Deep Variational Implicit Processes

Luis A. Ortega
Simón Rodriguez Santana
Daniel Hernández-Lobato

April 10, 2023

International Conference of Learning Representations

An **implicit stochastic process**[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}),\ \forall n = 1, \ldots, N\,.$$

[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

An implicit stochastic process[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}),\ \forall n = 1, \dots, N\,.$$

Gaussian process

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \textit{and} \quad f(\mathbf{x}_n) = \boldsymbol{L}(\mathbf{x}_n)^T \mathbf{z},\ \forall n = 1, \dots, N\,.$$

[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

An implicit stochastic process[1] (IP) is a collection of random variables $f(\cdot)$ such that any finite collection $\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}$ is implicitly defined by the following generative process:

$$\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z}) \quad \textit{and} \quad f(\mathbf{x}_n) = g_\theta(\mathbf{x}_n, \mathbf{z}), \; \forall n = 1, \ldots, N \,.$$

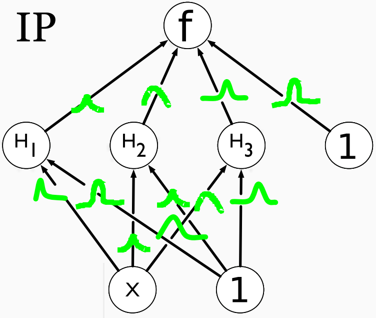
Bayesian Neural Networks.

$$(\mathbf{z}_1, \mathbf{z}_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$$

$$\boldsymbol{h} = r((\boldsymbol{\mu}_1 + \boldsymbol{\sigma}_1 \mathbf{z}_1)^T \mathbf{x}_n) \,.$$

$$g_\theta(\mathbf{x}_n, \mathbf{z}) = (\boldsymbol{\mu}_2 + \boldsymbol{\sigma}_2 \mathbf{z}_2)^T \boldsymbol{h}$$



[1] Ma, C., Li, Y. & Hernandez-Lobato, J.M.. (2019). Variational Implicit Processes.

Variational Implicit Processes

Approximate $P(\mathbf{f})$ with a GP $P_{\mathcal{GP}}(\mathbf{f})$ based on samples $f_1(\cdot), \ldots, f_S(\cdot)$. Setting a standard Gaussian prior $P(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\mathbf{0}, \mathbf{I})$.

$$f(\mathbf{x}) = \hat{m}(\mathbf{x}) + \mathbf{a}^T \hat{\phi}(\mathbf{x}) \implies P_{\mathcal{GP}}(\mathbf{f}) = \mathcal{N}(\hat{m}(\mathbf{x}), \hat{\phi}(\mathbf{x})^T \hat{\phi}(\mathbf{x}))\,.$$

$$\hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{S}} \Big(f_1(\mathbf{x}) - \hat{m}(\mathbf{x}), \ldots, f_S(\mathbf{x}) - \hat{m}(\mathbf{x}) \Big)^{\mathsf{T}}\,.$$

Variational Implicit Processes

Approximate $P(\mathbf{f})$ with a GP $P_{\mathcal{GP}}(\mathbf{f})$ based on samples $f_1(\cdot), \ldots, f_S(\cdot)$. Setting a standard Gaussian prior $P(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\mathbf{0}, \mathbf{I})$.

$$f(\mathbf{x}) = \hat{m}(\mathbf{x}) + \mathbf{a}^T \hat{\phi}(\mathbf{x}) \implies P_{\mathcal{GP}}(\mathbf{f}) = \mathcal{N}(\hat{m}(\mathbf{x}), \hat{\phi}(\mathbf{x})^T \hat{\phi}(\mathbf{x}))\,.$$

$$\hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{S}} \Big(f_1(\mathbf{x}) - \hat{m}(\mathbf{x}), \ldots, f_S(\mathbf{x}) - \hat{m}(\mathbf{x}) \Big)^{\mathsf{T}}\,.$$

- Defines a Gaussian process with a rich tunable kernel.
- Approximates the distribution of an implicit process.

Variational Implicit Processes

Approximate $P(\mathbf{f})$ with a GP $P_{\mathcal{GP}}(\mathbf{f})$ based on samples $f_1(\cdot), \ldots, f_S(\cdot)$. Setting a standard Gaussian prior $P(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\mathbf{0}, \mathbf{I})$.

$$f(\mathbf{x}) = \hat{m}(\mathbf{x}) + \mathbf{a}^T \hat{\phi}(\mathbf{x}) \implies P_{\mathcal{GP}}(\mathbf{f}) = \mathcal{N}(\hat{m}(\mathbf{x}), \hat{\phi}(\mathbf{x})^T \hat{\phi}(\mathbf{x})) \,.$$

$$\hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{S}} \Big(f_1(\mathbf{x}) - \hat{m}(\mathbf{x}), \ldots, f_S(\mathbf{x}) - \hat{m}(\mathbf{x}) \Big)^\mathsf{T} \,.$$

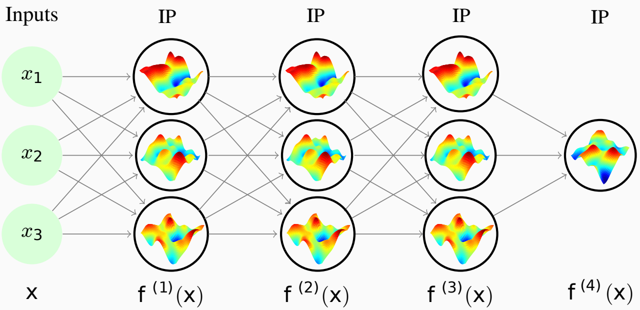
Using a variational distribution $Q(\mathbf{a}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ induces a variational distribution over functions

$$Q(\mathbf{f}) = \mathcal{N}\Big(\hat{m}(\mathbf{x}) + \hat{\phi}(\mathbf{x})^T \mathbf{m}, \hat{\phi}(\mathbf{x})^T \mathbf{S} \hat{\phi}(\mathbf{x})\Big) \,.$$

Deep Variational Implicit Processes

Deep variational implicit processes (DVIPs) are models that consider a **deep implicit process** as the prior for the latent function.

They are a **multi-layer generalization** of IPs.



The input of a layer is the output of the previous one.

ELBO

The evaluation of the ELBO,

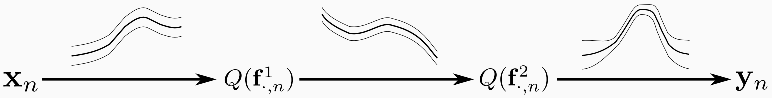
$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})} \left[\log P\big(y_n | \mathbf{f}^L_{\cdot,n}\big) \right] - \sum_{l=1}^{L} \sum_{h=1}^{H_l} \mathsf{KL}\big(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\big) \,.$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})}\left[\log P\left(y_n | \mathbf{f}^L_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathsf{KL}\left(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\right).$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

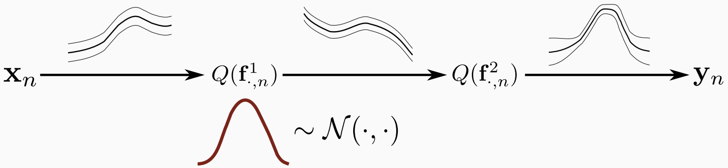


ELBO

The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})}\left[\log P\left(y_n | \mathbf{f}^L_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathrm{KL}\left(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\right).$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

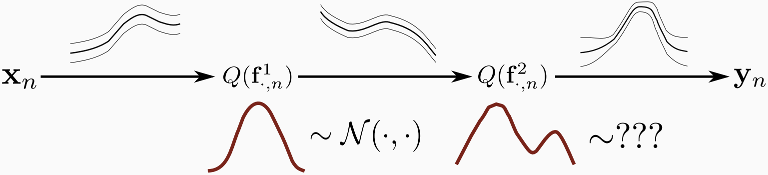


ELBO

The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})}\left[\log P\left(y_n|\mathbf{f}^L_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathsf{KL}\left(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\right).$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

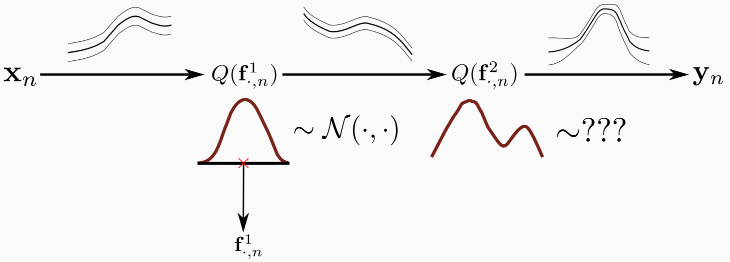


ELBO

The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})}\left[\log P\left(y_n | \mathbf{f}^L_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathrm{KL}\left(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\right).$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

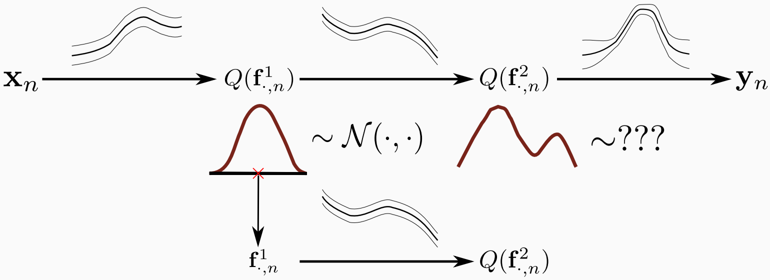


ELBO

The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^L_{\cdot,n})}\left[\log P\left(y_n | \mathbf{f}^L_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathrm{KL}\left(Q(\mathbf{a}^l_h) \mid P(\mathbf{a}^l_h)\right).$$

requires $Q(\mathbf{f}^L_{\cdot,n})$ which is **intractable**.

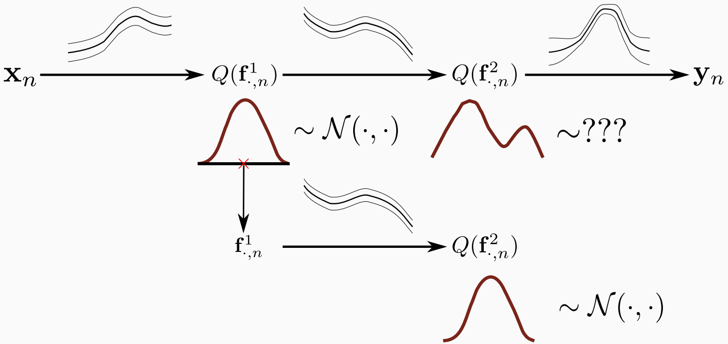


ELBO

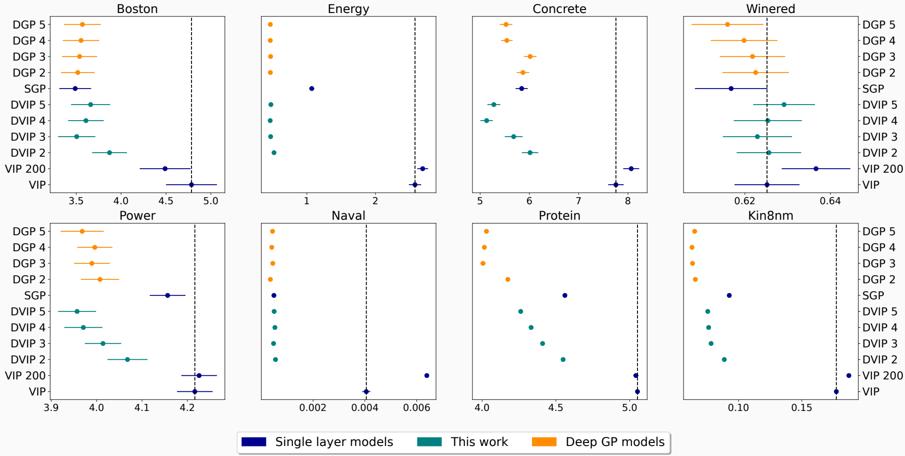
The evaluation of the ELBO,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{Q(\mathbf{f}^{L}_{\cdot,n})}\left[\log P\left(y_n | \mathbf{f}^{L}_{\cdot,n}\right)\right] - \sum_{l=1}^{L}\sum_{h=1}^{H_l} \mathsf{KL}\left(Q(\mathbf{a}_h^l) \mid P(\mathbf{a}_h^l)\right).$$

requires $Q(\mathbf{f}^{L}_{\cdot,n})$ which is **intractable**.

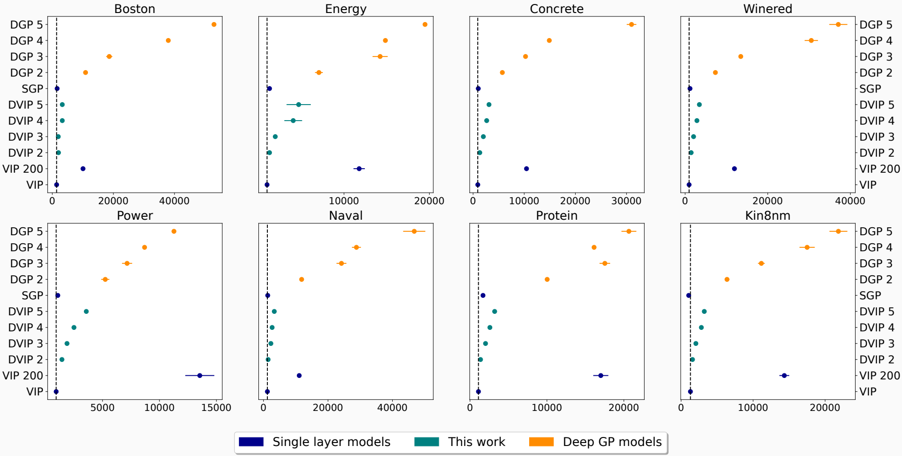


UCI Regression Benchmark (RMSE)



DGPs using the implementation from Salimbeni, H. & Deisenroth, M. (2017). Doubly Stochastic Variational Inference for Deep Gaussian Processes.

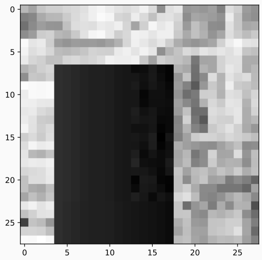
UCI Regression Benchmark (CPU Training Time)



DGPs using the implementation from Salimbeni, H. & Deisenroth, M. (2017). Doubly Stochastic Variational Inference for Deep Gaussian Processes.

Image Classification

Changed IP prior so that the first layer uses **deterministic convolutional layers** and a **Bayesian fully connected layer**.



	SGP	VIP	DVIP 2	DVIP 3	DGP 3
Accuracy (%)	73.64	85.50	87.92	**88.40**	77.18
Likelihood	−0.526	−0.349	−0.294	**−0.280**	−0.472
AUC	0.826	0.931	0.952	**0.956**	0.857

Conclusions

- Implicit processes define a richer distribution over functions.

Conclusions

- Implicit processes define a **richer distribution over functions**.
- DVIPs can be used on a variety of regression and classification problems with **no need of hand-tuning**.

Conclusions

- Implicit processes define a **richer distribution over functions**.
- DVIPs can be used on a variety of regression and classification problems with **no need of hand-tuning**.
- DVIPs **surpass or match the performance** of single layer VIPs and DGPs.

Conclusions

- Implicit processes define a **richer distribution over functions**.
- DVIPs can be used on a variety of regression and classification problems with **no need of hand-tuning**.
- DVIPs **surpass or match the performance** of single layer VIPs and DGPs.
- DVIPs do not over-fit by increasing the depth.

Conclusions

- Implicit processes define a **richer distribution over functions**.
- DVIPs can be used on a variety of regression and classification problems with **no need of hand-tuning**.
- DVIPs **surpass or match the performance** of single layer VIPs and DGPs.
- DVIPs **do not over-fit** by increasing the depth.
- **Increasing the number of layers is far more effective than increasing the complexity** of the prior of single-layer VIPs.

Conclusions

- Implicit processes define a **richer distribution over functions**.
- DVIPs can be used on a variety of regression and classification problems with **no need of hand-tuning**.
- DVIPs **surpass or match the performance** of single layer VIPs and DGPs.
- DVIPs **do not over-fit** by increasing the depth.
- **Increasing the number of layers is far more effective than increasing the complexity** of the prior of single-layer VIPs.
- The use of **domain specific priors** has demonstrated to give outstanding results compared to other GP methods.

Thank you for your attention!