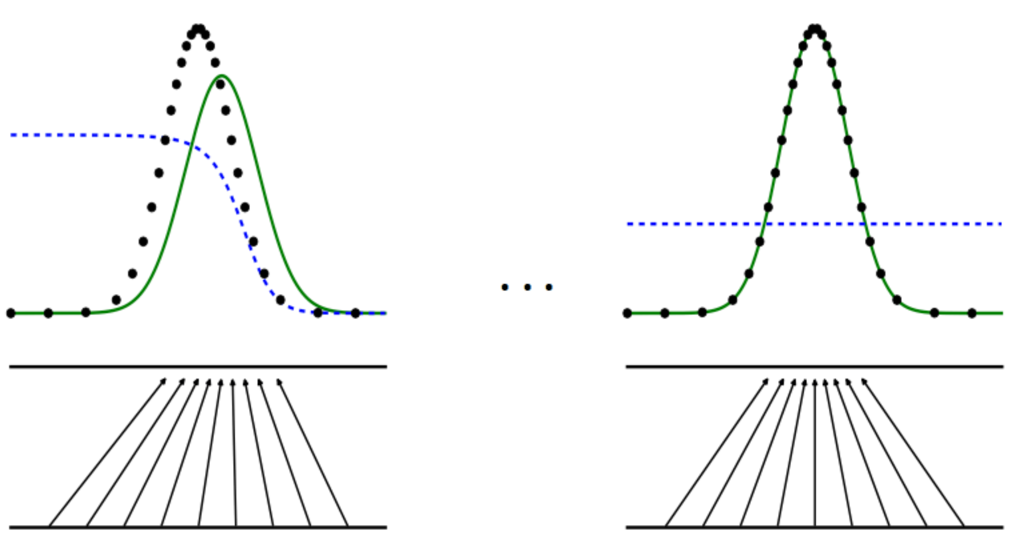
Implicit Posterior Variational Inference for Deep Gaussian Processes (IPVI DGP)



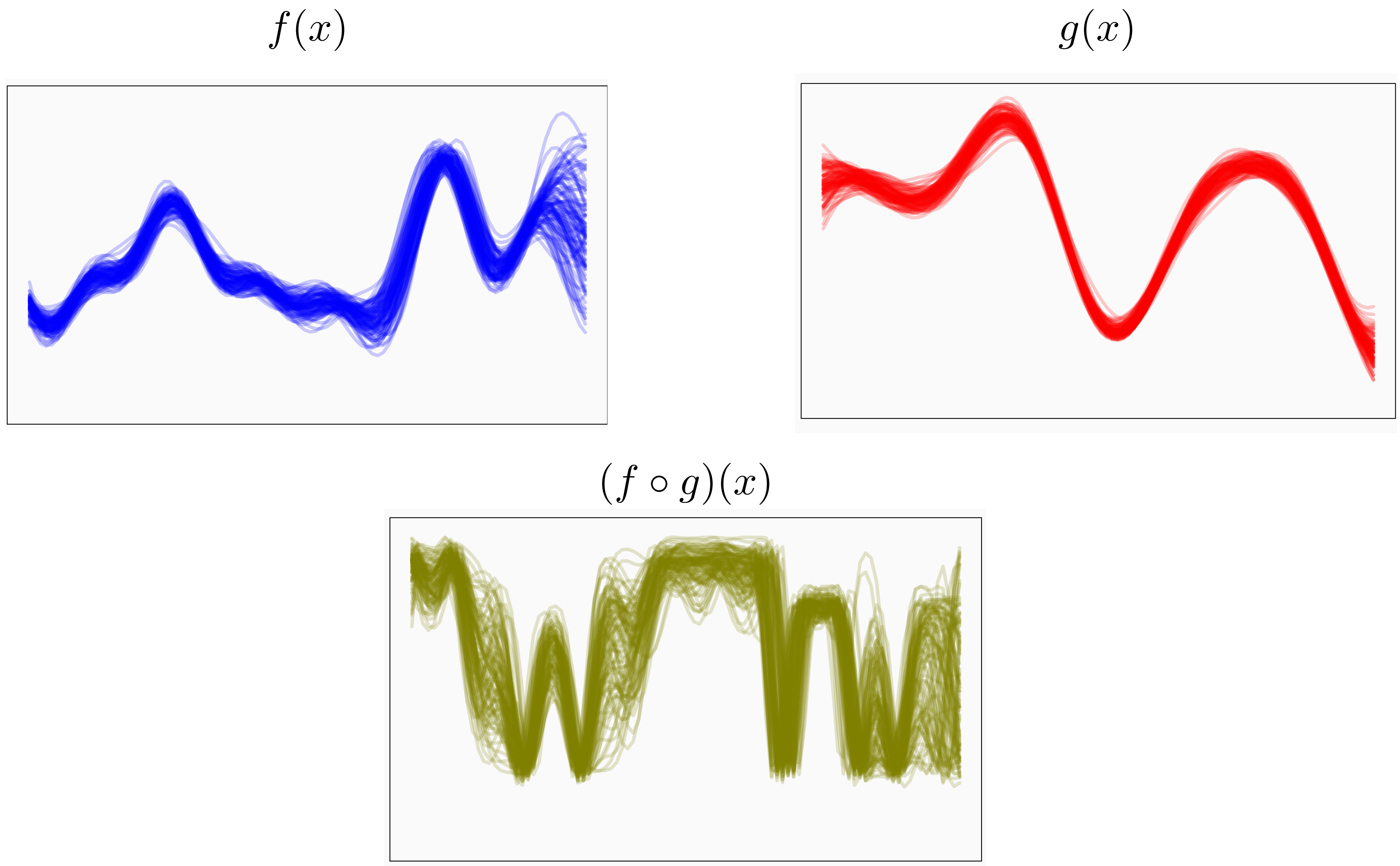
Haibin Yu*, Yizhou Chen*
Zhongxiang Dai
Bryan Kian Hsiang Low and Patrick Jaillet
Department of Computer Science
National University of Singapore
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

* indicates equal contribution

Gaussian Processes (GP) vs. Deep Gaussian Processes (DGP)

- A GP is fully specified by its kernel function
  - **RBF:** universal approximator
  - Matern
  - Brownian
  - Linear
  - Polynomial
  - ……

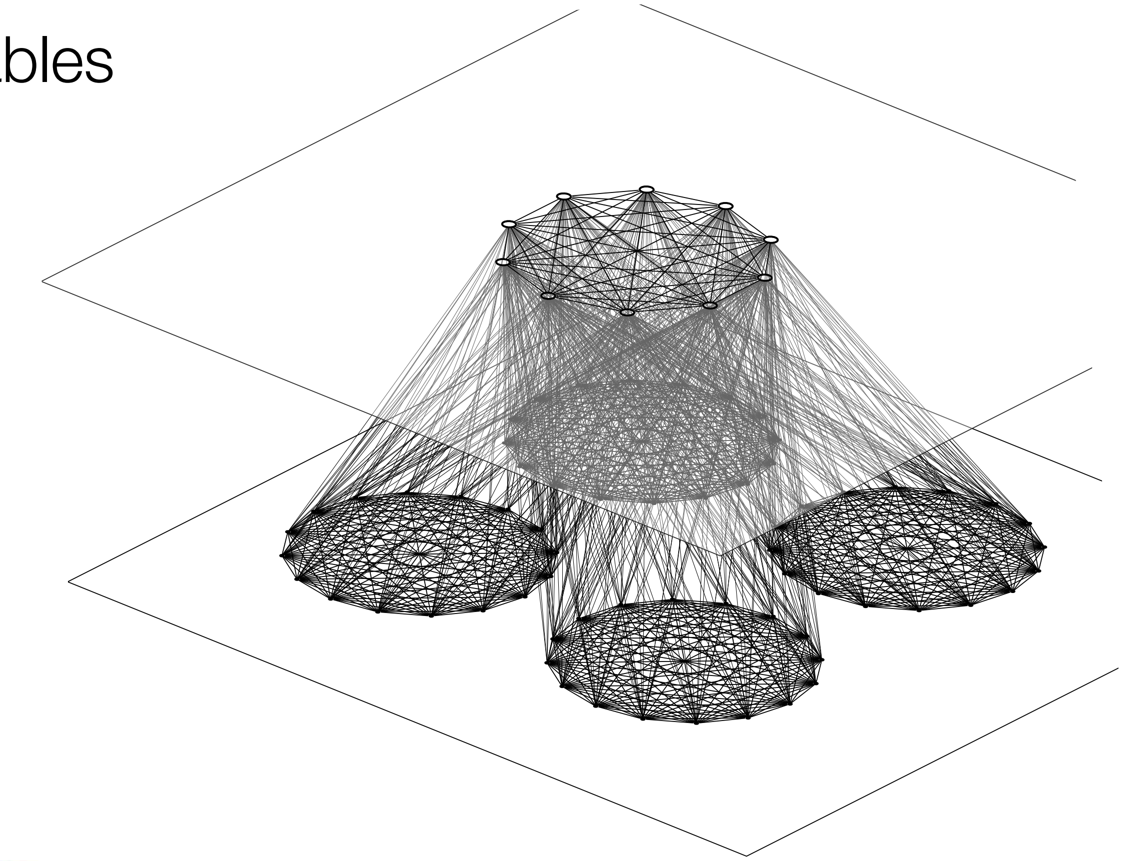
Gaussian Processes (GP) vs. Deep Gaussian Processes (DGP)

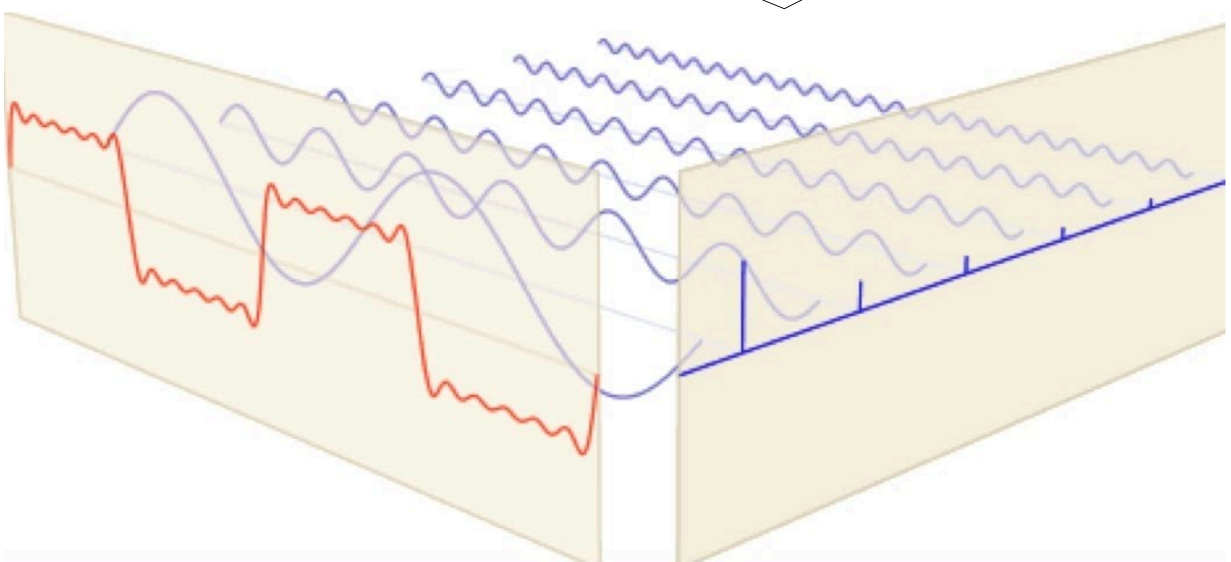


- Composition of GPs significantly boosts the expressive power

Existing DGP models

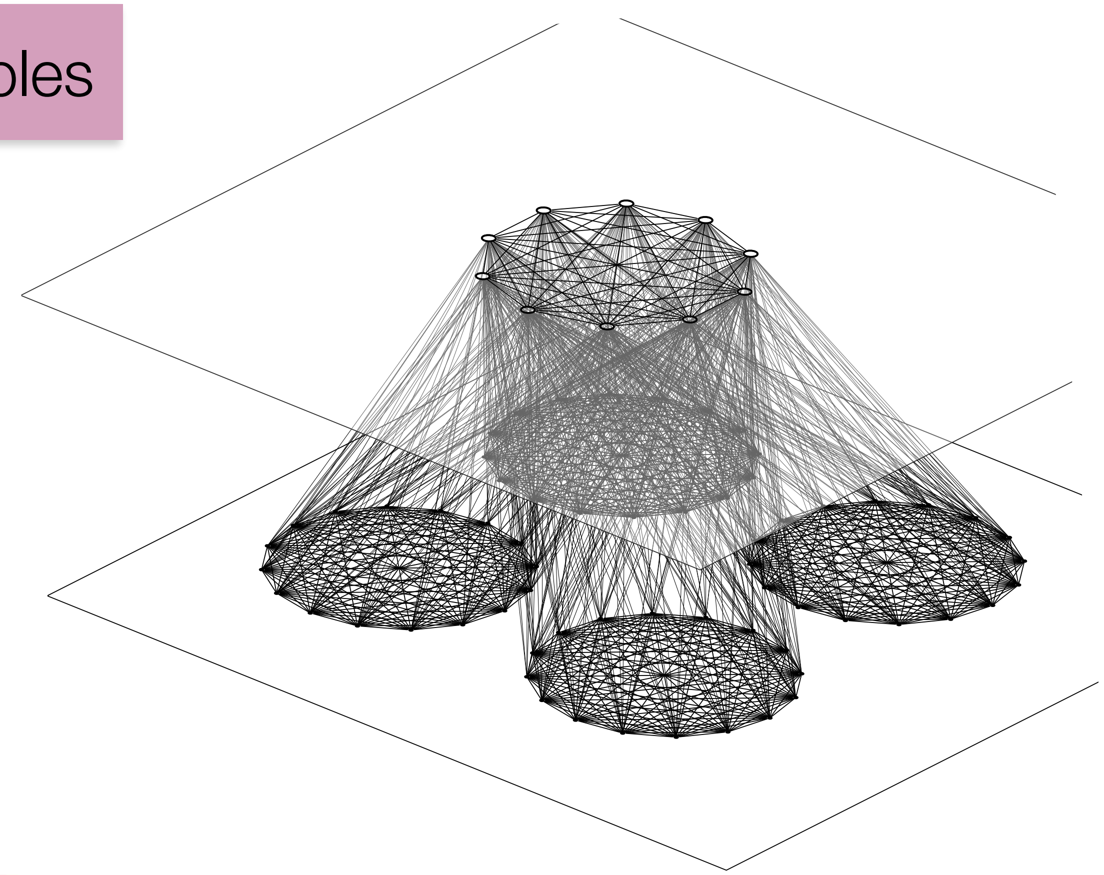
- Approximation methods based on inducing variables
    - Variational Inference
        - Damianou and Lawrence, AISTATS, 2013
        - Hensman and Lawrence, arXiv, 2014
        - Salimbeni and Deisenroth, NeurIPS, 2017
    - Expectation Propagation
        - Bui, ICML, 2016
    - MCMC
        - Havasi et al, NeurIPS 2018
- Random feature approximation methods
    - Cutajar et al, ICML 2017

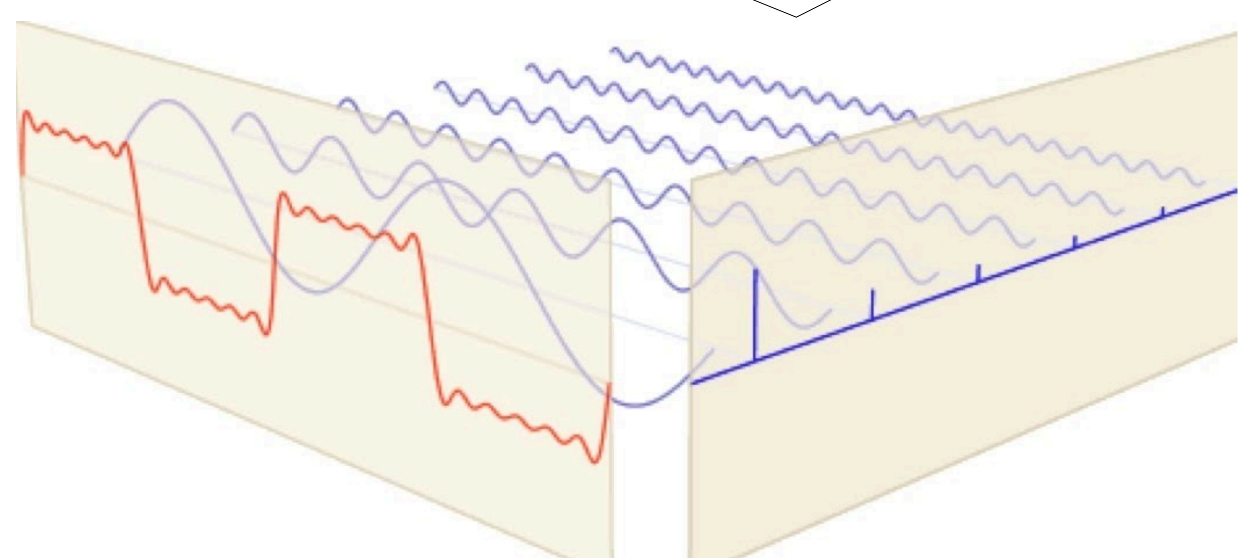




Existing DGP models

- Approximation methods based on inducing variables
  - Variational Inference
    - Damianou and Lawrence, AISTATS, 2013
    - Hensman and Lawrence, arXiv, 2014
    - Salimbeni and Deisenroth, NeurIPS, 2017
  - Expectation Propagation
    - Bui, ICML, 2016
  - MCMC
    - Havasi et al, NeurIPS 2018
- Random feature approximation methods
  - Cutajar et al, ICML 2017

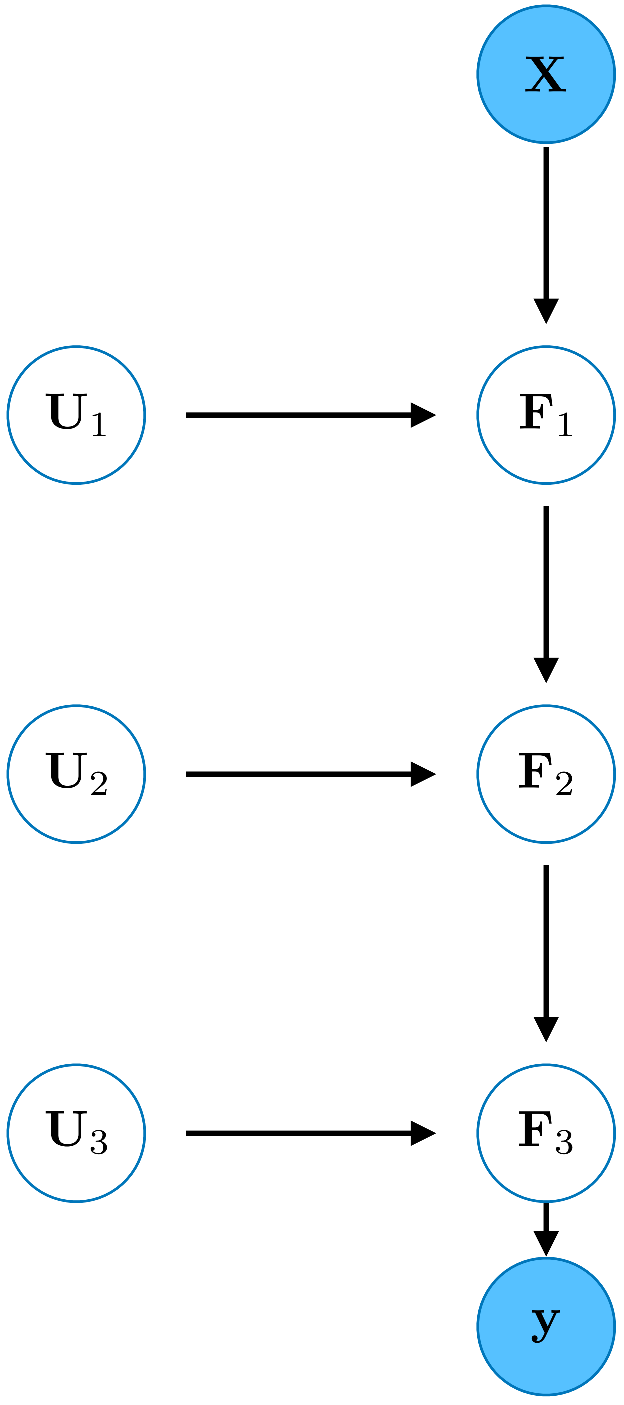




Deep Gaussian Processes (DGP)

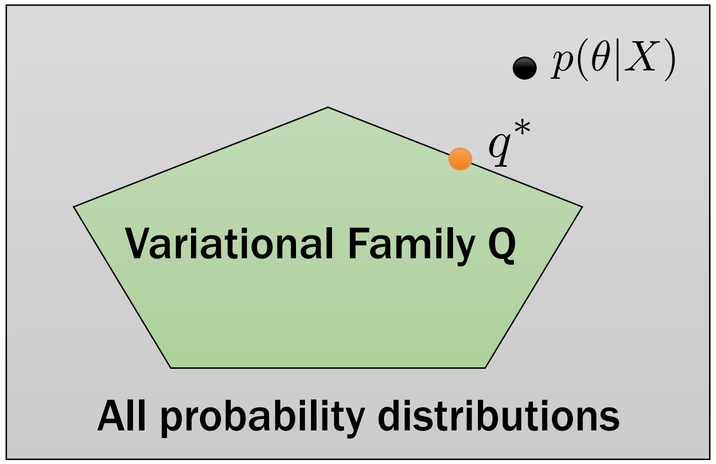
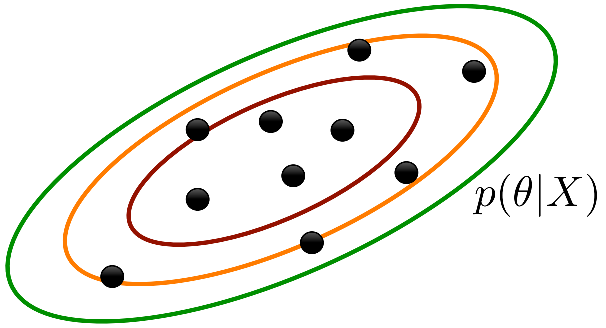
- Input $\mathbf{X}$
- Output $\mathbf{y}$
- Inducing variables $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_1, \ldots, \mathbf{U}_L\}$

- **Posterior $p(\boldsymbol{\mathcal{U}}|\mathbf{y})$ is intractable!**



DGP Inference

- Exact inference is intractable in DGP

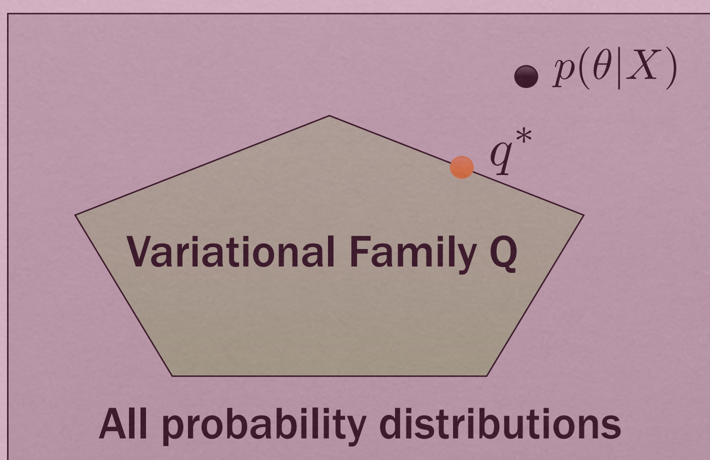
Variational Inference	Sampling
$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) \| p(\theta \| X)]$	$\mathbb{E}_{p(\theta\|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta\|X)$
	

DGP Inference: Variational Inference

- Exact inference is intractable in DGP

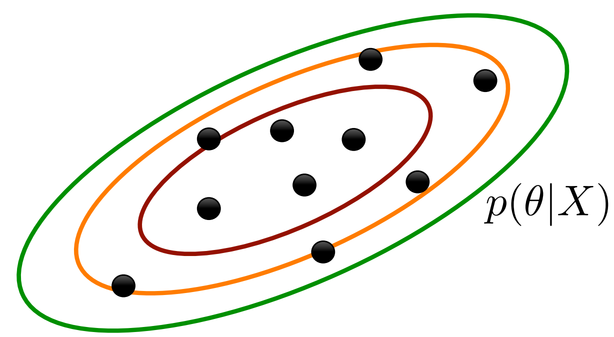
Variational Inference

$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) || p(\theta|X)]$$



Sampling

$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$



Variational Inference

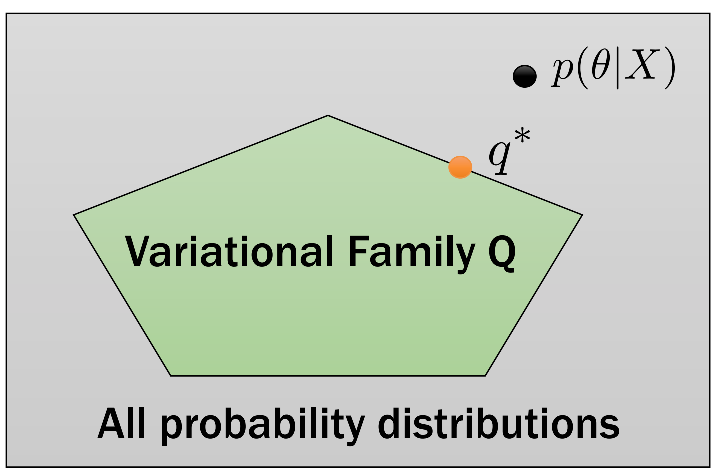
Gaussian approximation

Mean field approximation

DGP Inference: Variational Inference

Variational Inference

$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) || p(\theta|X)]$$

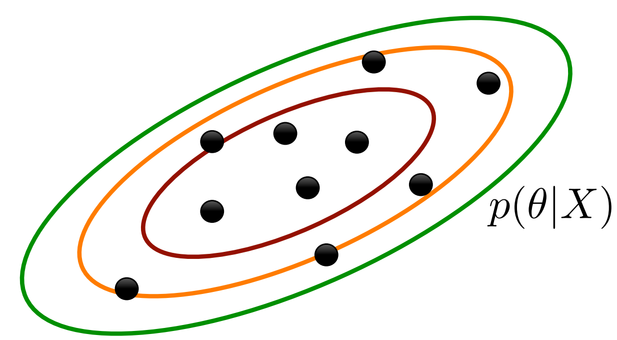


efficient

biased

Sampling

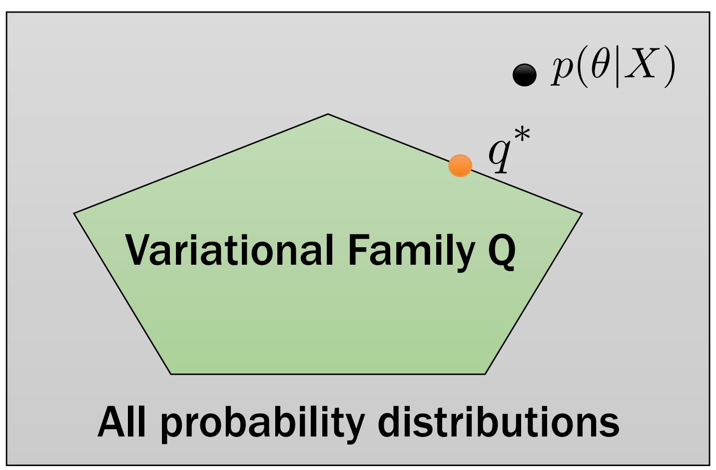
$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$



DGP Inference: Sampling

Variational Inference

$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) || p(\theta|X)]$$

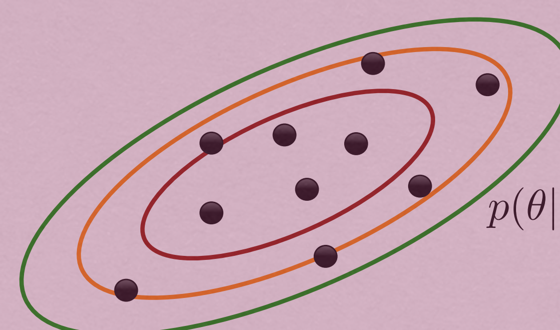


efficient

biased

Sampling

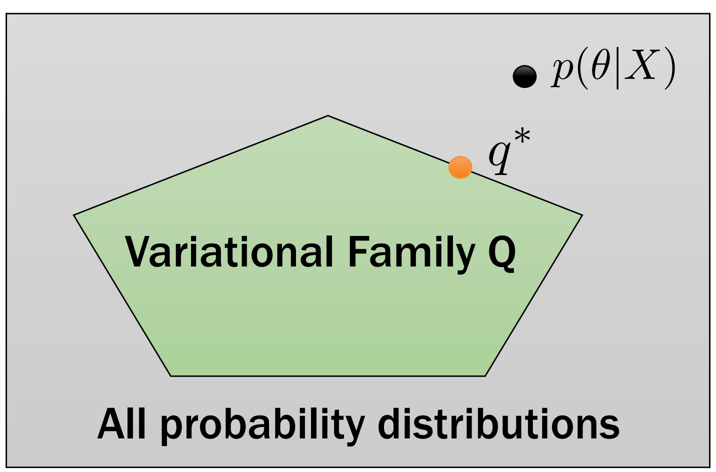
$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$



DGP Inference: Sampling

Variational Inference

$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) || p(\theta|X)]$$

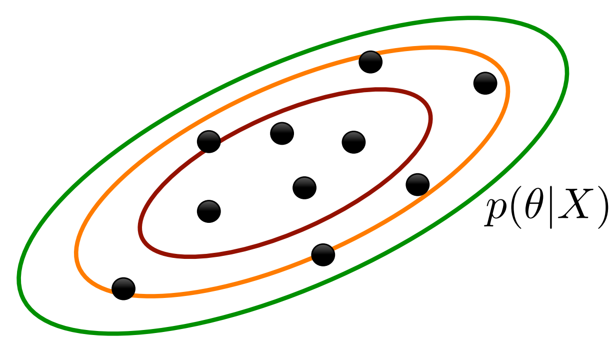


efficient

biased

Sampling

$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$



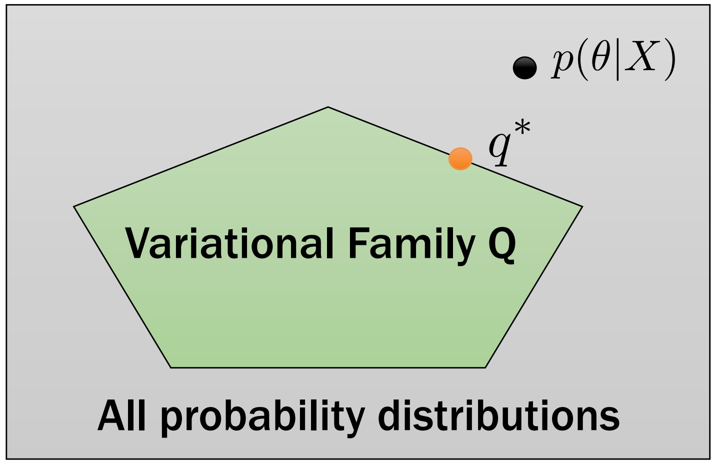
ideally unbiased

not efficient

DGP: Variational Inference vs. Sampling

Variational Inference

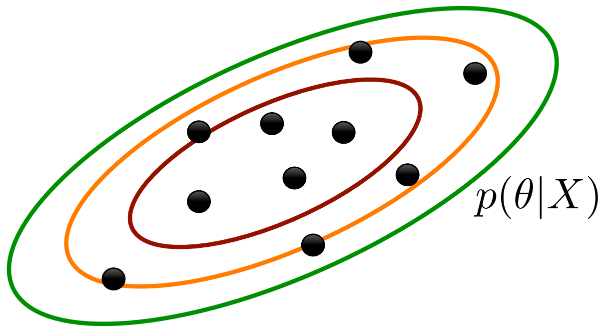
$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta) || p(\theta|X)]$$



efficiency

Sampling

$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$

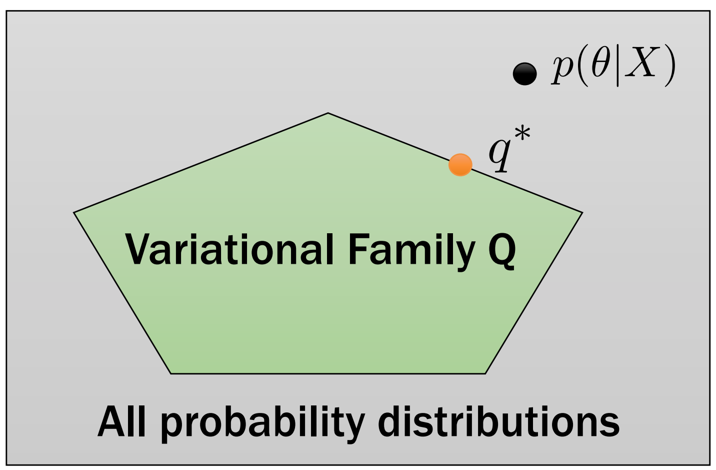


ideally unbiased

DGP: Variational Inference vs. Sampling

Variational Inference

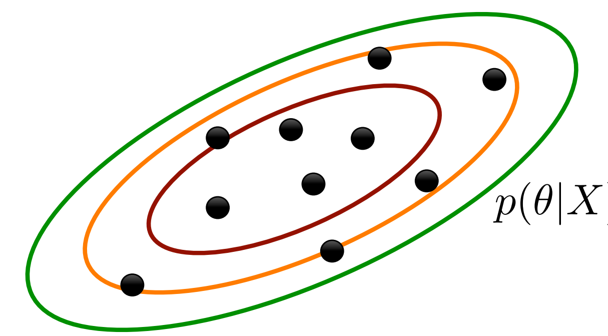
$$q^* = \min_{q \in Q} \mathrm{KL}[q(\theta)||p(\theta|X)]$$



efficiency

Sampling

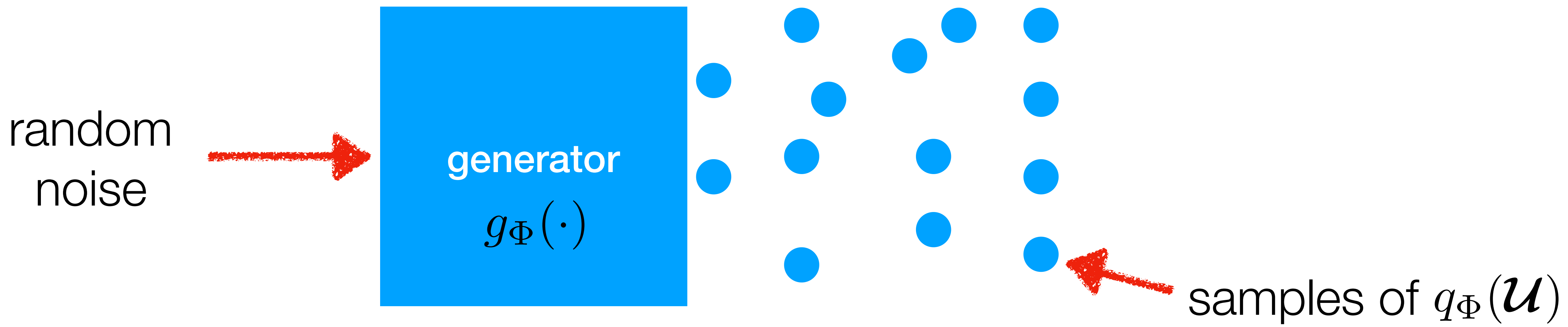
$$\mathbb{E}_{p(\theta|X)}[f(\theta)] \approx \frac{1}{T}\sum_{t=1}^{T} f(\theta_t) \quad : \quad \theta_t \sim p(\theta|X)$$



ideally unbiased

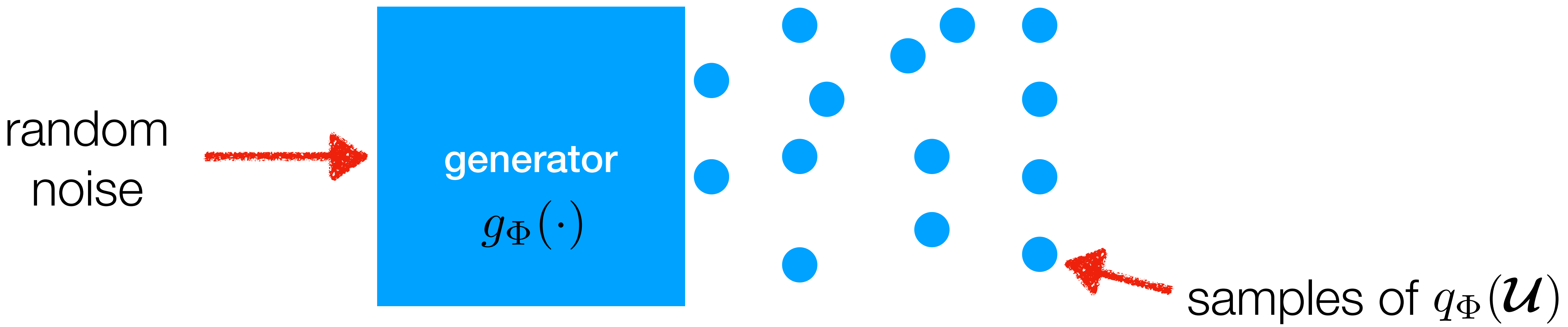
unbiased posterior & efficiency

Implicit Posterior Variational Inference



$$\mathrm{ELBO} = \mathbb{E}_{q(\mathbf{F}_L)}[\log p(\mathbf{y}|\mathbf{F}_L)] - \mathrm{KL}\left[q_\Phi(\mathcal{U})||p(\mathcal{U})\right]$$

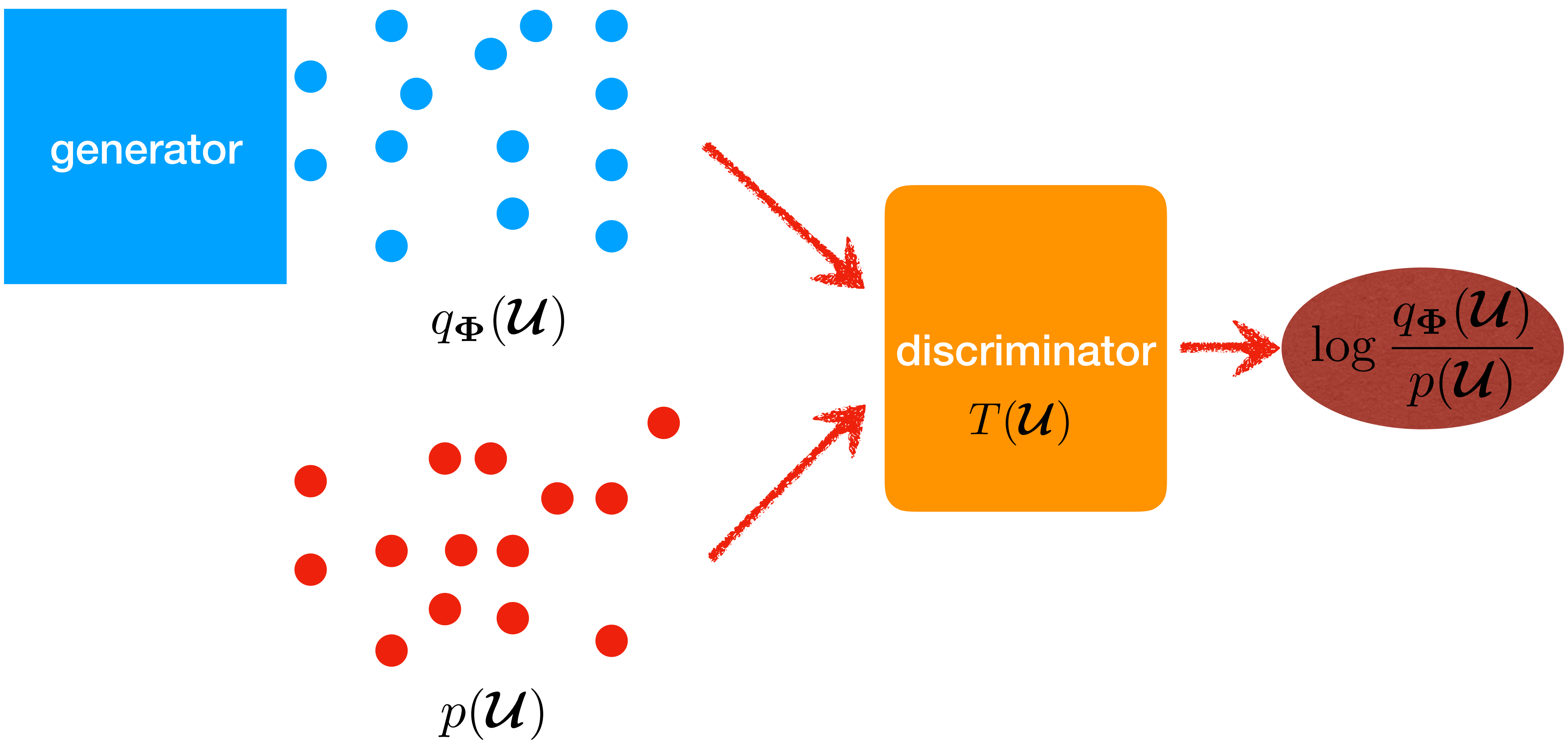
Implicit Posterior Variational Inference



$$\text{ELBO} = \mathbb{E}_{q(\mathbf{F}_L)}[\log p(\mathbf{y}|\mathbf{F}_L)] - \text{KL}\left[q_\Phi(\mathcal{U})||p(\mathcal{U})\right]$$

$$\text{KL}[q_\Phi(\mathcal{U})||p(\mathcal{U})] = \mathbb{E}_{q_\Phi(\mathcal{U})}\left[\log \frac{q_\Phi(\mathcal{U})}{p(\mathcal{U})}\right]$$

Implicit Posterior Variational Inference



Proposition 1. The optimal discriminator exactly recovers the log-density ratio

Implicit Posterior Variational Inference

- Two-player game

$$\textbf{Player [1]:} \quad \max_{\{\Psi\}} \mathbb{E}_{p(\mathcal{U})}[\log(1-\sigma(T_\Psi(\mathcal{U}))] + \mathbb{E}_{q_\Phi(\mathcal{U})}[\log \sigma(T_\Psi(\mathcal{U}))],$$

discriminator

$$\textbf{Player [2]:} \quad \max_{\{\theta,\Phi\}} \mathbb{E}_{q_\Phi(\mathcal{U})}[\mathcal{L}(\theta, \mathbf{X}, \mathbf{y}, \mathcal{U}) - T_\Psi(\mathcal{U})]$$

generator & DGP hyperparameters

- Best-response dynamics (BRD) to search for a Nash equilibrium

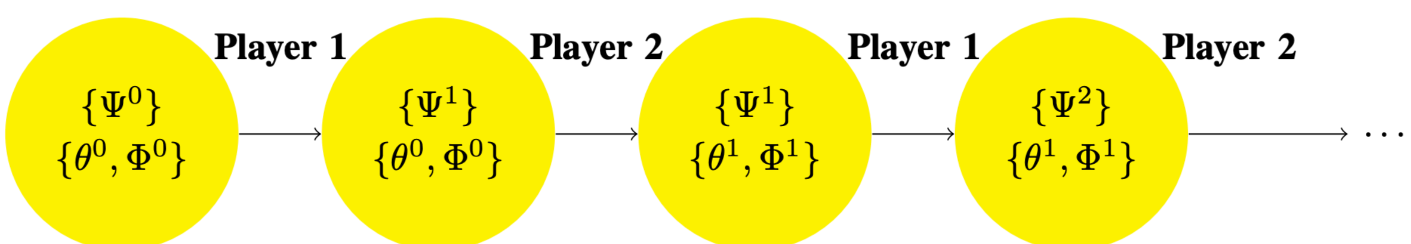
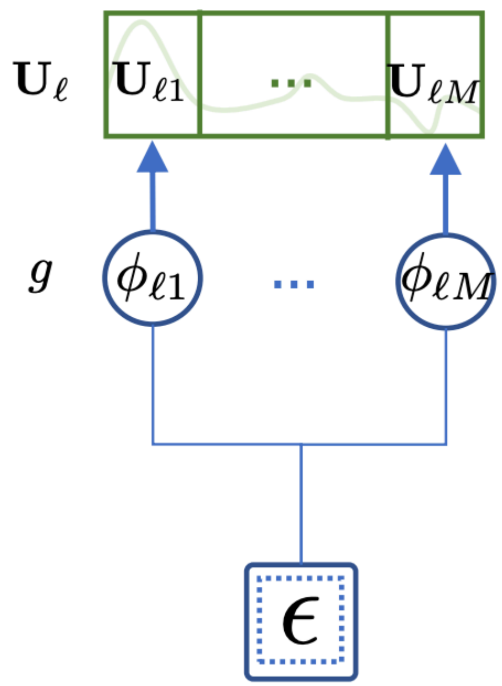


Figure 1: *Best-response dynamics* (BRD) algorithm

Proposition 2. Nash equilibrium recovers the true posterior $p(\mathcal{U}|\mathbf{y})$

Architecture of the generator and discriminator

- Naive design for layer ℓ



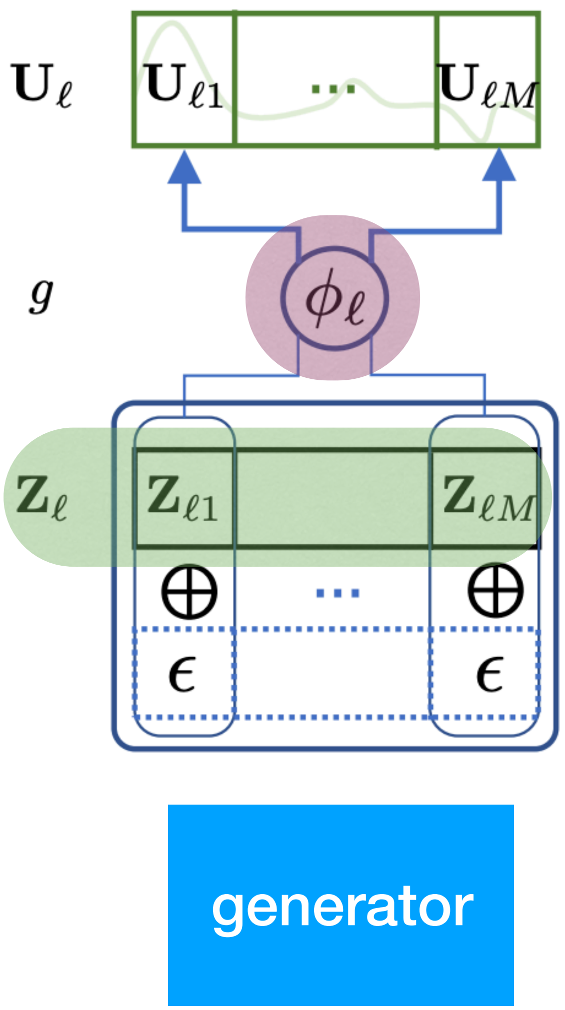
(a)

generator (naive)

- Fail to adequately capture the dependency of the inducing output variables $\mathcal{U} = \{\mathbf{U}_1, \ldots, \mathbf{U}_L\}$ on the corresponding inducing inputs $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_L\}$
- Relatively large number of parameters, resulting in overfitting, optimization difficulty, etc.

Architecture of Generator and Discriminator for DGP

- Our parameter-tying design for layer ℓ



- Concatenates the inducing inputs $\mathbf{Z}_\ell$
- Posterior samples are generated based on single shared parameter setting ϕ_ℓ

NUS National University of Singapore

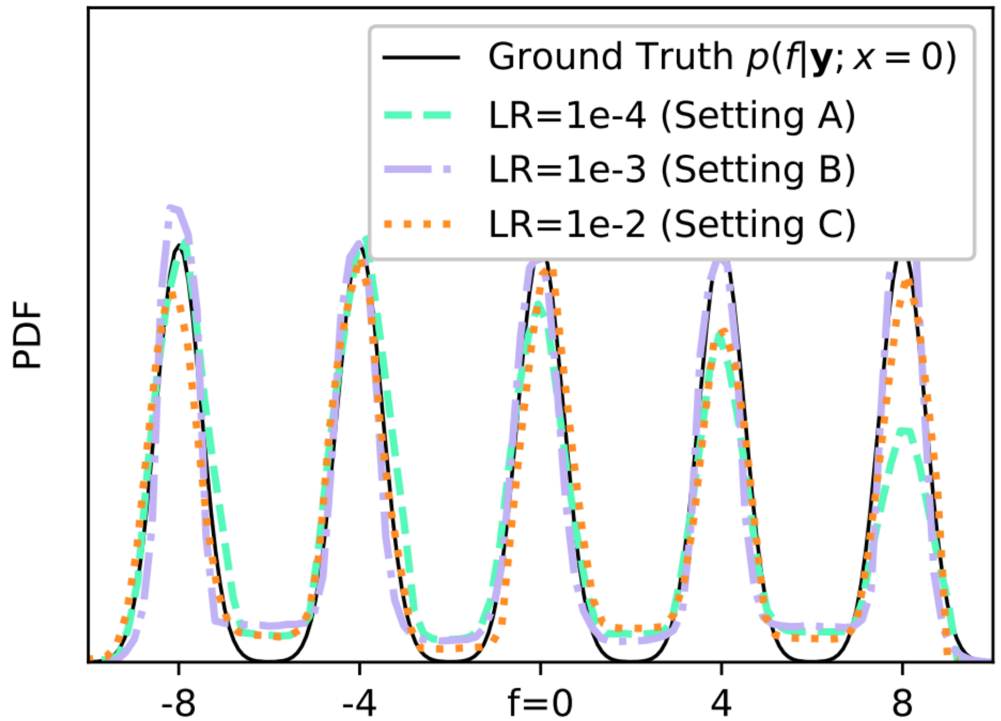
Experimental Results

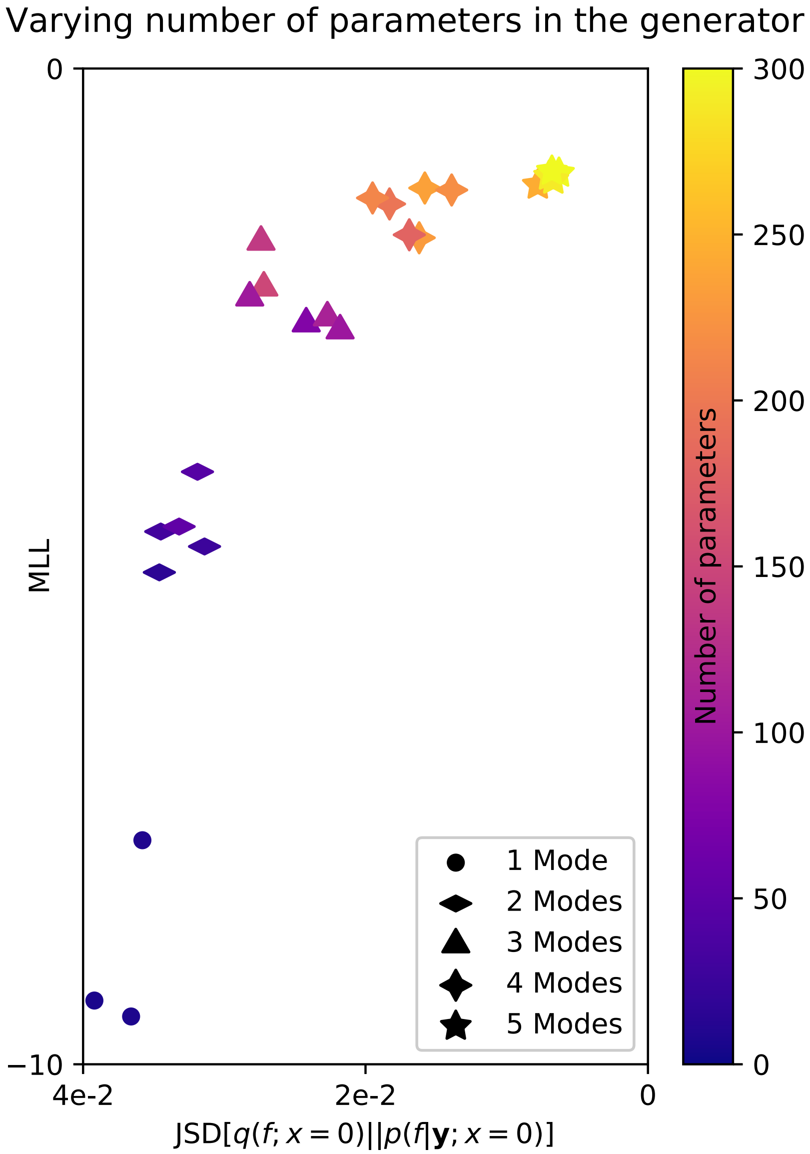
- Metric for evaluation
  - MLL (mean log likelihood)

- Algorithms for comparison
  - **DSVI DGP**: Doubly stochastic variational inference DGP [Salimbeni and Deisenroth, 2017]
  - **SGHMC DGP**: Stochastic gradient Hamilton Monte Carlo DGP [Havasi et al, 2018]

Experimental Results

- Synthetic Experiment: Learning a Multi-Modal Posterior Belief

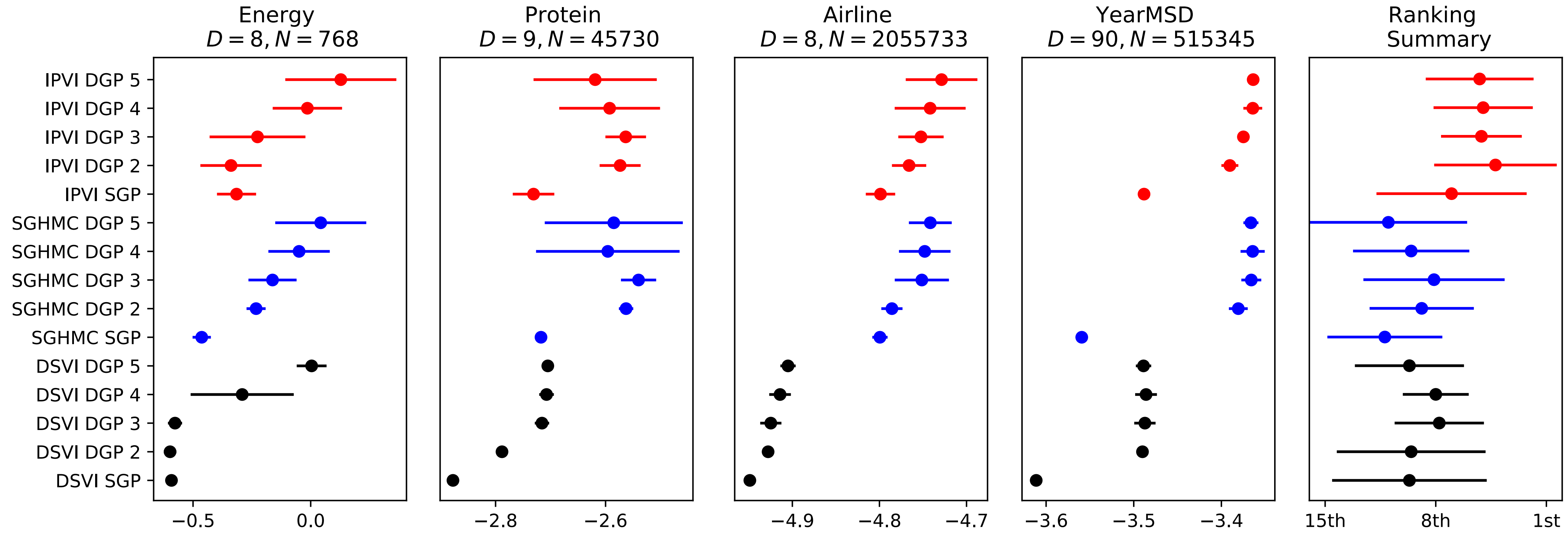




- IPVI is robust under different hyperparameter settings
- Expressive power of IPVI increases as the number of parameters in the generator increase

Experimental Results

- MLL on UCI Benchmark Regression & Real World Regression



Our IPVI DGP SGHMC DGP DSVI DGP

- Our IPVI DGP generally performs the best.

Experimental Results

- Mean test accuracy (%) for 3 classification datasets

Dataset	MNIST		Fashion-MNIST		CIFAR-10	
	SGP	DGP 4	SGP	DGP 4	SGP	DGP 4
DSVI	**97.32**	97.41	86.98	87.99	47.15	51.79
SGHMC	96.41	97.55	85.84	87.08	47.32	52.81
IPVI	97.02	**97.80**	**87.29**	**88.90**	**48.07**	**53.27**

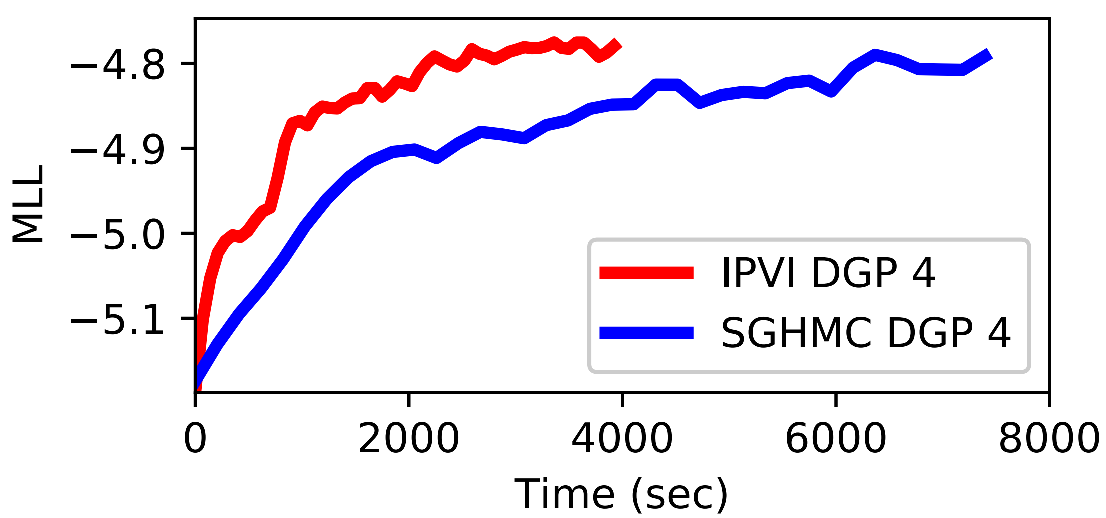
- Our IPVI DGP generally performs the best.

Experimental Results

- Time Efficiency

	IPVI	SGHMC
Average training time (per iter.)	0.35 sec.	3.18 sec.
$\mathcal{U}$ generation (100 samples)	0.28 sec.	143.7 sec.

Time incurred by sampling from a 4-layer DGP model for Airline dataset.



MLL vs. total incurred time to train a 4-layer DGP model for the Airline dataset.

- IPVI is much faster than SGHMC in terms of training as well as sampling.

Conclusion

- A novel IPVI DGP framework
  - Can ideally recover an unbiased posterior belief.
  - Preserve time efficiency.
- Cast the DGP inference into a two-player game
  - Search for Nash equilibrium using BRD
- Parameter-tying architecture
  - Alleviate overfitting
  - Speed up training and prediction

- More details of our paper
  - Detailed architecture of generator and discriminator.
  - Detailed analysis of our BRD algorithm.
  - More experimental results.