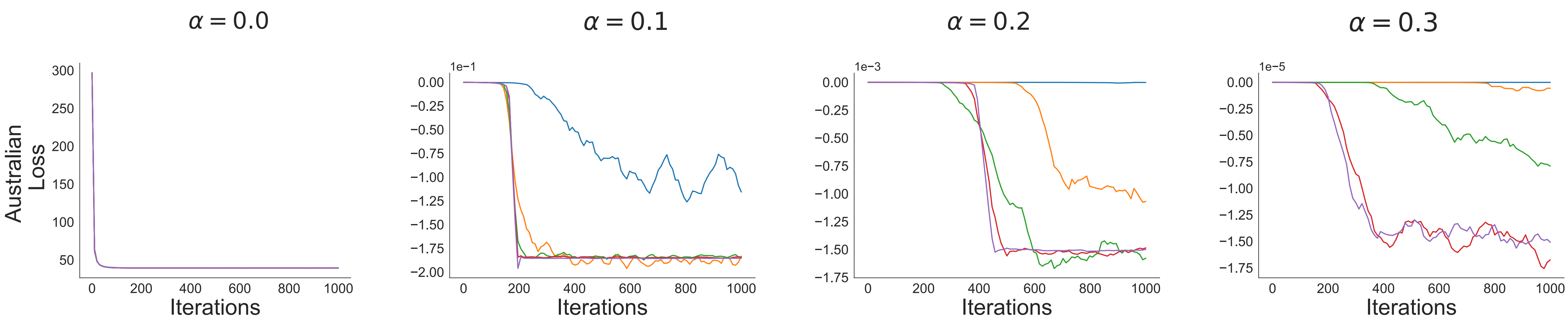
On the Difficulty of Unbiased Alpha Divergence Minimization

Tomas Geffner and Justin Domke
University of Massachusetts, Amherst



ICML 2021

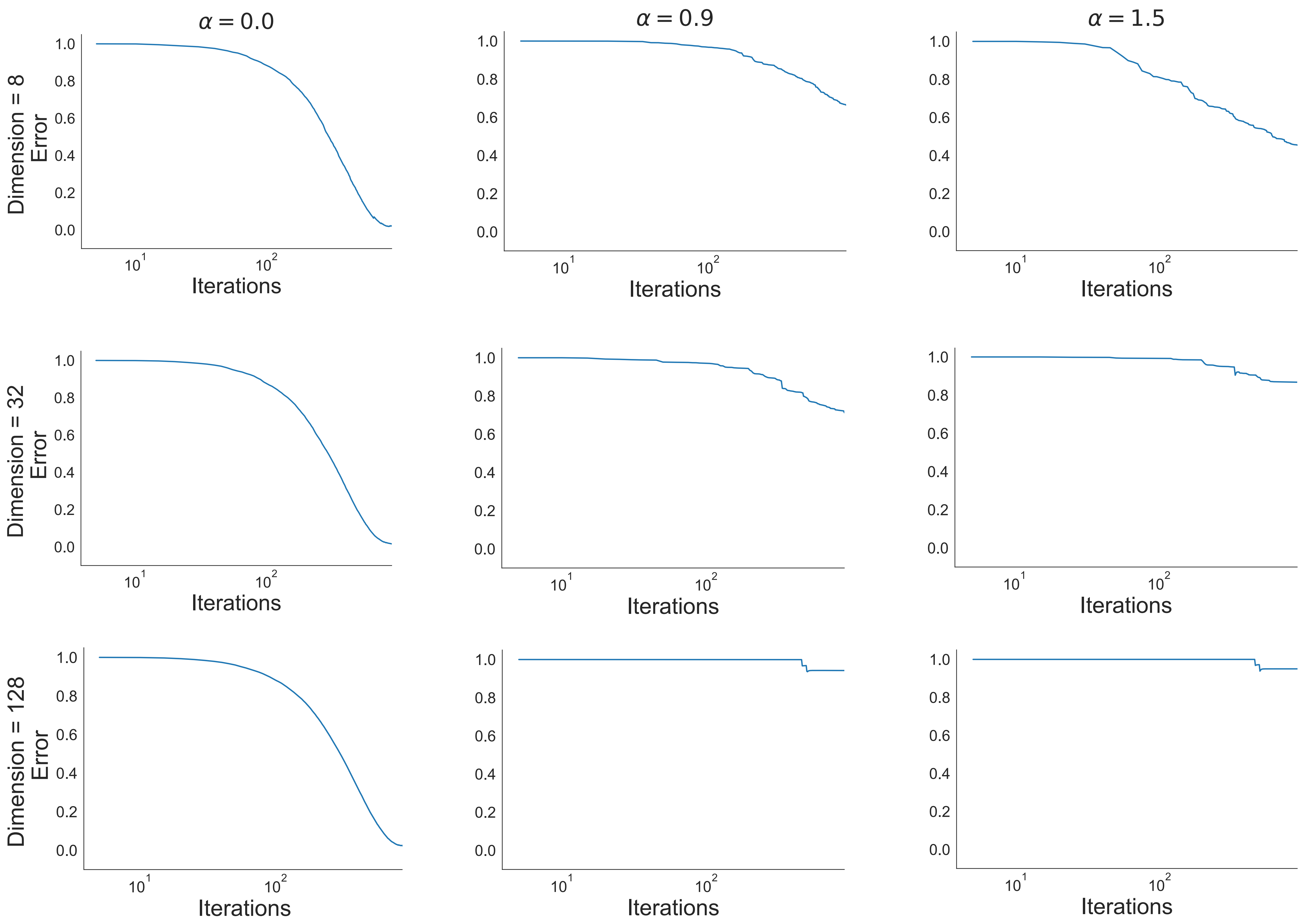
- Alpha divergence between target $\mathbf{p}$ and approximation $\mathbf{q}$ parameterized by $\mathbf{w}$:

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

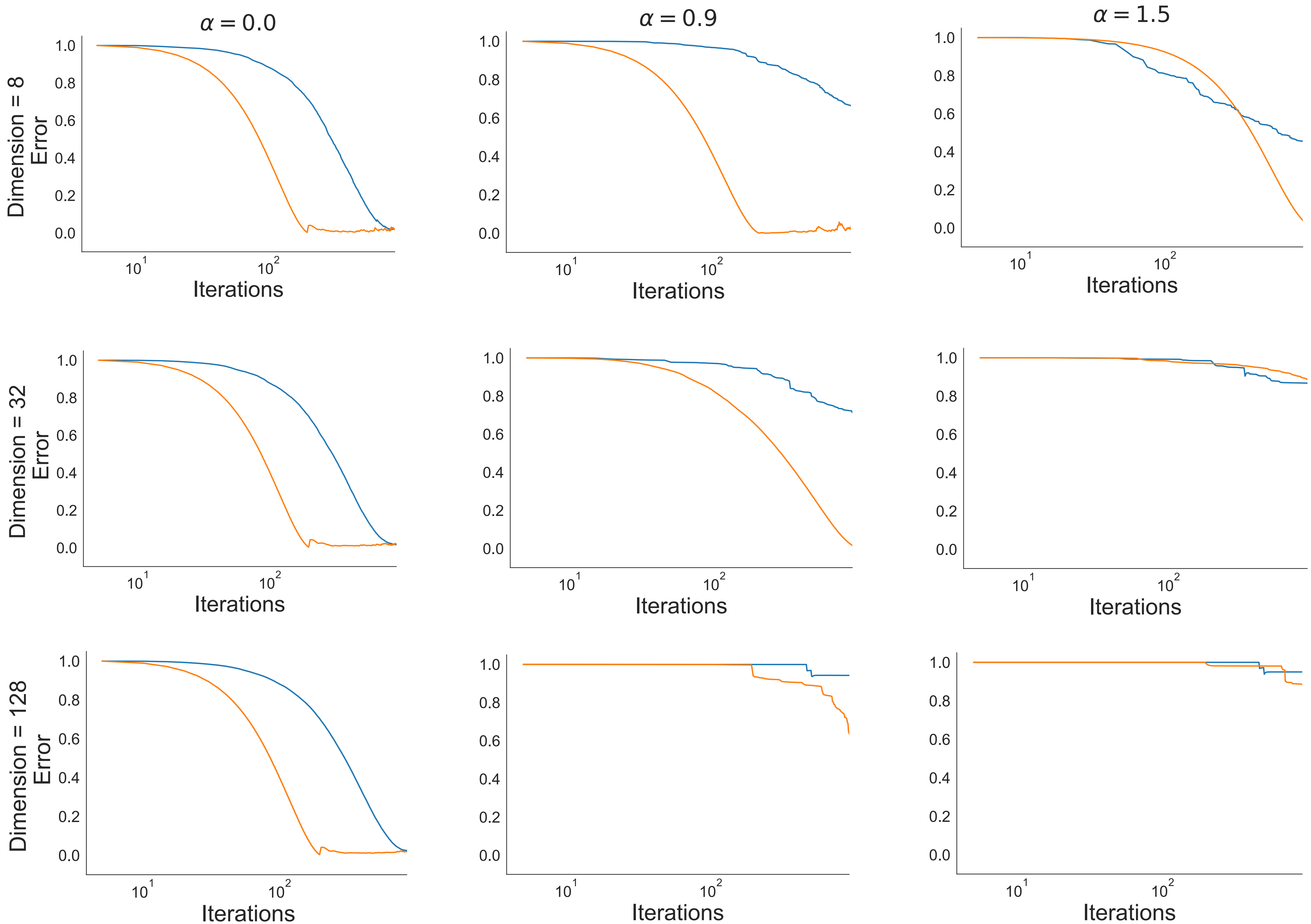
- Can compute unbiased reparameterization gradient wrt parameters $\mathbf{w}$ of approximation (or novel “double reparameterization” variant).

- We observe: Moderately high dimensionality and alpha very problematic.

p and **q** are mean zero Gaussians with different variance.
(1 sample to estimate gradient at each step.)



p and **q** are mean zero Gaussians with different variance.
(1 sample / 10000 samples to estimate gradient at each step.)



Our results:

- Gradient estimator SNR decreases exponentially with problem dimension.
    - Increasing alpha worsens the effect.

- We prove this for any fully factorized distribution and Gaussians.
    - We observe this empirically in other scenarios.

- An efficient and widely applicable alpha divergence minimization algorithm based on these gradient estimators may be unachievable.

- We study two cases for $\mathbf{p}$ and $\mathbf{q}$:
  1. Any fully factorized distributions,
  2. Gaussians.

- SNR gets exponentially worse unless $\alpha = 0$ or $\mathbf{p}$ is <u>very</u> close to $\mathbf{q}$.

- Roughly, for both cases we get

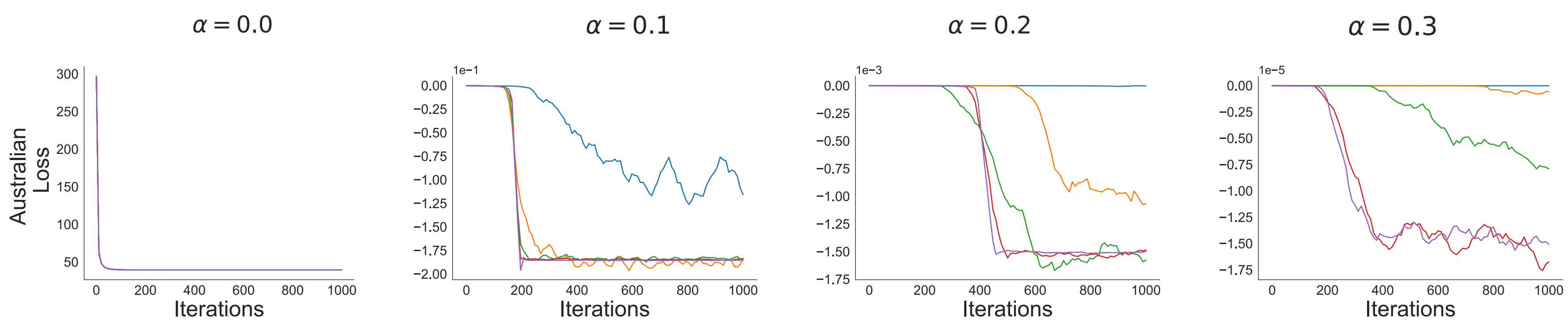
$$\text{SNR} \propto \prod_{d=1}^{D} C_d(\alpha), \qquad \text{where} \qquad C_d(\alpha) \leq 1.$$

  - $C_d(\alpha) = 1$ if $\alpha = 0$ (traditional VI).
  - $C_d(\alpha) = 1$ if $\mathbf{p}(z_d) = \mathbf{q}(z_d)$.
  - Otherwise $C_d(\alpha) < 1$.

- Gets worse as alpha increases or $\mathbf{p}(z_d)$ becomes more different from $\mathbf{q}(z_d)$.

(**For Gaussians** $C_d(\alpha) \propto \frac{1}{\sqrt{1+|\alpha|}}$.)

- Tested empirically on Bayesian Logistic Regression (dim=14), observed similar effect:



- Rigorous results in the paper:

Theorem 4. *Let* $p(z) = \mathcal{N}(z|0, \Sigma_p)$, $q(z) = \mathcal{N}(z|0, \Sigma_q)$, *and* S *be a matrix such that* $SS^\top = \Sigma_q$. *Let* $\lambda_1, \ldots, \lambda_d$ *be the eigenvalues of* $\Sigma_p^{-1}\Sigma_q$ *and* $g_\alpha = g_\alpha^{\text{drep}}$. *If* $1 + 2\alpha(\lambda_i - 1) \leq 0$ *for any* i, *then the estimator* g_α *has infinite variance. Otherwise, if* $\Sigma_p \neq \Sigma_q$, *then*

$$\text{SNR}[g_\alpha(p, q_w, \epsilon)] = \frac{\|BU^{-1}\|_F^2}{\text{tr}(V^{-1})\text{tr}(BV^{-1}B^\top) + 2\|BV^{-1}\|_F^2} \prod_{i=1}^{d} f(\lambda_i, \alpha) \qquad \text{if } \alpha \neq 0 \qquad (11)$$

$$\text{SNR}[g_\alpha(p, q_w, \epsilon)] = \frac{1}{d+2} \qquad \text{if } \alpha \to 0, \qquad (12)$$

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

Corollary 5. *Take the setting of Theorem 4 with* $\Sigma_p \neq \Sigma_q$ *and* $1 + 2\alpha(\lambda_i - 1) > 0$ *for all* i. *Then,*

$$\text{SNR}[g_\alpha(p, q_w, \epsilon)] \leq \left(\frac{1-\alpha+\alpha\lambda_{\min}}{1-2\alpha+2\alpha\lambda_{\max}}\right)^2 \prod_{i=1}^{d} f(\lambda_i, \alpha) \qquad \text{if } \alpha > 0 \qquad (13)$$

$$\text{SNR}[g_\alpha(p, q_w, \epsilon)] \leq \left(\frac{1-\alpha+\alpha\lambda_{\max}}{1-2\alpha+2\alpha\lambda_{\min}}\right)^2 \prod_{i=1}^{d} f(\lambda_i, \alpha) \qquad \text{if } \alpha < 0, \qquad (14)$$

Questions?