Approximate Inference: An Intro

Yingzhen Li

yingzhen.li@imperial.ac.uk

The Central Computation for Inference

- Inference: infer the unknowns
  - Unobserved/latent variables in the model
  - Quantities depending on the latent variables in the model

Random variable (unobserved)

prob. density

$$\int \underbrace{F(\theta)}_{\text{integrand function}} \underbrace{\pi(\theta)d\theta}_{\text{probability measure}}$$

(For discrete probability measures, integration becomes discrete sum.)

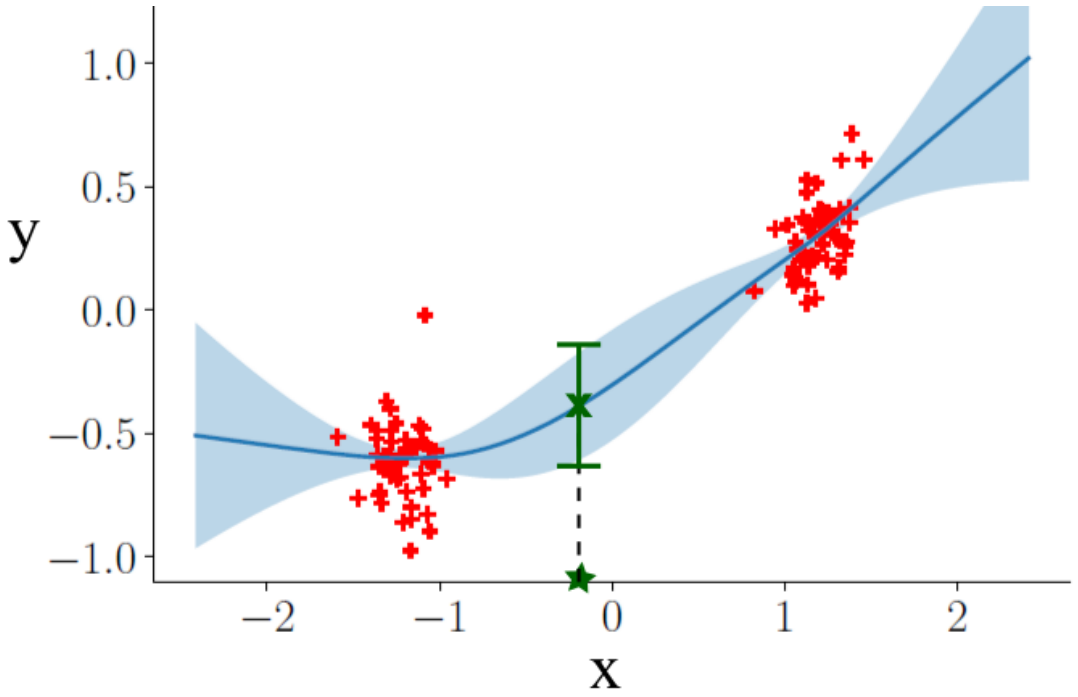
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the prediction distribution of the test output given a test input?"

$F(\theta) = p(y|x,\theta), \pi(\theta) = p(\theta \mid D),$
D = observed datapoints



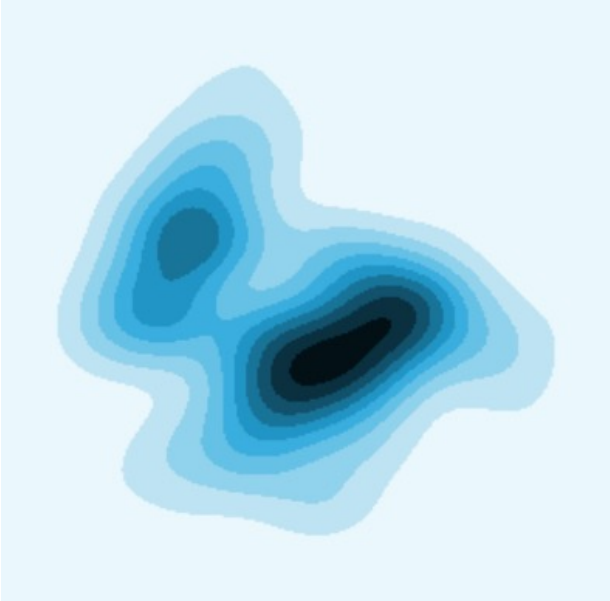
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the mean of this distribution?"

$F(\theta) = \theta, \pi(\theta)$ can be complicated and high dimensional



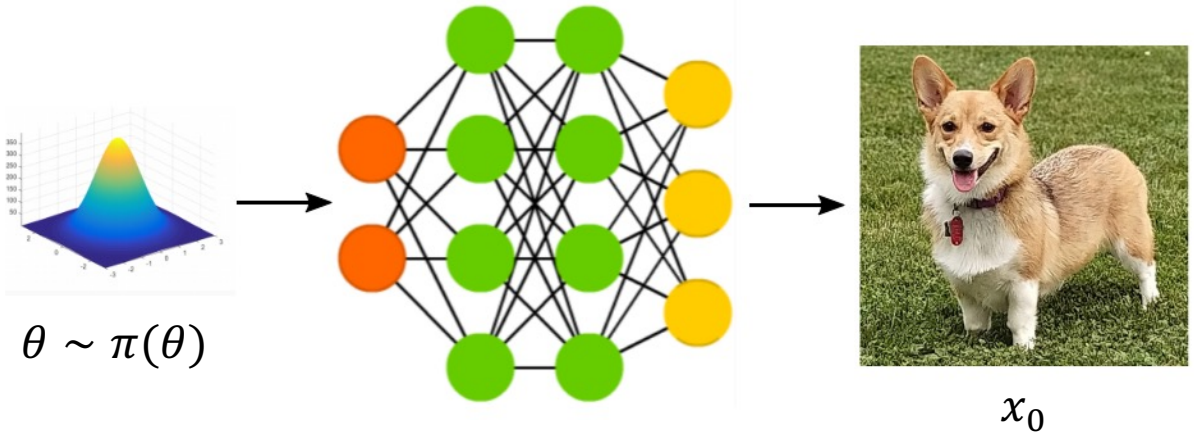
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the probability of generating this image?"

$F(\theta) = \delta(NN(\theta) = x_0), \pi(\theta) = N(0, I)$



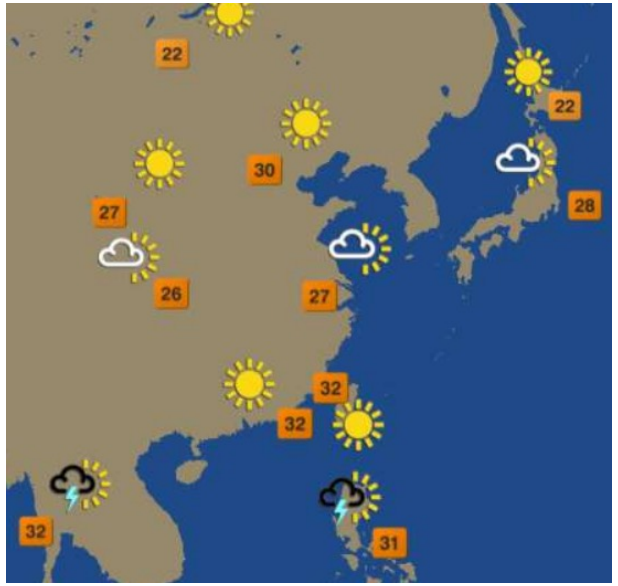
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the weather forecast for tomorrow?"

Answering this in a Bayesian way:
θ: forecasting simulator settings
D: historical weather record
$F(\theta) = Simulator(\theta), \pi(\theta) = p(\theta \mid D)$



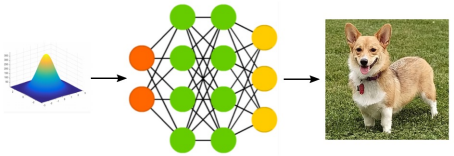
Nature laughs at the difficulties of integration.

--Pierre-Simon Laplace

Gordon and Sorkin. *The Armchair Science Reader.* New York 1959

Integration in Bayesian Computation

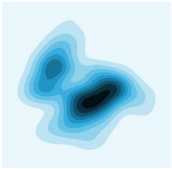
The integrand $F(\theta)$ is intractable

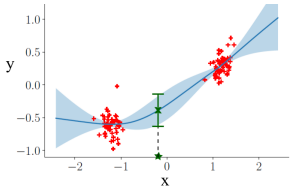


Implicit Models

Bayesian Optimization, Probabilistic Numerics

The probability distribution $\pi(\theta)$ is intractable





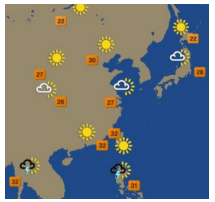
Approximate Inference

(in a strict sense)

This lecture

$$\int F(\theta)\pi(\theta)d\theta$$

Both $F(\theta)$ and $\pi(\theta)$ are intractable



Approximate Bayesian Computation

Approximate Inference

- Central task: approximate $\pi(\theta)$


$$q(\theta) \approx \pi(\theta)$$

- Different from fitting a model p to data with dataset $D \sim p_{data}$:
  - Cannot directly sample from $\pi(\theta)$
  - Sometimes know the form of unnormalized density for $\pi(\theta)$
  - Sometimes even the unnormalized density of $\pi(\theta)$ is expensive to compute

(Assumed $\int F(\theta)q(\theta)d\theta$ can be computed or approximated efficiently.)

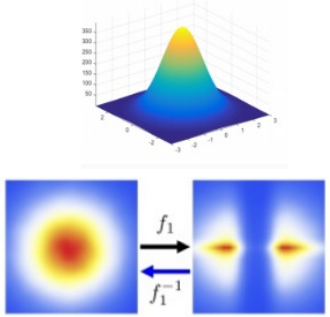
Approximate Inference

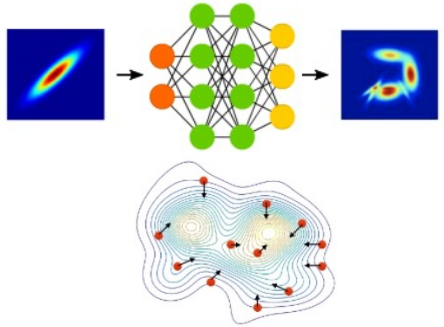
- Central task: approximate $\pi(\theta)$

$$\underline{q(\theta)} \underline{\approx} \pi(\theta)$$

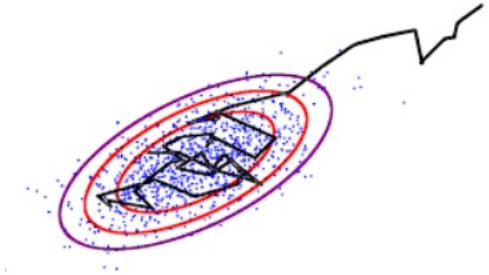
Approximate distribution design

Algorithm for fitting $q(\theta)$ to $\pi(\theta)$





$$\min\ Loss(q(\theta), \pi(\theta))$$



Explicit distributions

Implicit distributions

Optimisation-based approaches

Sampling-based approaches

(Assumed $\int F(\theta)q(\theta)d\theta$ can be computed or approximated efficiently.)

Tutorial Outline

Basics

Variational inference

Scalable variational inference

Monte Carlo techniques

Amortized inference

Advances

q distribution design

Optimization objective design

Future directions

Bayesian Inference

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$

- $P(\theta)$: prior
- $P(D \mid \theta)$: likelihood
- $P(\theta \mid D)$: posterior
- $P(D)$: marginal

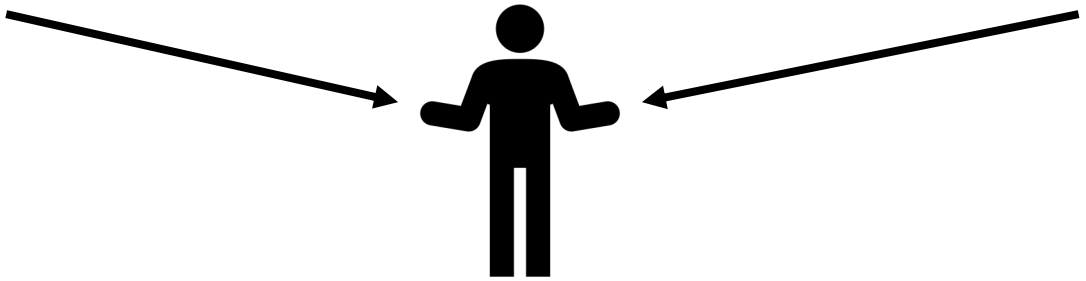
Variational Inference (VI)

The posterior

$$p(\theta|D) = p(D\,|\theta)p(\theta)/p(D)$$

The variational distribution

$$q_\phi(\theta)$$



Kullback-Leibler (KL) divergence

Kullback-Leibler Divergence

$$KL[q(\theta)||p(\theta)] = -\int q(\theta) \log \frac{p(\theta)}{q(\theta)} \, d\theta = E_{q(\theta)}[\log \frac{p(\theta)}{q(\theta)}]$$

- When $p = q$, KL is 0
- Otherwise, KL > 0
- It measures how similar are these two distributions

Let's Derive the Objective of VI

- Minimize $KL[q(\theta)||p(\theta|D)]$

$$KL[q(\theta)||p(\theta|D)] = -E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta|D)}{q(\theta)}\right]$$

$$= -E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta,D)}{p(D)q(\theta)}\right] = -E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta,D)}{q(\theta)} - \log p(D)\right]$$

$$= \boxed{\log p(D)} - E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta,D)}{q(\theta)}\right]$$

Model Evidence

Let's Derive the Objective of VI

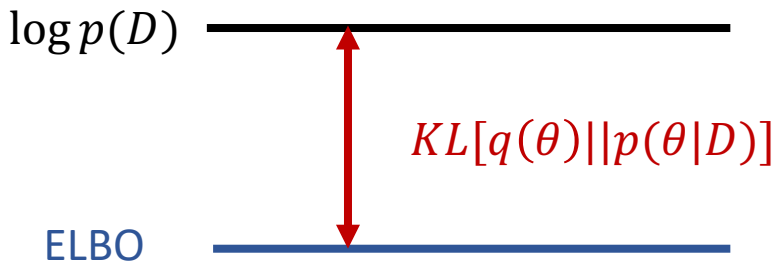
Minimize $KL[q(\theta)||p(\theta|D)]$

$$KL[q(\theta)||p(\theta|D)] = \boxed{\log p(D)} - E_{q(\theta)}\left[\log \frac{\mathrm{p}(\theta, D)}{q(\theta)}\right]$$

Model Evidence

Maximize $L = E_{q(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q(\theta)}\right]$

Evidence Lower Bound (ELBO)



"Model Evidence = ELBO + KL"

Variational Inference (VI)

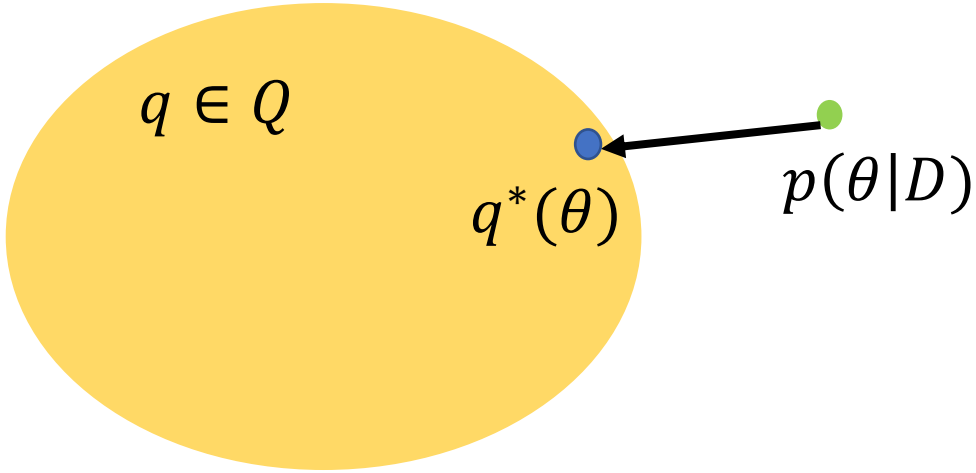
The posterior

$$p(\theta|D) = p(D|\theta)p(\theta)/p(D)$$

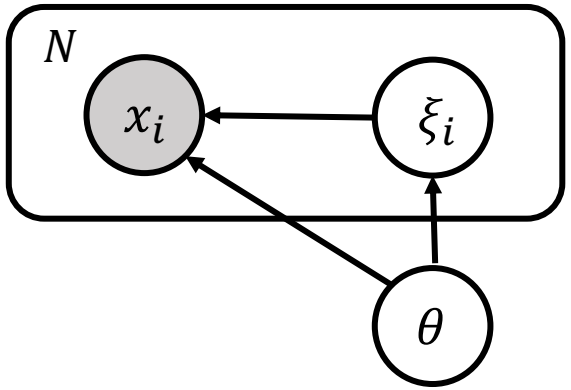
The variational distribution

$$q_\phi(\theta)$$

$$L = E_{q_{\phi(\theta)}}\left[\log\frac{\mathrm{p(D},\theta)}{q_\phi(\theta)}\right] = \log p(D) - KL[\,q_\phi(\theta)||p(\theta)]$$



Stochastic Variational Inference



$$\mathrm{p}(\theta, \boldsymbol{\xi}, \boldsymbol{x}) = p(\theta) \prod_{i=1}^{N} p(\xi_i|\theta) p(x_i|\xi_i, \theta)$$

$$L = E_q\left[\log \frac{\mathrm{p}(\theta, \boldsymbol{\xi}, \boldsymbol{x})}{q\,(\theta, \xi)}\right]$$

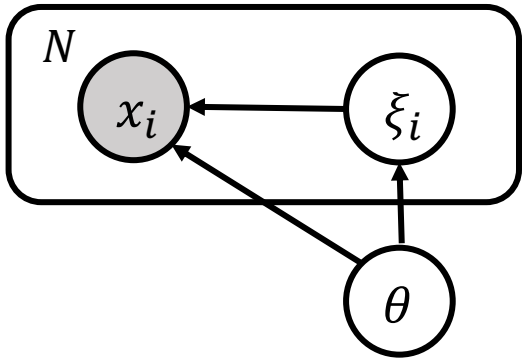
$$= E_q\left[\log \frac{p(\theta)\ \prod_{i=1}^{N} p(\xi_i|\theta) p(x_i|\xi_i, \theta)}{q(\theta) \prod_{i=1}^{N} q(\xi_i)}\right]$$

$$= E_q\left[\log \frac{p(\theta)}{q(\theta)}\right] + \sum_{i=1}^{N} E_q\left[\log \frac{p(\xi_i|\theta) p(x_i|\xi_i)}{q(\xi_i)}\right]$$

- O(N) time to compute in each update iteration
- N can be extremely large
- Even one iteration might not be affordable

Hoffman et al. Stochastic Variational Inference. JMLR 2013.

Stochastic Variational Inference



$$L = E_q\left[\log\frac{p(\theta)}{q(\theta)}\right] + \sum_{i=1}^{N} E_q\left[\log\frac{p(\xi_i|\theta)p(x_i|\xi_i)}{q(\xi_i)}\right] \xrightarrow{\text{gradient}} \nabla L = \nabla E_q\left[\log\frac{p(\theta)}{q(\theta)}\right] + \sum_{i=1}^{N} \nabla E_q\left[\nabla\log\frac{p(\xi_i|\theta)p(x_i|\xi_i)}{q(\xi_i)}\right]$$

Stochastic approximation with $M \ll N$

Stochastic approximation with $S \ll N$

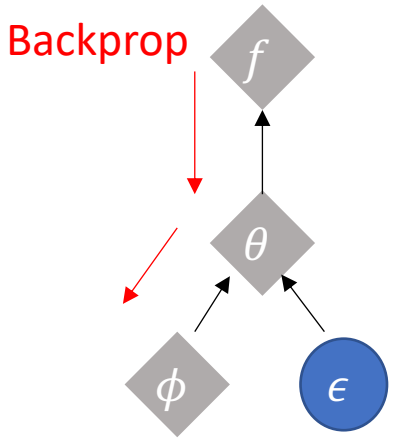
$$\hat{L} = E_q\left[\log\frac{p(\theta)}{q(\theta)}\right] + \frac{N}{M}\sum_{i=1}^{M} E_q\left[\log\frac{p(\xi_i|\theta)p(x_i|\xi_i)}{q(\xi_i)}\right] \xrightarrow{\text{gradient}} \nabla\hat{L} = \nabla E_q\left[\log\frac{p(\theta)}{q(\theta)}\right] + \frac{N}{M}\sum_{i=1}^{M} \nabla E_q\left[\nabla\log\frac{p(\xi_i|\theta)p(x_i|\xi_i)}{q(\xi_i)}\right]$$

Stochastic Gradient

How to compute the gradients?

Hoffman et al. Stochastic Variational Inference. JMLR 2013.

Reparameterization Trick



Express $q_\phi(\theta)$ as $\epsilon \sim r(\epsilon),\ \theta = g(\epsilon, \phi)$

$$\theta \sim N(\mu, \sigma^2)$$
$$\epsilon \sim N(0,1), \theta = \mu + \sigma\epsilon$$

ELBO $L = E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right] = E_{r(\epsilon)}\left[\log \frac{\mathrm{p}(g(\epsilon,\phi),D)}{q_\phi(g(\epsilon,\phi))}\right]$

Gradient $\nabla_\phi L = \nabla_\phi E_{r(\epsilon)}\left[\log \frac{\mathrm{p}(g(\epsilon,\phi),D)}{q_\phi(g(\epsilon,\phi))}\right] = E_{r(\epsilon)}\left[\nabla_\phi \log \frac{\mathrm{p}(g(\epsilon,\phi),D)}{q_\phi(g(\epsilon,\phi))}\right]$

Kingma and Welling. Auto-encoding variational bayes. ICLR 2014
Salimans and Knowles. Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. Bayesian Analysis 2013

REINFORCE Gradients

Log-derivative trick

$$\nabla q_\phi(\theta) = q_\phi(\theta)\, \nabla_\phi \log q_\phi(\theta)$$

ELBO $L = E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right]$

Gradient of the ELBO

$$\nabla_\phi L = \nabla_\phi E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right] = \int \nabla_\phi \{ q_\phi(\theta) \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)} \}\, d\theta$$

$$= \int \underline{\nabla_\phi q_\phi(\theta)} \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta + \int q_\phi(\theta)\, \nabla_\phi \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta$$

$$= \int \underline{q_\phi(\theta) \nabla_\phi \log q_\phi(\theta)}\ \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta - \underline{\int \nabla_\phi q_\phi(\theta)\, d\theta}$$

$$= \nabla_\phi \int q_\phi(\theta)\, d\theta = \nabla_\phi 1 = 0$$

Glynn (1990). Likelihood ratio gradient estimation for stochastic systems. Communications of the ACM, 33(10), 75–84.
Williams (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3-4), 229–256.
Fu (2006). Gradient estimation. Handbooks in Operations Research and Management Science, 13, 575–616.

REINFORCE Gradients

Log-derivative trick

$$\nabla q_\phi(\theta) = q_\phi(\theta)\, \nabla_\phi \log q_\phi(\theta)$$

ELBO $L = E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right]$

Gradient of the ELBO

$$\nabla_\phi L = \nabla_\phi E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right] = \int \nabla_\phi \{ q_\phi(\theta) \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)} \}\, d\theta$$

$$= \int \nabla_\phi q_\phi(\theta) \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta + \int q_\phi(\theta)\, \nabla_\phi \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta$$

$$= \int q_\phi(\theta) \nabla_\phi \log q_\phi(\theta)\ \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\, d\theta$$

$$= \mathrm{E}_{q_\phi(\theta)}\left[\underbrace{\nabla_\phi \log q_\phi(\theta)}_{\text{Score function}} \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right]$$

Glynn (1990). Likelihood ratio gradient estimation for stochastic systems. Communications of the ACM, 33(10), 75–84.
Williams (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3-4), 229–256.
Fu (2006). Gradient estimation. Handbooks in Operations Research and Management Science, 13, 575–616.

Monte Carlo Approximation

- To approximate: $E_{p(x)}[f(x)]$
- MC Approximation:
  1. Sample $x_1, x_2, \ldots, x_K \sim p(x)$
  2. Evaluate $f(x_i)$ for each sample
  3. Compute $E[f(x)] \approx \underline{\frac{1}{K}\sum_{i=1}^{K} f(x_i)}$

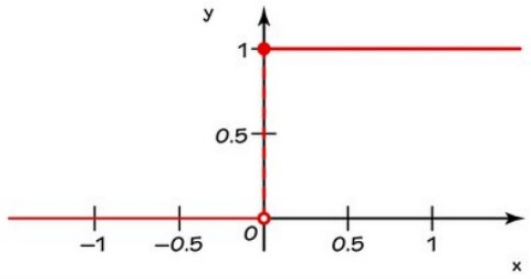
Unbiased Monte Carlo estimate

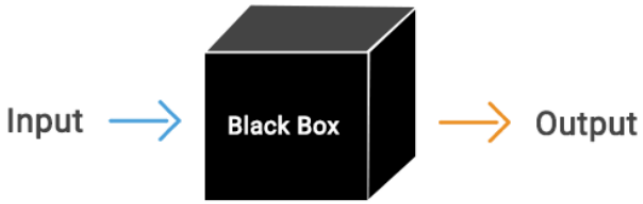
- Reparam. gradient: $E_{r(\epsilon)}\left[\nabla_\phi \log \frac{\mathrm{p}(g(\epsilon,\phi),D)}{q_\phi(g(\epsilon,\phi))}\right] \approx \frac{1}{K}\sum_{k=1}^{K} \nabla_\phi \log \frac{\mathrm{p}(g(\epsilon_k,\phi),D)}{q_\phi(g(\epsilon_k,\phi))}, \epsilon_k \sim r(\epsilon)$
- REINFORCE gradient: $E_{\mathrm{q}_\phi(\theta)}\left[\nabla_\phi \log q_\phi(\theta) \log \frac{\mathrm{p}(\theta,D)}{q_\phi(\theta)}\right] \approx \frac{1}{K}\sum_{k=1}^{K} \nabla_\phi \log q_\phi(\theta_k) \log \frac{\mathrm{p}(\theta_k,D)}{q_\phi(\theta_k)}, \theta_k \sim q_\phi(\theta)$

Variance Reduction Techniques in MCVI

- When non-differentiable, falls back to REINFORCE gradient

High variance!





- Solutions to high variance REINFORCE gradients:
  - Low variance unbiased estimators with control variates
  - Biased estimators to enable reparam. trick (potentially low variance)

For an incomplete list of variance reduced gradient estimators, see http://yingzhenli.net/home/en/?page_id=1262

Scalable variational inference: Summary

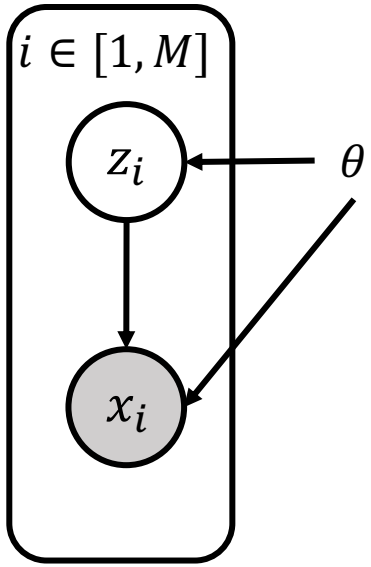
- Scalable variational inference:
  - Stochastic optimisation – using minibatches
  - Monte Carlo estimation – computing intractable expectations
  - Gradient estimation – Reparam. Trick or REINFORCE

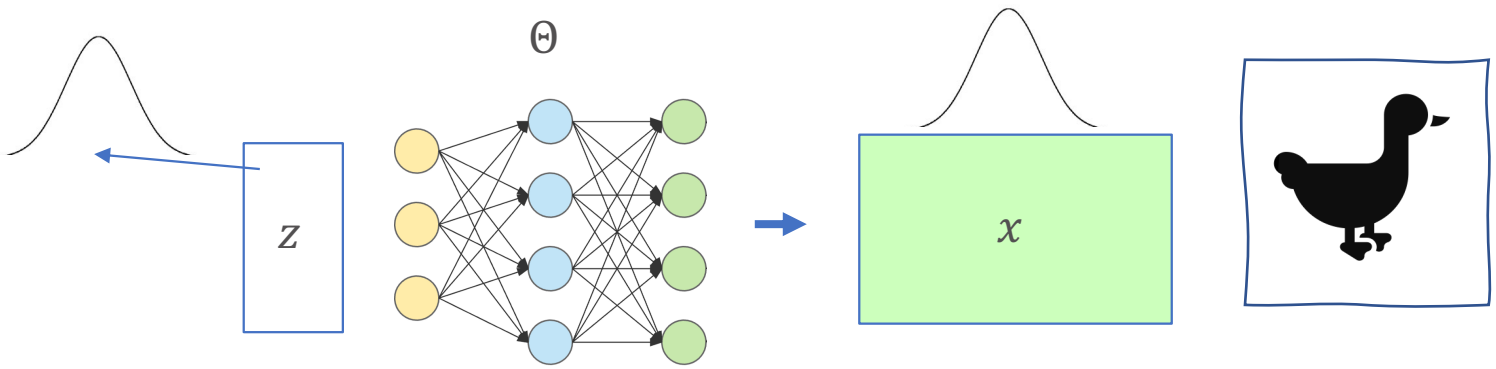
$$L = E_{q_\phi(\theta)}\left[\log\frac{\mathrm{p}(\theta, D)}{q_\phi(\theta)}\right] = E_{q_\phi(\theta)}[\log p(D|\theta) + \log\frac{p(\theta)}{q_\phi(\theta)}]$$

$$L \approx \frac{1}{K}\sum_{k=1}^{K}\frac{N}{M}\sum_{m=1}^{N}\log p(x_m|\theta_k) + \log\frac{p(\theta_k)}{q_\phi(\theta_k)}$$

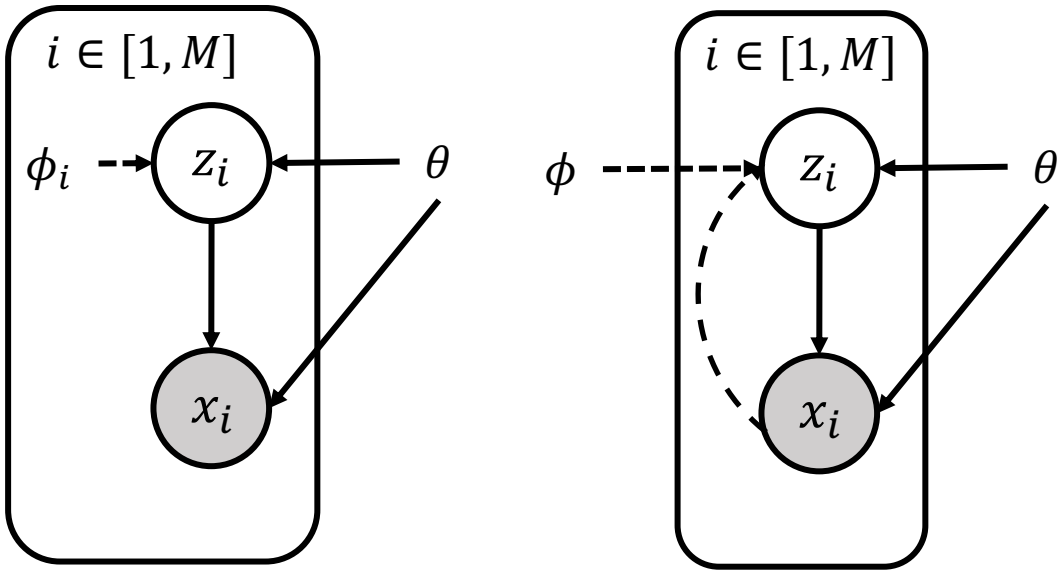
$$x_1, \dots, x_M \sim D$$
$$\theta_1, \dots, \theta_K \sim q_\phi(\theta)$$
$$\Leftrightarrow \theta_k = g_\phi(\epsilon_k), \epsilon_k \sim r(\epsilon)$$

Deep Latent Variable Model





Amortized Inference



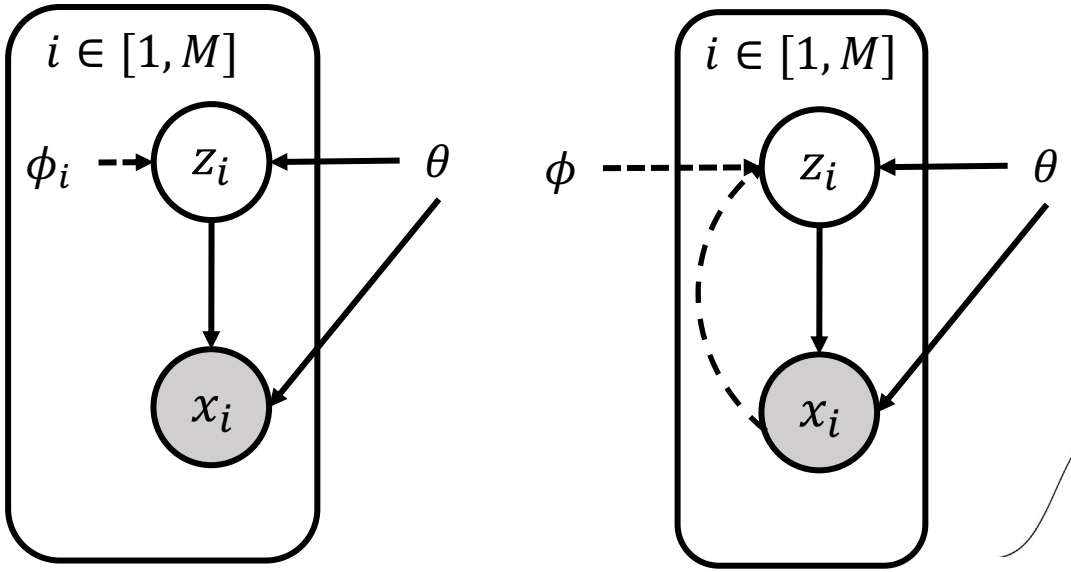
$$L_i = E_{q(z_i)}[\log \mathrm{p}_\theta(\boldsymbol{x_i}|\boldsymbol{z_i})] - KL[q(\boldsymbol{z}_i)||p(z_i)]$$

$$L = E_{q(z|x)}[\log \mathrm{p}_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL[q(\boldsymbol{z}|x)||p(z)]$$

ϕ parameter for amortized q distribution

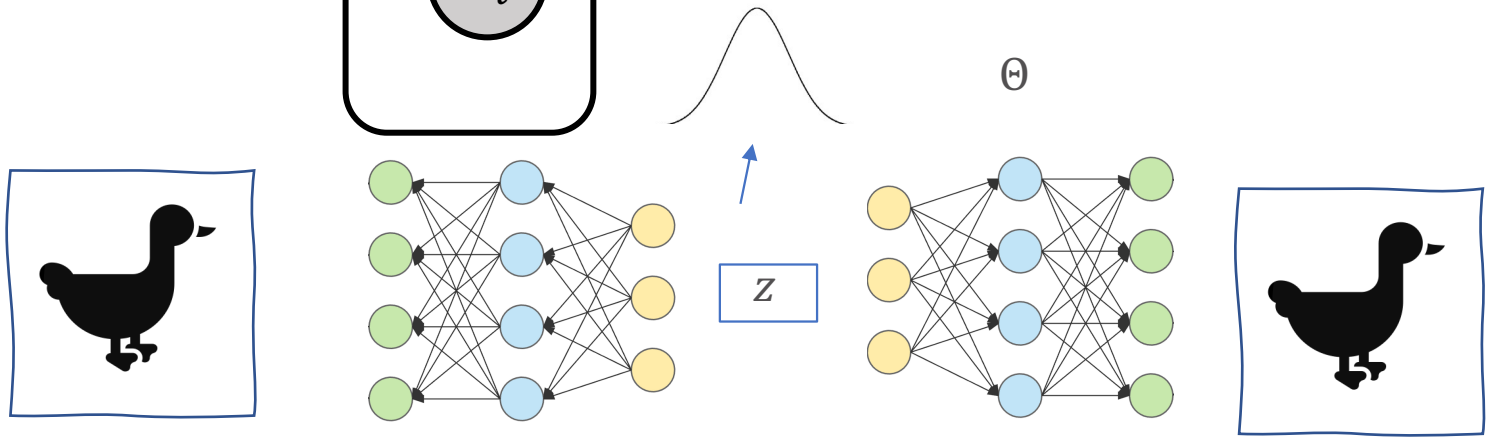
θ decoder parameter

Variational Auto-Encoders (VAE)



$$L_i = E_{q(z_i)}[\log \mathrm{p}_\theta(\boldsymbol{x_i}|\boldsymbol{z_i})] - KL[q(\boldsymbol{z}_i)||p(z_i)]$$

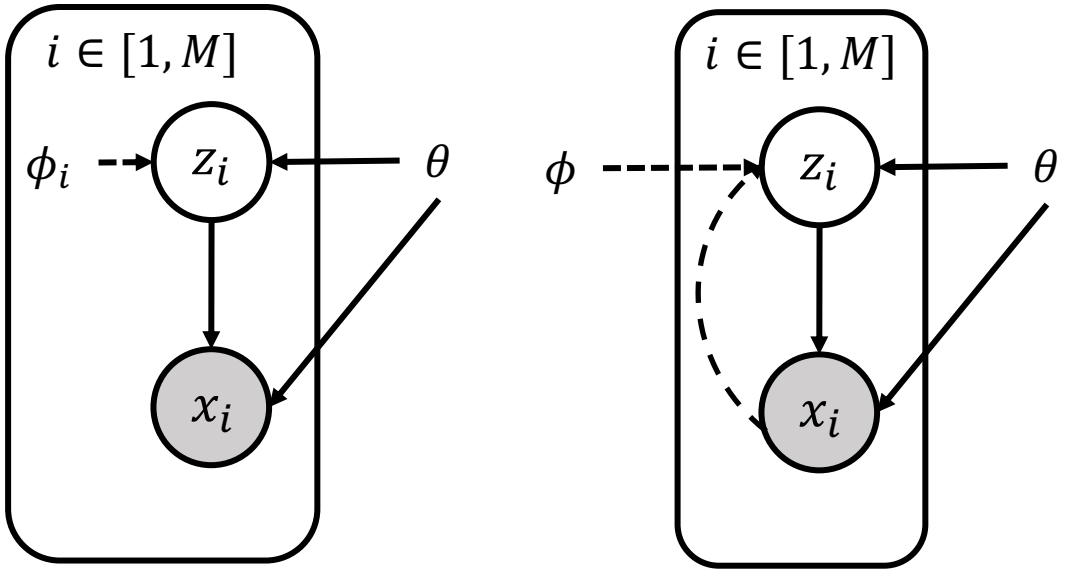
$$L = E_{q(z|x)}[\log \mathrm{p}_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL[q(\boldsymbol{z}|x)||p(z)]$$



Encoder/
inference network

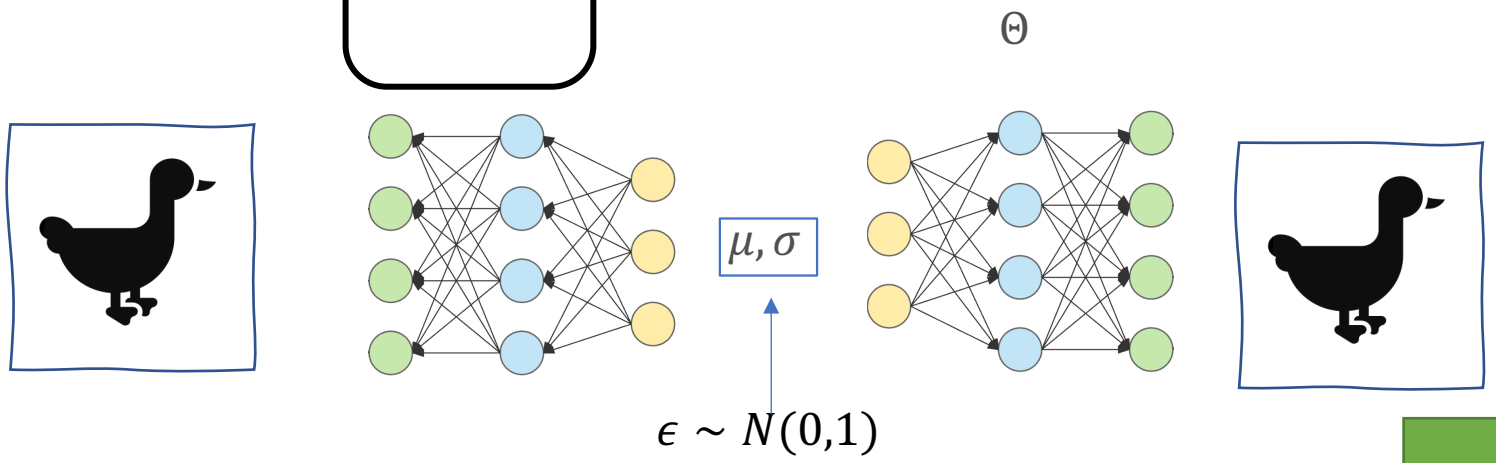
Decoder/
generator

Variational Auto-Encoders (VAE)



$$L_i = E_{q(z_i)}[\log \mathrm{p}_\theta(\boldsymbol{x_i}|\boldsymbol{z_i})] - KL[q(\boldsymbol{z}_i)||p(z_i)]$$

$$L = E_{q(z|x)}[\log \mathrm{p}_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL[q(\boldsymbol{z}|x)||p(z)]$$

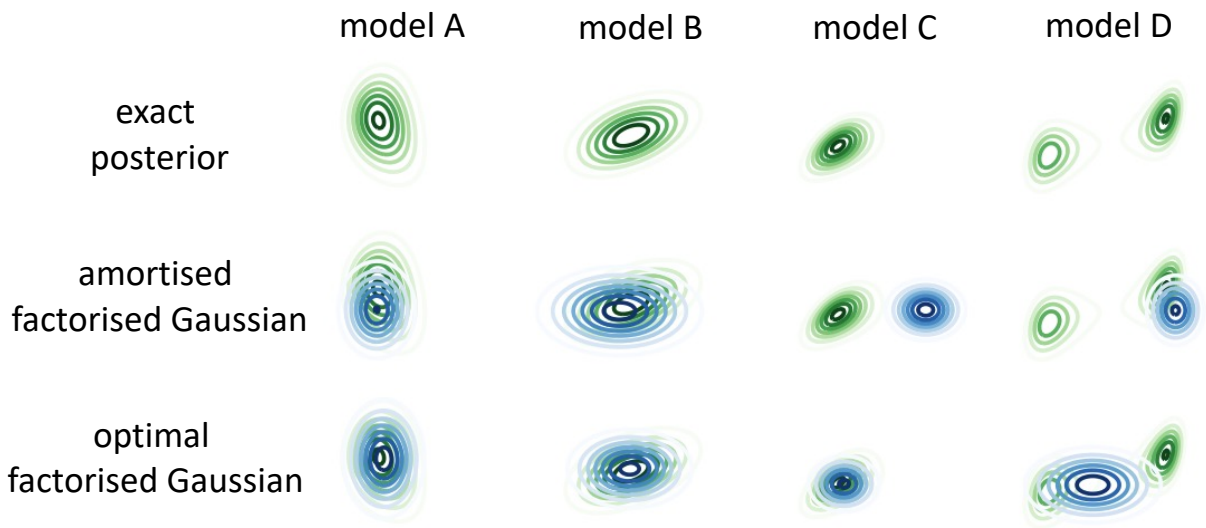


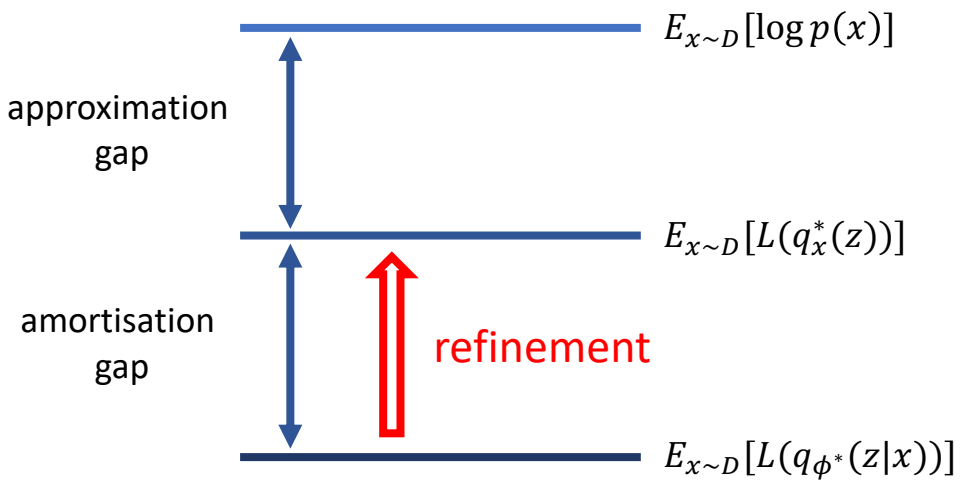
$$z \sim N(\mu, \sigma^2)$$
$$\epsilon \sim N(0,1), \theta = \mu + \sigma\epsilon$$

Kingma and Welling. Auto-encoding variational bayes. ICLR 2014.
Rezende et al. Stochastic backpropagation and approximate inference in deep generative models. ICML 2014.

Amortized Inference: Limitations

- Amortised approximate posteriors in practice are sub-optimal





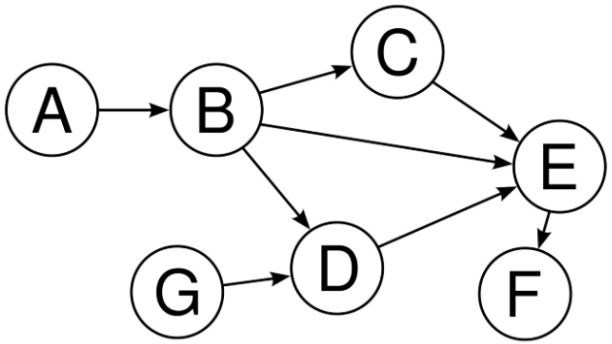
- The "refinement" idea:
  - Initialise $q(z|x) = N(z; \mu, \sigma^2)$ with the amortised solution $\mu \leftarrow \mu_\phi(x), \sigma \leftarrow \sigma_\phi(x)$
  - Then run T more VI gradient steps to update μ, σ

Cremer et al. Inference Suboptimality in Variational Autoencoders. ICML 2018
Marino et al. Iterative Amortized Inference. ICML 2018
Kim et al. Semi-Amortized Variational Autoencoders. ICML 2018

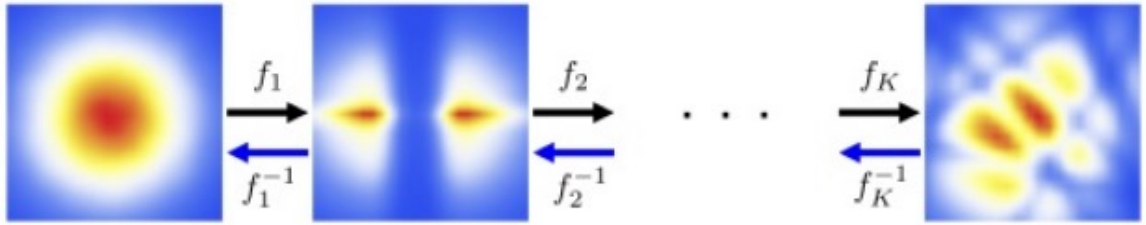
Part II: Advances

- **Approximate distribution design**
- **Optimization objective design**
- **Future directions**

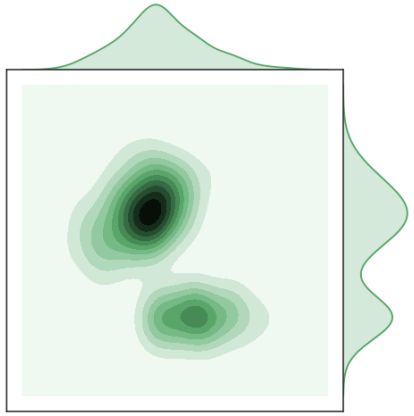
Designing q Distributions



Structured approximations



Normalizing flows

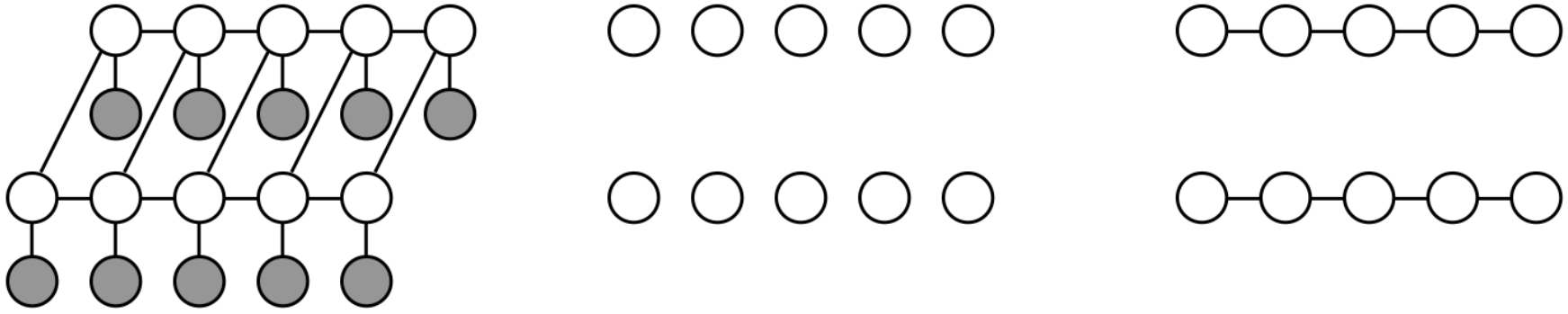


Auxiliary variables & mixture distributions

Implicit approximate posteriors

Structured Approximations

- introduce dependencies between random variables for q:



Hidden Markov Model | Mean-field approximation | Structured approximation

Exact posterior $p(z \mid x)$
$z_i \not\perp z_j \mid x$

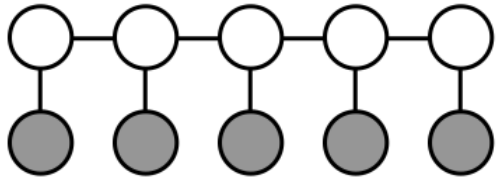
$$q(z) = \prod_i q(z_i)$$

$$q(z) = \prod_s q(z_s)$$
$$q(z_s) = q(\{z_i\}_{i \in s})$$

Main design question: the grouping and conditional dependency structure

Structured Approximations

- Auto-regressive distributions (as a specific dependency structure)



Hidden Markov Model

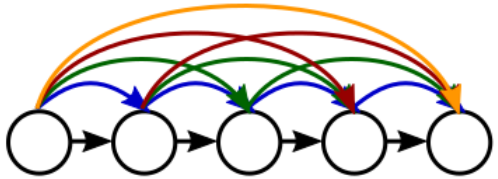
Exact posterior $p(z \mid x)$

$z_i \not\perp z_j \mid x$

Structured approximation

$q(z) = \prod_s q(z_s)$

$q(z_s) = q(\{z_i\}_{i \in s})$



Auto-regressive approximation

$$q(z) = \prod_i q(z_i \mid z_{<i})$$

$$q(z_1 \mid z_{<1}) = q(z_1)$$

Main design question:
the ordering of the
latent variables

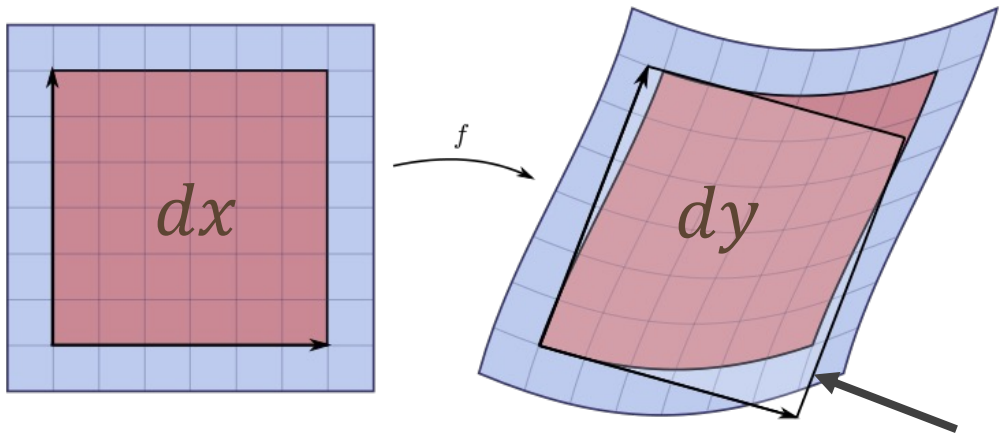
Normalizing Flows

- Change-of-variable formula:
  - x is a random variable with probability density function (PDF) $p_X(x)$
  - $\mathrm{y} = f(x)$ is an invertible mapping
  - The probability mass is preserved, and the PDF for $y = f(x)$ satisfies

$$\underbrace{p_Y(y)dy}_{\text{prob. mass of region around } y} = \underbrace{p_X(x)dx}_{\text{prob. mass of region around } x}$$

$$p_Y(y) = p_X(x)|\det(\frac{dx}{dy})|$$

$$p_X(x) = p_Y(y)|\det(\frac{dy}{dx})|$$



hypercube approximation
(exact when $\mathrm{vol}(dx) \to 0$)

Normalizing Flows

- Variational inference with Normalizing flow
  - Assume $q_0(z_0) = N(z_0; 0, I)$
  - Define $z = f_\phi(z_0)$ where $f_\phi(\cdot)$ is an invertible mapping parameterized by ϕ

$$q(z) = q_0(z_0) \left| \det\left(\frac{dz}{dz_0}\right) \right|^{-1} \qquad \text{with } z_0 = f_\phi^{-1}(z)$$

(change of variable: $q(z)dz = q_0(z_0)dz_0$)

  - Fit $q(z)$ to $p(x \mid z)$ with VI:

by def. of $q(z)$

$$\begin{aligned} L(q(z)) &= E_{q(z)}[\log p(x \mid z) + \log p(z) - \log q(z)] \\ &= E_{q(z)}\left[\log p(x, z) - \boxed{\log q_0(z_0 = f_\phi^{-1}(z)) \left| \det\left(\frac{dz}{dz_0}\right) \right|^{-1}}\right] \\ &= \underline{E_{q_0(z_0)}\left[\log p(x, f_\phi(z_0)) - \log q_0(z_0) + \log \left| \det\left(\frac{df_\phi}{dz_0}\right) \right|\right]} \end{aligned}$$

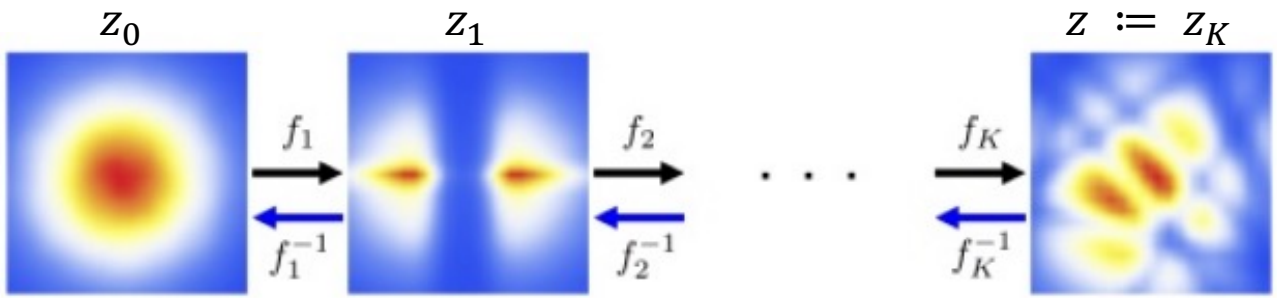
reparam. trick:
$z \sim q(z) \Leftrightarrow z_0 \sim q_0(z_0), z = f_\phi(z_0)$

  - Computing ELBO requires $\log \left| \det\left(\frac{df_\phi}{dz_0}\right) \right|$

Rezende and Mohamed. Variational Inference with Normalizing Flows. ICML 2015

Normalizing Flows

- Variational inference with Normalizing flow
  - Idea: define f_ϕ such that $\log|\det\left(\frac{df_\phi}{dz_0}\right)|$ is easy to compute!
    - Chain simple invertible mappings together to make a flexible mapping



$$f_\phi = f_K \circ f_{K-1} \circ \cdots \circ f_1,\ f_k(\cdot) \coloneqq f_{\phi_k}(\cdot),\ \phi = \{\phi_k\}_{k=1}^K$$

    - For each simple mapping, hopefully the Jacobian log-determinant is easy to compute

$$\Rightarrow\ \log|\det\left(\frac{df_\phi}{dz_0}\right)| = \sum_{k=1}^{K} \log|\det(\frac{dz_k}{dz_{k-1}})|$$

Rezende and Mohamed. Variational Inference with Normalizing Flows. ICML 2015

Auxiliary Variables & Mixture Distributions

- Construct $q(\theta)$ as a (hierarchical) mixture distribution

$$q(\theta) = \int q(\theta \mid a)\, q(a)\, da$$

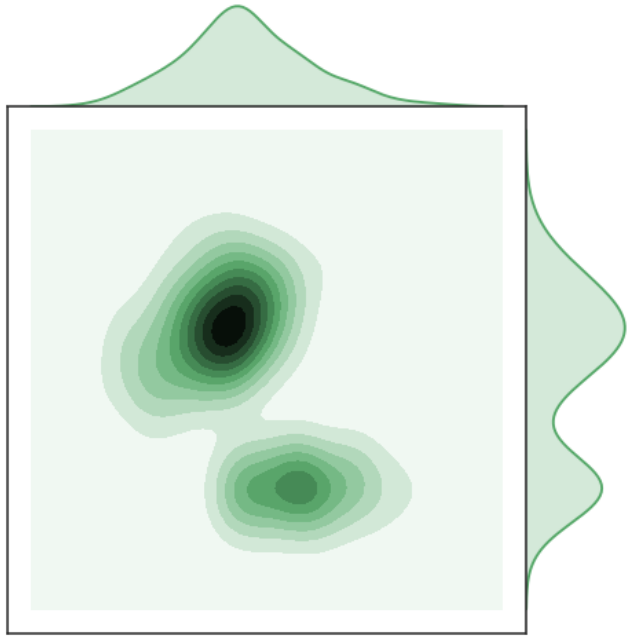
  - a is the auxiliary variable used to enrich the approximate posterior

  - Example: Mixture of Gaussians

$$a \sim q(a) = Categorical(\pi_1, \dots, \pi_K)$$

$$\theta \sim q(\theta \mid a) = N(\theta; m_a, \Sigma_a)$$

Can be very flexible with many components!



Auxiliary Variables & Mixture Distributions

- Construct $q(\theta)$ as a (hierarchical) mixture distribution

$$q(\theta) = \int q(\theta \mid a)\, q(a)\, da$$

  - a is the auxiliary variable used to enrich the approximate posterior
  - Now the variational lower-bound becomes intractable:

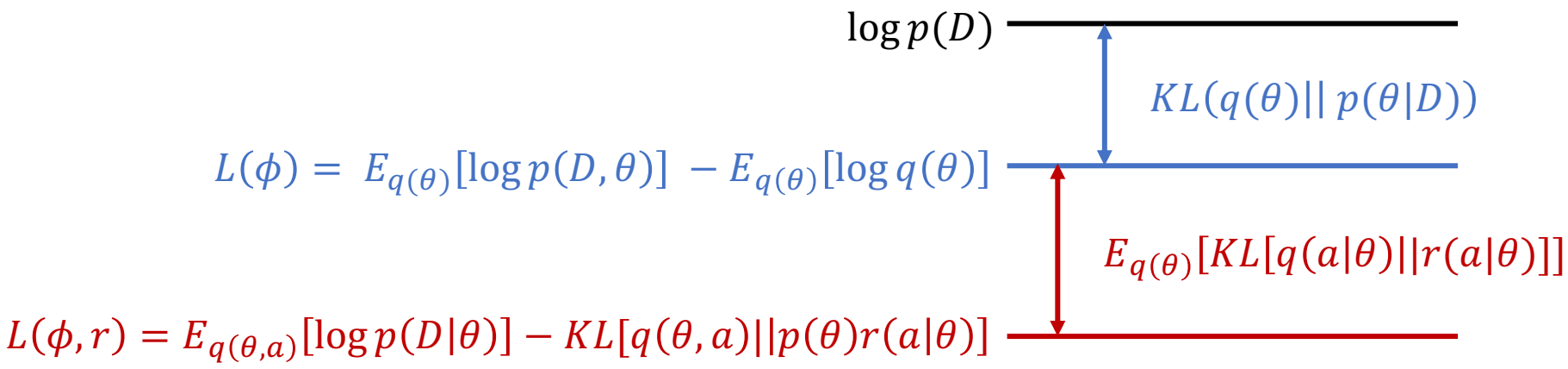
$$L(\phi) = \underline{E_{q(\theta)}[\log p(D, \theta)]} - E_{q(\theta)}[\underline{\log q(\theta)}]$$

Estimated by Monte Carlo:
$a_k \sim q(a), \theta_k \sim q(\theta \mid a_k)$

Intractable density
$q(\theta) = \int q(\theta|a) q(a)\, da$

Auxiliary Variables & Mixture Distributions

- Solution: introducing an auxiliary variational lower-bound $L(\phi, r)$ with an auxiliary distribution $r(a|\theta)$:

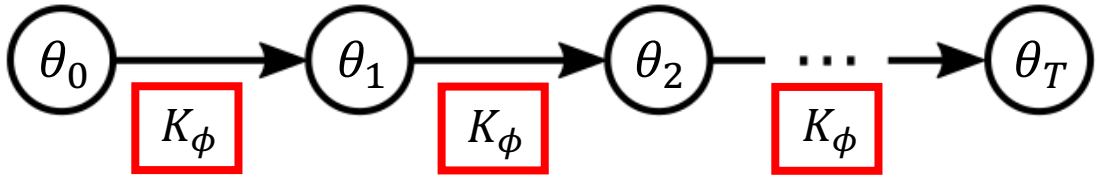


- Optimize $r(a|\theta)$ to close the gap!
- $L(\phi, r)$ estimated by Monte Carlo: $a_k \sim q(a), \theta_k \sim q(\theta \mid a_k)$

Agakov and Barber. An Auxiliary Variational Method. ICONIP 2004
Salimans et al. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. ICML 2015
Ranganath et al. Hierarchical Variational Models. ICML 2016

Auxiliary Variables & Mixture Distributions

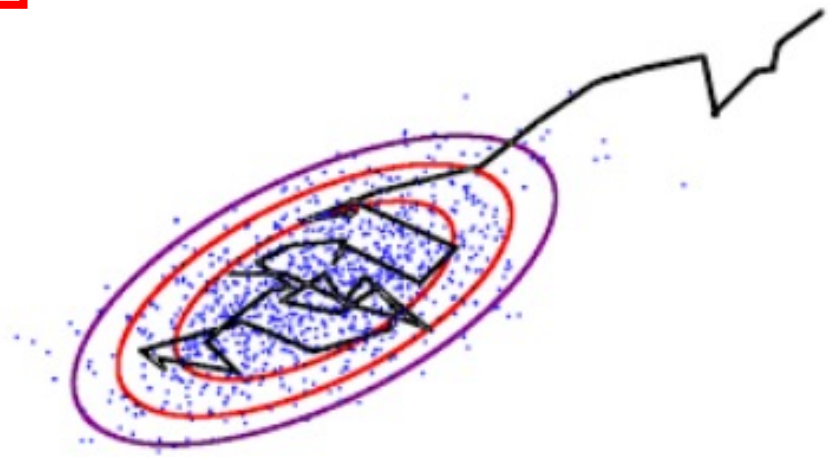
- Hierarchical mixture distributions for $q(\theta, a)$
  - VI-MCMC hybrid: build $q(\theta)$ with a Markov Chain:



learn the transition kernel with VI:

$$\theta := \theta^T, a = \{\theta^{0:T-1}\}$$

$$q(\theta^T) = \int q_0(\theta^0) \prod_{t=1}^{T} K_\phi\,(\theta^t \,|\theta^{t-1}) d\theta^{0:T-1}$$



Salimans et al. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. ICML 2015
Huang et al. Improving Explorability in Variational Inference with Annealed Variational Objectives. NeurIPS 2018

Auxiliary Variables: Continuous-Time Limit

- SDE for defining $q_\phi(\theta)$:

$$d\theta = f_\phi(\theta, t)dt + \sigma_\phi(\theta, t)dW_t, \theta_0 \sim q_0(\theta), q_\phi(\theta) \coloneqq q_T(\theta)$$

Design choices:
- SDEs in "complete SG-MCMC framework"
- Annealed Importance Sampling in continuous-time limit

Fitting objectives:
- ELBO with marginal $q_T(\theta)$ (by solving the corresponding prob. ODE)
- Auxiliary ELBO with discretisation + r distribution constructed by backward SDE

Ma et al. A Complete Recipe for Stochastic Gradient MCMC. NIPS 2015
Doucet et al. Score-Based Diffusion meets Annealed Importance Sampling. NeurIPS 2022
Geffner and Domke. Langevin Diffusion Variational Inference. ICML 2023

Implicit Approximate Posteriors

- Two quantities computed in (approximate) Bayesian inference:

approximate Bayesian predictive

$$p(y^*|x^*, D) \approx E_{q(\theta)}[p(y^*|x^*, \theta)]$$

$$\approx \frac{1}{K}\sum_k^K p(y^*|x^*, \theta_k),\ \theta_k \sim q(\theta)$$

approximate posterior moments

$$E_{q(\theta)}[F(\theta)]$$

$$\approx \frac{1}{K}\sum_k^K F(\theta_k),\ \theta_k \sim q(\theta)$$

Computed with Monte Carlo estimates

Only require fast sampling from q!
(no need for analytic form of the q distribution)

implicit distributions

Mohamed and Lakshminarayanan. Learning in Implicit Generative Models. arXiv 2016
Li and Liu. Wild Variational Inference. AABI 2016
Huszár. Variational Inference using Implicit Distributions. arXiv 2017

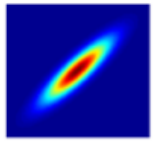
Implicit Approximate Posteriors

$$L(\phi) = \underline{E_{q(\theta)}[\log p(D|\theta)]} - E_{q(\theta)}[\log \boxed{\frac{p(\theta)}{q(\theta)}}]$$

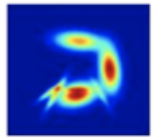
estimated by Monte Carlo

Approximated by using a discriminator (AVB):

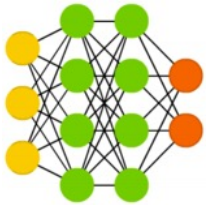
$\theta \sim p(\theta)$



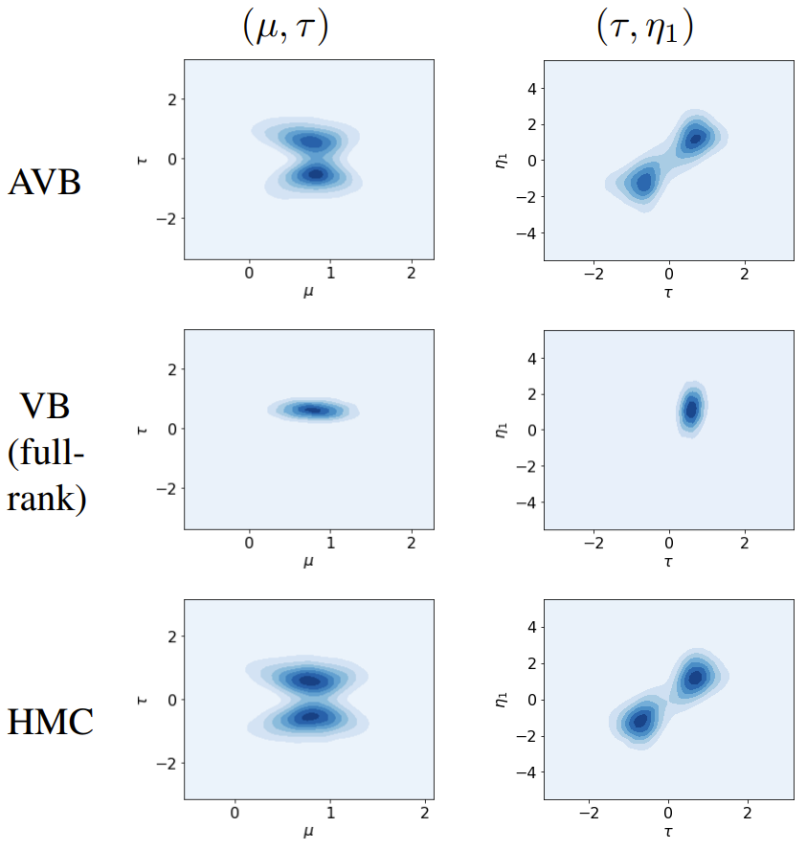
$\theta \sim q(\theta)$



discriminator



θ sampled from p or q?

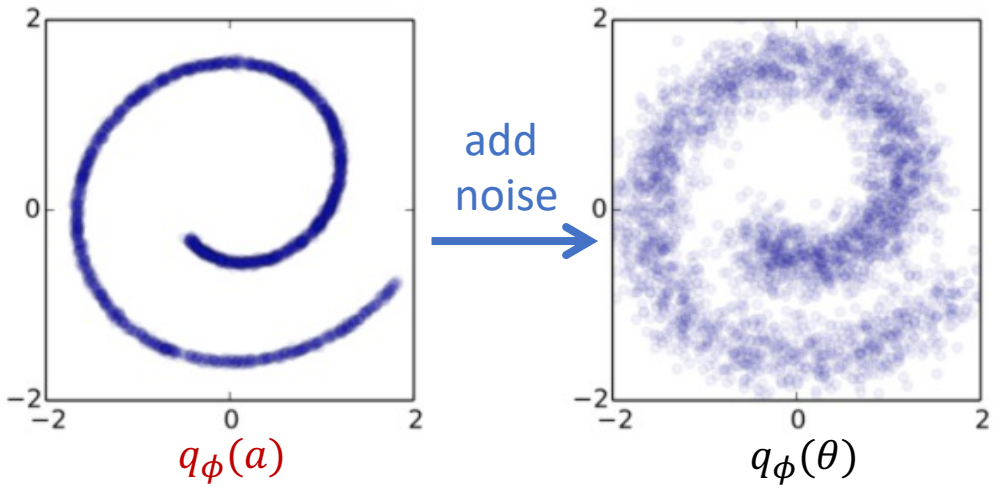


Mescheder et al. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. ICML 2017
Tran et al. Hierarchical Implicit Models and Likelihood-Free Variational Inference. NeurIPS 2017
Li and Turner. Gradient Estimators for Implicit Models. ICLR 2018
Yin and Zhou. Semi-Implicit Variational Inference. ICML 2018

Semi-Implicit Approximate Posteriors

- A combination of "Implicit Posterior" & "Auxiliary Variables":

$$q_\phi(\theta) = \int \underbrace{q_\phi(\theta|a)}_{\text{tractable}} \underbrace{q_\phi(a)}_{\text{implicit}} da,$$



$$L(\phi) = E_{q_\phi(\theta)}[\log p(D|\theta)] - E_{q_\phi(a)}[KL[q_\phi(\theta|a)\|p(\theta)]]$$

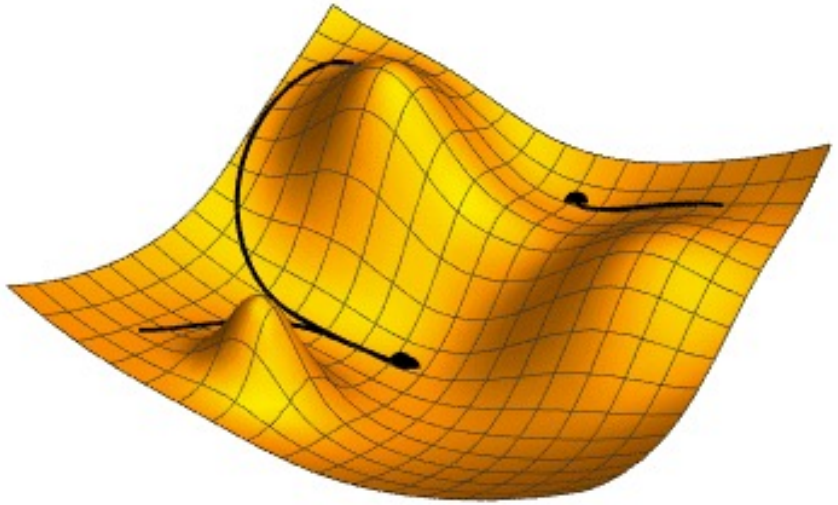
(equiv. to setting $r(a|\theta) = q(a)$)

- Direct optimisation of $L(\phi)$ leads to point mass for $q_\phi(a)$
- Solution: regularisers and/or alternative objectives to ensure $q_\phi(a)$ has non-zero entropy

Yin et.al. Semi-Implicit Variational Inference. ICML 2018
Yu et.al. Semi-Implicit Variational Inference via Score Matching. ICLR 2023

Objective Functions

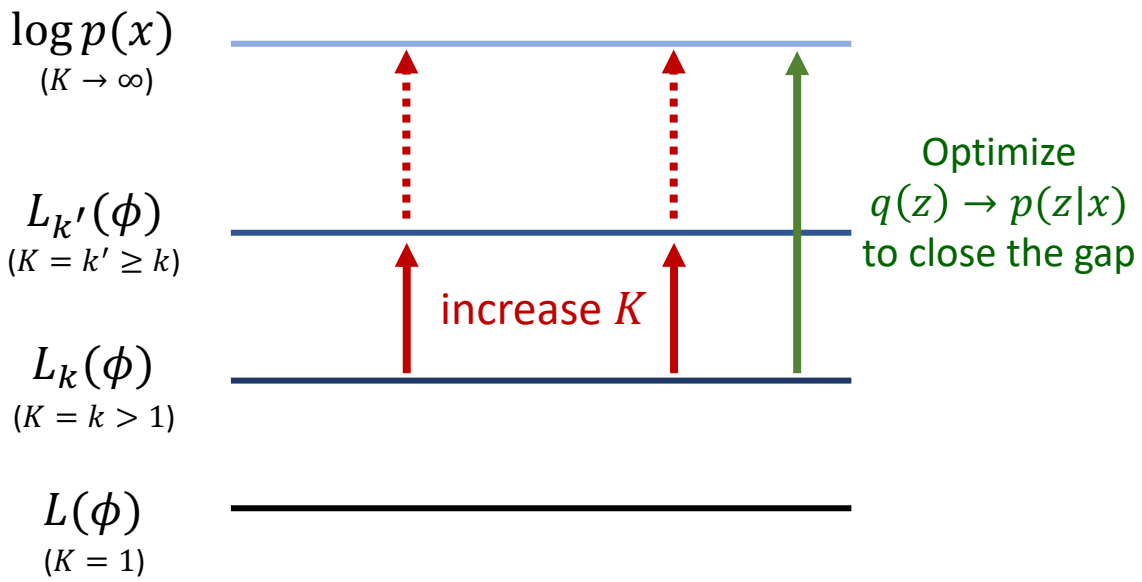
For fitting the approximate posterior



Improved Monte Carlo Bounds

- Importance weighted auto-encoder (IWAE) bound:

$$L_K(\phi) = E_{z_1,\ldots,z_K \sim q(z)}\left[\log \underbrace{\frac{1}{K}\sum_{k=1}^{K}\frac{p(x,z_k)}{q(z_k)}}_{\text{Importance sampling estimate of } p(x)}\right]$$



Burda et al. Importance Weighted Auto-encoders. ICLR 2016
Naesseth et al. Variational Sequential Monte Carlo. AISTATS 2018
Maddison et al. Filtering Variational Objectives. NeurIPS 2017
Le et al. Auto-encoding Sequential Monte Carlo. ICLR 2018
Masrani et al. The Thermodynamic Variational Objective. NeurIPS 2019

Improved Monte Carlo Bounds

- Constructing lower-bounds from an estimator R of the marginal:

$$E_{q(h)}[R(h,x)] = p(x) \quad \Rightarrow \quad \underline{E_{q(h)}[\log R(h,x)] \leq \log p(x)}$$

Jensen's inequality

  - Variational lower-bound: $h = z,\ R(z,x) = \frac{p(x,z)}{q(z)}$
  - IWAE bound: $h = (z_1, \dots, z_K), R(h,x) = \frac{1}{K}\sum_{k=1}^{K} \frac{p(x,z_k)}{q(z_k)}$

- Caveat: better MC estimator R doesn't necessarily lead to a better q
  - Example: make $K \to \infty$, then an OKish Gaussian q proposal can give very tight IWAE bound

Domke and Sheldon. Divide and Couple: Using Monte Carlo Variational Objectives for Posterior Approximation. NeurIPS 2019
Rainforth et al. Tighter Variational Bounds are Not Necessarily Better. ICML 2018

(Rényi) α-Divergence

$\alpha > 0, \alpha \neq 1$

$$D_\alpha[p||q] = \frac{1}{\alpha - 1} \log \int p(\theta)^\alpha q^{1-\alpha} d\,\theta$$

$\alpha = 1$

$$D_1[p||q] = \lim_{\alpha \to 1} D_{\alpha(p|q)} = KL(p||q)$$

VI with α-Divergence

ELBO

$$L = E_{\theta \sim q_\phi}\left[\log \frac{\mathrm{p}(\mathrm{D},\theta)}{q_\phi(\theta)}\right] = \log p(D) - \boxed{KL[\, q_\phi || p \,]}$$

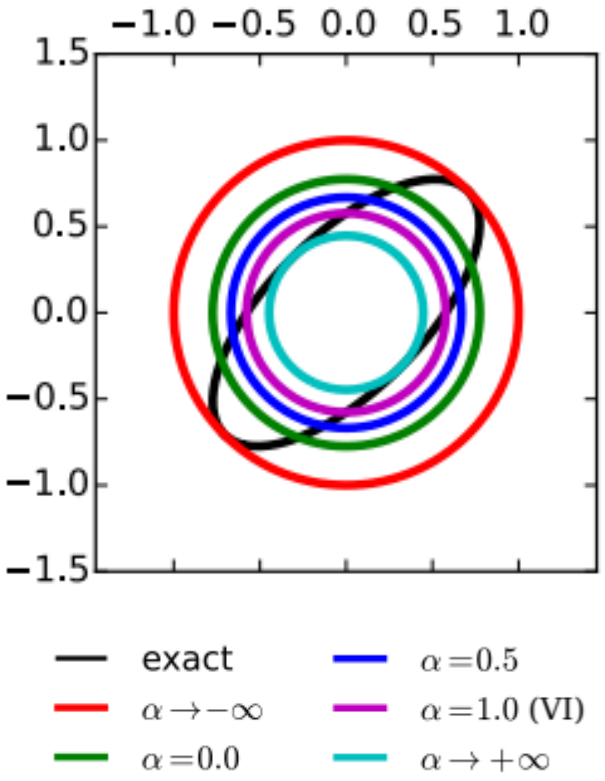
Variational Rényi bound:

$$L_\alpha = \frac{1}{1-\alpha} E_{\theta \sim q_\phi}\left[\left(\log \frac{\mathrm{p}(\mathrm{D},\theta)}{q_\phi(\theta)}\right)^{1-\alpha}\right] = \log p(D) - \boxed{D_\alpha[\, q_\phi || p \,]}$$

$$\lim_{\alpha \to 1} L_\alpha = L$$

Li and Turner. Rényi Divergence Variational Inference. NeurIPS 2016
Dieng et al. Variational Inference via χ-Upper Bound Minimization. NeurIPS 2017
Minka, Tom. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

Does It Work?



$$L_\alpha = \frac{1}{1-\alpha} E_{\theta \sim q_\phi}\left[\left(\log \frac{\mathrm{p}(\mathrm{D}, \theta)}{q_\phi(\theta)}\right)^{1-\alpha}\right]$$

Bayesian linear regression example:
Approximating the posterior with mean-field q

Li and Turner. Rényi Divergence Variational Inference. NeurIPS 2016
Dieng et al. Variational Inference via χ-Upper Bound Minimization. NeurIPS 2017
Minka, Tom. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

F-Divergence

$$D_f[p||q_\phi] = E_{\theta\sim q_\phi}[f\left(\frac{p(\theta)}{q_\phi(\theta)}\right) - f(1)]$$

$$f(t) = -\log t \longrightarrow KL(q||p)$$

$$f(t) = t\log t \longrightarrow KL(p||q)$$

$$f(t) = \frac{t^\alpha}{\alpha(\alpha-1)} \longrightarrow D_\alpha(p||q)$$

Wang et.al. Variational Inference with Tail-adaptive f-Divergence. NeurIPS 2018
Wan et.al. f-Divergence Variational Inference. NeurIPS 2020

Integral Probability Metric (IPM)

- Using a test function to describe difference:

$$D[q(z), p(z|x)] = \sup_{f \in F} |E_{q(z)}[f(z)] - E_{p(z|x)}[f(z)]|$$

  - Stein discrepancy: only requires $z \sim q(z)$ and $\nabla_z \log p(z|x) = \nabla_z \log p(z, x)$

$$S[q(z), p(z|x)] = \sup_{f \in F_q} |E_{q(z)}[\nabla_z \log p(z, x)^\top f(z) + \nabla_z^\top f(z)]|$$

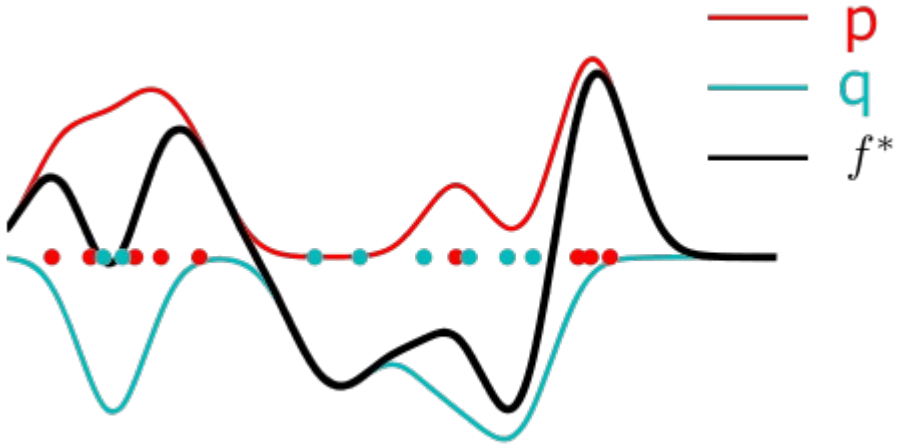


Figure adapted, source: Danica Sutherland

Gorham and Mackey. Measuring Sample Quality with Stein's Method. NeurIPS 2015
Ranganath et al. Operator Variational Inference. NeurIPS 2016
Liu and Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. NeurIPS 2016

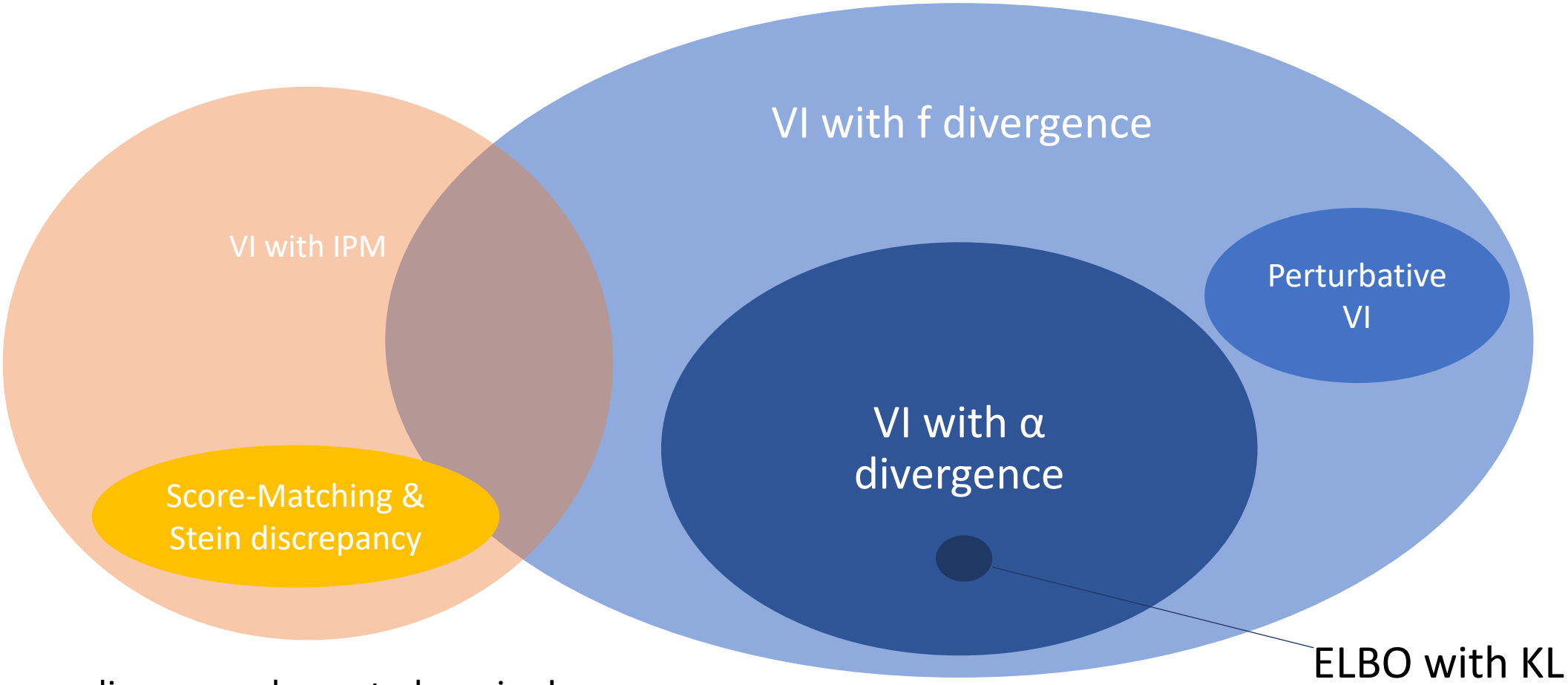
Score Matching

- Use Fisher divergence for q distribution fitting:

$$L(\phi) = E_{q_\phi(\theta)}[\|\nabla_\theta \log q(\theta) - \nabla_\theta \log p(D,\theta)\|_2^2]$$

Zhang et al. (2018). Variational Hamiltonian Monte Carlo via Score Matching. Bayesian Analysis 13(2) 485 - 506
Elkhalil et al. Fisher Auto-Encoders. AISTATS 2021.
LK Wenliang. On the failure of variational score matching for VAE models. arXiv:2210.13390

How to Choose the Inference Algorithm?

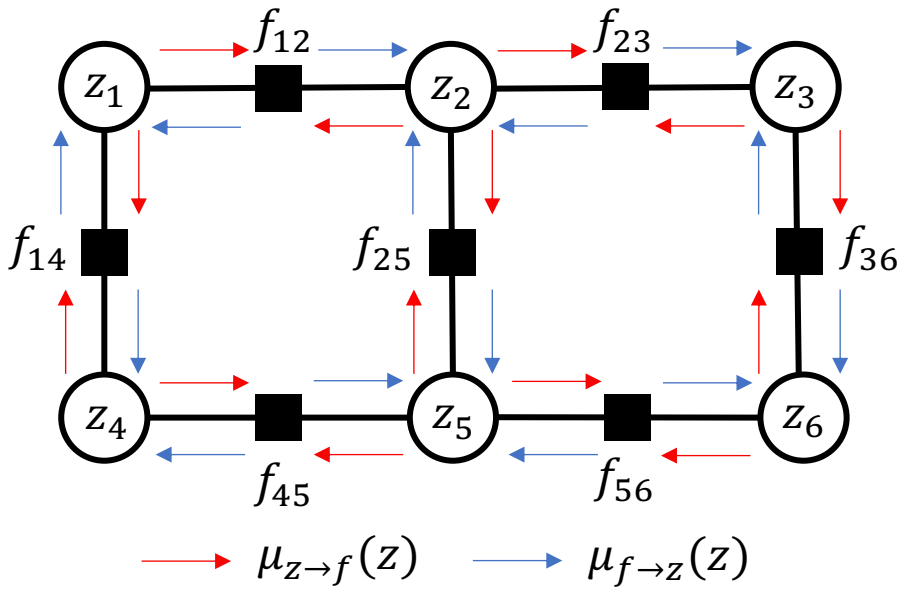


Choose divergence by meta-learning!

Zhang et al. Meta-Learning for Variational Inference. AABI 2019

Free-energy as an Objective

- Bethe free-energy & message passing:



- Both q and the inference algorithm are defined by the *factor graph*
- Optimal q achieved at the fixed point of the *Bethe free energy*

Wainwright and Jordan. Graphical Models, Exponential Families, and Variational Inference. 2008.
Li and Turner. A Unifying Approximate Inference Framework from Variational Free Energy Relaxation. AABI 2016

Landscape of Advances

q design

e.g. mean-field: $q(\theta) = \prod_i q(\theta_i)$

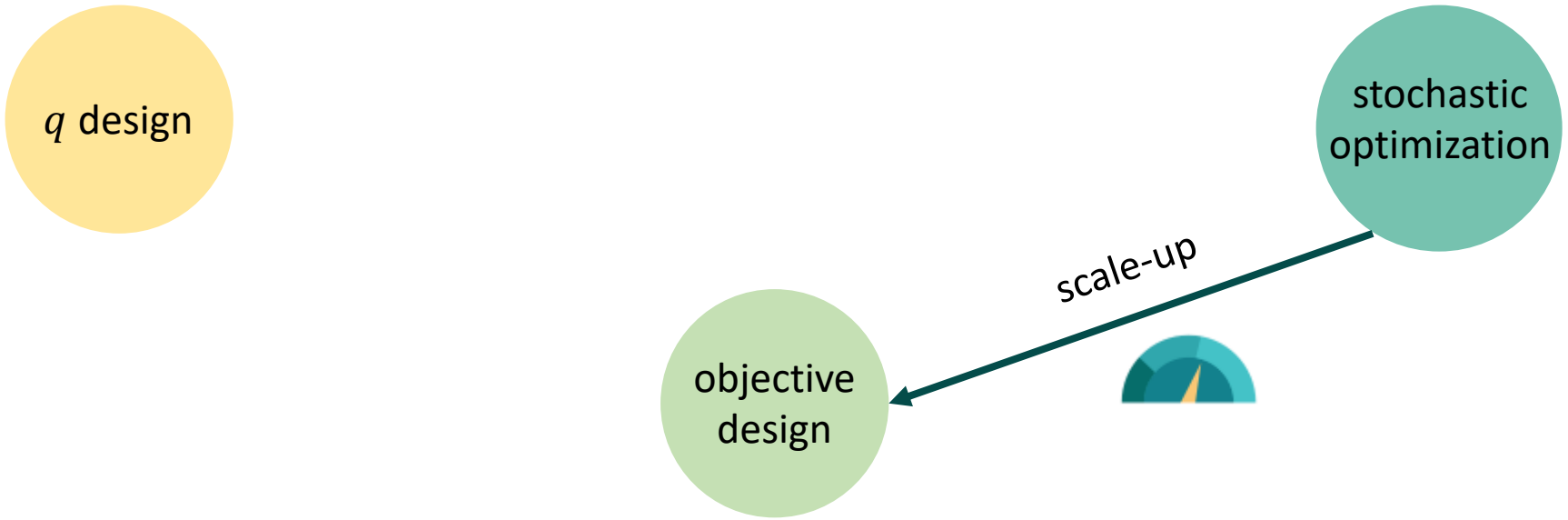
objective design

variational lower-bound:
$L(\phi) = E_{q(\theta)}[\log p(D|\theta)] - KL[q(\theta)||p(\theta)]$

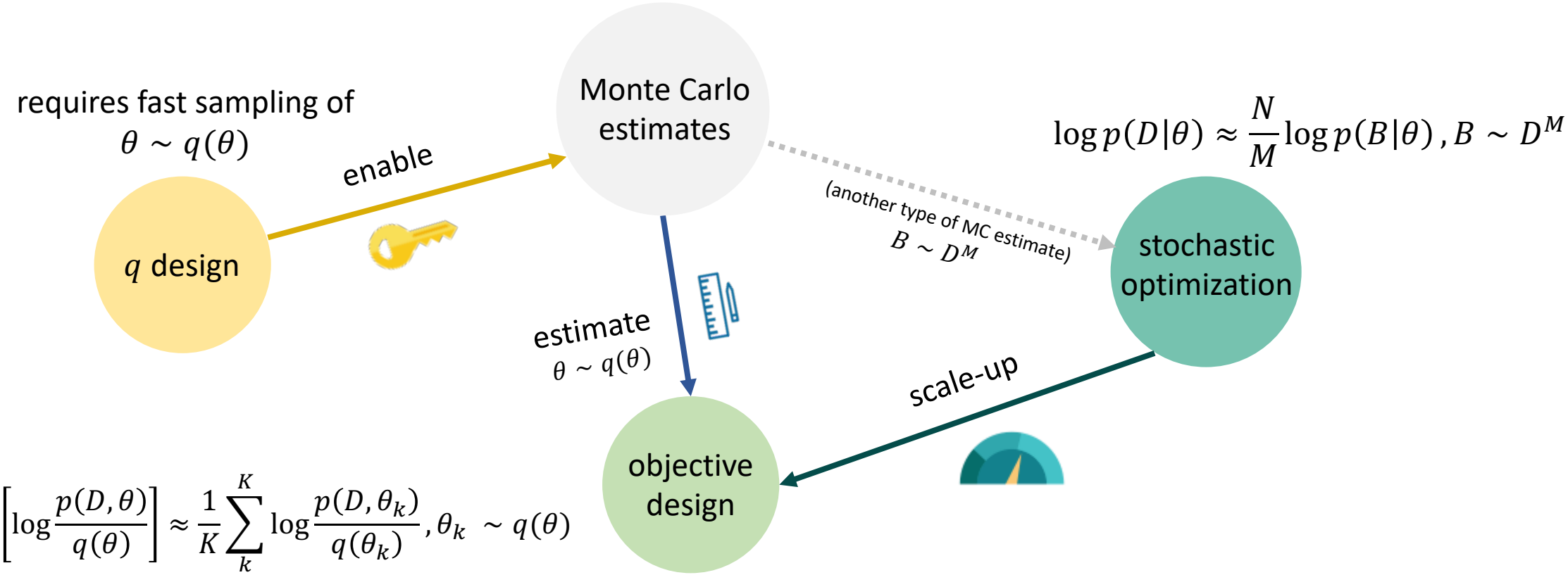
Landscape of Advances

mini-batch training:

$$\log p(D|\theta) \approx \frac{N}{M} \log p(B|\theta)\,, B \sim D^M$$



Landscape of Advances



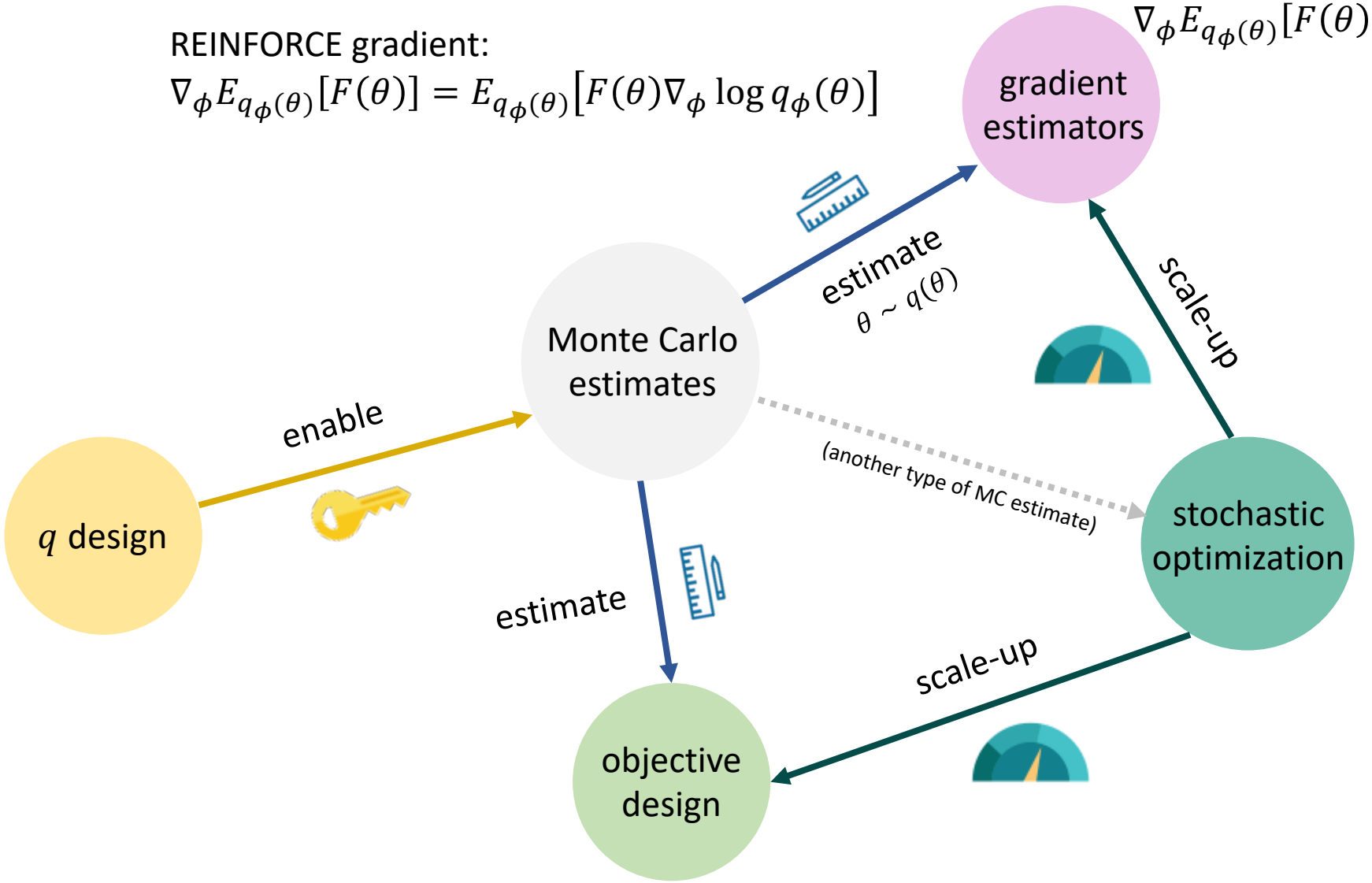
Landscape of Advances

REINFORCE gradient:

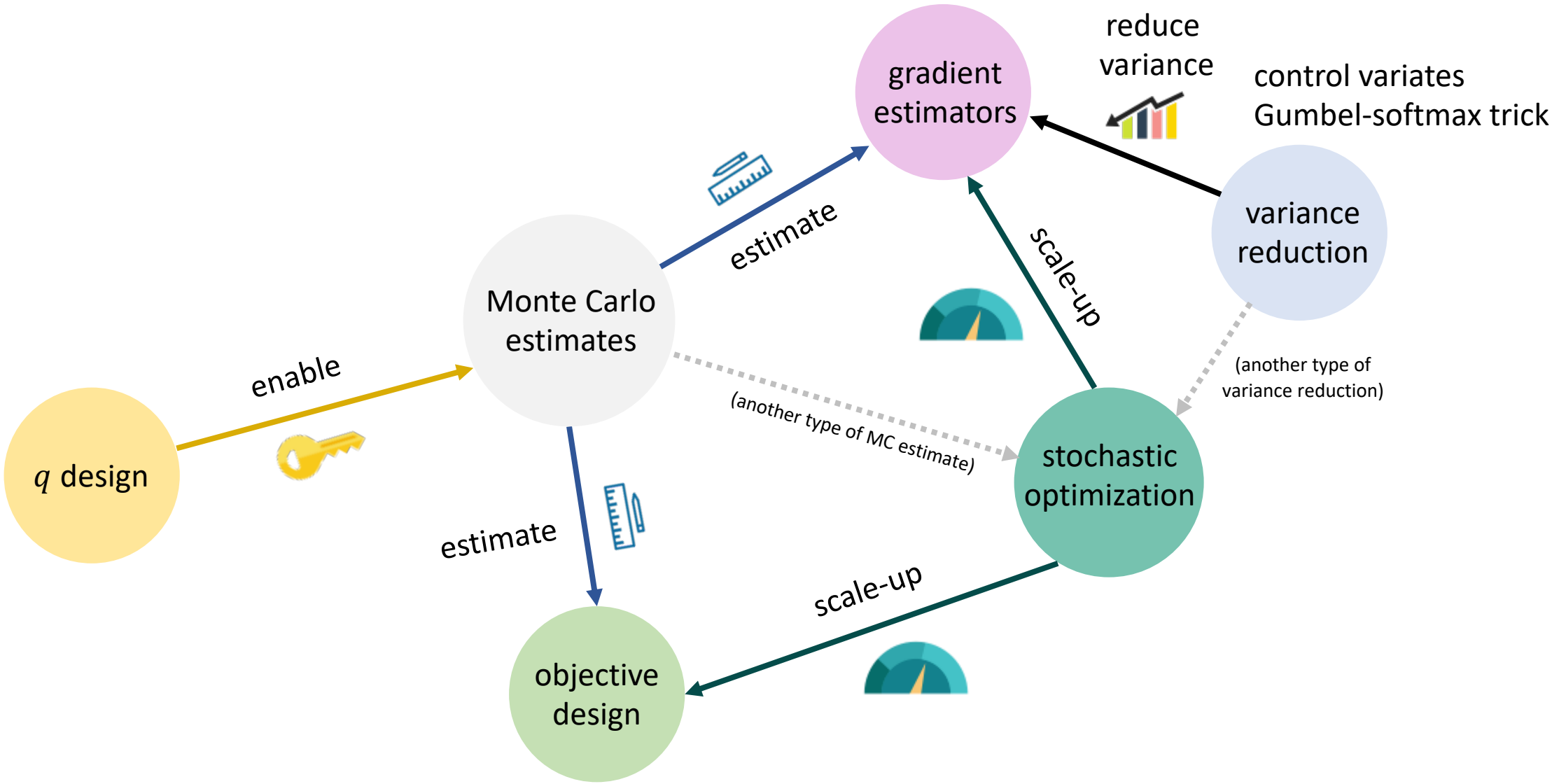
$$\nabla_\phi E_{q_\phi(\theta)}[F(\theta)] = E_{q_\phi(\theta)}\left[F(\theta)\nabla_\phi \log q_\phi(\theta)\right]$$

reparam. trick:

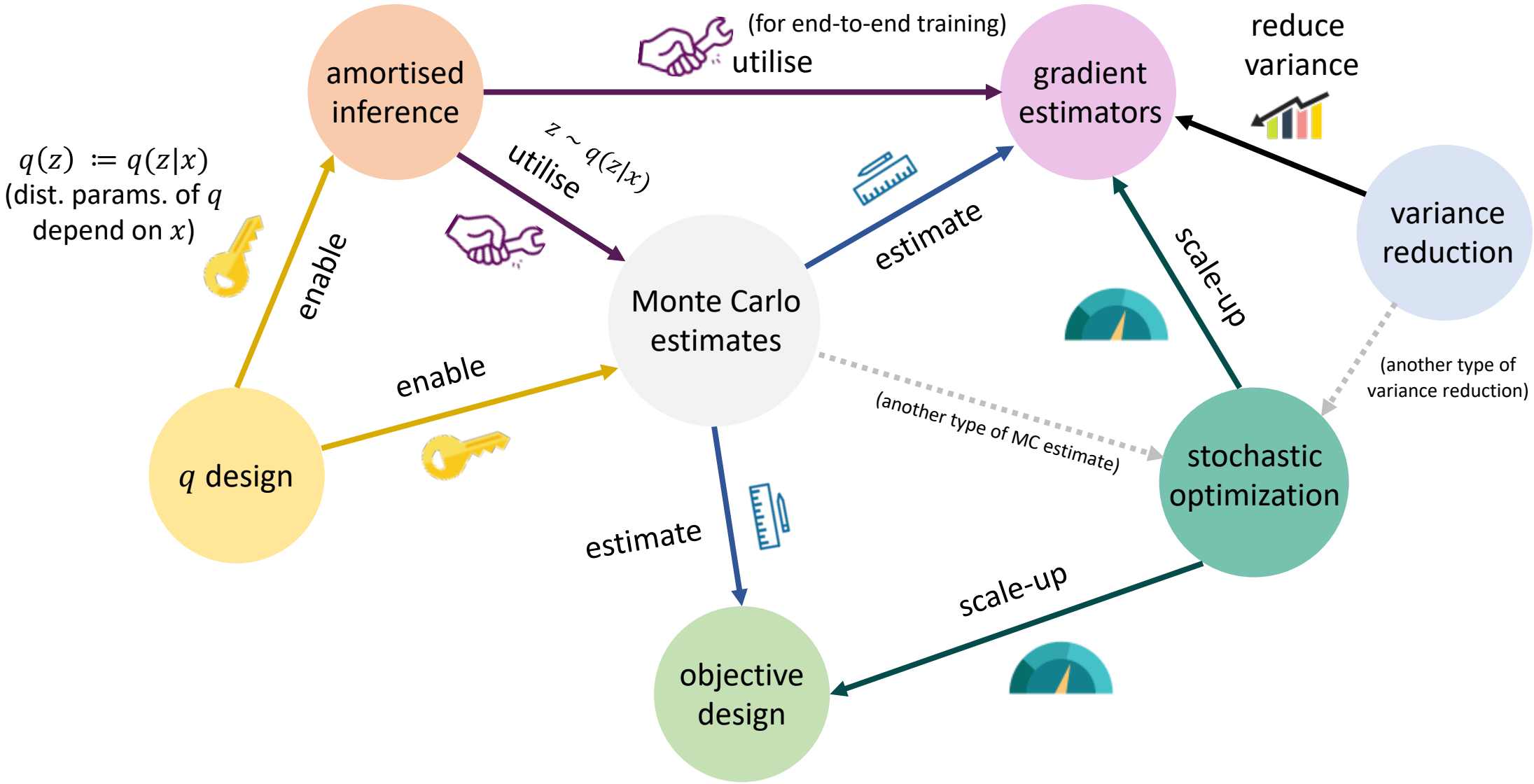
$$\nabla_\phi E_{q_\phi(\theta)}[F(\theta)] = E_{p(\epsilon)}[\nabla_\phi F(g_\phi(\epsilon))]$$



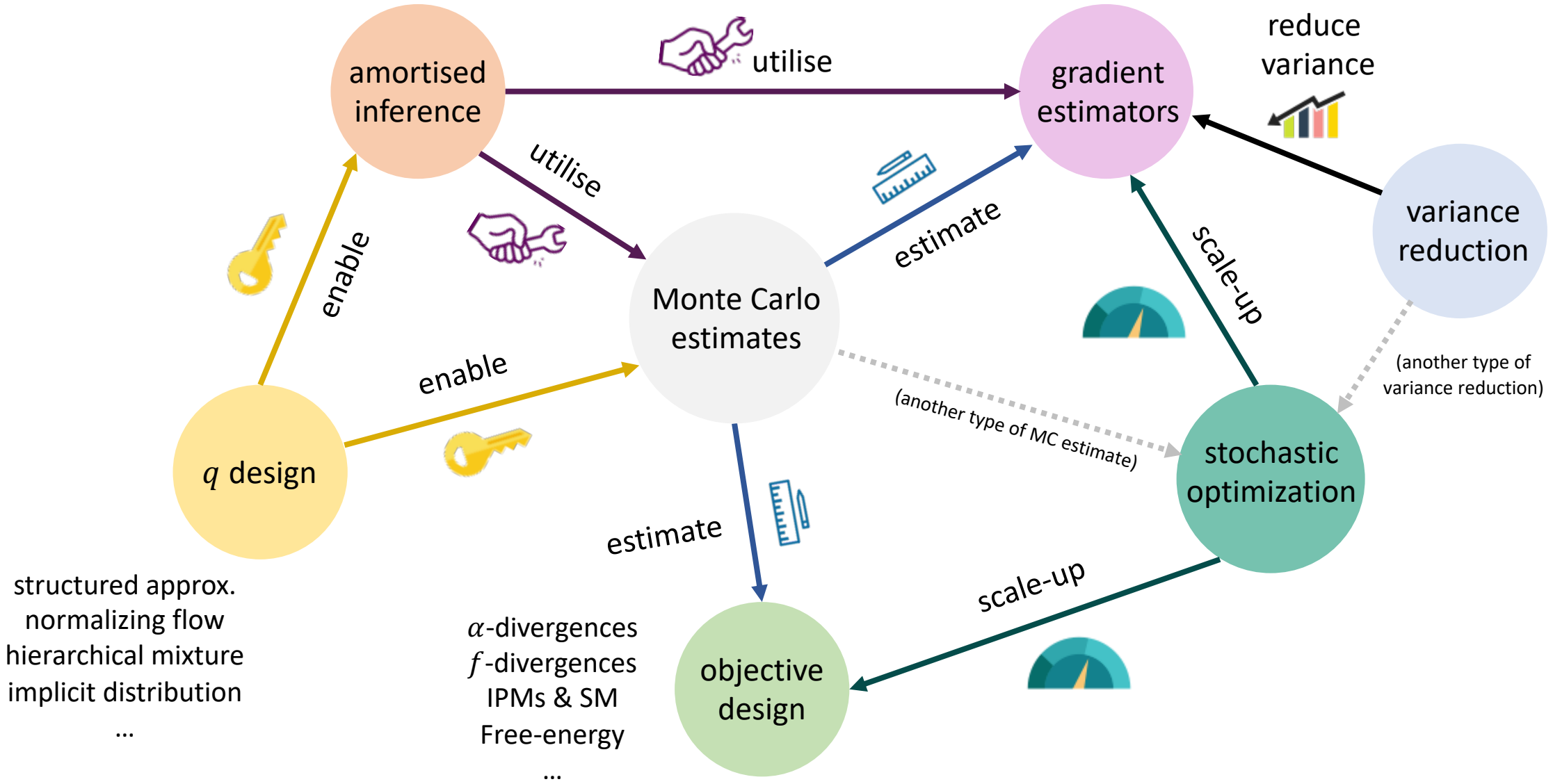
Landscape of Advances



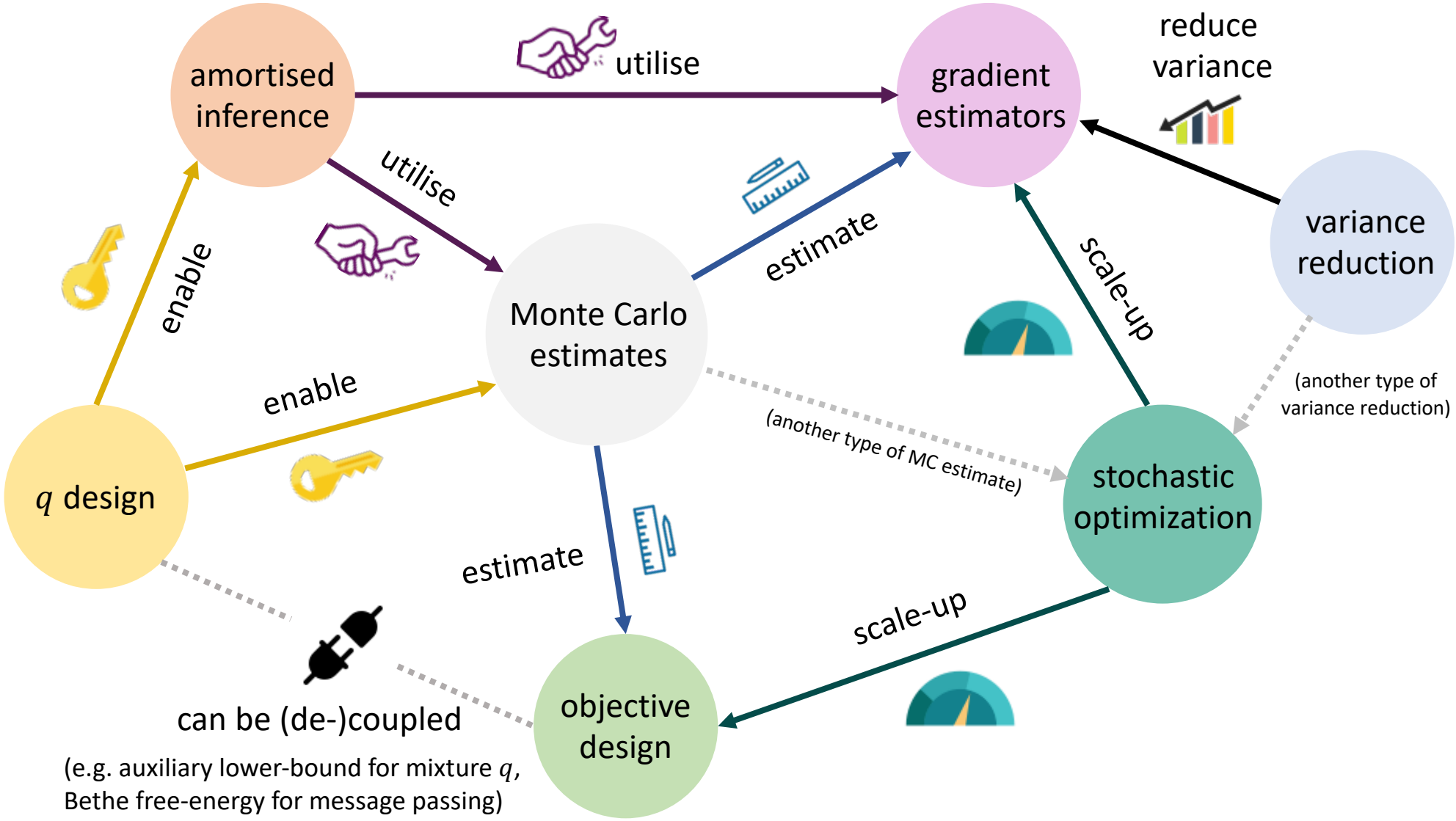
Landscape of Advances



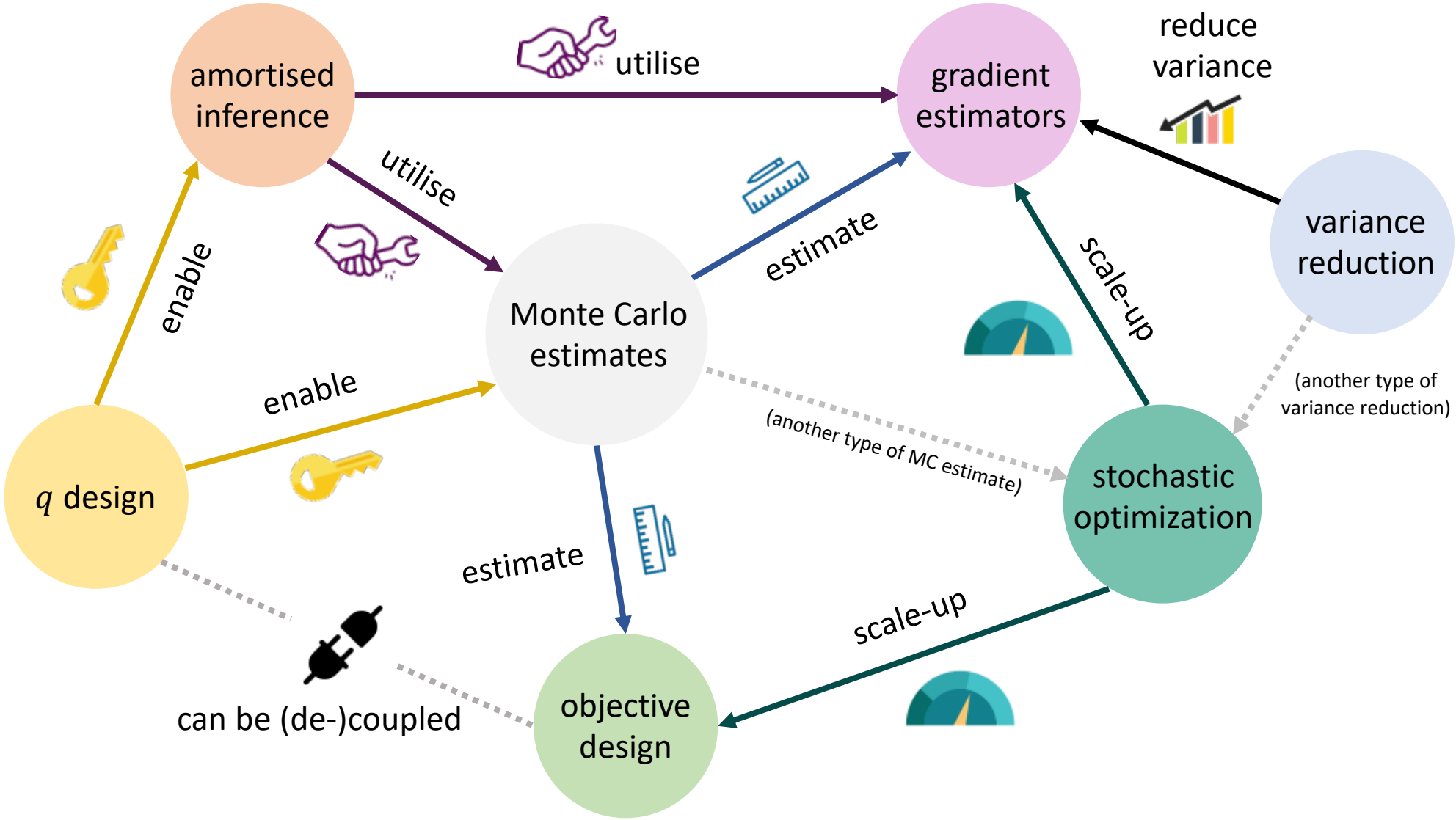
Landscape of Advances



Landscape of Advances



Landscape of Advances



Future Directions - Theory

- Understanding variational inference
  - Optimisation issues
  - Properties of the bounds (for model selection)

- Understanding encoder design in VAE:
  - Amortization gap
  - Inductive bias of q
  - How would approximate inference impact on p model learning
    - e.g., estimation of causal deep generative models
    - e.g., why “bottom-up” encoder in Hierarchical VAE doesn’t work very well

Future Directions - Methodology

- Scaling up
  - Memory efficiency in Bayesian neural network context
  - Faster test-time computation
- Connecting with sampling methods
  - Translating theory of MCMC to variational inference
  - Optimising design choices in sampling with variational inference

Bayesian Neural Network
Tutorial @ ProbAI 2022

Thank You!

Questions? Ask at:
yingzhen.li@imperial.ac.uk