Justin Domke, University of Massachusetts Amherst

these slides: t.ly/9vZvk or people.cs.umass.edu/domke/diffusion.pdf

do probabilistic inference

There are only two ways I know to make money in business:

bundling and unbundling.

variables variables out

Jim Barksdale



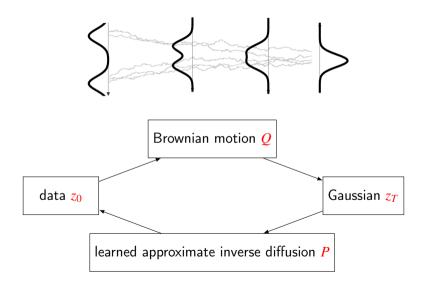


meanwhile...

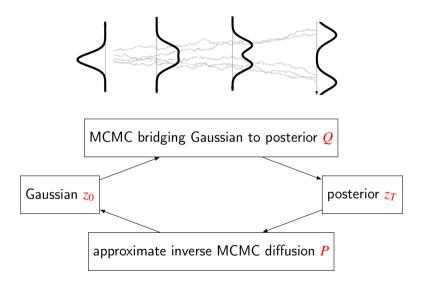
Variational Inference with Hamiltonian Monte Carlo	Stochastic Normalizing Flows	Monte Carlo Variational	d Auto-Encoders	Differentiable Annealed Importance Sampling and the Perils of Gradient Noise	MCMC Variational Inference via Uncorrected Hamiltonian Annealing	Bayesian Inference via Sparse Hamiltonian Flows
have sumped not primited and absolute but to one primary of the p	A common options in a table to having and studies in this special copyrights on a physics in the special configuration of any special configuration of the studies of the studies of the studies in the same and the same plane of the studies of the studies of the same and the same plane of the studies of the same and the same plane of t	Section 2014 and 1014	where the same bound of MT should as $M(M) = M(M) $	through pains," A feeling is been a second to the control of the c	Cologne of the state of the colonial co	clears member, enabling the use of scalable stochastic optimization algorithms (16, 17). While past work involved complex manuscrip families, record work has developed from families based on Markov
in 1900 K. Konjan and Welling (2001) apply it in the generation of honors. That the variety of reason and Marindian Keel are also prefer approximation. Additionally, we prevent specific approximation. Additionally, we prevent representations of the ISSN algorithm, which can be in- cluded with the thirt, Adv. Intrinsiche to Honor.	network, probing forward a probability drawity even a latest as "point" quar Z instant of the target X . Unlike the training of unlike the train much required result desirable that distinct point a real training the finite format is a factor of promoted samples, thereing them in the trained by soften maximizing the finite format and ΔM , or maximizing the Killeholt Lefelds description. (Maximized training the distribution of the ΔM and ΔM are maximized to the Killeholt Lefelds description.) It is a finite format in the ΔM and ΔM are the soften distribution of the ΔM and ΔM are the soften distribution of the ΔM and ΔM are the soften distribution of the ΔM and ΔM are the soften distribution of the ΔM and ΔM are the soften distribution of the ΔM and ΔM are the soften distribution of the soften distribu	nie Piets Andry Preuw CIDEL Bachtern bestehn of Krimser dem All Technology, Misserie Beisel, Techno Meissel, Beigheiser Brief, Raday, Preum "BEE Lab, HEE University, Misserie Beisel, and "University of Outstan Correspondence to: Adultic Beise and "University of Outstan Correspondence to: Adultic Beise Adulti deiter polymologies odes." Preumodium of the 16 th International Conference on Bachter Preumodium of the 16 th International Conference on Bachter 10 pp. 10	nor. These algorithms only on an extended target distri- tion for which an efficient importance distribution can be found using non-homogeneous Machor Leonds. In boson suggested in various contributions that AER could such in team NAE (Submune et al., 2015; We et al., by Macklows et al., 2015; We et al.,	(SMC) [Dissort et al., 2001] and needed sampling [Midling et al., 2006]. Under some assumptions, ANI is able to produce accounts estimates of marginal likelihood price enough computation time (it converges to the true ML value quickly by adding some intermediate distributions).	consume of instructible denotice [15, 44]. All greates a supersect of denotics the bridge from a tractable initial approximation 4 to the super a. These, the supersect violational distribution is given by a supersect of MCMP learnest negating rush bridging density, while the supersect supersect reversals of these invention. It leaves out that the safe of these supersects distributions are be computed using only architecture of the bridging densities. Combining Benthisman MCMC learnest with AIC	which is $(18, 19)$ —and is particular, from board on Langeron and Hamiltonian dynamics (20.20) . However, housest from Marker shines in prigorilly designed in larger the proteins distribution, each trip again requires a composition in ordinary of the data, under M Lange M Langeron M

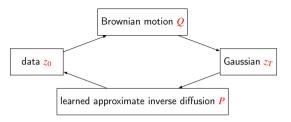
This talk

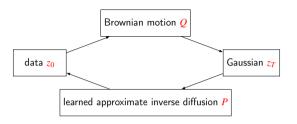
- What's the relationship?
- A unified view of these inference methods.



Diffusion-esque inference





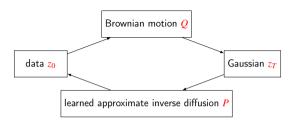


Q: sample z_0 from data, add noise until z_T

P: sample z_T from Gaussian, denoise until z_0

minimize KL(Q||P)

 $\nabla \log q_t(z)$ defines ideal denoising



Q: sample
$$z_0$$
 from data, add noise until z_T $dz_t = -\beta_t z_t dt + \sqrt{2\beta_t} dw_t$

P: sample
$$z_T$$
 from Gaussian, denoise until z_0 $dz_t = -\beta_t z_t dt - 2\beta_t s_\theta(z_t, t) dt + \sqrt{2\beta_t} d\bar{w}_t$

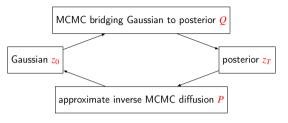
$$\mathsf{minimize}\ \mathit{KL}(\mathit{Q} \| \mathit{P})$$

 $KL(Q||P) = KL(q_T||p_T) + \mathbb{E}_{\mathsf{z}_T \sim q_T} KL(Q(\cdot|\mathsf{z}_T)||P(\cdot|\mathsf{z}_T))$

$$abla \log q_t(z)$$
 defines ideal denoising

oising if
$$s_{\theta}(z,t) = \nabla \log q_t(z)$$
 then $KL(Q||P) = KL(q_T||p_T)$

Diffusion-esque inference



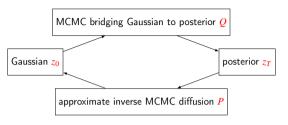
Q: sample z_0 from Gaussian, do Langevin on π_t where $\pi_0 =$ (Gaussian) and $\pi_T =$ (posterior).

P: sample z_T from posterior, do inverse MCMC.

minimize KL(Q||P)

 $\nabla \log q_t(z)$ defines optimal inverse dynamics

Diffusion-esque inference



Q: sample z_0 from Gaussian, do Langevin on π_t $dz_t = \nabla \log \pi_t(z_t) dt + \sqrt{2} \ dw_t$ where $\pi_0 =$ (Gaussian) and $\pi_T =$ (posterior).

P: sample z_T from posterior, do inverse MCMC. $d\mathbf{z}_t = \nabla \log \pi_t(\mathbf{z}_t) dt - 2s_{\theta}(\mathbf{z}_t, t) dt + \sqrt{2} d\bar{\mathbf{w}}_t$

minimize
$$KL(Q||P)$$

$$KL(Q||P) = KL(q_T||p_T) + \mathbb{E}_{\mathbf{z}_T \sim q_T} KL(Q(\cdot|\mathbf{z}_T)||P(\cdot|\mathbf{z}_T))$$

$$\nabla \log q_t(z)$$
 defines optimal inverse dynamics

if
$$s_{\theta}(z,t) = \nabla \log q_t(z)$$
 then $KL(Q||P) = KL(q_T||p_T)$

	Diffusion models	Diffusion-based VI
z_0	data (samples)	Gaussian
z_T	Gaussian	posterior (distribution)

	Diffusion models	Diffusion-based VI
z_0	data (samples)	Gaussian
z_T	Gaussian	posterior (distribution)
Q	noising	MCMC on bridging densities
P	learned denoising	inverse MCMC

	Diffusion models	Diffusion-based VI
z_0	data (samples)	Gaussian
ZT	Gaussian	posterior (distribution)
Q P	noising learned denoising	MCMC on bridging densities inverse MCMC
P	learned denoising	inverse ivicivic
goal	good P to model z_0	good ${\it Q}$ to model z_T

	Diffusion models	Diffusion-based VI
<i>z</i> ₀	data (samples)	Gaussian
Z_T	Gaussian	posterior (distribution)
Q P	noising learned denoising	MCMC on bridging densities inverse MCMC
goal	good P to model z_0	good Q to model z_T
marginals $q(z_t z_0)$ tractable scores $ abla \log q(z_t)$ tractable	yes no	no ③ no, but can approximate using bridging densities ⑤

Variational Inference

$$p(z) = \bar{p}(z)/Z$$

$$\min_{q \in \mathsf{Family}} \!\! \mathit{KL} \left(q(\mathbf{z}) \| p(\mathbf{z}) \right)$$

$$\max_{q \in \mathsf{Family}} \mathsf{ELBO}(q(\mathsf{z}) \| \bar{p}(\mathsf{z})) := \underset{q(\mathsf{z})}{\mathbb{E}} \log \left(\bar{p}(\mathsf{z}) / q(\mathsf{z}) \right)$$

- Q: $z_0 \sim q_0$, run MCMC diffusion on π_t $(q_0 = \pi_0 \leadsto \pi_T = p_T)$ until z_T .
- P: $z_T \sim p_T$, run reverse diffusion on π_t until z_0 .
- Convert *Q* and *P* to discrete time.
- Optimize $KL(Q||P) \ge KL(q_T||p_T)$.

- Q: $z_0 \sim q_0$, run MCMC diffusion on π_t $(q_0 = \pi_0 \leadsto \pi_T = p_T)$ until z_T .
- P: $z_T \sim p_T$, run reverse diffusion on π_t until z_0 .
- Convert Q and P to discrete time.
- Optimize $KL(Q||P) \ge KL(q_T||p_T)$.

Many instances (Wu et al 2020; Thin et al 2021; Geffner & D 2021, 2023; Zhang et al 2021; Doucet et al 2022).

- Q: $z_0 \sim q_0$, run MCMC diffusion on π_t $(q_0 = \pi_0 \leadsto \pi_T = p_T)$ until z_T .
- P: $z_T \sim p_T$, run reverse diffusion on π_t until z_0 .
- Convert Q and P to discrete time.
- Optimize $KL(Q||P) \ge KL(q_T||p_T)$.

Many instances (Wu et al 2020; Thin et al 2021; Geffner & D 2021, 2023; Zhang et al 2021; Doucet et al 2022).

Design choices:

- Starting distribution q_0
 - Bridging distributions π_t
 - ullet Forward process Q
 - Backward process P
 - ullet Numerical simulation of Q and P
 - Optimizer

(Standard Gaussian? Learned Gaussian?)

(Fixed? Learned?)

(Langevin? Underdamped Langevin?)

(Fixed? Learned score network?)

(Splitting? Euler-Maruyama?)

(SGD? Adam? Step sizes?)

Starting distribution

 q_0 closer to $p_T \Longrightarrow$ less distance for π_t to travel

Starting distribution

 q_0 closer to $p_T \Longrightarrow$ less distance for π_t to travel

$$\begin{array}{lll} q_0 & = & \mathcal{N}(0,I) \\ q_0 & = & \mathcal{N}(\mu, \operatorname{diag}(\sigma^2)), \; (\mu,\sigma^2) \; \text{from Gaussian VI} \\ q_0 & = & \mathcal{N}(\mu,\Sigma), & (\mu,\Sigma) \; \text{from Gaussian VI} \\ q_0 & = & \mathcal{N}(\mu,\Sigma), & (\mu,\Sigma) \; \text{optimized as part of } Q \end{array}$$

Starting distribution

 q_0 closer to $p_T \Longrightarrow$ less distance for π_t to travel

$$\begin{array}{lll} q_0 & = & \mathcal{N}(0,I) \\ q_0 & = & \mathcal{N}(\mu, \operatorname{diag}(\sigma^2)), \; (\mu,\sigma^2) \; \text{from Gaussian VI} \\ q_0 & = & \mathcal{N}(\mu,\Sigma), & (\mu,\Sigma) \; \text{from Gaussian VI} \\ q_0 & = & \mathcal{N}(\mu,\Sigma), & (\mu,\Sigma) \; \text{optimized as part of } Q \end{array}$$

Normalizing flows?

Bridging distributions

better path π_t between q_0 and $p_T \Longrightarrow q_t$ closer to π_t

Bridging distributions

better path π_t between q_0 and $p_T \Longrightarrow q_t$ closer to π_t

$$\begin{array}{lcl} \bar{\pi}_t(z) & = & q_0(z)^{1-\beta_t}\bar{p}(z)^{\beta_t}, & \beta_t \text{ fixed} \\ \bar{\pi}_t(z) & = & q_0(z)^{1-\beta_t}\bar{p}(z)^{\beta_t}, & \beta_t \text{ optimized} \end{array}$$

Bridging distributions

better path $\pi_{\!\scriptscriptstyle t}$ between q_0 and $p_T \Longrightarrow q_t$ closer to $\pi_{\!\scriptscriptstyle t}$

$$ar{\pi}_t(z) = q_0(z)^{1-eta_t}ar{p}(z)^{eta_t}, \quad eta_t ext{ fixed} \ ar{\pi}_t(z) = q_0(z)^{1-eta_t}ar{p}(z)^{eta_t}, \quad eta_t ext{ optimized}$$

Something with more parameters?

better diffusion $\Longrightarrow q_t$ closer to π_t

overdamped Langevin
$$d\mathbf{z}_t = \nabla \log \pi_t(\mathbf{z}_t) dt + \sqrt{2} \ d\mathbf{w}_t$$
 $q_t(z) \approx \pi_t(z)$

better diffusion $\Longrightarrow q_t$ closer to π_t

overdamped Langevin
$$dz_t = \nabla \log \pi_t(z_t) dt + \sqrt{2} \ dw_t$$
 $q_t(z) \approx \pi_t(z)$ underdamped Langevin $dz_t = \rho_t dt$ $d\rho_t = \nabla \log \pi_t(z_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ dw_t$ $q_t(z,\rho) \approx \pi_t(z,\rho) = \pi_t(z) \mathscr{N}(\rho|0,I)$

better diffusion $\Longrightarrow q_t$ closer to π_t

overdamped Langevin
$$dz_t = \nabla \log \pi_t(z_t) dt + \sqrt{2} \ dw_t$$
 $q_t(z) \approx \pi_t(z)$ underdamped Langevin $dz_t = \rho_t dt$ $d\rho_t = \nabla \log \pi_t(z_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ dw_t$ $q_t(z,\rho) \approx \pi_t(z,\rho) = \pi_t(z) \mathcal{N}(\rho|0,I)$...with mass-matrix $dz_t = M^{-1}\rho_t dt$ $d\rho_t = \nabla \log \pi_t(z_t) dt - \gamma M^{-1}\rho_t dt + \sqrt{2\gamma} \ dw_t$

 $q_t(z, \rho) \approx \pi_t(z, \rho) = \pi_t(z) \mathcal{N}(\rho|0, M)$

better diffusion $\Longrightarrow q_t$ closer to π_t

overdamped Langevin
$$dz_t = \nabla \log \pi_t(z_t) dt + \sqrt{2} \ dw_t$$
 $q_t(z) \approx \pi_t(z)$ underdamped Langevin $dz_t = \rho_t dt$ $d\rho_t = \nabla \log \pi_t(z_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ dw_t$ $q_t(z,\rho) \approx \pi_t(z,\rho) = \pi_t(z) \mathscr{N}(\rho|0,I)$...with mass-matrix $dz_t = M^{-1}\rho_t dt$ $d\rho_t = \nabla \log \pi_t(z_t) dt - \gamma M^{-1}\rho_t dt + \sqrt{2\gamma} \ dw_t$ $q_t(z,\rho) \approx \pi_t(z,\rho) = \pi_t(z) \mathscr{N}(\rho|0,M)$

Higher-order Langevin? Time-dependent momentum distribution?

better $P \Longrightarrow \mathit{KL}(Q \| P)$ closer to $\mathit{KL}(q_T \| p_T)$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt + \sqrt{2} \ d\mathsf{w}_t$$

better $P \Longrightarrow \mathit{KL}(Q \| P)$ closer to $\mathit{KL}(q_T \| p_T)$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt + \sqrt{2} \ d\mathsf{w}_t$$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt - 2\nabla \log q_t(\mathsf{z}_t) dt + \sqrt{2} \ d\bar{\mathsf{w}}_t$$

better $P \Longrightarrow KL(Q||P)$ closer to $KL(q_T||p_T)$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt + \sqrt{2} \ d\mathsf{w}_t$$

Ideal reversal P:

$$dz_t = \nabla \log \pi_t(z_t) dt - 2\nabla \log q_t(z_t) dt + \sqrt{2} d\bar{w}_t$$

Simplest option:

$$d\mathsf{z}_t = -
abla \log \pi_t(\mathsf{z}_t) dt + \sqrt{2} \ dar{\mathsf{w}}_t$$
 (Tight if $q_t = \pi_t$)

better $P \Longrightarrow KL(Q||P)$ closer to $KL(q_T||p_T)$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt + \sqrt{2} \ d\mathsf{w}_t$$

 $dz_t = -\nabla \log \pi_t(z_t) dt + 2s_{\theta}(z_t, t) + \sqrt{2} d\bar{w}_t$

Ideal reversal
$$P$$
:

$$dz_t = -\nabla \log \pi_t(z_t)dt + \sqrt{2} \ d\bar{w}_t$$

$$d\mathsf{z}_t = \nabla \log \pi_t(\mathsf{z}_t) dt$$

$$dz_t = \nabla \log \pi_t(z_t) dt - 2\nabla \log q_t(z_t) dt + \sqrt{2} d\bar{w}_t$$

$$+$$
 $\sqrt{2}$ u w_t

(Tight if
$$q_t = \pi_t$$
)

(Tight if $s_{\theta} = \nabla \log \frac{\pi_t}{a_t}$)

$$dz_t = \rho_t dt$$

$$d\rho_t = \left[\nabla \log \pi_t(z_t) - \gamma \rho_t\right] dt + \sqrt{2\gamma} dw_t$$

$$dz_{t} = \rho_{t}dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) - \gamma \rho_{t}\right]dt + \sqrt{2\gamma} dw_{t}$$

Ideal reversal P:

$$d\mathbf{z}_{t} = \boldsymbol{\rho}_{t} dt$$

$$d\boldsymbol{\rho}_{t} = \left[\nabla \log \boldsymbol{\pi}_{t}(\mathbf{z}_{t}) - \gamma \boldsymbol{\rho}_{t} - 2\gamma \nabla_{\boldsymbol{\rho}} \log q_{t}(\boldsymbol{\rho}_{t}, \mathbf{z}_{t})\right] dt + \sqrt{2\gamma} d\bar{w}_{t}$$

$$dz_{t} = \rho_{t}dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) - \gamma \rho_{t}\right]dt + \sqrt{2\gamma} dw_{t}$$

Ideal reversal P:

$$dz_{t} = \rho_{t} dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) - \gamma \rho_{t} - 2\gamma \nabla_{\rho} \log q_{t}(\rho_{t}, z_{t})\right] dt + \sqrt{2\gamma} d\bar{w}_{t}$$

Simplest option

$$d\mathsf{z}_t =
ho_t \ dt$$
 (Tight if $q_t = \pi_t$, so $abla_{
ho} \log q_t = -
ho$)
 $d
ho_t = \left[
abla \log \pi_t(\mathsf{z}_t) + \gamma
ho_t \right] dt + \sqrt{2\gamma} dar{w}_t$

Backward Process P (underdamped Langevin)

$$dz_{t} = \rho_{t}dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) - \gamma \rho_{t}\right]dt + \sqrt{2\gamma} dw_{t}$$

ldeal reversal P:

$$dz_{t} = \rho_{t} dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) - \gamma \rho_{t} - 2\gamma \nabla_{\rho} \log q_{t}(\rho_{t}, z_{t})\right] dt + \sqrt{2\gamma} d\bar{w}_{t}$$

Simplest option

$$dz_t = \rho_t dt$$
 (Tight if $q_t = \pi_t$, so $\nabla_\rho \log q_t = -\rho$)
$$d\rho_t = \left[\nabla \log \pi_t(\mathsf{z}_t) + \gamma \rho_t\right] dt + \sqrt{2\gamma} d\bar{w}_t$$

Corrective score network:

$$dz_{t} = \rho_{t} dt$$

$$d\rho_{t} = \left[\nabla \log \pi_{t}(z_{t}) + \gamma \rho_{t} + 2\gamma s_{\theta}(t, z_{t}, \rho_{t})\right] dt + \sqrt{2\gamma} d\bar{w}_{t}$$
(Tight if $s_{\theta} = \nabla_{\rho} \log \frac{\pi_{t}}{q_{t}}$)

Discretization

$$q(z_{1:K}) = \underbrace{q(z_1)}_{\mathsf{Gaussian}} \prod_{k=1}^{K-1} F_k(z_{k+1}|z_k)$$

$$\bar{p}(z_{1:K}) = \underbrace{\bar{p}(z_K)}_{\mathsf{Target}} \prod_{k=1}^{K-1} B_k(z_k|z_{k+1})$$

$$\min KL(q(z_{1:K}) || p(z_{1:K}))$$

Splitting

Forward SDE

$$\begin{bmatrix} d\mathsf{z}_t \\ d\rho_t \end{bmatrix} = \begin{bmatrix} \rho_t dt \\ \nabla \log \pi_t(\mathsf{z}_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ d\mathsf{w}_t \end{bmatrix}$$

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \begin{bmatrix} \rho_t dt \\ [\nabla \log \pi_t(\rho_t) dt - \gamma \rho_t dt - 2\gamma s_{\theta}(t, z_t, \rho_t)] dt + \sqrt{2\gamma} d\bar{w}_t \end{bmatrix}$$

Splitting

Forward SDE

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{\mathbf{U}} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(\mathsf{z}_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ d\mathsf{w}_t \end{bmatrix}}_{\mathbf{V}}$$

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{\mathbf{U}'} + \underbrace{\begin{bmatrix} \nabla \log \pi_t(\rho_t) - \gamma \rho_t - 2\gamma s_{\theta}(t, z_t, \rho_t) \end{bmatrix} dt + \sqrt{2\gamma} d\bar{w}_t}_{\mathbf{V}'} \end{bmatrix}$$

Splitting

Forward SDE

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{\mathbf{U}} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(\mathsf{z}_t) dt - \gamma \rho_t dt + \sqrt{2\gamma} \ d\mathsf{w}_t \end{bmatrix}}_{\mathbf{V}}$$

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{\mathbf{U}'} + \underbrace{\begin{bmatrix} 0 \\ [\nabla \log \pi_t(\rho_t) - \gamma \rho_t - 2\gamma s_{\theta}(t, z_t, \rho_t)] dt + \sqrt{2\gamma} d\bar{w}_t \end{bmatrix}}_{\mathbf{V}'}$$

$$F_k = V U$$
 (U exact, V Euler-Maruyama, both for time δ) $B_k = U'V'$ (U' exact, V' Euler-Maruyama, both for time δ)

Better Splitting

Forward SDE

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{A} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(z_t) dt \end{bmatrix}}_{B} + \underbrace{\begin{bmatrix} 0 \\ -\gamma \rho_t dt + \sqrt{2\gamma} \ dw_t \end{bmatrix}}_{0}$$

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{A'} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(\rho_t) dt \end{bmatrix}}_{B'} + \underbrace{\begin{bmatrix} (-\gamma \rho_t - 2\gamma s_\theta(t, z_t, \rho_t)) dt + \sqrt{2\gamma} d\bar{w}_t \end{bmatrix}}_{O'}$$

Better Splitting

Forward SDE

$$\begin{bmatrix} dz_t \\ d\rho_t \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_t dt \\ 0 \end{bmatrix}}_{A} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(z_t) dt \end{bmatrix}}_{B} + \underbrace{\begin{bmatrix} 0 \\ -\gamma \rho_t dt + \sqrt{2\gamma} \ dw_t \end{bmatrix}}_{0}$$

$$\begin{bmatrix} d\mathbf{z}_t \\ d\boldsymbol{\rho}_t \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\rho}_t \ dt \\ 0 \end{bmatrix}}_{\mathbf{A}'} + \underbrace{\begin{bmatrix} 0 \\ \nabla \log \pi_t(\boldsymbol{\rho}_t) dt \end{bmatrix}}_{\mathbf{B}'} + \underbrace{\begin{bmatrix} (-\gamma \boldsymbol{\rho}_t - 2\gamma s_{\theta}(t, \mathbf{z}_t, \boldsymbol{\rho}_t)) dt + \sqrt{2\gamma} d\bar{w}_t \end{bmatrix}}_{\mathbf{0}'}$$

$$F_k = 0$$
 B A B (A exact, B Euler-Maruyama, O exact, B for $\delta/2$, others for δ) $B_k = B'A'B'O'$ (A' exact, B Euler-Maruyama, O' Euler-Maruyama)

Algorithm

Algorithm 1 Forward transition $F_k(z_{k+1}, \rho_{k+1}|z_k, \rho_k)$

Require: z_k , ρ_k , step-size δ

Re-sample momentum $\rho_k' \sim m_F(\rho_k'|\rho_k, \gamma, \delta)$

 $\begin{array}{l} \text{Update } \rho_k'' = \rho_k' + \frac{\delta}{2} \nabla \log \pi^{k\delta}(z_k) \\ \text{Update } z_{k+1} = z_k + \delta \rho_k'' \\ \text{Update } \rho_{k+1} = \rho_k'' + \frac{\delta}{2} \nabla \log \pi^{k\delta}(z_{k+1}) \end{array} \right\}$

return (z_{k+1}, ρ_{k+1})

Algorithm 3 Generating the augmented ELBO (eq. (5)).

Sample $(z_1, \rho_1) \sim q(z_1, \rho_1)$.

Initialize estimator as $\mathcal{L} \leftarrow -\log q(z_1, \rho_1)$.

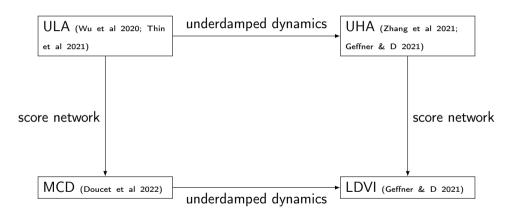
for $k = 1, 2, \dots, K - 1$ do

Run F_k (alg. 1) on (z_k, ρ_k) , store $\rho'_k, z_{k+1}, \rho_{k+1}$.

Update $\mathcal{L} \leftarrow \mathcal{L} + \log \frac{m_B(\rho_k|\rho_k',z_k,\gamma,\delta)}{m_B(\rho_k'|\rho_k,\gamma,\delta)}$.

Update $\mathcal{L} \leftarrow \mathcal{L} + \log \bar{p}(z_K, \rho_K)$. return \mathcal{L}

Results



Results

- Tune:
 - ▶ Initial distribution q_0 (diagonal Gaussian)
 - Discretization step-size
 - ightharpoonup Bridging densities' parameters eta
 - Damping coefficient
 - Score network (two hidden layers with residual connections)
- Train with $K \in \{8, 16, 32, 64, 128, 256\}$ bridging distributions.
- \bullet Optimize with Adam for 150k iters w/ learning rates of 10^{-3} , 10^{-4} , 10^{-5} , keep best.

Logistic Regression (1)

	Logistic regression (Ionosphere)					
	ULA	MCD	UHA	LDVI	UHA_{EM}	$LDVI_{\mathrm{EM}}$
K = 8	-116.4	-114.6	-115.6	-114.4	-117.7	-115.5
K = 16	-115.4	-113.6	-114.4	-113.1	-115.9	-113.8
K = 32	-114.5	-112.9	-113.4	-112.4	-114.6	-112.9
K = 64	-113.8	-112.5	-112.8	-112.1	-113.6	-112.4
K = 128	-113.1	-112.2	-112.3	-111.9	-113.1	-112.1
K = 256	-112.7	-112.1	-112.1	-111.7	-112.5	-111.9

Plain VI: −124.1

Logistic Regression (2)

	Logistic regression (Sonar)					
	ULA	MCD	UHA	LDVI	UHA_{EM}	$LDVI_{\mathrm{EM}}$
K = 8	-122.4	-117.2	-120.1	-116.3	-124.1	-118.5
K = 16	-119.9	-114.4	-116.8	-112.6	-119.9	-114.4
K = 32	-117.4	-112.4	-113.9	-110.6	-116.4	-111.7
K = 64	-115.3	-111.1	-111.9	-109.7	-113.8	-110.3
K = 128	-113.5	-110.2	-110.6	-109.1	-111.9	-109.6
K = 256	-112.1	-109.7	-109.7	-108.9	-110.7	-109.1

Plain VI: −138.6

 ${\sf ULA-Overdamped+Score\ net\ /\ UHA-Underdamped+Score\ net\$

Time Series (1)

	Brownian motion					
	ULA	MCD	UHA	LDVI	$\mathrm{UHA}_{\mathrm{EM}}$	$LDVI_{\mathrm{EM}}$
K = 8	-1.9	-1.4	-1.6	-1.1	-2.8	-2.8
K = 16	-1.5	-0.8	-1.1	-0.5	-2.2	-1.4
K = 32	-1.1	-0.4	-0.5	0.1	-1.6	-0.5
K = 64	-0.7	-0.1	0.1	0.5	-0.9	0.1
K = 128	-0.3	0.2	0.4	0.7	-0.4	0.4
K = 256	-0.1	0.5	0.6	0.9	0.1	0.6

Plain VI: -4.4

Time Series (2)

	Lorenz system					
	ULA	MCD	UHA	LDVI	UHA_{EM}	$LDVI_{\mathrm{EM}}$
K = 8	-1168.2	-1168.1	-1166.3	-1166.1	-1170.5	-1170.5
K = 16	-1165.7	-1165.6	-1163.1	-1162.2	-1169.8	-1166.8
K = 32	-1163.2	-1163.3	-1160.3	-1157.6	-1167.9	-1162.9
K = 64	-1160.9	-1161.1	-1157.7	-1153.7	-1161.3	-1161.4
K = 128	-1158.9	-1158.9	-1155.4	-1153.1	-1158.1	-1163.4
K = 256	-1157.2	-1157.1	-1153.3	-1151.1	-1163.1	-1154.6

Plain VI: −1187.8

Hierarchical

	Random effect regression (seeds)					
	ULA	MCD	UHA	LDVI	UHA_{EM}	$LDVI_{\mathrm{EM}}$
K = 8	-75.5	-75.1	-74.9	-74.9	-75.9	-75.5
K = 16	-75.2	-74.6	-74.6	-74.5	-75.1	-75.1
K = 32	-74.9	-74.3	-74.2	-74.2	-74.8	-74.8
K = 64	-74.6	-74.1	-74.1	-73.9	-74.4	-74.4
K = 128	-74.3	-73.9	-73.8	-73.7	-74.1	-74.1
K = 256	-74.1	-73.7	-73.7	-73.6	-73.9	-73.7

Plain VI: -77.1

Conclusions

Compared to diffusion models...

- Harder since marginals of Q not tractable
- Easier because have a good guess for score network

Experimentally...

- Underdamped dynamics help
- Learning a score network helps
- Better discretization helps
- Tuning more stuff helps

There's a lot we still don't understand...

Conclusions

Compared to diffusion models...

- ullet Harder since marginals of Q not tractable
- Easier because have a good guess for score network

Experimentally...

- Underdamped dynamics help
- Learning a score network helps
- Better discretization helps
- Tuning more stuff helps

There's a lot we still don't understand...

Thank you!

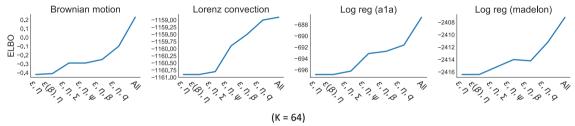
Joint work with Tomas Geffner



Langevin Diffusion Variational Inference AISTATS 2023. arXiv:2208.07743

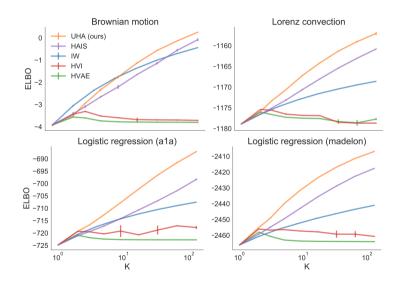
these slides: t.ly/9vZvk or
people.cs.umass.edu/domke/diffusion.pdf

Optimizing more stuff helps



- ε step size of HMC dynamics
- n damping coefficient
- Σ moment covariance
- β temperature schedule
- ψ "full rank" temperature schedule
- a initial distribution

Better than importance-weighted VI



UHA	Our algorithm
HAIS	Annealed Importance Sampling using HMC dynamics
IW	Importance Weighting
HVI	"Bridging the gap" using HMC dynamics
HVAE	(Recent algorithm)

VAEs

ELBO on test set

K = 8K = 32K = 64K = 1K = 8K = 16K = 32K = 64K = 1K = 16UHA -88.5-87.5-87.2-87.0-86.9UHA -93.4-89.8-88.8-88.1-87.6mnist mnist IW -93.4-90.5-89.9-89.4-89.0IW -88.5-87.6-87.5-87.3-87.2-129.9UHA -137.9-133.5-132.3-131.5-130.9UHA -131.9-130.7-130.3-130.1letters letters IW-137.9-134.6-133.9-133.2-132.7IW-131.9-130.9-130.7-130.6-130.4-184.2-176.6-174.6-173.2-171.6-174.3-172.2-171.6-171.2-170.2UHA UHA kmnist kmnist IW -184.2-179.7-178.7-177.8-177.0IW -174.3-173.0-172.6-172.4-172.2

log-likelihood on test set

Looseness

Finite time decomposition:

$$\underbrace{\mathit{KL}(q(z_{1:K}) \| p(z_{1:K}))}_{\text{what you optimize}} = \underbrace{\mathit{KL}(q(z_K) \| p(z_K))}_{\text{what you care about}} + \underbrace{\mathit{KL}(q(z_{1:K-1} | \mathbf{z}_K) \| p(z_{1:K-1} | \mathbf{z}_K))}_{\text{looseness}}$$

Analogous to continuous time decomposition:

$$\underbrace{\mathit{KL}(Q\|P)}_{\text{what you optimize}} = \underbrace{\mathit{KL}(q_T\|p_T)}_{\text{what you care about}} + \underbrace{\underset{\mathsf{z}_T \sim q_T}{\mathbb{E}} \mathit{KL}(Q(\cdot|\mathsf{z}_T)\|P(\cdot|\mathsf{z}_T))}_{\text{looseness}}$$

Ideal (intractable) transitions

$$B_k(z_k|z_{k+1}) = F_k(z_{k+1}|z_k) \frac{q(z_k)}{q(z_{k+1})}$$

would give $KL(q(z_{1:K})||p(z_{1:K})) = KL(q(z_K)||p(z_K))$.

Approximating these transitions analogous to approximating score function

SDEs

$$d\mathbf{z}_t = f(\mathbf{z}_t, t)dt + g(t)d\mathbf{w}_t$$
 $\mathbf{z}_{t+\delta} = \mathbf{z}_t + \delta f(\mathbf{z}_t) + g(t)\varepsilon_t$ $\varepsilon_0, \varepsilon_\delta, \dots \varepsilon_T \sim \mathcal{N}(0, \delta)$ $d\mathbf{z}_t = f(\mathbf{z}_t, t)dt + g(t)d\bar{\mathbf{w}}_t$ $\mathbf{z}_{t-\delta} = \mathbf{z}_t - \delta f(\mathbf{z}_t) + g(t)\varepsilon_t$

 $\varepsilon_T, \varepsilon_{T-\delta}, \cdots, \varepsilon_0, \cdots \sim \mathcal{N}(0, \delta)$