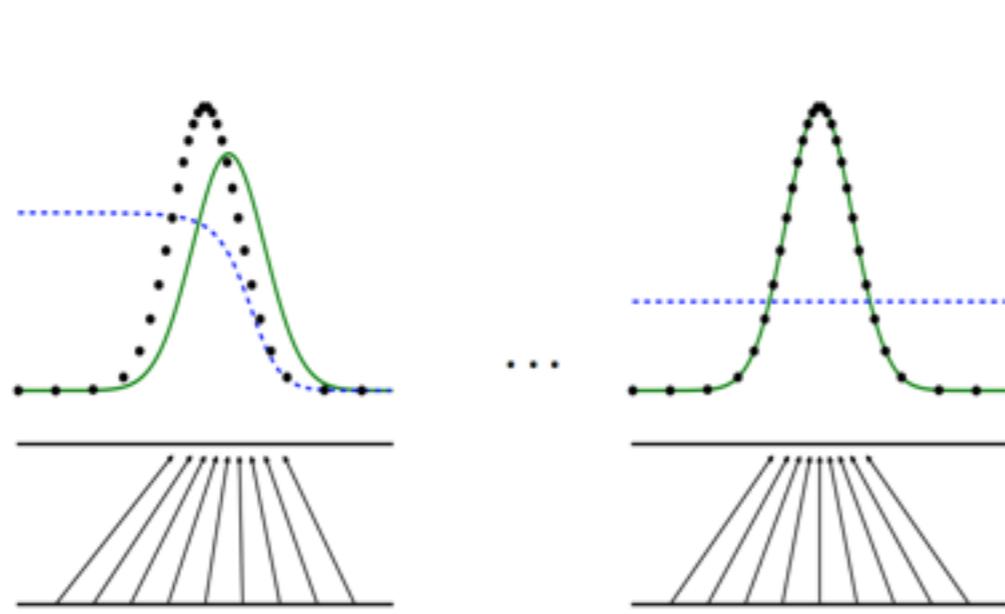


Implicit Posterior Variational Inference for Deep Gaussian Processes (IPVI DGP)



Haibin YU*, Yizhou Chen*,
Bryan Kian Hisang Low,
Patrick Jaillet and Zhongxiang Dai
Department of Computer Science
National University of Singapore
Massachusetts Institute of Technology

* indicates equal contribution

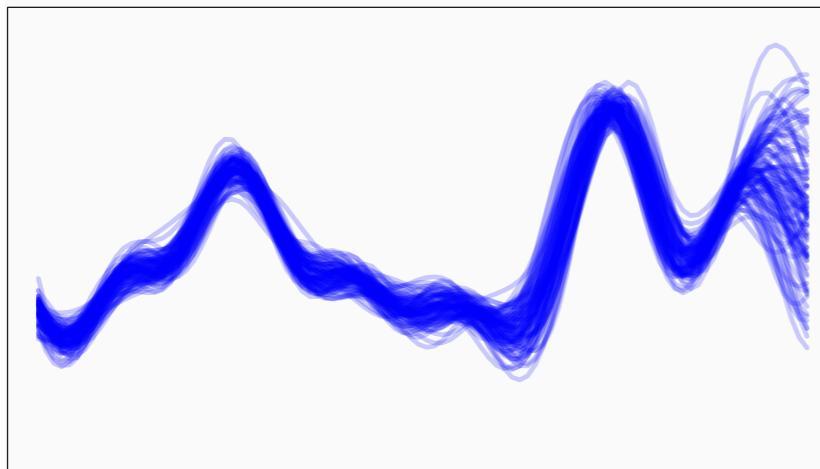
GP V.S. Deep GP

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

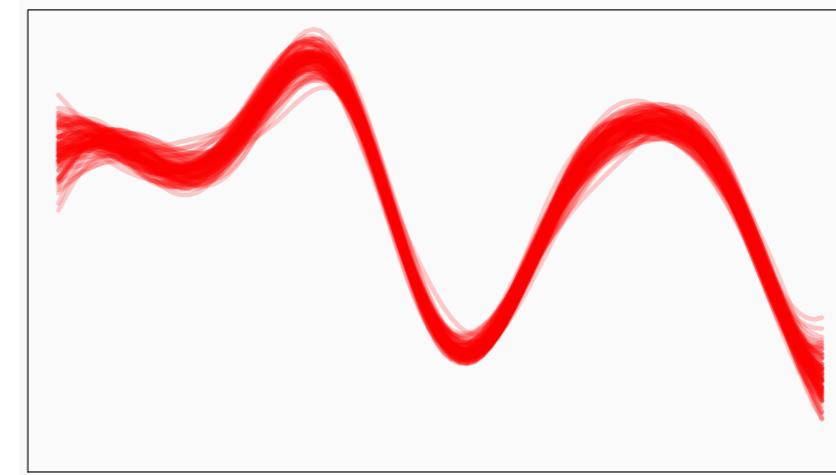


GP V.S. Deep GP

$f(x)$



$g(x)$



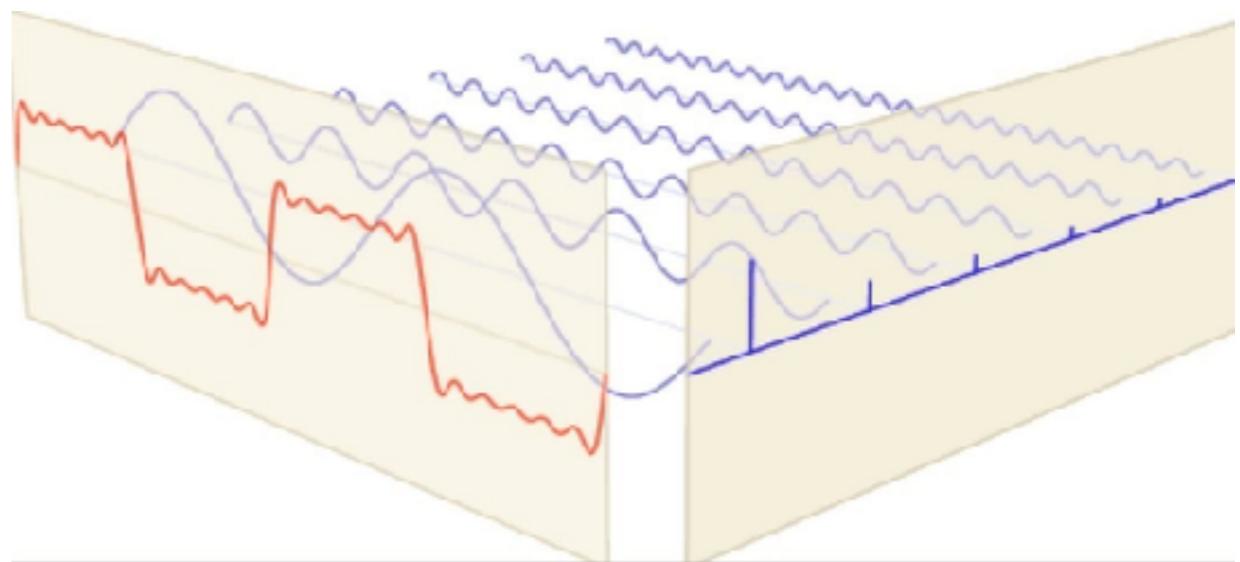
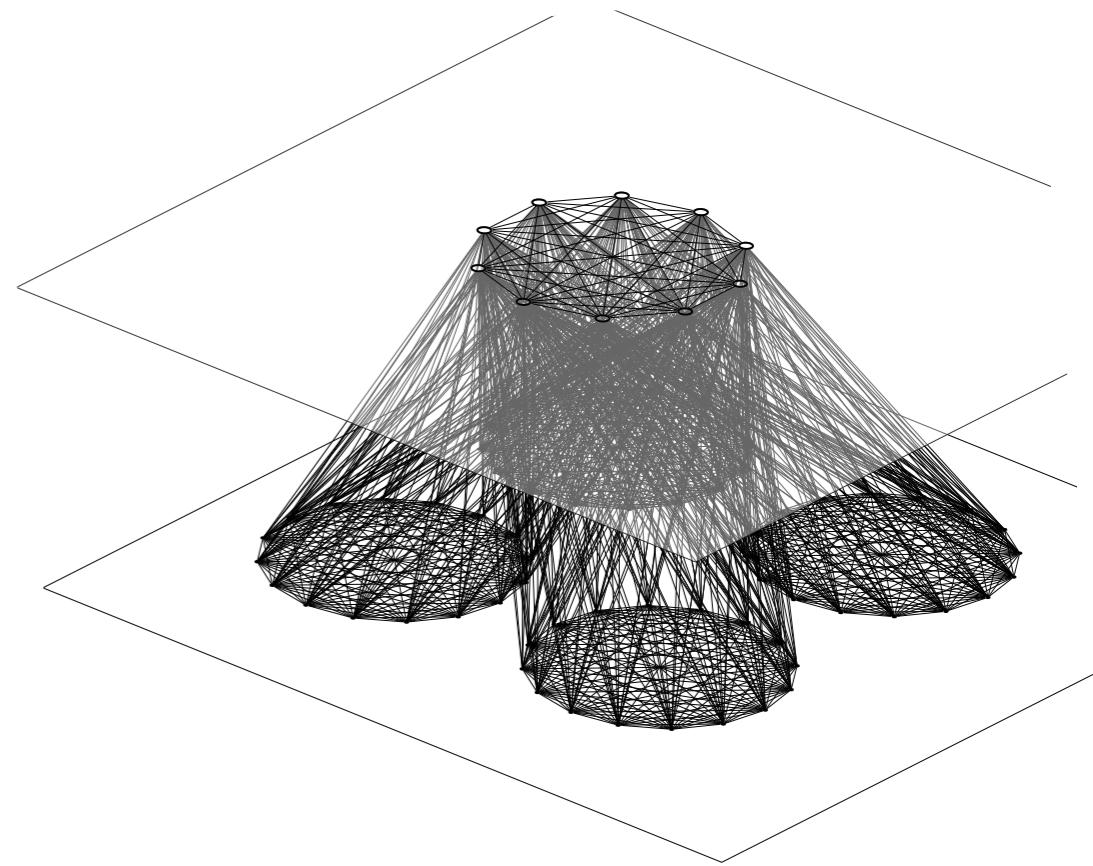
$(f \circ g)(x)$



- Composition of Gaussian processes yields very complex function

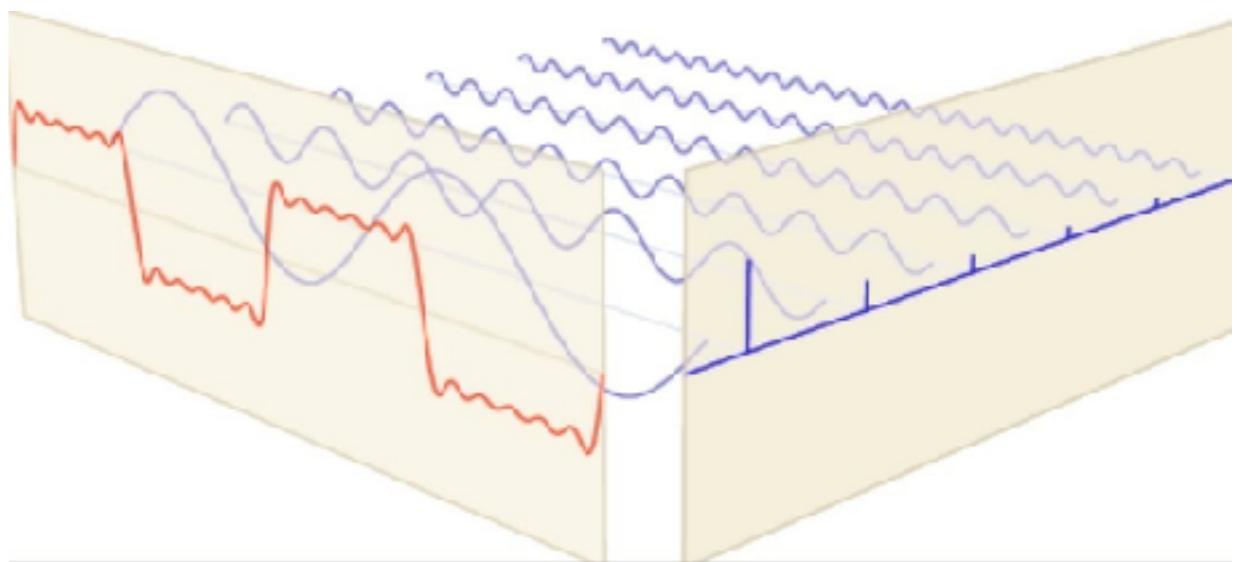
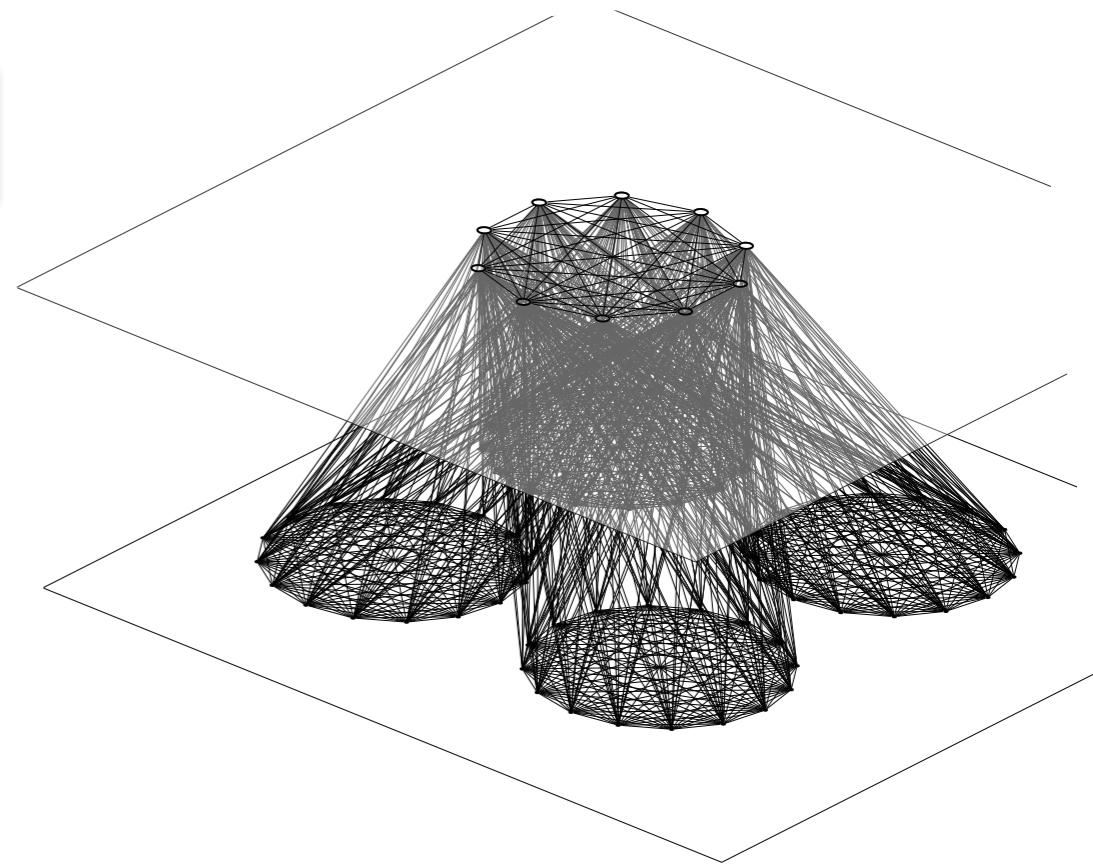
Deep Gaussian Process (DGP)

- Inducing variables based approximations
 - Variational Inference
 - Damianou and Lawrence, AISTATS, 2013
 - Hensman and Lawrence, arXiv, 2014
 - Salimbeni and Deisenroth, NIPS, 2017
 - Expectation Propogation
 - Bui, ICML, 2016
 - MCMC
 - Havasi et al, NIPS 2018
- Random feature approximations
 - Cutajar et al, ICML 2017



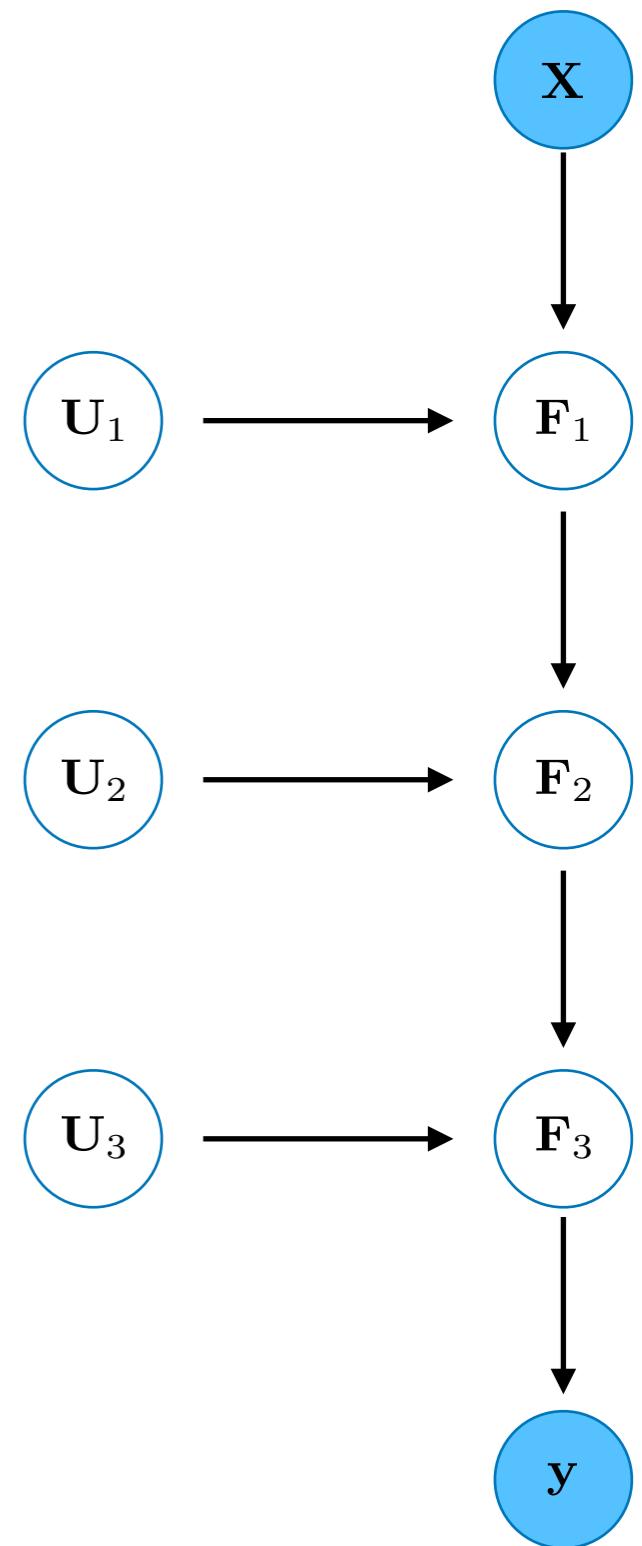
Deep Gaussian Process (DGP)

- Inducing variables based approximations
 - Variational Inference
 - Damianou and Lawrence, AISTATS, 2013
 - Hensman and Lawrence, arXiv, 2014
 - Salimbeni and Deisenroth, NIPS, 2017
 - Expectation Propogation
 - Bui, ICML, 2016
 - MCMC
 - Havasi et al, NIPS 2018
- Random feature approximations
 - Cutajar et al, ICML 2017



Deep GP (DGP)

- Inducing variables $\mathcal{U} = \{\mathbf{U}_1, \dots, \mathbf{U}_L\}$
- Posterior is intractable $p(\mathcal{U}|\mathbf{y})$

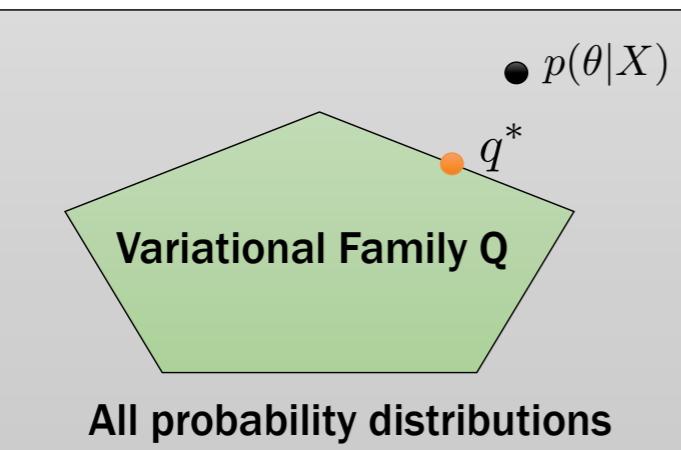


Deep GP (DGP)

- Exact inference is intractable in DGP

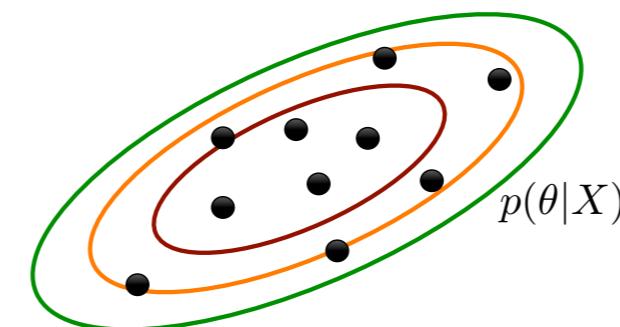
Variational Inference

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



Sampling

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

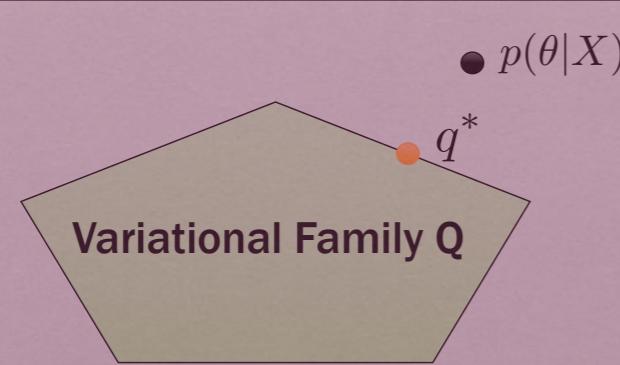


Deep GP (DGP)

- Exact inference is impossible in DGP

Variational Inference

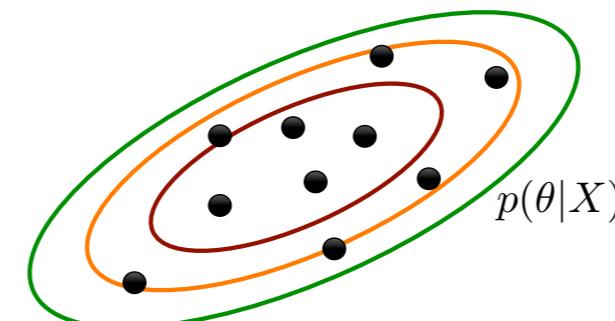
$$q^* = \min_{q \in Q} \text{KL}[q(\theta) || p(\theta|X)]$$



All probability distributions

Sampling

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



Deep GP (DGP)

- Inducing variables $\mathcal{U} = \{\mathbf{U}_1, \dots, \mathbf{U}_L\}$

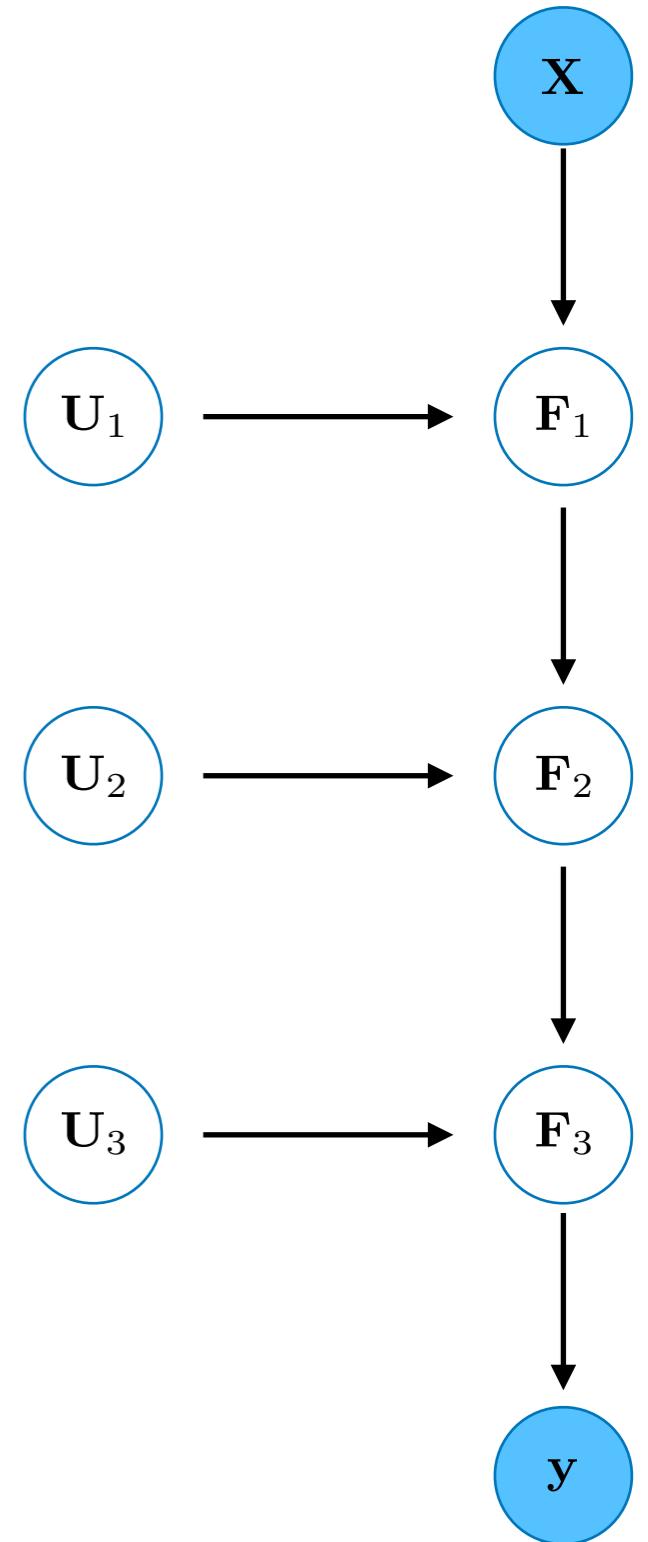
- Posterior is intractable $p(\mathcal{U}|\mathbf{y})$

- Gaussian approximation: $q(\mathbf{U}_l)$

- mean field approximation: $q(\mathcal{U}) = \prod_{l=1}^L q(\mathbf{U}_l)$

$$\text{ELBO} = \int q(\mathbf{F}_L) \log p(\mathbf{y}|\mathbf{F}_L) d\mathbf{F}_L - \text{KL}[q(\mathcal{U})||p(\mathcal{U})]$$

$$q(\mathbf{F}_L) = \int \prod_{l=1}^L p(\mathbf{F}_l|\mathbf{U}_l, \mathbf{F}_{l-1}) q(\mathcal{U}) d\mathbf{F}_1 \dots d\mathbf{F}_{L-1} d\mathcal{U}$$

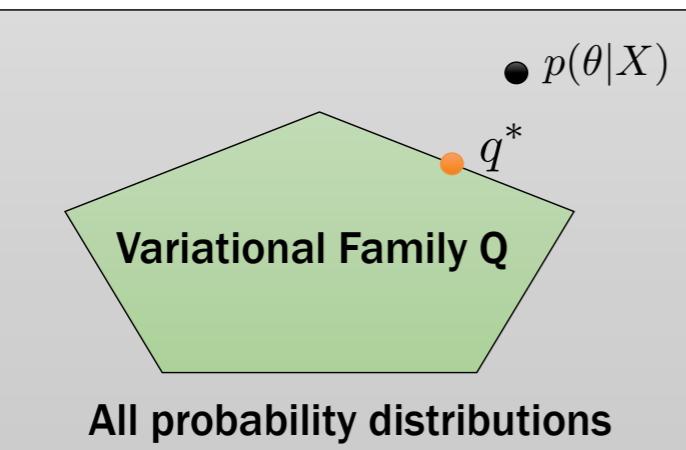


Deep GP (DGP)

- Exact inference is impossible in DGP

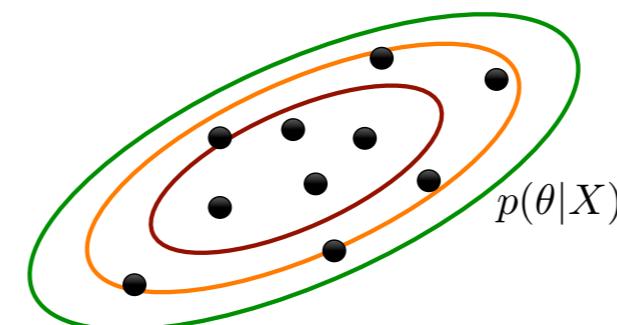
Variational Inference

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



Sampling

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

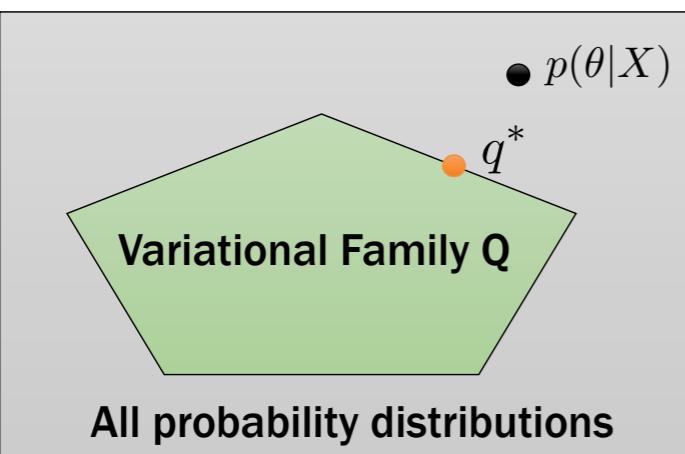


1. biased
2. local minima

Variational Inference

Variational Inference

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

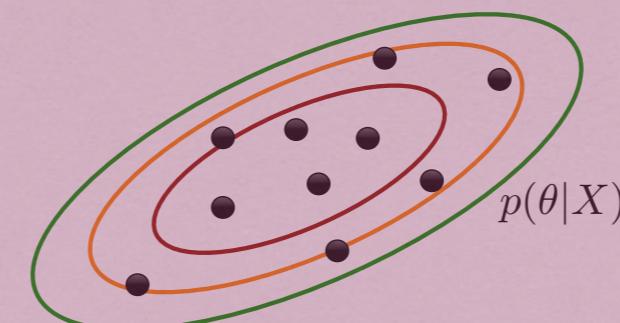


- 1.
- 2.

biased
local minima

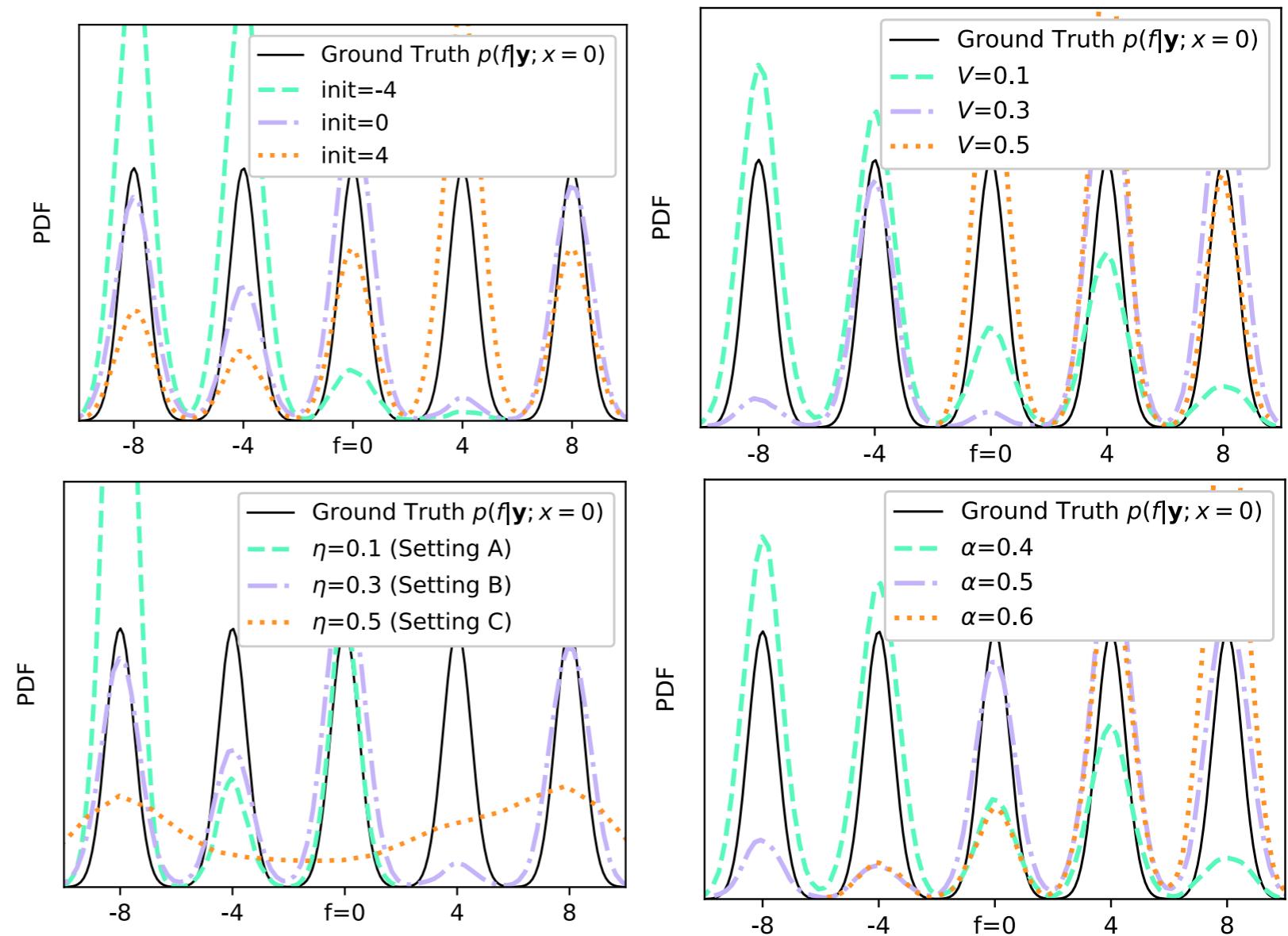
Sampling

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



MCMC

- 5 mode posterior distribution
- Various parameter setting for MCMC method
 - init
 - V
 - η
 - α

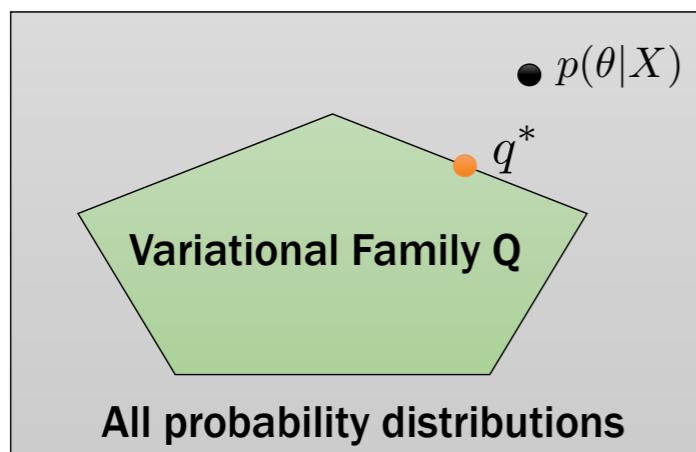


Deep GP (DGP)

- Stochastic Gradient Markov Chain Monte Carlo

Variational Inference

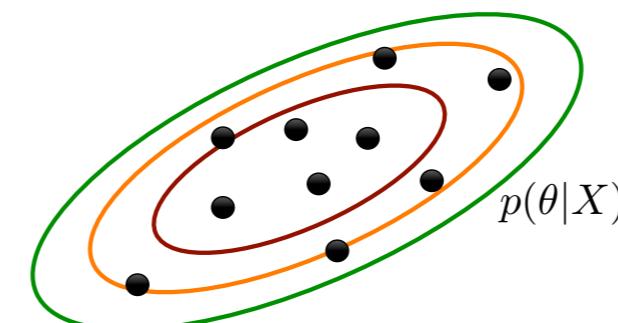
$$q^* = \min_{q \in Q} \text{KL}[q(\theta) || p(\theta|X)]$$



1. biased
2. local minima

Sampling

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

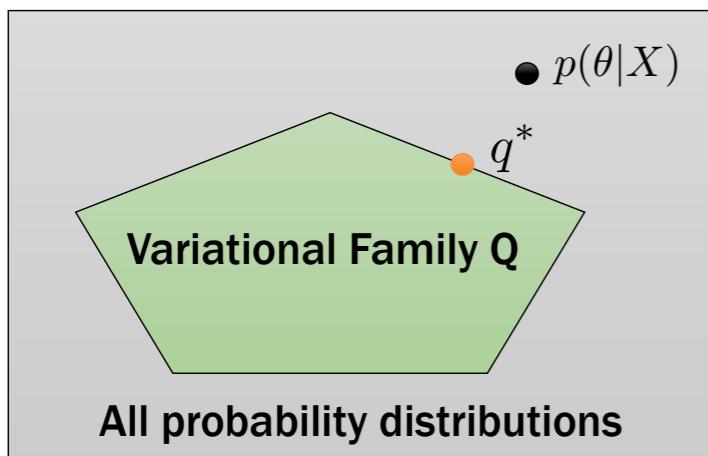


1. local modes
2. efficiency

Deep GP (DGP)

Variational Inference

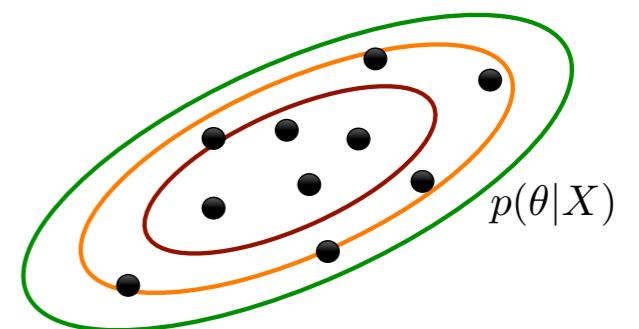
$$q^* = \min_{q \in Q} \text{KL}[q(\theta) || p(\theta|X)]$$



efficiency

Sampling

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

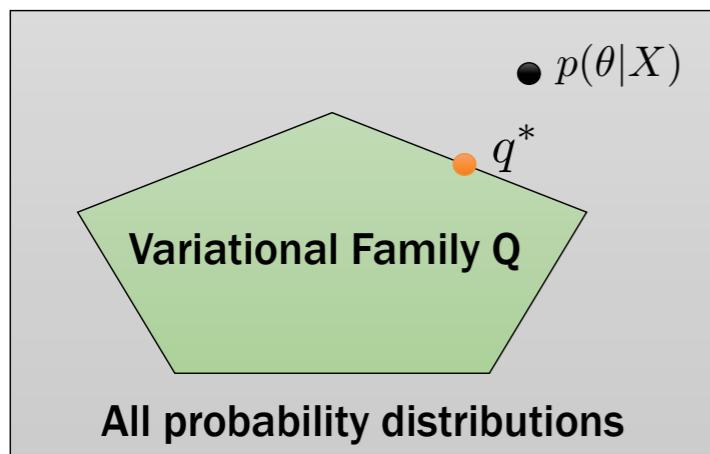


unbiased

Deep GP (DGP)

Variational Inference

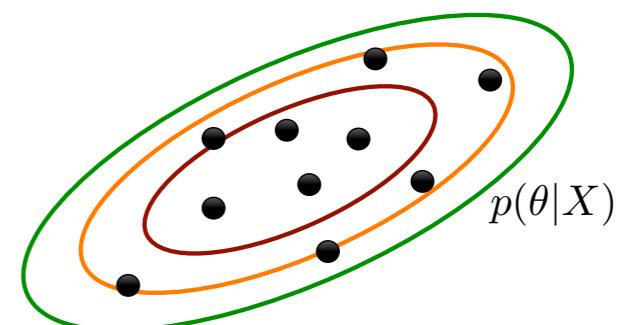
$$q^* = \min_{q \in Q} \text{KL}[q(\theta) || p(\theta|X)]$$



efficiency

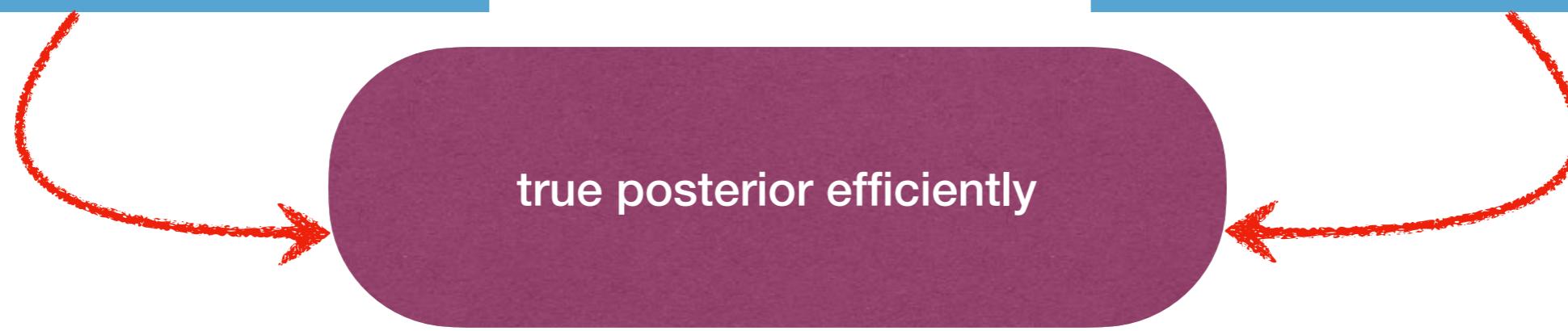
Sampling

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

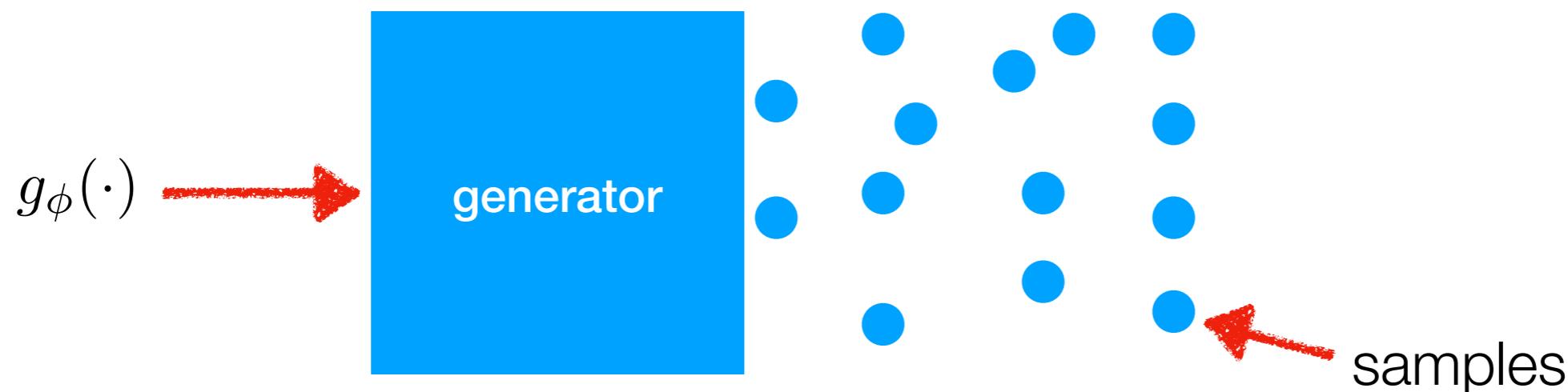


unbiased

true posterior efficiently

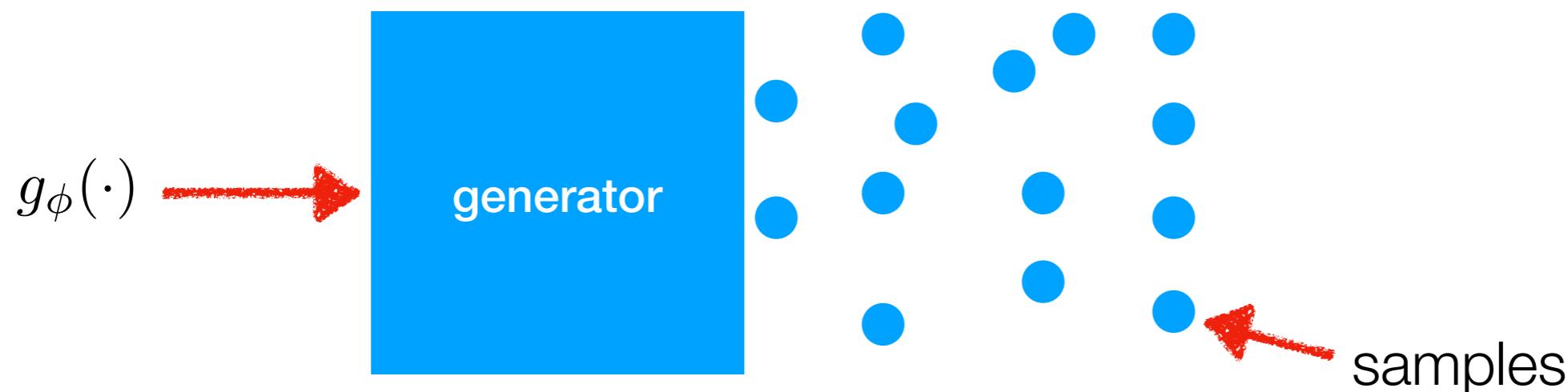


Implicit Posterior Variational Inference



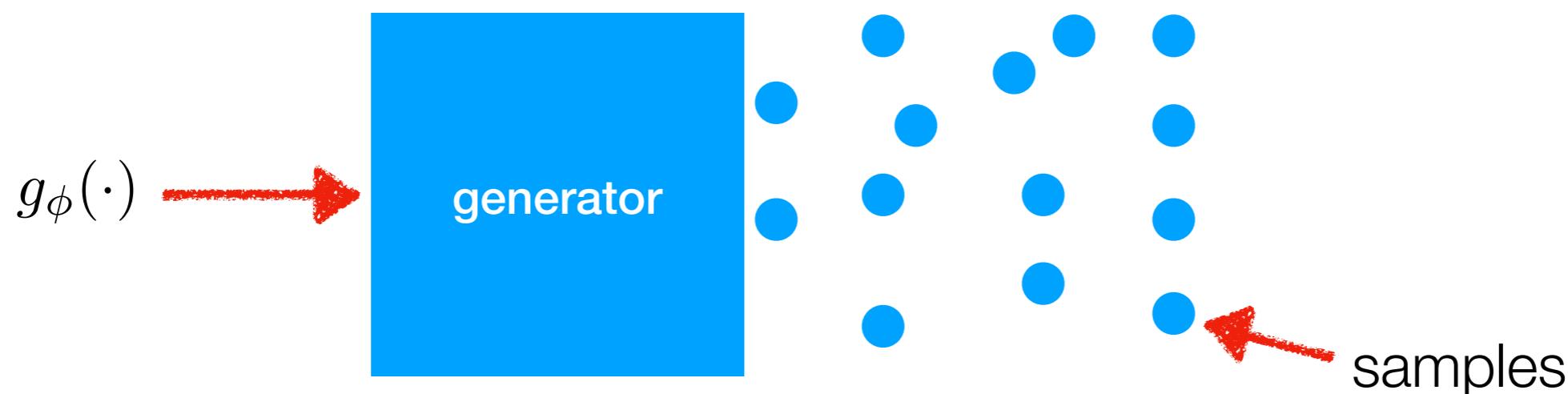
$$\text{ELBO} = \mathbb{E}_{q(\mathbf{F}_L)}[\log p(\mathbf{y}|\mathbf{F}_L)] - \text{KL}[q_\Phi(\mathcal{U})||p(\mathcal{U})]$$

Implicit Posterior Variational Inference



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

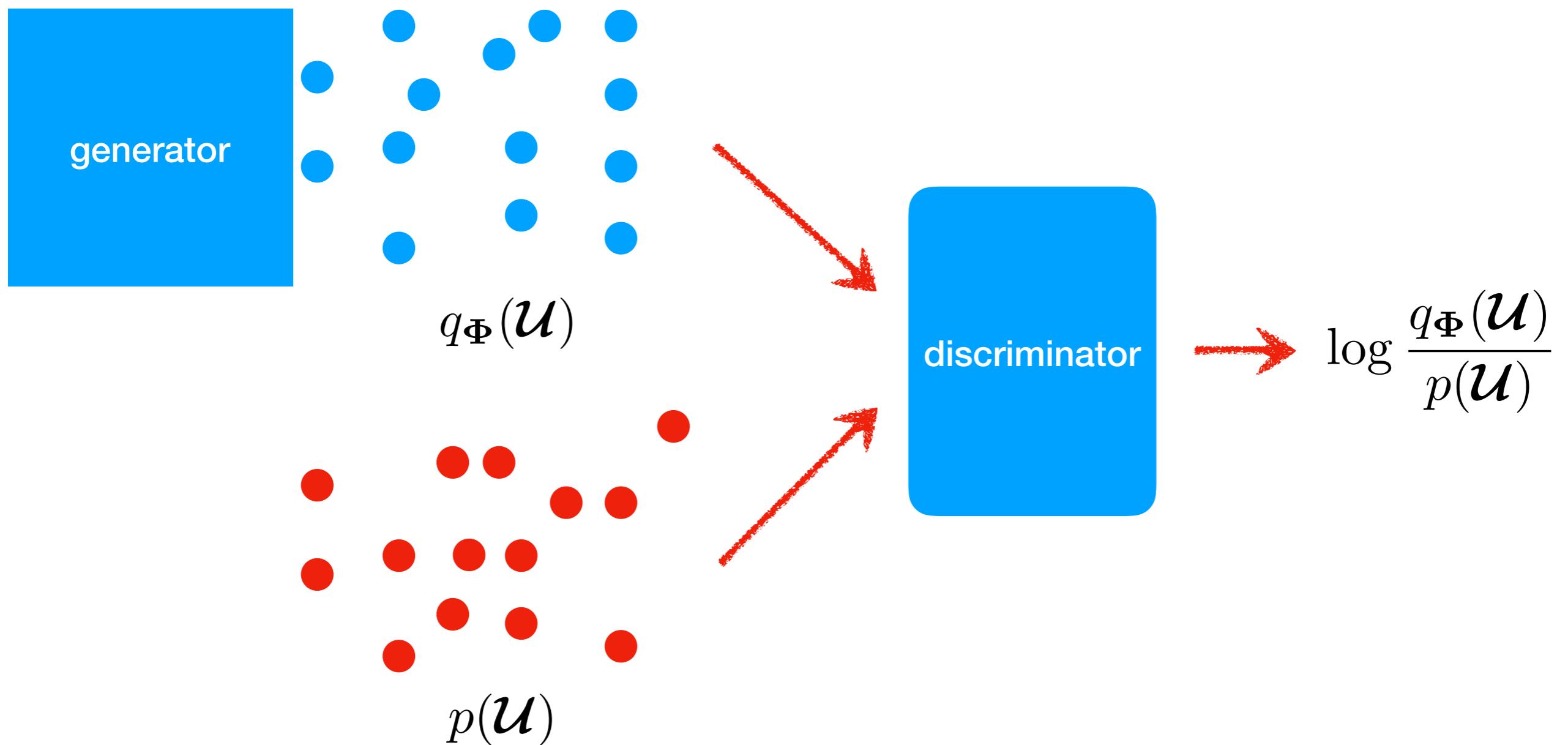
Implicit Posterior Variational Inference



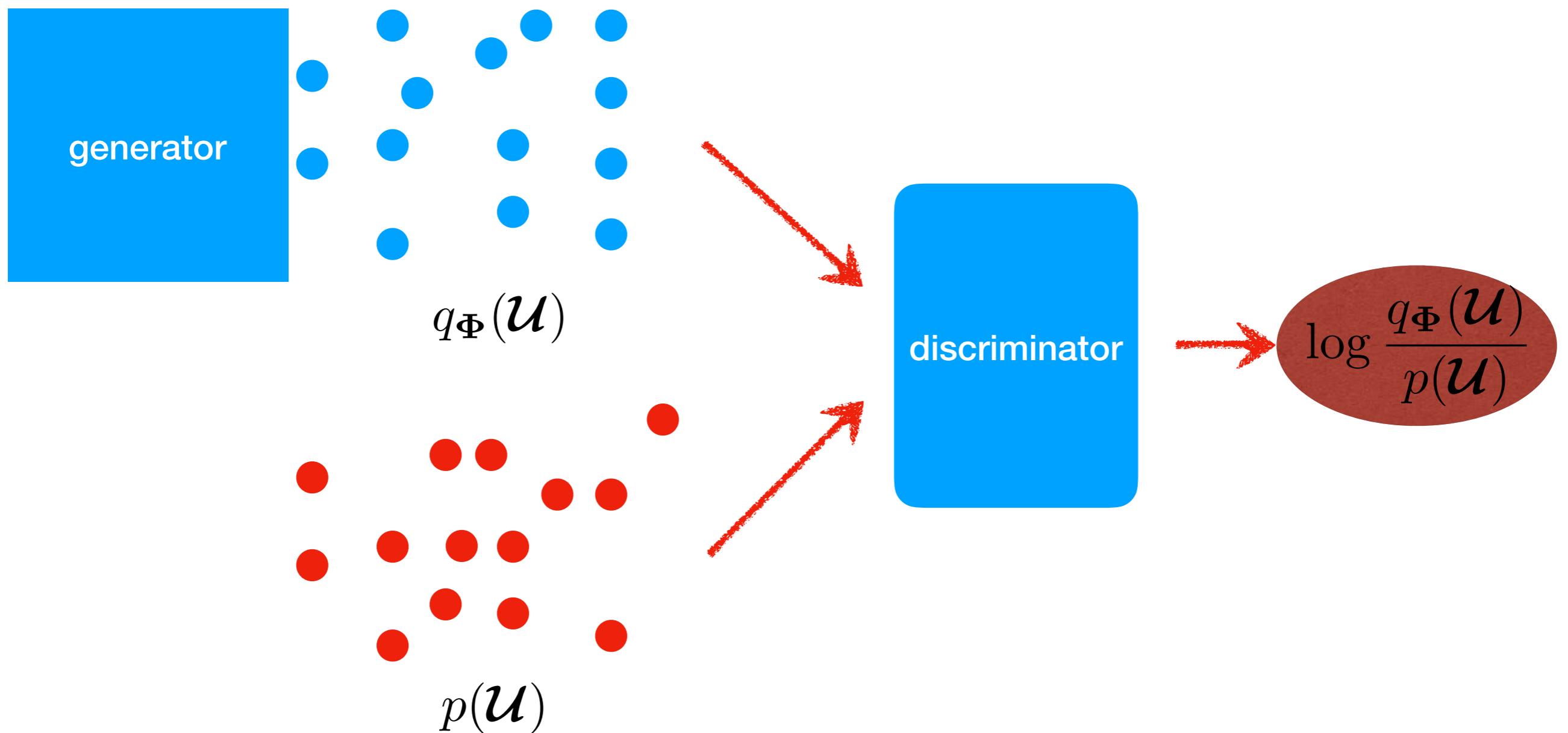
$$\text{ELBO} = \mathbb{E}_{q(\mathbf{F}_L)}[\log p(\mathbf{y}|\mathbf{F}_L)] - \text{KL}[q_\Phi(\mathcal{U})||p(\mathcal{U})]$$

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

Implicit Posterior Variational Inference



Implicit Posterior Variational Inference



Implicit Posterior Variational Inference

Proposition Let $\sigma(x) \triangleq 1/(1 + \exp(-x))$. Consider the following maximization problem:

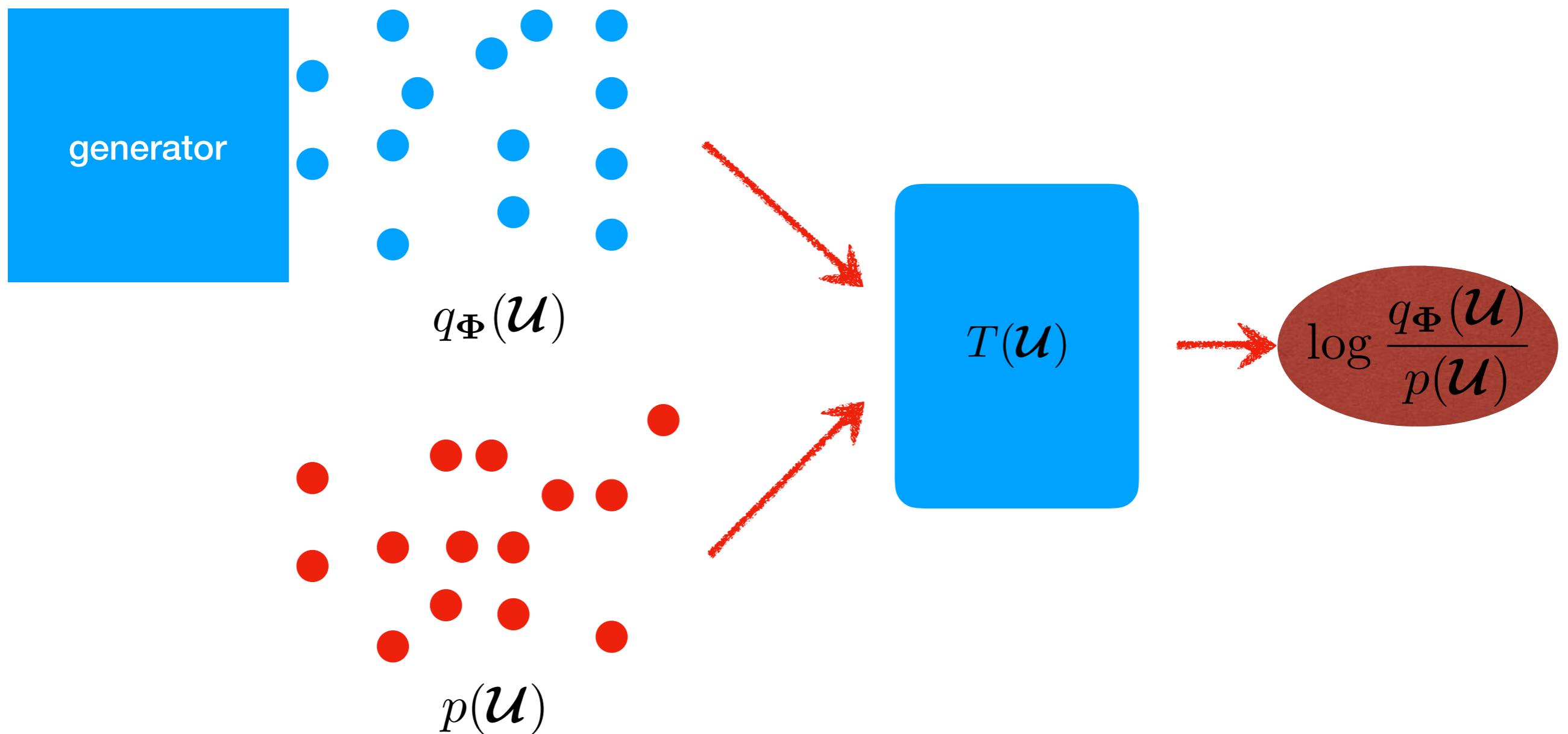
$$\max_T \mathbb{E}_{p(\mathcal{U})}[\log(1 - \sigma(T(\mathcal{U})))] + \mathbb{E}_{q_\Phi(\mathcal{U})}[\log \sigma(T(\mathcal{U}))] .$$

If $p(\mathcal{U})$ and $q_\Phi(\mathcal{U})$ are known, then the optimal T^* is the log-density ratio:

$$T^*(\mathcal{U}) = \log q_\Phi(\mathcal{U}) - \log p(\mathcal{U}) .$$



Implicit Posterior Variational Inference



Implicit Posterior Variational Inference

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

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

IBR: Iterative Best Response

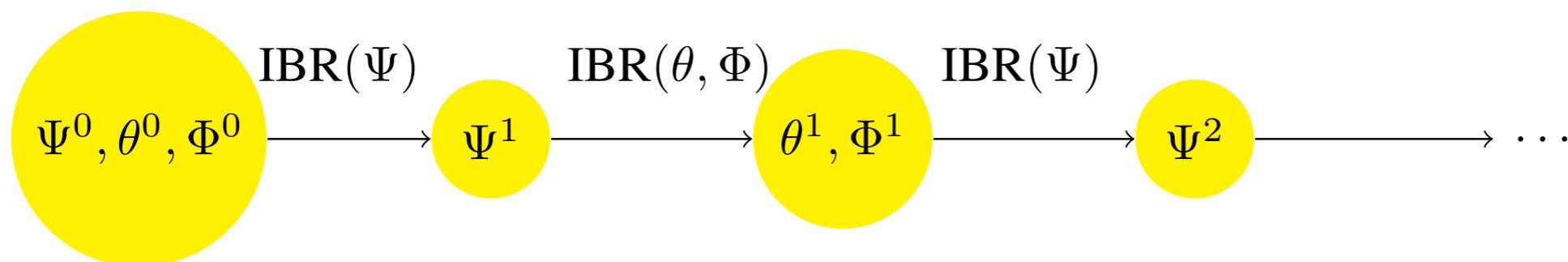


Figure 1: IBR between two players

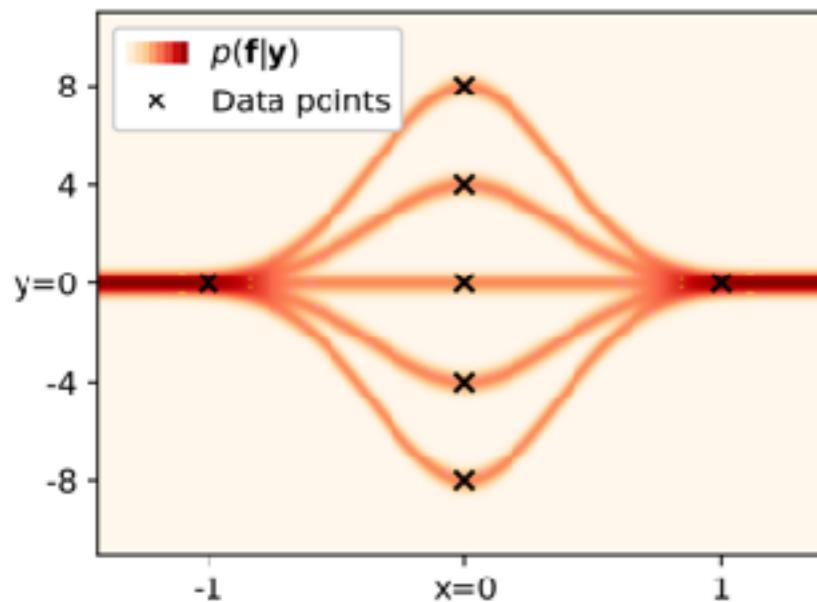
Experimental Results

- Metrics for evaluation
 - RMSE
 - MNLP
 - MLL: mean log likelihood
 - Jenson Shannon Divergence: measure the divergence between distributions

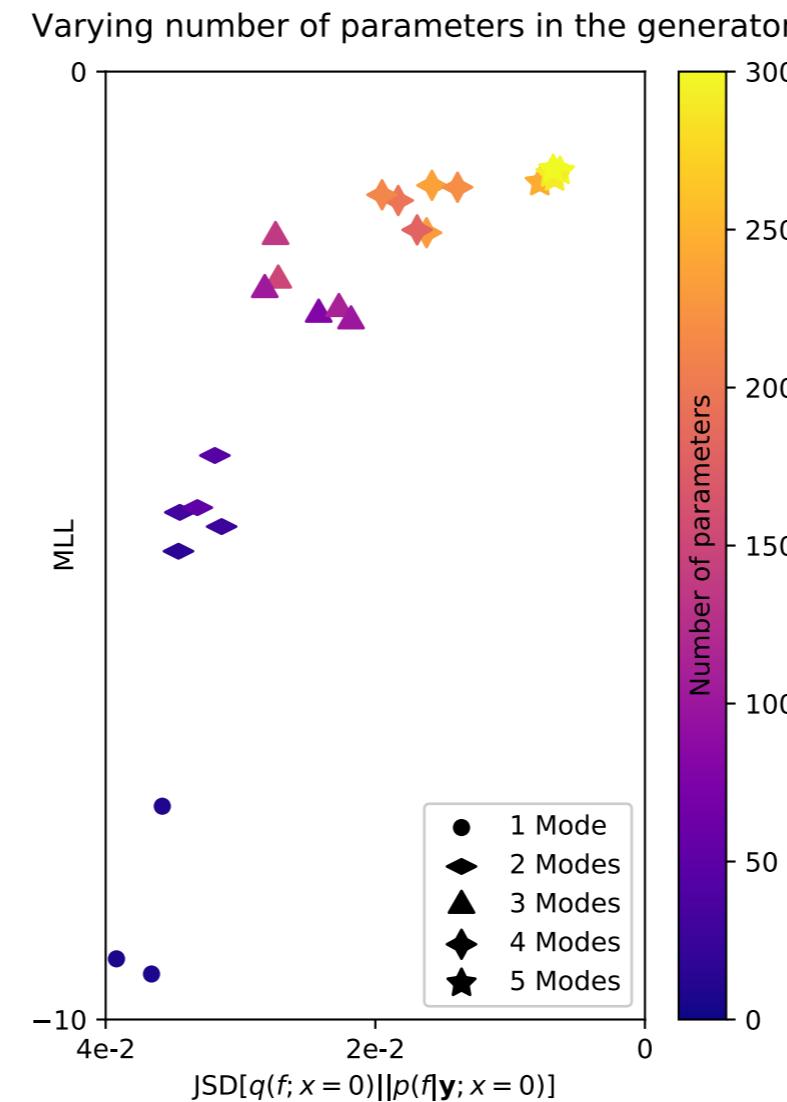
- Algorithms for comparison
 - DSVI DGP: Doubly stochastic variational inference DGP [Salimbeni and Deisenroth, 2017]
 - SGHMC DGP: Stochastic gradient Hamilton Monte Carlo DGP [Havasi et al, 2018]

Experimental Results

- Synthetic Experiment: Learning a Multi-Modal Posterior Belief



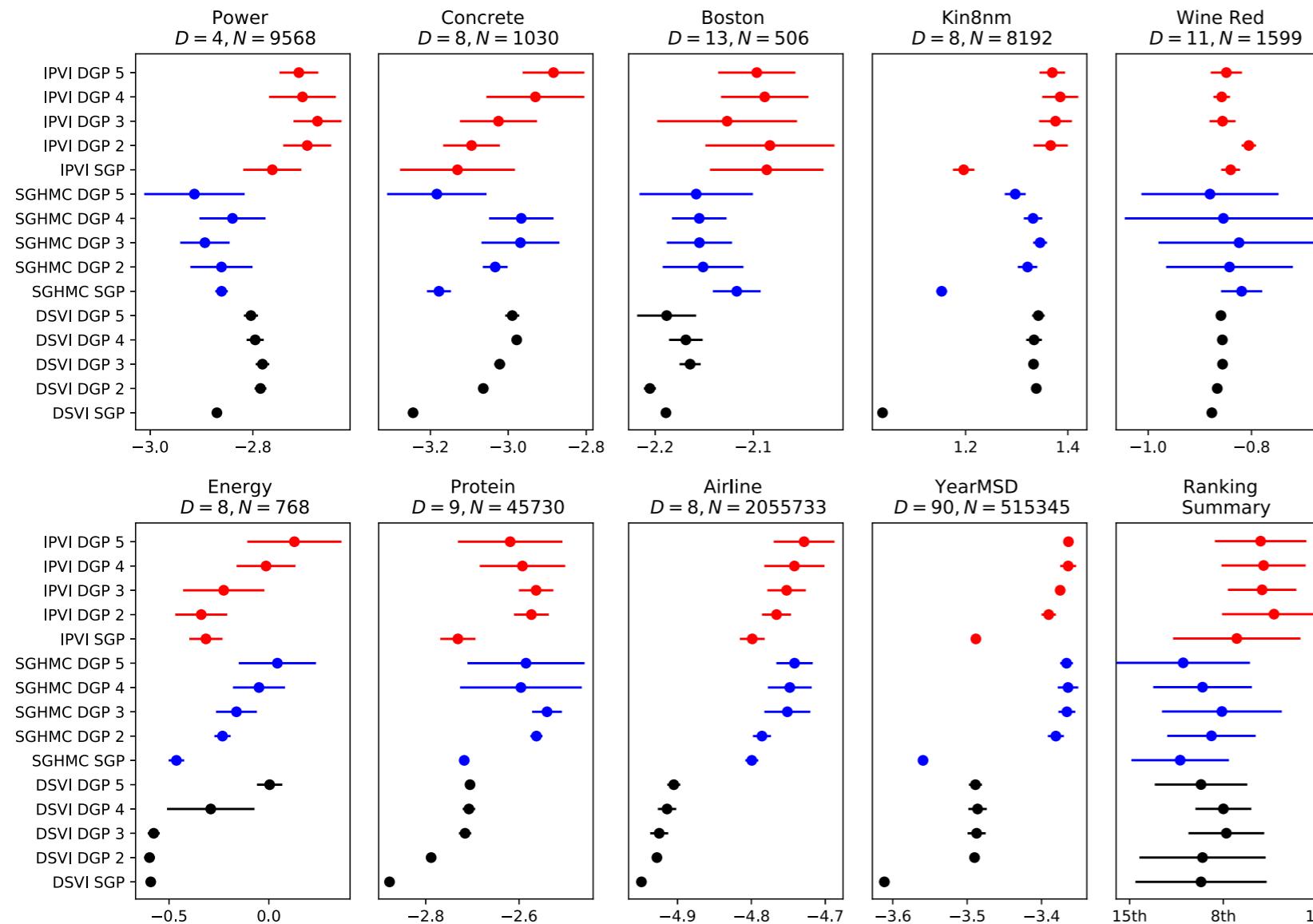
Setting	JSD	MLL
IPVI A ($LR = 1e^{-4}$)	$1.0e^{-2}$	-1.15
IPVI B ($LR = 1e^{-3}$)	$8.3e^{-3}$	-0.99
IPVI C ($LR = 1e^{-2}$)	$8.6e^{-3}$	-1.02
SGHMC A ($\eta = 0.1$)	$2.1e^{-2}$	-2.36
SGHMC B ($\eta = 0.3$)	$1.2e^{-2}$	-1.10
SGHMC C ($\eta = 0.5$)	$7.5e^{-2}$	-2.83



- IPVI is better than SGHMC
- Expressiveness of IPVI increases w.r.t the increase of parameters.

Experimental Results

- MNLP: UCI Benchmark Regression & Real World Regression



- Our IPVI DGP
- SGHMC DGP: [Havasi et al, 2018]
- DSVI DGP:[Salimbeni and Deisenroth, 2017]
- Our IPVI DGP performs the best.

Experimental Results

- Classification

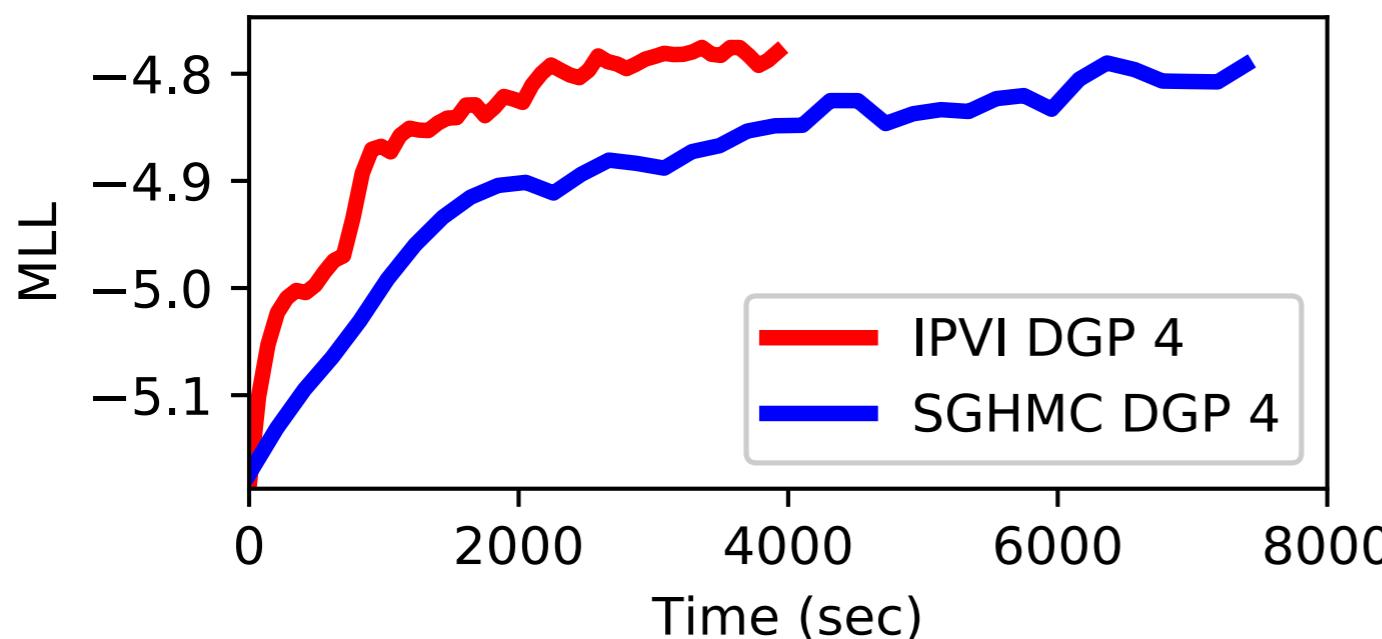
Dataset	MNIST		Fashion-MNIST		CIFAR-10	
	SGP	DGP 4	SGP	DGP 4	SGP	DGP 4
DSVI	97.32	97.41	86.98	87.99	47.15	51.79
SGHMC	96.41	97.55	85.84	87.08	47.32	52.81
IPVI	97.02	97.80	87.29	88.90	48.07	53.27

- Mean test accuracy
- 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

- 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 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

- Present a novel IPVI DGP framework
 - Boost expressive power of GP
 - Recover unbiased posterior belief
 - Preserve time efficiency

