Distilling Intractable Generative Models

George Papamakarios
g.papamakarios@ed.ac.uk

Iain Murray
i.murray@ed.ac.uk

University of Edinburgh

Intractable generative models

$$p(\mathbf{x}) = \frac{1}{Z}\,\bar{p}(\mathbf{x}) \quad \text{with} \quad Z = \int \bar{p}(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

tractable	intractable
$\bar{p}(\mathbf{x})$	Z, $p(\mathbf{x})$

Why Z?

Bayesian inference

$$p(\mathbf{w} \mid D, M) = \frac{p(D \mid \mathbf{w}, M)\, p(\mathbf{w} \mid M)}{p(D \mid M)}$$

- $p(D \mid M)$ measures model fit
- it can be used for **model comparison**

Why Z?

Likelihood-based comparison

$$\log p(D \mid \mathbf{w}_1, M_1) > \log p(D \mid \mathbf{w}_2, M_2)$$

- best generative models typically intractable
  - RBM, DBN, VAE, GAN, LAPGAN, ...
- need Z to **compare likelihoods**

How to calculate Z?

- If we know $p(\mathbf{x})$, then we know $Z = \bar{p}(\mathbf{x})/p(\mathbf{x})$.
- If we can **estimate** $p(\mathbf{x})$, then we can **estimate** Z.
- **Idea**: distil intractable model to a tractable one.
- Distil $\Rightarrow$ train a flexible tractable model $q_\theta(\mathbf{x})$ to approximate $p(\mathbf{x})$ as closely as possible.

How to distil: loss functions

Loss function $E(\boldsymbol{\theta}) \Rightarrow$ measure disagreement between $p(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\mathbf{x})$.

KL divergence

$$E_{\mathrm{KL}}(\boldsymbol{\theta}) = D_{\mathrm{KL}}(p(\mathbf{x}) \,\|\, q_{\boldsymbol{\theta}}(\mathbf{x})) = -\left\langle \log q_{\boldsymbol{\theta}}(\mathbf{x}) \right\rangle_{p(\mathbf{x})} + \mathrm{const}$$

Square error

$$E_{\mathrm{SE}}(\boldsymbol{\theta}) = \left\langle \tfrac{1}{2} \left\| \log q_{\boldsymbol{\theta}}(\mathbf{x}) - \log p(\mathbf{x}) \right\|^2 \right\rangle_{p(\mathbf{x})}$$

$$E_{\mathrm{SE}}(\boldsymbol{\theta}) = \left\langle \tfrac{1}{2} \left\| \log q_{\boldsymbol{\theta}}(\mathbf{x}) - \log \bar{p}(\mathbf{x}) + c \right\|^2 \right\rangle_{p(\mathbf{x})} \text{ with } c \leq \log Z$$

How to distil: minimization

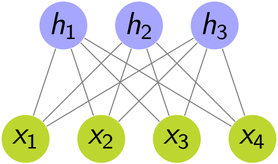
Both loss functions are of the form

$$E(\boldsymbol{\theta}) = \langle E(\mathbf{x}, \boldsymbol{\theta}) \rangle_{p(\mathbf{x})}$$

Minimize it stochastically:

1. MCMC $\Rightarrow$ sample $\{\mathbf{x}_s\}$ from $p(\mathbf{x})$
2. stochastic gradient $\frac{1}{S} \sum_s \frac{\partial}{\partial \boldsymbol{\theta}} E(\mathbf{x}_s, \boldsymbol{\theta})$
3. update $\boldsymbol{\theta}$
4. iterate

Case study: distilling an RBM into a NADE



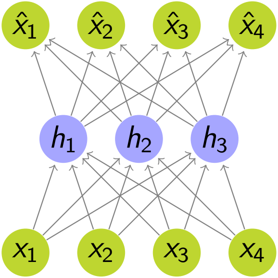
RBM

$$p(\mathbf{x}) = \frac{1}{Z} \prod_j \left(1 + \exp \mathbf{x}^T \mathbf{w}_j\right)$$

500 hidden units

trained on MNIST

$\Rightarrow$



NADE

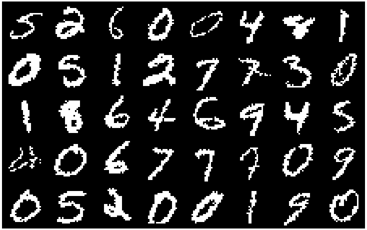
$$q_\theta(\mathbf{x}) = \prod_i \hat{x}_i$$

500 hidden units

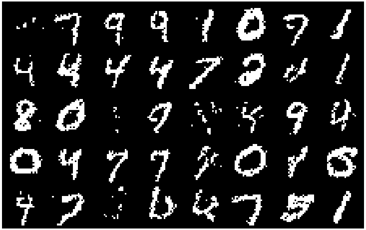
distilled from RBM

Distillation: results

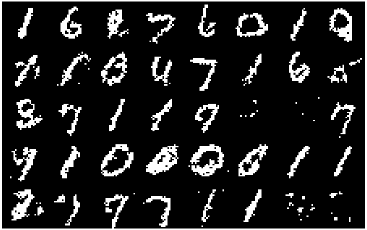
MNIST



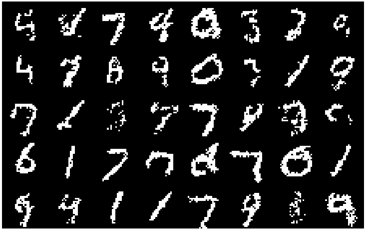
RBM



NADE (KL divergence)



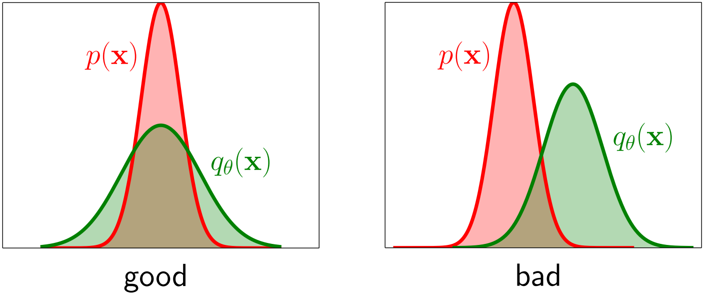
NADE (square error)



Estimating Z by simple Monte Carlo

Importance sampling

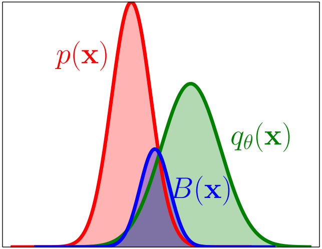
$$Z = \left\langle \frac{\bar{p}(\mathbf{x})}{q_{\boldsymbol{\theta}}(\mathbf{x})} \right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x})} \approx \frac{1}{S} \sum_{s} \frac{\bar{p}(\mathbf{x}_s)}{q_{\boldsymbol{\theta}}(\mathbf{x}_s)} \quad \text{with} \quad \mathbf{x}_s \sim q_{\boldsymbol{\theta}}(\mathbf{x}_s)$$



Estimating Z by simple Monte Carlo

Bridge sampling

- bridge distribution $B(\mathbf{x})$



- importance sampling estimates

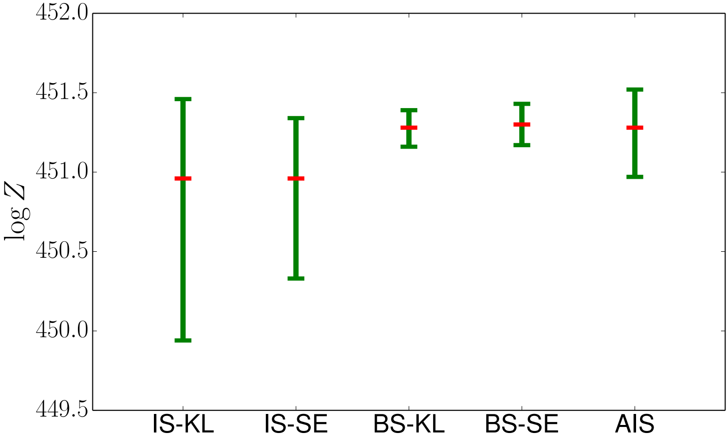
$$Z_1 = \left\langle \frac{B(\mathbf{x})}{q_\theta(\mathbf{x})} \right\rangle_{q_\theta(\mathbf{x})} \qquad Z_2 = \left\langle \frac{B(\mathbf{x})}{\bar{p}(\mathbf{x})} \right\rangle_{p(\mathbf{x})} \qquad Z = \frac{Z_1}{Z_2}$$

Estimating $\log Z$: results

Sampling method	Loss function	
	KL divergence	Square error
Importance sampling	450.96	450.96
Bridge sampling	451.28	451.30

cf. Annealed Importance Sampling: $\log Z \approx 451.28$.
(Salakhutdinov & Murray, 2008)

Estimating $\log Z$: results



To sum up: how to estimate Z

- Choose a flexible tractable model (such as NADE).
- Distil the intractable model into it.
- Use simple Monte Carlo (such as bridge sampling) with the tractable model as proposal.
- For more information: `http://arxiv.org/abs/1510.02437`

Appendix

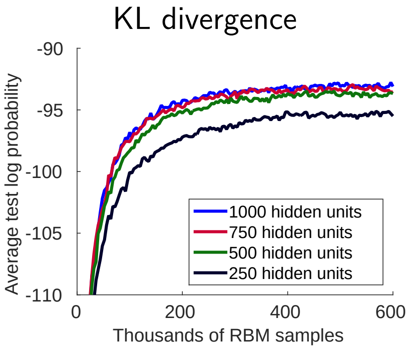
Why Z?

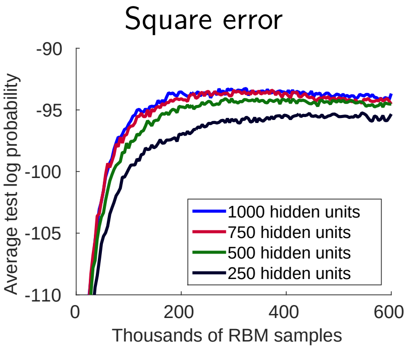
Graphical models

$$p(\mathbf{x} \mid \mathbf{w}) = \frac{1}{Z(\mathbf{w})} \exp(-U(\mathbf{x}, \mathbf{w}))$$

- typically defined by energy function $U(\mathbf{x}, \mathbf{w})$
- $Z(\mathbf{w})$ and its derivatives $\Rightarrow$ **useful information** about the model

Distillation: results





What next?

- Continuous distributions: RNADE.
- Score matching: match derivatives w.r.t. $\mathbf{x}$.
- Model distillation as an alternative to variational inference.