2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–9

Structured Semi-Implicit Variational Inference

Iuliia Molchanova[1,*] YUKOVA.31@GMAIL.COM
Dmitry Molchanov[1,3,*] DMOLCH111@GMAIL.COM
Novi Quadrianto[2] NOVI.QUADRIANTO@GMAIL.COM
Dmitry Vetrov[1,3] VETROVD@YANDEX.RU
[1] *Samsung AI Center Moscow* [2] *Predictive Analytics Lab, University of Sussex* [*] *Equal contribution*
[3] *National Research University Higher School of Economics, Samsung-HSE Joint Lab*

Abstract

In this work we construct flexible joint distributions from low-dimensional conditional semi-implicit distributions. Explicitly defining the structure of the approximation allows to make the variational lower bound tighter, resulting in more accurate inference.

1. Introduction

Many recent advances in variational inference have been focused on different ways to estimate or bound the KL divergence between two complicated distributions. They made it possible to perform variational inference with hierarchical distributions (Ranganath et al., 2016; Titsias and Ruiz, 2018; Sobolev and Vetrov, 2019), semi-implicit distributions (Yin and Zhou, 2018; Molchanov et al., 2019) and even fully implicit distributions (Mescheder et al., 2017; Shi et al., 2017; Huszár, 2017). While these methods work well for low-dimensional cases, they can misbehave when the dimensionality of the problem grows.

In this work, we focus on semi-implicit variational inference, and consider structured multi-dimensional distributions. We show that taking this structure into account, we can obtain a much tighter entropy bound and, consequentially, a much tighter evidence lower bound. We also demonstrate that structured semi-implicit variational inference can successfully capture the multi-modal nature of the posterior distribution in deep Gaussian processes, and show a way to construct and learn an autoregressive semi-implicit model.

2. Semi-Implicit Variational Inference

Variational inference provides a way to approximate the generally intractable posterior distribution $p(z \mid \mathcal{D})$ in a probabilistic model with a parametric approximation $q_\phi(z)$. It typically requires the variational distribution $q_\phi(z)$ to be reparameterizable and have a tractable log-density (Kingma and Welling, 2013).

Semi-implicit variational inference (Yin and Zhou, 2018; Molchanov et al., 2019) extends this framework to so-called *semi-implicit* distributions. By mixing a simple explicit distribution $q_\phi(z \mid \epsilon)$ with an implicit distribution $q(\epsilon)$ one obtains a so-called semi-implicit distribution with a generally intractable marginal density $q_\phi(z)$.

$$q_\phi(z) = \int q_\phi(z \mid \epsilon) q(\epsilon) d\epsilon. \tag{1}$$

A typical example of a semi-implicit model uses a Gaussian conditional distribution $q_\phi(z \mid \epsilon) = \mathcal{N}(z \mid \mu_\phi(\epsilon), \text{diag}(\sigma^2_\phi(\epsilon)))$, parameterized by neural networks $\mu_\phi(\epsilon)$ and $\sigma^2_\phi(\epsilon)$, whereas $q(\epsilon)$ can be any fixed distribution that allows for efficient sampling[1]. Unlike methods such as HVI, UIVI or IWHVI, SIVI does not need to access the density $q(\epsilon)$.

The main idea behind semi-implicit variational inference, or SIVI, is to use $K+1$-sample estimates in order to obtain a lower bound on the entropy of such distribution. Since all variables follow distribution $q(\cdot)$, we omit it for brevity.

$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}^{\text{SIVI}}_K[q_\phi(z)] = -\mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{z \mid \epsilon^0} \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \epsilon^k). \tag{2}$$

This entropy bound can be used to construct a proper variational objective by bounding the KL-term in the evidence lower bound.

3. Structured Semi-Implicit Variational Inference

As with most implicit variational inference algorithms, the performance of semi-implicit variational inference can quickly degrade as the number of dimensions grows. As SIVI essentially approximates a multi-dimensional distribution $q_\phi(z)$ with a mixture of $K+1$ Gaussian distributions in order to bound its entropy, it may require an exponentially large mixture size K to obtain an adequate approximation. In order to solve this problem, we propose to factorize a high-dimensional joint semi-implicit distribution into a product of low-dimensional conditional semi-implicit distributions. Here and after we abuse the notation and assume $z_{1..0}$ to denote an empty set.

$$q_\phi(z) = q_\phi(z_1) \prod_{i=2}^{d} q_\phi(z_i \mid z_{1..i-1}), \quad q_\phi(z_i \mid z_{1..i-1}) = \int q_\phi(z_i \mid z_{1..i-1}, \epsilon_i) q(\epsilon_i) d\epsilon_i. \tag{3}$$

In this case the entropy bounds can be written as follows:

$$\underline{\mathcal{H}}^{\text{SSIVI}}_K[q_\phi(z)] = -\sum_{i=1}^{d} \mathbb{E}_{z_{1..i-1}} \mathbb{E}_{\epsilon_i^{0..K}} \mathbb{E}_{z_i \mid z_{1..i-1}, \epsilon_i^0} \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z_i \mid z_{1..i-1}, \epsilon_i^k). \tag{4}$$

This way we would only need to model low-dimensional semi-implicit distributions while still recovering a non-trivial joint distribution. We provide two examples of models that follow such structure in Section 4.

It can be shown that given the same joint distribution $q_\phi(z)$, taking the structure into account results in a tighter bound:

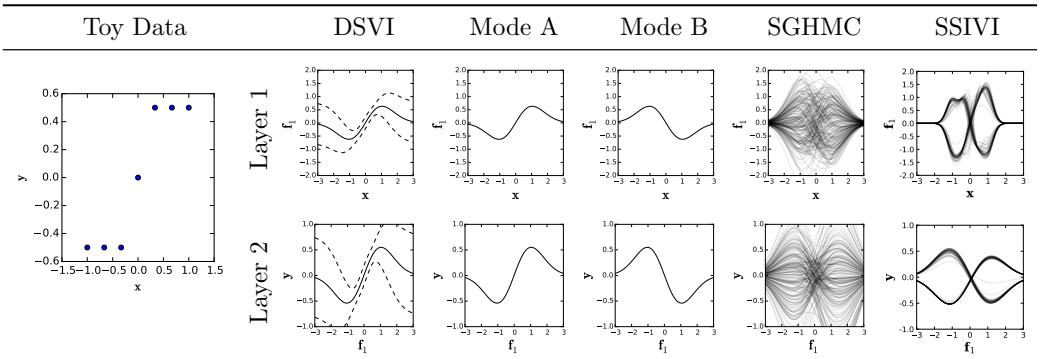
Theorem 1 *For a structured semi-implicit distribution (3), the following inequalities hold:*

$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}^{\text{SSIVI}}_K[q_\phi(z)] \geq \underline{\mathcal{H}}^{\text{SIVI}}_K[q_\phi(z)]. \tag{5}$$

We provide the proof of the theorem in appendix A. The main idea is to show that the structured SIVI bound with $K+1$ samples of the mixing variable ϵ essentially approximates the marginal distribution $q_\phi(z)$ with an exponentially large mixture of $(K+1)^d$ distributions, placed on a d-dimensional grid.

1. In general, $q(\epsilon)$ can be any parametric reparameterizable distribution. For simplicity, we incorporate all its parameters as well as the reparameterizing transformation into the conditional distribution $q_\phi(z \mid \epsilon)$.

Figure 1: Visualisation of layers in a two-layer DGP. All columns except "SSIVI" are taken from Havasi et al. (2018). They show the mean and the standard deviation of the variational posterior under DSVI, two MAP solutions under Mode A and Mode B, and the posterior function samples under SGHMC and SSIVI.



4. Experiments

4.1. Deep Gaussian Processes

We apply SSIVI to deep Gaussian processes (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017). For detailed formulation of the deep GP model follow Appendix B.

Conventional variational inference techniques for DGPs assume a Gaussian posterior approximation $q_\phi(u^{1..L}) = \prod_{l=1}^{L} \mathcal{N}(u^l \,|\, \mu_l, \Sigma_l)$ over the inducing values $u^{1..L}$ that factorizes across the layers. Havasi et al. (2018) show that in practice the true posterior over the inducing values does not in fact factorize across layers, is non-Gaussian and is multimodal. We get rid of all these limiting assumptions by using the following structured semi-implicit posterior approximation:

$$q_\phi(u^{1..L}) = q_\phi(u^1) \prod_{l=2}^{L} q_\phi(u^l \,|\, u^{l-1}), \quad q_\phi(u^l \,|\, u^{l-1}) = \int q_\phi(u^l \,|\, \epsilon^l, u^{l-1}) q(\epsilon^l) d\epsilon^l. \tag{6}$$

We use a fully-factorized Gaussian conditional distribution $q_\phi(u^l \,|\, \epsilon^l, u^{l-1})$ with means $\mu_\phi^l(\epsilon^l, u^{l-1})$ and scales $\sigma_\phi^l(\epsilon^l, u^{l-1})$ parameterized by neural networks. Using the structured SIVI bound (4), we obtain the final variational objective for SSIVI-DGP (see Appendix C for more details).

To demonstrate that SSIVI allows to recover multimodal posteriors with cross-layer dependencies, we consider the toy problem, proposed by Havasi et al. (2018). It is a noise-free (the likelihood variance is set to zero) regression problem consisting of seven training datapoints. There are two natural modes in the posterior space, denoted Mode A and Mode B in the plots in Figure 1.

We use SSIVI bound (35) with $K = 100$ and perform 3000 Adam (Kingma and Ba, 2014) updates with default hyperparameters and the learning rate set to 5×10^{-3}. We use seven inducing inputs on the first layer, fixed at the training point locations, and two

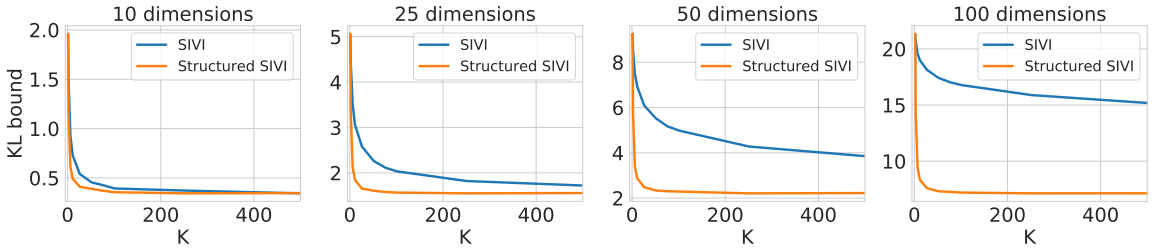


Figure 2: SIVI and SSIVI KL bounds for an autoregressive semi-implicit model and a synthetic multi-dimensional distribution. As expected, SSIVI always outperforms SIVI, and the gap increases with the number of dimensions.

inducing inputs on the second layers, fixed at 1 and -1. Means and variances of the Gaussian conditional distributions $q_\phi(u^l \,|\, \epsilon^l, u^{l-1})$ are modeled by fully-connected neural networks with three hidden layers of 100 neurons each, and mixing variables ϵ^l are sampled from 100-dimensional standard Gaussian distributions.

As shown in Figure 1, DSVI (Salimbeni and Deisenroth, 2017) converges into one of them depending on the randomness in the initialization and the stochastic optimization process. On the contrary, both SGHMC and SSIVI allow to capture all modes and the inter-layer dependencies.

4.2. Auto-Regressive Semi-Implicit Generator

We implement the structured semi-implicit distribution (3) in a general case using a recurrent neural network. The generative process looks as follows:

$$h_1 = h(0, 0), \qquad \epsilon_1 \sim \mathcal{N}(\epsilon_1 \,|\, 0, 1), \qquad z_1 \sim \mathcal{N}(z_1 \,|\, \mu(h_1, \epsilon_1), \sigma^2(h_1, \epsilon_1)), \tag{7}$$

$$h_i = h(z_{i-1}, h_{i-1}), \qquad \epsilon_i \sim \mathcal{N}(\epsilon_i \,|\, 0, 1), \qquad z_i \sim \mathcal{N}(z_i \,|\, \mu(h_i, \epsilon_i), \sigma^2(h_i, \epsilon_i)). \tag{8}$$

In our experiments $h(\cdot, \cdot)$ is defined by two stacked GRU cells, and $\mu(\cdot, \cdot)$ and $\sigma^2(\cdot, \cdot)$ are defined as a fully-connected neural network with three hidden layers that outputs the mean and the log-scale of the one-dimensional Gaussian distribution. All mixing variables ϵ_i are scalar and follow the standard Gaussian distribution. The width of all layers (both recurrent and fully-connected) is 100.

We train this model to generate samples from a synthetic multi-dimensional structured distribution $p(z) = \text{Laplace}(z_1 \,|\, 0, 1) \prod_{i=2}^{d} \text{Laplace}(z_i \,|\, z_{i-1}, 1)$. To do this, we minimize the structured SIVI bound on the KL divergence $\text{KL}\,(q_\phi(z) \,\|\, p(z))$ with $K = 100$. We use the SSIVI entropy bound (4) and estimate the cross-entropy using the reparameterization trick. We perform 10000 steps with Adam with standard hyperparameters.

As one can see from Figure 2, SSIVI provides a much tighter bound, and the gap between SIVI and SSIVI increases as the number of dimensions grows.

Acknowledgements

Novi Quadrianto has been supported by the UK EPSRC project EP/P03442X/1. Dmitry Vetrov has been supported by the Russian Science Foundation grant no.19-71-30020.

References

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Marton Havasi, José Miguel Hernández-Lobato, and Juan José Murillo-Fuentes. Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. In *Advances in Neural Information Processing Systems*, pages 7517–7527, 2018.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2391–2400. JMLR. org, 2017.

Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. *AISTATS*, 2019.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

Artem Sobolev and Dmitry Vetrov. Importance weighted hierarchical variational inference. *arXiv preprint arXiv:1905.03290*, 2019.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*, 2018.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.

Appendix A. Structured Semi-Implicit Variaitonal Inference: entropy bound

Theorem 1 *For a structured semi-implicit distribution (3), the following inequalities hold:*

$$\mathcal{H}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] \geq \underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)] \tag{9}$$

Proof The first inequality, namely the fact that $\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)]$ is indeed a lower bound on the entropy $\mathcal{H}[q_\phi(z)]$, can be proven in exactly the same fashion as the corresponding proof for the original SIVI objective by Molchanov et al. (2019).
The following proves the second inequality. Firstly, let's rewrite the SIVI bound $\underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)]$.

$$-\underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)] = \mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{\underbrace{\prod_{j=1}^d q_\phi\left(z_j \,|\, z_{1..j-1}, \epsilon_j^0\right)}_{q_\phi(z \,|\, \epsilon^0)}} \log \underbrace{\frac{1}{K+1} \sum_{k=0}^K \underbrace{\prod_{i=1}^d q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^k\right)}_{q_\phi\left(z \,|\, \epsilon^k\right)}}_{\tilde{q}_\phi(z \,|\, \epsilon^{0..K})} \tag{10}$$

$$= \mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{\tilde{q}_\phi(z \,|\, \epsilon^{0..K})} \log \tilde{q}_\phi(z \,|\, \epsilon^{0..K}) \tag{11}$$

Transition to line (11) holds since ϵ^k are independent and identically distributed, making the expectations $\mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{q_\phi(z \,|\, \epsilon^0)}[\cdot]$ and $\mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{q_\phi(z \,|\, \epsilon^i)}[\cdot]$ identical for all $i = 0..K$ and equal to $\mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{\tilde{q}_\phi(z \,|\, \epsilon^{0..K})}[\cdot]$.

Now let's expand the SSIVI bound $\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)]$.

$$-\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] = \sum_{i=1}^d \mathbb{E}_{z_{1..i-1}} \mathbb{E}_{\epsilon_i^{0..K}} \mathbb{E}_{z_i | z_{1..i-1}, \epsilon_i^0} \log \frac{1}{K+1} \sum_{k=0}^K q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^k\right) \tag{12}$$

$$= \mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{z_1 | \epsilon_1^0;\ z_2 | z_1, \epsilon_2^0;\ \ldots;\ z_d | z_{1..d-1}, \epsilon_d^0} \sum_{i=1}^d \log \frac{1}{K+1} \sum_{k=0}^K q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^k\right) \tag{13}$$

$$= \mathbb{E}_{\epsilon^{0..K}} \mathbb{E}_{z_1 | \epsilon_1^0;\ z_2 | z_1, \epsilon_2^0;\ \ldots;\ z_d | z_{1..d-1}, \epsilon_d^0} \log \frac{1}{(K+1)^d} \prod_{i=1}^d \sum_{k=0}^K q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^k\right) \tag{14}$$

We can expand the product of d sums in eq. (14) into a sum of $(K+1)^d$ products of form $\prod_{i=1}^d q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^{\omega_i}\right)$, where $\omega \in \{0..K\}^d$. We thus obtain a mixture of $(K+1)^d$ distributions that we denote as $\hat{q}_\phi(z \,|\, \epsilon^{0..K})$. Similarly to eq. (11), we can rewrite the expectation in eq. (14) as an expectation over $\tilde{q}_\phi(z \,|\, \epsilon^{0..K})$ since $\hat{q}_\phi(z \,|\, \epsilon^{0..K})$ is also invariant

to permutation of ϵ^k. This way we can rewrite the SSIVI bound as follows:

$$-\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] = \mathbb{E}_{\epsilon^{0..K}}\mathbb{E}_{z_1|\epsilon_1^0;\ z_2|z_1,\epsilon_2^0;\ \ldots;\ z_d|z_{1..d-1},\epsilon_d^0} \log \underbrace{\frac{1}{(K+1)^d}\sum_{\substack{\omega\in\\ \{0..K\}^d}}\prod_{i=1}^{d} q_\phi\left(z_i \,|\, z_{1..i-1}, \epsilon_i^{\omega_i}\right)}_{\hat{q}_\phi(z\,|\,\epsilon^{0..K})} \tag{15}$$

$$= \mathbb{E}_{\epsilon^{0..K}}\mathbb{E}_{\tilde{q}_\phi(z\,|\,\epsilon^{0..K})} \log \hat{q}_\phi\left(z \,|\, \epsilon^{0..K}\right) \tag{16}$$

We can finally write down the gap between the SIVI and the SSIVI bounds:

$$\underline{\mathcal{H}}_K^{\text{SSIVI}}[q_\phi(z)] - \underline{\mathcal{H}}_K^{\text{SIVI}}[q_\phi(z)] = \mathbb{E}_{\epsilon^{0..K}}\mathbb{E}_{\tilde{q}_\phi(z\,|\,\epsilon^{0..K})}\left(-\log \hat{q}_\phi\left(z \,|\, \epsilon^{0..K}\right) + \log \tilde{q}_\phi\left(z \,|\, \epsilon^{0..K}\right)\right) \tag{17}$$

$$= \mathbb{E}_{\epsilon^{0..K}}\mathrm{KL}\left(\tilde{q}_\phi\left(z \,|\, \epsilon^{0..K}\right) \,\|\, \hat{q}_\phi\left(z \,|\, \epsilon^{0..K}\right)\right) \geq 0 \tag{18}$$

This concludes proof of the theorem. ■

Appendix B. Sparse Deep Gaussian Processes

This section provides a brief overview of the definition of the DGP model, closely following DSVI. A conventional single-output Gaussian Process model is defined as follows:

$$p(y, f \,|\, x, \theta) = \underbrace{p(f \,|\, x, \theta)}_{\text{GP prior}} \prod_{i=1}^{N} \underbrace{p(y_i \,|\, f_i, \theta)}_{\text{Likelihood}} \tag{19}$$

Here $p(f \,|\, x, \theta)$ is the Gaussian process prior, which is typically a zero-mean Gaussian distribution with the covariance matrix defined by a covariance function $k_\theta(\cdot, \cdot)$ (we denote it as $K_{\cdot,\cdot}$ for brevity); $y \in \mathbb{R}^N, f \in \mathbb{R}^N, x \in \mathbb{R}^{N\times D}$. The training of the parameters θ of the prior and the likelihood is performed using maximum marginal likelihood:

$$\log p(y \,|\, x, \theta) = \log \int p(y, f \,|\, x, \theta) df \to \max_\theta \tag{20}$$

In order to reduce the required complexity, the sparse GP model (Titsias, 2009) introduces auxiliary variables, inducing inputs $z \in \mathbb{R}^{M\times D}$ and values $u \in \mathbb{R}^M$:

$$p(y, f, u \,|\, x, z, \theta) = p(f \,|\, u, x, z, \theta) p(u \,|\, z, \theta) \prod_{i=1}^{N} p(y_i \,|\, f_i, \theta) \tag{21}$$

In sparse GPs, direct maximization of the marginal likelihood is replaced with maximization of its lower bound (ELBO):

$$\mathbb{E}_{q_\phi(f,u\,|\,x,z,\theta)} \log p(y \,|\, f, \theta) - \mathrm{KL}(q_\phi(f, u \,|\, x, z, \theta) \,\|\, p(f, u \,|\, x, z, \theta)) \to \max_{z,\theta,\phi} \tag{22}$$

The approximate posterior $q_\phi(f, u \,|\, x, z, \theta)$ is specifically designed to reduce the computational complexity by cancelling out the most computation-heavy term $p(f \,|\, u, x, z, \theta)$:

$$q_\phi(f, u \,|\, x, z, \theta) = p(f \,|\, u, x, z, \theta) q_\phi(u) \tag{23}$$

$$q_\phi(u) = \mathcal{N}(u \,|\, m, S); \quad \phi = \{m, S\} \tag{24}$$

$$q_\phi(f_i \,|\, x_i, z, \theta) = \int p(f_i \,|\, u, x_i, z, \theta) q_\phi(u) du = \mathcal{N}(f_i \,|\, \mu_i, \sigma_i^2) \tag{25}$$

$$\mu_i = K_{x_i z} K_{zz}^{-1} m \tag{26}$$

$$\sigma_i^2 = K_{x_i x_i} - K_{x_i z} K_{zz}^{-1} (K_{zz} - S) K_{zz}^{-1} K_{z x_i} \tag{27}$$

$$\mathbb{E}_{q_\phi(f, u \,|\, x, z, \theta)} \log p(y_i \,|\, f_i, \theta) = \mathbb{E}_{q_\phi(f_i \,|\, x_i, z, \theta)} \log p(y_i \,|\, f_i, \theta) \tag{28}$$

This reduces the lower bound (22) to the following sparse GP ELBO:

$$\mathcal{L}_{SGP} = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(f_i \,|\, x_i, z, \theta)} \log p(y_i \,|\, f_i, \theta) - \mathrm{KL}(q_\phi(u \,|\, z, \theta) \,\|\, p(u \,|\, z, \theta)) \to \max_{z, \theta, \phi}, \tag{29}$$

which allows for doubly stochastic optimization.

A deep Gaussian process Damianou and Lawrence (2013); Salimbeni and Deisenroth (2017) is constructed as a chain of multi-output sparse Gaussian processes, or GP layers. The output of each GP is considered as an input to the next GP. The deep GP probabilistic model is defined similarly to conventional sparse GPs. For each GP layer l, we have an output variable f^l and a set of values u^l corresponding to the inducing inputs z^l The joint distribution over these variables is defined as follows (for brevity, we denote $f^0 := x$):

$$p(y, f^{1..L}, u^{1..L} \,|\, x, z^{1..L}, \theta) = \prod_{l=1}^{L} p(f^l \,|\, f^{l-1}, u^l, z^l, \theta) p(u^l \,|\, z^l, \theta) \prod_{i=1}^{N} p(y_i \,|\, f_i^L, \theta) \tag{30}$$

Similarly to sparse GPs, assuming a specific posterior approximation $q_\phi(f^l, u^l \,|\, f^{l-1}, z^l, \theta) = q_\phi(u^l) p(f^l \,|\, u^l, f^{l-1}, z^l, \theta)$, training the DGP involves bounding the data log marginal likelihood $\log p(y \,|\, x, \theta)$ with the following variational lower bound, and then maximizing it w.r.t. both the variational parameters ϕ and model parameters θ and z:

$$\mathcal{L}_{DGP} = \sum_{i=1}^{N} \mathbb{E}_{\prod_{l=1}^{L} q_\phi(f_i^l \,|\, f_i^{l-1}, z^l, \theta)} \log p(y_i \,|\, f_i^L, \theta) - \sum_{l=1}^{L} \mathrm{KL}(q_\phi(u^l \,|\, z^l, \theta) \,\|\, p(u^l \,|\, z^l, \theta)) \to \max_{z^{1..L}, \theta, \phi} \tag{31}$$

Appendix C. SSIVI for Sparse DGP

We substitute the factorized Gaussian approximation used in DSVI with a structured semi-implicit distribution:

$$q_\phi(u^{1..L}) = q_\phi(u^1) \prod_{l=2}^{L} q_\phi(u^l \,|\, u^{l-1}), \tag{32}$$

$$q_\phi(u^l \,|\, u^{l-1}) = \int \mathcal{N}\left(u^l \,\middle|\, \mu_\phi^l(\epsilon^l, u^{l-1}), \mathrm{diag}(\sigma_\phi^l(\epsilon^l, u^{l-1}))^2\right) \mathcal{N}(\epsilon^l \,|\, 0, I) d\epsilon^l \tag{33}$$

Conventional variational inference for DGPs allows to integrate out the inducing values u analytically and obtain the marginal variational posteriors for $q_\phi(f^l \,|\, f^{l-1}, z^l, \theta)$. This is not possible in SSIVI-DPGs, as now we have to explicitly condition the variational model on the inducing values from the previous layer. Therefore we have to resort to plain MC estimation of the expected log-likelihood by sampling from the joint distribution $q_\phi(f, u, \epsilon \,|\, z, \theta, x)$.

$$\begin{aligned} &q_\phi(f_i, u, \epsilon \,|\, z, \theta, x) = \\ &\quad = q_\phi(f_i^1 \,|\, u^1, x_i, z, \theta) q_\phi(u^1 \,|\, \epsilon^1) q(\epsilon^1) \prod_{l=2}^{L} \Big[q_\phi(f_i^l \,|\, u^l, f_i^{l-1}, z, \theta) q_\phi(u^l \,|\, u^{l-1}, \epsilon^l) q(\epsilon^l) \Big] \end{aligned} \tag{34}$$

Now we modify lower bound (31) taking into account dependencies of inducing values between the layers and obtain the final objective for training SSIVI-DGPs:

$$\begin{aligned} \mathcal{L}^K(\phi, \theta) = &\underbrace{\sum_{i=1}^{N} \mathbb{E}_{q_\phi(f_i, u, \epsilon \,|\, z, \theta, x)} \log p_\theta(y_i \,|\, f_i^L)}_{\text{Expected log-likelihood}} - \\ &\underbrace{- \sum_{l=1}^{L} \mathbb{E}_{q_\phi(u^{l-1})} \mathbb{E}_{\epsilon_{0..1}^l} \mathbb{E}_{q_\phi(u^l \,|\, \epsilon_0^l, u^{l-1})} \log \frac{\frac{1}{K+1} \sum_{k=0}^{K} q_\phi(u^l \,|\, \epsilon_k^l, u^{l-1})}{p(u^l \,|\, z^l, \theta)}}_{\text{KL bound}} \to \max_{\phi, \theta} \end{aligned} \tag{35}$$