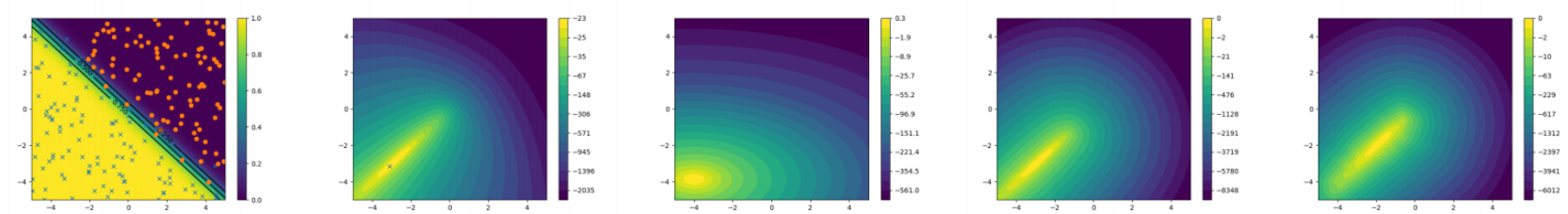
Implicit Variational Inference

David Burt and Andrew Foong

March 2019



(a) Training data (b) True posterior (c) VI (factorized) (d) HMC (e) KIVI

Variational Inference: Advantages and Limitations

- Variational inference converts inference into optimization.
- Amount of bias depends on the expressiveness variational posterior.
- Posterior is commonly assumed to be in exponential family, often mean-field Gaussian.

What if we want to use a more flexible variational posterior?

More Flexible VI Families

There have been many proposals for more flexible posteriors, including:

- Structured variational families (Saul & Jordan, 1996).
- Mixture distributions (Bishop et. al. 1998).
- Hierarchical posteriors (Ranganath et. al. 2016).
- Normalising flows (Rezende & Mohamed, 2015).
- Sampler-based variational posteriors (Salimans et. al. 2015).

Implicit variational inference is a recently developing field that offers a very flexible posterior family.

Implicit Distributions

Implicit distributions are distributions where we can:

- Sample from them easily.
- Take gradients of the samples with respect to the parameters of the distribution.

Example: Let $\epsilon \sim \mathcal{N}(0, I)$ and let $z = f_W(\epsilon)$ be the output of a neural network that takes ϵ as input. The distribution $q(z)$ is an implicit distribution.

- We can differentiate z wrt W by backpropagation.
- We don't have an explicit expression for $q(z)$.
- $q(z)$ can be very flexible.

Could we use implicit distributions as variational families for VI?

VI for latent variable models

Consider the following latent variable model, where x are observed:

$$p_\theta(x) = \int p(z)p_\theta(x|z)dz$$

- Want posterior $p(z|x)$ and also maximum likelihood estimate of θ.
- Intractable since we usually can't compute $p_\theta(x)$.
- Introduce a variational distribution $q_\phi(z)$ and minimise $KL(q_\phi(z)||p(z|x))$.

ELBO

It is easy to show that:

$$\begin{aligned}\mathcal{F}_{\theta,\phi} &\equiv \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - KL(q_\phi(z)||p(z)) \\ &= \log p_\theta(x) - KL(q_\phi(z)||p(z|x)) \\ &\leq \log p_\theta(x).\end{aligned}$$

$\mathcal{F}_{\theta,\phi}$ is known as the variational free energy or evidence lower bound (ELBO).

If we can evaluate $\mathcal{F}_{\theta,\phi}$:

- minimise $KL(q_\phi(z)||p(z|x))$ by maximising wrt ϕ.
- approximate max likelihood learning by maximising wrt θ.

Get inference and learning from a single objective!

ELBO

$$\begin{aligned}\mathcal{F}_{\theta,\phi} &\equiv \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - KL(q_\phi(z)||p(z)) \\ &= \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(z)}{p(z)}\right]\end{aligned}$$

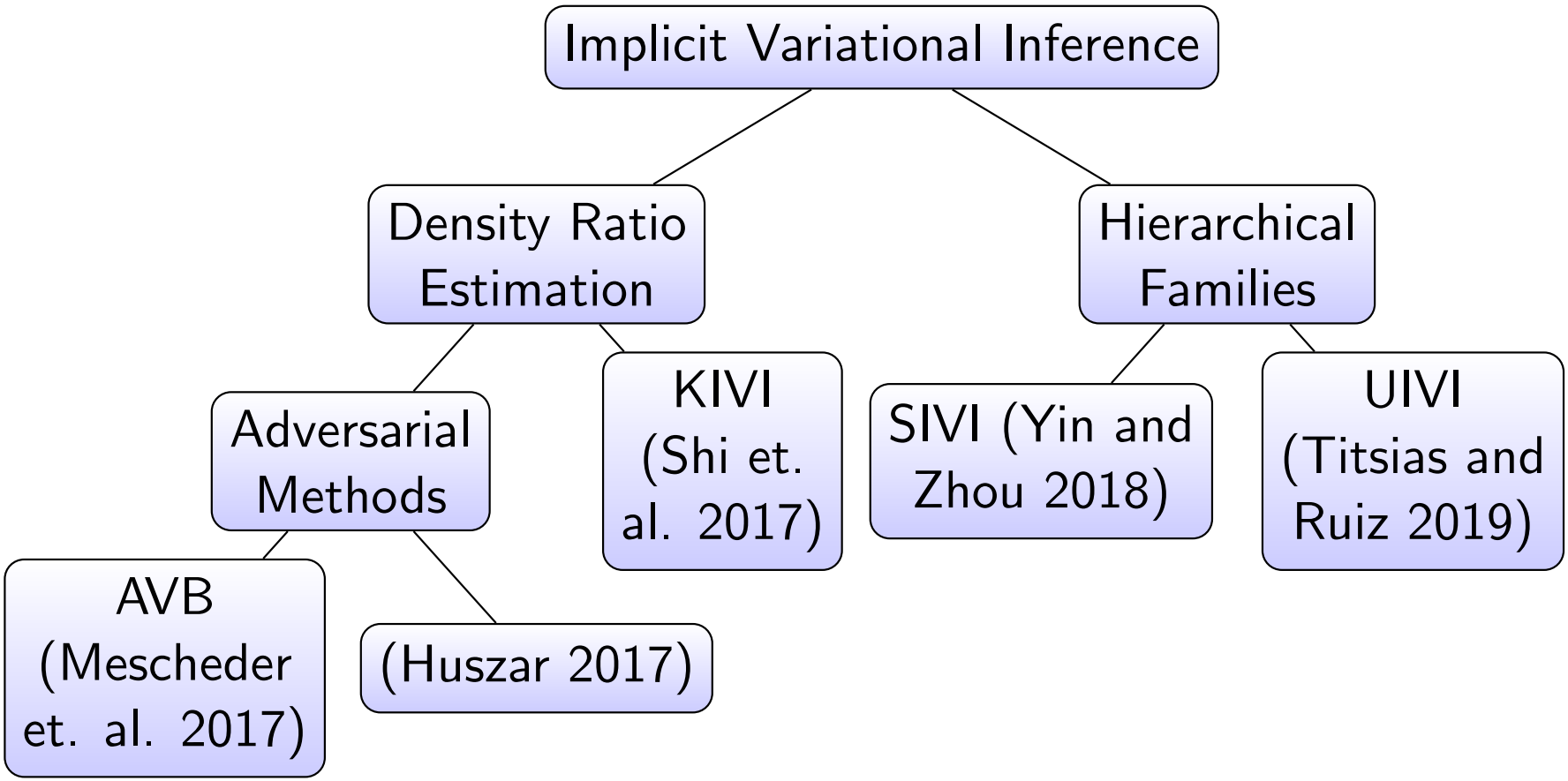
The first term can be estimated by Monte Carlo sampling.

- Gradients calculated using the reparameterisation trick.
- Tractable as long as can sample from q_ϕ and evaluate $\log p_\theta(x|z)$

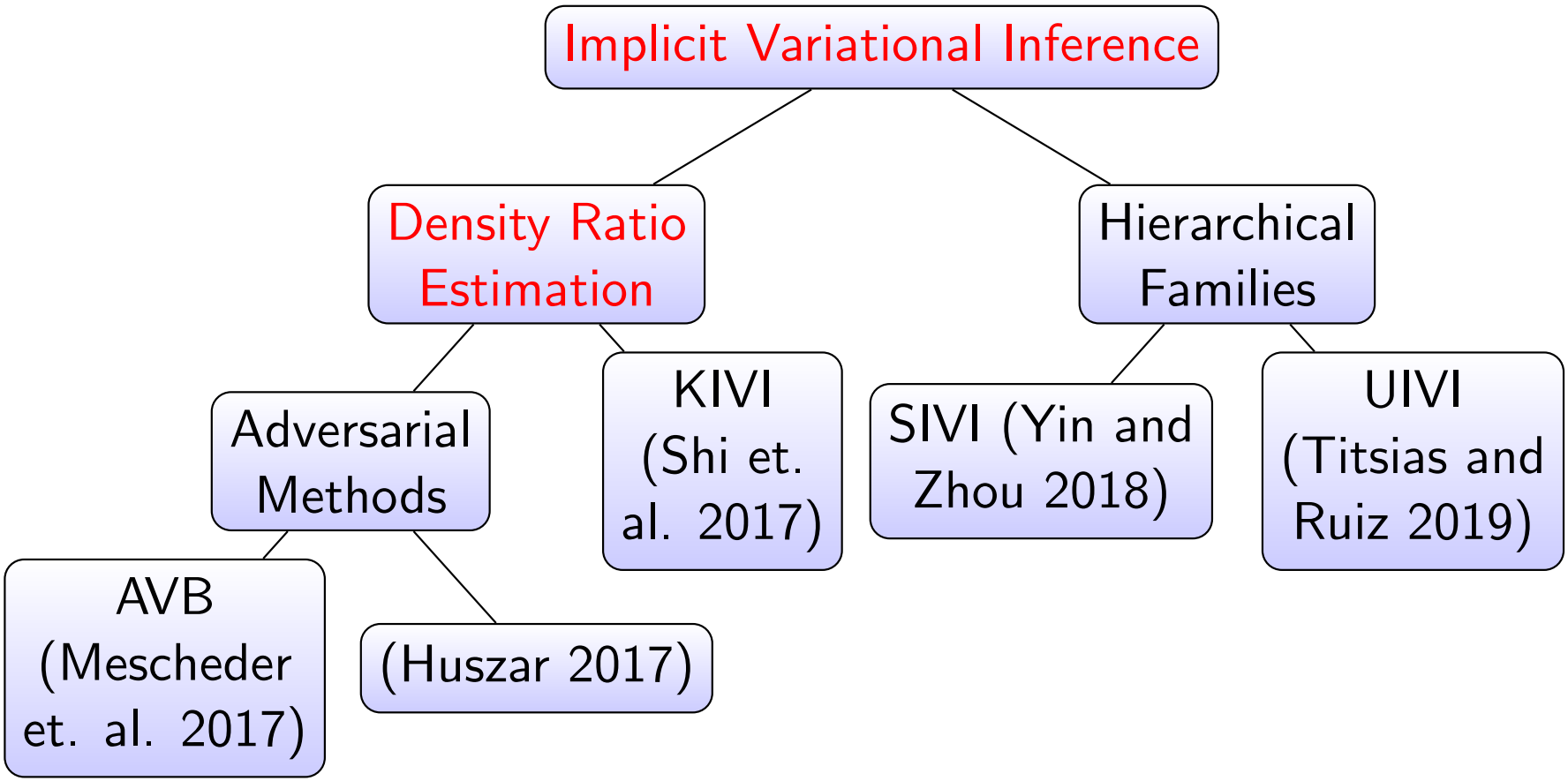
The KL term depends on the form of the distributions:

- If q_ϕ and p in exponential family, analytically tractable.
- If $\log \frac{q_\phi}{p}$ can be evaluated, can use reparameterisation trick.
- if q_ϕ is implicit, $\log \frac{q_\phi}{p}$ cannot be evaluated.

Methods for Implicit Variational Inference



Methods for Implicit Variational Inference



Density Ratio Estimation

$$\mathcal{F}_{\theta,\phi} = \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(z)}{p(z)}\right]$$

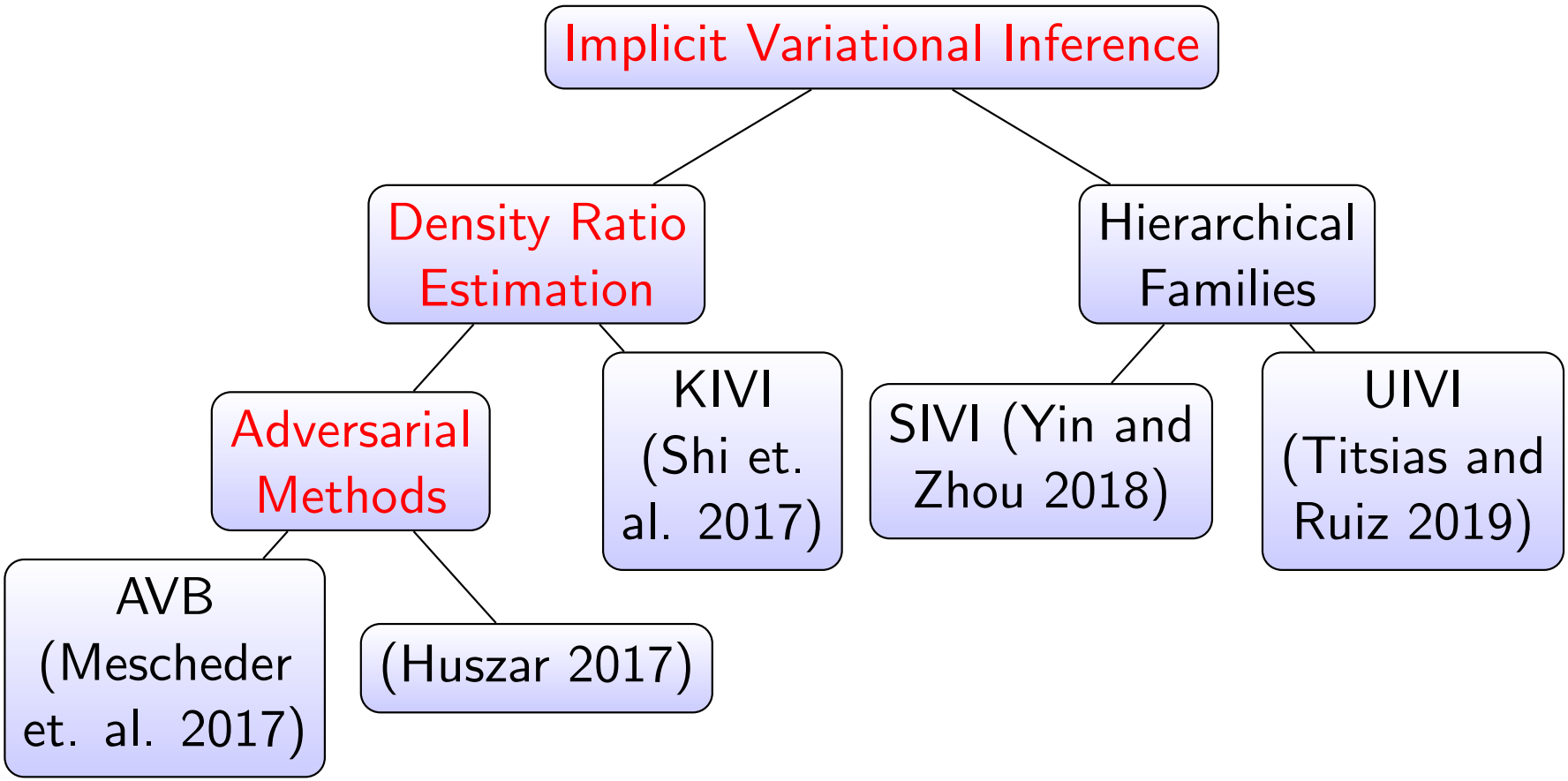
In order to evaluate the ELBO, need to estimate $\log \frac{q_\phi(z)}{p(z)}$ given only samples from q_ϕ and p.

The general problem of estimating $\frac{p_1(z)}{p_2(z)}$ given only samples $z_1 \sim p_1$ and $z_2 \sim p_2$ is known as density ratio estimation (DRE).

Here we focus on two very different approaches:

- Discriminator based/adversarial methods.
- Kernel methods.

Methods for Implicit Variational Inference



Discriminators for Density Ratio Estimation

Idea: We can translate the DRE problem into a supervised learning problem.

Let $D(z)$ be a discriminator network. Train $D(z)$ to maximise the objective function:

$$\mathbb{E}_{q_\phi}\left[\log D(z)\right] + \mathbb{E}_{p}\left[\log(1 - D(z))\right]$$

This has the interpretation:

- Draw $z_i \sim q_\phi$ with probability $1/2$ and $z_i \sim p$ with probability $1/2$ for $i = 1, 2, ...N$.
- Let the *label* $y_i = 1$ if z_i was drawn from q_ϕ, and 0 otherwise.
- This is the expected reward of a *logarithmic scoring rule*.

Proper Scoring Rules

Consider a classifier that returns a probability vector $r(z)$ where $r_y(z)$ is the probability of class y given observation z.

- A proper scoring rule defines a reward variable that takes the value $S(r(z), y)$ if y is the true class for z.
- The expected reward $\mathbb{E}_{z,y}\left[S(r(z), y)\right]$ is uniquely maximised by the true probabilities $r_y(z) = p(y|z)$
- The logarithmic scoring rule $S(r(z), y) = \log r_y(z)$ is strictly proper.

Hence the objective is maximised when

$$\begin{aligned} D^*(z) &= p(y = 1|z) \\ &= \frac{q_\phi(z)}{q_\phi(z) + p(z)} \end{aligned}$$

Approximate ELBO

If we assume the discriminator $D(z)$ globally maximises the objective, then

$$D^*(z) = \frac{q_\phi(z)}{q_\phi(z) + p(z)}$$

$$\log \frac{q_\phi(z)}{p(z)} = \log \frac{D^*(z)}{1 - D^*(z)}$$

Therefore we can approximate the ELBO as:

$$\begin{aligned}\mathcal{F}_{\theta,\phi} &= \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(z)}{p(z)}\right] \\ &\approx \mathbb{E}_{q_\phi}\left[\log p_\theta(x|z)\right] - \mathbb{E}_{q_\phi}\left[\log D(z) - \log(1 - D(z))\right]\end{aligned}$$

where gradients of all terms are obtained using the reparameterisation trick.

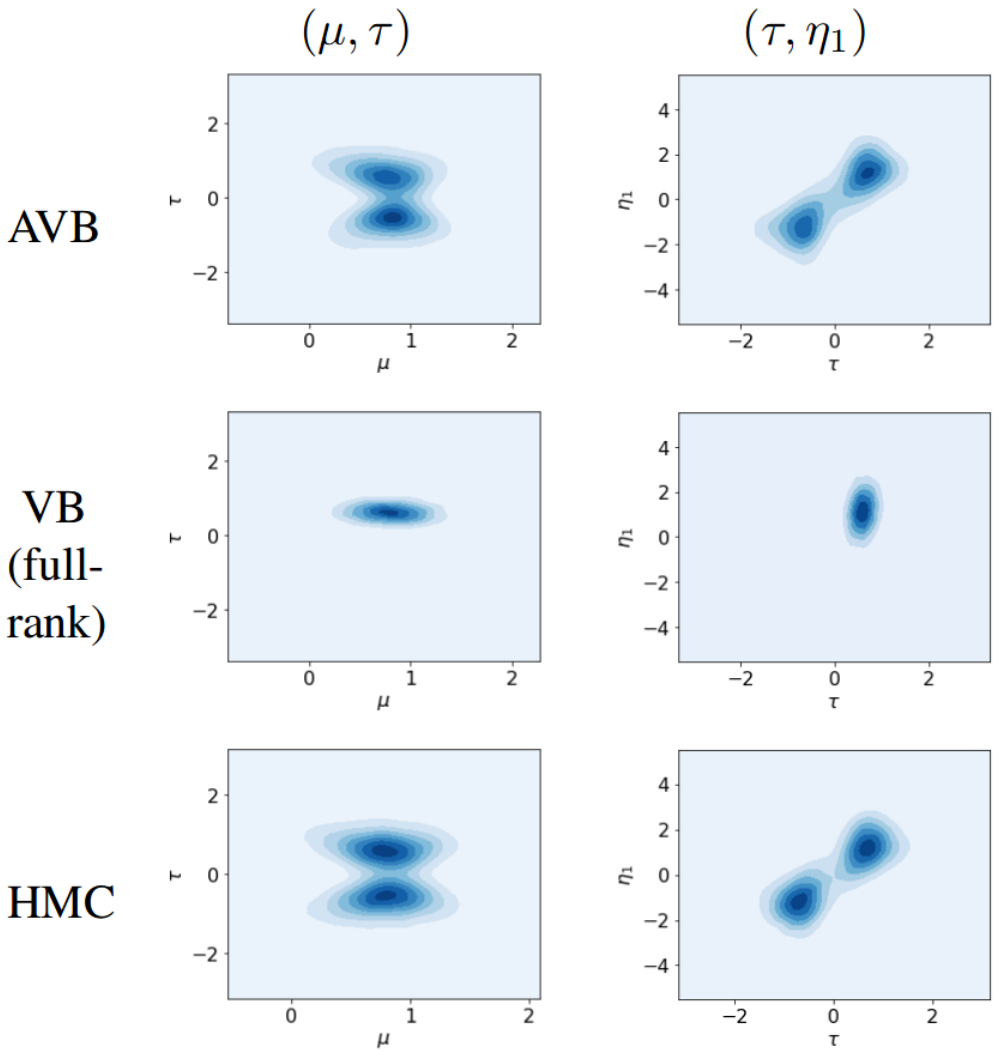
Case study: Adversarial Variational Bayes (AVB)

Mescheder et. al. 2017 perform approximate inference on the 'eight schools' example.

- Two layer network for the implicit posterior and 5 layer ResNet for the discriminator.
- For every posterior update step, perform 2 steps for discriminator.

They compare AVB with full rank Gaussian VI and HMC.

Eight Schools Posterior



Adversarial Variational Bayes for VAEs

Mescheder et. al. 2017 also train a VAE with an implicit distribution as the encoder.

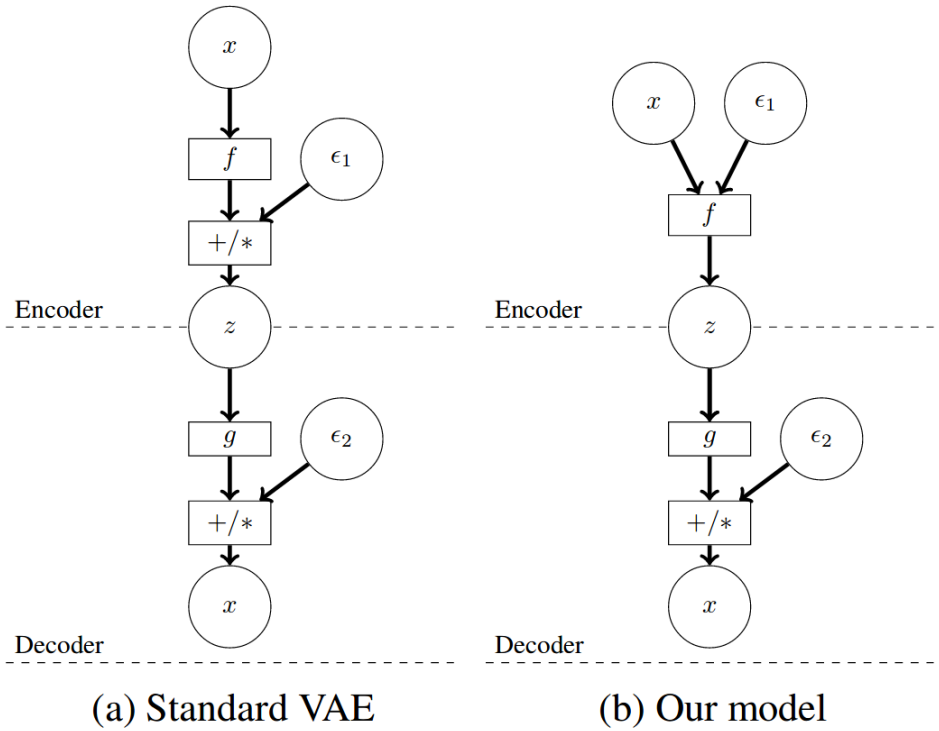


Figure 2. Schematic comparison of a standard VAE and a VAE with black-box inference model, where ϵ_1 and ϵ_2 denote samples from some noise distribution. While more complicated inference models for Variational Autoencoders are possible, they are usually not as flexible as our black-box approach.

Adversarial Variational Bayes for VAEs

(a) Training data (b) Random samples

Figure 7. Independent samples for a model trained on MNIST

	$\log p(x) \geq$	$\log p(x) \approx$	
AVB (8-dim)	$(\approx -83.6 \pm 0.4)$	-91.2 ± 0.6	
AVB + AC (8-dim)	$\approx -96.3 \pm 0.4$	-89.6 ± 0.6	
AVB + AC (32-dim)	$\approx -79.5 \pm 0.3$	-80.2 ± 0.4	
VAE (8-dim)	-98.1 ± 0.5	-90.9 ± 0.6	
VAE (32-dim)	-87.2 ± 0.3	-81.9 ± 0.4	
VAE + NF (T=80)	-85.1	$-$	(Rezende & Mohamed, 2015)
VAE + HVI (T=16)	-88.3	-85.5	(Salimans et al., 2015)
convVAE + HVI (T=16)	-84.1	-81.9	(Salimans et al., 2015)
VAE + VGP (2hl)	-81.3	$-$	(Tran et al., 2015)
DRAW + VGP	-79.9	$-$	(Tran et al., 2015)
VAE + IAF	-80.8	-79.1	(Kingma et al., 2016)
Auxiliary VAE (L=2)	-83.0	$-$	(Maaløe et al., 2016)

Table 2. Log-likelihoods on binarized MNIST for AVB and other methods improving on VAEs. We see that our method achieves state of the art log-likelihoods on binarized MNIST. The approximate log-likelihoods in the lower half of the table were not obtained with AIS but with importance sampling.

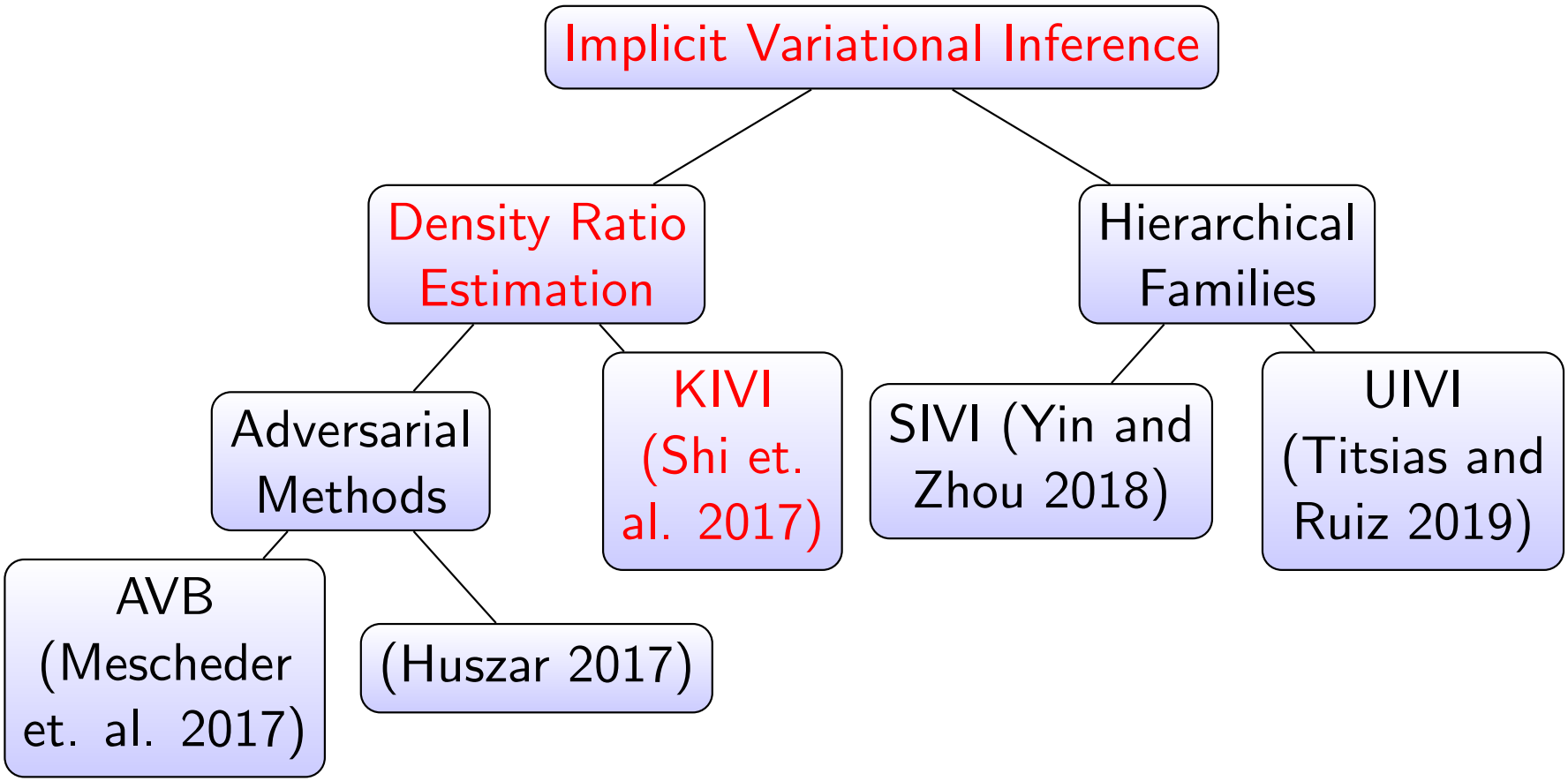
Pitfalls of Discriminator-based Implicit VI

Get correct ELBO if the discriminator is optimal. However in practice:

- Discriminator may not be sufficiently flexible.
- Discriminator objective is Monte Carlo sampled.
- q_ϕ changes with every training iteration - discriminator needs to 'catch up'.

Hence the approximate ELBO is biased and there is no way to estimate/bound the error - not a true lower bound.

Methods for Implicit Variational Inference



Kernel Implicit Variational Inference

Recall,

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}\left[p(x \mid z)\right] + \mathbb{E}_{q_\phi}\left[\log\left(\frac{p(z)}{q_\phi(z)}\right)\right]$$

As before, the first term can be handled directly with MC methods.

Main Idea: Define $r(z) = \frac{p(z)}{q_\phi(z)}$. Approximate r using kernel methods.

Approximation to r

Solve the minimization problem:

$$\min_{\hat{r}\in\mathcal{H}} \frac{1}{2}\left(\int (\hat{r}(z) - r(z))^2 q_\phi(z) dz\right)$$

Note that,

$$\frac{1}{2}\left(\int (\hat{r}(z) - r(z))^2 q_\phi(z) dz\right) = \frac{1}{2}\mathbb{E}_{q_\phi}\left[\hat{r}(z)^2\right] - \mathbb{E}_p\left[\hat{r}(z)\right] + C$$

We can sample from both p and q in order to approximate the loss function.

RKHS and Representer Theorem

A regularization term is introduced, and the objective function that is minimized is:

$$\min_{\hat{r}\in\mathcal{H}} \left(\frac{1}{2}\sum_{i=1}^{n_q} \hat{r}(z_i^q)^2 - \sum_{j=1}^{n_p} \hat{r}(z_j^p) + \frac{\lambda}{2}\|\hat{r}(z)\|_{\mathcal{H}}^2 \right)$$

The representer theorem tells us that there exists a set of coefficients $\{\alpha_i, \beta_j\}$ such that

$$\hat{r}(z) = \sum_{i=1}^{n_q} \alpha_i k(z_i^q, \cdot) + \sum_{j=1}^{n_p} \beta_j k(z_j^p, \cdot).$$

Optimal Solution

The proposed optimization function is convex, and has closed form solution,

$$\boldsymbol{\beta} = \frac{1}{\lambda n_p}\mathbf{1}_{n_p} \text{ and } \boldsymbol{\alpha} = -\frac{1}{\lambda n_p n_q}\left(\mathbf{K_q} + \lambda \mathbf{I}\right)^{-1}\mathbf{K_{qp}}\mathbf{1}_{n_p}.$$

This estimator for $\hat{r}(z)$ is used in place of $\frac{q_\phi(z)}{p(z)}$ to estimate the density ratio in the ELBO[1].

[1] It may need to be clipped in order to be nonnegative

KIVI Algorithm

Algorithm 1 Algorithm for Approximate ELBO

Sample $z_i^p \sim p(z_i)$
Sample $z_i^q \sim q_\phi(z_i)$
Compute $\log\left(\hat{r}(z_i^q)\right)$ according to the formulas given on the previous two slides.
Use this estimator in place of the density ratio. The first M samples from q are used to estimate the likelihood term.

Application: Training A BNN

- Reduces number of parameter in inference hypernetwork with a *matrix multiplication network*
  $\Rightarrow \mathbf{X}^{(i+1)} = \text{ReLU}\left(\mathbf{A}_1^{(i)}\mathbf{X}^{(i)}\mathbf{A}_2^{(i)}\right) + \mathbf{B}^{(i)}$.
- Inference in a BNN with AVB (or similar adversarial methods) is not generally feasible due to the high dimensional input space for the discriminator.

KIVI on BNN for MNIST

Method	# Hidden	# Weights	Test err.
SGD (Simard et al., 2003)	800	1.3m	1.6%
Dropout (Srivastava et al., 2014)			$\approx$ 1.3%
Dropconnect (Wan et al., 2013)	800	1.3m	**1.2%***
Bayes B. (Blundell et al., 2015),	400	500k	1.82%
with Gaussian posterior	800	1.3m	1.99%
	1200	2.4m	2.04%
Bayes B. (Blundell et al., 2015),	400	500k	1.36%*
with scale mixture prior	800	1.3m	1.34%*
	1200	2.4m	1.32%*
KIVI	400	500k	**1.29%**
	800	1.3m	**1.22%**
	1200	2.4m	**1.27%**

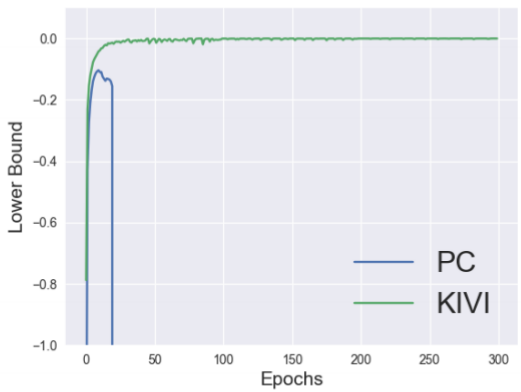
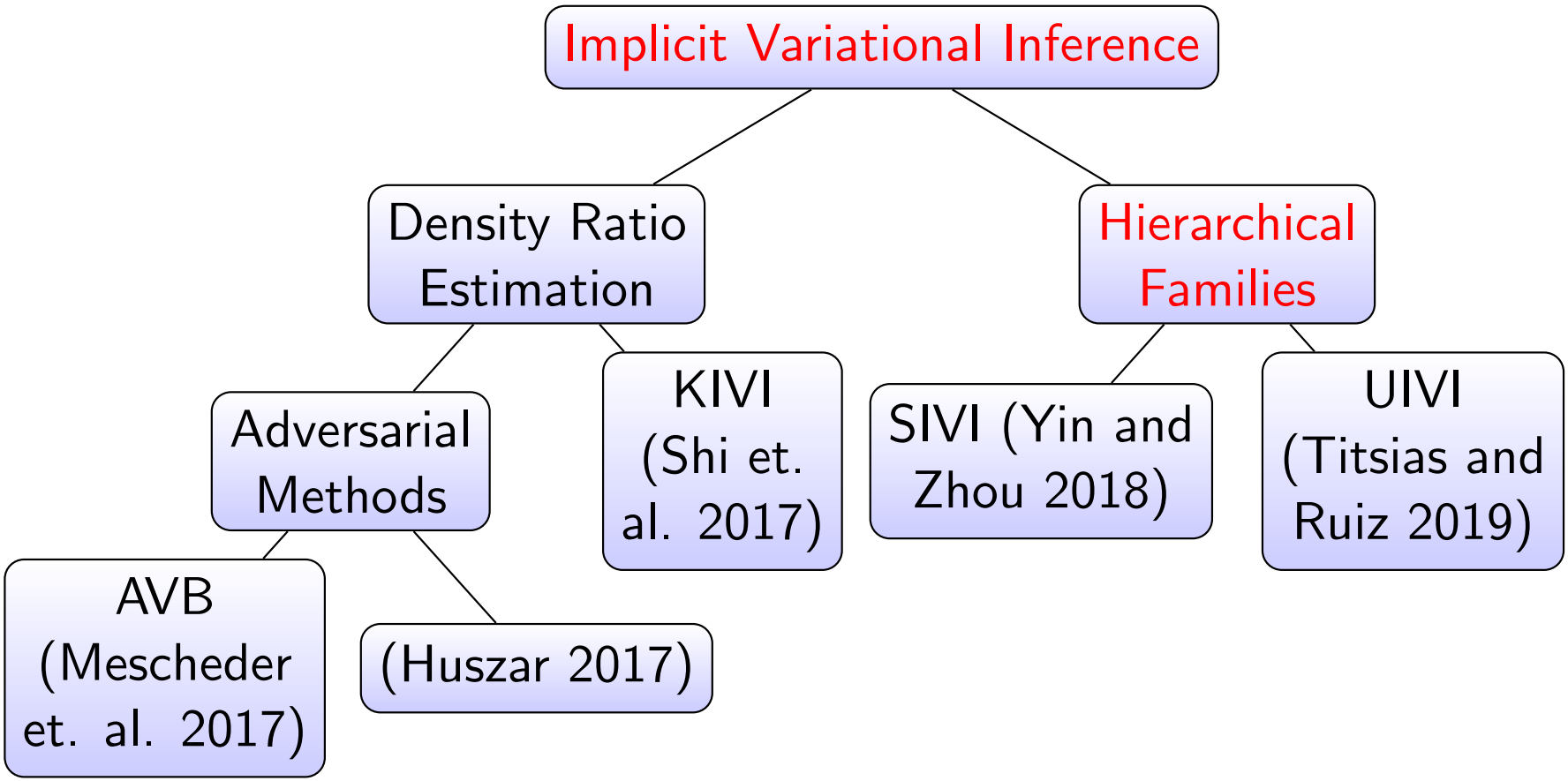


Figure 2: Results for MNIST classification. The left table shows the test error rates. $\star$ indicates results that are not directly comparable to ours: Wan et al. (2013) used an ensemble of 5 networks, and the second part of Blundell et al. (2015) changed the prior to a scale mixture. The plot on the right shows training lower bound in MNIST classification with prior-contrastive and KIVI.

Methods for Implicit Variational Inference



Hierarchical Variational Distributions

Define the posterior by the sampling process:

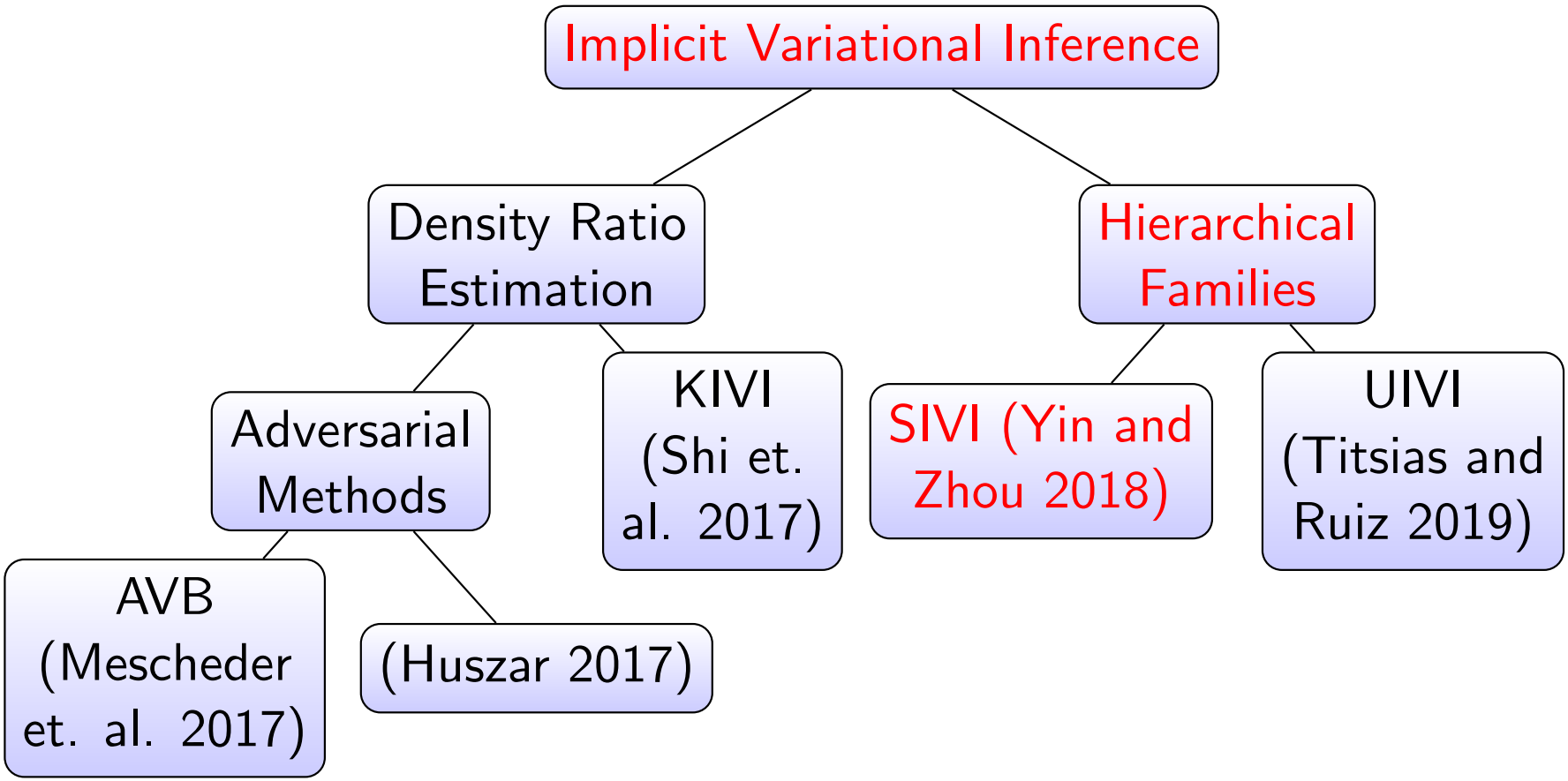
$$\epsilon \sim q(\epsilon)$$
$$z \sim q_\phi(z|\epsilon)$$

Or equivalently:

$$q_\phi(z) = \int q_\phi(z|\epsilon) q(\epsilon)\, d\epsilon$$

$q_\phi(z|\epsilon)$ is a simple distribution (e.g. exponential family) whose parameters are a complicated function of ϵ (e.g. neural network with weights ϕ).
This is called a semi-implicit distribution - still very flexible.

Methods for Implicit Variational Inference



Semi-Implicit Variational Inference (Yin and Zhou 2018)

A tractable lower bound on the ELBO can be derived via the chain rule of KL-divergences:

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{z\sim q(z)}\left[\log\left(\frac{p(x,z)}{\mathbb{E}_{\epsilon\sim q(\epsilon)}\left[q(z|\epsilon)\right]}\right)\right] \\ \geq& \mathbb{E}_{\epsilon\sim q(\epsilon)}\left[\mathbb{E}_{z\sim q(z|\epsilon)}\left[\log\left(\frac{p(x,z)}{q(z|\epsilon)}\right)\right]\right] =: \mathcal{L}_{\text{lower}}\end{aligned}$$

Issue: Using this bound will lead to $q_\phi(z)$ to be a member of the same family of distributions as $q_\phi(z|\epsilon)$

A Sequence of lower bounds

Solution: Sequence of lower bounds such that $\mathcal{L}_{\text{lower}} = \mathcal{L}_0$ and $\lim_{K\to\infty} \mathcal{L}_K = \mathcal{L}$.

$$\mathcal{L}_K = \mathbb{E}_{\epsilon^{0:K}\sim q(\epsilon)}\left[\mathbb{E}_{z\sim q^K(z|\epsilon^{0:K})}\left[\log\left(\frac{p(x,z)}{q^K(z|\epsilon^{(0:K)})}\right)\right]\right]$$

where

$$q^K(z|\epsilon^{0:K}) := \frac{1}{K+1}\sum_{k=0}^{K} q(z|\epsilon^k)$$

These can be shown to be monotonically increasing.

Shortcomings of Implicit Inference

- Many more hyperparameteres to choose/tune than exponential family VI.
- Potentially unstable (particularly using adversarial approaches).
- Unclear how suitable approximate ELBO is for model comparison (except maybe with SIVI)
- Introduces bias into estimation of gradient of ELBO (except maybe UIVI)

Summary of Implicit VI

- Implicit VI allows us to use arbitrarily complicated variational posteriors
- This flexibility comes at a cost of needing to tune many hyperparameters, additional computational cost, or a risk of unstable algorithms.

References I

[1] Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[2] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2391–2400. JMLR. org, 2017.

[3] Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. *arXiv preprint arXiv:1810.02789*, 2018.

[4] Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference, 2017.

[5] Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*, 2018.

References II

[6] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[7] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5646–5655, 2018.