Overview and Summary

- Variational Inference minimizes KL(q||p).
- Could target alpha-divergence

\[ D_{\alpha}(p||q) = \frac{1}{\alpha(\alpha-1)} E_q \left[ \frac{(p(x)/q(x))^{\alpha-1}}{1 + (p(x)/q(x))^{\alpha-1}} \right] \]

- 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.

Motivating example

- p and q factorized Gaussians with mean zero and variances \( \sigma_p^2 = 1, \sigma_q^2 = 4 \).
- Find parameters \( \sigma_p^2 \) that minimize alpha-divergence.
- [Solution easy, just want simple empirical test for estimator]

\[ g \sim \mathcal{N}(0, \Sigma), \quad q \sim \mathcal{N}(0, \Sigma_q) \]

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

\[ \text{SNR}(g(p, q)) = \text{SNR}(g(p_i, q_i)) \quad \text{if} \quad \alpha \to 0 \]

\[ \text{SNR}(g(p, q)) = \prod_{i=1}^d \text{SNR}(D_{\alpha}(p_i, q_i)) \quad \text{if} \quad \alpha \neq 0, \]

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 variances \( \sigma_p^2, \sigma_q^2 \). Let \( \lambda_i \sim \sigma_i^2 \). Then, if all expectations exist,

\[ \text{SNR}(g(p, q)) = \frac{1 + 2\alpha(\lambda_i - 1)}{\text{SNR}(g(p_i, q_i))} \prod_{i=1}^d f(\lambda_i, \alpha) \]

where

\[ f(\lambda, \alpha) = \frac{1}{\lambda^{2\alpha} \text{SNR}(D_{\alpha}(p_i, q_i))} \]

**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 \( \Sigma_p \) and \( \Sigma_q \). Let \( \lambda_1, \ldots, \lambda_d \) be the eigenvalues of \( \Sigma_p \). Then, if all expectations exist,

\[ \text{SNR}(g(p, q)) = \frac{1}{\lambda^{2\alpha}} \quad \text{if} \quad \alpha \to 0, \]

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

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.