Variational Noise Contrastive Estimation — Supplementary Materials —

Benjamin Rhodes
University of Edinburgh
ben.rhodes@ed.ac.uk

Michael Gutmann
University of Edinburgh
michael.gutmann@ed.ac.uk

1 Convexity result for NCE lower bound

For non-negative real numbers a, b and u, the function

$$f(u) = \log(a + bu^{-1}) \tag{S1}$$

is convex. We see this by differentiating f twice:

$$f'(u) = -\frac{b}{au^2 + bu} \qquad f''(u) = \frac{b(2au + b)}{(au^2 + bu)^2}, \tag{S2}$$

and observing that $f''(u) \geq 0$ since a, b and u are non-negative.

2 Proof of Lemma 1

Key to this proof is the following factorisation

$$\phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = \phi_{\boldsymbol{\theta}}(\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}), \tag{S3}$$

where the conditional distribution is normalised and the factorisation holds because the unnormalised distributions on either side of the equation have the same partition function

$$\int \int \phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{x} = \int \phi_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{S4}$$

With this factorisation at hand, we now consider the difference between the NCE objective: $J_{\mathrm{NCE}}(\boldsymbol{\theta})$ in (4) and the VNCE objective: $J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q)$ in (13). Each objective consists of two terms: the first is an expectation with respect to the data, the second an expectation with respect to the noise distribution $p_{\mathbf{y}}$. The second terms of J_{NCE} and J_{VNCE} are identical, so their difference equals the

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

difference between their first terms

$$
\begin{aligned}
&J_{\mathrm{NCE}}(\boldsymbol{\theta}) - J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) \\
&= \mathbb{E}_{\mathbf{x}} \log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} \right) - \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) + \nu p_{\mathbf{y}}(\mathbf{x}) q(\mathbf{z} \mid \mathbf{x})} \right) && \text{(S5)} \\
&= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} \right) + \log \left(1 + \frac{\nu p_{\mathbf{y}}(\mathbf{x}) q(\mathbf{z} \mid \mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S6)} \\
&= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} + \frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} \frac{\nu p_{\mathbf{y}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x})} \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S7)} \\
&= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} + \frac{\nu p_{\mathbf{y}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S8)} \\
&= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} + \left(1 - \frac{\phi_{\boldsymbol{\theta}}(\mathbf{x})}{\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})} \right) \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S9)} \\
&= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\kappa_{\mathbf{x}} + (1 - \kappa_{\mathbf{x}}) \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S10)} \\
&= \mathbb{E}_{\mathbf{x}} \left[D_{f_{\mathbf{x}}}(p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z} \mid \mathbf{x})) \right], && \text{(S11)}
\end{aligned}
$$

where $f_{\mathbf{x}}(u) = \log(\kappa_{\mathbf{x}} + (1 - \kappa_{\mathbf{x}})u^{-1})$. To ensure that that $D_{f_{\mathbf{x}}}$ is a valid f-divergence, we need to prove that f is convex and $f_{\mathbf{x}}(1) = 0$. The latter is trivial, since $f_{\mathbf{x}}(1) = \log(\kappa_{\mathbf{x}} + (1 - \kappa_{\mathbf{x}})) = \log(1) = 0$, and convexity follows directly from Supplementary Materials 1.

We now prove that this f-divergence can be expressed as the difference of two KL-divergences as in (17) in the main text. To do this, we pull q/p outside of the log in (S10),

$$
\begin{aligned}
&D_{f_{\mathbf{x}}}(p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z} \mid \mathbf{x})) \\
&= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right] + \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\kappa_{\mathbf{x}} \frac{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z} \mid \mathbf{x})} + (1 - \kappa_{\mathbf{x}}) \right) \right] && \text{(S12)} \\
&= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \frac{q(\mathbf{z} \mid \mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})} \right] - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} \left[\log \left(\frac{q(\mathbf{z} \mid \mathbf{x})}{\kappa_{\mathbf{x}} p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) + (1 - \kappa_{\mathbf{x}}) q(\mathbf{z} \mid \mathbf{x})} \right) \right] && \text{(S13)} \\
&= D_{KL}(q(\mathbf{z} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})) - D_{KL}(q(\mathbf{z} \mid \mathbf{x}) \parallel m_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x})). && \text{(S14)}
\end{aligned}
$$

where $m_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x}) = \kappa_{\mathbf{x}} p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) + (1 - \kappa_{\mathbf{x}}) q(\mathbf{z} \mid \mathbf{x})$.

3 Proof of Theorem 1

We first show that

$$
J_{\mathrm{NCE}}(\boldsymbol{\theta}) = J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) \quad \Leftrightarrow \quad q(\mathbf{z} \mid \mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}). \tag{S15}
$$

We could obtain this result directly from the lower bound in Section 3.1 in the main text. However, for brevity, we make use of the Lemma 1, where we obtained the equality

$$
J_{\mathrm{NCE}}(\boldsymbol{\theta}) - J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{x}} \left[D_{f_{\mathbf{x}}}(p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z} \mid \mathbf{x})) \right]. \tag{S16}
$$

The f-divergence on the right-hand side is non-negative and equal to zero if and only if the two posteriors coincide. Hence, $J_{\mathrm{NCE}}(\boldsymbol{\theta}) = J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q)$ if and only if $q(\mathbf{z} \mid \mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$.

We now show that

$$
D_{f_{\mathbf{x}}}(p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z} \mid \mathbf{x})) \;\rightarrow\; D_{KL}(q(\mathbf{z} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})) \tag{S17}
$$

as $\kappa_{\mathbf{x}} = \phi_{\boldsymbol{\theta}}(\mathbf{x})/(\phi_{\boldsymbol{\theta}}(\mathbf{x}) + \nu p_{\mathbf{y}}(\mathbf{x})) \to 0$. Again, this follows quickly from Lemma 1. Specifically, in (S10), we obtained

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}) - J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\mid\mathbf{x})}\left[\log\left(\kappa_{\mathbf{x}} + (1-\kappa_{\mathbf{x}})\frac{q(\mathbf{z}\mid\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}\mid\mathbf{x})}\right)\right]. \tag{S18}$$

As $\kappa_{\mathbf{x}} \to 0$, we obtain the standard KL-divergence.

4 Proof of Theorem 2

Our goal is to show that

$$\max_{\boldsymbol{\theta}} J_{\mathrm{NCE}}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \max_{q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q). \tag{S19}$$

We know from Theorem 1 that:

$$p_{\boldsymbol{\theta}}(\mathbf{z}\mid\mathbf{x}) = \arg\max_{q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q), \tag{S20}$$

and that, plugging this optimal q into J_{VNCE} makes the variational lower bound tight,

$$J_{\mathrm{VNCE}}(\boldsymbol{\theta}, p_{\boldsymbol{\theta}}(\mathbf{z}\mid\mathbf{x})) = J_{\mathrm{NCE}}(\boldsymbol{\theta}). \tag{S21}$$

Hence,

$$\max_{\boldsymbol{\theta}} \max_{q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) = \max_{\boldsymbol{\theta}} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, p_{\boldsymbol{\theta}}(\mathbf{z}\mid\mathbf{x})) = \max_{\boldsymbol{\theta}} J_{\mathrm{NCE}}(\boldsymbol{\theta}). \tag{S22}$$

5 Proof of Corollary 1

Let $k \in \mathbb{N}$. After the E-step of optimisation, we have $q_k(\mathbf{z}\mid\mathbf{x}) = p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta}_k)$ and so, by Lemma 1,

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}_k) - J_{\mathrm{VNCE}}(\boldsymbol{\theta}_k, q_k) = \mathbb{E}_{\mathbf{x}}\left[D_{f_{\mathbf{x}}}(p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta}_k) \parallel p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta}_k))\right] = 0, \tag{S23}$$

implying that $J_{\mathrm{VNCE}}(\boldsymbol{\theta}_k, q_k) = J_{\mathrm{NCE}}(\boldsymbol{\theta}_k)$. Now, in the M-step of optimisation, we have

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q_k) \implies J_{\mathrm{VNCE}}(\boldsymbol{\theta}_{k+1}, q_k) \geq J_{\mathrm{VNCE}}(\boldsymbol{\theta}_k, q_k), \tag{S24}$$

finally, by using Lemma 1 again, we see that $J_{\mathrm{NCE}}(\boldsymbol{\theta}_{k+1}) \geq J_{\mathrm{VNCE}}(\boldsymbol{\theta}_{k+1}, q_k)$. Putting everything together,

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}_{k+1}) \geq J_{\mathrm{VNCE}}(\boldsymbol{\theta}_{k+1}, q_k) \geq J_{\mathrm{VNCE}}(\boldsymbol{\theta}_k, q_k) = J_{\mathrm{NCE}}(\boldsymbol{\theta}_k). \tag{S25}$$

6 Optimal proposal distribution in the second term of the VNCE objective

We know from Theorem 1 that the optimal variational distribution is the true posterior, $q(\mathbf{z}\mid\mathbf{y}) = p(\mathbf{z}\mid\mathbf{y};\boldsymbol{\theta})$. Thus, we simply need to show that the true posterior is the optimal proposal distribution for the importance sampling (IS) estimate in the second term of the VNCE objective.

As shown in Supplementary Materials 2, the following factorisation holds

$$\phi_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}) = \phi_{\boldsymbol{\theta}}(\mathbf{y}) p_{\boldsymbol{\theta}}(\mathbf{z}\mid\mathbf{y}). \tag{S26}$$

Using this factorisation of ϕ, we get

$$\phi(\mathbf{y};\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\mid\mathbf{y})}\left[\frac{\phi(\mathbf{y},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z}\mid\mathbf{y})}\right] \tag{S27}$$

$$= \phi(\mathbf{y};\boldsymbol{\theta})\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}\mid\mathbf{y})}\left[\frac{p(\mathbf{z}\mid\mathbf{y};\boldsymbol{\theta})}{q(\mathbf{z}\mid\mathbf{y})}\right]. \tag{S28}$$

Hence, the variance of a Monte Carlo estimate of the expectation in (S27) will equal the variance of a Monte Carlo estimate of the expectation in (S28). When $q(\mathbf{z}|\mathbf{y}) = p(\mathbf{z}|\mathbf{y};\boldsymbol{\theta})$, the latter expectation equals one, yielding a zero-variance—and thus optimal—Monte Carlo estimate.

We have therefore shown that the use of IS is optimal when we have access to $p(\mathbf{z}\mid\mathbf{y};\boldsymbol{\theta})$. More generally, it will still be sensible when we have access to a parameterised approximate posterior $q(\mathbf{z}\mid\mathbf{y};\boldsymbol{\alpha})$, which is close to the true posterior. However, one potential issue that could arise in practice is that q is only close to the true posterior when conditioning on data $\mathbf{x}$, but not when conditioning on noise samples $\mathbf{y}$. This is because we only optimise the parameters of q with respect to the first term of the VNCE objective, in which we only condition on data $\mathbf{x}$. In our experiments, we did not observe such an issue. However, we expect that if $\mathbf{z}$ is high-dimensional and the noise distribution is sufficiently different from the data distribution, then this could become an issue.

7 Experimental settings for toy approximate inference problem

In Section 4.1 we approximated a posterior $p(\mathbf{z} \mid \mathbf{x})$ with a variational distribution $q(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{z};\ \boldsymbol{\mu}(\mathbf{x}; \boldsymbol{\alpha}), \boldsymbol{\Sigma}(\mathbf{x}; \boldsymbol{\alpha}))$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are parametrised by a single 2-layer feed-forward neural network with weights $\boldsymbol{\alpha}$.

The output layer of the neural network has 4 dimensions, containing the concatenated vectors $\boldsymbol{\mu}$ and $\log(\mathrm{diag}(\boldsymbol{\Sigma}))$. The input to the network is a 2 dimensional vector $\mathbf{x}$ of observed data. In each hidden layer there are 100 hidden units, generated by an affine mapping composed with a $\tanh$ non-linearity applied to the previous layer. The weights of the network are initialised from $\mathcal{U}(-0.05, 0.05)$ and optimised with stochastic gradient ascent in minibatches of 100 and learning rate of 0.0001 for a total of 50 epochs.

8 Experimental settings for toy parameter estimation (Figure 4)

Figure 4 shows the accuracy of VNCE for parameter estimation using a population analysis over multiple sample sizes, comparing to NCE and MLE. To produce it, we generated 500 distinct ground-truth values for the standard deviation parameter in the unnormalised MoG, sampling uniformly from the interval $[2, 6]$. For each of the 500 sampled values of θ^*, we estimate θ using all three estimation methods and with a range of sample sizes. Every run was initialised from five random values and the best result out of the five was kept in order to avoid local optima which exist since both the likelihood and NCE objective functions are bi-modal.

9 Estimation of noise distribution for undirected graphical model experiments

Assume the observed data are organised in a matrix X with each column containing all observations of a single variable. We want to fit a univariate truncated Gaussian to each column. To do so, we could estimate the means μ_i and variances σ_i^2 of the *pre-truncated* Gaussians using the following equations (Burkardt, 2014), where x_i denotes a column of X with empirical mean $\bar{\mu}_i$ and variance $\bar{\sigma}_i^2$:

$$\bar{\mu_i} = \mu_i + \frac{\psi(\alpha)}{1 - \Phi(\alpha)}\sigma_i, \qquad \bar{\sigma_i^2} = \left[1 + \frac{\alpha\psi(\alpha)}{1 - \Phi(\alpha)} - \left(\frac{\psi(\alpha)}{1 - \Phi(\alpha)}\right)^2\right]\sigma_i^2, \tag{S29}$$

where ψ is the pdf of a standard normal and Φ is its cdf. These pairs on non-linear simultaneous equations can then be solved with a variety of methods, such as Newton-Krylov (Knoll and Keyes, 2004). However, whenever $\alpha = \frac{-\mu_i}{\sigma_i} \gg 0$, computing the fractions $\frac{\alpha\psi(\alpha)}{1-\Phi(\alpha)}$, $\frac{\psi(\alpha)}{1-\Phi(\alpha)}$ becomes numerically unstable. In a short note available on GitHub, Fernandez-de-cossio Diaz (2018) explains how to fix this using the more numerically stable scaled complementary error function $\mathrm{erfcx}(x) = \exp(x^2)\,\mathrm{erf}(x)$, where $\mathrm{erf}(x)$ is the error function. Introducing the notation

$$F_1(x) = \frac{1}{\mathrm{erfcx}(x)}, \qquad F_2(x) = \frac{x}{\mathrm{erfcx}(x)}, \tag{S30}$$

we can then re-express the required fractions in a numerically stable form,

$$\frac{\alpha\psi(\alpha)}{1 - \Phi(\alpha)} = \frac{2}{\sqrt{\pi}}F_2(\frac{\alpha}{\sqrt{2}}), \qquad \frac{\psi(\alpha)}{1 - \Phi(\alpha)} = \frac{2}{\sqrt{\pi}}F_2(\frac{\alpha}{\sqrt{2}}) - \frac{2}{\pi}\left[F_1(\frac{\alpha}{\sqrt{2}})\right]^2. \tag{S31}$$

10 Experimental settings for the undirected graphical model experiments

For VNCE and NCE we set $\nu = 10$, and optimise with the `BFGS` optimisation method of Python's `scipy.optimize.minimize`, capping the number of iterations at 80. In the case of VNCE, we use variational-EM, alternating every 5 iterations, and approximating expectations with respect to the variational distribution with 5 samples per datapoint. Derivatives with respect to the variational parameters are computed using the reparametrisation trick (Kingma and Welling, 2013; Rezende et al., 2014), using a standard normal as the base distribution.

For MC-MLE, we apply stochastic gradient ascent for 80 epochs with minibatches of 100 datapoints. The Monte-Carlo expectations with respect to the posterior distribution and joint distribution use 5 samples per datapoint. These samples are obtained with the `tmvtnorm` Gibbs sampler, using the Gibbs sampler from the `tmvtnorm` package in `R` with a burnin period of 100 samples and thinning factor of 10.

For VNCE and NCE, we do not enforce positive semi-definiteness of the matrix $\mathbf{K}$ in (28), in line with Lin et al. (2016). For MCMLE, we do enforce it, since `tmvtnorm` requires it.

References

Burkardt, J. (2014). The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*.

Fernandez-de-cossio Diaz, J. (2018). Moments of the univariate truncated normal distribution.

Kingma, D. P. and Welling, M. (2013). Stochastic gradient VB and the variational auto-encoder. *The 2nd International Conference on Learning Representations*.

Knoll, D. A. and Keyes, D. E. (2004). Jacobian-free Newton–Krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193(2):357–397.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10(1):806.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning*.