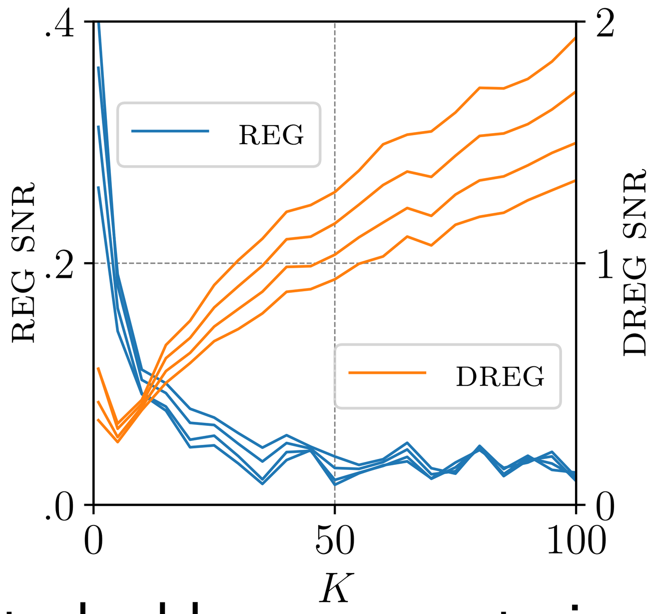
On Signal-to-Noise Ratio Issues in Variational Inference for Deep Gaussian Process

Tim G. J. Rudner[†] *, Oscar Key[‡] *, Yarin Gal[†], Tom Rainforth[†]

[†] University of Oxford [‡] University College London * Equal contribution. Corresponding author: tim.rudner@cs.ox.ac.uk. @timrudner

TL;DR

- We show that the gradient estimates used in training Deep GPs (DGPs) with importance-weighted variational inference are susceptible to signal-to-noise ratio (SNR) issues.
- We demonstrate both theoretically and empirically that the SNR of the gradient estimates for the latent variable's variational parameters *decreases* as the number of importance samples *increases*.



- To address this pathology, adapt doubly-reparameterized gradient estimators to DGP models and show that the resultant estimators completely remedy the SNR issue, thereby providing more reliable training and improved performance.

Model & Inference

$$p(y_n \mid f^{(1)}, f^{(2)}, z_n; x_n) = \mathcal{N}(y \mid f^{(2)}(f^{(1)}([x, z])), \sigma^2 I_P),$$

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

where $\mathcal{F}(x_n, y_n, f_k^{(1)}, z_{n,k}) \stackrel{\text{def}}{=} \exp\left(\mathbb{E}_{q(f^{(2)})}\left[\log p(y_n \mid f^{(2)}, f_k^{(1)}, z_{n,k})\right]\right)$,

$$\Delta^{\mathrm{DGP}}_{n,M,K}(\phi) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \nabla_\phi \log \frac{1}{K} \sum_{k=1}^{K} w_{n,m,k}, \quad (2)$$

where $w_{n,m,k} \stackrel{\text{def}}{=} \frac{\mathcal{F}(x_n, y_n, f_{m,k}^{(1)}, z_{n,m,k}) p(z_{n,m,k})}{q_\phi(z_{n,m,k})}$,

$$z_{n,m,k} \sim q_\phi(z_n), \quad f_{m,k}^{(1)} \sim q(f^{(1)}).$$

SNR Issues in Deep GPs

Theorem 1 (Asymptotic SNR in IWVI for DGPs)**.** *Let $w_{n,m,k}$ be as defined as in $\hat{Z}_{n,m,K} \stackrel{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K} w_{n,m,k}$. Assume that when $M = K = 1$, the expectation and variance of the gradients estimates in Equation (2) are non-zero, and that the first four moments of $w_{n,1,1}$ and $\nabla_\phi w_{n,1,1}$ are all finite and that their variances are also non-zero. Then the signal-to-noise ratio of each $\Delta^{DGP}_{n,M,K}(\phi)$ converges at the following rate*

$$\mathrm{SNR}^{\mathrm{DGP}}_{n,M,K}(\phi) = \sqrt{M} \left| \frac{\nabla_\phi \operatorname{Var}\left[w_{n,1,1}\right] + \mathcal{O}\left(\frac{1}{K}\right)}{2 Z_n \sqrt{K} \sqrt{\operatorname{Var}\left[\nabla_\phi w_{n,1,1}\right]} + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)} \right| = \mathcal{O}\left(\sqrt{M/K}\right), \quad (3)$$

where $Z_n \stackrel{\text{def}}{=} \mathbb{E}[w_{n,1,1}]$ is a lower bound on the marginal likelihood of the n^{th} data point.

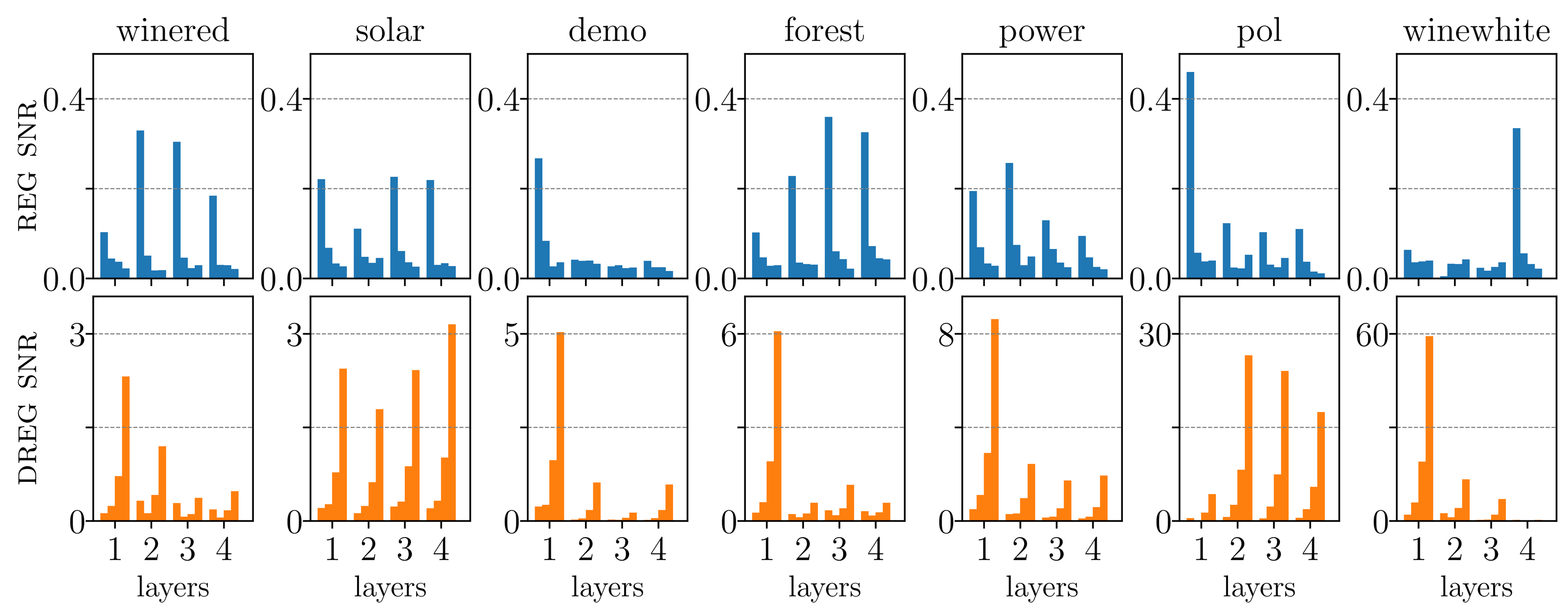


Figure 1: SNR of reparameterization (top row) and doubly reparameterized (bottom row) gradient estimates for shallow GPs and DGPs of 2-4 layers on a selection of real-world datasets. The labels on the x-axes correspond to the depths of the models. The bars for each depth show the SNR for increasing numbers of importance samples, $K = 1, 10, 100, 1000$, from left to right. In the top row, for (D)GPs of any depth, larger K tends to correspond to lower SNRs. In the bottom row, for (D)GPs of any depth, larger K tends to correspond to higher SNRs. Note the difference in y-axis scales across plots in the bottom row.

→ The SNR issue is confirmed theoretically and empirically.

Full paper: https://arxiv.org/abs/2011.00515

Quantifying & Fixing the SNR Pathology

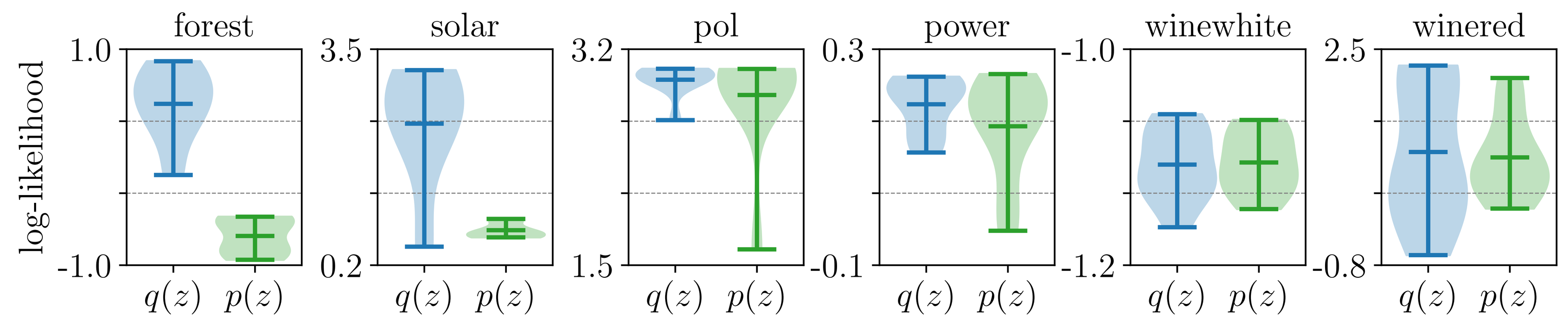


Figure 2: Comparison of predictive performance of 2-layer DGPs with a learned variational distribution over the latent variable (left of each pair, blue) and a variational distribution over the latent variable fixed to the prior (right of each pair, green). The shaded area shows the range of test log-likelihoods over 10 train–test splits, with the width indicating the distribution over the range. The central horizontal lines in each plot show the mean.

Effect of SNR Issue

The more non-Gaussian the data

- the larger the improvement in performance from learning $q_\phi(z)$.
- the larger the improvement in performance from fixing the SNR issue.

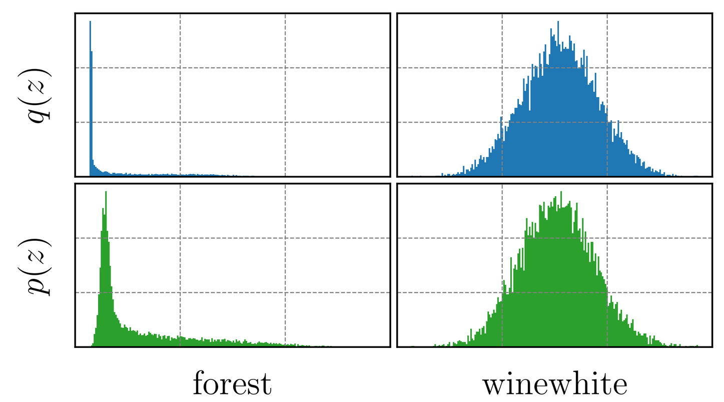


Figure 3: Marginal predictive distributions of 2-layer DGPs with a learned variational distribution over the latent variable (top row, blue) and a variational distribution over the latent variable fixed to the prior (bottom row, green) for randomly selected test points from the 'forest' and 'winewhite' datasets.

Table 1. Comparison of predictive performance of two-layer DGPs trained with REG and DREG estimators. For each dataset, we provide the mean ELBOs on the training dataset and log-likelihoods on the test dataset over 20 random train-test splits as well as the corresponding standard errors. Boldface indicates higher means. The rightmost column shows p-values for one-sided Wilcoxon signed-rank hypothesis tests on the log-likelihoods.

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%

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

This work is based on Salimbeni et al. *Deep Gaussian Processes with Importance-Weighted Variational Inference*. ICML, 2019.