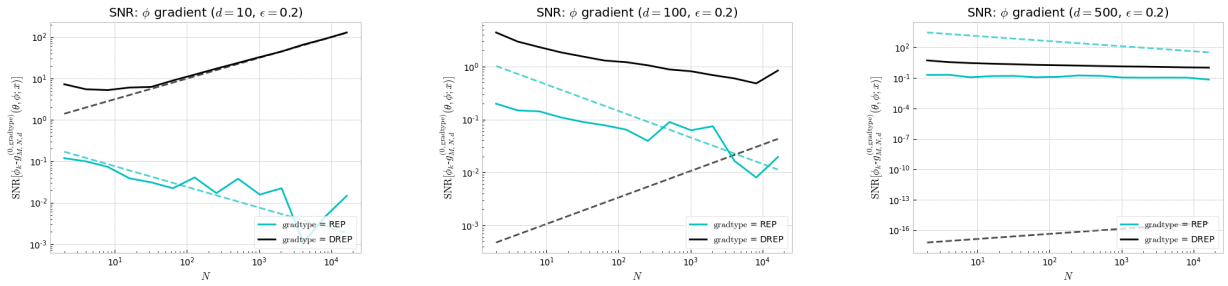

Figure 5: Plotted are $\mathrm{SNR}[\phi_k\text{-}g^{(0,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ and $\mathrm{SNR}[\phi_k\text{-}g^{(0,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Gaussian example described in Section 4.1 as a function of N and with $\epsilon = 0.2$. The solid lines correspond to the SNRs, while the dashed lines correspond to predictions of the form (208) and (209).

D.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 4: as $N \to \infty$, for all $\alpha \in [0,1)$,

$$\mathrm{SNR}[\theta_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N}(\theta,\phi;x)] = \sqrt{MN}\,\frac{\left|\frac{x_k-\theta_k}{2}+\frac{3\epsilon\alpha}{4-\alpha}\right|(1+o(1))}{\frac{(4-\alpha)^{d/2}}{(15-6\alpha)^{d/4}}\exp\left(\frac{12(1-\alpha)^2}{(4-\alpha)(5-2\alpha)}d\epsilon^2\right)\sqrt{\frac{2}{5-2\alpha}+\left(\frac{12(1-\alpha)\epsilon}{(5-2\alpha)(4-\alpha)}\right)^2}} \tag{210}$$

$$\mathrm{SNR}[\tilde{b}_k\text{-}g^{(0,\mathrm{DREP})}_{M,N}(\theta,\phi;x)] = \sqrt{MN}\,\frac{\frac{24\epsilon 4^{d-1}\exp\left(\frac{24d\epsilon^2}{4\cdot 5}\right)}{3^{d/2}5^{\frac{d}{2}+1}}}{\sqrt{\tilde{b}_k\text{-}V^{(0,\mathrm{DREP})}(\theta,\phi;x)}}(1+o(1)) \tag{211}$$

We let $d \in \{10, 100, 500\}$, $\epsilon \in \{0.2, 1\}$, $N \in \{2^j, 1 \ldots 15\}$, $M = 1$ and our results are plotted on Figures 6 and 7. 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 (210) and (211) 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 (210).

References

Gordon, R. D. (1941). "Values of Mills' Ratio of Area to Bounding Ordinate and of the Normal Probability Integral for Large Values of the Argument". In: *The Annals of Mathematical Statistics* 12.3, pp. 364–366.

Pickands III, J. (1968). "Moment Convergence of Sample Extremes". In: *The Annals of Mathematical Statistics* 39.3, pp. 881–889.

Billingsley, P. (1971). *Weak Convergence of Measures*. Society for Industrial and Applied Mathematics.

Cressie, N., A. S. Davis, J. L. Folks, and G. E. Policello II (1981). "The moment-generating function and negative integer moments". In: *Amer. Statist.* 35.3, pp. 148–150.

L'Ecuyer, P. (1995). "On the Interchange of Derivative and Expectation for Likelihood Ratio Derivative Estimators". In: *Management Science* 41.4, pp. 738–748.

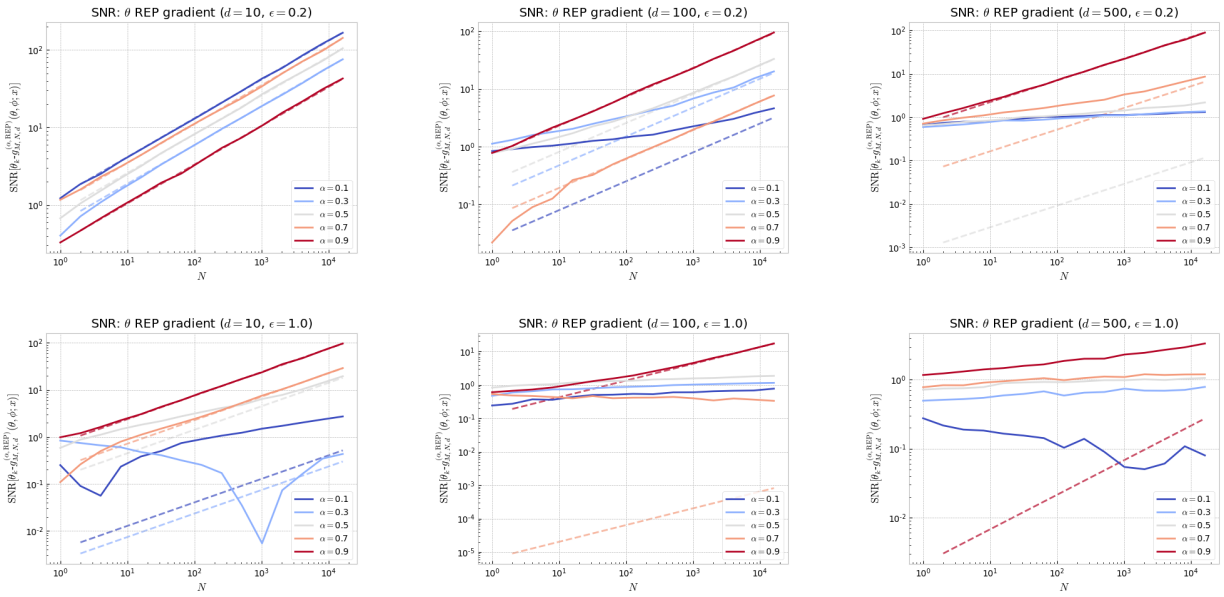

Figure 6: Plotted is $\mathrm{SNR}[\theta_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, for varying values of (α, d, ϵ) and a randomly selected datapoint x. The solid lines correspond to $\mathrm{SNR}[\theta_k\text{-}g^{(\alpha,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$, while the dashed lines correspond to predictions of the form (210).

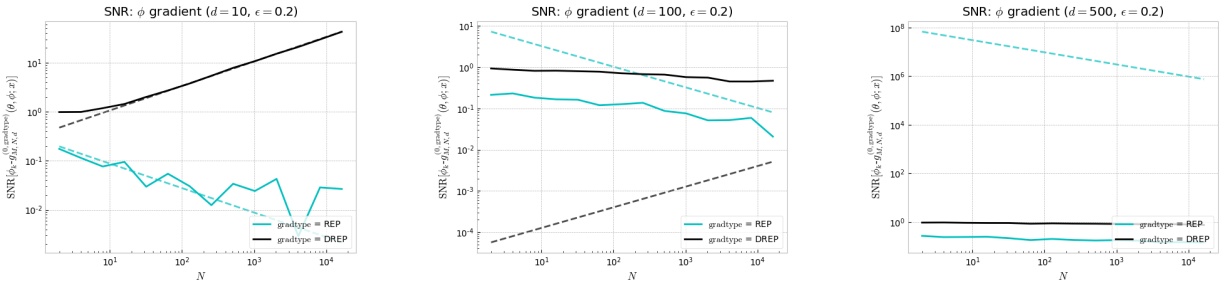

Figure 7: Plotted are $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(0,\mathrm{REP})}_{M,N,d}(\theta,\phi;x)]$ and $\mathrm{SNR}[\tilde{b}_k\text{-}g^{(0,\mathrm{DREP})}_{M,N,d}(\theta,\phi;x)]$ computed over 2000 Monte Carlo samples for the Linear Gaussian example described in Section 4.2 as a function of N, with $\epsilon = 0.2$ and for a randomly selected datapoint x. The solid lines correspond to the SNRs, while the dashed lines correspond to predictions of the form (210) and (211).

Petrov, V. V. (1995). *Limit theorems of probability theory*. Vol. 4. Oxford Studies in Probability. Sequences of independent random variables, Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, pp. xii+292.

Jordan, M., Z. Ghahramani, T. Jaakkola, and L. Saul (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37, pp. 183–233.

Haan, L. de and A. Ferreira (2007). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer New York. ISBN: 9780387344713.

Kingma, D. P. and M. Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*.

Burda, Y., R. Grosse, and R. Salakhutdinov (2016). "Importance weighted autoencoders". In: *4th International Conference on Learning Representations (ICLR)*.

Li, Y. and R. E. Turner (2016). "Rényi Divergence Variational Inference". In: *Advances in Neural Information Processing Systems*. Vol. 29.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877.

Dieng, A. B., D. Tran, R. Ranganath, J. Paisley, and D. Blei (2017). "Variational Inference via χ-Upper Bound Minimization". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Li, Y. and Y. Gal (2017). "Dropout Inference in Bayesian Neural Networks with Alpha-divergences". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70, pp. 2052–2061.

Maddison, C. J., J. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. Teh (2017). "Filtering Variational Objectives". In: *Advances in Neural Information Processing Systems 30*, pp. 6573–6583.

Roeder, G., Y. Wu, and D. K. Duvenaud (2017). "Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Domke, J. and D. R. Sheldon (2018). "Importance Weighting and Variational Inference". In: *Advances in Neural Information Processing Systems 31*, pp. 4470–4479.

Douc, R., E. Moulines, P. Priouret, and P. Soulier (2018). *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, pp. xviii+757.

Rainforth, T., A. Kosiorek, T. A. Le, C. Maddison, M. Igl, F. Wood, and Y. W. Teh (2018). "Tighter Variational Bounds are Not Necessarily Better". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80, pp. 4277–4285.

Wang, D., H. Liu, and Q. Liu (2018). "Variational Inference with Tail-adaptive f-Divergence". In: *Advances in Neural Information Processing Systems*. Vol. 31.

Tucker, G., D. Lawson, S. S. Gu, and C. J. Maddison (2019). "Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives". In: *Proceedings of the 7th International Conference on Learning Representations*.

Geffner, T. and J. Domke (2020). "Empirical Evaluation of Biased Methods for Alpha Divergence Minimization". In: *3rd Symposium on Advances in Approximate Bayesian Inference*.

Daudel, K. and R. Douc (2021). "Mixture weights optimisation for Alpha-Divergence Variational Inference". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 4397–4408.

Daudel, K., R. Douc, and F. Portier (2021). "Infinite-dimensional gradient-based descent for alpha-divergence minimisation". In: *The Annals of Statistics* 49.4, pp. 2250–2270.

Dhaka, A. K., A. Catalina, M. Welandawe, M. R. Andersen, J. H. Huggins, and A. Vehtari (2021). "Challenges and Opportunities in High-dimensional Variational Inference". In: vol. 34. Neural information processing systems foundation, pp. 7787–7798.

Geffner, T. and J. Domke (2021). "On the difficulty of unbiased alpha divergence minimization". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139, pp. 3650–3659.

Rudner, T. G., O. Key, Y. Gal, and T. Rainforth (2021). "On signal-to-noise ratio issues in variational inference for deep Gaussian processes". In: *International Conference on Machine Learning*. PMLR, pp. 9148–9156.

Guilmeau, T., E. Chouzenoux, and V. Elvira (2022). "Regularized R\'enyi divergence minimization through Bregman proximal gradient algorithms". In: *arXiv:2211.04776*.

Knoblauch, J., J. Jewson, and T. Damoulas (2022). "An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference". In: *Journal of Machine Learning Research* 23.132, pp. 1–109.

Rodríguez-Santana, S. and D. Hernández-Lobato (2022). "Adversarial α-divergence minimization for Bayesian approximate inference". In: *Neurocomputing* 471, pp. 260–274.

Daudel, K., J. Benton, Y. Shi, and A. Doucet (2023a). "Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics". In: *Journal of Machine Learning Research* 24.243, pp. 1–83.

Daudel, K., R. Douc, and F. Roueff (2023b). "Monotonic Alpha-divergence Minimisation for Variational Inference". In: *Journal of Machine Learning Research* 24.62, pp. 1–76.

Guilmeau, T., N. Branchini, E. Chouzenoux, and V. Elvira (2024). "Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization". In: *International Conference on Artificial Intelligence and Statistics*, pp. 3871–3879.

Margossian, C. C., L. Pillaud-Vivien, and L. K. Saul (2024). "Variational Inference for Uncertainty Quantification: an Analysis of Trade-offs". In: *arXiv:2403.13748*.