Inference in Nonlinear Dynamical Systems
PhD Candidacy Exam

Antonio Moretti

June 5, 2017
Two Photon Calcium Imaging of Mouse Hippocampus

Deep

Superficial

2 min of activity at 5x speed
State Space Models

\begin{align}
\{Z_t\}_{t\geq 0} & \quad \{X_t\}_{t\geq 0} \\
Z_t | (Z_{0:t-1}) & \sim f_\theta(z_t | z_{t-1}) \\
X_t | (Z_{0:t}, X_{0:t-1}) & \sim g_\theta(x_t | z_t)
\end{align}

\[f_\theta\]

\[g_\theta\]

\begin{array}{cccc}
Z_0 & \quad \rightarrow & \quad Z_1 & \quad \rightarrow & \quad Z_2 & \quad \rightarrow & \quad \ldots \\
\downarrow & & \downarrow & & \downarrow & & \\
X_0 & & X_1 & & X_2 & & \ldots
\end{array}
Bayesian Inference

Posterior:

\[ p(z|x, \theta) = \frac{p(x, z|\theta)}{p(x|\theta)} = \frac{p(x|z, \theta)p(z|\theta)}{\int p(x, z|\theta)dz} \] (4)
Bayesian Inference

Posterior:

\[ p(z|x, \theta) = \frac{p(x, z|\theta)}{p(x|\theta)} = \frac{p(x|z, \theta)p(z|\theta)}{\int p(x, z|\theta)dz} \] 

(5)

Two approaches:

- MCMC
- VI

Can VI outperform MCMC in (nonlinear) dynamical systems?
Inference Challenges

Likelihood:

\[ p_\theta(x_{0:t}) = \int p_\theta(z_{0:t}, x_{0:t}) dz_{0:t} \quad (6) \]

Joint:

\[ p_\theta(z_{0:t}, x_{0:t}) = \mu_\theta(z_0) \prod_{k=1}^{t} f_\theta(z_k | z_{k-1}) \prod_{k=0}^{t} g_\theta(x_k | z_k) \quad (7) \]

Posterior:

\[ p(\theta|x_{0:t}) = \frac{p_\theta(x_{0:t})p(\theta)}{\int p_\theta(x_{0:t})p(\theta)d\theta} \quad (8) \]
Two Key Recursions

Likelihood and posterior satisfy the recursions:

\[
p_\theta(z_{0:t}|x_{0:t}) = p_\theta(z_{0:t-1}|x_{0:t-1}) \frac{f_\theta(z_t|z_{t-1})g_\theta(x_t|z_t)}{p_\theta(x_t|x_{0:t-1})} \tag{9}\]

and

\[
p_\theta(x_t|x_{0:t-1}) = \int f_\theta(z_t|z_{t-1})g_\theta(x_t|z_t)p_\theta(z_{t-1}|x_{0:t-1})\,dz_{t-1:t} \tag{10}\]

Particle filters are algorithms for numerical approximation.
Importance function $q_\theta$:

$$q_\theta(z_t, x_t|z_{t-1}) = \underbrace{q_\theta(z_t|x_t, z_{t-1})}_{\text{easy dist}} \cdot q_\theta(x_t|z_{t-1})$$  (11)

$q_\theta(x_t|z_{t-1})$ is non-negative on the same support as $\mathcal{Z} \times \mathcal{X}$.

$$w_0(z_0) = \frac{g_\theta(x_0|z_0) \mu_\theta(z_0)}{q_\theta(z_0|x_0)}$$  (12)

and

$$w_t(z_{t-1:t}) = \frac{g_\theta(x_t|z_t) f_\theta(z_t|z_{t-1})}{q_\theta(z_t, x_t|z_{t-1})}$$  (13)

Weights $w_t$ compensate for sampling from $q_\theta$ instead of $p_\theta$. 

Auxiliary Particle Filtering (Pitt, 1998)
Auxiliary Particle Filtering (Pitt, 1998)

Auxiliary particle filter (Pitt and Shepard, 1998) includes many special cases of particle algorithms.

- Bootstrap particle filter (Gordon, 1993)
  \[ q_\theta(z_t|x_t, z_{t-1}) = f_\theta(z_t|z_{t-1}) \]  

- Sequential Importance Sampling Resampling (Doucet, 2000)
  \[ q_\theta(x_t|z_{t-1}) = 1 \]

Choose importance function \( q_\theta \) to get the above.
Poisson Linear Dynamical System

\[ z_t = A z_{t-1} + \Sigma^{1/2} \epsilon \]
\[ x_t = Pois \left( \exp(B z_t) \right) \]

\[ z \sim N(A z_{t-1}, \Sigma^{1/2}) \]  \hspace{1cm} (16)
\[ x \sim Pois \left( \exp(B z_t) \right) \]  \hspace{1cm} (17)
Path Measures
PLDS: Overcoming Nonlinear Measurement Models
Particle Quadrature

Nodes are roots of Hermite polynomial $H_n(x)$, weights $w_i$ given:

$$\int_{-\infty}^{\infty} \exp(-x^2)f(x)dx \approx \sum_{i=1}^{n} w_if(x_i) \quad w_i = \frac{2^{n-1}n!\sqrt{\pi}}{n^2[H_{n-1}(x_i)^2]} \quad (18)$$

<table>
<thead>
<tr>
<th>Function $f(x, y)$</th>
<th>Quadrature Integration</th>
<th>Monte Carlo</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{100}(y - x^2)$</td>
<td>$0.1$</td>
<td>$0.0894$</td>
<td>$2.367 \times 10^{-5}$</td>
</tr>
<tr>
<td>$x^2 + y^2$</td>
<td>$7.0$</td>
<td>$6.99779$</td>
<td>$0.00220$</td>
</tr>
<tr>
<td>$\exp(-x^2)$</td>
<td>$0.41368$</td>
<td>$0.4139$</td>
<td>$0.00026$</td>
</tr>
<tr>
<td>$\cos(x) + \cos(y)$</td>
<td>$0.07530$</td>
<td>$0.0754$</td>
<td>$0.00015$</td>
</tr>
<tr>
<td>$1 - x^2 + \frac{1}{100}(y - x^2)^2$</td>
<td>$1$</td>
<td>$0.9985$</td>
<td>$0.0014$</td>
</tr>
</tbody>
</table>
Feynman Kac Flow

$F(z_1, \cdots, z_n)$ is function bounded on set of paths up to time.

$$\int dz_0 \int dz_1 \cdots \int dz_n F(z_0, \cdots z_n) p(z_0, \cdots, z_n | x_0, \cdots, x_n)$$

We’re approximating a functional integral over the set of paths:

$$\propto \int_{z(t) \in \mathbb{R}^n} F[z(t)] \prod_{k=0}^{n} p(x_k | z_k) p(z_0, \cdots, z_n) \, dz(t) \quad (19)$$
Feynman Kac Path Measure

Let $G_k(z_k) = p(x_k|z_k)$. We’re computing an expectation:

$$
\mathbb{E}\left(F(Z_0, \ldots, Z_n) \prod_{k=0}^{n} G_k(Z_k)\right)
\frac{\mathbb{E}\left(\prod_{k=0}^{n} G_k(Z_k)\right)}{
\mathbb{E}\left(\prod_{k=0}^{n} G_k(Z_k)\right)}
$$

(20)

$Z_n$ is an MC with trans prob $M_n$ on $E_n$ (seq of measurable spaces).

$$
d\mathcal{Q}_n = \frac{1}{Z_n} \left\{ \prod_{0 \leq p < n} G_p(Z_p) \right\} d\mathbb{P}_n
$$

(21)

$G_n$ is a sequence of potential functions on $E_n$, $\mathcal{Q}_n$ is a prob measure on pairs $(M_n, G_n)$ and $\mathbb{P}_n$ is a distribution on paths of $Z_p$. 
Markov Chain Monte Carlo

- Random Walk Metropolis

\[ T(\theta, \theta') = N(\theta' | \theta, \sigma^2) \times \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right) \]

- Gibbs Sampler

\[ T(\theta, \theta') = \prod_i \pi(\theta_i | \theta_{\setminus i}) \]

- Metropolis Hastings

\[ \alpha(y|x) = \min\left[\frac{q(x|y)\pi(y)}{q(y|x)\pi(x)}, 1\right] \]

Standard MCMC is inefficient and requires parameterization.
Concentration of Measure

Typical set?
Concentration of Measure

Annulus of width $\varepsilon = \mathcal{O}(r/d)$ near boundary

$$\frac{Vol((1 - \varepsilon)A)}{Vol(A)} = (1 - \varepsilon)^d \leq e^{-\varepsilon d}$$ (22)
Hamiltonian Dynamics

- $H$ is function of position $q$ and momenta $p$ or $z = (q, p)$:

$$H(q, p) = U(q) + K(p)$$  \hspace{1cm} (23)

- EoM:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$  \hspace{1cm} (24)

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$  \hspace{1cm} (25)

- Deterministic mapping $T_s$ from state at time $t$ to $t + s$

$$\frac{dz}{dt} = \begin{bmatrix} 0_{d \times d} & I_{d \times d} \\ -I_{d \times d} & 0_{d \times d} \end{bmatrix} \nabla H(z)$$  \hspace{1cm} (26)
Illustrating Hamiltonian Flow

Quadratic kinetic and potential energies in one dimension:

\[ H(q, p) = \frac{q^2}{2} + \frac{p^2}{2} \]  \hspace{1cm} (27)

Then

\[ \frac{dq}{dt} = p \quad \frac{dp}{dt} = -q \]  \hspace{1cm} (28)

Solving:

\[ q(t) = r \cos(a + t) \quad p(t) = -r \sin(a + t) \]  \hspace{1cm} (29)

Quadratic potential gives standard Gaussian distribution for \( q \).
Hybrid (Hamiltonian) Monte Carlo (Duane, 1987)

Sample deterministic paths for coherent exploration of state space.

- Vector field defined by mapping \((q, p) \rightarrow (q, p, \dot{q}, \dot{p})\)

\[
\begin{align*}
\dot{q} &= p \\
\dot{p} &= -q
\end{align*}
\]

- Analogy with Hamiltonian mechanics via momenta \(p = mv\).
Hamiltonian Flow (Duane, 1987)

Perform measure preserving mapping:

\[ \rho(p, q) \rightarrow e^{-H(p, q)} d^n p d^n q \]  
(30)

Markov transition:

\[ T(q', q) = \pi(p) \delta((p', q') - \phi_T(p, q)) \]  
(31)

Take log:

\[ H(p, q) = -\log \pi(p, q) \]  
(32)

\[ = - \underbrace{\log \pi(p|q)}_{\text{Kinetic Energy } T} - \underbrace{\log \pi(q)}_{\text{Potential Energy } U} \]  
(33)

KE is the conditional distribution and PE is the target distribution. Must integrate to get \( \pi(q) \).
Euclidean HMC (Duane, 1987)

\[ H(p, q) = -\log \pi(p, q) \]
\[ = -\log \pi(p|q) - \log \pi(q) \]

\[ \begin{align*}
\text{Kinetic Energy } T & = \frac{1}{2} p_i p_j (M^{-1})_{ij} \quad \text{Potential Energy } U \\
& = \frac{1}{2} p^T M^{-1} p
\end{align*} \]

- Quadratic T with constant metrics gives dynamics of Euclidean manifold.

\[ \pi(p|q) = \mathcal{N}(0, M) \]

- When kinetic energy has a quadratic form \( \Rightarrow \) Euclidean HMC.
Hamiltonian constant ⇒ \[ \Delta U = \Delta T = d/2. \]

- Keep quadratic KE but allow position dependent covariance metric \( M(q) \):

  \[
  \pi(p|q) = \mathcal{N}(0, M(q))
  \]
  \[
  T = \frac{1}{2} p_i p_j (M^{-1}(q))^{ij} + \frac{1}{2} \log |M(q)|
  \]

- Inverse covariance matrix corrects curvature (Fisher information) emulating dynamics of a Riemannian manifold.

- Normalization acts like a capacitor.
  - High curvature ⇔ absorb energy.
  - Low curvature ⇔ release energy.
Difficulties of Hamiltonian Monte Carlo

HMC is difficult to tune correctly.

\[
\frac{dq}{dt} = M^{-1}p \tag{40}
\]

\[
\frac{dp}{dt} = -\frac{\partial U}{\partial q} \tag{41}
\]

\[
q \rightarrow q + \epsilon M^{-1}p \tag{42}
\]

\[
p \rightarrow p - \epsilon \frac{\partial U}{\partial q} \tag{43}
\]

Define the acceptance rule as follows:

\[
\pi(\text{accept}) = \min \left(1, \frac{\pi(\phi_\tau(p, q))}{\pi(p, q)} \right) \tag{44}
\]

Requires specifying the target distribution \( U \) and its derivative \( \frac{\partial U}{\partial q} \), the metric \( M^{-1} \), the step size \( \epsilon \) and trajectory length \( \tau \).
Why MCMC?

MCMC works well in some dynamical systems...

- Asymptotically correct.
- Requires skill, fails often.
- When will it converge?

Can VI beat MCMC in dynamical systems?
VI in a Nutshell

- Approx intractable \( p(z|x, \theta) \) w/ tractable \( q(z) \).

  \[ \Rightarrow \text{Trade integration for optimization} \]

- Jensen’s inequality (EM).

  Duality: Min KL \( \Leftrightarrow \) Max ELBO.
A Hidden Symmetry

\[
\log p(x|\theta) = \int q(z) \log \left( \frac{p(x|\theta) q(z)}{q(z)} \right) dz
\]  
(45)

Replace \( p(x|\theta) \) with \( p(z, x|\theta)/p(z|x, \theta) \):

\[
\log p(x|\theta) = \int q(z) \log \frac{p(x, z|\theta)}{q(z)} dz + \int q(z) \log \frac{q(z)}{p(z|x, \theta)} dz
\]  
(46)

\[
\log p(x|\theta) = \mathbb{E} \left[ \log p(x, z|\theta) \right] + \mathcal{H}(q) + KL(q(z)||p(z|x, \theta))
\]  
(47)

Maximizing ELBO (first two terms) equivalent to minimizing \( D_{KL} \):

\[
\arg\min_q KL(q(z)||p(z|x, \theta)) = \arg\max_q \mathcal{L}_{ELBO}
\]  
(48)
Stochastic Variational Inference (Hoffman, 2013)

Don’t cycle through all data. Speed up w/ existing ideas:

- Stochastic optimization (Robbins and Monro, 1951)
- Natural gradient (Amari, 1998)

Scalable inference w/ one or several data points per iteration.
Motivating Natural Gradients (Desjardins, 2015)

Rate of change could be different for different parameters.

⇒ Euclidean distance not good measure.

\[
\theta^{t+1} = \theta^t - \eta F^{-1}_\theta \left( \frac{\partial I(\theta, J)}{\partial \theta} \right)^T
\]  

(49)

- Riemannian metric \( g \) on manifold \( M \): inner product on tangent space s.t. each pt varies smoothly.

\[
g_p : T_pM \times T_pM \to \mathbb{R} \quad \forall X, Y \text{ on } M \quad p \to g_p(X(p), Y(p))
\]

- Riemannian manifold has *Fisher information* as metric tensor.

\[
F_{ij}(\theta) = -\int f(x, \theta) \frac{\partial^2 \log f(x, \theta)}{\partial \theta_i \partial \theta_j} dx
\]

(50)
Stochastic Optimization (Robbins, 1951)

Cannot directly observe $M(\theta)$ but can get samples of R.V.:

$$\mathbb{E}[N(\theta)] = M(\theta) \quad (51)$$

Pick a sequence $a_n$ s.t.

$$\sum_{n=0}^{\infty} a_n = \infty \quad \sum_{n=0}^{\infty} a_n^2 < \infty$$

Perform the update:

$$\theta_{n+1} = \theta_n - a_n(N(\theta_n) - \alpha) \quad (52)$$

Pick $\epsilon_t = t^{-\kappa}$ with $\kappa \in (0.5, 1]$, $M(\theta)$ is second term in ELBO.
Variational Auto Encoder (Kingma, 2014)

\[
p(z|x, \theta) = \frac{p(x|z, \theta)p(z|\theta)}{\int p(x, z|\theta)dz}
\]

ELBO does not include posterior explicitly:

\[
\mathcal{L}_{ELBO} = \int q(z)ln \frac{p(x, z|\theta)}{q(z)}dz = \int q(z)ln \frac{p(x|z, \theta)p(z)}{q(z)}dz
\]

\[
= \mathbb{E}_q \left[ ln p(x|z, \theta) \right] - KL(q(z)||p(z))
\]

ELBO approx \( p \) if search space for \( q \) is large enough.

- Idea: let’s not hard code the structure of \( p(x|z, \theta) \).
  
  ⇒ Use NN to learn generative model.
Variational Auto Encoder (Kingma, 2014)

* What is $p(x|z, \theta)$ w/o knowing $z$?

- Map $z$ onto a dist over $x$ parameterized by $\theta$.
  
  ⇒ Use NN to learn the generative model:
  
  $$p(x|z, \theta) = \mathcal{N}(x|\mu_\theta(z), \Sigma_\theta(z)) \quad (55)$$

- Map $x$ onto a distribution over $z$ w/ parameters $\phi$.
  
  ⇒ Use NN for the recognition model:
  
  $$q(z|x, \phi) = \mathcal{N}(z|\mu_\phi(x), \Sigma_\phi(x)) \quad (56)$$
Variational Auto Encoder (Kingma, 2014)

- Backprop needs partial wrt $\mu, L$; $\epsilon$ not included in derivation:

$$z = \mu + L\epsilon$$

$$\epsilon \sim \mathcal{N}(0, I)$$

$$\leftarrow f(\epsilon, (\mu, L))$$

$$q(z|x, \phi)$$

$$p(x|z, \theta)$$

$$ELBO = \mathbb{E}_q\left[\ln p(x|z, \theta)\right] - KL(q(z)||p(z))$$

(57)
Reparameterization (Kingma, 2014) and (Rezende, 2014)

- Take derivatives though random samples, back-propagate

\[
\nabla_\phi \mathbb{E}_{q(z|x,\phi)}[f(z, \theta)] = \int \nabla_\phi q(z|x, \phi)f(z, \theta)dz \quad (58)
\]

- Don’t directly sample \( z \) - instead sample \( \mathcal{N}(0, 1) \) and do an affine transformation.

\[
\tilde{z} \sim \mathcal{N}(\mu, \Sigma) \Rightarrow z = \mu + L\epsilon \quad \epsilon \sim \mathcal{N}(0, I) \quad \Sigma = LL^T \quad (59)
\]

- When taking the derivative bottom up requires the partial with respect to \( \mu, L \) and \( \epsilon \) is not included in the derivation.
VI with Normalizing Flow (Rezende, 2016)

Transform pdf through sequence of invertible smooth mappings:

\[ f : \mathbb{R}^d \rightarrow \mathbb{R}^d \quad f^{-1} = g \quad g \circ f(z) = g \] (60)

Need Jacobian for change of variables:

\[ q(z') = q(z) \left| \det \frac{\partial f^{-1}}{\partial z'} \right| = q(z) \left| \det \frac{\partial f}{\partial z} \right|^{-1} \] (61)

Repeatedly applying transformation:

\[ z_k = f_k \circ f_{k-1} \circ \cdots \circ f_1(z_0) \] (62)

\[ \ln q_k(z_k) = \ln q_0(z_0) - \sum_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|^{-1} \] (63)

The path of the r.v. is called a flow.
VI with Normalizing Flow (Rezende, 2016)

Idea:

- Pick an approx posterior \( q(z|x) \) and choose transformations \( f_k \)
  - Use simple factorized distributions (mean field)
- Apply sequence of transformations (normalizing flow) for expressive posterior.

LOTUS

- Expectations wrt \( q_k \) computed w/o \( q_k \) (no log-det Jacobian terms when \( h(z) \perp \!
\!
\perp q_k \)):

\[
\mathbb{E}_{q_k} [h(z)] = \mathbb{E}_{q_0} [h(f_k \circ f_{k-1} \circ \cdots \circ f_1(z_0))]
\] (64)
Langevin Stochastic Differential Equation

As length of norm flow $\rightarrow \infty \Rightarrow$ PDE:

\[
\frac{\partial}{\partial t} q_t(z) = - \sum_i \frac{\partial}{\partial z_i} \left[ F_i(z, t) q_t \right] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial z_i \partial z_j} \left[ D_{ij}(z, t) q_t \right] \quad (65)
\]
Stationary Distribution for Diffusion Process

Stationary distribution for $q_t(z)$:

$$\lim_{t \to \infty} q_t(z) \to q_\infty(z) \propto e^{-\mathcal{L}(z)} \quad (66)$$

Start w/ some initial distribution $q_0(z)$ let it evolve to get above.

- Hamiltonian Monte Carlo

$$\tilde{z} = (z, \omega) \quad H(z, \omega) = -\mathcal{L}(z) - \frac{1}{2} \omega^T M \omega \quad (67)$$
Planar Flows (Rezende, 2016)

\[ f(z) = z + uh(w^T z + b) \tag{68} \]
\[ \psi(z) = h'(w^T z + b)w \tag{69} \]

Log det Jacobian is \( \mathcal{O}(d) \)

\[
\left| \det \frac{\partial f}{\partial z} \right| = \left| \det (I + u\psi(z)^T) \right| \tag{70}
\]
\[ = \left| 1 + u^T \psi(z) \right| \tag{71} \]

\( f(z) \): family of transformations, \( \lambda = \{w \in \mathbb{R}^d, u \in \mathbb{R}^d, b \in \mathbb{R}\} \): free parameters, \( h(\cdot) \): smooth element wise nonlinearity.

\[ z_k = f_k \circ f_{k-1} \circ \cdots \circ f_1(z) \tag{72} \]

\[
\ln q_k(z_k) = \ln q_0(z) - \sum_{k=1}^{K} \ln \left| 1 + u_k^T \psi_k(z_k - 1) \right| \tag{73}
\]
Flow Based Free Energy Bound (Rezende, 2016)

$k$ : length of flow, $q_\phi(z|x) = q_k(z_k)$.

$$F(z) = \mathbb{E}_{q_\phi(z|x)} \left[ \ln q_\phi(z|x) - \ln p(x, z) \right]$$

$$= \mathbb{E}_{q_0(z_0)} \left[ \ln q_k(z_k) - \ln p(x, z_k) \right]$$

$$= \mathbb{E}_{q_0(z_0)} \left[ \ln q_0(z_0) \right] - \mathbb{E}_{q_0(z_0)} \left[ \ln p(x, z_k) \right]$$

Apply matrix determinant lemma:

$$- \mathbb{E}_{q_0(z)} \left[ \sum_{k=1}^{K} \ln \left| 1 + u_k^T \psi_k(z_k - 1) \right| \right]$$

Can be used with any model (VAE or AEVB with deep networks)

- Sampling and computing log-det Jacobian: $O(LN^2) + O(kd)$
- $L$ : # layers, $N$: avg layer size, $d$: dimension of $z$
Hamiltonian Variational Inference (Salimans, 2015)

Synthesize MCMC and VI?

\[ z_t \sim q(z_t | z_{t-1}, z) \]  \hspace{1cm} (77)

- MC \( z_t \rightarrow p(z|x) \).

- Converged MC is variational approx in state space.

\[ q(z|x) = q(z_0|x) \prod_{t=1}^{T} q(z_t | z_{t-1}, x) \]  \hspace{1cm} (78)

- \( y = z_0, z_1, \ldots z_t \) are auxiliary r.v.’s:
Hamiltonian Variational Inference (Salimans, 2015)

Write auxiliary r.v.'s \( y = z_0, \cdots, z_n \) into the lower bound:

\[
\mathcal{L}_{aux} = \mathbb{E}_{q(y,z_T|x)} \left[ \log[p(x,z_T) r(y|x,z_T)] - \log[q(y,z_T|x)] \right]
\]  

(79)

Iterated expectation \((D_{KL})\):

\[
= \mathcal{L} - \mathbb{E}_{q(z_T|x)} \left\{ D_{KL}(q(y|z_T,x) \| r(y|z_T,x)) \right\}
\]

(80)

Note \( D_{KL} \geq 0 \)

\[
\leq \mathcal{L} \leq \log p(x)
\]

(81)

• Free to choose \( r(y|x,z_T) \) as aux inference dist.
• Marginal approx \( q(z_T|x) = \int q(y,z_T|x)dy \) is mix of distn’s \( q(z_T|x,y) \)
Free to choose $r(y|x, z_T)$ as aux inference dist.

Marginal approx $q(z_T|x) = \int q(y, z_T|x)dy$ is mix of distn’s $q(z_T|x, y)$

$r(y|x, z_T) = q(y|x, z_T)$ is exact

→ Compromise and trade-off

* $r(y|x, z_T)$ flexible parametric form, opt lower bound of params
Gaussian Approximate Posterior (Archer, 2016)

Gaussian approx posterior (generative model is Gaussian Process).

\[ q(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \] \hspace{1cm} (82)

Recall our SGVB lower bound:

\[ ELBO = \mathcal{H}(q_\phi(z|x)) + \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x, z) \right] \] \hspace{1cm} (83)

Gaussian entropy above:

\[ \mathcal{H}(q_\phi(z|x)) = -\mathbb{E}_{q_\phi(z|x)} \left[ \log q_\phi(z|x) \right] \] \hspace{1cm} (84)

\[ = \frac{nT}{2} (1 + \log 2\pi) + \frac{1}{2} \log \det(\Sigma) \] \hspace{1cm} (85)

\( \Sigma \in \mathbb{R}^{nT \times nT} \Rightarrow \) number of params scales quadratically with \( T \).
Gaussian Approximate Posterior (Archer, 2016)

- Tradeoff between expressiveness and complexity:
  - Choosing diagonal $\Sigma$: *fully factorized* or *mean field*
- Instead use block tridiagonal inverse covariance:
  \[
  q(z|x) = \mathcal{N}\left(\mu_\phi(x), [R_\phi(x) \cdot R_\phi(x)^T]^{-1}\right)
  \]
  \[
  \Sigma^{-1} = R \cdot R^T \quad \text{R is lower block bidiagonal}
  \]
- Can find R linear in time and space:
  \[
  z = \mu + R^{-T} \epsilon
  \]
  \[
  \forall \ R \in \mathbb{R}^{nT \times nT}, \quad R^{-T} \epsilon \text{ cubic in dimensionality of } R
  \]
  \[\Rightarrow \text{But } R^{-1} \text{ linear (Trefethan, 1997)}\]
- $\log(|\Sigma|) = -2 \log(|R|) = -2 \sum_{i=1}^{T} \log(R_{ii})$
Parameterization of Smoothing Posterior (Archer, 2016)

Trick: never represent $\Sigma$ explicitly when learning $\phi, \theta$.

- $\text{cov}(z_t, z_{t+1})$ correspond to block diagonal and block off diagonal components of $\Sigma$.
- Differentiate $\phi$ through $\Sigma^{-1} = RR^T$

*Diagonal and off-diagonal parameterization* via three NN:

- $\mu_t = \text{NN}_{\phi_\mu}(x_t)$
- $D_t = \text{NN}_{\phi_D}(x_t)$
- $B_t = \text{NN}_{\phi_B}(x_t, x_{t+1})$

$$
\Sigma_{\phi}(x)^{-1} = \begin{bmatrix}
D_0 & B_0^T \\
B_0 & D_1 & B_1^T \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & B_{T-1}^T \\
& B_{T-1} & & B_T \\
& & D_T
\end{bmatrix}
$$
Product of Gaussian Approx Posterior (Archer, 2016)

Another approach is product of Gaussian factors:

\[ q(z|x) \propto r_0(z) \times r_1(z|x) \]

\[ = \mathcal{N}(z|0, D) \times \mathcal{N}(z|M_\phi(x), C_\phi(x)) \]  

- \( D, C \in \mathbb{R}^{nT \times nT}, M \in \mathbb{R}^{nT} \)
- Expressing the posterior as a Gaussian:

\[ \Sigma_\phi(x) = (D^{-1} + C_\phi^{-1}(x))^{-1} \]  
\[ M_\phi(x) = \Sigma_\phi(x)C_\phi(x)^{-1}M_\phi(x) \]

- \( D^{-1}, C^{-1} \) block tridiagonal, multiplicative interaction b/t posterior mean and covariance.
Linear Dynamical Population Models through Nonlinear Embeddings (Gao, 2016)

Temporally correlated approximate posterior:

\[ q_\phi(z_{r1}) \sim \mathcal{N}(\mu_1, Q_1) \]
\[ q_\phi(z_{rt} | z_{r(t-1)}) \sim \mathcal{N}(Az_{r(t-1)}, Q) \]
\[ q_\phi(z_{rt} | x_{rt}) \sim \mathcal{N}(m_\phi(x_{rt}), c_\phi(x_{rt})) \]

- NN parameterizes matrix valued function \( r_\phi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times m} \)

\[ c_\phi(x_{rt}) = (r_\phi(x_{rt})r_\phi(x_{rt})^T)^{-1} \quad (91) \]

- Full approximate posterior:

\[ q_\phi(z_r | x_r) \propto \prod_{t=1}^{T} q_\phi(z_{rt} | z_{r(t-1)}) q_\phi(z_{rt} | x_{rt}) q_\phi(z_{r1}) \quad (92) \]
Linear Dynamical Population Models: PLDS

\[ z_t = Az_{t-1} + \Sigma^{1/2} \epsilon \]
\[ x_t = \text{Pois} \left( \exp(Bz_t) \right) \]

\[ z \sim N(Az_{t-1}, \Sigma^{1/2}) \quad (93) \]
\[ x \sim \text{Pois} \left( \exp(Bz_t) \right) \quad (94) \]
Linear Dynamical Population Models: PLDS
Summary

- VI and MCMC use flow to evolve density.
  - Must specify dynamics for MCMC.
- Learn large class of $f$ and $g$ with VAE
  - Often restricted to GP
- Quantify uncertainty in VI?
Some Open Questions

ML:

- VI for nonlinear systems governed by differential equations...
- Learn flow to guide HMC.
- Approx MCMC chain with VI.
- Adapt normalizing flow for SSMs...
- Dimensionality reduction that preserves Riemannian manifolds.

Neuroscience:

- Apply to Ca+ and voltage imaging in hippocampus or V1
- Reproduce results only found in physiology.
- Hierarchical models of deep layers (dendritic trees).
Conclusion

Thank you!


References IV

