Natural-Gradient Stochastic Variational Inference for Non-Conjugate Structured Variational Autoencoder

Abstract

We propose a new variational inference method which uses recognition models for amortized inference in graphical models that contain deep generative models. Unlike many existing approaches, our method can handle non-conjugacy in both the latent graphical model and the deep generative model, and enables fully amortized inference at test time. Our method is based on an extension of a recently proposed mirror-descent algorithm and employs natural-gradient updates for all three components of the model, i.e. the latent graphical model, the deep generative model, and the recognition model. We also propose structured recognition models to capture posterior correlations among local latent variables. We show that our method has computational advantages over existing approaches in two classes of non-conjugate models, namely, latent mixture models and nonlinear state-space models. An additional advantage of our method is that it can be implemented by reusing existing software for graphical models and deep models.

1. Introduction

In this paper, we develop a new amortized inference method for graphical models that contain deep generative models. Such models merge two important lines of work, namely deep learning and probabilistic inference. Several works have recently proposed these types of models (Archer et al., 2015; Krishnan et al., 2015; Johnson et al., 2016). The first two of these works have considered modeling of time-series data with neural networks, while Johnson et al. (2016) propose to compose a general class of conjugate latent graphical models with neural networks and call it structured variational auto-encoders (SVAE). Inspired by recent works on variational inference and deep learning, Johnson et al. (2016) derive an inference scheme that combines ideas from message passing, stochastic variational inference (SVI), and back-propagation using the reparamaterization trick.

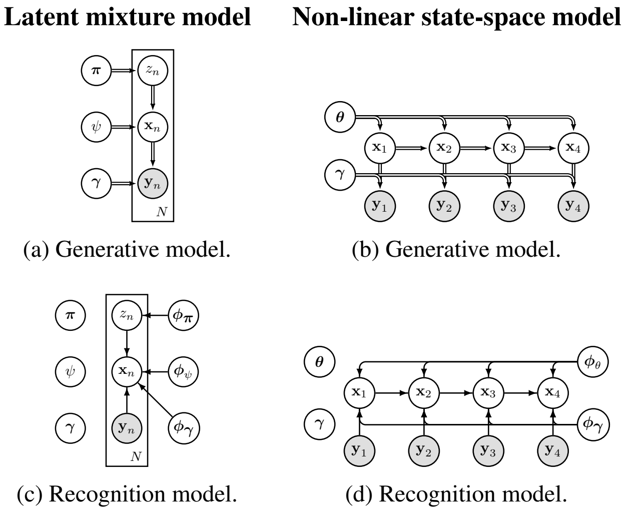
There are several issues with the work of Johnson et al. (2016). The first issue is that their method requires a strong conditional-conjugacy structure in the latent graphical model. The second issue is that their method only performs natural-gradient updates for some global latent variables. The third issue is that their method does not support amortized inference and therefore needs to run inference over some of the local variables at test time. The final issue is that their method does not model posterior correlations among all the local latent variables.

In this paper, we propose a method to solve these issues and generalize the method of Johnson et al. (2016) to a larger class of models. Instead of using message passing or standard SVI, we use a more general method called Conjugate-Computation Variational Inference (CVI) (Khan & Lin, 2017). We extend this method to handle recognition models by showing that its updates can be expressed as an adaptive-gradient method. This results in a natural-gradient method that does not require conditional-conjugacy, solving the first two issues. We propose structured recognition models for local variables and show that amortized inference can be performed while preserving the correlation between all local variables. This solves the third and fourth issue. Finally, our method can be implemented by reusing existing software for graphical models and deep models.

2. Models and Related work

In this section, we describe the generative model as well as our variational approximations. Figure 1 gives two examples of the model classes. We consider models that employ at most two layers of *local* latent variables to model N observed outputs $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$. The first layer contains continuous variables denoted by $\mathbf{x}_n$ and the second layer contains finite-discrete variables z_n. We denote by $\mathbf{x}$ and $\mathbf{z}$ the sets of these two local variables for all n. We assume that the relationship between $\mathbf{x}$ and $\mathbf{z}$ is specified by using a graphical model denoted by $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the set of all global variables. Following previous works (Johnson et al., 2016; Archer et al., 2015; Krishnan et al., 2015), we model $\mathbf{y}$ given $\mathbf{x}$ using a neural-network likelihood with $\boldsymbol{\gamma}$ being the neural-network parameters:

$$p(\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \underbrace{\Big[p(\boldsymbol{\theta})p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\Big]}_{\text{Latent graphical model}} \underbrace{\Big[p(\boldsymbol{\gamma}) \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\gamma})\Big]}_{\text{Deep generative model}}$$



(a) Generative model. (b) Generative model.

(c) Recognition model. (d) Recognition model.

Figure 1: Two examples of model classes. The left column shows the latent mixture model and the right column shows the non-linear state-space model. The top row shows the model while the bottom row shows the recognition model. Double lines indicate non-conjugate relationship, while the solid line indicate conjugate ones. In our framework, all distributions in the generative model can be non-conjugate, but the recognition model is conjugate. Our recognition models preserve structural dependencies among the local variables that are present in the original model.

We assume all factors of the above distribution to be minimal exponential family distributions. This model class is a more general than the one considered by Johnson et al. (2016) because we do *not* require the joint distribution $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ to be conditionally-conjugate, i.e., $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$, $p(\mathbf{z}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ need not be conditionally conjugate.

Our goal is to estimate an approximation to the posterior distribution of the $\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. In this paper, we propose the following variational approximation with a *structured* recognition model:

$$p(\mathbf{z}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}) \approx q_{\phi}(\mathbf{x}, \mathbf{z}|\mathbf{y})q(\boldsymbol{\theta})q(\boldsymbol{\gamma}) \qquad (1)$$

The recognition model facilitates fast *amortized* inference which reduces computations at test time by avoiding inference over the local variables. Our proposal in this paper is slightly more general than that of Johnson et al. (2016) who consider $q_{\phi}(\mathbf{x}, \mathbf{z}|\mathbf{y}) = q_{\phi}(\mathbf{x}|\mathbf{y})q(\mathbf{z})$, i.e., an amortize inference is used for $\mathbf{x}$, but $q(\mathbf{z})$ still need to be inferred at test time. Our structured-recognition models maintain local structure present in the model, i.e., *the structure among the local variables in the variational approximation is the same as that in the generative model*. Moreover, our recognition models can re-use existing conjugate factors in the generative model.

Denoting the natural parameters of $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\gamma})$ specified in (1) by $\boldsymbol{\lambda}_{\theta}$ and $\boldsymbol{\lambda}_{\gamma}$ respectively (and $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}_{\gamma}, \boldsymbol{\lambda}_{\theta}\}$), the lower bound to be optimized is given as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}) &:= \mathbb{E}_q \log\left[p(\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})/q(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})\right] \\ &= \sum_{n=1}^{N} \mathbb{E}_{q_{\phi}(x,z|y)q(\gamma)}\left[\log p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\gamma})\right] \\ &\quad - \mathbb{E}_{q(\theta)}\left\{\mathbb{D}_{KL}[q_{\phi}(\mathbf{x}, \mathbf{z}|\mathbf{y}) \,\|\, p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]\right\} \\ &\quad - \mathbb{D}_{KL}[q(\boldsymbol{\theta})q(\boldsymbol{\gamma}) \,\|\, p(\boldsymbol{\theta})p(\boldsymbol{\gamma})] \qquad (2)\end{aligned}$$

Below, we give two examples of various graphical models where our work is applicable. In all these examples, we only specify $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ and assume that $\mathbf{y}$ is modeled using a neural network.

Latent Mixture Models: For finite mixture models with K mixture components, we assume a discrete assignment vector $\mathbf{z}_n$ whose k'th element is $z_{n,k} \in \{0, 1\}$ for every vector $\mathbf{x}_n$. The distribution is shown below:

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\psi}, \boldsymbol{\pi}) = \prod_{n=1}^{N}\left[\prod_{k=1}^{K}[\pi_k p(\mathbf{x}_n|\boldsymbol{\psi}_k)]^{z_{nk}}\right], \qquad (3)$$

$$\{\boldsymbol{\psi}_k, \pi_k\}_{k=1}^{K} \sim p(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = \{\boldsymbol{\psi}_k, \pi_k\}_{k=1}^{K} \qquad (4)$$

where $\boldsymbol{\psi}_k$ is the parameter of the distribution of the k'th mixture and π_k is the mixture proportion which sums to 1. The Gaussian mixture model is a member of this family, but we also can handle non-conjugate mixture models.

Nonlinear State-Space Models: Consider the following state-space model (Kokkala et al., 2015)

$$\mathbf{x}_n = \mathbf{f}(\mathbf{x}_{n-1}, \boldsymbol{\theta}) + \mathbf{q}_n, \quad \mathbf{q}_n \sim \mathcal{N}(\mathbf{q}_n|0, \mathbf{Q}(\boldsymbol{\theta})) \qquad (5)$$

where $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_0|\mathbf{m}_0(\boldsymbol{\theta}), \mathbf{P}_0(\boldsymbol{\theta}))$ and each $\mathbf{x}_n$ is Gaussian with its mean being a non-linear function $f(\mathbf{x}_{n-1})$ and $\boldsymbol{\theta}$ are model parameters distributed according to $p(\boldsymbol{\theta})$. The prior need not be conjugate to $\mathbf{x}$. Archer et al. (2015) propose a variational inference method for these types of models, but use a Gaussian posterior distribution with tri-diagonal covariance which is quite restricted. Johnson et al. (2016) consider a linear state-space model with a conjugate prior $p(\boldsymbol{\theta})$ and our work generalizes their work to nonlinear state-transitions. Krishnan et al. (2015) use RNN to model the transitions and the likelihoods and their recognition model is a bit more general than ours.

3. Main Contributions

We build upon the conjugate-computation variational inference (CVI) method of Khan & Lin (2017). This method can handle non-conjugate factors, but it does not work with recognition models. Our first contribution is to extend CVI to handle recognition models and our second contribution is to propose a structured-recognition model to simplify inference. We first give an overview of CVI and then describe our two contributions in the subsequent sections.

3.1. Conjugate-computation Variational Inference

The CVI method is based on a mirror-descent formulation of the variational lower bound optimization in the mean-parameter space. This reformulation enables natural-gradient updates for the mean-field variational inference in general non-conjugate graphical models. Stochastic variational inference (SVI) and variational message passing (VMP) can be obtained as special cases when the model is conditionally-conjugate.

We give a summary of the method below. Given a Bayesian network over N nodes $\mathbf{u}_i$, we wish to obtain the mean-field approximation $q(\mathbf{u})$, where each component $q_i(\mathbf{u}_i)$ is a minimal exponential family (denote its natural parameter by $\boldsymbol{\lambda}_i$ and mean parameter, which is the expectation of sufficient statistics, by $\boldsymbol{\mu}_i$). CVI assumes that conditional distributions of $\mathbf{u}_i$ given the rest of the nodes $\mathbf{u}_{/i}$ can be expressed as a product between a factor that is conjugate to $q(\mathbf{u}_i)$ (denoted by $\tilde{p}_c^i$) and a factor that is non-conjugate (denoted by $\tilde{p}_{nc}^i$), i.e.,

$$p(\mathbf{u}_i|\mathbf{u}_{/i}) \propto \; \tilde{p}_c^i(\mathbf{u}_i, \mathbf{u}_{/i}) \times \tilde{p}_{nc}^i(\mathbf{u}_i, \mathbf{u}_{/i}). \qquad (6)$$

CVI employs the following mirror-descent update in the mean-parameter space:

$$\boldsymbol{\mu}_{i,t+1} = \arg\max_{\mu_i} \langle \boldsymbol{\mu}_i, \widehat{\nabla}_{\mu_i}\mathcal{L}(\boldsymbol{\mu}_t)\rangle - \frac{1}{\beta_t}\mathbb{B}_{A^*}(\boldsymbol{\mu}_i \| \boldsymbol{\mu}_{i,t}), \qquad (7)$$

where $\mathcal{L}(\boldsymbol{\mu})$ is a reparameterization of the lower bound $\mathcal{L}$ in terms of the mean parameter, $A^*(\boldsymbol{\mu})$ is the convex-conjugate of the log-partition function of $q(\mathbf{u}_i)$, $\mathbb{B}_{A^*}$ is the Bregman divergence defined by A^*, $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{i,t}$ denote the values of $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_i$, respectively, at iteration t, and $\beta_t > 0$ is the step-size.

An important feature of the above update is that there is a closed-form solution. This update separately performs conjugate and non-conjugate computation as shown below:

$$\boldsymbol{\lambda}_{i,t+1} = (1-\beta_t)\boldsymbol{\lambda}_{i,t} + \beta_t\left[\boldsymbol{\lambda}^*_{i,t+1} + \widehat{\nabla}_{\mu_i}\mathbb{E}_q(\log \tilde{p}^i_{nc})|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t}\right], \qquad (8)$$

where $\boldsymbol{\lambda}^*_{i,t+1}$ is the mean-field update obtained using only the conjugate term $\tilde{p}_c^i$. The above update is a natural-gradient update which exploits the geometry of the variational distribution. Therefore, CVI is a generalization of SVI to non-conjugate models.

3.2. Contribution 1: CVI as an adaptive-gradient method

CVI does not directly apply to the estimation of deterministic parameters such as the parameters of the recognition model. In this section, we present a framework that enables such application of CVI. Under our framework, CVI updates can be expressed as an adaptive-gradient method, very similar to methods such as AdaGrad and RMSprop. The following claim summarizes our results.

Claim 1. *Defining a Gaussian variational distribution for $\boldsymbol{\phi}$ at iteration t as $q_t(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\phi}_t, \mathbf{S}_t^{-1})$, the CVI update, shown below, approaches a local maximum of $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda})$ for a fixed $\boldsymbol{\lambda}$.*

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \beta_t \mathbf{S}_{t+1}^{-1}\hat{\mathbf{g}}_t, \qquad (9)$$

where $\mathbf{S}_{t+1} = \mathbf{S}_t - \beta_t\hat{\mathbf{H}}_t$ with $\hat{\mathbf{g}}_t$ and $\hat{\mathbf{H}}_t$ being the sample approximations to the average gradient $\mathbb{E}_{q(\phi)}\left[\nabla_\phi\mathcal{L}\right]$ and Hessian $\mathbb{E}_{q(\phi)}[\nabla^2_{\phi\phi}\mathcal{L}]$ at $q_t(\boldsymbol{\phi})$.

The update is obtained by optimizing an expectation of the lower bound $\mathbb{E}_{q(\phi)}[\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda})]$, followed by some reparameterization tricks. The proof of convergence to a local maximum of $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda})$ is obtained by using a result given in Appendix A of (Maaløe et al., 2016). These updates are very similar to existing adaptive-gradient methods and we can establish a connection by approximating the Hessian by the following diagonal approximation (Martens, 2014): $\nabla^2_{\phi\phi}\mathcal{L} \approx -\text{diag}(\hat{\mathbf{g}}^2)$. The method most similar to ours is AROW (Crammer et al., 2009) which was originally proposed for supervised online-learning with a hinge loss. If we use $\mathbf{S}_t^{1/2}$ instead of $\mathbf{S}_t$ in the update of $\boldsymbol{\phi}_t$, our update becomes equivalent to a noisy version of AdaGrad (Duchi et al., 2011). We can also show that, by using a different choice of the posterior $q(\boldsymbol{\phi})$, CVI updates arrive arbitrary close to the updates of RMSprop. Finally, $q(\boldsymbol{\phi})$ can be non-Gaussian, which could be used for exploiting conjugacy in recognition models.

3.3. Contribution 2: Structured-Recognition Models and Natural-Gradient Updates

In this section, we describe how to construct structured-recognition models that preserve the structural dependencies of the true posterior distribution of the local variables. We also show how to obtain efficient natural gradient updates that reuse existing software for implementation. Due to space limitations, we describe our method on one particular example of latent mixture model, although our method is more generally applicable.

We choose a recognition model $q_\phi(\mathbf{x}, \mathbf{z}|\mathbf{y})$ that consists of three conditional distributions where the first and second term are conjugate to each other with respect to $\mathbf{x}_n$:

$$q_\phi(\mathbf{x}_n, \mathbf{z}_n|\mathbf{y}_n) \propto q_{\phi_\gamma}(\mathbf{x}_n|\mathbf{y}_n) q_{\phi_\psi}(\mathbf{x}_n|\mathbf{z}_n) q_{\phi_\pi}(\mathbf{z}_n), \quad (10)$$

where $\boldsymbol{\phi}_\gamma, \boldsymbol{\phi}_\psi$, and $\boldsymbol{\phi}_\pi$ are the parameters of recognition models that mimic the role of $\boldsymbol{\gamma}, \boldsymbol{\psi}$, and $\boldsymbol{\pi}$ in the latent MM. This can also be seen in Figure 1. The three terms need to be chosen such that computing the marginal $q_\phi(\mathbf{z}_n|\mathbf{y}_n)$ and sampling the conditional $q_\phi(\mathbf{x}_n, \mathbf{z}_n|\mathbf{y}_n)$ is easy.

One possible way to construct such recognition models is by choosing q_ϕ to be a conditionally-conjugate model, e.g., the following distribution, a GMM with a Gaussian observation, can be used as $q_\phi(\mathbf{x}_n, \mathbf{z}_n|\mathbf{y}_n)$:

$$\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_{\phi_\gamma}(\mathbf{y}_n), \boldsymbol{\Sigma}_{\phi_\gamma}(\mathbf{y}_n)\right)\mathcal{N}\left(\mathbf{x}_n|\bar{\boldsymbol{\mu}}_{z_n}, \bar{\boldsymbol{\Sigma}}_{z_n}\right)\mathcal{M}(\mathbf{z}_n;\bar{\boldsymbol{\pi}}),$$

where the second and third term constitute a GMM with parameters $\boldsymbol{\phi}_\psi = \{\bar{\boldsymbol{\mu}}_{1:K}, \bar{\boldsymbol{\Sigma}}_{1:K}\}$ and $\boldsymbol{\phi}_\pi = \bar{\pi}_{1:K}$, while the first term is a recognition model similar to VAE and corresponds to a Gaussian measurement $\boldsymbol{\mu}_{\phi_\gamma}(\mathbf{y}_n)$ with mean $\mathbf{x}_n$ and covariance $\boldsymbol{\Sigma}_{\phi_\gamma}(\mathbf{y}_n)$. Since this recognition model is conjugate, we can easily sample from it and also compute the marginal $q_\phi(\mathbf{z}_n|\mathbf{y}_n)$. In this case these steps can be implemented using the E-step in a GMM (we skip the details due to space constraints). In general, these steps can be performed by reusing inference on a conditionally-conjugate model.

Given the marginal $q_\phi(\mathbf{z}_n|\mathbf{y}_n)$ and samples $\mathbf{x}_n^*$ from $q_\phi(\mathbf{x}_n, \mathbf{z}_n|\mathbf{y}_n)$, updates of the global variables are simplified and can be implemented using a combination of the CVI update of (8) and (9). Figure 1c shows all the global variables in the variational approximation. For the global variables of the neural networks, i.e., $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}_\gamma$, the update (9) is very similar to that of a VAE with only one difference – the prior distribution over $\mathbf{x}_n$ is a mixture instead of a single Gaussian prior. For the global variables of the mixture model, i.e., $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$, the update (8) can be used. This update simplifies if part of the model is conjugate, e.g., when the latent graphical model is a latent GMM, the update can be implemented using variational Bayes updates (the M-step). Finally, the update for the global variables of the recognition model, i.e., $\boldsymbol{\phi}_\psi$ and $\boldsymbol{\phi}_\pi$, can be obtained using (9) where the gradients are computed using the computation graph of the recognition model. This is also easy since the recognition model is conjugate.

4. Discussion

We proposed a new variational inference method that uses structured-recognition model for local variables. Our method simplifies inference by pushing all the difficult computation to the recognition model. By choosing a conditionally-conjugate recognition model, we simplify the difficult computation which ultimately reduces to sampling from the recognition model. This sampling plays a role very similar to the E-step in the EM algorithm or a local variable update in SVI. Given samples from the recognition model, we greatly simplify the update of global variables. For example, if part of the model is conditionally-conjugate, then mean-field implementation can be used to perform natural-gradient updates (which is essentially a step in SVI). For the non-conjugate parts, we use stochastic gradients. Moreover, for deterministic parameters, we recover the adaptive-gradient like updates.

We also established that CVI can be used as an adaptive-gradient method, thereby connecting natural-gradients to adaptive-gradients. We discussed application to a specific type of local variable structure. It is possible that our method is useful for more general structures. For example, this work could be extended to capture dependencies between global and local variables in graphical models. It is also possible to obtain natural-gradient updates for $\boldsymbol{\phi}$ by imposing a distribution on $\boldsymbol{\phi}$ that takes the same form as $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

References

Archer, Evan, Park, Il Memming, Buesing, Lars, Cunningham, John, and Paninski, Liam. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*, 2015.

Crammer, Koby, Kulesza, Alex, and Dredze, Mark. Adaptive regularization of weight vectors. In *Advances in neural information processing systems*, pp. 414–422, 2009.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Johnson, Matthew, Duvenaud, David K, Wiltschko, Alex, Adams, Ryan P, and Datta, Sandeep R. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, pp. 2946–2954, 2016.

Khan, Mohammad Emtiyaz and Lin, Wu. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*, 2017.

Kokkala, Juho, Solin, Arno, and Särkkä, Simo. Sigma-point filtering and smoothing based parameter estimation in nonlinear dynamic systems. *arXiv preprint arXiv:1504.06173*, 2015.

Krishnan, Rahul G, Shalit, Uri, and Sontag, David. Deep Kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

Maaløe, Lars, Sønderby, Casper Kaae, Sønderby, Søren Kaae, and Winther, Ole. Auxiliary deep generative models. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1445–1453, 2016.

Martens, James. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.