On the Difficulty of Unbiased Alpha Divergence Minimization

Tomas Geffner [1] Justin Domke [1]

[1]University of Massachusetts, Amherst

Overview and Summary

- **Variational Inference** minimizes $\mathrm{KL}(q||p)$.
- Could target **alpha-divergence**

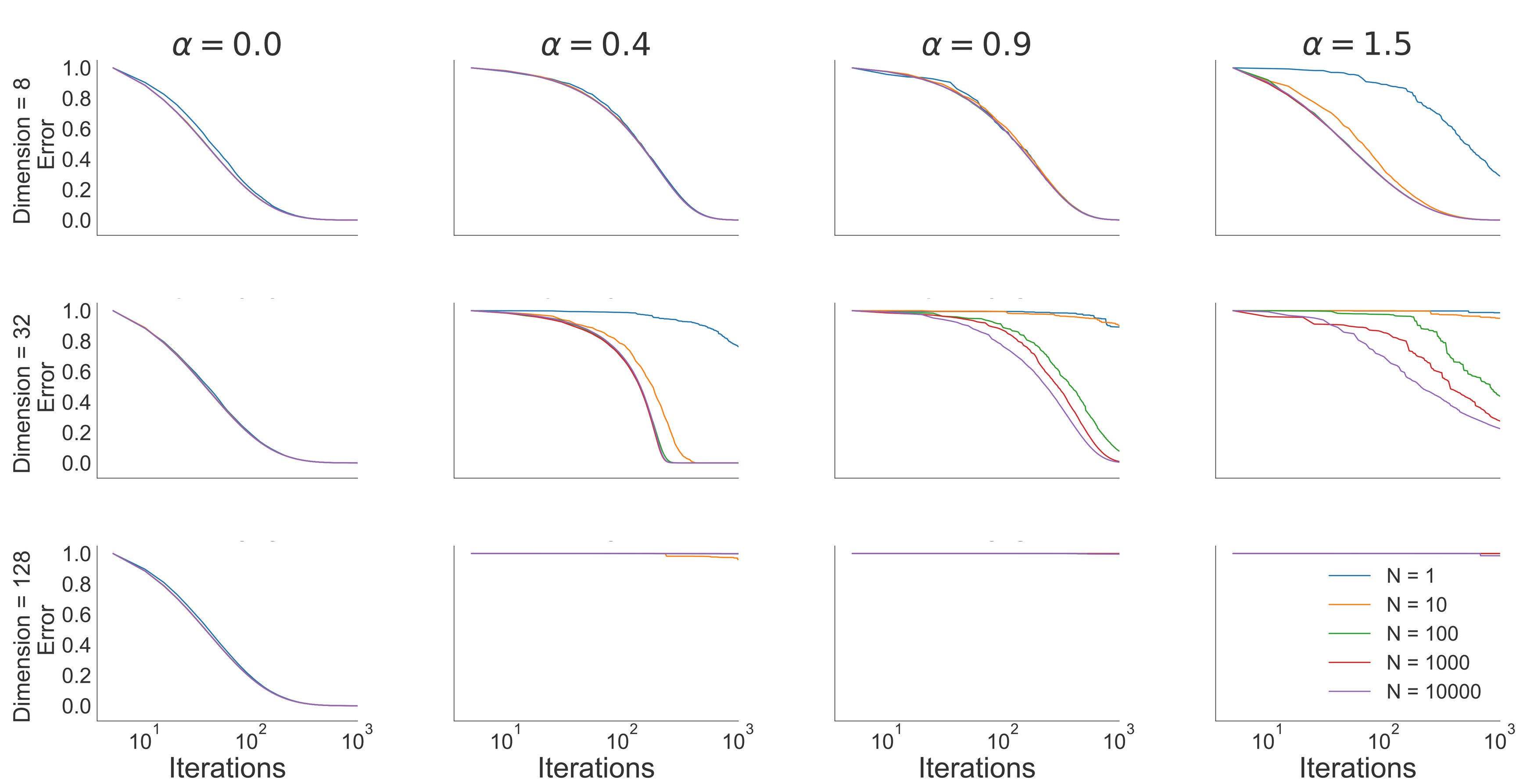
$$D_\alpha(p||q) = \frac{1}{\alpha(\alpha-1)} \mathbb{E}_{q(z)} \left[\left(\frac{p(z)}{q(z)} \right)^\alpha - 1 \right].$$

- Previous work: use unbiased **reparameterization gradients**.
- We observe this often **fails in high dimensions and high alpha**.
- Why? **Estimator's SNR decreases exponentially with dimension**.

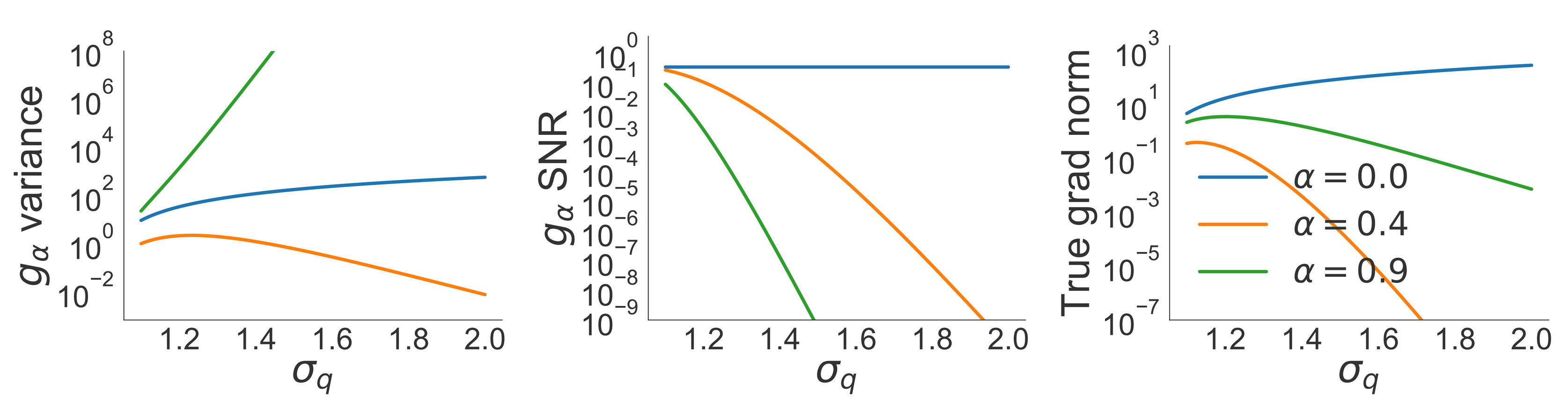
We use $\mathrm{SNR}[X] = \frac{\|\mathbb{E}X\|^2}{\mathbb{E}\|X\|^2} \leq 1.$

Motivating example

- p and q factorized Gaussians with mean zero and variances $\sigma_{pi}^2 = 1$, $\sigma_{qi}^2 = 4$.
- Find parameters σ_{qi} that minimize alpha-divergence.
- (Solution easy, just want simple empirical test for estimator.)



Variance alone does not explain failure, SNR does:



Fully Factorized Distributions

Theorem. Let $p(z) = \prod_{i=1}^d p_i(z_i)$ and $q(z) = \prod_{i=1}^d q_i(z_i)$, and let $g(p,q)$ be the unbiased reparameterization estimator of the alpha-divergence between p and q. Then

$$\mathrm{SNR}[g_j(p,q)] = \mathrm{SNR}[g(p_j,q_j)] \qquad \text{if } \alpha \to 0$$

$$\mathrm{SNR}[g_j(p,q)] = \mathrm{SNR}[g(p_j,q_j)] \prod_{i=1, i\neq j}^{d} \mathrm{SNR}[\tilde{D}_\alpha(p_i,q_i)] \qquad \text{if } \alpha \neq 0,$$

where $D_\alpha(p_i,q_i)$ is an unbiased estimator (up to constants) of $D_\alpha(p_i,q_i)$.

Simply put:

- If $\alpha \to 0$ the SNR is just the SNR of the gradient estimator of a divergence between two 1-dimensional distributions.
- If $\alpha \neq 0$ the SNR includes the product of d terms, all less than one (unless $p_i = q_i$).

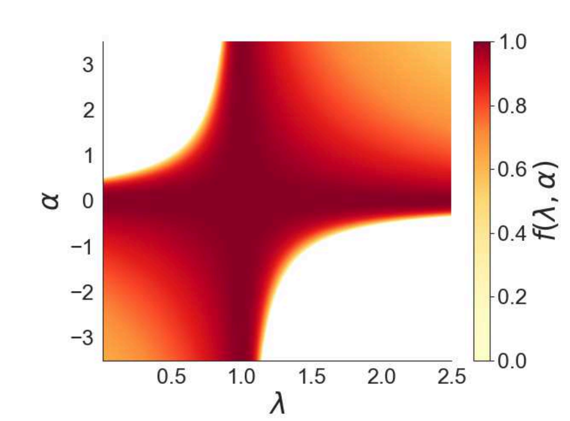
Corollary: Let p and q be mean-zero factorized Gaussians with vars σ_{pi}^2, σ_{qi}^2. Let $\lambda_i = \frac{\sigma_{qi}^2}{\sigma_{pi}^2}$. Then, if all expectations exist,

$$\mathrm{SNR}[g_j(p,q)] = \underbrace{\frac{1+2\alpha(\lambda_j-1)}{3} f(\lambda_j,\alpha)^3}_{\mathrm{SNR}[g(p_j,q_j)]} \prod_{i=1, i\neq j}^{d} \underbrace{f(\lambda_i,\alpha)}_{\mathrm{SNR}[\tilde{D}_\alpha(p_i,q_i)]},$$

where

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

Simply put, the SNR contains the product of d terms all less than one, which get smaller for alpha far from zero and for p and q very different.



Full Rank Gaussians

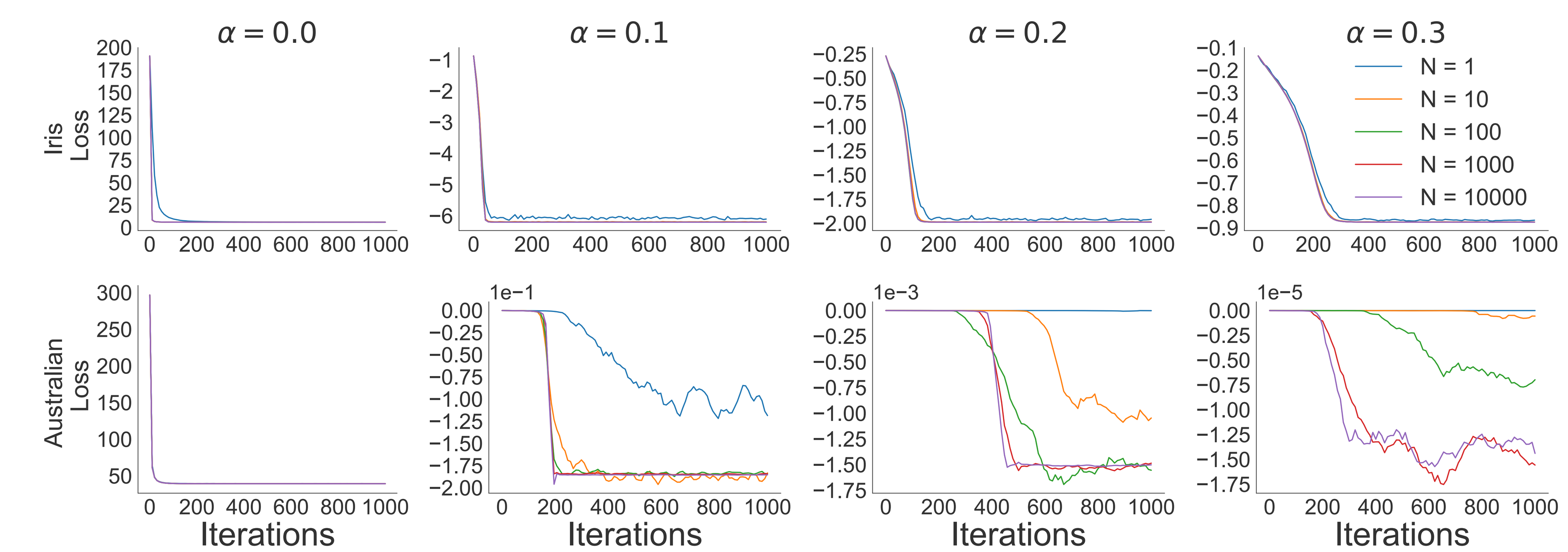
Theorem. Let p and q be mean-zero Gaussians with covariances Σ_p and Σ_q. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\Sigma_p^{-1}\Sigma_q$. Then, if all expectations exist,

$$\mathrm{SNR}[(p,q)] = \frac{1}{d+2} \qquad \text{if } \alpha \to 0,$$

$$\mathrm{SNR}[(p,q)] \leq \left(\frac{1+\alpha(\lambda_{\min}-1)}{1+2\alpha(\lambda_{\max}-1)} \right)^2 \prod_{i=1}^{d} f(\lambda_i,\alpha) \qquad \text{if } \alpha > 0.$$

Empirical Evaluation

- Bayesian logistic regression.
- Two datasets: *iris* ($d=4$) and *australian* ($d=14$).



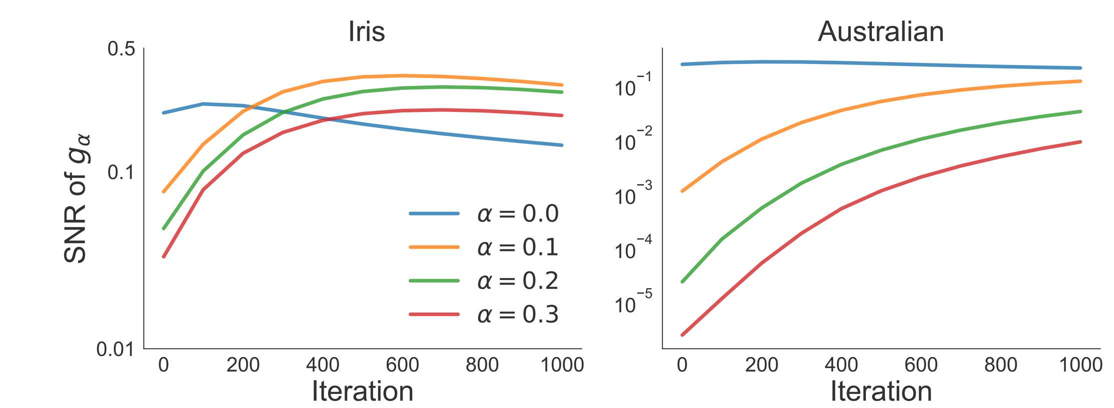


Figure 1: SNR along a single optimization trace.

Final thoughts

- Optimization theory suggests that an exponential amount of computation time would be needed to optimize the objectives.
- One might hope to guarantee a good SNR under some assumptions about the target. For example, if the log-posterior were fully-factorized, concave, strongly concave, Lipschitz smooth, or Gaussian. Our results show that, for general alpha-divergences, no such guarantee is possible.
- A general-purpose algorithm for optimizing an alpha-divergence based on currently available unbiased gradient estimators may be unachievable.
- Other optimizers (e.g. Adam) do not fix the issue.