On the Difficulty of Unbiased Alpha Divergence Minimization

Tomas Geffner [1] **Justin Domke** [1]

Abstract

Several approximate inference algorithms have been proposed to minimize an alpha-divergence between an approximating distribution and a target distribution. Many of these algorithms introduce bias, the magnitude of which becomes problematic in high dimensions. Other algorithms are unbiased. These often seem to suffer from high variance, but little is rigorously known. In this work we study unbiased methods for alpha-divergence minimization through the Signal-to-Noise Ratio (SNR) of the gradient estimator. We study several representative scenarios where strong analytical results are possible, such as fully-factorized or Gaussian distributions. We find that when alpha is not zero, the SNR worsens exponentially in the dimensionality of the problem. This casts doubt on the practicality of these methods. We empirically confirm these theoretical results.

1. Introduction

Variational inference (VI) typically minimizes the KL-divergence from an approximating distribution q_w to a target distribution p (Jordan et al., 1999; Blei et al., 2017; Zhang et al., 2017). While computationally convenient, the use of this objective may lead to distributions q_w with undesirable statistical properties (e.g. variance underestimation (Minka, 2005)). To avoid this, recent methods instead attempt to minimize an alpha-divergence (Amari, 1985). This is a class of divergences indexed by a parameter α, that reduces to the typical KL-divergence when $\alpha \to 0$. The α parameter determines the divergence's properties. For instance, for $\alpha \gg 0$, the divergence penalizes distributions q_w that place no mass in regions where p does, penalizing variance underestimation. For many use-cases, approximating distributions minimizing these divergences would be more useful.

[1]College of Information and Computer Science, University of Massachusetts, Amherst, MA, USA. Correspondence to: Tomas Geffner <tgeffner@cs.umass.edu>, Justin Domke <domke@cs.umass.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Existing alpha-divergence minimization algorithms can be classified into two broad groups: biased methods (Li & Turner, 2016; Regli & Silva, 2018) and unbiased methods (Kuleshov & Ermon, 2017; Dieng et al., 2017). While some positive empirical results have been obtained using biased methods, it has been recently observed that, in problems with high dimensionality, these often fail to minimize the target alpha-divergence (Geffner & Domke, 2021).[1] Thus, in this paper we turn our attention to unbiased methods. These attempt to minimize an alpha-divergence by running stochastic optimization algorithms with unbiased estimators of the divergence's gradient.

The major open question for unbiased methods concerns the difficulty of the optimization problem. Too much variability in the gradient estimator would force a very small step-size and thus a huge number of optimization steps. While some positive empirical results have been observed, authors have cautioned the the gradient estimators used can be temperamental (Kuleshov & Ermon, 2017; Dieng et al., 2017). It is currently unclear when these methods will succeed, and how their performance depends on α or the dimensionality of the problem.

We address this question. Informally, our main conclusion is that methods based on unbiased gradient estimates of an alpha-divergence will often require catastrophically large amounts of computation to scale to high dimensional models for any $\alpha \neq 0$. This is not always due to high variance gradients, but to an extremely low Signal-to-Noise ratio (SNR).

In Section 3 we present two gradient estimators used by unbiased methods, one obtained using reparameterization, and a novel one obtained by applying "double" reparameterization. An empirical evaluation shows that, even in simple scenarios, these methods do not seem scale to problems of moderately high dimension ($d \approx 100$), nor to moderately large values of α ($\alpha \approx 0.4$). Curiously, optimization fails even in cases where the gradient estimator has provably low variance. Instead, we propose that this failure is best explained by the estimator's (SNR), which is known to be related to optimization convergence (Section 4.3).

[1]In fact, it was observed that, in high dimensions, biased methods often just minimize the typical VI objective, the KL divergence from q_w to p, regardless of the target alpha-divergence chosen.

The main contribution of this paper is a theoretical analysis of the gradient estimators' SNR, given in Section 4. We analyze two representative scenarios: The first is when the target and approximating family are arbitrary fully-factorized distributions. The second is when the target and approximating family are both full-rank Gaussians. We give exact results and bounds for the SNR in these cases. We show that for *any* $\alpha \neq 0$, under mild assumptions, the SNR decreases *exponentially* in the dimensionality of the problem. Thus, even in these seemingly "easy" scenarios, unbiased methods will not scale to high dimensional problems. Finally, in Section 5 we empirically confirm that the same phenomena seems to occur in real problems.

Our results are pessimistic. Ideally, one might hope to guarantee a good SNR under some favorable assumptions about the target. For example, one might hope that the SNR could be guaranteed to be reasonably large if the log-posterior obeyed some common regularity conditions, e.g. that it were fully-factorized, concave, strongly concave, Lipschitz smooth, Gaussian, or even a fully-factorized Gaussian. Our results show that, for general alpha-divergences, no such guarantee is possible.

We do not rule out the possibility of a good SNR guarantee under some other assumptions about the target. However, these assumptions would have to be *stronger* than any of those listed above, and also *prohibit* the cases we typically think of as easy, e.g. fully-factorized or Gaussian distributions. This suggests that a general-purpose algorithm for optimizing an alpha-divergence based on currently available unbiased gradient estimators may be unachievable.

2. Preliminaries

Variational Inference (VI) is an approximate inference algorithm that finds w such that the approximating distribution $q_w(z)$ is close to some target distribution $p(z)$. This is usually done by minimizing $\mathrm{KL}(q_w||p)$. In most cases the gradient of this objective with respect to w cannot be computed exactly. However, unbiased estimates are often available. Two popular alternatives are reparameterization (Titsias & Lázaro-Gredilla, 2014; Kingma & Welling, 2013; Rezende et al., 2014) and the "sticking the landing" (STL) estimator (Roeder et al., 2017). Both require a mapping $\mathcal{T}_w$ that transforms a base density q_0 into q_w. Then, the estimators are computed as

$$\begin{aligned} g^{\mathrm{rep}}(p, q_w, \epsilon) &= \nabla_w \log \tfrac{q_w(\mathcal{T}_w(\epsilon))}{p(\mathcal{T}_w(\epsilon))} \\ g^{\mathrm{STL}}(p, q_w, \epsilon) &= \nabla_w \log \tfrac{q_v(\mathcal{T}_w(\epsilon))}{p(\mathcal{T}_w(\epsilon))}\Big|_{v=w}, \end{aligned} \tag{1}$$

where $\epsilon \sim q_0(\epsilon)$. While both estimators have shown good empirical performance, it has been observed that g^{STL} often leads to better results (Roeder et al., 2017). Also, it has the desirable property of being deterministically zero at the optimum $p = q_w$, which is not true for the reparameterization estimator.

While minimizing $\mathrm{KL}(q_w||p)$ is computationally convenient, it may lead to distributions q_w with undesirable properties. For instance, the resulting q_w tends to underestimate the variance of p (Minka, 2005). This is problematic, for instance, if q_w will be used as a proposal distribution to estimate the expectation $\mathbb{E}_p f(z)$ with importance sampling. An under-dispersed distribution q_w leads to importance weights with high variance (Owen, 2013). For reasons like this, recent work has developed methods to minimize other divergence measures (Minka, 2004; Hernández-Lobato et al., 2016; Li & Turner, 2016; Kuleshov & Ermon, 2017; Dieng et al., 2017; Regli & Silva, 2018; Wang et al., 2018; Wan et al., 2020; Naesseth et al., 2020).

Alpha-divergences may be used as an objective for VI. The alpha-divergence between distributions p and q_w is given by

$$D_\alpha(p||q_w) = \frac{1}{\alpha(\alpha-1)} \mathop{\mathbb{E}}_{q_w}\left[\left(\frac{p(z)}{q_w(z)}\right)^\alpha - 1\right], \quad \alpha \in \mathbb{R}\backslash\{0,1\}. \tag{2}$$

For $\alpha \gg 0$ the divergence will penalize distributions q_w that place no mass in regions where p does, penalizing under-dispersion. For instance, minimizing $D_\alpha(p||q_w)$ for $\alpha = 2$ is equivalent to minimizing the variance of the importance weights $p(z)/q_w(z)$. Also, alpha-divergences recover several well known divergences for different values of α (Cichocki & Amari, 2010): the χ^2-divergence for $\alpha = 2$, and the Hellinger distance for $\alpha = 0.5$. For $\alpha = 0$ it is defined as the limit $\alpha \to 0$, which result in $\mathrm{KL}(q_w||p)$, the divergence typically used for VI. Algorithms to minimize alpha-divergences can be classified into two groups:

Biased methods (Minka, 2004; Bornschein & Bengio, 2014; Hernández-Lobato et al., 2016; Li & Turner, 2016; Regli & Silva, 2018). These either minimize local surrogates and/or use biased gradient estimators. Therefore, the distributions q_w returned by these methods are not minimizers of $D_\alpha(p||q_w)$. Geffner & Domke (2021) present an extensive empirical evaluation of methods based on biased gradients showing that, in high dimensions, these methods often fail to minimize the target alpha-divergence; they return suboptimal distributions q_w that heavily underestimate the variance of p. This is problematic, since one of the goals of using alpha-divergences is to avoid under-dispersion.

Unbiased methods (Kuleshov & Ermon, 2017; Dieng et al., 2017). These methods were developed for the χ^2-divergence. However, as mentioned by the authors, the same approach can be used for $\alpha \neq 2$. These methods attempt to minimize $D_\alpha(p||q_w)$ exactly by running SGD with an unbiased estimator of $\nabla_w D_\alpha(q_w||p)$. Under an appropriate choice for the step-size this is guaranteed to converge

(Bottou et al., 2018). As mentioned previously, these methods often present scalability issues, and it is unclear when they may be successfully applied.[2]

The **signal-to-noise ratio** (SNR) can be used as a measure of an estimator's quality. We define the SNR of a random vector X as

$$\mathrm{SNR}[X] = \frac{\|\mathbb{E}X\|^2}{\mathbb{E}\|X\|^2} = \frac{\|\mathbb{E}X\|^2}{\|\mathbb{E}X\|^2 + \mathbb{V}\|X\|}, \tag{3}$$

which is always less than or equal to one, with equality holding if and only if the variance of the random variable is zero. If some of the expectations do not exist, the SNR is not defined. We will use this quantity to evaluate the quality of an estimator. This has been previously done in the context of importance weighted auto-encoders (Rainforth et al., 2018).

3. Gradient Estimators

We present two unbiased estimators for $\nabla_w D_\alpha(p||q_w)$. The first one, obtained via reparameterization, is given by (Dieng et al., 2017)

$$g_\alpha^{\mathrm{rep}}(p, q_w, \epsilon) = \begin{cases} \frac{1}{\alpha^2 - \alpha} \nabla_w \left(\frac{p(\mathcal{T}_w(\epsilon))}{q_w(\mathcal{T}_w(\epsilon))} \right)^\alpha & \text{if } \alpha \notin \{0, 1\} \\ \nabla_w \log \frac{q_w(\mathcal{T}_w(\epsilon))}{p(\mathcal{T}_w(\epsilon))} & \text{if } \alpha \to 0, \end{cases} \tag{4}$$

where $\epsilon \sim q_0(\epsilon)$. For $\alpha \to 0$ this estimator recovers g^{rep} from eq. 1. As presented, this estimator is not defined for $\alpha \to 1$. A second unbiased estimator, obtained by applying reparameterization twice, is given by

$$g_\alpha^{\mathrm{drep}}(p, q_w, \epsilon) = \begin{cases} -\frac{1}{\alpha} \nabla_w \left(\frac{p(\mathcal{T}_w(\epsilon))}{q_v(\mathcal{T}_w(\epsilon))} \right)^\alpha \Big|_{v=w} & \text{if } \alpha \neq 0 \\ \nabla_w \log \frac{q_v(\mathcal{T}_w(\epsilon))}{p(\mathcal{T}_w(\epsilon))} \Big|_{v=w} & \text{if } \alpha \to 0. \end{cases} \tag{5}$$

This estimator is novel. We show its derivation in Appendix C. For $\alpha \to 0$ it recovers g^{STL} from eq. 1. For $\alpha \neq 0$ it is derived by applying the reparameterization trick twice. This is similar to the derivation for the "doubly-reparameterized" gradient estimator (Tucker et al., 2018) for importance weighted auto-encoders (Burda et al., 2016). A third unbiased estimator may be obtained via the score function method (Williams, 1992). In this work we do not focus on this one since it has been consistently observed that reparameterization estimators outperform their score function counterparts.

In practice, we observed that g_α^{drep} often works better than g_α^{rep} (Appendix A shows an empirical comparison). This may not be surprising since (i) g_α^{drep} is a natural extension of g^{STL}, which often works better than g^{rep} when $\alpha \to 0$; (ii) g_α^{drep} has the property of being deterministically zero at the optimum $p = q_w$, which is not true for g_α^{rep}; (iii) The use of double reparameterization has led to significant improvements over "plain" reparameterization for multi-samples objectives (Tucker et al., 2018).

[2] While in their simulations (Dieng et al., 2017) use a biased algorithm, their whole formulation was carried out for unbiased divergence minimization.

Empirical evaluation. We now present empirical results that motivate this work. These demonstrate two important phenomena. First, for larger α, optimization scales poorly to high dimensions. Understanding this is the central goal of this paper. Second, this may happen even when the gradient estimator's variance is very small. This explains why we study SNR rather than "raw" variance.

We set p to be a standard Gaussian in d dimensions and q_w to be a mean-zero fully-factorized Gaussian. The parameters of q_w are $w = \sigma \in \lambda^d$, representing the standard deviation of each dimension of q_w. We initialize $\sigma_i = 2$, and optimize $D_\alpha(p||q_w)$ from eq. 2. We do so by running SGD with the gradient estimator g_α^{drep} for 1000 steps. (Appendix A shows results using g_α^{rep}, which are worse.)

We perform this optimization for three different dimensionalities, $d \in \{8, 32, 128\}$, for $\alpha \in \{0, 0.4, 0.9, 1.5\}$, and for gradient estimators obtained averaging N samples, for $N \in \{1, 10, 10^2, 10^3, 10^4\}$. For each triplet (d, α, N) we tuned the step-size; we ran simulations for all step-sizes in the set $\{10^i\}_{i=-7}^{7}$ and selected the one that lead to the best final performance. All results are averages over 15 simulations.

Fig. 1 shows the results. Optimization succeeds when the dimensionality d is small or α is small. Indeed, for $\alpha \to 0$, optimization converges in approximately 30 steps, regardless of the dimensionality and the number of samples used to estimate each gradient. However, when α is larger, increasing d seems to cause major difficulties. For instance, for $d = 32$ and $\alpha = 1.5$, optimization does not meaningfully converge within 1000 steps, even using 10^4 samples to estimate gradients. Furthermore, for $d = 128$ and $\alpha \neq 0$ optimization barely makes any progress regardless of the number of samples used to estimate gradients.

Optimization results using Adam (Kingma & Ba, 2014) are shown in Fig. 7 (Appendix B). The same effect is observed; optimization converges properly when the dimensionality or α are low, but fails in high dimensions for larger values of α. (We include a brief discussion of Adam's performance in Section 6.)

For $\alpha = 2$, previous work has attributed the scaling issues of unbiased methods to the use of gradient estimates with high variance (Kuleshov & Ermon, 2017; Dieng et al., 2017). While correct in spirit, care is needed. Some of the failures in Fig. 1 happen despite *low*-variance gradients, because the true gradient is even smaller. Take $d = 128$, and let $\sigma_i = \sigma$

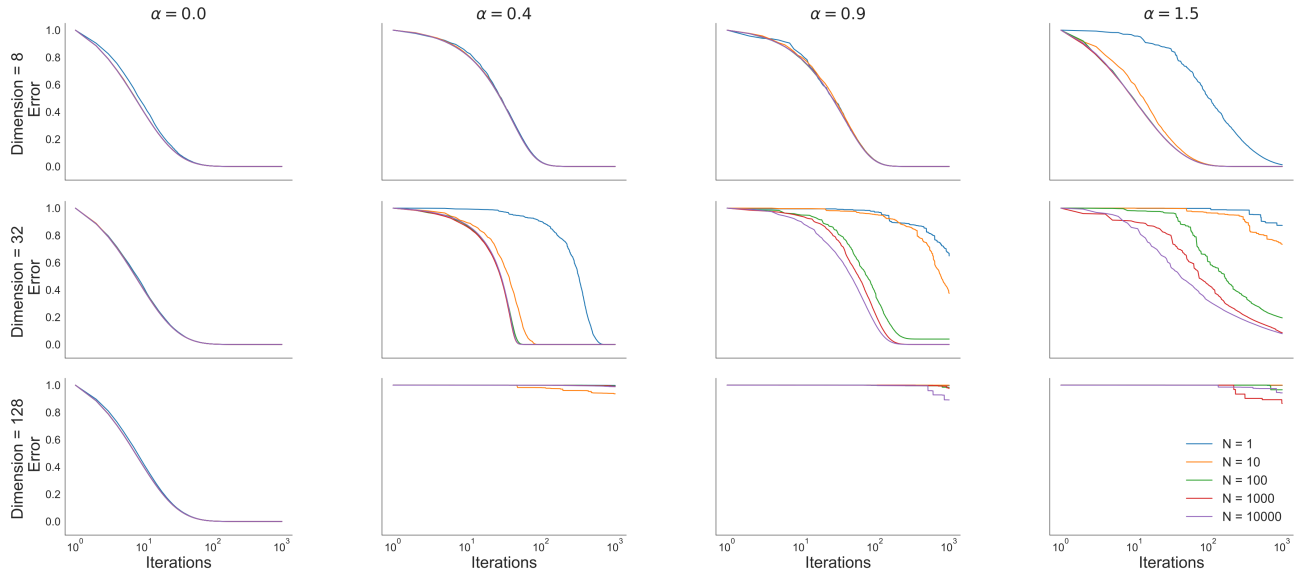

Figure 1. **Running SGD with unbiased gradient estimates becomes inefficient for moderately large dimensions and α.** Each plot shows the optimization results for each pair (d, α) for all values of N considered. Plots show "Error vs. Iteration", where error is the normalized distance between the scale parameters of p and q_w, computed as $(1/d)\sum_i(\sigma_{qi} - 1)^2$.

for all i, so that q_w is isotropic. Fig. 2 (left) shows the variance of the gradient estimator for different values of σ and α. We see that increasing α sometimes decreases gradient variance. Instead, we attribute optimization's scaling issues to the estimator's SNR. As shown in Fig. 2 (center) we see that this decreases *very* rapidly with higher α.[3]

4. SNR Analysis

In this section we present a detailed analysis of the SNR of gradient estimators for two general and representative scenarios. First, we consider the case where p and q_w are arbitrary fully-factorized distributions. Second, we consider the case where p and q_w are Gaussians with an arbitrary full-rank covariance matrices. This case is particularly relevant, since Gaussians are a good approximation of a huge range of posteriors (Bayesian central limit theorem). We show that, in both cases, for $\alpha \neq 0$ the gradient estimator's SNR becomes very small for problems with high dimensionality d. In fact, we present examples for which the SNR decreases *exponentially* in d. In contrast to this, we show that, for $\alpha \to 0$ (typical VI), the SNR of the estimator decreases at most as $1/d$, and does not depend on d if both p and q_w are factorized. Intuitively, a low SNR means that the level of noise present in the estimator is considerably larger than the "learning signal", which difficults optimization.

[3] In Appendix E we give an upper bound for the variance, and show that the variance of g_α^{drep} for $\alpha = 0.4$ becomes "small" for high dimensional problems. Surprisingly, the variance of each component of the estimator decreases as the dimension increases.

We proceed in a similar way for all scenarios considered. We first present a rigorous result, and then give a simple and intuitive interpretation and examples.

4.1. Fully-Factorized Distributions

We begin by studying the case where both p and q_w are arbitrary fully-factorized distributions. Of course, if we knew that p is fully factorized, it would make sense to perform inference on each component separately. The point of examining this case is the insight it gives us into how gradient estimators behave as dimensionality changes. For simplicity, we assume that there is one parameter w_i for each coordinate z_i. This assumption is used to simplify notation and can be removed, as long as each component is determined by disjoint sets of parameters.

Theorem 1. *Let* $p(z) = \prod_{i=1}^d p_i(z_i)$, $q_w(z) = \prod_{i=1}^d q_{w_i}(z_i)$, *and* $g_\alpha \in \{g_\alpha^{\text{rep}}, g_\alpha^{\text{drep}}\}$.

If $p_j \neq q_{w_j}$ *and* g_α *has finite variance, the SNR of the* j*-th component of the estimator,* $g_{\alpha j}$*, is given by*

$$\text{SNR}[g_{\alpha j}(p, q_w, \epsilon)] = \text{SNR}\left[g_\alpha(p_j, q_{w_j}, \epsilon_j)\right] \tag{6}$$

for $\alpha \to 0$*, and by*

$$\text{SNR}[g_{\alpha j}(p, q_w, \epsilon)] = \text{SNR}\left[g_\alpha(p_j, q_{w_j}, \epsilon_j)\right] \prod_{\substack{i=1 \\ i\neq j}}^{d} \text{SNR}\left[\tilde{D}_\alpha(p_i, q_{w_i}, \epsilon_i)\right] \tag{7}$$

for $\alpha \neq 0$*, where* $\tilde{D}_\alpha(p_i, q_{w_i}, \epsilon_i) = \left(\frac{p_i(\mathcal{T}_{w_i}(\epsilon_i))}{q_{w_i}(\mathcal{T}_{w_i}(\epsilon_i))}\right)^\alpha$ *is an unbiased estimator of* $\alpha(\alpha-1)D_\alpha(p_i||q_{w_i}) + 1$.

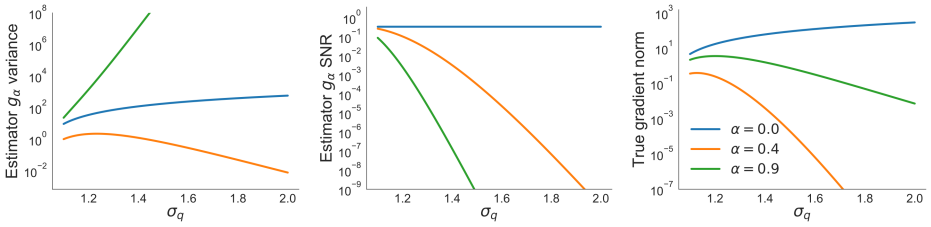

Figure 2. **Optimization difficulty is explained not by high variance but by low SNR.** p is a standard Gaussian in $d = 128$ dimensions and q_w is an isotropic Gaussian with standard deviation σ_q. Left: Variance of gradient estimator. Center: SNR. Right: Squared norm of the true gradient, $\mathbb{E}[g]$.

If $p_j = q_{w_j}$, the SNR is 0 for g_α^{rep} and is undefined for g_α^{drep} (because g_α^{drep} is deterministically zero).

To clarify the Theorem's notation, $g_{\alpha j}(p, q_w, \epsilon)$ is the j-th component of the estimator g_α for the vector $\nabla_w D_\alpha(p\|q_w)$. On the other hand, $g_\alpha(p_j, q_{w_j}, \epsilon_j)$ is the estimator for the scalar quantity $\nabla_{w_j} D_\alpha(p_j\|q_{w_j})$, the derivative (with respect to w_j) of the divergence between the one dimensional distributions p_j and q_{w_j}.

What does the theorem say? If $\alpha \to 0$ (eq. 6), the SNR of each component of the estimator consists on a single term, which is the same as if inference were performed on each dimension of p separately. That is, the SNR of the estimator's j-th component only depends on p_j and q_{w_j}, and is not affected by the dimensionality of the problem d in any way. In contrast, if $\alpha \neq 0$ (eq. 7), there are $d-1$ additional terms. These determine how the SNR scales with dimensionality. Since these terms can be expressed as the SNR of an estimator for $D_\alpha(p_i\|q_{w_i})$ (for each $i \neq j$, up to scaling constants), each of them is at most one, with equality only if $p_i = q_{w_i}$. Thus, for $\alpha \neq 0$, discrepancies in several dimensions of p and q_w accumulate (as products of terms strictly smaller than one), leading to a large detrimental effect on the estimator's SNR. Intuitively, the larger the dimensionality of the problem d, the worse this effect becomes.

An example that clearly illustrates this curse of dimensionality is given by the case where p and q are isotropic distributions. Suppose each component of p and q_w are the same, that is, $p_i = p_1$ and $q_{w_i} = q_{w_1}$ for all i. Following eq. 7, if $\alpha \neq 0$, the SNR of the j-th component of the gradient estimator is given by

$$\text{SNR}\left[g_\alpha(p_1, q_{w_1}, \epsilon_1)\right] \left(\text{SNR}\left[\tilde{D}_\alpha(p_1, q_{w_1}, \epsilon_1)\right]\right)^{d-1}, \quad (8)$$

which worsens *exponentially* in d. In contrast, if $\alpha \to 0$, the SNR does not depend on d at all (eq. 6).

4.1.1. Fully-Factorized Gaussians

As a second example of fully-factorized distributions, we consider the case where p and q_w are d-dimensional diagonal Gaussians with mean zero. The parameters are the standard deviations of each component of q_w, i.e. $w = \{\sigma_{q1}, \ldots, \sigma_{qd}\}$. In this case we can compute each term in eq. 7 in closed form.

Corollary 2. *Let p and q be two fully-factorized d-dimensional Gaussian distributions with mean zero and variances σ_{pi}^2 and σ_{qi}^2. Let $\lambda_i = \sigma_{qi}^2/\sigma_{pi}^2$ and $g_\alpha = g_\alpha^{\text{drep}}$.*

If $\lambda_j \neq 1$ and $1 + 2\alpha(\lambda_i - 1) > 0$ for all i,

$$\text{SNR}[g_{\alpha j}(p, q_w, \epsilon)] = \underbrace{\frac{1 + 2\alpha(\lambda_j - 1)}{3} f(\lambda_j, \alpha)^3}_{\text{SNR}\left[g_\alpha(p_j, q_{w_j}, \epsilon_j)\right]} \prod_{\substack{i=1 \\ i \neq j}}^{d} \underbrace{f(\lambda_i, \alpha)}_{\text{SNR}\left[\tilde{D}_\alpha(p_i, q_{w_i}, \epsilon_i)\right]} \quad (9)$$

where

$$f(\lambda, \alpha) = \frac{1}{\sqrt{1 + \alpha^2 \frac{(\lambda-1)^2}{1+2\alpha(\lambda-1)}}}. \quad (10)$$

Otherwise, the SNR is not defined. If $\lambda_j = 1$ this is because the estimator $g_{\alpha j}$ is 0 deterministically. If $1+2\alpha(\lambda_i-1) \leq 0$ for any i, this is because the estimator has infinite variance.

Corollary 2 gives conditions under which the SNR of the gradient estimator is well-defined (i.e. estimator has finite variance), and gives an expression for the SNR in such cases. In order to understand this expression, it is important to understand the behavior of the function $f(\lambda, \alpha)$. Fig. 3 shows a visualization. It can be observed that (i) $f(\lambda, \alpha)$ achieves its maximum value of 1 if and only if $\lambda = 1$ or $\alpha = 0$; and (ii) $f(\lambda, \alpha)$ decreases as α moves aways from 0 or λ moves away from 1. We present a formal characterization of this function in Lemma 5 (Appendix D.1).

Again, the behavior of the SNR for problems with high dimensionalities is determined by the d-term product $\prod_i f(\lambda_i, \alpha)$ in eq. 9. We see that, if $\alpha \to 0$, each term in this product is just one (because $f(\lambda, 0) = 1$), and thus the SNR is just $1/3$. On the other hand, if $\alpha \neq 0$, each of the terms is at most one (with equality only if the corresponding $\lambda_i = 1$). Therefore, if $\alpha \neq 0$, discrepancies in several dimensions of p and q_w accumulate, leading to a large detrimental effect on the SNR of *every* component

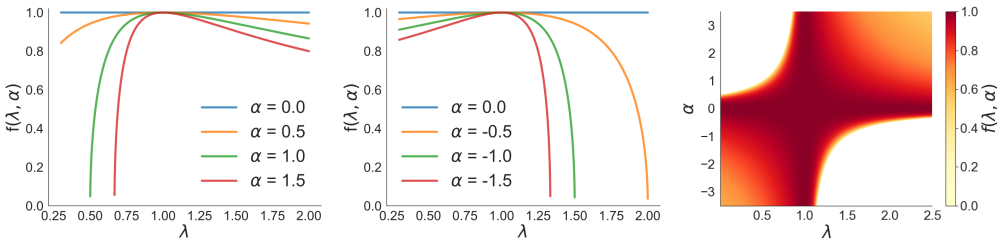

Figure 3. $f(\lambda, \alpha)$. White (rightmost plot) indicates regions where $1 + 2\alpha(\lambda - 1) > 0$ is not satisfied (in this region the estimator has infinite variance, and thus the SNR is not defined).

of the estimator. In addition, in this case we can exactly quantify this deterioration in terms of α and λ_i: The SNR worsens for α values with large absolute value and when discrepancies between components is large (i.e. λ_i far from 1), since for these cases $f(\lambda, \alpha)$ is significantly less than one (see Fig. 3).

In addition, there are entirely non-pathological cases for which the estimator has *infinite* variance. This occurs whenever the condition $1 + 2\alpha(\lambda_i - 1) > 0$ is not satisfied for some i. For instance, this happens if we set $\alpha = 1$, $\sigma_{pi} = 1$ and $\sigma_{qi} = 0.7$. More generally, the condition is equivalent to $\alpha\sigma_q^2 > (\alpha - 1/2)\sigma_p^2$. This is always satisfied for $0 \leq \alpha < 1/2$. If $\alpha \geq 1/2$, this means that the variance of q_w cannot be much smaller than that of p. If $\alpha < 0$, this means that the variance of q_w cannot be much larger than that of p.

Example. Consider the case where $\alpha = 0.4$, p is a standard Gaussian with dimension $d = 128$, and q_w is a mean zero factorized Gaussian with $\sigma_{qi} = 2$ for all i. Eq. 9 yields $\mathrm{SNR}[g_{\alpha j}] \approx 1.2 \times 10^{-10}$. This means that the variance of the estimator is approximately 8×10^9 times larger than the actual signal. In contrast, for $\alpha \to 0$, the SNR is just $1/3$. Obtaining an estimator with a similar SNR for $\alpha = 0.4$ would require averaging $N \approx 4 \times 10^9$ independent samples (this quantity grows exponentially if problems of larger dimensionality are considered).

4.2. Gaussians with arbitrary Covariances

We now move away from factorized distributions and consider the case in which both p and q_w are d-dimensional Gaussians with mean zero and arbitrary full-rank covariances Σ_p and Σ_q. The set of parameters is given by $w = S$, where S is a matrix such that $SS^\top = \Sigma_q$, and reparameterization is given by $\mathcal{T}_w(\epsilon) = S\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

Theorem 3. *Let $p(z) = \mathcal{N}(z|0, \Sigma_p)$ and $q(z) = \mathcal{N}(z|0, \Sigma_q)$. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\Sigma_p^{-1}\Sigma_q$ and $g_\alpha = g_\alpha^{\mathrm{drep}}$.*

If $\Sigma_p \neq \Sigma_q$ and $1 + 2\alpha(\lambda_i - 1) > 0$ for all i we get

$$\mathrm{SNR}[g_\alpha(p, q_w, \epsilon)] = \frac{1}{d+2} \tag{11}$$

for $\alpha \to 0$,

$$\mathrm{SNR}[g_\alpha(p, q_w, \epsilon)] \leq \left(\frac{1 + \alpha(\lambda_{\min} - 1)}{1 + 2\alpha(\lambda_{\max} - 1)}\right)^2 \prod_{i=1}^{d} f(\lambda_i, \alpha) \tag{12}$$

for $\alpha > 0$, and

$$\mathrm{SNR}[g_\alpha(p, q_w, \epsilon)] \leq \left(\frac{1 + \alpha(\lambda_{\max} - 1)}{1 + 2\alpha(\lambda_{\min} - 1)}\right)^2 \prod_{i=1}^{d} f(\lambda_i, \alpha) \tag{13}$$

for $\alpha < 0$, where $\lambda_{\max} = \max_i \lambda_i$, $\lambda_{\min} = \min_i \lambda_i$ (these are both positive), and $f(\lambda, \alpha)$ is defined in eq. 10.

Otherwise, the SNR is not defined. If $\Sigma_p = \Sigma_q$, this is because the estimator is zero deterministically. If $1 + 2\alpha(\lambda_i - 1) \leq 0$ for any i, this is because the estimator has infinite variance.

The results in Theorem 3 can be interpreted similarly to those in Corollary 2. If $\alpha \to 0$, the SNR is just $1/(d+2)$, independent of p and q_w. If $\alpha \neq 0$, the SNR's upper bound contains the product of d terms, all at most 1, with equality only if the corresponding $\lambda_i = 1$. As with factorized distributions, for $\alpha \neq 0$, discrepancies between several dimensions of p and q_w accumulate, leading to a small SNR. As with fully-factorized Gaussians, this deterioration worsens for α values with large magnitude and for λ_i far from one. The condition that must be satisfied to get an estimator with finite variance is similar to the one for factorized Gaussians. The only difference is that, in this case, λ_i represents an eigenvalue of $\Sigma_p^{-1}\Sigma_q$, instead of the ratio $\sigma_{qi}^2/\sigma_{pi}^2$.

Example. Consider the case where p and q_w are d dimensional isotropic Gaussians with covariances $\sigma_p^2 I$ and $\sigma_q^2 I$, with $\sigma_p \neq \sigma_q$. If $\alpha \neq 0$, the estimator's SNR is upper bounded by $\propto f(\lambda, \alpha)^d$, where $f(\lambda, \alpha)$ is strictly less than 1. This upper bound goes to zero *exponentially* as a function of d. In contrast, for $\alpha \to 0$, the SNR decreases as $1/d$.

It is worth mentioning that the bounds in Theorem 3 are obtained as a relaxation of an exact but much more technical result, shown in Section 6 (Theorem 4). While this latter result is fully precise (it gives an exact expression for the SNR) it is hard to interpret, so we do not include it here.

4.3. Effect of SNR on Optimization

We presented general and representative scenarios for which the gradient estimator's SNR becomes extremely small as the dimensionality of the problem increases. How does this affect optimization convergence? Under some regularity assumptions, an SGD convergence guarantee assuming a bound on the SNR is known. See, for instance, Theorem 4.8 by (Bottou et al., 2018). Their eq. 4.9 is equivalent to an SNR bound. They show that (under some assumptions) SGD requires a number of iterations that is $\mathcal{O}(1/\text{SNR})$ to converge. Thus, an exponentially small SNR translates to an exponentially large number of SGD iterations. Intuitively, this is because a small SNR leads to a small step-size, which in turn leads to a large number of SGD iterations.

In addition, there are papers that analyze this from a more empirical perspective. For instance, (Shalev-Shwartz et al., 2017) relate gradient estimators with extremely low SNRs to complete failures of gradient-based optimization methods. They mention that when the SNR approaches small values, the noise can completely mask the signal, and thus gradients are not sufficiently informative for optimization to succeed.

5. Experiments and Results

Results in this work show that, for the scenarios considered, if $\alpha \neq 0$ unbiased estimates of $\nabla_w D_\alpha(p||q_w)$ suffer from a low SNR, which worsens fast with the dimensionality of the problem. Thus, methods based on these estimates will not scale to high dimensional problems. In this section we empirically show similar severe scalability issues for Bayesian logistic regression models.

We use two datasets: *Iris* and *Australian*, which have dimensionalities 4 and 14, respectively. For both datasets we used a subset of 100 samples. For *Iris* this reduced to keeping only data-points from two classes (out of the original three), while for *Australian* we subsampled 100 data-points. We use a diagonal Gaussian as variational distribution q_w, initialized to have mean zero and covariance identity.

When p is a posterior, we cannot directly estimate gradients of the alpha-divergence since $p(z|x)$ is intractable. However, if we define the "α-ELBO"

$$\mathcal{L}_\alpha(w) = \frac{1}{\alpha(1-\alpha)} \mathop{\mathbb{E}}_{q_w(z)} \left[\left(\frac{p(x,z)}{q_w(z)} \right)^\alpha - 1 \right], \qquad (14)$$

then it's easy to show that maximizing $\mathcal{L}_\alpha$ is equivalent to minimizing the alpha-divergence, since

$$f_\alpha(p(x)) = \mathcal{L}_\alpha(w) + p(x)^\alpha D_\alpha(p(z|x)||q(z)) \qquad (15)$$

for $f_\alpha(x) = \frac{1}{\alpha(1-\alpha)}(x^\alpha - 1)$. Thus a gradient of $\mathcal{L}_\alpha$ is equal to a gradient of D_α up to a sign change and a multiplication by the constant factor of $p(x)^\alpha$. Eq. 15 gives a lower-bound on $p(x)$ for $\alpha < 1$ and an upper-bound for $\alpha > 1$ (corresponding to the cases where f_α is increasing and decreasing, respectively).

We optimize $\mathcal{L}_\alpha$ by running SGD with unbiased gradient estimates for 1000 steps. We do this for $\alpha \in \{0, 0.1, 0.2, 0.3\}$ and for $N \in \{1, 10, 10^2, 10^3, 10^4\}$ (number of samples used to estimate the gradient at each step). For each pair (α, N) we tuned the step-size; we ran simulations for all step-sizes in the set $\{10^i\}_{i=-7}^{7}$ and selected the top-performing one. All results shown are averages over 15 simulations.

Optimization results are shown in Fig. 4. For the smaller dataset, *Iris* ($d = 4$), optimization converges properly for all values of α considered regardless of the number of samples N used to estimate the gradient. The situation is different for the *Australian* dataset ($d = 14$). In this case, optimization converges properly for $\alpha \to 0$, but as α is increased a much larger number of samples N is required to retain convergence ($N \geq 1000$ for $\alpha = 0.3$). This shows that, even for a simple logistic regression model of low dimension ($d = 14$), alpha-divergence minimization methods based on unbiased gradient estimates scale *very* poorly with the dimensionality of the problem when $\alpha \neq 0$. We also ran simulations with larger datasets ($d \approx 40$), for which optimization barely made any progress at all regardless of the number of samples N used. Similar results are obtained using the Adam optimizer (shown in Fig. 8, Appendix B).

Fig. 5 shows the SNR of the estimator for different values of α along a single optimization path, obtained by minimizing $\mathcal{L}_\alpha$ for $\alpha \to 0$ (equivalent to maximizing the ELBO). For the smaller dataset, *Iris* ($d = 4$), all values of α considered lead to comparable SNRs. In contrast, for the *Australian* dataset ($d = 14$), the SNR descreases rapidly as α is increased.

6. Discussion

We study unbiased methods for alpha-divergence minimization with the goal of understanding when these methods may be successfully applied. We present a detailed analysis of the SNR of unbiased gradient estimates for different scenarios. Our results are pessimistic. Suppose the variational family is any fully-factorized family, or the set of full-rank Gaussians. We show that in the favorable case where the posterior is inside the variational family, the SNR degrades catastrophically in the dimensionality. Optimization theory suggests that an exponential amount of computation time would be needed to optimize the objectives.

Interestingly, results in this work rule out some potential "intuitive" solutions. For instance, one might think that, for $\alpha > 0$, using an over-dispersed distribution q_w could mitigate variance issues. However, Theorem 1 shows this is not the case. While an over-dispersed distribution might help avoid estimators with infinite variance, it would not

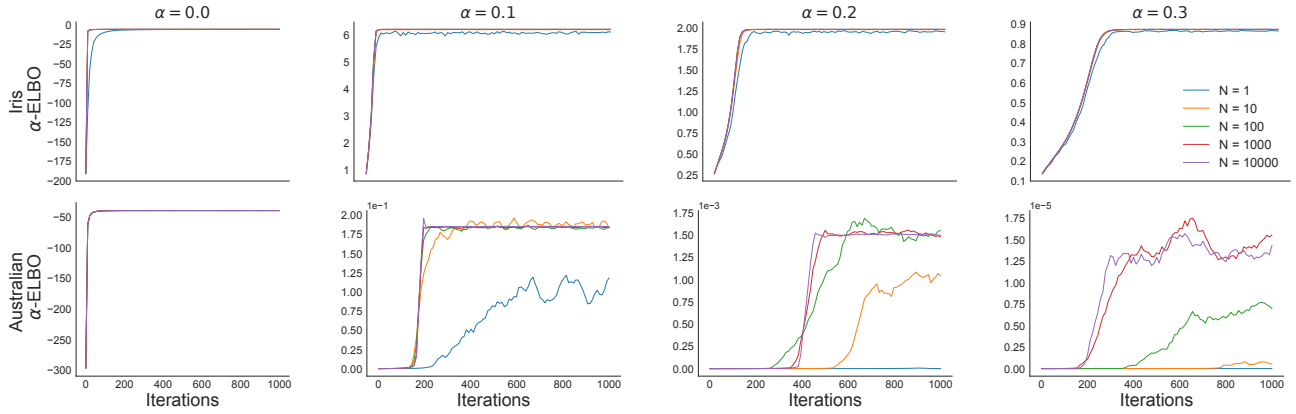

Figure 4. Optimization results for each α for all values of N considered. The loss at each step (eq. 14) is estimated using 2.5×10^5 samples for both datasets.

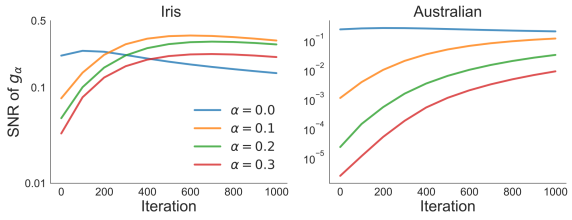

Figure 5. Direct test of SNR for different values of α along a single shared optimization path.

avoid the exponential deterioration of the SNR in terms of the dimensionality of the problem.

In addition, one could also consider using an adaptive gradient optimization method, such as Adam (Kingma & Ba, 2014). As it can be observed in Figs. 7 and 8 in Appendix B, this does not solve the issue. This is not really surprising. Intuitively, this is because the extremely low SNR is a property of the gradient estimator, and thus will difficult optimization regardless of the gradient-based optimizer used. Indeed, Section 2.1 of the original Adam paper (Kingma & Ba, 2014) states[4] "With a smaller SNR the effective stepsize Δt will be closer to zero. This is a desirable property, since a smaller SNR means that there is greater uncertainty about whether the direction of m corresponds to the direction of the true gradient." In addition, the theoretical results by Shalev-Shwartz et al. (2017) are independent of the optimization algorithm used, and some of their negative empirical results were obtained using Adam.

Why is the behavior for $\alpha \neq 0$ so different to $\alpha \to 0$? Our understanding is that that the problematic terms vanish in the limit of $\alpha \to 0$. For example, eq. 7 becomes eq. 6, due to the fact that $\tilde{D}_\alpha \to 1$ as $\alpha \to 0$. Similarly, in eq. 9, we get that $f(\lambda, \alpha) \to 1$ as $\alpha \to 0$, meaning that only $\mathrm{SNR}[g_\alpha]$ remains. For full-rank Gaussians, it is probably easiest to understand via the exact result in Thm. 4 (see below). There, if $\alpha \to 0$, we have that $f(\lambda, \alpha) \to 1$, $U \to I$, $V \to I$, meaning the overall SNR becomes $1/(d+2)$ (exactly eq. 11) which has no problematic exponential dependence on dimensionality.

Given the failure of unbiased methods, one could consider using some biased alternative. However, it has been recently observed that, in high dimensions, these methods return suboptimal solutions that fail to minimize the target alpha-divergence (Geffner & Domke, 2021). An analysis analogous to the one presented in this work is needed to understand this failure. We believe that such an analysis might be related to the curse of dimensionality for self-normalized importance sampling, which conjectures that to get meaningful results the number of samples used from the proposal distribution should be exponential in the dimensionality of the problem (Bengtsson et al., 2008; Bugallo et al., 2017).

Exact result for Gaussians. Theorem 3 bounds the gradient estimator's SNR for Gaussians with arbitrary covariances. While it admits a nice interpretation, it is not tight. As mentioned previously, an exact result is possible. We include it here. While this result is fully precise, it is harder to interpret than the bounds from Theorem 3. It may be possible to find tighter bounds that are still "simple", or to find an intuitive interpretation of the exact result. We believe a step in these directions may further increase our understanding of these methods.

Theorem 4. *Take the setting of Theorem 3 with $\Sigma_p \neq \Sigma_q$ and $1 + 2\alpha(\lambda_i - 1) > 0$ for all i. Let S be a matrix such that $\Sigma_q = SS^\top$ and let $\alpha \neq 0$. Then,*

$$\mathrm{SNR}[g_\alpha(p, q_w, \epsilon)] = \frac{\|BU^{-1}\|_F^2 \times \prod_{i=1}^d f(\lambda_i, \alpha)}{\mathrm{tr}(V^{-1})\mathrm{tr}(BV^{-1}B^\top) + 2\|BV^{-1}\|_F^2},$$

[4] v9 on arxiv.

where $B = (\Sigma_p^{-1} - \Sigma_q^{-1})S$, $U = (1-\alpha)I + \alpha S^\top \Sigma_p^{-1} S$, *and* $V = (1-2\alpha)I + 2\alpha S^\top \Sigma_p^{-1} S$.

Acknowledgements

We would like to thank Javier Burroni and Juan Cristi for useful feedback and suggestions.

References

Amari, S.-i. Differential-geometrical methods in statistics. *Lecture Notes on Statistics*, 28:1, 1985.

Bengtsson, T., Bickel, P., Li, B., et al. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and statistics: Essays in honor of David A. Freedman*, pp. 316–334. Institute of Mathematical Statistics, 2008.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bornschein, J. and Bengio, Y. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. Adaptive importance sampling: the past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *Proceedings of the International Conference on Learning Representations*, 2016.

Cichocki, A. and Amari, S.-i. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2732–2741, 2017.

Geffner, T. and Domke, J. Empirical evaluation of biased methods for alpha divergence minimization. In *Symposium on Advances in Approximate Bayesian Inference*, 2021.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T., and Turner, R. Black-box α-divergence minimization. In *Proceedings of the 33nd International Conference on Machine Learning (ICML-16)*. International Machine Learning Society, 2016.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2013.

Kuleshov, V. and Ermon, S. Neural variational inference and learning in undirected graphical models. In *Advances in Neural Information Processing Systems*, pp. 6734–6743, 2017.

Li, Y. and Turner, R. E. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pp. 1073–1081, 2016.

Minka, T. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

Minka, T. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

Naesseth, C. A., Lindsten, F., and Blei, D. Markovian score climbing: Variational inference with kl (p—— q). *arXiv preprint arXiv:2003.10374*, 2020.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning (ICML-18)*, 2018.

Regli, J.-B. and Silva, R. Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*, 2018.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1278–1286, 2014.

Roeder, G., Wu, Y., and Duvenaud, D. K. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pp. 6925–6934, 2017.

Shalev-Shwartz, S., Shamir, O., and Shammah, S. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML-17)*, 2017.

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Wan, N., Li, D., and Hovakimyan, N. f-divergence variational inference. *arXiv preprint arXiv:2009.13093*, 2020.

Wang, D., Liu, H., and Liu, Q. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pp. 5737–5747, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. Advances in variational inference. *arXiv preprint arXiv:1711.05597*, 2017.