On Signal-to-Noise Ratio Issues in Variational Inference for Deep Gaussian Processes




Tim G. J. Rudner* Oscar Key* Yarin Gal Tom Rainforth

— ICML 2021 —

Correspondence to

tim.rudner@cs.ox.ac.uk

Latent-Variable Deep Gaussian Processes

Conventional Deep GPs (Damianou et al., [2013]; Salimbeni et al. [2017])

$$y_n = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}\left(x_n\right)\right)\dots\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

Latent-Variable Deep Gaussian Processes

Conventional Deep GPs (Damianou et al., [2013]; Salimbeni et al. [2017])

$$y_n = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}\left(x_n\right)\right)\dots\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

Latent-Variable Deep GPs (Salimbeni et al. [2018])

$$y_n = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}\left([x_n, z_n]\right)\right)\dots\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

$$z_n \sim \mathcal{N}\left(0, I_{\tilde{D}}\right)$$

Latent-Variable Deep Gaussian Processes

Conventional Deep GPs (Damianou et al., [2013]; Salimbeni et al. [2017])

$$y_n = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}\left(x_n\right)\right)\dots\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

Latent-Variable Deep GPs (Salimbeni et al. [2018])

$$y_n = f^{(L)}\left(f^{(L-1)}\left(\dots f^{(1)}\left([x_n, z_n]\right)\right)\dots\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

$$z_n \sim \mathcal{N}\left(0, I_{\tilde{D}}\right)$$

Latent Variable

2-Layer Latent-Variable Deep GP Model

$$y_n = f^{(2)}\left(f^{(1)}\left([x_n, z_n]\right)\right) + \varepsilon_n$$

$$\varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

$$z_n \sim \mathcal{N}\left(0, I_{\tilde{D}}\right)$$

Latent-Variable Deep Gaussian Processes

2-Layer Latent-Variable Deep GP Model

$$y_n = f^{(2)}\left(f^{(1)}\left([x_n, z_n]\right)\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

$$z_n \sim \mathcal{N}\left(0, I_{\tilde{D}}\right)$$

Variational Objective (Salimbeni et al. [2018])

$$\mathcal{L}_K \overset{\text{def}}{=} \mathbb{E}_{f^{(1)}_{1:K}, z_{n,1:K}}\left[\sum_n \log \frac{1}{K}\sum_{k=1}^{K} \frac{\mathcal{F}\left(x_n, y_n, f^{(1)}_k, z_{n,k}\right) p\left(z_{n,k}\right)}{q_\phi\left(z_{n,k}\right)}\right] - \sum_{\ell=1}^{2} D_{\mathrm{KL}}\left(q(f^{(\ell)}) \middle\| p(f^{(\ell)})\right)$$

Latent-Variable Deep Gaussian Processes

2-Layer Latent-Variable Deep GP Model

$$y_n = f^{(2)}\left(f^{(1)}\left([x_n, z_n]\right)\right) + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

$$z_n \sim \mathcal{N}\left(0, I_{\tilde{D}}\right)$$

Variational Objective (Salimbeni et al. [2018])

$$\mathcal{L}_K \overset{\text{def}}{=} \mathbb{E}_{f_{1:K}^{(1)}, z_{n,1:K}}\left[\sum_n \log \frac{1}{K}\sum_{k=1}^{K} \frac{\mathcal{F}\left(x_n, y_n, f_k^{(1)}, z_{n,k}\right) p\left(z_{n,k}\right)}{q_\phi\left(z_{n,k}\right)}\right] - \sum_{\ell=1}^{2} D_{\mathrm{KL}}\left(q(f^{(\ell)}) \,\middle\|\, p(f^{(\ell)})\right)$$

Importance Weight = $w_{n,k}$

Main Result: SNR Decreases as K Increases

Theorem 1.

Assume mild regularity conditions and let

$$\Delta^{\mathrm{DGP}}_{n,M,K}(\phi) \overset{\mathrm{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \nabla_\phi \log \frac{1}{K}\sum_{k=1}^{K} w_{n,m,k}.$$

Then, the signal-to-noise ratio of gradient estimator $\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)$ is

$$\mathrm{SNR}^{\mathrm{DGP}}_{n,M,K}(\phi) = \frac{\left|\mathbb{E}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]\right|}{\sqrt{\mathrm{Var}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]}} = \mathcal{O}\left(\sqrt{M/K}\right)$$

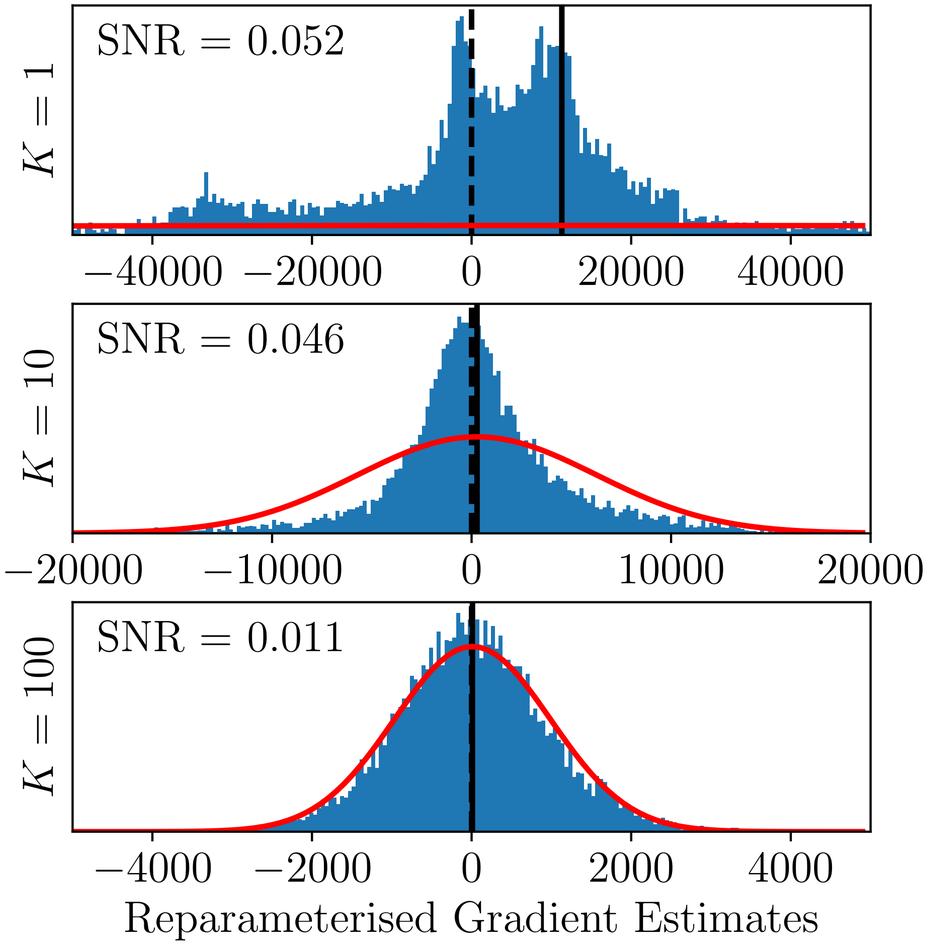
Empirical Confirmation of SNR Deterioration

- Gradient Estimate

$$\Delta^{\mathrm{DGP}}_{n,M,K}(\phi) \overset{\mathrm{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \nabla_\phi \log \frac{1}{K}\sum_{k=1}^{K} w_{n,m,k}$$

- Signal-to-Noise Ratio

$$\mathrm{SNR}^{\mathrm{DGP}}_{n,M,K}(\phi) = \frac{\left|\mathbb{E}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]\right|}{\sqrt{\mathrm{Var}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]}} = \mathcal{O}\left(\sqrt{M/K}\right)$$



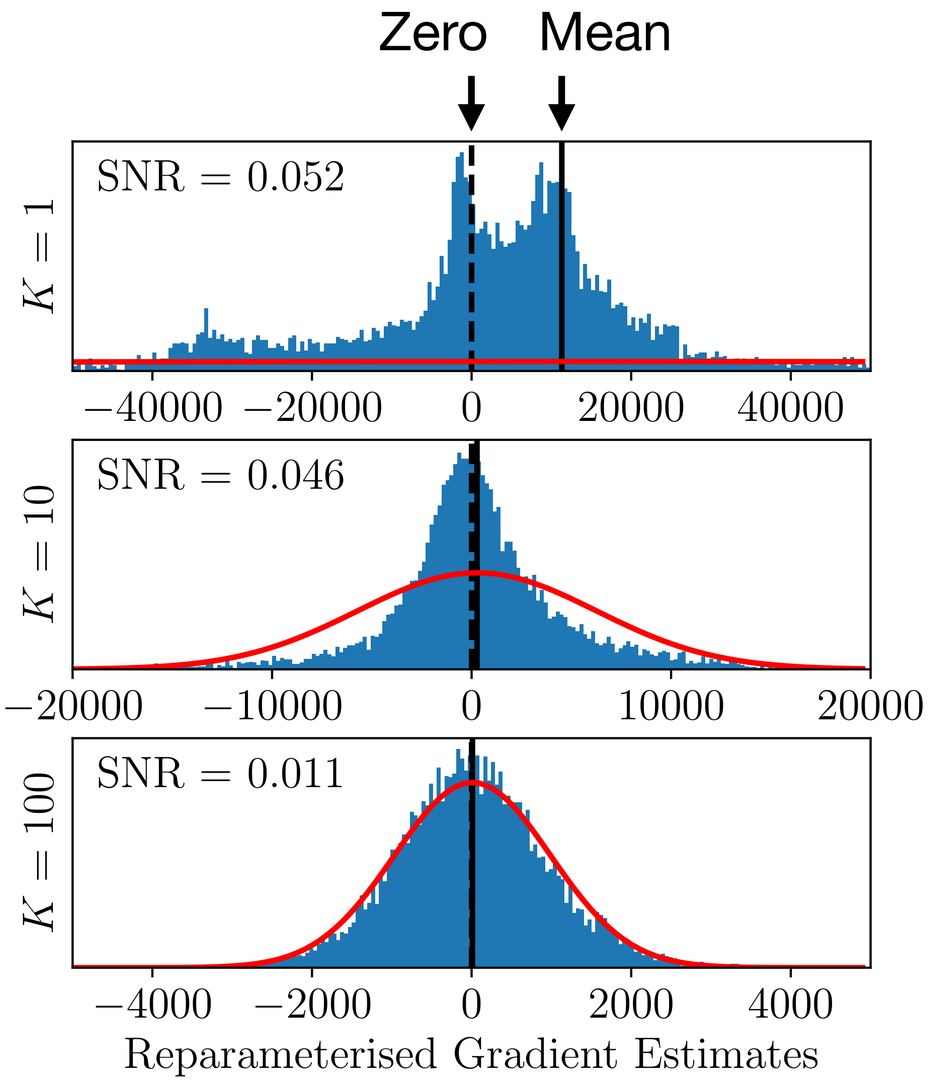
Empirical Confirmation of SNR Deterioration

- Gradient Estimate

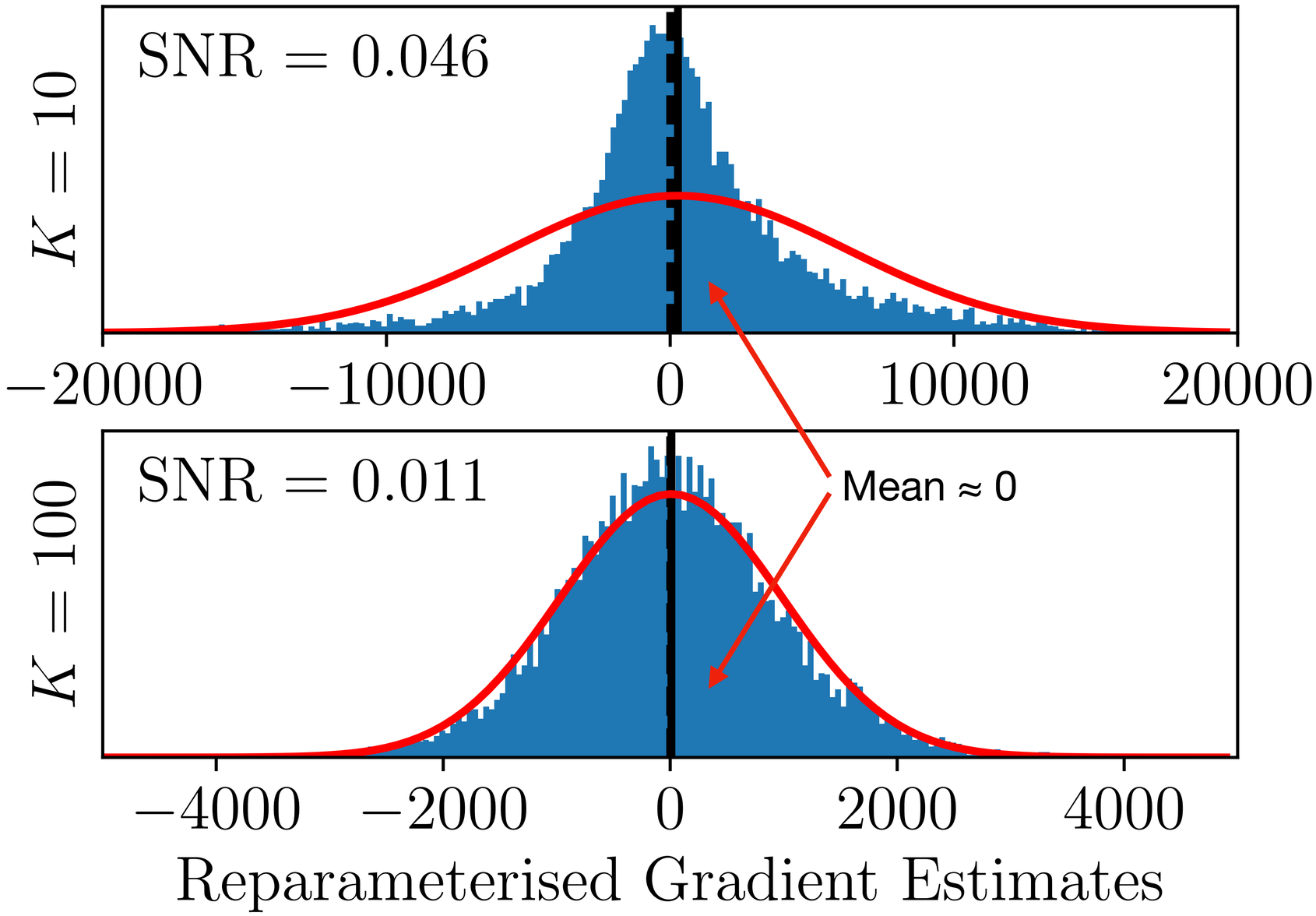
$$\Delta^{\mathrm{DGP}}_{n,M,K}(\phi) \overset{\mathrm{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \nabla_\phi \log \frac{1}{K}\sum_{k=1}^{K} w_{n,m,k}$$

- Signal-to-Noise Ratio

$$\mathrm{SNR}^{\mathrm{DGP}}_{n,M,K}(\phi) = \frac{\left|\mathbb{E}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]\right|}{\sqrt{\mathrm{Var}\left[\Delta^{\mathrm{DGP}}_{n,M,K}(\phi)\right]}} = \mathcal{O}\left(\sqrt{M/K}\right)$$

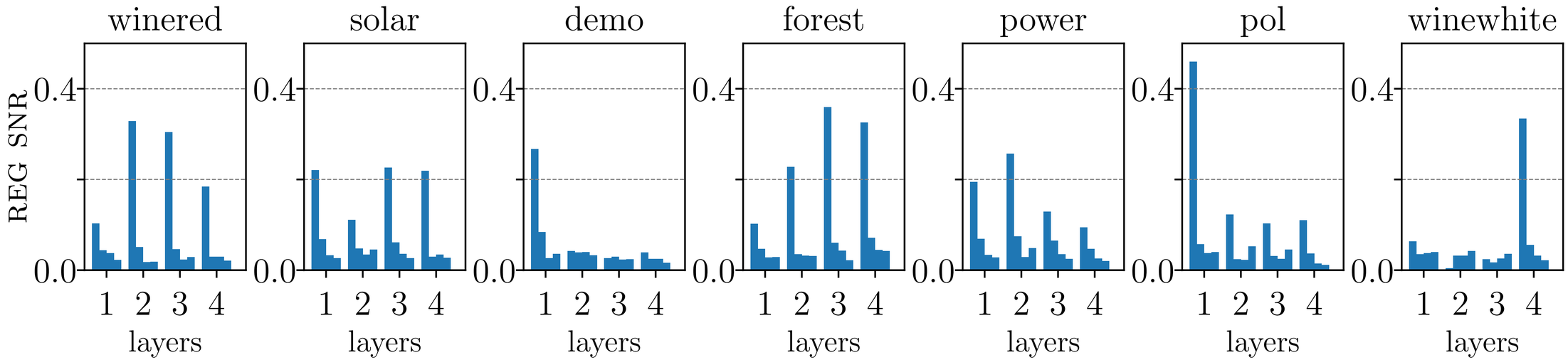


Empirical Confirmation of SNR Deterioration



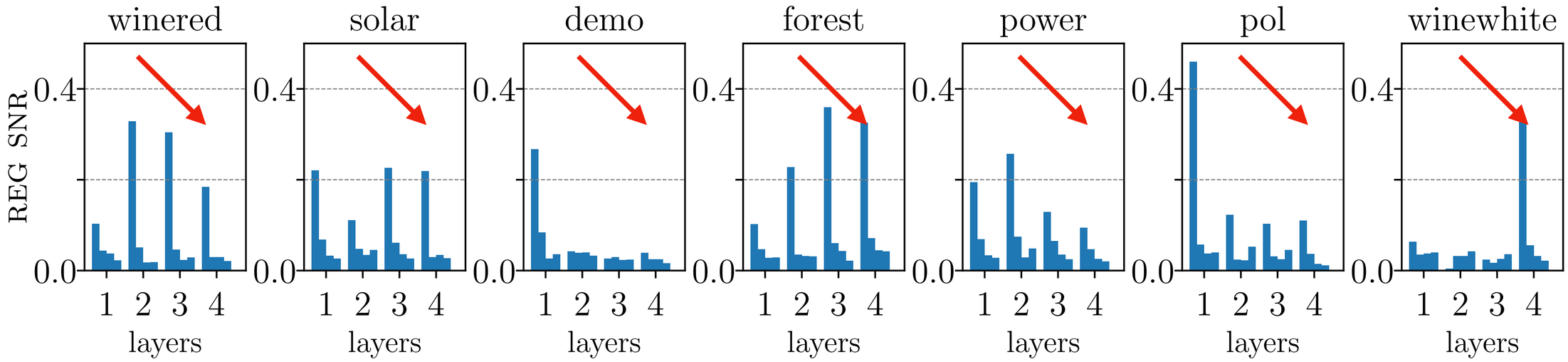
Signal-to-Noise Ratio Issues in Deep GPs

Empirical Confirmation of SNR Deterioration



Signal-to-Noise Ratio Issues in Deep GPs

Empirical Confirmation of SNR Deterioration



Fixing the Pathology

Deep GP Doubly Reparameterized Gradient Estimator

Fixing the Pathology

Deep GP Doubly Reparameterized Gradient Estimator

- Adapted to deep GPs from Tucker et al. [2018]:

$$\widetilde{\Delta}^{\mathrm{DGP}}_{n,M,K}(\phi) \overset{\mathrm{def}}{=} \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\frac{w_{n,m,k}}{\sum_{j=1}^{K} w_{n,m,j}}\right)^2 \frac{\partial \log w_{n,m,k}}{\partial z_{n,m,k}} \frac{\partial z_{n,m,k}}{\partial \phi}$$

Fixing the Pathology

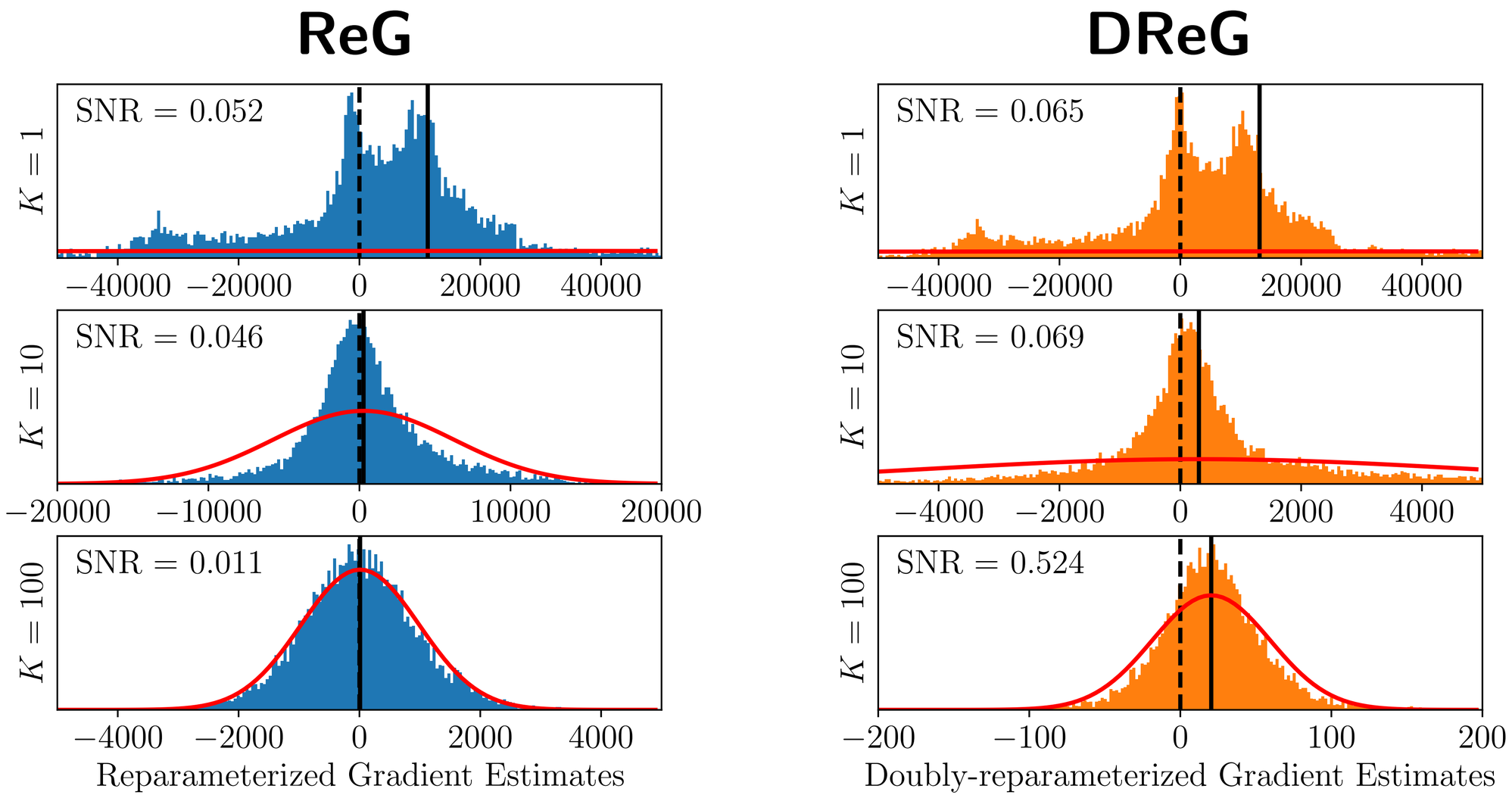
Deep GP Doubly Reparameterized Gradient Estimator

- Adapted to deep GPs from Tucker et al. [2018]:

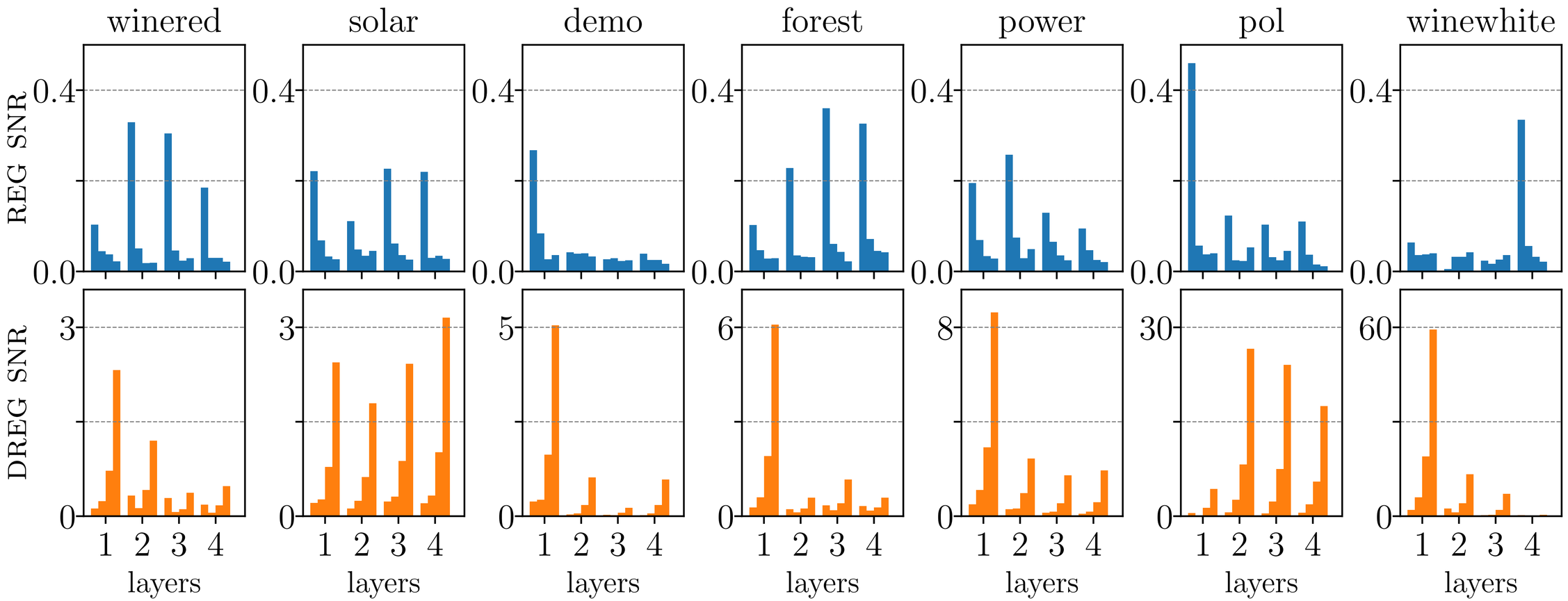
$$\widetilde{\Delta}^{\mathrm{DGP}}_{n,M,K}(\phi) \overset{\mathrm{def}}{=} \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\frac{w_{n,m,k}}{\sum_{j=1}^{K} w_{n,m,j}}\right)^2 \frac{\partial \log w_{n,m,k}}{\partial z_{n,m,k}} \frac{\partial z_{n,m,k}}{\partial \phi}$$

- More importance samples $K \longrightarrow$ higher SNR

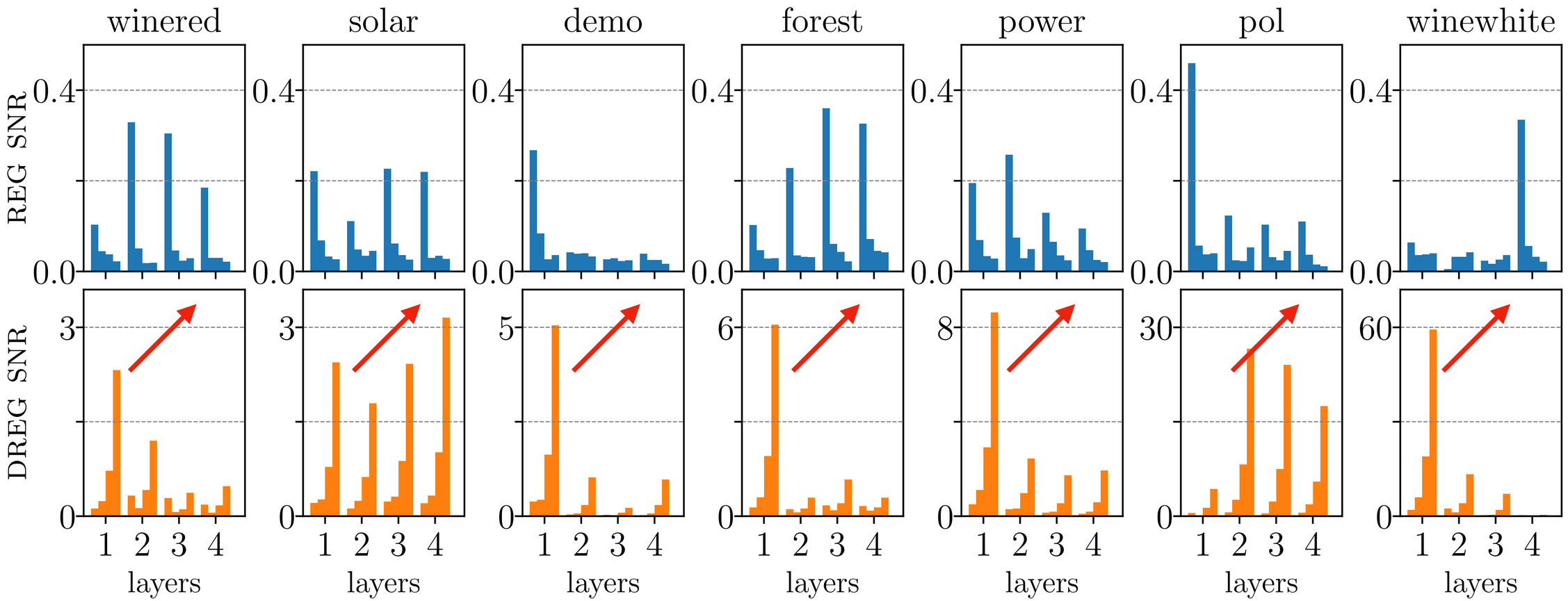
Fixing the Pathology: Empirical Confirmation



Fixing the Pathology: Empirical Confirmation



Fixing the Pathology: Empirical Confirmation



Fixing the Pathology: Empirical Confirmation

Improvement in Predictive Performance

Dataset	Train ELBO ($K = 50$)				Test log-likelihood				Wilcoxon Test
	REG		DREG		REG		DREG		
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	p-value
forest	-97.56	(11.04)	**-92.53**	(10.42)	0.59	(0.08)	**0.63**	(0.08)	0.1%
solar	1657.41	(27.56)	**1707.75**	(42.20)	2.33	(0.17)	**2.57**	(0.11)	2.8%
pol	34610.49	(66.18)	**34665.08**	(70.34)	2.99	(0.01)	2.99	(0.01)	24.7%
power	1510.50	(10.62)	**1515.60**	(10.16)	0.21	(0.01)	0.21	(0.01)	67.3%
winewhite	**-4701.26**	(4.92)	-4703.14	(4.98)	-1.11	(0.01)	-1.11	(0.01)	50.0%
winered	**447.91**	(249.81)	314.75	(216.32)	0.57	(0.27)	**0.61**	(0.20)	41.1%
							Across Datasets:		1.2%

Fixing the Pathology: Empirical Confirmation

Improvement in Predictive Performance

Dataset	Train ELBO ($K = 50$)				Test log-likelihood				Wilcoxon Test
	REG		DREG		REG		DREG		
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	p-value
forest	-97.56	(11.04)	**-92.53**	(10.42)	0.59	(0.08)	**0.63**	(0.08)	0.1%
solar	1657.41	(27.56)	**1707.75**	(42.20)	2.33	(0.17)	**2.57**	(0.11)	2.8%
pol	34610.49	(66.18)	**34665.08**	(70.34)	2.99	(0.01)	2.99	(0.01)	24.7%
power	1510.50	(10.62)	**1515.60**	(10.16)	0.21	(0.01)	0.21	(0.01)	67.3%
winewhite	**-4701.26**	(4.92)	-4703.14	(4.98)	-1.11	(0.01)	-1.11	(0.01)	50.0%
winered	**447.91**	(249.81)	314.75	(216.32)	0.57	(0.27)	**0.61**	(0.20)	41.1%
							Across Datasets:		1.2%

On SNR Issues in Variational Inference for Deep GPs

Summary

- **Deteriorating gradient SNR** in variational inference for deep GPs.
- Fixing the SNR deterioration leads to **improved performance**.
- Improvement comes at **no additional computational cost**.

Thank You!

Tim G. J. Rudner
@timrudner

Oscar Key

Yarin Gal
@yaringal

Tom Rainforth
@tom_rainfroth

Correspondence: tim.rudner@cs.ox.ac.uk

Code: https://github.com/timrudner/snr_issues_in_deep_gps