Variational Gaussian Process

Dustin Tran[†‡], Rajesh Ranganath[*], David Blei[‡]

[†]Harvard University, [‡]Columbia University, [*]Princeton University

Summary

- Deep generative models provide complex representations of data.
- Variational inference methods require a rich family of approximating distributions.
- We develop a powerful variational model—the variational Gaussian process (VGP).
- We prove a universal approximation theorem: the VGP can capture any continuous posterior distribution.
- We derive an efficient black box algorithm.

Variational Models

- We want to compute the posterior $p(\mathbf{z}\mid\mathbf{x})$, for latent variables $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_d)$ and data $\mathbf{x}$.
- Variational inference seeks to minimize $\mathrm{KL}(q(\mathbf{z};\boldsymbol{\lambda})\,\|\,p(\mathbf{z}\mid\mathbf{x}))$ for a family $q(\mathbf{z};\boldsymbol{\lambda})$.
- Equivalent to maximizing evidence lower bound (ELBO)

$$\mathcal{L}=\mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}[\log p(\mathbf{x}\mid\mathbf{z})]-\mathrm{KL}(q(\mathbf{z};\boldsymbol{\lambda})\|p(\mathbf{z})).$$

- Interpret the family as a *variational model* for posterior latent variables $\mathbf{z}$ [3].
- *Hierarchical variational models*: prior $q(\boldsymbol{\lambda};\boldsymbol{\theta})$ and likelihood $\prod_i q(\mathbf{z}_i\mid\boldsymbol{\lambda}_i)$,

$$q(\mathbf{z};\boldsymbol{\theta})=\int\left[\prod_i q(\mathbf{z}_i\mid\boldsymbol{\lambda}_i)\right]q(\boldsymbol{\lambda};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\lambda}.$$

- Their expressiveness is determined by the complexity of the prior $q(\boldsymbol{\lambda})$.

Gaussian Processes

- Consider a data set of m source-target pairs $\mathcal{D}=\{(\mathbf{s}_n,\mathbf{t}_n)\}_{n=1}^m$, with source $\mathbf{s}_n\in\mathbb{R}^c$ paired with a target $\mathbf{t}_n\in\mathbb{R}^d$.
- We aim to learn a function over all source-target pairs, $\mathbf{t}_n=f(\mathbf{s}_n)$, where $f:\mathbb{R}^c\to\mathbb{R}^d$ is unknown.
- Let $f=(f_1,\ldots,f_d)$, where each $f_i:\mathbb{R}^c\to\mathbb{R}$. Gaussian process (GP) regression estimates f by placing a prior

$$p(f)=\prod_{i=1}^d\mathcal{GP}(f_i;\mathbf{0},\mathbf{K}_{ss}),$$

- Given data $\mathcal{D}$, the conditional distribution of the GP is

$$p(f\mid\mathcal{D})=\prod_{i=1}^d\mathcal{GP}(f_i;\mathbf{K}_{\xi s}\mathbf{K}_{ss}^{-1}\mathbf{t}_i,\mathbf{K}_{\xi\xi}-\mathbf{K}_{\xi s}\mathbf{K}_{ss}^{-1}\mathbf{K}_{\xi s}^\top).$$

Variational Gaussian Process

Let $\mathcal{D}=\{(\mathbf{s}_n,\mathbf{t}_n)\}_{n=1}^m$ be *variational data*, comprising input-outputs which are parameters. Let $\boldsymbol{\theta}$ be kernel parameters.

The variational Gaussian process (VGP) specifies the following generative process:

- Draw latent input $\xi\in\mathbb{R}^c$: $\xi\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.
- Draw non-linear mapping $f:\mathbb{R}^c\to\mathbb{R}^d$ conditioned on $\mathcal{D}$: $f\sim\prod_{i=1}^d\mathcal{GP}(\mathbf{0},\mathbf{K}_{\xi\xi})\,|\,\mathcal{D}$.
- Draw approximate posterior samples $\mathbf{z}\in\mathrm{supp}(p)$: $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_d)\sim\prod_{i=1}^d q(f_i(\xi))$.

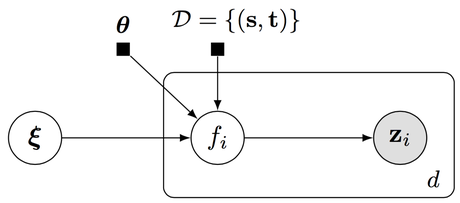

Figure 1: Variational Gaussian process (VGP).

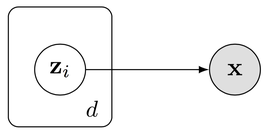

Figure 2: Class of generative models the VGP can learn.

The density of the VGP is

$$q_{\mathrm{VGP}}(\mathbf{z};\boldsymbol{\theta},\mathcal{D})=\iint\left[\prod_{i=1}^d q(\mathbf{z}_i\,|\,f_i(\xi))\right]\left[\prod_{i=1}^d\mathcal{GP}(f_i;\mathbf{0},\mathbf{K}_{\xi\xi})\,|\,\mathcal{D}\right]\mathcal{N}(\xi;\mathbf{0},\mathbf{I})\,\mathrm{d}f\,\mathrm{d}\xi,$$

- The VGP forms an *infinite* ensemble of mean-field distributions, with "weights" specified by a Bayesian nonparametric prior on mean-field parameters.
- The variational data anchors the random non-linear mappings at certain input-ouput pairs.
- **Special cases of the VGP.** The discrete mixture of mean-field distributions [2] is a form of VGP without a kernel; factor analysis in the variational space is a form of VGP without variational data.

Universal Approximation Theorem

Theorem. *For any posterior distribution $p(\mathbf{z}\mid\mathbf{x})$ with a finite number of latent variables and continuous inverse CDF, there exist a set of parameters $(\boldsymbol{\theta},\mathcal{D})$ such that*

$$\mathrm{KL}(q(\mathbf{z};\boldsymbol{\theta},\mathcal{D})\,\|\,p(\mathbf{z}\mid\mathbf{x}))=0.$$

Any posterior distribution with strictly positive density can be represented by a VGP.

Variational Lower Bound

The ELBO is analytically intractable due to $\log q_{\mathrm{VGP}}(\mathbf{z})$. We present a new variational lower bound:

$$\widetilde{\mathcal{L}}=\mathbb{E}_{q_{\mathrm{VGP}}}[\log p(\mathbf{x}\mid\mathbf{z})]-\mathbb{E}_{q_{\mathrm{VGP}}}\left[\mathrm{KL}\left(q(\mathbf{z}\,|\,f(\xi))\middle\|p(\mathbf{z})\right)+\mathrm{KL}\left(q(\xi,f)\middle\|r(\xi,f\,|\,\mathbf{z})\right)\right].$$

Auto-encoder interpretation: maximize the expected negative reconstruction error, regularized by expected divergences. It is a nested VAE bound.

Auto-Encoding Variational Models

We specify inference networks to parameterize both the variational and auxiliary models:

$$\mathbf{x}_n\mapsto q(\mathbf{z}_n\,|\,\mathbf{x}_n;\boldsymbol{\theta}_n),\qquad\mathbf{x}_n,\mathbf{z}_n\mapsto r(\xi_n,f_n\,|\,\mathbf{x}_n,\mathbf{z}_n;\boldsymbol{\phi}_n),$$

where q has local variational parameters given by the variational data $\mathcal{D}_n$, and r is specified as a fully factorized Gaussian.

Black Box Stochastic Optimization

We aim to maximize $\widetilde{\mathcal{L}}$ over variational parameters $\boldsymbol{\theta}$ and auxiliary parameters $\boldsymbol{\phi}$.

Analytic KL terms.

- $\mathrm{KL}\left(q(\xi,f)\middle\|r(\xi,f\,|\,\mathbf{z})\right)$: This is analytic as we've specified both distributions to be jointly Gaussian.
- $\mathrm{KL}\left(q(\mathbf{z}\,|\,f(\xi))\middle\|p(\mathbf{z})\right)$: This is standard in VAEs—it is analytic for deep generative models such as the deep latent Gaussian model [4] and DRAW [1].

Reparameterization.

- Latent inputs $\xi\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.
- Apply location-scale transform $\mathbf{f}(\xi;\boldsymbol{\theta})=\mathbf{L}\xi+\mathbf{K}_{\xi s}\mathbf{K}_{ss}^{-1}\mathbf{t}_i$. This acts as an evaluation $f(\xi)$ with random f from the GP.
- Suppose the mean-field $q(\mathbf{z}\,|\,f(\xi))$ is also reparameterizable: let $\boldsymbol{\epsilon}\sim w$ such that $\mathbf{z}(\boldsymbol{\epsilon};\mathbf{f})\sim q(\mathbf{z}\,|\,f(\xi))$.

The reparameterized variational lower bound is

$$\widetilde{\mathcal{L}}(\boldsymbol{\theta},\boldsymbol{\phi})=\mathbb{E}_{\mathcal{N}(\xi)}\left[\mathbb{E}_{w(\epsilon)}\left[\log p(\mathbf{x}\mid\mathbf{z}(\boldsymbol{\epsilon};\mathbf{f}(\xi;\boldsymbol{\theta})))\right]\right]-\mathbb{E}_{\mathcal{N}(\xi)}\left[\mathbb{E}_{w(\epsilon)}\left[\mathrm{KL}(q(\mathbf{z}\,|\,\mathbf{f}(\xi;\boldsymbol{\theta}))\|p(\mathbf{z}))+\mathrm{KL}(q(\xi,f;\boldsymbol{\theta})\|r(\xi,f\,|\,\mathbf{z}(\boldsymbol{\epsilon};\mathbf{f}(\xi;\boldsymbol{\theta}));\boldsymbol{\phi}))\right]\right].$$

Gradients backpropagate inside the expectations. Stochastic gradients exhibit low variance due to analytic KL terms and reparameterization.

Experiments

We compare the VGP for inferring the deep latent Gaussian model [4] and DRAW [1]. We report predictive likelihood on held-out data.

Binarized MNIST

Model	$-\log p(\mathbf{x})$	$\leq$
DLGM + VAE		86.76
DLGM + HVI (8 leapfrog steps)	85.51	88.30
DLGM + NF ($k=80$)		85.10
EoNADE-5 2hl (128 orderings)	84.68	
DBN 2hl	84.55	
DARN 1hl	84.13	
Convolutional VAE + HVI	81.94	83.49
DLGM 2hl + IWAE ($k=50$)		82.90
DRAW		80.97
DLGM 1hl + VGP (**this paper**)		83.64
DLGM 2hl + VGP (**this paper**)		81.90
DRAW + VGP (**this paper**)		**80.11**

The VGP achieves the highest known results using DRAW, and the highest among non-structure exploiting models using the DLGM.

Sketch

Model	Epochs	$\leq-\log p(\mathbf{x})$
DRAW	100	526.8
	200	479.1
	300	464.5
DRAW + VGP	100	475.9
	200	430.0
	300	425.4

The VGP (top) learns texture and sharpness, able to sketch more complex shapes than the standard DRAW (bottom).

References

[1] Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. (2015). DRAW: A recurrent neural network for image generation. In *International Conference on Machine Learning*.

[2] Lawrence, N. (2000). *Variational Inference in Probabilistic Models*. PhD thesis.

[3] Ranganath, R., Tran, D., and Blei, D. M. (2015). Hierarchical variational models. *arXiv preprint arXiv:1511.02386*.

[4] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.