SUPPLEMENTARY MATERIALS:NEURAL OPERATOR VARIATIONAL INFERENCE BASED ON REGULARIZED STEIN DISCREPANCY FOR DEEP GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

A THE SOLUTION TO SVGP AND DSVI

Due to the Gaussian mean-field assumptions, the solution to SVGP has an analytical solution

$$\begin{aligned} & q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{where} \quad & \boldsymbol{\mu} = \boldsymbol{K}_{\boldsymbol{XZ}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{m} \\ & \boldsymbol{\Sigma} = \boldsymbol{K}_{\boldsymbol{XX}} - \boldsymbol{K}_{\boldsymbol{XZ}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}(\boldsymbol{K}_{\boldsymbol{ZZ}} - \boldsymbol{S})\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}} \end{aligned} \tag{1}$$

While performing similarly in DSVI, they have a analytical form for $q(\mathbf{F})$

$$\begin{aligned} & q(\{\mathbf{F}_\ell\}_{\ell=1}^L) = \\ & \prod_{\ell=1}^L \prod_{d=1}^{D_\ell} \int q\left(\mathbf{f}_{\ell,d}|\mathbf{F}_{\ell-1}, \mathbf{u}_{\ell,d}\right) q\left(\mathbf{u}_{\ell,d}\right) d\mathbf{u}_{\ell,d} = \prod_{\ell=1}^L \prod_{d=1}^{D_\ell} \mathcal{N}\left(\mathbf{f}_{\ell,d}|\boldsymbol{\mu}_{\ell,d}, \boldsymbol{\Sigma}_{\ell,d}\right), \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{\ell,d}, \boldsymbol{\Sigma}_{\ell,d}$ is defined as Equation (1).

B PROOF OF THEOREM 1

Theorem 1. *The score function $\nabla_{\mathbf{U}} \log p(\mathbf{U}|\mathcal{D}, \boldsymbol{\nu})$ in Equation (18) can be evaluated by Monte Carlo sampling:*

$$\nabla_{\mathbf{U}} \log p(\mathbf{U}|\mathcal{D}, \boldsymbol{\nu}) = -\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_\ell, \ldots, \boldsymbol{\Delta}_L\right) + \nabla_{\mathbf{U}} \log \sum\nolimits_{s=1}^S p(\boldsymbol{y}|\widehat{\mathbf{F}}_L^{(s)}) \tag{3}$$

where $\boldsymbol{\Delta}_\ell = (\boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,1}, ..., \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,d}, ..., \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,D_\ell})$ and $\widehat{\mathbf{f}}_{\ell,d}^{(s)} \sim \mathcal{N}(\boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\boldsymbol{Z}_\ell}\boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,d}, \boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\widehat{\mathbf{F}}_{\ell-1}} - \boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\boldsymbol{Z}_\ell}\boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\boldsymbol{K}_{\boldsymbol{Z}_\ell\widehat{\mathbf{F}}_{\ell-1}})$ for $\ell = 1, \ldots, L$, S is the number of samples involved in estimation.

Proof. From Bayes' Rule:

$$\log p(\mathbf{U}|\mathcal{D}, \boldsymbol{\nu}) = \log \frac{p(\mathbf{U})p(\mathcal{D}|\mathbf{U}, \boldsymbol{\nu})}{p(\mathcal{D})} = \log p(\mathbf{U}) + \log p(\mathcal{D}|\mathbf{U}, \boldsymbol{\nu}) - \log p(\mathcal{D}), \tag{4}$$

since the prior term $p\left(\mathbf{u}_{\ell,d}\right) = \mathcal{N}\left(0, \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}\right)$, the gradient with $\mathbf{U}$ is a long vector and is tractable:

$$\begin{aligned} & \nabla_{\mathbf{U}} \log p\left(\mathbf{U}\right) \\ & = \nabla_{\mathbf{U}} \log \prod\nolimits_{\ell=1}^L \prod\nolimits_{d=1}^{D_\ell} p\left(\mathbf{u}_{\ell,d}\right) = -\frac{1}{2} \sum\nolimits_{\ell=1}^L \sum\nolimits_{d=1}^{D_\ell} \nabla_{\mathbf{U}} \mathbf{u}_{\ell,d}^T \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \mathbf{u}_{\ell,d} \\ & = -\frac{1}{2}\left(\nabla_{\mathbf{U}_1} \sum\nolimits_{d=1}^{D_1} \mathbf{u}_{1,d}^T \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \mathbf{u}_{1,d}, ..., \nabla_{\mathbf{U}_\ell} \sum\nolimits_{d=1}^{D_\ell} \mathbf{u}_{\ell,d}^T \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \mathbf{u}_{\ell,d}, ..., \nabla_{\mathbf{U}_L} \sum\nolimits_{d=1}^{D_L} \mathbf{u}_{L,d}^T \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \mathbf{u}_{L,d}\right) \\ & = -\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_\ell, \ldots, \boldsymbol{\Delta}_L\right) \end{aligned} \tag{5}$$

where $\boldsymbol{\Delta}_\ell = (\boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,1}, ..., \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,d}, ..., \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1}\mathbf{u}_{\ell,D_\ell})$. The third term of Equation (4) is a constant w.r.t $\mathbf{U}$. We compute the second data likelihood term $\log p(\mathcal{D}|\mathbf{U}, \boldsymbol{\nu})$ using re-parameterization

trick and Monte Carlo method over each layer:

$$\nabla_{\mathbf{U}} \log p(\mathcal{D}|\mathbf{U}, \boldsymbol{\nu}) = \nabla_{\mathbf{U}} \log \int p(\boldsymbol{y}|\mathbf{F}_L) \prod_{\ell=1}^{L} p(\mathbf{F}_\ell|\mathbf{F}_{\ell-1}, \mathbf{U}_\ell) d\mathbf{F}_{\ell-1}$$

$$= \nabla_{\mathbf{U}} \log \mathbb{E}_{p(\mathbf{F}_L|\mathbf{U})}\ p(\boldsymbol{y}|\mathbf{F}_L) \approx \nabla_{\mathbf{U}} \log \sum\nolimits_{s=1}^{S} p(\boldsymbol{y}|\widehat{\mathbf{F}}_L^{(s)}) \tag{6}$$

we draw S samples $\widehat{\mathbf{f}}_{\ell,d}^{(s)}$ from $\widehat{\mathbf{f}}_{\ell,d} \sim p(\mathbf{f}_{\ell,d}|\widehat{\mathbf{F}}_{\ell-1}, \mathbf{u}_{\ell,d})$ for $\ell = 1, \ldots, L$ as

$$\widehat{\mathbf{f}}_{\ell,d} = \boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\boldsymbol{Z}_\ell} \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \mathbf{u}_{\ell,d} + \boldsymbol{\epsilon}_\ell \odot \sqrt{\operatorname{diag}\left(\boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\widehat{\mathbf{F}}_{\ell-1}} - \boldsymbol{K}_{\widehat{\mathbf{F}}_{\ell-1}\boldsymbol{Z}_\ell} \boldsymbol{K}_{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell}^{-1} \boldsymbol{K}_{\boldsymbol{Z}_\ell\widehat{\mathbf{F}}_{\ell-1}}\right)}$$

where $\boldsymbol{\epsilon}^\ell \sim \mathcal{N}(0, \boldsymbol{I}_{D^\ell})$. As a result, we obtain the score function via automatic differentiation:

$$\nabla_{\mathbf{U}} \log p(\mathbf{U}|\mathcal{D}, \boldsymbol{\nu}) = -\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_\ell, \ldots, \boldsymbol{\Delta}_L\right) + \nabla_{\mathbf{U}} \log \sum\nolimits_{s=1}^{S} p(\boldsymbol{y}|\widehat{\mathbf{F}}_L^{(s)}) \tag{7}$$

Moreover, for regression task, let $S = 1$, Equation (7) has a simpler form:

$$\nabla_{\mathbf{U}} \log p\left(\mathbf{U}|\mathcal{D}, \boldsymbol{\nu}\right) = -\left(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_\ell, \ldots, \boldsymbol{\Delta}_L\right) + \frac{1}{\sigma^2} \nabla_{\mathbf{U}} \widehat{\mathbf{F}}_L^{(s)T} \left(\boldsymbol{y} - \widehat{\mathbf{F}}_L^{(s)}\right) \tag{8}$$

where σ^2 is the noise variance. □

C Proof of Theorem 2 and Theorem 3

Definition 1. *Let $p(\boldsymbol{x})$ be a probability density supported on $\mathcal{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$ be a differentiable function, we define Langevin-Stein Operator(LSO) Ranganath et al. (2016)*

$$\mathcal{A}_p \boldsymbol{\phi}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}) + \operatorname{Tr}(\nabla_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x})). \tag{9}$$

Lemma 1. *(Stein's Identity) Liu & Wang (2016) Suppose $p(\boldsymbol{x})$ is a probability density supported on $\mathcal{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$ is a differentiable function satisfying $\int_{\partial\mathcal{X}} p(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) d\boldsymbol{x} = \mathbf{0}$ where $\partial\mathcal{X}$ denotes the boundary of $\mathcal{X}$. If $\mathcal{X}$ is instead all of $\mathbb{R}^d$, then the condition must hold in the limit $r \to \infty$ for integral over the ball B_r of radius r centered at the origin. Then*

$$\mathbb{E}_{\boldsymbol{x}\sim p}\left[\mathcal{A}_p \boldsymbol{\phi}(\boldsymbol{x})\right] = 0 \tag{10}$$

Proof.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}\sim p}[\mathcal{A}_p \boldsymbol{\phi}(\boldsymbol{x})] &= \mathbb{E}_{\boldsymbol{x}\sim p}[\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}) + \operatorname{Tr}(\nabla_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x}))] \\ &= \operatorname{Tr}(\mathbb{E}_{\boldsymbol{x}\sim p}[\boldsymbol{\phi}(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T + \nabla_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x})]) \end{aligned} \tag{11}$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}\sim p}[\boldsymbol{\phi}(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T + \nabla_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x})] &= \int_{\mathcal{X}} p(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^T + p(\boldsymbol{x}) \nabla_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \nabla_{\boldsymbol{x}} (p(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})) d\boldsymbol{x} \end{aligned} \tag{12}$$

From Divergence Theorem:

$$\operatorname{Tr}(\int_{\mathcal{X}} \nabla_{\boldsymbol{x}} (p(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})) d\boldsymbol{x}) = \int_{\mathcal{X}} \operatorname{div}(p(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})) d\boldsymbol{x} = \int_{\partial\mathcal{X}} p(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{n}(\boldsymbol{x}) d\boldsymbol{x} = 0 \tag{13}$$

where $\boldsymbol{n}(\boldsymbol{x})$ is the outward-pointing unit vector on the boundary of $\mathcal{X}$. □

Lemma 2. *Suppose $p(\boldsymbol{x})$, $q(\boldsymbol{x})$ are probability densities supported on $\mathcal{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$ is a differentiable function satisfying $\int_{\partial\mathcal{X}} p(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) d\boldsymbol{x} = \mathbf{0}$ and $\int_{\partial\mathcal{X}} q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) d\boldsymbol{x} = \mathbf{0}$, then*

$$\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p \boldsymbol{\phi}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim q}[(\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}))^T \boldsymbol{\phi}(\boldsymbol{x})] \tag{14}$$

Proof. By Lemma 1,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}\sim q}\left[\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x})+\operatorname{Tr}\left(\nabla_{\boldsymbol{x}}\boldsymbol{\phi}(\boldsymbol{x})\right)\right]=0 \\ &\Rightarrow \mathbb{E}_{\boldsymbol{x}\sim q}\left[\operatorname{Tr}\left(\nabla_{\boldsymbol{x}}\boldsymbol{\phi}(\boldsymbol{x})\right)\right]=-\mathbb{E}_{\boldsymbol{x}\sim q}\left[\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x})\right] \end{aligned} \tag{15}$$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{\phi}(\boldsymbol{x})\right] &= \mathbb{E}_{\boldsymbol{x}\sim q}[\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x})+\operatorname{Tr}(\nabla_{\boldsymbol{x}}\boldsymbol{\phi}(\boldsymbol{x}))] \\ &= \mathbb{E}_{\boldsymbol{x}\sim q}[(\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})-\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x}))^T\boldsymbol{\phi}(\boldsymbol{x})] \end{aligned} \tag{16}$$

□

Lemma 3. *For any $\boldsymbol{a},\boldsymbol{y}\in\mathbb{R}^d$ and $\lambda>0$, the function $\boldsymbol{y}\mapsto\boldsymbol{a}^T\boldsymbol{y}-\lambda\boldsymbol{y}^T\boldsymbol{y}$ achieves its maximum $\frac{1}{4\lambda}\boldsymbol{a}^T\boldsymbol{a}$ if and only if $\boldsymbol{y}=\frac{1}{2\lambda}\boldsymbol{a}$.*

Proof. From Cauchy-Schwarz inequality:

$$\boldsymbol{a}^T\boldsymbol{y}-\lambda\boldsymbol{y}^T\boldsymbol{y}\leqslant\|\boldsymbol{a}\|_2\|\boldsymbol{y}\|_2-\lambda\|\boldsymbol{y}\|_2^2=\frac{1}{4\lambda}\|\boldsymbol{a}\|_2^2-\lambda(\|\boldsymbol{y}\|_2-\frac{1}{2\lambda}\|\boldsymbol{a}\|_2)^2\leqslant\frac{1}{4\lambda}\|\boldsymbol{a}\|_2^2. \tag{17}$$

The equality holds iff $\boldsymbol{y}=\frac{1}{2\lambda}\boldsymbol{a}$. □

Definition 2. *The Fisher divergence Sriperumbudur et al. (2017) between two suitably smooth density functions is defined as*

$$F\left(q,p\right)=\int_{\mathbb{R}^d}\left\|\nabla\log q\left(\boldsymbol{x}\right)-\nabla\log p\left(\boldsymbol{x}\right)\right\|_2^2 q\left(\boldsymbol{x}\right)d\boldsymbol{x}. \tag{18}$$

Theorem 2. *Supposed that the discriminator and the generator network has enough capacity. Training the generator with the optimal discriminator corresponds to minimizing the fisher divergence between p_θ and q. The corresponding optimal loss is*

$$\mathcal{L}\left(\boldsymbol{\theta},\boldsymbol{\nu}\right)=\frac{1}{4\lambda}F\left(q_{\boldsymbol{\theta}}\left(\mathbf{U}\right),p\left(\mathbf{U}|\mathcal{D},\boldsymbol{\nu}\right)\right) \tag{19}$$

Proof. Let our loss function be $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\nu})$, by Lemma 2,

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\nu})=\sup_{\boldsymbol{\eta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{U})}[\mathcal{A}_p\boldsymbol{\phi}_{\boldsymbol{\eta}}\left(\mathbf{U}\right)-\lambda\boldsymbol{\phi}_{\boldsymbol{\eta}}\left(\mathbf{U}\right)^T\boldsymbol{\phi}_{\boldsymbol{\eta}}(\mathbf{U})] \\ &=\sup_{\boldsymbol{\eta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{U})}[(\nabla_{\mathbf{U}}\log p(\mathbf{U}|\mathcal{D},\boldsymbol{\nu})-\nabla_{\mathbf{U}}q_{\boldsymbol{\theta}}\left(\mathbf{U}\right))^T\boldsymbol{\phi}_{\boldsymbol{\eta}}(\mathbf{U})-\lambda\boldsymbol{\phi}_{\boldsymbol{\eta}}\left(\mathbf{U}\right)^T\boldsymbol{\phi}_{\boldsymbol{\eta}}(\mathbf{U})] \end{aligned} \tag{20}$$

The above integral is maximal in function $\boldsymbol{\phi}_{\boldsymbol{\eta}}$ if and only if the integrand is maximal in $\boldsymbol{\phi}_{\boldsymbol{\eta}}\left(\mathbf{U}\right)$ for every $\mathbf{U}$, by Lemma 3,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\nu}) &= \frac{1}{4\lambda}\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{U})}[\|\nabla_{\mathbf{U}}\log p(\mathbf{U}|\mathcal{D},\boldsymbol{\nu})-\nabla_{\mathbf{U}}q_{\boldsymbol{\theta}}\left(\mathbf{U}\right)\|_2^2] \\ &= \frac{1}{4\lambda}F\left(q_\theta\left(\mathbf{U}\right),p\left(\mathbf{U}|\mathcal{D},\boldsymbol{\nu}\right)\right) \end{aligned} \tag{21}$$

The optimal discriminator is:

$$\boldsymbol{\phi}_{\boldsymbol{\eta}^\star}(\mathbf{U})=\frac{1}{2\lambda}\left(\nabla_{\mathbf{U}}\log p(\mathbf{U}|\mathcal{D},\boldsymbol{\nu})-\nabla_{\mathbf{U}}q_{\boldsymbol{\theta}}\left(\mathbf{U}\right)\right) \tag{22}$$

□

Lemma 4. *Suppose $p(\boldsymbol{x})$, $q(\boldsymbol{x})$ are probability densities on $\mathbb{R}^d$ and $\boldsymbol{\phi}:\mathbb{R}^d\to\mathbb{R}^d$ is a differentiable function satisfying $\lim_{\|\boldsymbol{x}\|\to\infty}q\left(\boldsymbol{x}\right)\boldsymbol{\phi}\left(\boldsymbol{x}\right)=\mathbf{0}$, we have*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{\phi}(\boldsymbol{x})\right]\right|\leqslant\sqrt{\mathbb{E}_{\boldsymbol{x}\sim q}\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2}\sqrt{F\left(q,p\right)} \tag{23}$$

Proof. By Lemma 2, we have:

$$|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{\phi}(\boldsymbol{x})\right]| = |\mathbb{E}_{\boldsymbol{x}\sim q}[\left(\nabla_{\boldsymbol{x}}\log p\left(\boldsymbol{x}\right) - \nabla_{\boldsymbol{x}}\log q\left(\boldsymbol{x}\right)\right)^T\boldsymbol{\phi}(\boldsymbol{x})]|. \tag{24}$$

From Cauchy-Schwarz inequality and Hölder's inequality:

$$\begin{aligned}|\mathbb{E}_{\boldsymbol{x}\sim q}[\left(\nabla_{\boldsymbol{x}}\log p\left(\boldsymbol{x}\right) - \nabla_{\boldsymbol{x}}\log q\left(\boldsymbol{x}\right)\right)^T\boldsymbol{\phi}(\boldsymbol{x})]| &\leqslant \mathbb{E}_{\boldsymbol{x}\sim q}\left[\|\nabla_{\boldsymbol{x}}\log p\left(\boldsymbol{x}\right) - \nabla_{\boldsymbol{x}}\log q\left(\boldsymbol{x}\right)\|_2\|\boldsymbol{\phi}(\boldsymbol{x})\|_2\right] \\ &\leqslant \sqrt{\mathbb{E}_{\boldsymbol{x}\sim q}\|\nabla_{\boldsymbol{x}}\log p\left(\boldsymbol{x}\right) - \nabla_{\boldsymbol{x}}\log q\left(\boldsymbol{x}\right)\|_2^2}\sqrt{\mathbb{E}_{\boldsymbol{x}\sim q}\|\boldsymbol{\phi}(\boldsymbol{x})\|_2^2} = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim q}\|\boldsymbol{\phi}(\boldsymbol{x})\|_2^2}\sqrt{F\left(q,p\right)}\end{aligned}$$

□

Definition 3. *Suppose $p(\boldsymbol{x})$ is probability densities on $\mathbb{R}^d$ and $\psi:\mathbb{R}^d\to\mathbb{R}$ is a function, we define $\boldsymbol{\phi}_\psi^p\left(\boldsymbol{x}\right)$ as a solution of the Stein equation $\mathcal{A}_p\boldsymbol{\phi}\left(\boldsymbol{x}\right) = \psi(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p}[\psi(\boldsymbol{x})]$.*

Lemma 5. *Suppose $\psi:\mathbb{R}^d\to\mathbb{R}$ is a bounded function, there exists a bounded solution of the Stein equation.*

Proof. Let $h(\boldsymbol{x}) = \psi(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p}[\psi(\boldsymbol{x})]$, $h(\boldsymbol{x})$ is obviously bounded, then

$$\phi_1\left(\boldsymbol{x}\right) = \frac{1}{p\left(\boldsymbol{x}\right)}\int_{-\infty}^{x_1} p\left(t,x_2,...,x_d\right)h\left(t,x_2,...,x_d\right)dt, \quad \phi_2\left(\boldsymbol{x}\right) = \cdots = \phi_d\left(\boldsymbol{x}\right) = 0 \tag{25}$$

is a bounded solution. □

Lemma 6. *Suppose $p(\boldsymbol{x})$, $q(\boldsymbol{x})$ are probability densities on $\mathbb{R}^d$ and $\boldsymbol{\psi}:\mathbb{R}^d\to\mathbb{R}^n$ is a bounded function. $\forall i\in(1,\ldots,n)$, let $\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)$ be a solution of the Stein equation, then we have*

$$\|\mathbb{E}_{\boldsymbol{x}\sim q}[\boldsymbol{\psi}\left(\boldsymbol{x}\right)] - \mathbb{E}_{\boldsymbol{x}\sim p}[\boldsymbol{\psi}\left(\boldsymbol{x}\right)]\|_2 \leqslant c_{\boldsymbol{\psi}}^{p,q}\sqrt{F\left(q,p\right)} \tag{26}$$

where $c_{\boldsymbol{\psi}}^{p,q} \triangleq \sqrt{\sum_{i=1}^n \mathbb{E}_{\boldsymbol{x}\sim q}\left\|\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)\right\|_2^2}$ is bounded.

Proof. By Lemma 4, we have

$$\begin{aligned}&|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\psi_i(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\psi_i(\boldsymbol{x})\right]| = |\mathbb{E}_{\boldsymbol{x}\sim q}\left[\psi_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\psi_i(\boldsymbol{x})\right]\right]| = \left|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\mathcal{A}_p\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)\right]\right| \\ &\leqslant \sqrt{\mathbb{E}_{\boldsymbol{x}\sim q}\left\|\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)\right\|_2^2}\sqrt{F\left(q,p\right)}.\end{aligned} \tag{27}$$

As a result,

$$\begin{aligned}&\|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\boldsymbol{\psi}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\boldsymbol{\psi}(\boldsymbol{x})\right]\|_2 = \sqrt{\sum\nolimits_{i=1}^n\left|\mathbb{E}_{\boldsymbol{x}\sim q}\left[\psi_i(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p}\left[\psi_i(\boldsymbol{x})\right]\right|^2} \\ &\leqslant \sqrt{\sum\nolimits_{i=1}^n \mathbb{E}_{\boldsymbol{x}\sim q}\|\boldsymbol{\phi}_{x_i}^p\left(\boldsymbol{x}\right)\|_2^2 F\left(q,p\right)} = c_{\boldsymbol{\psi}}^{p,q}\sqrt{F\left(q,p\right)},\end{aligned} \tag{28}$$

where

$$c_{\boldsymbol{\psi}}^{p,q} \triangleq \sqrt{\sum\nolimits_{i=1}^n \mathbb{E}_{\boldsymbol{x}\sim q}\left\|\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)\right\|_2^2} \leqslant \sqrt{\sum\nolimits_{i=1}^n\left\|\boldsymbol{\phi}_{\psi_i}^p\left(\boldsymbol{x}\right)\right\|_\infty^2} \tag{29}$$

is bounded by Lemma 5. □

Theorem 3. *The bias of the estimate of the prediction $\widehat{\mathbf{F}}_L^\star$ in Equation (21) from the DGPs exact evaluation can be bounded by the square root of the Fisher divergence between $q_\theta(\mathcal{U})$ and $p\left(\mathcal{U}|\mathcal{D},\nu\right)$ up to multiplying a constant.*

Proof. From Law of Large Numbers, we have

$$\widehat{\mathbf{F}}_L^\star = \frac{1}{S}\sum\nolimits_{s=1}^S \widehat{\mathbf{F}}_L^{\star(s)} \approx \mathbb{E}_{q\left(\mathbf{F}_L^\star\right)}[\mathbf{F}_L^\star] \tag{30}$$

, where S denotes the number of samples involved in the estimation and $q\left(\mathbf{F}_L^\star\right)$ is represented as:

$$q(\mathbf{F}_L^\star) = \int\prod\nolimits_{\ell=1}^L\prod\nolimits_{d=1}^{D_\ell} p(\mathbf{f}_{\ell,d}^\star|\mathbf{F}_{\ell-1}^\star,\mathbf{u}_{\ell,d})q_{\theta^\star}\left(\mathbf{U}_\ell\right)d\mathbf{F}_{\ell-1}^\star d\mathbf{u}_{\ell,d}. \tag{31}$$

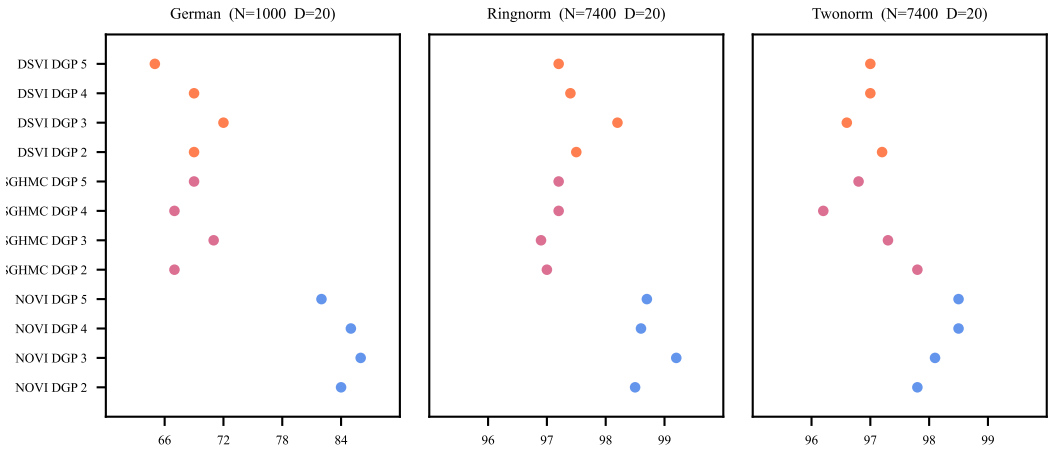

Figure 1: Classification mean test accuracy (%) by our NOVI method (blue), SGHMC (pink) and DSVI (orange) for DGPs on three UCI benchmark datasets. Higher is better.

The DGPs exact evaluation can be written as:

$$\widetilde{\mathbf{F}}_L^\star = \mathbb{E}_{p(\mathbf{F}_L^\star|\mathcal{D},\boldsymbol{\nu})}[\mathbf{F}_L^*]. \tag{32}$$

Similarly:

$$p(\mathbf{F}_L^\star|\mathcal{D},\boldsymbol{\nu}) = \int \prod_{\ell=1}^{L}\prod_{d=1}^{D\ell} p(\mathbf{F}_\ell^\star \mid \mathbf{F}_{\ell-1}^\star, \mathbf{u}_{\ell,d}) p\left(\mathbf{U}_\ell|\mathcal{D},\boldsymbol{\nu}\right) d\mathbf{F}_{\ell-1}^\star d\mathbf{u}_{\ell,d}. \tag{33}$$

By Lemma 6:

$$\begin{aligned}
&\left\|\widehat{\mathbf{F}}_L^\star - \widetilde{\mathbf{F}}_L^\star\right\|_2 \\
=&\| \mathbb{E}_{q(\mathbf{F}_L^\star)}[\mathbf{F}_L^\star] - \mathbb{E}_{p(\mathbf{F}_L^\star|\mathcal{D},\boldsymbol{\nu})}[\mathbf{F}_L^\star] \|_2 \\
=&\| \mathbb{E}_{q(\mathbf{U})}[\int \mathbf{F}_L^\star \prod_{\ell=1}^{L}\prod_{d=1}^{D_\ell} p(\mathbf{f}_{\ell,d}^\star|\mathbf{F}_{\ell-1}^\star,\mathbf{u}_{\ell,d}) d\mathbf{F}_{\ell-1}^\star d\mathbf{F}_L^\star] \\
&- \mathbb{E}_{p(\mathbf{U}|\mathcal{D},\boldsymbol{\nu})}[\int \mathbf{F}_L^\star \prod_{\ell=1}^{L}\prod_{d=1}^{D_\ell} p(\mathbf{f}_{\ell,d}^\star|\mathbf{F}_{\ell-1}^\star,\mathbf{u}_{\ell,d}) d\mathbf{F}_{\ell-1}^\star d\mathbf{F}_L^\star] \|_2 \\
=&\left\|\mathbb{E}_{q(\mathbf{U})}\left[\boldsymbol{\psi}\left(\mathbf{U}\right)\right] - \mathbb{E}_{p(\mathbf{U}|\mathcal{D},\boldsymbol{\nu})}\left[\boldsymbol{\psi}\left(\mathbf{U}\right)\right]\right\|_2 \leqslant c_{\boldsymbol{\psi}}^{p,q}\sqrt{F\left(q\left(\mathbf{U}\right), p\left(\mathbf{U}|\mathcal{D},\boldsymbol{\nu}\right)\right)},
\end{aligned} \tag{34}$$

where $\boldsymbol{\psi}\left(\mathbf{U}\right) = \int \mathbf{F}_L^\star \prod_{\ell=1}^{L}\prod_{d=1}^{D_\ell} p(\mathbf{f}_{\ell,d}^\star|\mathbf{F}_{\ell-1}^\star,\mathbf{u}_{\ell,d}) d\mathbf{F}_{\ell-1}^\star d\mathbf{F}_L^\star$ is obviously bounded.

□

D Additional Results

D.1 UCI Classification Benchmark

We performed classification task on three UCI benchmark datasets, with size ranging from 1000 to 7400. Results are reported in Figure 1 compared through test accuracy as performance metric. It can be seen that NOVI achieves the best results in different sizes of datasets and shows competitive performance within different layers.

D.2 Ablation study on Classification datasets

We also performed ablation study on clasification datasets and reported its results by test accuracy in Figure 2. From which it can be seen that NOVI not only achieves better results on large scale datasets, which demonstrates its scalability, but also the results on the test set have far exceeded the performance of the Monte Carlo log-likelihood maximization method on the training set, suggesting the feasibility of adversarial training.

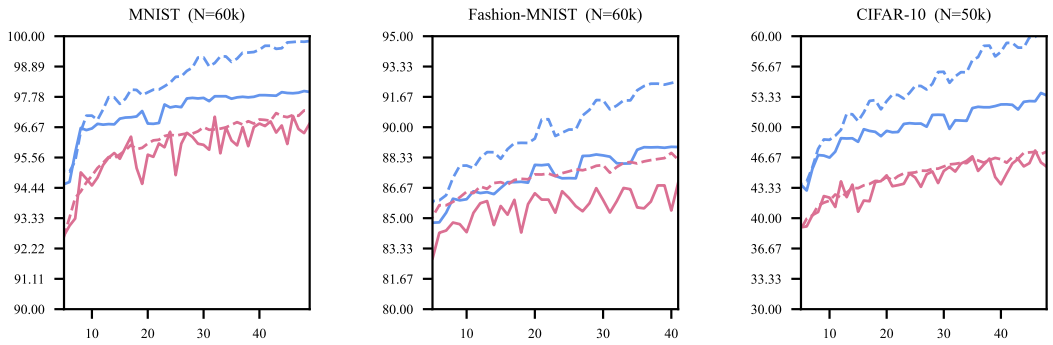

Figure 2: The mean accuracy comparison of NOVI (blue) with Monta Carlo log-likelihood maximization method (pink) using 3-layer DGP model on three image classification datasets. The results of the training and test sets are shown by dashed and solid lines, respectively.

	Concrete	**Energy**	**Boston**	**Kin8nm**
Iteration	500	600	300	500
RMSE (M=50)	0.28 (0.00)	0.04 (0.00)	0.23 (0.00)	0.26 (0.00)
Time (M=50)	0.397s	0.404s	0.380s	0.600s
RMSE (M=100)	0.24 (0.00)	0.04 (0.00)	0.20 (0.01)	0.24 (0.00)
Time (M=100)	0.403s	0.420s	0.400s	0.613s
RMSE (M=200)	0.20 (0.00)	0.03 (0.00)	0.20 (0.00)	0.24 (0.00)
Time (M=200)	0.408s	0.450s	0.410s	0.646s
RMSE (M=400)	0.19 (0.00)	0.03 (0.00)	0.18 (0.01)	0.23 (0.00)
Time (M=400)	0.408s	0.450s	0.420s	0.658s

Table 1: Comparison of number of inducing points (50, 100, 200 and 400) using 2-layer DGP model on 4 UCI regression datasets. M denotes the number of inducing points per layer.

D.3 Tabular version of Figure 2 in the main text

Tabular version of Figure 2 in the main text can be seen in Table 2.

D.4 Comparison about inducing points

In order to investigate the robustness of NOVI at different numbers of induced points, we have performed ablation study to compare accuracy and training time on 4 UCI regression datasets using 2-layer DGP model. For each dataset, number of iteration is set to be the same for fair comparison. Results are shown in Table 1. From which it can be seen that the performance increases gradually with the number of induction points, while the time fluctuates only slightly, which shows the robustness of NOVI to the number of inducing points.

D.5 Additional experiments

We have also performed additional regression experiments for two real-world datasets: Estate and Elevators. Results are shown in Table 3. From these two datasets, it can be seen that NOVI has achieved better PMSE value than other two methods (which demonstrates its robustness to complex real-world problems).

E Training Details

E.1 UCI datasets

Training We have performed a random $0.9/0.1$ train/test split and normalized features to $[-1, 1]$. The depth L of DGP models are varied from 2 to 5 with 100 inducing points per layer, which are

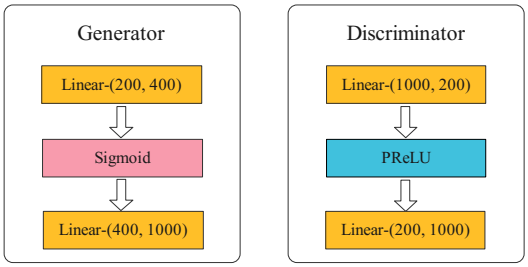

Figure 3: Semantic diagram of generator and discriminator network. The numbers in parentheses indicate the number of input and output neurons respectively.

initialized by sampling from isotrophic Gaussian distribution. The output dimension for each hidden layer is set to 1 for final layer and 10 for others. We have utilized RQ kernel for all tasks. For all datasets, we have optimized hyper-parameters and network parameters jointly and utilized different learning rate, 0.02 for hyper-parameters and 0.001 for network parameters using Adam optimizer Kingma & Ba (2014). The dimension of noise ϵ used to generate $\mathbf{U}$ is set to 200 for all datasets. We train for almost 500 iterations for all datasets. DSVI and SGHMC methods are initialized the same as NOVI to obtain a fair comparison.

Network Settings The generator and discriminator network are all constructed by fully-connected layer, activated by Sigmoid and PReLU function respectively. The schemantic diagram is shown in Figure 3.

E.2 Image datasets

Training We have followed the division of the original dataset and normalized pixel values to $[-1, 1]$. The depth L of DGP models are varied from 3 to 4 with 100 inducing points per layer, which are initialized by sampling from isotrophic Gaussian distribution. The output dimension for each hidden layer is set to be 10 for final layer (which is the exact number of class to predict), and 60 for others. We have utilized RQ kernel for all tasks. For all datasets, we have optimized hyper-parameters and network parameters jointly and utilized different learning rate, 0.02 for hyper-parameters and 0.001 for network parameters using Adam optimizer Kingma & Ba (2014). The dimension of noise ϵ used to generate $\mathbf{U}$ is set to 200 for all datasets. We train for almost $10k$ iterations for all datasets. DSVI and SGHMC methods are initialized the same as NOVI to obtain a fair comparison.

Network Settings The network is constructed in the same way when applied to the UCI dataset, which also can be seen in Figure 3.

E.3 Dataset introduction

We now give a brief introduction to the datasets we have used. All regression datasets are derived from real-world specific problems.

Boston The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts.

Energy It performs energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. It simulates various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses.

Power The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

Concrete Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. It has 1030 instances and 9 variables with 8 quantitative input variables, and 1 quantitative output variable.

Yacht Yacht dataset is used to predict the hydodynamic performance of sailing yachts from dimensions and velocity. It comprises 308 instances and 6 features performed at the Delft Ship Hydromechanics Laboratory.

Qsar This dataset was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of chemicals. to predict acute aquatic toxicity towards Daphnia Magna. LC50 data, which is the concentration that causes death in 50% of test D. magna over a test duration of 48 hours, was used as model response. The model comprised 8 molecular descriptors: TPSA(Tot) (Molecular properties), SAacc (Molecular properties), H-050 (Atom-centred fragments), MLOGP (Molecular properties), RDCHI (Connectivity indices), GATS1p (2D autocorrelations), nN (Constitutional indices), C-040 (Atom-centred fragments).

Protein This is a data set of Physicochemical Properties of Protein Tertiary Structure. The data set is taken from CASP 5-9. There are 45730 decoys and size varying from 0 to 21 armstrong.

Kin8nm This is a data set concerned with the forward kinematics of an 8 link robot arm. Among the existing variants of this data set it has used the variant 8nm, which is known to be highly non-linear and medium noisy.

Estate The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan. The real estate valuation is a regression problem. The data set was randomly split into the training data set ($2/3$ samples) and the testing data set ($1/3$ samples).

Elevators The problem has 18 attributes and this data set is obtained from the task of controlling a F16 aircraft, although the target variable and attributes are different from the ailerons domain. In this case the goal variable is related to an action taken on the elevators of the aircraft.

MNIST The MNIST database, an extension of the NIST database, is a low-complexity data collection of handwritten digits used to train and test various supervised machine learning algorithms. The database contains 70000 28×28 black and white images representing the digits zero through nine. The data is split into two subsets, with 60000 images belonging to the training set and 10000 images belonging to the testing set. The separation of images ensures that given what an adequately trained model has learned previously, it can accurately classify relevant images not previously examined.

Fashion-MNIST Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60000 examples and a test set of 10000 examples. Each example is a 28×28 grayscale image, associated with a label from 10 classes. Zalando intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

CIFAR-10 The CIFAR-10 dataset consists of 60000 32×32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

REFERENCES

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

Data	DSVI 2	DSVI 3	DSVI 4	DSVI 5	SGHMC 2	SGHMC 3	SGHMC 4	SGHMC 5	IPVI 2	IPVI 3	IPVI 4	IPVI 5	NOVI 2	NOVI 3	NOVI 4	NOVI 5
Boston	0.32 (0.02)	0.32 (0.02)	0.32 (0.02)	0.32 (0.02)	0.37 (0.07)	0.38 (0.08)	0.35 (0.09)	0.39 (0.07)	0.35 (0.06)	0.34 (0.05)	0.33 (0.06)	0.32 (0.04)	**0.20 (0.01)**	0.34 (0.02)	0.40 (0.03)	0.38 (0.02)
Energy	0.05 (0.00)	0.05 (0.00)	0.05 (0.00)	0.05 (0.00)	0.13 (0.01)	0.13 (0.01)	0.09 (0.04)	0.13 (0.01)	0.13 (0.01)	0.13 (0.01)	0.12 (0.03)	0.11 (0.04)	**0.04 (0.00)**	0.05 (0.00)	0.06 (0.00)	0.06 (0.00)
Power	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.23 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.01)	0.22 (0.01)	0.22 (0.01)	0.21 (0.01)	0.22 (0.00)	**0.21 (0.00)**	**0.21 (0.00)**	**0.21 (0.00)**
Concrete	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.35 (0.03)	0.33 (0.03)	0.31 (0.02)	0.31 (0.02)	0.32 (0.02)	0.30 (0.03)	0.31 (0.03)	0.30 (0.04)	0.24 (0.00)	0.25 (0.00)	0.24 (0.00)	**0.23 (0.00)**
Yacht	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)	0.03 (0.01)	0.03 (0.01)	**0.02 (0.01)**	0.03 (0.01)	0.03 (0.02)	0.03 (0.02)	0.04 (0.03)	0.03 (0.01)	0.03 (0.00)	0.09 (0.01)	0.08 (0.00)	0.06 (0.00)
Power	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.23 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.01)	0.22 (0.01)	0.22 (0.01)	0.21 (0.01)	0.22 (0.00)	**0.21 (0.00)**	**0.21 (0.00)**	**0.21 (0.00)**
Qsar	0.57 (0.00)	0.50 (0.00)	0.47 (0.00)	**0.42 (0.00)**	0.56 (0.00)	0.56 (0.00)	0.56 (0.00)	0.56 (0.00)	0.56 (0.01)	0.54 (0.01)	0.54 (0.01)	0.54 (0.01)	0.51 (0.00)	0.46 (0.01)	0.45 (0.01)	0.44 (0.01)
Protein	0.81 (0.00)	0.77 (0.00)	0.79 (0.00)	0.73 (0.00)	0.72 (0.01)	0.71 (0.01)	0.70 (0.01)	0.69 (0.00)	0.68 (0.01)	0.65 (0.01)	0.65 (0.01)	0.65 (0.01)	0.67 (0.00)	**0.65 (0.00)**	0.66 (0.00)	0.66 (0.00)
Kin8nm	0.39 (0.00)	0.37 (0.00)	0.34 (0.00)	0.30 (0.00)	0.26 (0.01)	0.25 (0.01)	0.25 (0.01)	0.24 (0.01)	0.25 (0.01)	0.25 (0.01)	0.25 (0.00)	0.26 (0.01)	**0.24 (0.00)**	0.28 (0.00)	0.26 (0.00)	0.27 (0.00)

Table 2: Tabular version of Figure 2 in the main text.

Method	Estate				Elevators			
	L=2	L=3	L=4	L=5	L=2	L=3	L=4	L=5
DSVI	0.65 (0.02)	0.66 (0.02)	0.50 (0.02)	0.64 (0.02)	0.37 (0.00)	0.36 (0.00)	0.37 (0.00)	0.36 (0.00)
SGHMC	0.54 (0.01)	0.50 (0.01)	0.53 (0.01)	0.61 (0.01)	0.36 (0.00)	0.36 (0.00)	**0.35 (0.00)**	**0.35 (0.00)**
NOVI	0.56 (0.02)	0.40 (0.02)	0.40 (0.01)	**0.39 (0.02)**	0.36 (0.00)	**0.35 (0.00)**	**0.35 (0.00)**	**0.35 (0.00)**

Table 3: Additional experiments for real-world datasets. It shows regression mean test RMSE values with its standard deviation on the round bracket. L denotes the number of layers in DGP models.