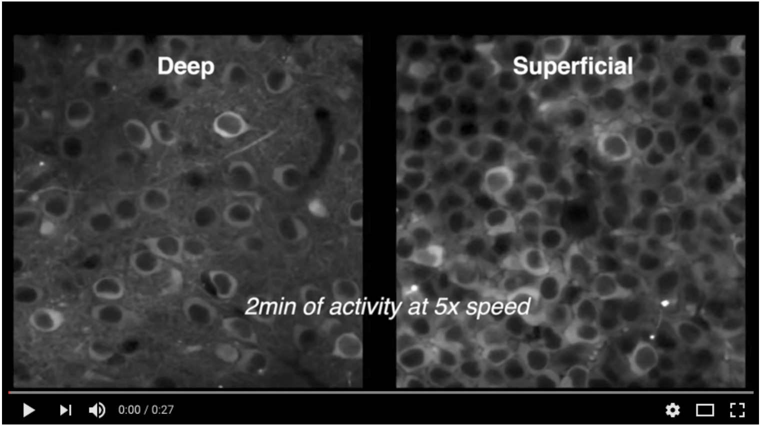
Inference in Nonlinear Dynamical Systems

PhD Candidacy Exam

Antonio Moretti

June 5, 2017

Two Photon Calcium Imaging of Mouse Hippocampus



State Space Models

$$\{Z_t\}_{t\geq 0} \qquad \{X_t\}_{t\geq 0} \tag{1}$$

$$Z_t|(Z_{0:t-1}) \sim f_\theta(z_t|z_{t-1}) \tag{2}$$

$$X_t|(Z_{0:t}, X_{0:t-1}) \sim g_\theta(x_t|z_t) \tag{3}$$

$$\begin{array}{ccccccc}
Z_0 & \xrightarrow{f_\theta} & Z_1 & \longrightarrow & Z_2 & \longrightarrow & \dots \\
g_\theta \big\downarrow & & \big\downarrow & & \big\downarrow & & \\
X_0 & & X_1 & & X_2 & & \dots
\end{array}$$

Bayesian Inference

Posterior:

$$p(z|x,\theta) = \frac{p(x,z|\theta)}{p(x|\theta)} = \frac{p(x|z,\theta)p(z|\theta)}{\int p(x,z|\theta)dz} \tag{4}$$

Bayesian Inference

Posterior:

$$p(z|x,\theta) = \frac{p(x,z|\theta)}{p(x|\theta)} = \frac{p(x|z,\theta)p(z|\theta)}{\int p(x,z|\theta)dz} \tag{5}$$

Two approaches:

- MCMC
- VI

Can VI outperform MCMC in (nonlinear) dynamical systems?

Inference Challenges

Likelihood:

$$p_\theta(x_{0:t}) = \int p_\theta(z_{0:t}, x_{0:t}) dz_{0:t} \tag{6}$$

Joint:

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

Posterior:

$$p(\theta|x_{0:t}) = \frac{p_\theta(x_{0:t})p(\theta)}{\int p_\theta(x_{0:t})p(\theta)d\theta} \tag{8}$$

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.

Auxiliary Particle Filtering (Pitt, 1998)

Importance function q_θ :

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

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

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

and

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

Weights w_t compensate for sampling from q_θ instead of p_θ.

Auxiliary Particle Filtering (Pitt, 1998)

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

- Bootstrap particle filter (Gordon, 1993)

$$q_\theta(z_t|x_t, z_{t-1}) = f_\theta(z_t|z_{t-1}) \tag{14}$$

- Sequential Importance Sampling Resampling (Doucet, 2000)

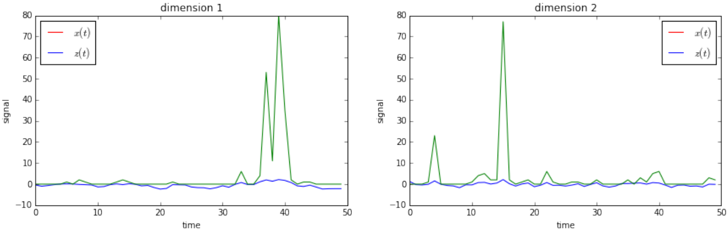
$$q_\theta(x_t|z_{t-1}) = 1 \tag{15}$$

Choose importance function q_θ to get the above.

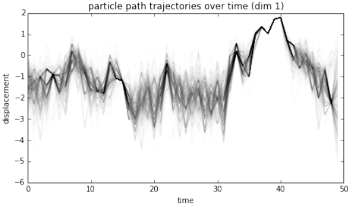
Poisson Linear Dynamical System

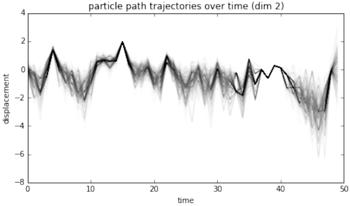
$$z_t = Az_{t-1} + \Sigma^{1/2}\epsilon \qquad z \sim N(Az_{t-1}, \Sigma^{1/2}) \tag{16}$$

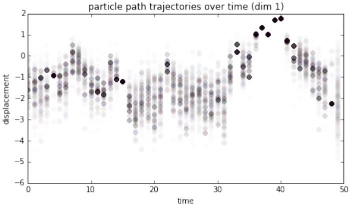
$$x_t = Pois\Big(exp(Bz_t)\Big) \qquad x \sim Pois\Big(exp(Bz_t)\Big) \tag{17}$$

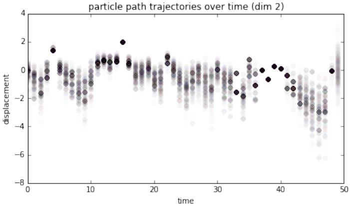


Path Measures

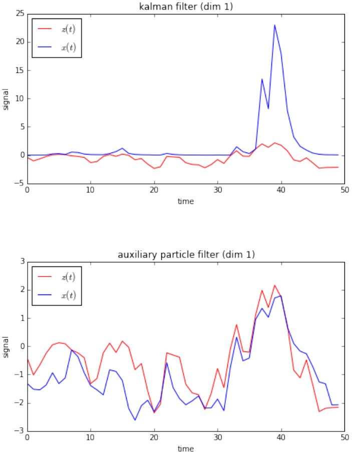




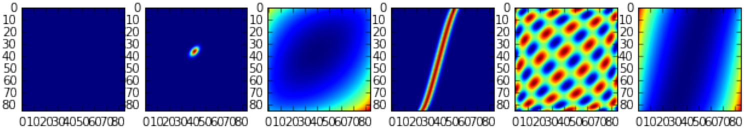




PLDS: Overcoming Nonlinear Measurement Models



Particle Quadrature



Function $f(x,y)$	Quadrature Integration	Monte Carlo	Difference
1	1	1	4.440e-16
$exp(-x^2-y^2)$	0.08939	0.0894	2.367e-05
x^2+y^2	7.0	6.99779	0.00220
$exp(-x^2)$	0.41368	0.4139	0.00026
$cos(x)+cos(y)$	0.07530	0.0754	0.00015
$(1-x)^2+\frac{1}{100}(y-x^2)^2$	1	0.9985	0.0014

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

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

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:

$$\frac{\mathbb{E}\Big(F(Z_0, \cdots, Z_n) \prod_{k=0}^{n} G_k(Z_k)\Big)}{\mathbb{E}\Big(\prod_{k=0}^{n} G_k(Z_k)\Big)} \tag{20}$$

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

$$d\mathbb{Q}_n = \frac{1}{\mathcal{Z}_n}\Big\{\prod_{0 \le p < n} G_p(Z_p)\Big\} d\mathbb{P}_n \tag{21}$$

G_n is a sequence of *potential functions* on E_n, $\mathbb{Q}_n$ 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

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

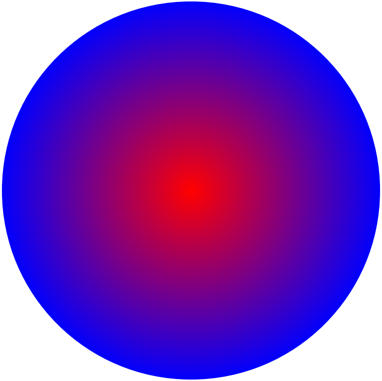
- Metropolis Hastings

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

Standard MCMC is inefficient and requires parameterization.

Concentration of Measure

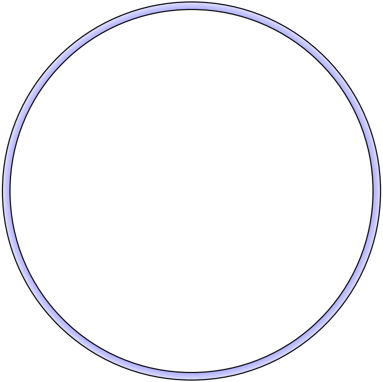
Typical set?



Concentration of Measure

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

$$\frac{Vol((1-\epsilon)A)}{Vol(A)} = (1-\epsilon)^d \leq e^{-\epsilon d} \tag{22}$$



Hamiltonian Dynamics

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

$$H(q,p) = U(q) + K(p) \tag{23}$$

- EoM:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{24}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{25}$$

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

$$\frac{dz}{dt} = \underbrace{\begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}}_{J} \nabla H(z) \tag{26}$$

Illustrating Hamiltonian Flow

Quadratic kinetic and potential energies in one dimension:

$$H(q,p) = \underbrace{\frac{q^2}{2}}_{U(q)} + \underbrace{\frac{p^2}{2}}_{K(p)} \tag{27}$$

Then

$$\frac{dq}{dt} = p \qquad \frac{dp}{dt} = -q \tag{28}$$

Solving:

$$q(t) = r\ cos(a+t) \qquad p(t) = -r\ sin(a+t) \tag{29}$$

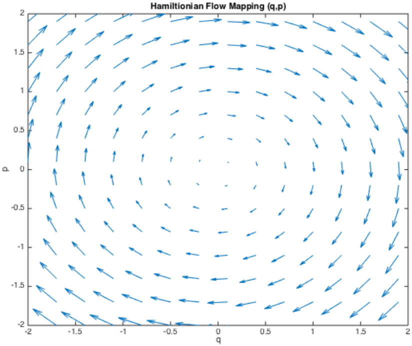
Quadratic potential gives standard Gaussian distribution for q.

Hybrid (Hamiltonian) Monte Carlo (Duane, 1987)

Sample deterministic paths for coherent exploration of state space.

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

$$\dot{q} = p \qquad \dot{p} = -q$$



- Analogy with Hamiltonian mechanics via momenta $p = mv$.

Hamiltonian Flow (Duane, 1987)

Perform measure preserving mapping:

$$\rho(p,q) \to e^{-H(p,q)} d^n p d^n q \tag{30}$$

Markov transition:

$$T(q',q) = \pi(p)\delta\Big((p',q') - \phi_\tau(p,q)\Big) \tag{31}$$

Take log:

$$H(p,q) = -log\ \pi(p,q) \tag{32}$$

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

KE is the conditional distribution and PE is the target distribution. Must integrate to get $\pi(q)$.

Euclidean HMC (Duane, 1987)

$$H(p,q) = -log\ \pi(p,q) \tag{34}$$
$$= -\underbrace{log\ \pi(p|q)}_{Kinetic\ \ Energy\ \ T} - \underbrace{log\ \pi(q)}_{Potential\ \ Energy\ \ U} \tag{35}$$

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

$$\pi(p|q) = \mathcal{N}(0, M) \tag{36}$$
$$T = \frac{1}{2} p_i p_j (M^{-1})^{ij} = \frac{1}{2} p^T M^{-1} p \tag{37}$$

- When kinetic energy has a quadratic form $\Rightarrow$ Euclidean HMC.

Riemann Manifold Langevin and Hamiltonian Monte Carlo (Girolami, 2015)

Hamiltonian constant $\Rightarrow \Delta U = \Delta T = d/2$.

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

$$\pi(p|q) = \mathcal{N}(0, M(q)) \tag{38}$$

$$T = \frac{1}{2} p_i p_j \left(M^{-1}(q)\right)^{ij} + \frac{1}{2} log \left|M(q)\right| \tag{39}$$

- Inverse covariance matrix corrects curvature (Fisher information) emulating dynamics of a Riemannian manifold.
- Normalization acts like a capacitor.
  - High curvature $\Leftrightarrow$ absorb energy.
  - Low curvature $\Leftrightarrow$ 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:

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

Requires specifying the target distribution U and its derivative $\frac{\partial U}{\partial q}$, the metric M^{-1}, the step size ϵ and trajectory length τ.

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 $q(z)$.

  $\Rightarrow$ Trade integration for optimization

- Jensen's inequality (EM).

  Duality: Min KL $\Leftrightarrow$ Max ELBO.

A Hidden Symmetry

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

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

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

$$log\ \ p(x|\theta) = \underbrace{\mathbb{E}\Big[log\ \ p(x,z|\theta)\Big] + \mathcal{H}(q)}_{ELBO(q)} + KL(q(z)||p(z|x,\theta)) \tag{47}$$

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

$$\underset{q}{\text{argmin}}\, KL(q(z)||p(z|x,\theta)) = \underset{q}{\text{argmax}}\, \mathcal{L}_{ELBO} \tag{48}$$

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.

$\Rightarrow$ Euclidean distance not good measure.

$$\theta^{t+1} = \theta^{t} - \eta F_{\theta}^{-1} \left(\frac{\partial l(\theta, J)}{\partial \theta} \right)^{T} \tag{49}$$

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

$$g_p : T_p M \times T_p M \rightarrow R \quad \forall X, Y \text{ on } M \;\; p \rightarrow g_p(X(p), Y(p))$$

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

$$F_{ij}(\theta) = -\int f(x, \theta) \frac{\partial^2 log \;\; f(x, \theta)}{\partial \theta_i \partial \theta_j} dx \tag{50}$$

Stochastic Optimization (Robbins, 1951)

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

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

Pick a sequence a_n s.t.

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

Perform the update:

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

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:

$$\mathcal{L}_{ELBO} = \int q(z) ln \frac{p(x,z|\theta)}{q(z)} dz = \int q(z) ln \frac{p(x|z,\theta)p(z)}{q(z)} dz \tag{53}$$

$$= \mathbb{E}_q\Big[ln\ p(x|z,\theta)\Big] - KL(q(z)||p(z)) \tag{54}$$

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

- Idea: let's not hard code the structure of $p(x|z,\theta)$.

  $\Rightarrow$ 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 θ.

  ⇒ Use NN to learn the generative model:

$$p(x|z, \theta) = \mathcal{N}(x|\mu_\theta(z), \Sigma_\theta(z)) \tag{55}$$

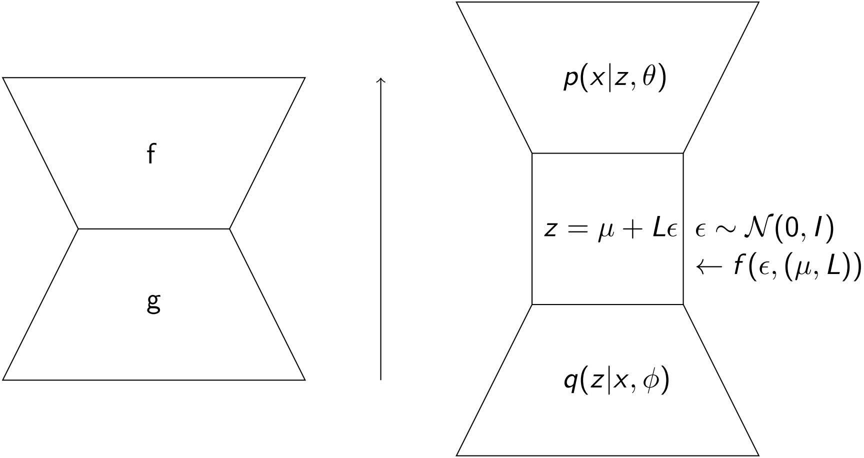
- Map x onto a distribution over z w/ parameters ϕ.

  ⇒ Use NN for the recognition model:

$$q(z|x, \phi) = \mathcal{N}(z|\mu_\phi(x), \Sigma_\phi(x)) \tag{56}$$

Variational Auto Encoder (Kingma, 2014)

- Backprop needs partial wrt μ, L; ϵ not included in derivation:



$$ELBO = \mathbb{E}_q\Big[ln\ p(x|z,\theta)\Big] - KL(q(z)||p(z)) \tag{57}$$

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

- Take derivatives though random samples, back-propagate

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

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

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

- When taking the derivative bottom up requires the partial with respect to μ, L and ϵ is not included in the derivation.

VI with Normalizing Flow (Rezende, 2016)

Transform pdf through sequence of invertible smooth mappings:

$$f : \mathbb{R}^d \to \mathbb{R}^d \qquad f^{-1} = g \qquad g \circ f(z) = g \tag{60}$$

Need Jacobian for change of variables:

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

Repeatedly applying transformation:

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

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

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.

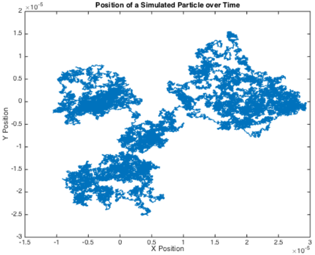
LOTUS

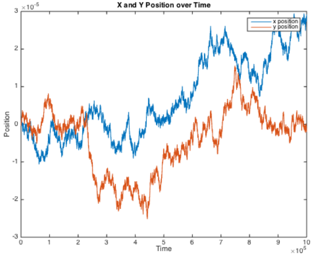
- Expectations wrt q_k computed w/o q_k (no log-det Jacobian terms when $h(z) \perp\!\!\!\perp q_k$):

$$\mathbb{E}_{q_k}\left[h(z)\right] = \mathbb{E}_{q_0}\left[h\left(f_k \circ f_{k-1} \circ \cdots \circ f_1(z_0)\right)\right] \tag{64}$$

Langevin Stochastic Differential Equation

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





$$\frac{\partial}{\partial t} q_t(z) = -\sum_i \frac{\partial}{\partial z_i}\Big[F_i(z,t)q_t\Big] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial z_i \partial z_j}\Big[D_{ij}(z,t)q_t\Big] \quad (65)$$

Stationary Distribution for Diffusion Process

Stationary distribution for $q_t(z)$:

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

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

⋆ Hamiltonian Monte Carlo

$$\tilde{z} = (z, \omega) \qquad H(z, \omega) = -\mathcal{L}(z) - \frac{1}{2}\omega^T M \omega \tag{67}$$

Planar Flows (Rezende, 2016)

$$f(z) = z + uh(w^T z + b) \quad (68)$$

$$\psi(z) = h'(w^T z + b)w \quad (69)$$

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

$$\left| det \frac{\partial f}{\partial z} \right| = \left| det(I + u\psi(z)^T) \right| \quad (70)$$

$$= \left| 1 + u^T \psi(z) \right| \quad (71)$$

$f(z)$: 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.

$$z_k = f_k \circ f_{k-1} \circ \cdots \circ f_1(z) \quad (72)$$

$$ln\ q_k(z_k) = ln\ q_0(z) - \sum_{k=1}^{K} ln \left| 1 + u_k^T \psi_k(z_k - 1) \right| \quad (73)$$

Flow Based Free Energy Bound (Rezende, 2016)

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

$$\mathcal{F}(z) = \mathbb{E}_{q_\phi(z|x)}\Big[ln\ q_\phi(z|x) - ln\ p(x,z)\Big] \tag{74}$$

$$= \mathbb{E}_{q_0(z_0)}\Big[ln\ q_k(z_k) - ln\ p(x,z_k)\Big] \tag{75}$$

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

Apply matrix determinant lemma:

$$-\mathbb{E}_{q_0(z)}\Big[\sum_{k=1}^{K} ln\big|1 + u_k^T \psi_k(z_k - 1)\big|\Big] \tag{76}$$

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

- Sampling and computing log-det Jacobian: $\mathcal{O}(LN^2) + \mathcal{O}(kd)$
- L : # layers, N: avg layer size, d: dimension of z

Hamiltonian Variational Inference (Salimans, 2015)

Synthesize MCMC and VI?

$$z_t \sim q(z_t | z_{t-1}, z) \tag{77}$$

- MC $z_t \longrightarrow p(z|x)$.

⋆ Converged MC is variational approx in state space.

$$q(z|x) = q(z_0|x) \prod_{t=1}^{T} q(z_t | z_{t-1}, x) \tag{78}$$

- $y = z_0, z_1, \cdots z_t$ 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:

$$\mathcal{L}_{aux} = \mathbb{E}_{q(y,z_T|x)}\left[log[p(x, z_T)r(y|x, z_T)] - log[q(y, z_T|x)]\right] \quad (79)$$

Iterated expectation (D_{KL}):

$$= \mathcal{L} - \mathbb{E}_{q(z_T|x)}\left\{D_{KL}\left(q(y|z_T, x)||r(y|z_T, x)\right)\right\} \quad (80)$$

Note $D_{KL} \geq 0$

$$\leq \mathcal{L} \leq log\ \ p(x) \quad (81)$$

- 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 $q(z_T|x, y)$

Hamiltonian Variational Inference (Salimans, 2015)

- 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

  $\rightarrow$ Compromise and trade-off

$\star$ $r(y|x, z_T)$ flexible parametric form, opt lower bound of params

Gaussian Approximate Posterior (Archer, 2016)

Gaussian approx posterior (generative model is Gaussian Process).

$$q(z|x) = \mathcal{N}\Big(\mu_\phi(x), \Sigma_\phi(x)\Big) \tag{82}$$

Recall our SGVB lower bound:

$$ELBO = \mathcal{H}\Big(q_\phi(z|x)\Big) + \mathbb{E}_{q_{\phi(z|x)}}\Big[log\ \ p_\theta(x,z)\Big] \tag{83}$$

Gaussian entropy above:

$$\mathcal{H}\Big(q_\phi(z|x)\Big) = -\mathbb{E}_{q_{\phi(z|x)}}\Big[log\ \ q_\phi(z|x)\Big] \tag{84}$$

$$= \frac{nT}{2}(1 + log\ \ 2\pi) + \frac{1}{2} log\ \ det\Big(\Sigma\Big) \tag{85}$$

$\Sigma \in \mathbb{R}^{nT \times nT} \Rightarrow$ number of params scales quadratically with T.

Gaussian Approximate Posterior (Archer, 2016)

- Tradeoff between expressiveness and complexity:

  $\Rightarrow$ Choosing diagonal Σ: *fully factorized* or *mean field*
- Instead use block tridiagonal inverse covariance:

$$q(z|x) = \mathcal{N}\left(\mu_\phi(x), [R_\phi(x) \cdot R_\phi(x)^T]^{-1}\right) \tag{86}$$

$$\Sigma^{-1} = R \cdot R^T \qquad R \text{ is lower block bidiagonal}$$

- Can find R linear in time and space:

$$z = \mu + R^{-T}\epsilon \tag{87}$$

$$\forall\ R \in \mathbb{R}^{nT \times nT}, \quad R^{-T}\epsilon \text{ cubic in dimensionality of R}$$

$$\Rightarrow \text{But } R^{-1} \text{ linear (Trefethan, 1997)}$$

- $log(|\Sigma|) = -2\ log(|R|) = -2 \sum_{i=1}^{T} log(R_{ii})$

Parameterization of Smoothing Posterior (Archer, 2016)

Trick: never represent Σ explicitly when learning ϕ, θ.

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

Diagonal and off-diagonal parameterization via three NN:

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

$$\Sigma_\phi(x)^{-1} = \begin{bmatrix} D_0 & B_0^T & & & \\ B_0 & D_1 & B_1^T & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & B_{T-1}^T \\ & & & B_{T-1} & D_T \end{bmatrix}$$

Product of Gaussian Approx Posterior (Archer, 2016)

Another approach is product of Gaussian factors:

$$q(z|x) \propto \underbrace{r_0(z)}_{} \times \underbrace{r_1(z|x)}_{} \tag{88}$$
$$= \mathcal{N}\big(z|0, D\big) \times \mathcal{N}\big(z|M_\phi(x), C_\phi(x)\big)$$

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

$$\Sigma_\phi(x) = \Big(D^{-1} + C_\phi^{-1}(x)\Big)^{-1} \tag{89}$$
$$M_\phi(x) = \Sigma_\phi(x) C_\phi(x)^{-1} M_\phi(x) \tag{90}$$

- 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:

$$
\begin{aligned}
q_\phi(z_{r1}) &\sim \mathcal{N}(\mu_1, Q_1) \\
q_\phi(z_{rt}|z_{r(t-1)}) &\sim \mathcal{N}(Az_{r(t-1)}, Q) \\
q_\phi(z_{rt}|x_{rt}) &\sim \mathcal{N}(m_\phi(x_{rt}), c_\phi(x_{rt}))
\end{aligned}
$$

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

$$c_\phi(x_{rt}) = \left(r_\phi(x_{rt})r_\phi(x_{rt})^T\right)^{-1} \tag{91}$$

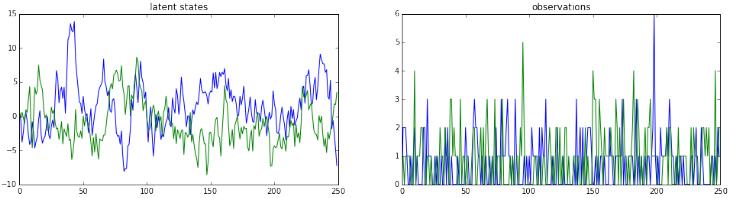
- Full approximate posterior:

$$q_\phi(z_r|x_r) \propto \prod_{t=1}^{T} q_\phi\left(z_{rt}|z_{r(t-1)}\right)q_\phi(z_{rt}|x_{rt})q_\phi(z_{r1}) \tag{92}$$

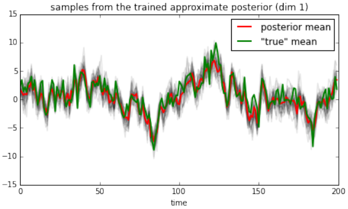
Linear Dynamical Population Models: PLDS

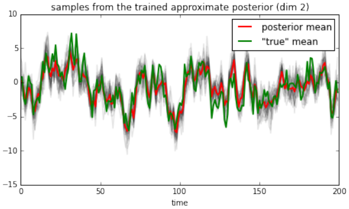
$$z_t = Az_{t-1} + \Sigma^{1/2}\epsilon \qquad z \sim N(Az_{t-1}, \Sigma^{1/2}) \tag{93}$$

$$x_t = Pois\Big(exp(Bz_t)\Big) \qquad x \sim Pois\Big(exp(Bz_t)\Big) \tag{94}$$



Linear Dynamical Population Models: PLDS





Summary

- VI and MCMC use flow to evolve density.

  $\Rightarrow$ Must specify dynamics for MCMC.

- Learn large class of f and g with VAE

  $\Rightarrow$ 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 I

Cdric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor, *Gaussian process approximations of stochastic differential equations.*, Gaussian Processes in Practice (Neil D. Lawrence, Anton Schwaighofer, and Joaquin Quionero Candela, eds.), JMLR Proceedings, vol. 1, JMLR.org, 2007, pp. 1–16.

Cdric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor, *Variational inference for diffusion processes.*, NIPS (John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, eds.), Curran Associates, Inc., 2007, pp. 17–24.

Tianqi Chen, Emily B. Fox, and Carlos Guestrin, *Stochastic gradient hamiltonian monte carlo.*, CoRR **abs/1402.4102** (2014).

References II

Roger Frigola, Yutian Chen, and Carl E. Rasmussen, *Variational gaussian process state-space models.*, NIPS (Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, eds.), 2014, pp. 3680–3688.

Yuanjun Gao, Evan W. Archer, Liam Paninski, and John P. Cunningham, *Linear dynamical neural population models through nonlinear embeddings.*, NIPS (Daniel D. Lee, Masashi Sugiyama, Ulrike V. Luxburg, Isabelle Guyon, and Roman Garnett, eds.), 2016, pp. 163–171.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley, *Stochastic variational inference*, Journal of Machine Learning Research **14** (2013), 1303–1347.

References III

- Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein, *A latent variable recurrent neural network for discourse relation language models.*, CoRR **abs/1603.01913** (2016).
- Rahul G. Krishnan, Uri Shalit, and David Sontag, *Deep kalman filters.*, CoRR **abs/1511.05121** (2015).
- Diederik P. Kingma and Max Welling, *Auto-encoding variational bayes.*, CoRR **abs/1312.6114** (2013).
- Seyed Mohammad Khansari-Zadeh and Aude Billard, *Learning stable nonlinear dynamical systems with gaussian mixture models.*, IEEE Trans. Robotics **27** (2011), no. 5, 943–957.
- Radford Neal, *Mcmc using hamiltonian dynamics*, CRC Press, May 2011.

References IV

Danilo Jimenez Rezende and Shakir Mohamed, *Variational inference with normalizing flows.*, CoRR **abs/1505.05770** (2015).

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, *Stochastic backpropagation and approximate inference in deep generative models.*, ICML, JMLR Workshop and Conference Proceedings, vol. 32, JMLR.org, 2014, pp. 1278–1286.

Tim Salimans, Diederik P. Kingma, and Max Welling, *Markov chain monte carlo and variational inference: Bridging the gap.*, ICML (Francis R. Bach and David M. Blei, eds.), JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, 2015, pp. 1218–1226.

References V

Alexander Y. Shestopaloff and Radford M. Neal, *MCMC for non-linear state space models using ensembles of latent sequences*, arXiv preprint arXiv:1305.0320 (2013).