The Variational Gaussian Process

Dmitry Protasov

MIPT

May 29, 2024

Introduction

- Variational inference is a powerful tool for approximate posterior inference.
- Traditional methods like mean-field approximation, which assume **independent distributions for each latent variable**, have limitations in capturing dependencies among them.
- **The Variational Gaussian Process (VGP)** is introduced as a flexible model that adapts to complex posterior distributions.
- VGP achieves state-of-the-art results in unsupervised learning, particularly with deep latent Gaussian models and DRAW.

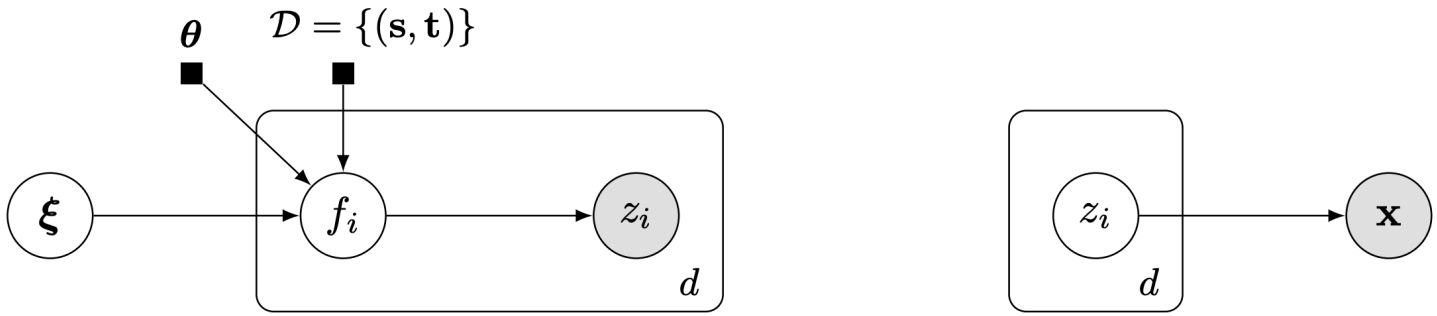
Gaussian processes

- GP regression estimates the function f given data pairs $\{(s_n, t_n)\}$.
- GP prior: $p(f) = \prod \mathrm{GP}(f_i; 0, K_{ss})$ where K_{ss} is the covariance function $k(s, s')$.
- ARD kernels: $k(s, s') = \sigma^2_{ARD} \exp\left(-\frac{1}{2} \sum \omega_j (s_j - s'_j)^2\right)$.
- Weights ω_j tune the importance of each dimension and can be driven to zero during inference, leading to automatic dimensionality reduction.
- Given data D, the conditional distribution of the GP interpolates between input-output pairs:

$$p(f|D) = \prod_{i=1}^{d} \mathrm{GP}\left(f_i; K_{\xi s} K_{ss}^{-1} t_i, K_{\xi\xi} - K_{\xi s} K_{ss}^{-1} K_{\xi s}^{\top}\right),$$

where $K_{\xi s}$ is the covariance function $k(\xi, s)$ for input ξ and all data inputs s_n.

Variational gaussian process



(a) VARIATIONAL MODEL **(b)** GENERATIVE MODEL

The VGP specifies the following generative process for posterior latent variables z

- Draw latent input $\xi \in \mathbb{R}^c$: $\xi \sim \mathcal{N}(0, I)$.
- Draw non-linear mapping $f : \mathbb{R}^c \to \mathbb{R}^d$ conditioned on $\mathcal{D}$: $f \sim \prod_{i=1}^{d} \mathcal{GP}(0, K_{\xi\xi}) \mid \mathcal{D}$.
- Draw approximate posterior samples $z \in \text{supp}(p)$: $\mathbf{z} = (z_1, \ldots, z_d) \sim \prod_{i=1}^{d} q(f_i(\xi))$.
- The VGP can capture complex dependencies and correlations between latent variables. VGP is parameterized by kernel hyperparameters θ and variational data

Universal approximation theorem

Theorem

Let $q(z;\theta,D)$ denote the variational Gaussian process. Consider a posterior distribution $p(z|x)$ with a finite number of latent variables and continuous quantile function (inverse CDF). There exists a sequence of parameters (θ_k, D_k) such that

$$\lim_{k\to\infty} \mathrm{KL}(q(z;\theta_k,D_k)\|p(z|x)) = 0.$$

- Theorem: any posterior distribution with strictly positive density can be represented by a VGP.
- This makes the VGP a flexible model for learning posterior distributions.

Black box inference

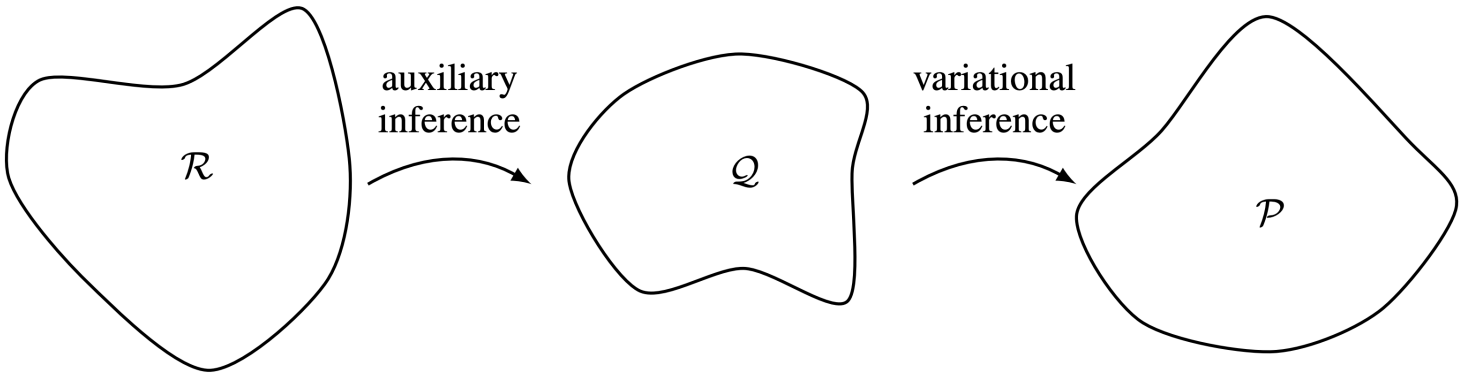


Figure: Sequence of domain mappings during inference, from variational latent variable space $\mathcal{R}$ to posterior latent variable space $\mathcal{Q}$ to data space $\mathcal{P}$

Unlike previous approaches, we rewrite this variational objective to connect to auto-encoders:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_{\text{VGP}}}[\log p(\mathbf{x} \mid \mathbf{z})] - \mathbb{E}_{q_{\text{VGP}}}[\text{KL}(q(\mathbf{z} \mid f(\boldsymbol{\xi})) \parallel p(\mathbf{z}))]$$

$$- \text{E}_{q_{\text{VGP}}}[\text{KL}(q(f \mid \boldsymbol{\xi}; \theta) \parallel r(f \mid \boldsymbol{\xi}, \mathbf{z}; \phi)) + \log q(\boldsymbol{\xi}) - \log r(\boldsymbol{\xi} \mid \mathbf{z})] \quad (1)$$

Black box inference

Algorithm 1 Black box inference with a variational Gaussian process

Input: Model $p(\mathbf{x}, \mathbf{z})$, Mean-field family $\prod_i q(z_i \mid f_i(\boldsymbol{\xi}))$.
Output: Variational and auxiliary parameters (θ, ϕ).
Initialize (θ, ϕ) randomly.
while not converged **do**
    Draw noise samples $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}), \epsilon \sim w$.
    Parameterize variational samples $\mathbf{z} = \mathbf{z}(\epsilon; f(\boldsymbol{\xi})), f(\boldsymbol{\xi}) = f(\boldsymbol{\xi}; \theta)$.
    Update (θ, ϕ) with stochastic gradients $\nabla_\theta \mathcal{L}, \nabla_\phi \mathcal{L}$.
end while

- Algorithm has $O(d + m^3 + LH^2)$ complexity, where d is the number of latent variables, m is the size of variational data, and L is the number of layers in neural networks.
- Unlike most GP literature, we require no low rank constraints, such as the use of inducing variables for scalable computation.

Related Work and Advantages of VGP

- VGP vs. Parametric Methods:
  - VGP uses a Bayesian nonparametric prior over all continuous mappings.
  - No need for costly Jacobian determinants ($O(d^3)$ complexity).
  - Fully Bayesian and flexible over the space of mappings.
- Efficiency and Flexibility:
  - VGP provides black box inference with lower variance gradients.
  - Applies location-scale transforms for reparameterization.
  - Efficient auxiliary inference for variational latent variables.
- Classic Techniques and Adaptability:
  - Builds on classic Bayesian inference techniques.
  - VGP adaptively learns transformations, avoiding discretization.

Experiments

- **Binarized MNIST**:
  - Deep latent Gaussian model (DLGM) and DRAW used for evaluation.
  - VGP achieves highest known results on log-likelihood using DRAW (-79.88).
  - VGP achieves highest known results among non-structure exploiting models using DLGM (-81.32).
- **Sketch Data Set**:
  - DRAW with VGP outperforms original DRAW on Sketch dataset.
  - Higher visual fidelity and better lower bound on log-likelihood.

Experiments

Model	$-\log p(\mathbf{x})$	$\leq$
DLGM + VAE [1]		86.76
DLGM + HVI (8 leapfrog steps) [2]	85.51	88.30
DLGM + NF ($k = 80$) [3]		85.10
EoNADE-5 2hl (128 orderings) [4]	84.68	
DBN 2hl [5]	84.55	
DARN 1hl [6]	84.13	
Convolutional VAE + HVI [2]	81.94	83.49
DLGM 2hl + IWAE ($k = 50$) [1]		82.90
DRAW [7]		80.97
DLGM 1hl + VGP		84.79
DLGM 2hl + VGP		81.32
DRAW + VGP		**79.88**

Table 1: Negative predictive log-likelihood for binarized MNIST. Previous best results are [1] (Burda et al., 2016), [2] (Salimans et al., 2015), [3] (Rezende & Mohamed, 2015), [4] (Raiko et al., 2014), [5] (Murray & Salakhutdinov, 2009), [6] (Gregor et al., 2014), [7] (Gregor et al., 2015).

Experiments

Figure 3: Generated images from DRAW with a VGP (top), and DRAW with the original variational auto-encoder (bottom). The VGP learns texture and sharpness, able to sketch more complex shapes.

Discussion

- The VGP is a powerful variational model for approximating complex posterior distributions.
- Future work includes exploring VGP in Monte Carlo methods and characterizing local optima in the optimization procedure.
- The VGP shows promise for efficient and flexible inference in a variety of generative models.

Literature

1. **Main article** The Variational Gaussian Process