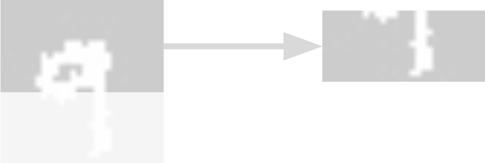
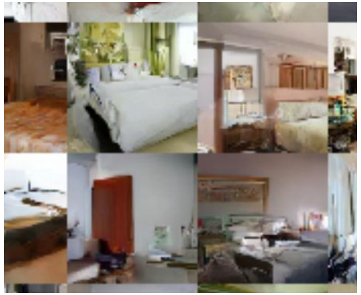


DReGs

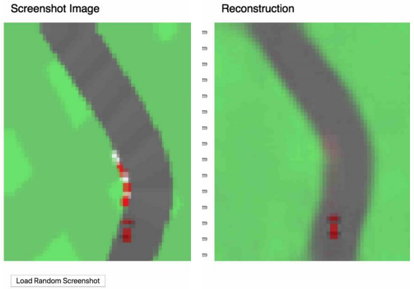
Doubly Reparameterized Gradient Estimators for latent variable models

George Tucker (gjt@google.com)
Dieterich Lawson, Shixiang Gu, Chris J. Maddison

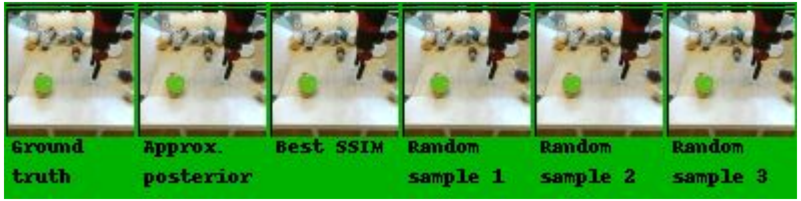
Generative Latent Variable Models



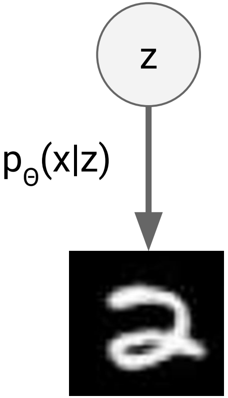
Samples from PixelVAE
(Gulrajani et al. 2016)



Model based RL
(Ha and Schmidhuber 2018)



Samples from SVG
(Denton & Fergus 2018)



z

$p_\Theta(x|z)$

z are latent factors
(e.g., number, stroke
width)

Model data *x* with a generative model

$$p_\Theta(x) = \int p_\Theta(x, z)\, dz$$

Would like to maximize log likelihood, log $p_\Theta(x)$, with stochastic gradient descent.

DReGs

Introduce an **unbiased, low variance gradient estimator for training latent variable models.**

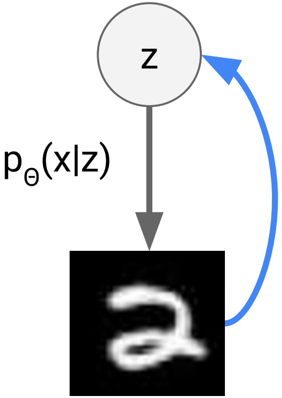
Show its applicability to three recent training techniques for latent variable models:

- **IWAE** (Burda et al. 2015)
- **Reweighted Wake-Sleep (RWS)** (Bornschein & Bengio, 2014)
- **Jackknife Variational Inference (JVI)** (Nowozin, 2018)

1. Background on IWAE
2. DReG estimators
3. Experiments
    a. Gaussian system
    b. Omniglot and MNIST

Latent Variable Models

We optimize a *variational* lower bound (**ELBO**) on the log likelihood



z are latent factors (e.g., number)

$q_\phi(z|x)$ *encodes* images into latent factors

$$\log p_\theta(x) = \log \int_z p_\theta(x, z)\ dz$$

Importance weighted bounds (**IWAE**)

We can improve the bound using K samples instead of 1

$$\mathrm{IWAE}(K) = \mathbb{E}_{z_{1:K}}\left[\log\left(\frac{1}{K}\sum_{i=1}^{K} w_i\right)\right] \leq \log p_\theta(x) \qquad w_i = w(z_i) = \frac{p_\theta(x, z_i)}{q_\phi(z_i|x)}$$

$$\mathrm{ELBO} = \mathrm{IWAE}(1) \leq \mathrm{IWAE}(2) \leq \cdots \leq \mathrm{IWAE}(K) \leq \log p_\theta(x)$$

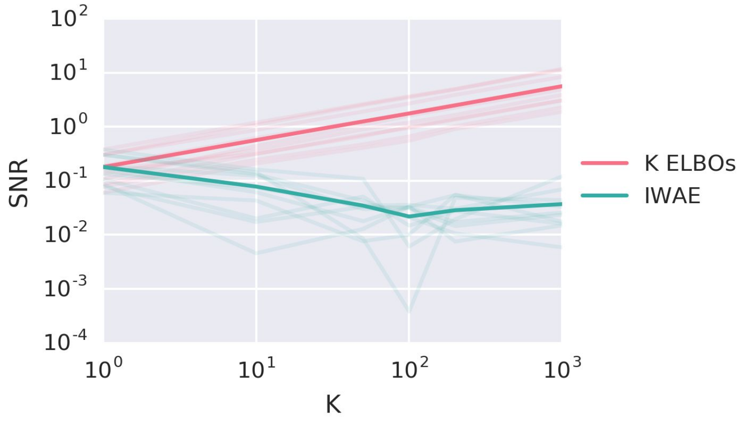
An increasingly tight bound, **so as K -> ∞, the effect of an overly simplistic q_ϕ family diminishes.**

IWAE (Burda 2015)

Importance weighted bounds (**IWAE**)

But the gradient wrt to ϕ gets worse as K increases ...

Gradient of K ELBOs vs. IWAE(K)



SNR = Mean / Standard deviation

As K increases:

- IWAE(K) becomes tighter.
- ϕ gradient estimator for IWAE(K) degrades.

Can we resolve this tension?

Google

Tighter lower bounds (Rainforth et al. 2018)

Double Reparameterized Gradient Estimator

$$\nabla_\phi \mathbb{E}_{z_{1:K}}\left[\log\left(\frac{1}{K}\sum_{i=1}^{K} w_i\right)\right] = \mathbb{E}_{\epsilon_{1:K}}\left[\sum_{i=1}^{K} \frac{w_i}{\sum_j w_j}\nabla_{\theta,\phi}\log w_i\right]$$

The single sample estimator is typically used.

$$= \mathbb{E}_{\epsilon_{1:K}}\Bigg[\boxed{\cancel{\sum_{i=1}^{K} \frac{w_i}{\sum_{j=1}^{K} w_j}\Bigg(-\frac{\partial}{\partial\phi}\log q_\phi(z_i|x)}}+\frac{\partial \log w_i}{\partial z_i}\frac{dz_i}{d\phi}\Bigg)\Bigg]$$

When K = 1, t t

(Apply reparameterization trick)

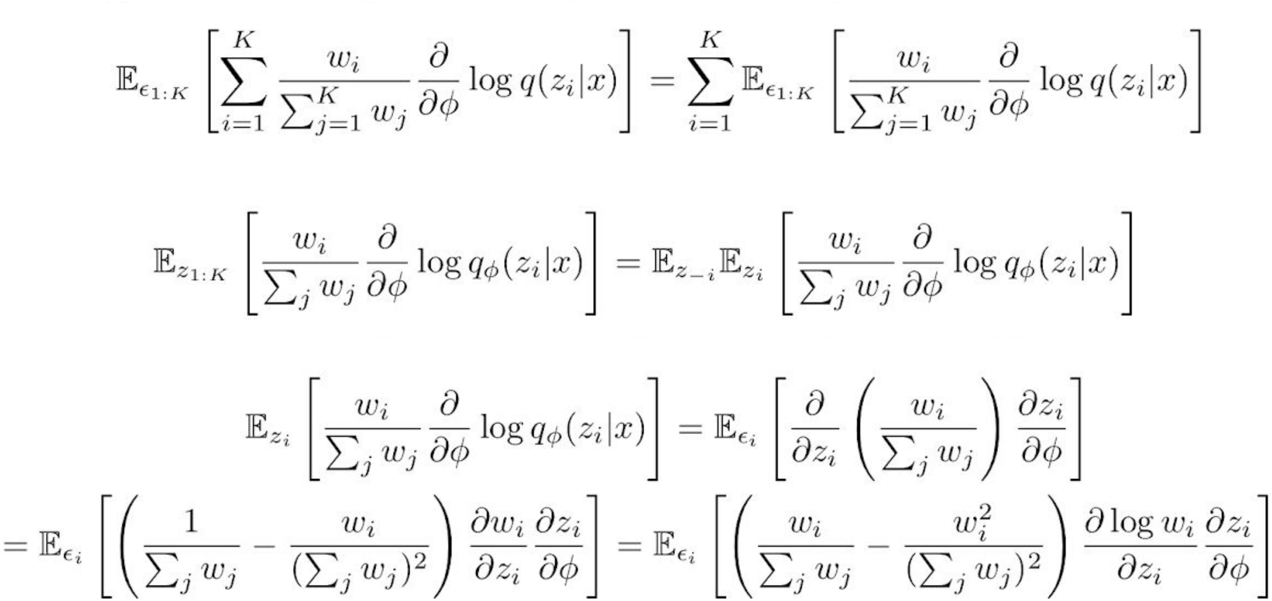
$$= \mathbb{E}_{\epsilon_{1:K}}\left[\sum_{i=1}^{K}\left(\frac{w_i}{\sum_j w_j}\right)^2 \frac{\partial \log w_i}{\partial z_i}\frac{\partial z_i}{\partial\phi}\right]$$

So drop it wh K

al. 2017)

The IWAE-DReG estimator is the single sample Monte Carlo estimator.

Double Reparameterized Gradient Estimator



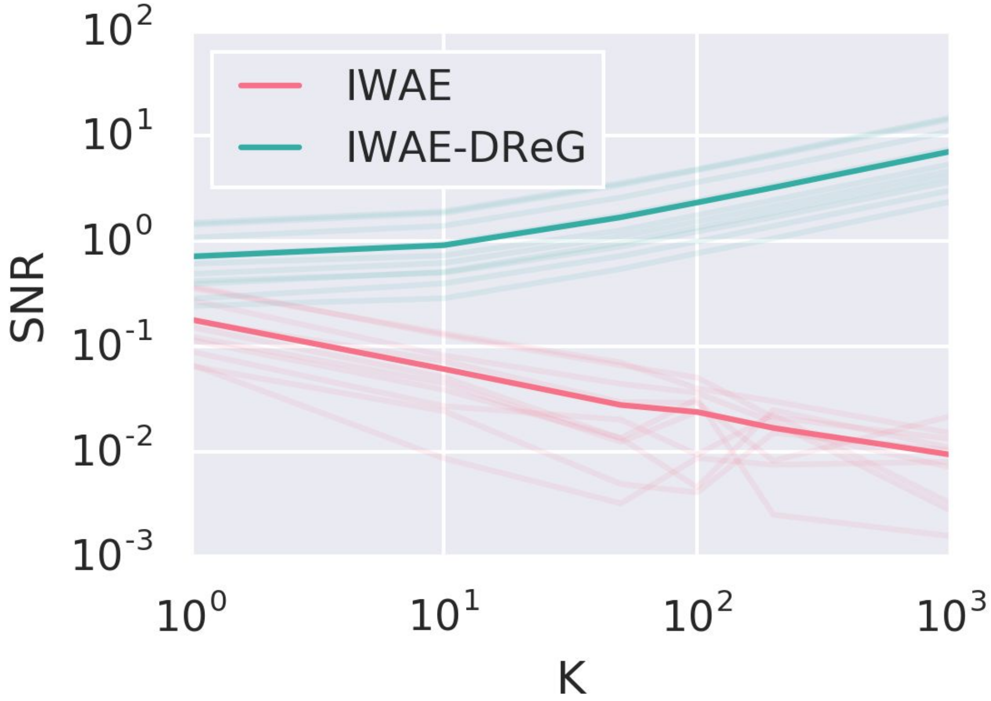
$$\mathbb{E}_{\epsilon_{1:K}}\left[\sum_{i=1}^{K}\frac{w_i}{\sum_{j=1}^{K} w_j}\frac{\partial}{\partial\phi}\log q(z_i|x)\right] = \sum_{i=1}^{K}\mathbb{E}_{\epsilon_{1:K}}\left[\frac{w_i}{\sum_{j=1}^{K} w_j}\frac{\partial}{\partial\phi}\log q(z_i|x)\right]$$

$$\mathbb{E}_{z_{1:K}}\left[\frac{w_i}{\sum_j w_j}\frac{\partial}{\partial\phi}\log q_\phi(z_i|x)\right] = \mathbb{E}_{z_{-i}}\mathbb{E}_{z_i}\left[\frac{w_i}{\sum_j w_j}\frac{\partial}{\partial\phi}\log q_\phi(z_i|x)\right]$$

$$\mathbb{E}_{z_i}\left[\frac{w_i}{\sum_j w_j}\frac{\partial}{\partial\phi}\log q_\phi(z_i|x)\right] = \mathbb{E}_{\epsilon_i}\left[\frac{\partial}{\partial z_i}\left(\frac{w_i}{\sum_j w_j}\right)\frac{\partial z_i}{\partial\phi}\right]$$

$$= \mathbb{E}_{\epsilon_i}\left[\left(\frac{1}{\sum_j w_j} - \frac{w_i}{(\sum_j w_j)^2}\right)\frac{\partial w_i}{\partial z_i}\frac{\partial z_i}{\partial\phi}\right] = \mathbb{E}_{\epsilon_i}\left[\left(\frac{w_i}{\sum_j w_j} - \frac{w_i^2}{(\sum_j w_j)^2}\right)\frac{\partial \log w_i}{\partial z_i}\frac{\partial z_i}{\partial\phi}\right]$$

Simple Gaussian System

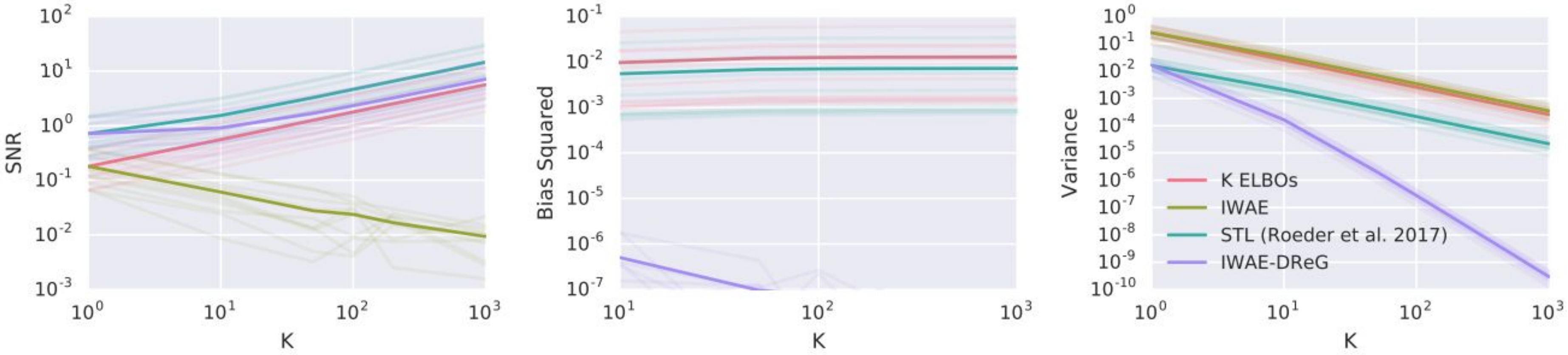


Delta method implies

IWAE SNR: $O(1/K^{0.5})$

IWAE-DreG SNR: $O(K^{0.5})$

Simple Gaussian System



Reweighted Wake-Sleep (RWS)

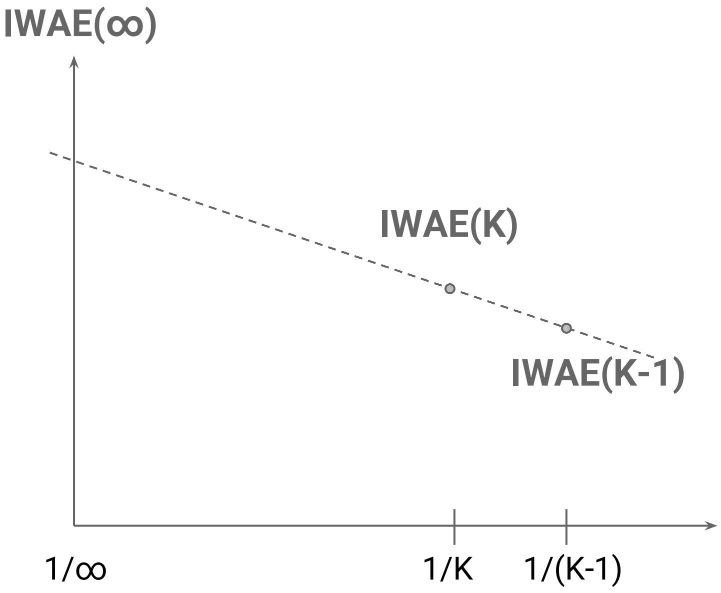
Instead, optimize ϕ by minimizing KL(p(z|x) || q(z|x))

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

$$-\mathbb{E}_{p_\theta(z|x)}\left[\frac{\partial}{\partial\phi}\log q_\phi(z|x)\right] \approx -\mathbb{E}_{z_{1:K}}\underbrace{\left[\sum_i \frac{w_i}{\sum_j w_j}\frac{\partial}{\partial\phi}\log q_\phi(z_i|x)\right]}_{\text{(Apply reparameterization trick)}}$$

RWS (Bornschein and Bengio 2014)

Jackknife Variational Inference (JVI)



Use a linear combination of IWAE(K) and IWAE(K-1) to cancel the first order bias term.

JVI(1)

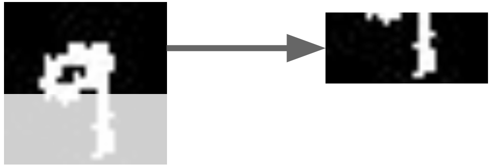
$$K \times \underbrace{\mathbb{E}_{z_1:K}\left[\log\left(\frac{1}{K}\sum_{i=1}^{K} w_i\right)\right]}_{\text{IWAE(K)}} - \frac{K-1}{K}\sum_{i=1}^{K} \underbrace{\mathbb{E}_{z_{-i}}\left[\log\left(\frac{1}{K-1}\sum_{j\neq i} w_j\right)\right]}_{\text{IWAE(K-1)}}$$

JVI (Nowozin 2018)

Experiments

MNIST & Omniglot Generative Modeling

MNIST Structured Prediction



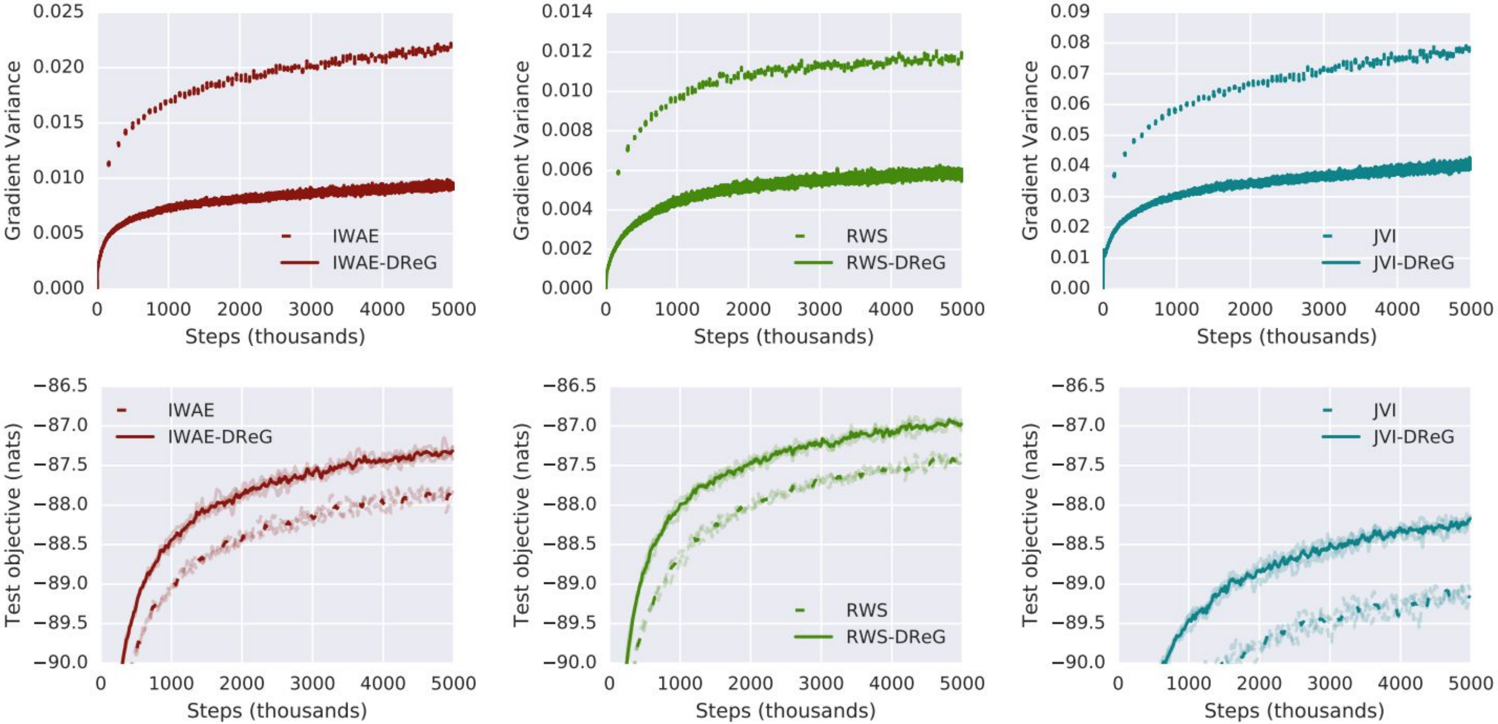
Network details

50 dimensional latent variable (z)

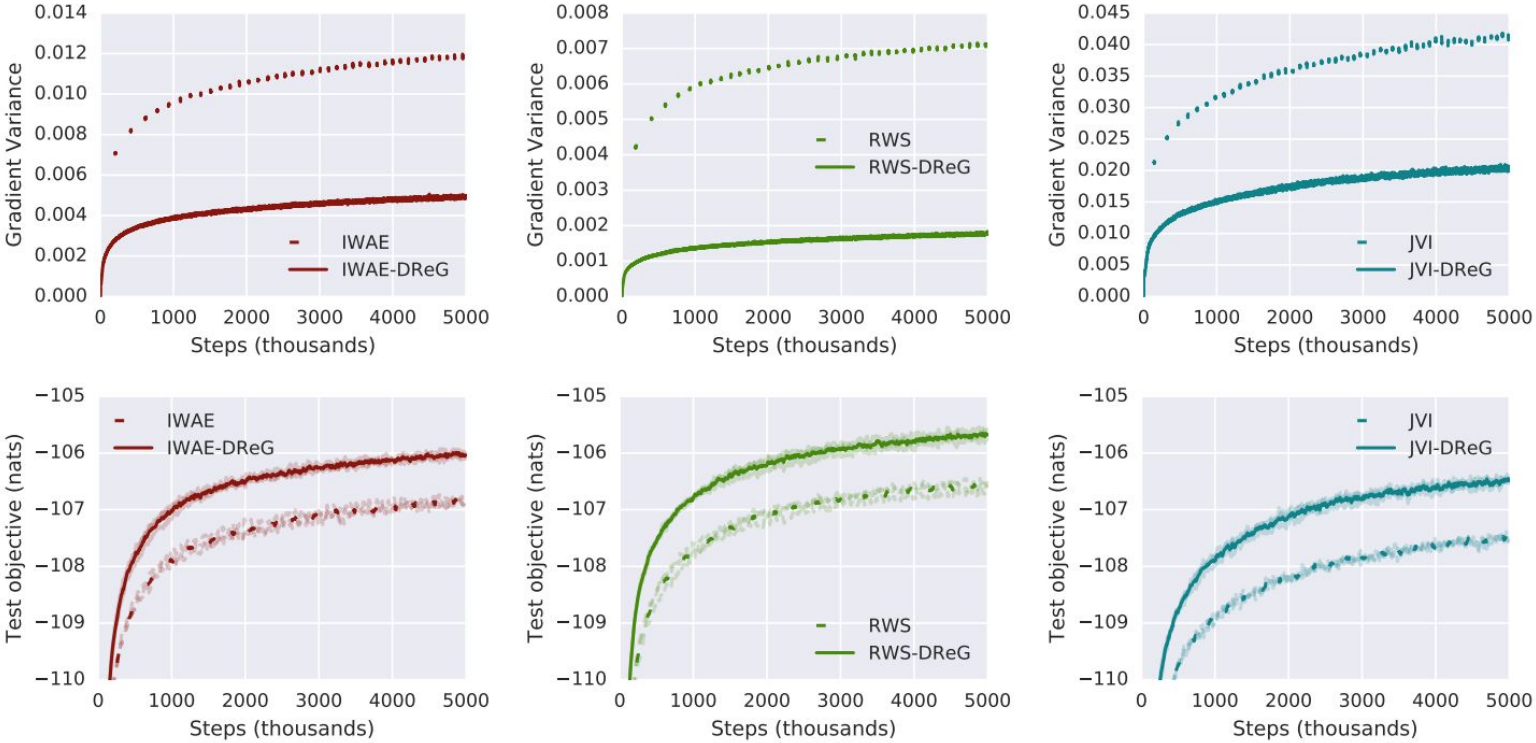
2 hidden layers 200 tanh units

Factorized Bernoulli and Gaussian distributions

MNIST Generative Modeling

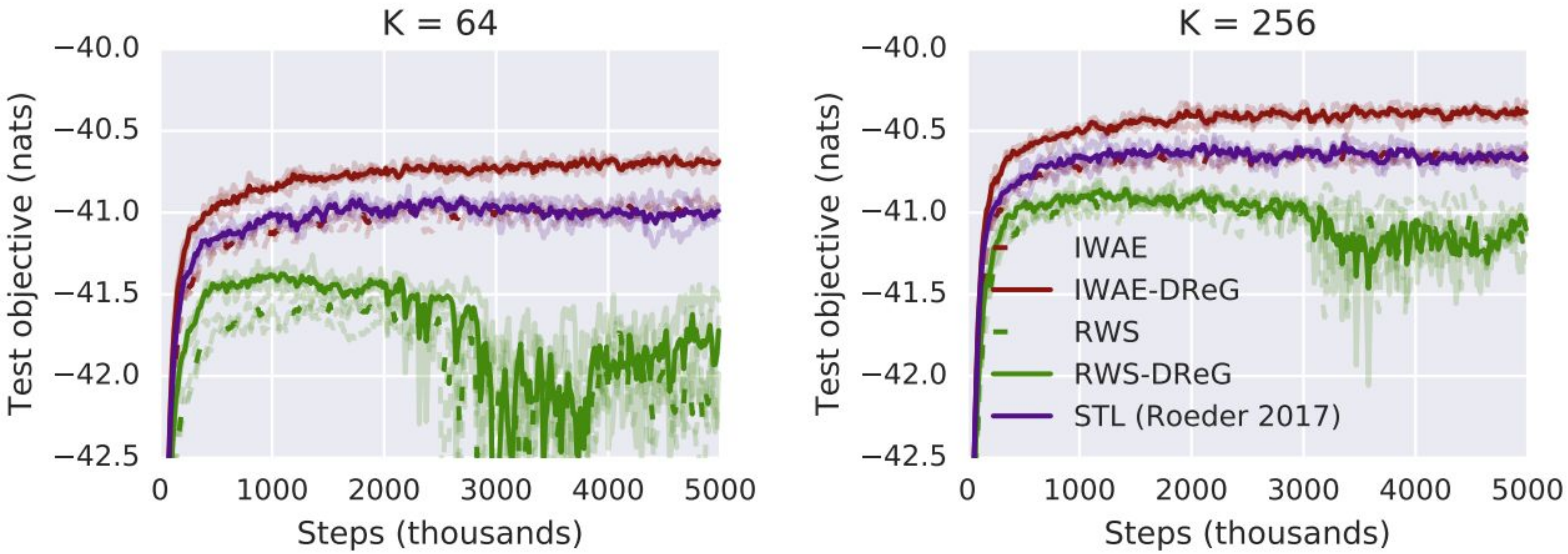


Omniglot Generative Modeling



Google

MNIST Structured Prediction



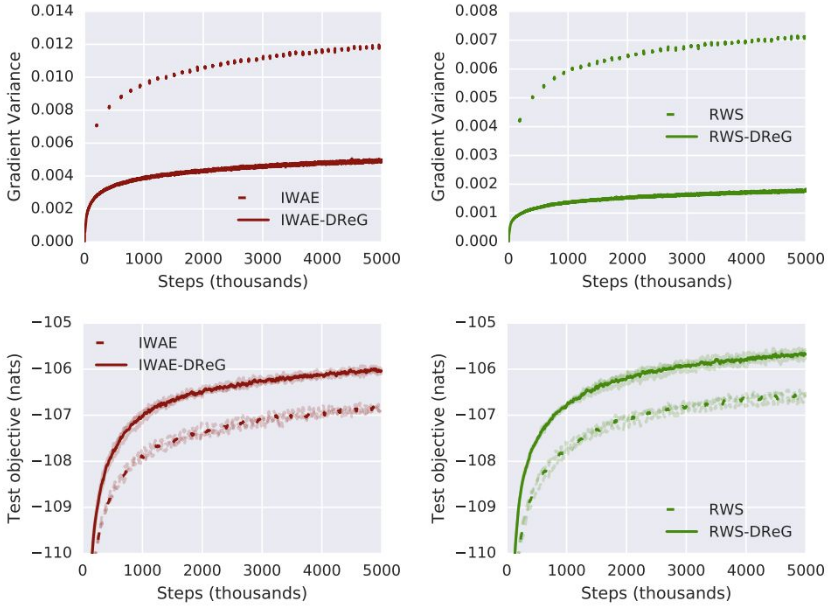
Summary & Future Work

DReG estimators are computationally efficient, unbiased, drop-in replacements for standard gradient estimators.

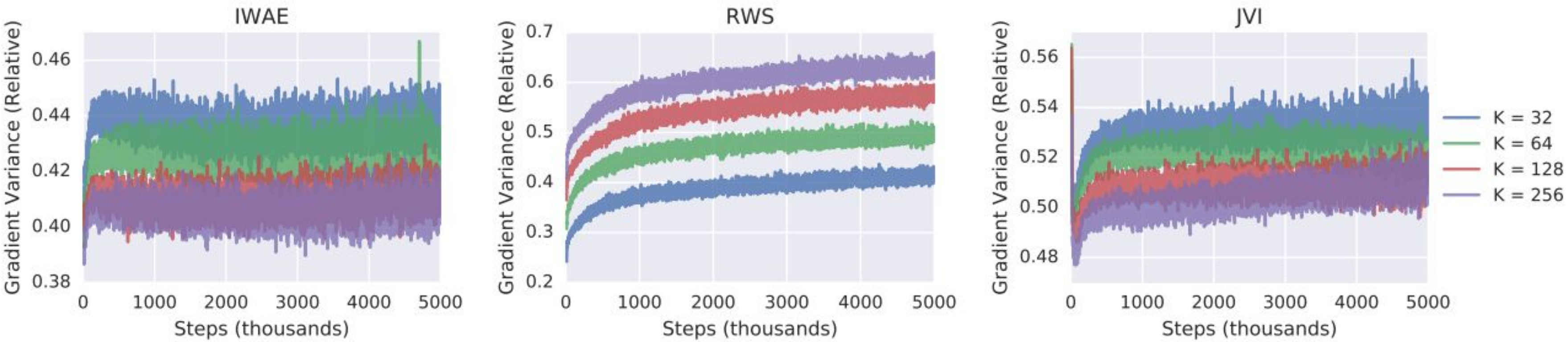
They rectify practical and asymptotic issues raised in Rainforth et al. 2018.

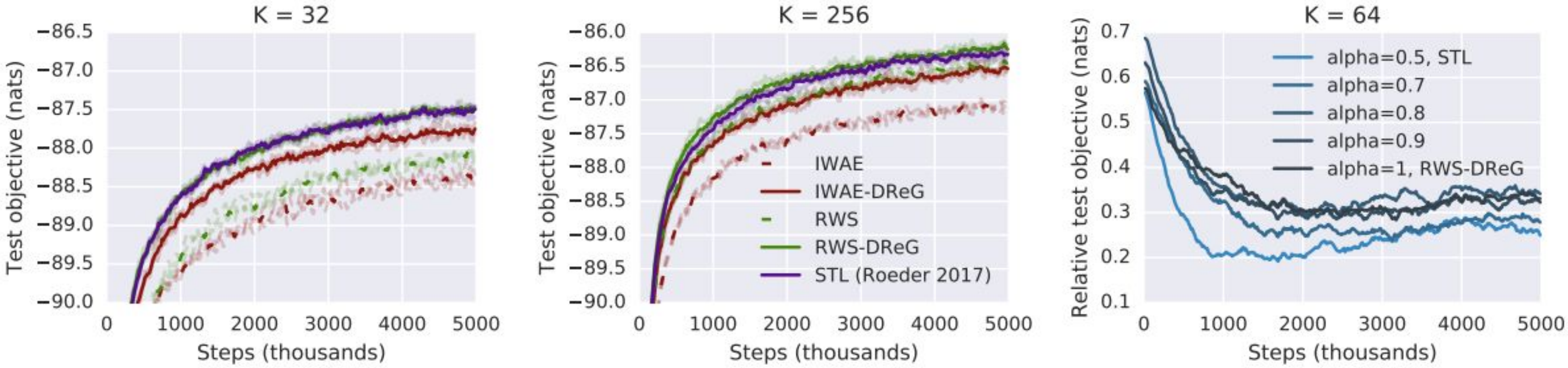
We plan to explore extensions to sequential models (e.g., Maddison et al. 2017, Naesseth et al. 2018, Le et al. 2018).

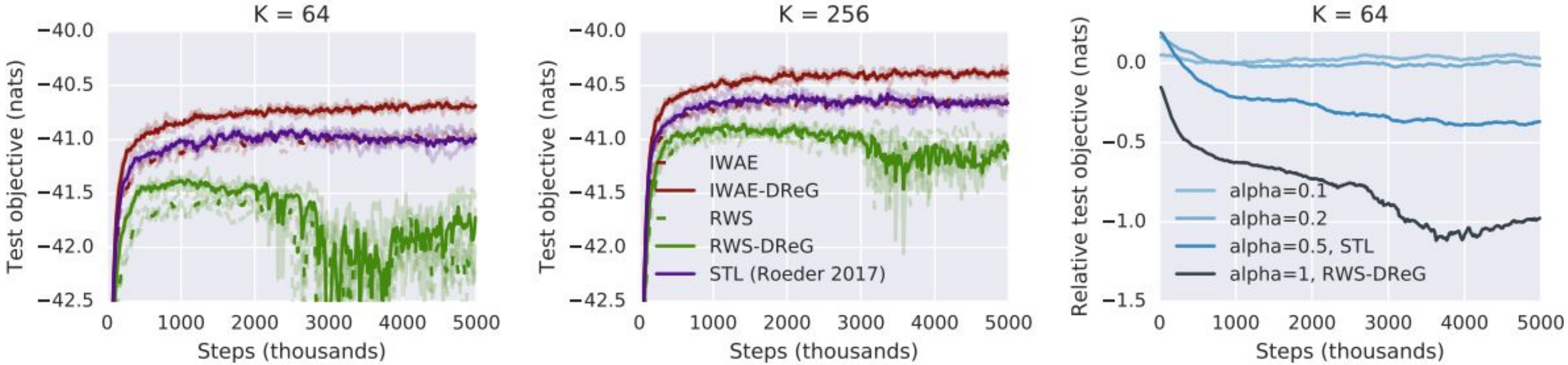
Paper, Slides, Code: sites.google.com/view/dregs



Appendix


IWAE
RWS
JVI
Gradient Variance (Relative)
0.46
0.44
0.42
0.40
0.38
0.7
0.6
0.5
0.4
0.3
0.2
0.56
0.54
0.52
0.50
0.48
0
1000
2000
3000
4000
5000
Steps (thousands)
K = 32
K = 64
K = 128
K = 256


K = 32
Test objective (nats)
−86.5
−87.0
−87.5
−88.0
−88.5
−89.0
−89.5
−90.0
0
1000
2000
3000
4000
5000
Steps (thousands)
K = 256
Test objective (nats)
−86.0
−86.5
−87.0
−87.5
−88.0
−88.5
−89.0
−89.5
−90.0
IWAE
IWAE-DReG
RWS
RWS-DReG
STL (Roeder 2017)
Steps (thousands)
K = 64
Relative test objective (nats)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
alpha=0.5, STL
alpha=0.7
alpha=0.8
alpha=0.9
alpha=1, RWS-DReG
Steps (thousands)

K = 64
Test objective (nats)
−40.0
−40.5
−41.0
−41.5
−42.0
−42.5
0
1000
2000
3000
4000
5000
Steps (thousands)
K = 256
Test objective (nats)
−40.0
−40.5
−41.0
−41.5
−42.0
−42.5
0
1000
2000
3000
4000
5000
Steps (thousands)
IWAE
IWAE-DReG
RWS
RWS-DReG
STL (Roeder 2017)
K = 64
Relative test objective (nats)
0.0
−0.5
−1.0
−1.5
0
1000
2000
3000
4000
5000
Steps (thousands)
alpha=0.1
alpha=0.2
alpha=0.5, STL
alpha=1, RWS-DReG