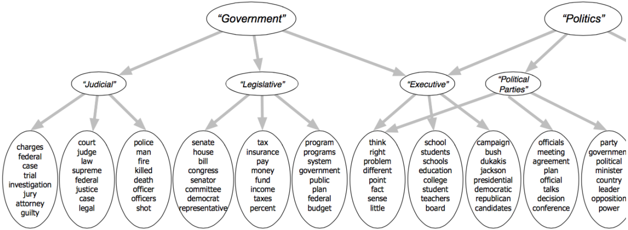
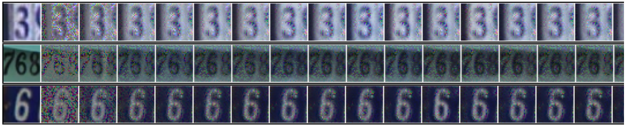
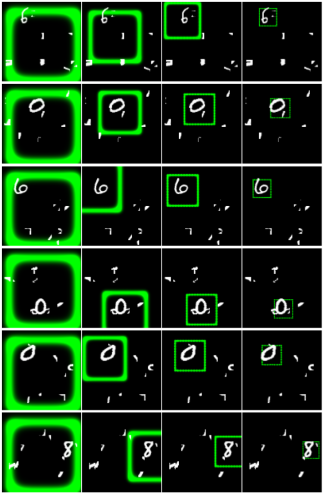
Variational models

Dustin Tran
Department of Computer Science
Columbia University

Joint work with:
Rajesh Ranganath, David M. Blei







[Rezende et al., 2014; Ranganath et al., 2015; Gregor et al., 2015]

- Deep generative models provide complex representations of data.
- Learning these representations are fundamentally tied to their inference method from data.
- With variational inference methods, the bottleneck is specifying a rich family of approximating distributions.

How we can build expressive variational families in a black box framework, which adapt to the model complexity at hand?

Background

Given:

- Data set $\mathbf{x}$.
- Joint probability model $p(\mathbf{x}, \mathbf{z})$, with latent variables $\mathbf{z}_1, \ldots, \mathbf{z}_d$.

Goal:

- Compute posterior $p(\mathbf{z} \mid \mathbf{x})$.

Variational inference:

- Posit a family of distributions $\{q(\mathbf{z}; \lambda) : \lambda \in \Lambda\}$.
- Minimize $\mathrm{KL}(q \parallel p)$, which is equivalent to maximizing the ELBO

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

- Commonly use a mean-field distribution $q(\mathbf{z}; \lambda) = \prod_{i=1}^{d} q(\mathbf{z}_i; \lambda_i)$.

Variational models

Interpret q as a *variational model* for posterior latent variables $\mathbf{z}$.

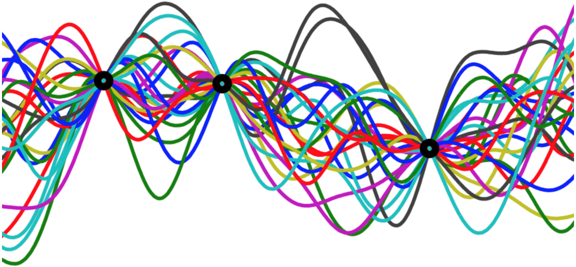
Hierarchical variational models: prior $q(\lambda; \theta)$, likelihood $\prod_i q(\mathbf{z}_i \mid \lambda_i)$.

$$q(\mathbf{z}; \theta) = \int \Big[\prod_i q(\mathbf{z}_i \mid \lambda_i) \Big] q(\lambda; \theta) \, \mathrm{d}\lambda$$

- Hierarchical variational models unify other expressive approximations (mixture, structured, MCMC, copula,...).
- Their expressiveness is determined by the complexity of the prior $q(\lambda)$.

[Ranganath et al., 2015]

Prior: Gaussian processes



Consider a data set of m source-target pairs $\mathcal{D} = \{(\mathbf{s}_n, \mathbf{t}_n)\}_{n=1}^{m}$, with $\mathbf{s}_n \in \mathbb{R}^c$ and $\mathbf{t}_n \in \mathbb{R}^d$.

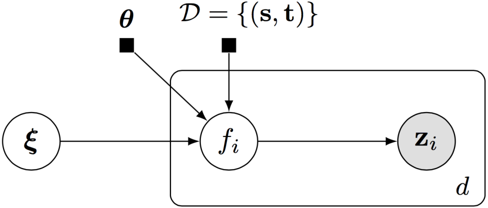
We aim to learn a function $f : \mathbb{R}^c \to \mathbb{R}^d$ over all source-target pairs,

$$\mathbf{t}_n = f(\mathbf{s}_n), \qquad p(f) = \prod_{i=1}^{d} \mathcal{GP}(f_i; \mathbf{0}, \mathbf{K}),$$

where each $f_i : \mathbb{R}^c \to \mathbb{R}$. Given data $\mathcal{D}$, the conditional $p(f \mid \mathcal{D})$ forms a distribution over mappings which interpolate between input-output pairs.

(fig. by Ryan Adams)

Variational Gaussian process



$\mathcal{D} = \{(\mathbf{s}_n, \mathbf{t}_n)\}_{n=1}^{m}$ is *variational data*, comprising input-output pairs.
θ are kernel hyperparameters.

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 $\mathcal{D}$:
  $f \sim \prod_{i=1}^{d} \mathcal{GP}(\mathbf{0}, \mathbf{K}) \,|\, \mathcal{D}$.
- Draw approximate posterior samples $\mathbf{z} \in \text{supp}(p)$:
  $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_d) \sim \prod_{i=1}^{d} q(\mathbf{z} \,|\, f_i(\xi))$.

Variational Gaussian process

The density of the VGP is

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

- The VGP forms an ensemble of mean-field distributions—“weights” (mixing measure) are specified by a Bayesian nonparametric prior.
- The variational data $\mathcal{D}$ anchors the random non-linear mappings at certain input-output pairs.

Variational Gaussian process

Universal Approximation 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* $(\theta, \mathcal{D})$ *such that*

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

The VGP's complexity grows efficiently and towards *any* distribution, adapting to the generative model's complexity at hand.

Black box inference

The ELBO is analytically intractable due to the density $q_{\mathrm{VGP}}(\mathbf{z})$.

We present a new variational lower bound:

$$\begin{aligned}\widetilde{\mathcal{L}} &= \mathbb{E}_{q_{\mathrm{VGP}}}[\log p(\mathbf{x} \mid \mathbf{z})] \\ &\quad - \mathbb{E}_{q_{\mathrm{VGP}}}\Big[\mathrm{KL}\Big(q(\mathbf{z} \mid f(\xi)) \Big\| p(\mathbf{z})\Big) + \mathrm{KL}\Big(q(\xi, f) \Big\| r(\xi, f \mid \mathbf{z})\Big)\Big],\end{aligned}$$

where r is an auxiliary distribution.

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

Black box inference

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

1. Inference networks. Specify inference networks to parameterize both the variational and auxiliary models:

$$\mathbf{x}_n \mapsto q(\mathbf{z}_n \mid \mathbf{x}_n; \mathcal{D}_n), \qquad \mathbf{x}_n, \mathbf{z}_n \mapsto r(\xi_n, f_n \mid \mathbf{x}_n, \mathbf{z}_n; \phi_n),$$

where r is specified as a fully factorized Gaussian with $\phi_n = (\mu_n, \sigma_n^2 \mathbf{I})$. This amortizes the cost of computation by making all parameters global.

Black box inference

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

2. Analytic KL terms.

- $\mathrm{KL}\left(q(\mathbf{z} \mid f(\xi)) \middle\| p(\mathbf{z}) \right)$: Standard in VAEs—it is analytic for deep generative models such as the DLGM and DRAW.
- $\mathrm{KL}\left(q(\xi, f) \middle\| r(\xi, f \mid \mathbf{z}) \right)$: Always analytic as we've specified both joint distributions to be Gaussian.

Black box inference

$$\widetilde{\mathcal{L}}(\theta, \phi) = \mathbb{E}_{q_{\text{VGP}}}[\log p(\mathbf{x} \mid \mathbf{z})] \\ - \mathbb{E}_{q_{\text{VGP}}}\Big[\mathrm{KL}\Big(q(\mathbf{z} \mid f(\xi)) \Big\| p(\mathbf{z})\Big) + \mathrm{KL}\Big(q(\xi, f; \theta) \Big\| r(\xi, f \mid \mathbf{z}; \phi)\Big)\Big]$$

3. Reparameterization.

- For $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, apply location-scale transform $\mathbf{f}(\xi; \theta)$. This implies $\mathbf{f}(\xi; \theta) = f(\xi)$ for $f \sim \prod_{i=1}^{d} \mathcal{GP}(\mathbf{0}, \mathbf{K}) \mid \mathcal{D}$.
- Suppose the mean-field is also reparameterizable: let $\epsilon \sim w$ such that $\mathbf{z}(\epsilon; \mathbf{f}) \sim q(\mathbf{z} \mid f(\xi))$.

Black box inference

The reparameterized variational lower bound is

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

Run stochastic optimization:

- Stochastic gradients exhibit low variance due to analytic KL terms and reparameterization.
- Complexity is linear in the number of latent variables, which is the same as a mean-field approximation!

Experiments: Binarized MNIST

Model	$-\log p(\mathbf{x})$	$\leq$
DLGM + VAE [Burda et al., 2015]		86.76
DLGM + HVI (8 leapfrog steps) [Salimans et al., 2015]	85.51	88.30
DLGM + NF ($k = 80$) [Rezende + Mohamed, 2015]		85.10
EoNADE-5 2hl (128 orderings) [Raiko et al., 2015]	84.68	
DBN 2hl [Murray + Salakhutdinov, 2009]	84.55	
DARN 1hl [Gregor et al., 2014]	84.13	
Convolutional VAE + HVI [Salimans et al., 2015]	81.94	83.49
DLGM 2hl + IWAE ($k = 50$) [Burda et al., 2015]		82.90
DRAW [Gregor et al. 2015]		80.97
DLGM 1hl + VGP		83.64
DLGM 2hl + VGP		81.90
DRAW + VGP		**80.11**

We also find richer latent representations than the VAE or IWAE.

Experiments: 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**

Data set of 20,000 human sketches equally distributed over 250 object categories.

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

Summary

- Introduced the framework of variational models.
- Developed the variational Gaussian process—proven to be a universal approximator.
- Derived scalable black box inference—three key ingredients with inference networks, analytic KL terms, and reparameterization.

Thanks!

Discussion

- Can we apply black box variational inference for more complicated posterior estimands of interest?
- Can we obtain theoretical guarantees similar to Monte Carlo methods?
- Can we extend these results to learning Bayesian nonparametric models (without truncation)?
- How do we solve the automation challenges when integrating the VGP into ADVI?