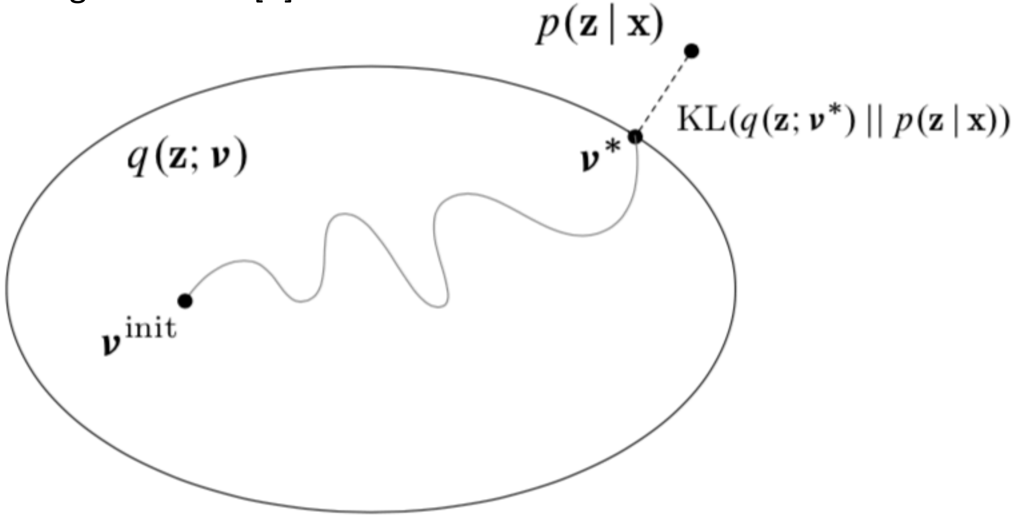
# Composing graphical models with neural networks for structured representations and fast inference

A Presentation by Tushar Agarwal

# Variational Inference

Image from Ref [1]




A probabilistic model is usually a joint pdf of the form $p(x, y)$ where $x, y$ refer to hidden and visible variables respectively.

- Inference about $x$ given $y$, $p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\int_x \, p(x,y)dx}$. The denominator of this posterior is an intractable integration/summation in many interesting due to curse of dimensionality and useful cases and so we resort to approximate inference.
- Converts this inference into optimization problem. Similar to the idea of subspace projection for eg. Finding a Linear model given data.
- We choose a distribution from a family of distributions $q(z; \nu)$ with some desired properties (for eg. Exponential family, conjugate priors, fully factorable joint distributions etc.).
- Then find the optimal variational parameters $\nu$ that minimize the KL divergence (a distance metric indicating how similar 2 pdfs are) between the true and assumed distributions.

$$KL + ELBO = \ln p(y) = \text{constant given } y$$

Due to this, $\min(\mathrm{KL}) \equiv \max(ELBO)$ and ELBO has $p(x, y)$ instead of the intractable $p(x|y)$

The running theme would be to:

Assume a family of distributions with nice properties and make them fit the distribution of real data by minimizing KL Divergence.

# Section 2.2

Content from Ref [2]

$x = \{x_n\}_{n=1}^N$, and data $y = \{y_n\}_{n=1}^N$,

$$p(\theta, x, y) = p(\theta) \prod_{n=1}^{N} p(x_n|\theta) p(y_n|x_n, \theta), \quad (1)$$

Always true

where $p(\theta)$ is the natural exponential family conjugate prior to the exponential family $p(x_n, y_n|\theta)$,

$$\ln p(\theta) = \langle \eta_\theta, t_\theta(\theta) \rangle - \ln Z_\theta(\eta_\theta) \quad (2)$$

$$\ln p(x_n, y_n|\theta) = \langle \eta_{xy}(\theta), t_{xy}(x_n, y_n) \rangle - \ln Z_{xy}(\eta_{xy}(\theta)) = \langle t_\theta(\theta), (t_{xy}(x_n, y_n), 1) \rangle. \quad (3)$$

$t_{xy}$ denotes sufficient statistic. Prop B.4 pg 15, v5. This part assumes that TRUE distributions also belong to exp. family.

Figure 2 shows the graphical model. The mean field variational inference problem is to approximate the posterior $p(\theta, x \,|\, y)$ with a tractable distribution $q(\theta, x)$ by finding a local minimum of the KL divergence $\mathrm{KL}(q(\theta, x) \| p(\theta, x \,|\, y))$ or, equivalently, using the identity

$$\ln p(y) = \mathrm{KL}(q(\theta, x) \| p(\theta, x \,|\, y)) + \mathbb{E}_{q(\theta,x)}\left[\ln \frac{p(\theta, x, y)}{q(\theta, x)}\right],$$

Same as before

to choose $q(\theta, x)$ to maximize the objective

$$\mathcal{L}[q(\theta, x)] \triangleq \mathbb{E}_{q(\theta,x)}\left[\ln \frac{p(\theta, x, y)}{q(\theta, x)}\right] \leq \ln p(y). \quad (4)$$

ELBO

Consider the mean field family $q(\theta)q(x) = q(\theta)\prod_n q(x_n)$. Because of the conjugate exponential family structure, the optimal global mean field factor $q(\theta)$ is in the same family as the prior $p(\theta)$,

**Idea 1**: Fully factorable mean-field family. Helps separate maximization over arguments.

Refer to Pg 15, v5

$$\ln q(\theta) = \langle \widetilde{\eta}_\theta, t_\theta(\theta) \rangle - \ln Z_\theta(\widetilde{\eta}_\theta). \quad (5)$$

**Conditionally conjugate models**

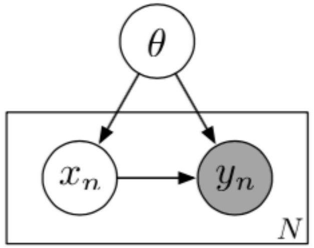



*Figure 2.* Prototypical graphical model for SVI.

The mean field objective on the global variational parameters $\widetilde{\eta}_\theta$, optimizing out the local variational factors $q(x)$, can then be written

Separable maximization: An advantage of mean field assumption.

$$\mathcal{L}(\widetilde{\eta}_\theta) \triangleq \max_{q(x)} \mathbb{E}_{q(\theta)q(x)}\left[\ln \frac{p(\theta, x, y)}{q(\theta)q(x)}\right] \leq \ln p(y) \quad (6)$$

and the natural gradient of the objective (6) decomposes into a sum of local expected sufficient statistics (Hoffman et al., 2013):

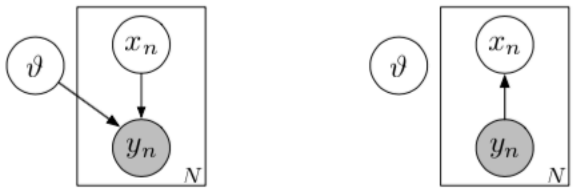
A curvature corrected version of gradient. Refer to Pg 18, v5, Section C.2

$$\widetilde{\nabla}_{\widetilde{\eta}_\theta} \mathcal{L}(\widetilde{\eta}_\theta) = \eta_\theta + \sum_{n=1}^{N} \mathbb{E}_{q^*(x_n)}(t_{xy}(x_n, y_n), 1) - \widetilde{\eta}_\theta, \quad (7)$$

where $q^*(x_n)$ is a locally optimal local mean field factor given $\widetilde{\eta}_\theta$. Thus we can compute a stochastic natural gradient update for our global mean field objective by sampling a data minibatch $y_n$, optimizing the local mean field factor $q(x_n)$, and computing scaled expected sufficient statistics.

# Section 2.3

Content from Ref [2]




(a) VAE generative model. (b) VAE variational family.

a) Achieved by decoder
b) Achieved by encoder

*Figure 3.* Graphical models for the variational autoencoder.

$$x_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \quad n = 1, 2, \ldots, N \tag{8}$$

$$y_n \mid x_n, \vartheta \sim \mathcal{N}(\mu(x_n; \vartheta), \Sigma(x_n; \vartheta)) \tag{9}$$

$$(\mu(x_n; \vartheta), \Sigma(x_n; \vartheta)) = \text{MLP}(x_n; \vartheta). \tag{14}$$

To approximate the posterior $p(\vartheta, x \mid y)$, the variational autoencoder uses the mean field family

$$q(\vartheta)q(x \mid y) = q(\vartheta) \prod_{n=1}^{N} q(x_n \mid y_n). \tag{15}$$

Another tractable family. Extension of Mean-field assumption to induce correlation as much as possible.

A key insight of the variational autoencoder is to use a conditional variational density $q(x_n \mid y_n)$, where the parameters of the variational distribution on $x_n$ depend on the corresponding data point $y_n$. In particular, we can take the mean and covariance parameters of $q(x_n \mid y_n)$ to be $\mu(y_n; \phi)$ and $\Sigma(y_n; \phi)$, respectively, where

$$(\mu(y_n; \phi), \Sigma(y_n; \phi)) = \text{MLP}(y_n; \phi) \tag{16}$$

**Idea 2**: Amortized Inference.

**Idea 3**: Reparameterization trick to make the objective differentiable with respect to $x_n$

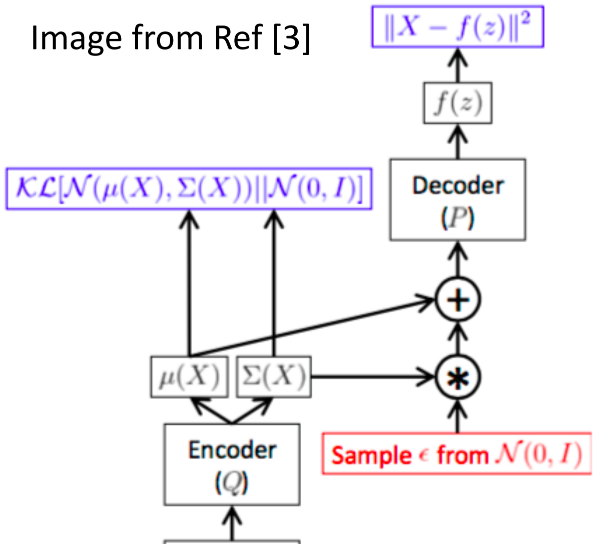
Encoder or Recognition Model.

$$\mathcal{L}(\vartheta^*, \phi) = \mathbb{E}_{q(x \mid y)} \ln p(y \mid x, \vartheta^*) - \text{KL}(q(x \mid y) \,\|\, p(x)).$$

Simplified Objective Function.

**Idea 4**: Switching order of gradient and expectation.

Image from Ref [3]




can be computed in closed form. To compute stochastic gradients of the expectation term, since a random variable $x_n \sim q(x_n \mid y_n)$ can be parameterized as

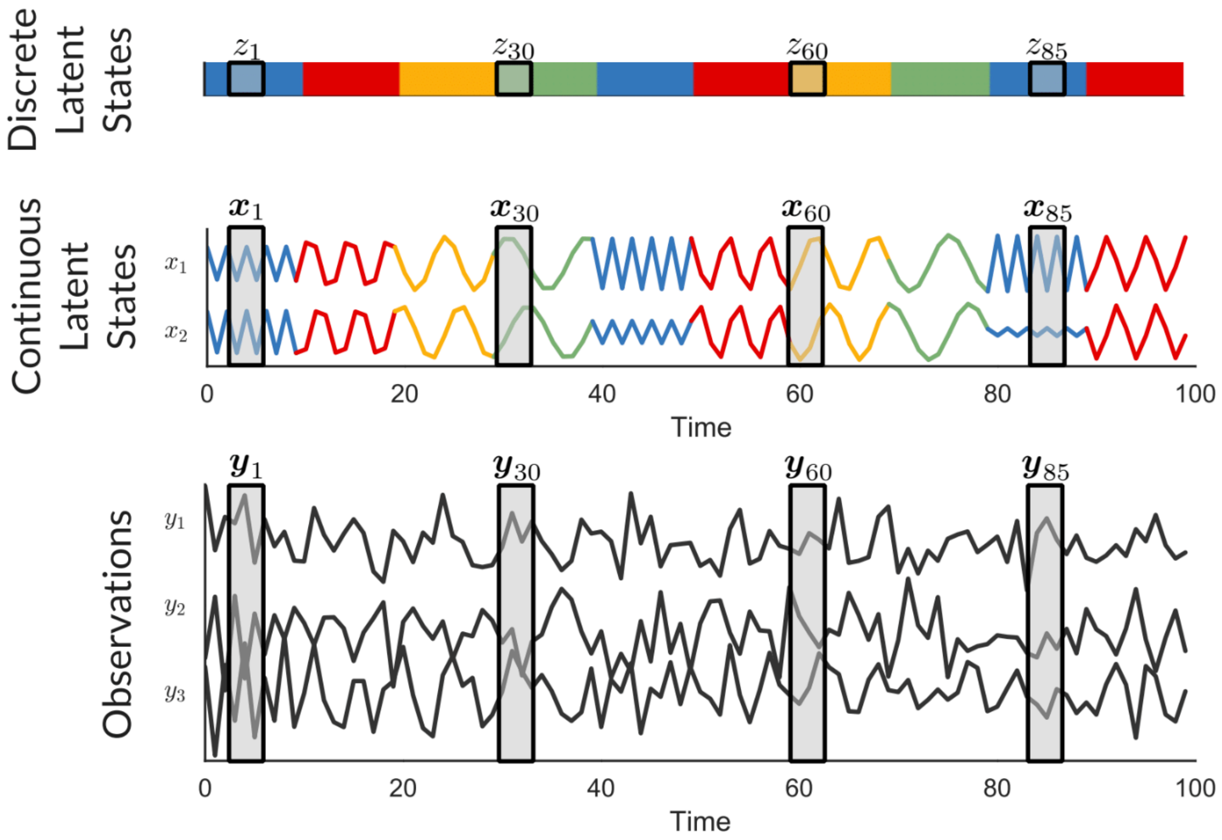
$$x_n = g(\phi, \epsilon) \triangleq \mu_q(y_n; \phi) + \Sigma_q(y_n; \phi)^{\frac{1}{2}} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

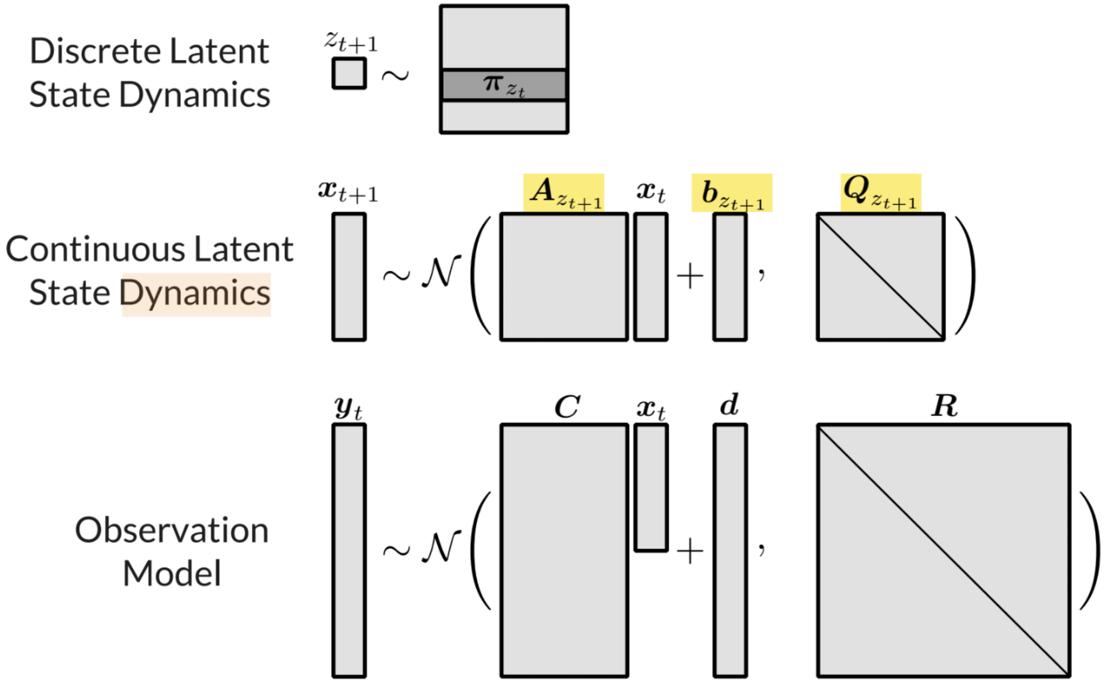
the expectation term can be rewritten in terms of $g(\phi, \epsilon)$ and its gradient approximated via Monte Carlo over $\epsilon$,

$$\nabla_{\vartheta^*, \phi} \mathbb{E}_q \ln p(y \mid x, \vartheta^*) \approx \sum_{n=1}^{N} \nabla_{\vartheta^*, \phi} \ln p(y_n \mid g(\phi, \hat{\epsilon}_n), \vartheta^*)$$

where $\hat{\epsilon}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$. Because $g(\phi, \epsilon)$ is a differentiable function of $\phi$, these gradients can be computed using standard backpropagation. For scalability, the sum over data points is also approximated via Monte Carlo. General non-

# Switching Linear Dynamical Systems

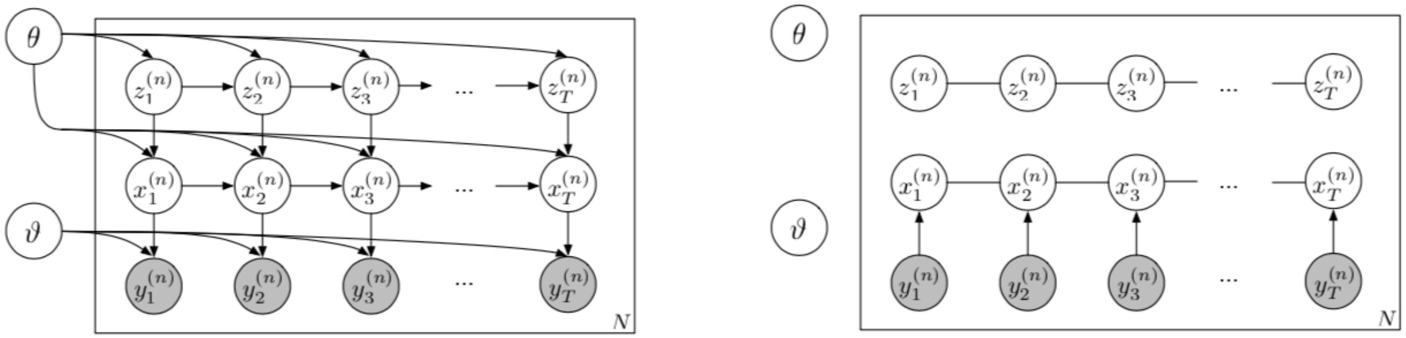



Images from Ref [4]

# Section 3

Content from Ref [2]



(a) SLDS generative model with nonlinear observation model parameterized by $\vartheta$.

(b) Structured CRF variational family with node potentials $\{\psi(x_t^{(n)}; y_t^{(n)}, \phi)\}_{t=1}^{T}$ parameterized by $\phi$.

*Figure 4.* Graphical models for the SLDS generative model and corresponding structured CRF variational family.

$$x_{t+1} = A^{(z_t)} x_t + B^{(z_t)} u_t, \quad u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{17}$$

where $A^{(k)}, B^{(k)} \in \mathbb{R}^{M \times M}$ for $k = 1, 2, \ldots, K$. The discrete latent state $z_t$ evolves according to Markov dynamics,

$$z_{t+1} \mid z_t, \pi \sim \pi^{(z_t)} \tag{18}$$

$$z_1 \mid \pi_{\text{init}} \sim \pi_{\text{init}}, \tag{19}$$

$$x_1 \mid z_1, \mu_{\text{init}}, \Sigma_{\text{init}} \sim \mathcal{N}(\mu_{\text{init}}^{(z_1)}, \Sigma_{\text{init}}^{(z_1)}). \tag{20}$$

$$\theta = (\pi, \pi_{\text{init}}, \{(A^{(k)}, B^{(k)}, \mu_{\text{init}}^{(k)}, \Sigma_{\text{init}}^{(k)})\}_{k=1}^{K}).$$

$$y_t \mid x_t, \vartheta \sim \mathcal{N}(\mu(x_t; \vartheta), \Sigma(x_t; \vartheta)). \tag{21}$$

$$(\mu(x_t; \vartheta), \Sigma(x_t; \vartheta)) = \text{MLP}(x_t; \vartheta). \tag{23}$$

$$q(\theta, \vartheta, z_{1:T}, x_{1:T}) = q(\theta) q(\vartheta) q(z_{1:T}) q(x_{1:T}). \tag{26}$$

To leverage bottom-up inference networks, we parameterize the factor $q(x_{1:T})$ as a conditional random field (CRF) (Murphy, 2012). That is, using the fact that the optimal factor $q(x_{1:T})$ is Markov according to a chain graph, we write

it terms of pairwise potentials and node potentials as

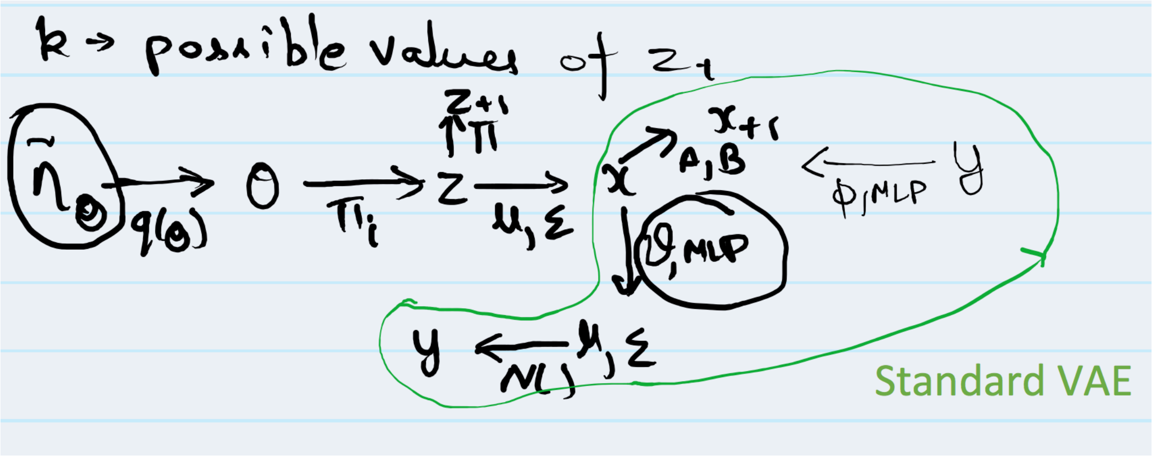
$$q(x_{1:T}) \propto \left( \prod_{t=1}^{T-1} \psi(x_t, x_{t+1}) \right) \left( \prod_{t=1}^{T} \psi(x_t; y_t, \phi) \right) \tag{27}$$

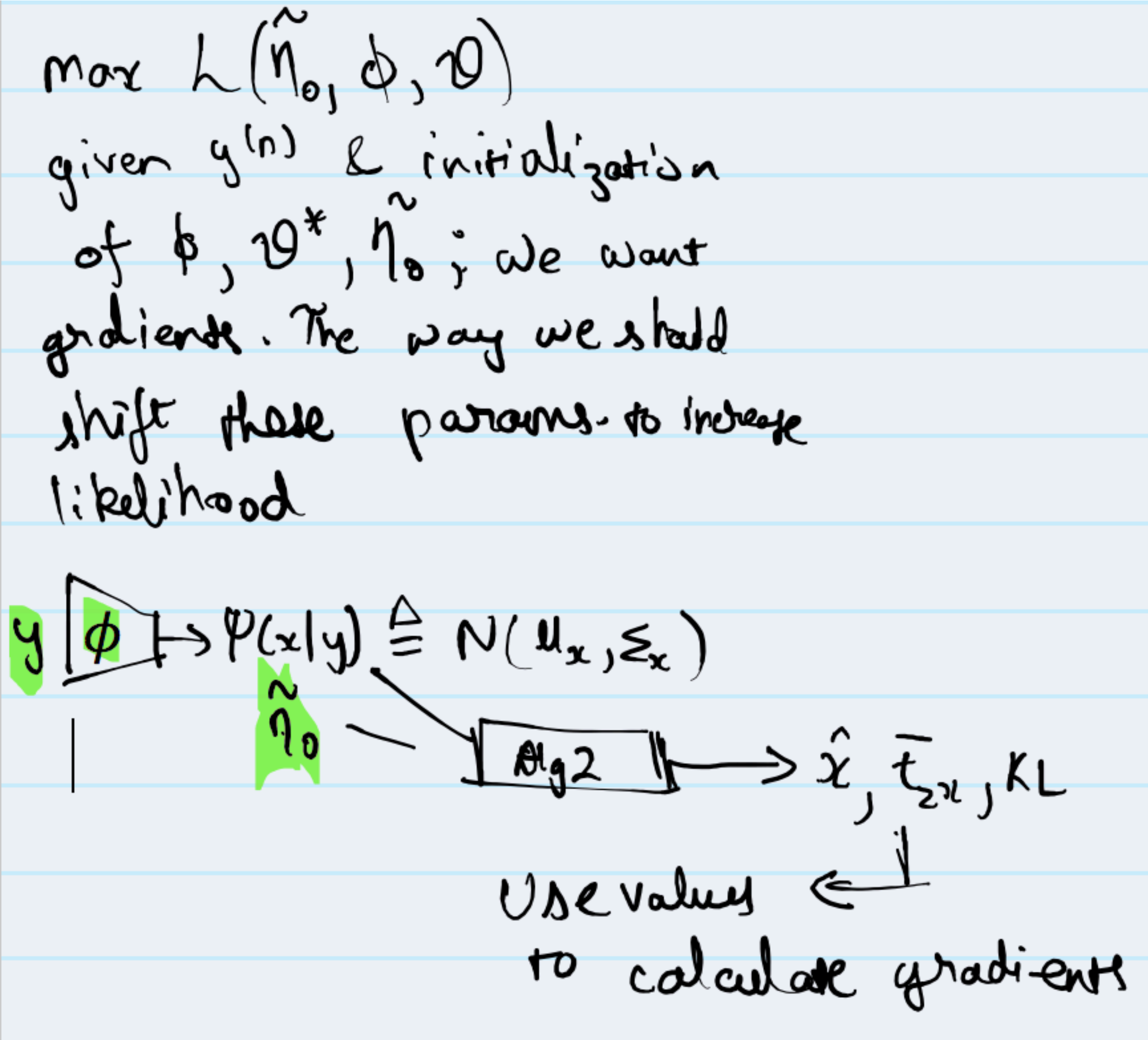
$$\psi(x_t; y_t, \phi) \propto \exp\left\{ -\frac{1}{2} x_t^{\mathsf{T}} J(y_t; \phi) x_t + h(y_t; \phi)^{\mathsf{T}} x_t \right\},$$

$$(h(y_t; \phi), J(y_t; \phi)) = \text{MLP}(y_t; \phi), \tag{28}$$

Notice that the conditional $\psi(x\_t \mid y\_t, \phi)$ is written in information form to allow for relationships between parameters of conditionals and conditioning rv's to be simple

# Workflow


k → possible values of z_t
q(θ)
Φ, MLP
θ, MLP
A, B
Standard VAE


max L(η̃₀, φ, ϑ)
given y(n) & initialization
of φ, ϑ*, η̃₀ ; we want
gradients. The way we should
shift these params. to increase
likelihood
Alg 2
KL
Use values
to calculate gradients

# Section 4

Content from Ref [2]

**Algorithm 1** Computing gradients of the SVAE objective

**Input:** Variational dynamics parameter $\widetilde{\eta}_\theta$ of $q(\theta)$, observation model parameter $\vartheta^*$, recognition network parameters $\phi$, sampled sequence $y^{(n)}$

**function** SVAEGRADIENTS($\widetilde{\eta}_\theta$, $\vartheta^*$, $\phi$, $y^{(n)}$)

$\psi \leftarrow$ RECOGNIZE$(y^{(n)}, \phi)$

$(\hat{x}^{(n)}(\phi),\ \bar{t}_{zx}^{(n)},\ \mathrm{KL}(\phi)) \leftarrow$ INFERENCE$(\widetilde{\eta}_\theta, \psi)$

$\widetilde{\nabla}_{\widetilde{\eta}_\theta}\mathcal{L} \leftarrow \eta_\theta + N(\bar{t}_{zx}^{(n)}, 1) - \widetilde{\eta}_\theta$

$\nabla_{\vartheta^*,\phi}\mathcal{L} \leftarrow \nabla_{\vartheta^*,\phi}\left[N \ln p(y^{(n)} \mid \hat{x}^{(n)}(\phi), \vartheta^*) - \mathrm{KL}(\phi)\right]$

**return**natural gradient $\widetilde{\nabla}_{\widetilde{\eta}_\theta}\mathcal{L}$, gradient $\nabla_{\vartheta^*,\phi}\mathcal{L}$

**end function**

The SVAE algorithm computes a natural gradient with respect to $\widetilde{\eta}_\theta$ and standard gradients with respect to $\vartheta^*$ and $\phi$. To compute these gradients, as in Section 2.2 we split the objective $\mathcal{L}(\widetilde{\eta}_\theta, \vartheta^*, \phi)$ as

$$\mathbb{E}_{q(x)} \ln p(y \mid x, \vartheta^*) - \mathrm{KL}(q(\theta, z, x) \,\|\, p(\theta, z, x)). \quad (31)$$

$$\hat{x}^{(n)}(\phi) = g(\phi, \epsilon) \triangleq J^{-1}(\phi)h(\phi) + J(\phi)^{-\frac{1}{2}}\epsilon \quad (34)$$

where $\epsilon \sim \mathcal{N}(0, I)$, $h(\phi) \in \mathbb{R}^{TM}$, and $J(\phi) \in \mathbb{R}^{TM \times TM}$. The matrix $J(\phi)$ is a block tridiagonal matrix corresponding to the Gaussian LDS of (27), the block diagonal of which depends on $\phi$. Since $g(\phi, \epsilon)$ is differentiable with

**Idea 3**: Reparameterization trick to make the objective differentiable with respect to $x_n$

**Algorithm 2** Model inference subroutine for the SLDS

**Input:** Variational dynamics parameter $\widetilde{\eta}_\theta$ of $q(\theta)$, node potentials $\{\psi(x_t; y_t)\}_{t=1}^T$ from recognition network

**function** INFERENCE($\widetilde{\eta}_\theta$, $\{\psi(x_t; y_t)\}_{t=1}^T$)

Initialize factor $q(x)$

**repeat**

$q(z) \propto \exp\{\mathbb{E}_{q(\theta)q(x)} \ln p(z, x \mid \theta)\}$

$q(x) \propto \exp\{\mathbb{E}_{q(\theta)q(z)} \ln p(x \mid z, \theta)\} \prod_t \psi(x_t; y_t)$

**until** $q(z)$ and $q(x)$ converge

$\hat{x} \leftarrow$ sample $q(x)$

$\bar{t}_{zx} \leftarrow \mathbb{E}_{q(z)q(x)} t_{zx}(z, x)$

$\mathrm{KL} \leftarrow \mathrm{KL}(q(\theta) \,\|\, p(\theta))$
$\quad + N\mathbb{E}_{q(\theta)}\, \mathrm{KL}(q(z)q(x) \,\|\, p(z, x \mid \theta))$

**return**sample $\hat{x}$, expected stats $\bar{t}_{zx}$, divergence KL

**end function**

$$\widetilde{\nabla}_{\widetilde{\eta}_\theta}\mathcal{L} = \eta_\theta + \sum_{n=1}^{N} \mathbb{E}_{q(z)q(x)}(t_{zx}(z^{(n)}, x^{(n)}), 1) - \widetilde{\eta}_\theta \quad (32)$$

where $q(z)$ and $q(x)$ are taken to be locally optimal local mean field factors as in Eq. (7). Therefore by sampling the sequence index $n$ uniformly at random, an unbiased estimate of the natural gradient is given by

$$\widetilde{\nabla}_{\widetilde{\eta}_\theta}\mathcal{L} \approx \eta_\theta + N\mathbb{E}_{q(z)q(x)}(t_{zx}(z^{(n)}, x^{(n)}), 1) - \widetilde{\eta}_\theta. \quad (33)$$

# References


1. David Blei, Rajesh Ranganath, Shakir Mohamed. NIPS 2016 Tutorial · December 5, 2016
2. Johnson, M., Duvenaud, D.K., Wiltschko, A., Adams, R.P. and Datta, S.R., 2016. Composing graphical models with neural networks for structured representations and fast inference. In Advances in neural information processing systems (pp. 2946-2954).
3. Doersch, C., 2016. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908.
4. Scott W. Linderman, Bayesian Time Series Analysis with Recurrent Switching Linear Dynamical Systems. April 12, 2016