Sampling on Constrained Spaces

Elena Bortolato

Universitat Pompeu Fabra
Barcelona School of Economics

SISBAYES 2025, Padova 4-5 September

Based on joint work with
Pierre E. Jacob (ESSEC), Robin J. Ryder (Imperial College London)

Overview

1. Motivation
2. Manifold MCMC methods
3. Coupled markov chains
4. Sampling on *artificial* submanifolds

Sampling on Constrained Spaces: Why?

Many statistical problems involve **constraints** — on parameters, on data, or jointly on both. Closed-form solutions are often unavailable, making *sampling methods* essential.

Sampling on Constrained Spaces: Why?

Many statistical problems involve **constraints** — on parameters, on data, or jointly on both. Closed-form solutions are often unavailable, making *sampling methods* essential.

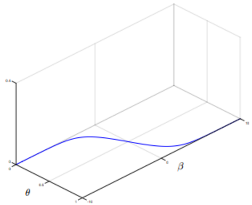
In some cases, complex sampling problems can be **reformulated** as problems with artificial constraints, opening the door to new methodological tools.

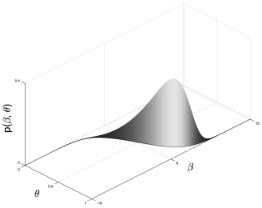
Overparametrized models [Bornn et al., 2019]

- Data with support $s_1, \ldots, s_J$, such that $Pr(y_i = s_j) = \theta_j$, nuisance parameters.
- β parameter of interest (e.g. regression)
- priors on both β and θ, overparametrized!

Moment-type conditions

$$E_y[g(y), \beta] = \int g(y, \beta) f(dy) = \sum_{i=1}^{J} \theta_j g(s_j, \beta) = 0$$





Example ($J = 2$): $y \in \{0, 1\}$
parameters $\theta, 1 - \theta$

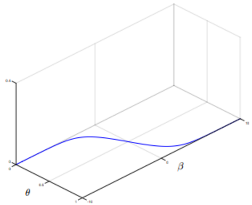
$$\left\{(\theta, \beta) \in \Theta \times \mathrm{B} | \beta = \log\left(\frac{\theta}{1-\theta}\right)\right\}$$

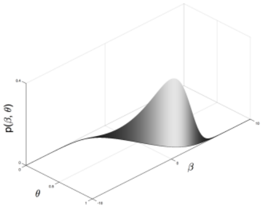
Overparametrized models [Bornn et al., 2019]

- Data with support $s_1, \ldots, s_J$, such that $Pr(y_i = s_j) = \theta_j$, nuisance parameters.
- β parameter of interest (e.g. regression)
- priors on both β and θ, overparametrized!

Moment-type conditions

$$E_y[g(y), \beta] = \int g(y, \beta) f(dy) = \sum_{i=1}^{J} \theta_j g(s_j, \beta) = 0$$





Example ($J = 2$): $y \in \{0, 1\}$
parameters $\theta, 1 - \theta$

$$\{(\theta, \beta) \in \Theta \times \mathrm{B} | \beta = \log\left(\frac{\theta}{1-\theta}\right)\}$$

... also in Empirical likelihood, Econometrics [Gallant, 2023]

Manifold from data generating equation

- Intractable likelihood: prior $\pi(\theta)$, random/model components $u \sim p(u|\theta)$, on

$$\mathcal{S} = \{(\theta,\ u_i) \in \Theta \times U | g(\theta, u_i) = y_i^{\text{obs}}, i = 1, \ldots, n\}.$$

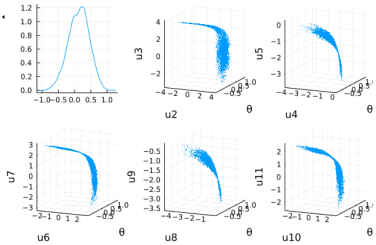
and keeping only θ, we sample $\pi(\theta|y^{\text{obs}})$.
[Graham and Storkey, 2017]

Manifold from data generating equation

- Intractable likelihood: prior $\pi(\theta)$, random/model components $u \sim p(u|\theta)$, on

$$\mathcal{S} = \{(\theta,\ u_i) \in \Theta \times U | g(\theta, u_i) = y_i^{\text{obs}}, i = 1, \ldots, n\}.$$

and keeping only θ, we sample $\pi(\theta|y^{\text{obs}})$.
[Graham and Storkey, 2017]



- Similar reasoning for models with intractable prior:

$$\mathcal{S}' = \{(\theta, u_i) \in \Theta \times U | y_i = g(u_i, \theta), q_j(y, \theta) + m_j(\theta) = 0,\ i = 1, \ldots, n,\ j = 1, \ldots, p\}.$$

[Bortolato and Ventura, 2024], q_j, m_j derivatives of loglikelihood/prior.

Overview

Sampling probability distribution functions on submanifolds

Object of interest:

$$\pi(x) = \frac{f(x)\mathbb{1}\{x \in \mathcal{S}\}}{Z}\sigma_{\mathcal{S}}, \quad f(x) \geq 0, \; Z \text{ normalizing constant},$$

$\mathcal{S} = \{x \in \mathbb{R}^D | q(x) = 0 \in \mathbb{R}^m\}$, submanifold

Sampling probability distribution functions on submanifolds

Object of interest:

$$\pi(x) = \frac{f(x)\mathbb{1}\{x \in \mathcal{S}\}}{Z}\sigma_{\mathcal{S}}, \quad f(x) \geq 0, \; Z \text{ normalizing constant},$$

$\mathcal{S} = \{x \in \mathbb{R}^D | q(x) = 0 \in \mathbb{R}^m\}$, submanifold

Difficulties:

1. Z unknown
2. Proposing points in $\mathbb{R}^D$ won't work.

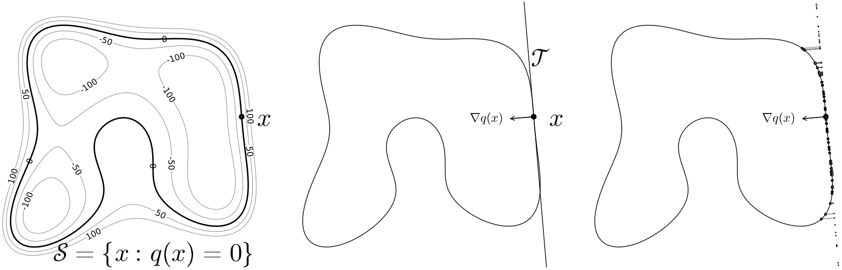
Andersen, 1983, Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations.

Metropolis–Rosenbluth–Teller–Hastings on submanifolds

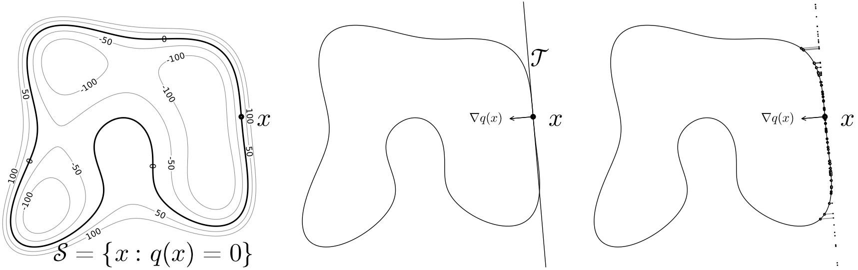
Random walk [Zappa et al., 2018]:
define $\nabla q(x)$ gradient, $\mathcal{T}_x = \{x^* \in \mathbb{R}^D | \nabla q(x)^\top (x^* - x) = 0\}$, $d = D - m$.

1. Start from $x \in \mathcal{S}$.
2. Compute U_x, an orthonormal basis of $\mathcal{T}_x$.
3. Draw d-dimensional $\nu \sim p_\nu$ and propose a step on $\mathcal{T}_x$: $x + U_x \nu$.
4. Follow the direction given by $\nabla q(x)$ to project on $\mathcal{S}$: $y = x + U_x \nu + \nabla q(x)\alpha$ for some α such that $y \in \mathcal{S}$ (Projection).
5. Check whether x can be reached from y (Reverse projection).
6. Accept/reject.

Projections



Projections

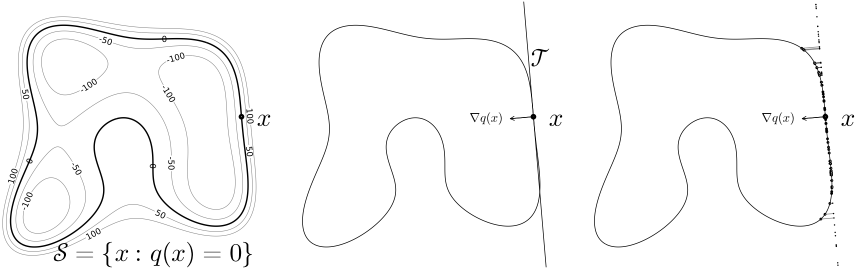


Projections (and reverse projections) employ Newton's method to find a root of

$$q(x + U_x \nu + \nabla q(x)\alpha)$$

moving $\alpha \in \mathbb{R}^m$.

Projections



Projections (and reverse projections) employ Newton's method to find a root of

$$q(x + U_x\nu + \nabla q(x)\alpha)$$

moving $\alpha \in \mathbb{R}^m$.

If a solution is not found, the chain remains at its current state.

A closer look at the proposal

The proposal distribution can be written as

$$q(x, dy) = r(x)\delta_x(dy) + (1 - r(x))\underbrace{|\det DG_x(y)|p_\nu(\nu)\sigma_{\mathcal{S}}(dy)}_{k(x,dy).}. \tag{1}$$

$DG_x(y) = U_x^\top U_y$ is the differential of the map G_x,

$$G_x(y) =: U_x^\top(y - x) = \nu. \tag{2}$$

defines a one-to-one relation among y and ν.

Acceptance probability

Acceptance ratio evaluated only when the projection steps succeed

$$\frac{f(y)|\det DG_y(x)|p_\nu(\nu')}{f(x)|\det DG_x(y)|p_\nu(\nu)},$$

The determinants cancel out: $|\det U_x^\top U_y| = |\det U_y^\top U_x|$.

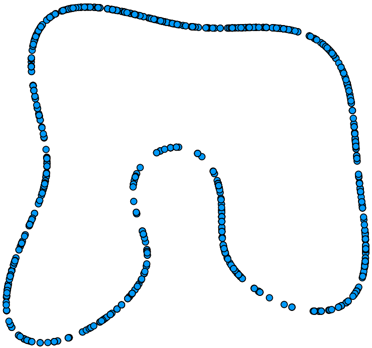
Efficiency - Convergence ?

1. Computational cost: $O(m^2 D)$

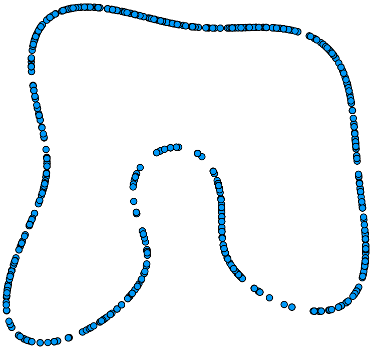
Efficiency - Convergence ?

1. Computational cost: $O(m^2 D)$
2. How long should the chain run?

Uniform distribution on $\mathcal{S}$

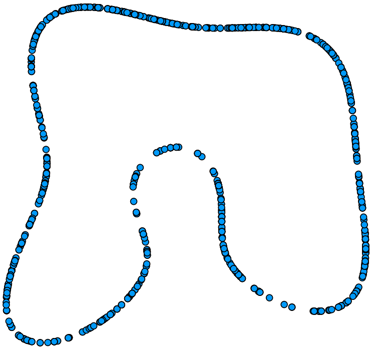


Uniform distribution on $\mathcal{S}$

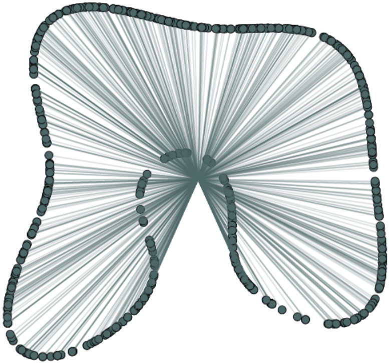


converged?

Uniform distribution on $\mathcal{S}$

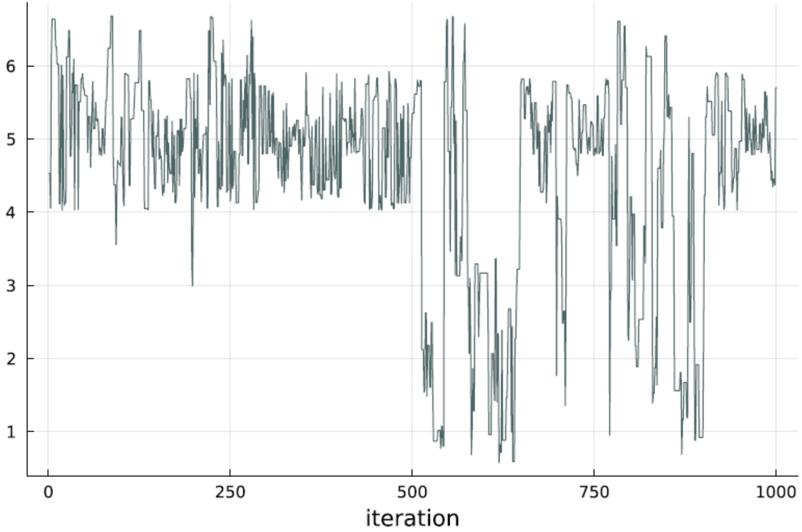


converged?



test function: $x \mapsto \|x\|$

Traceplot of $||x||$



Overview

Coupled Markov chains on submanifolds

Produce pairs of MCMC chains (X_t) and $(\tilde{X}_t)$, with $\text{law}(X_t) = \text{law}(\tilde{X}_t)$ that after a random meeting time $\tau \in \mathbb{N}$ for all $t \geq \tau$, $X_t = \tilde{X}_{t-L}$.

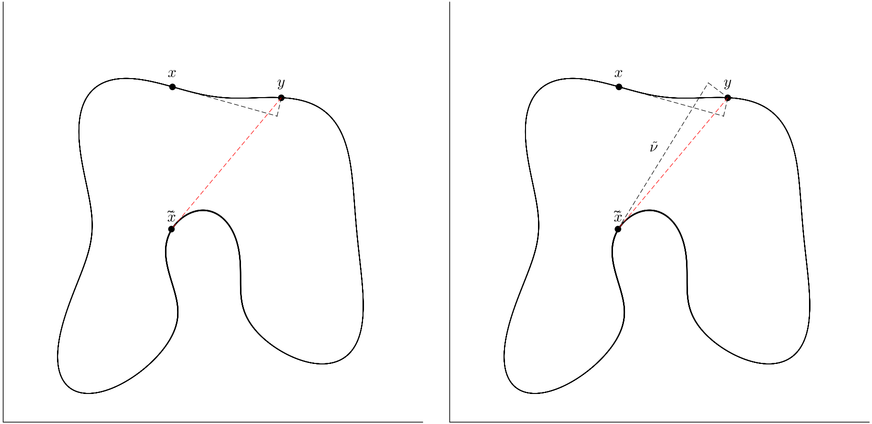
Coupled Markov chains on submanifolds

Produce pairs of MCMC chains (X_t) and $(\tilde{X}_t)$, with $\text{law}(X_t) = \text{law}(\tilde{X}_t)$ that after a random meeting time $\tau \in \mathbb{N}$ for all $t \geq \tau$, $X_t = \tilde{X}_{t-L}$.

Monte Carlo estimates based on independent copies of τ:

- Bounds [Biswas et al., 2019] $|\pi_t - \pi|_{TV} \leq \mathbb{E}\left[\max\left(0, \left\lceil \frac{\tau-L-t}{L} \right\rceil\right)\right], \forall t.$
- Unbiased estimates of functions $h(X)$ [Glynn and Rhee, 2014][Jacob et al., 2020].
- Asymptotic variance of the chains [Douc et al., 2022].

Proposing the same point



$$\tilde{\nu} = G_{\tilde{x}}(y) = U_{\tilde{x}}^{\top}(y - \tilde{x})$$

Coupling of proposals with point masses

Coupled transition kernels

$$q(x, dy) = r(x)\delta_x(dy) + (1 - r(x))k(x, dy),$$
$$q(\tilde{x}, dy) = r(\tilde{x})\delta_{\tilde{x}}(dy) + (1 - r(\tilde{x}))k(\tilde{x}, dy),$$

$Y \sim q(x, dy)$ and $\tilde{Y} \sim q(\tilde{x}, dy)$ s.t. $Y = \tilde{Y}$ sometimes

without evaluating $r(x)$.

Coupling of proposals with point masses

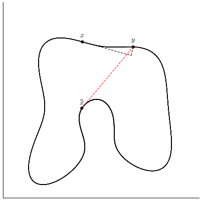
Coupled transition kernels

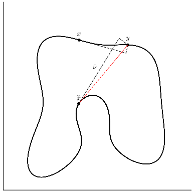
$$q(x, dy) = r(x)\delta_x(dy) + (1 - r(x))k(x, dy),$$
$$q(\tilde{x}, dy) = r(\tilde{x})\delta_{\tilde{x}}(dy) + (1 - r(\tilde{x}))k(\tilde{x}, dy),$$

$Y \sim q(x, dy)$ and $\tilde{Y} \sim q(\tilde{x}, dy)$ s.t. $Y = \tilde{Y}$ sometimes

without evaluating $r(x)$.

- Draw $Y \sim q(x, dy)$, $W \sim \text{Uniform}(0, 1)$.
- If $Y \neq x$ and $W \leq k(\tilde{x}, Y)/k(x, Y)$ then (Y, Y)
- Else, enter while loop:
  - Draw $\tilde{Y} \sim q(\tilde{x}, dy)$.
  - If $\tilde{Y} = \tilde{x}$, return $(Y, \tilde{Y})$.
  - Else draw $W^* \sim \text{Uniform}(0, 1)$.
  - If $W^* > k(x, \tilde{Y})/k(\tilde{x}, \tilde{Y})$, return $(Y, \tilde{Y})$.





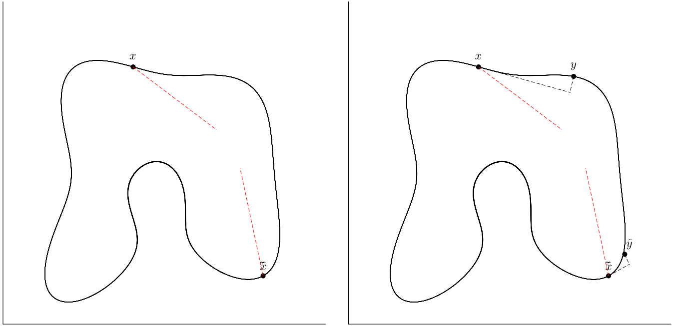
$\tilde{v}$ used in $k(\tilde{x}, dy)$

Reflection-contractive couplings

With the previous coupling...

- If chains are distant, they evolve independently
- Reflection couplings in the ambient space + rotation / projections help in obtaining meeting times faster

(Intuition)

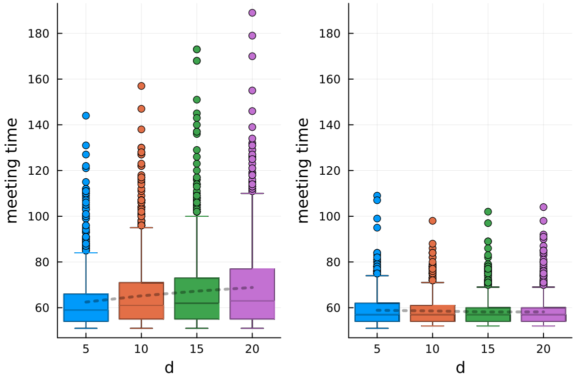


Scaling properties: sequence of Hyperspheres

Setting: Uniform distribution on $\mathcal{HS}^d = \{x \in \mathbb{R}^D \mid \sum_{i=1}^{D} x_i^2 = 1\}, d = D-1 \in \{5, 10, 15, 20\}$

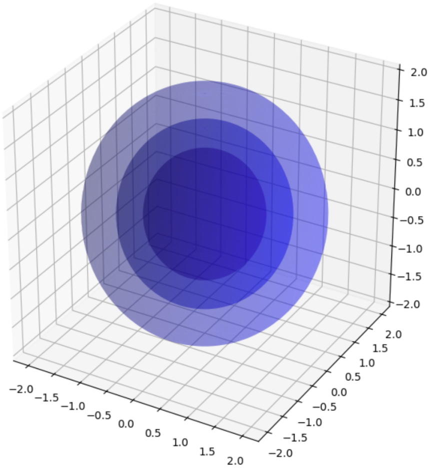
M Maximal coupling only

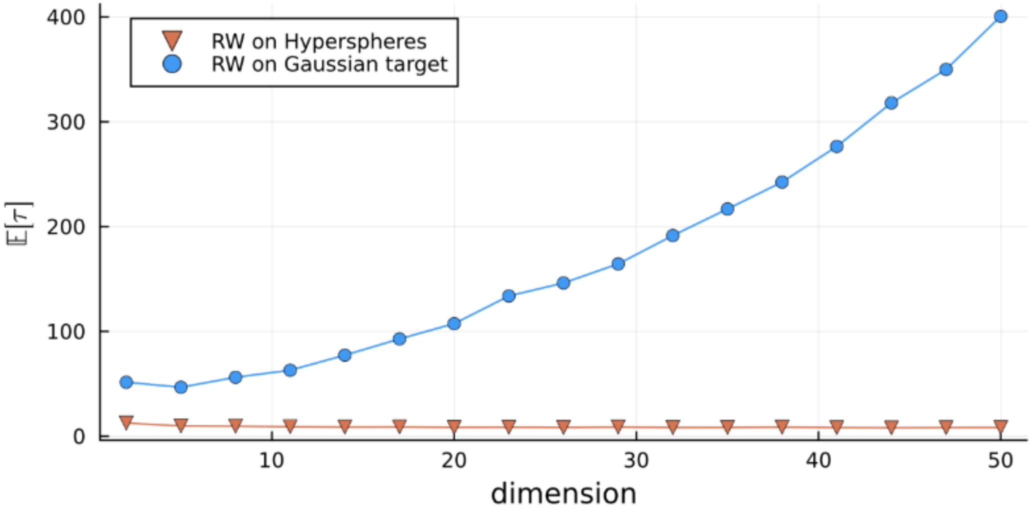
M+R Maximal coupling + reflections if $\|x - \tilde{x}\|_2^2 > 1/\sqrt{d} = \sigma$ (scale of the proposal kernel)



Can we use that beyond sampling on hyperspheres?

Can we use that beyond sampling on hyperspheres?
Hypershperes with different radii are level sets of the Gaussian...




RW on Hyperspheres
RW on Gaussian target
400
300
200
100
0
$\mathbb{E}[\tau]$
10
20
30
40
50
dimension

Overview

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Algorithm

- fix $q_0 = q(x)$
- move on $\mathcal{S} = \{x \in \mathcal{X} | q(x) - q_0 = 0\}$ leaving π_t invariant.
- change level set leaving π_a invariant (e.g. random walk).

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Algorithm

- fix $q_0 = q(x)$
- move on $\mathcal{S} = \{x \in \mathcal{X} | q(x) - q_0 = 0\}$ leaving π_t invariant.
- change level set leaving π_a invariant (e.g. random walk).

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Algorithm

- fix $q_0 = q(x)$
- move on $\mathcal{S} = \{x \in \mathcal{X} | q(x) - q_0 = 0\}$ leaving π_t invariant.
- change level set leaving π_a invariant (e.g. random walk).

Intuition

Moves on contour sets follow "relevant directions" for the target π_a (large steps).

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Algorithm

- fix $q_0 = q(x)$
- move on $\mathcal{S} = \{x \in \mathcal{X} | q(x) - q_0 = 0\}$ leaving π_t invariant.
- change level set leaving π_a invariant (e.g. random walk).

Intuition

Moves on contour sets follow "relevant directions" for the target π_a (large steps).
explore "fast" the $D-1$ contour sets of π_a,

Contour walk

Taking $q(\cdot) := \pi_a(\cdot)$ the disintegration defines a distribution on the level sets.

Algorithm

- fix $q_0 = q(x)$
- move on $\mathcal{S} = \{x \in \mathcal{X} | q(x) - q_0 = 0\}$ leaving π_t invariant.
- change level set leaving π_a invariant (e.g. random walk).

Intuition

Moves on contour sets follow "relevant directions" for the target π_a (large steps).
explore "fast" the $D - 1$ contour sets of π_a,

Connections with:

Andersen and Diaconis [2007] Hit and run as a unifying device.
Ludkin and Sherlock [2023] Hug and hop: a discrete-time, nonreversible Markov chain Monte Carlo algorithm.

Disintegration [Chang and Pollard, 1997]

For π_a on $\mathbb{R}^D$ and $q : \mathbb{R}^D \to \mathbb{R}^m$ measurable, let $q\sharp\pi_a$ be the distribution of $q(X)$ where $X \sim \pi_a$.

There exists a unique set of measures $(\pi_t)_{t\in\mathbb{R}^m}$ on $\mathbb{R}^D$, called a q-disintegration of π_a, such that:

- $\pi_t(\{x : q(x) \neq t\}) = 0$ for $q_\sharp\pi_a$-a.e. t (**concentration**)
- for $f : \mathbb{R}^D \to \mathbb{R}_+$ non-negative and measurable,

$$\int f(x)\pi_a(dx) = \int \left(\int f(x)\pi_t(dx) \right) q_\sharp\pi_a(dt).$$

Disintegration [Chang and Pollard, 1997]

For π_a on $\mathbb{R}^D$ and $q : \mathbb{R}^D \to \mathbb{R}^m$ measurable, let $q\sharp\pi_a$ be the distribution of $q(X)$ where $X \sim \pi_a$.
There exists a unique set of measures $(\pi_t)_{t\in\mathbb{R}^m}$ on $\mathbb{R}^D$, called a q-disintegration of π_a, such that:

- $\pi_t(\{x : q(x) \neq t\}) = 0$ for $q_\sharp\pi_a$-a.e. t (**concentration**)
- for $f : \mathbb{R}^D \to \mathbb{R}_+$ non-negative and measurable,

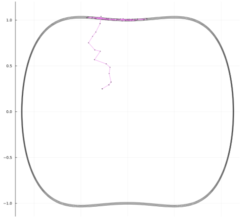
$$\int f(x)\pi_a(dx) = \int \left(\int f(x)\pi_t(dx) \right) q_\sharp\pi_a(dt).$$

By the co-area formula

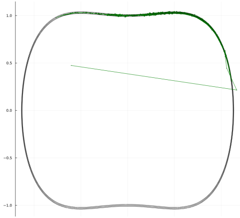
$$\pi_t(x) \propto \pi_a(x)\det(\nabla q(x)\nabla q(x)^\top)^{1/2}\sigma_{q(t)^{-1}}.$$

Example taken from Au et al. (2023): 1000 iterations, initializing from $N_2(0,1)$.

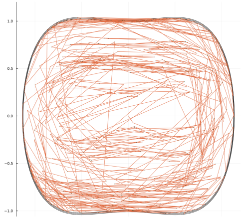
Random walk



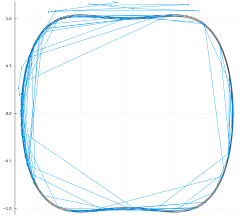
HMC leapfrog



Hug and Hop



Walking on curves



Thanks for your attention!

Under the support of CNS2022-135963 - MCIN/AE

Some references I

H. C. Andersen and P. Diaconis. Hit and run as a unifying device. *Journal de la societe francaise de statistique & revue de statistique appliquee*, 148(4):5–28, 2007.

K. X. Au, M. M. Graham, and A. H. Thiery. Manifold lifting: scaling markov chain monte carlo to the vanishing noise regime. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad023, 2023.

N. Biswas, P. E. Jacob, and P. Vanetti. Estimating convergence of Markov chains with L-lag couplings. *Advances in Neural Information Processing Systems*, 32, 2019.

L. Bornn, N. Shephard, and R. Solgi. Moment conditions and Bayesian non-parametrics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):5–43, 2019.

E. Bortolato. Advancement in distribution sampling for statistical inference, Chapters 4 and 5. *PhD Thesis*, 2024.

E. Bortolato and L. Ventura. Objective priors for invariant e-values in the presence of nuisance parameters. *Entropy*, 26(1):58, 2024.

Some references II

R. Douc, P. E. Jacob, A. Lee, and D. Vats. Solving the poisson equation using coupled markov chains. *arXiv preprint arXiv:2206.05691*, 2022.

A. R. Gallant. Variance–covariance from a Metropolis chain on a curved, singular manifold. *Journal of Econometrics*, 235(2):843–861, 2023.

P. W. Glynn and C.-h. Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

M. Graham and A. Storkey. Asymptotically exact inference in differentiable generative models. In *Artificial Intelligence and Statistics*, pages 499–508. PMLR, 2017.

P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3), 2020.

M. Ludkin and C. Sherlock. Hug and hop: a discrete-time, nonreversible markov chain monte carlo algorithm. *Biometrika*, 110(2):301–318, 2023.

E. Zappa, M. Holmes-Cerfon, and J. Goodman. Monte carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, 71(12):2609–2647, 2018.

Context: submanifolds in likelihood-free inference

- Models with intractable posterior [Graham and Storkey, 2017]
- Link with Approximate Bayesian Computation: Sampling $\theta^* \sim \pi(\theta)$, $u \sim p(u|\theta)$, $y^* = g(u, \theta^*)$, retain θ^* s.t. y^* close to y^{obs},

$$\pi(\theta|y^{\text{obs}})^{\text{ABC}} \propto \pi(\theta)p(u|\theta)\mathbb{1}_{\{|g(u,\theta)-y^{\text{obs}}|\leq\epsilon\}}.$$

As $\epsilon \to 0$ is defined on the submanifold

$$\mathcal{S} = \{(\theta,\ u_i) \in \Theta \times U | g(\theta, u_i) = y_i^{\text{obs}}, i = 1, \ldots, n\}.$$

By sampling on $\mathcal{S}$ and keeping only θ, we sample $\pi(\theta|y^{\text{obs}})$.

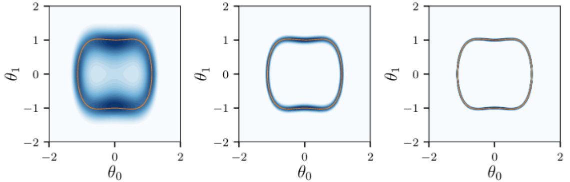
ABC-type problem: sum of lognormals

- $\theta \sim \text{Normal}(0, 1)$ (parameter)
- $u_\ell \sim \text{Normal}(0, 1)$ for $\ell = 1, \ldots, K$ (random input)
- $y = \sum_{\ell=1}^{K} \exp(\theta + u_\ell)$ (relation between observation, parameter and input)
- With $g : (u_{i1}, \ldots, u_{iK}, \theta) \mapsto \sum_{\ell=1}^{K} \exp(\theta + u_{i\ell}), \quad i = 1, \ldots, n$
  $g(u_{i1}, \ldots, u_{iK}, \theta) - y_i^{\text{obs}} = 0, \quad i = 1, \ldots, n$ (manifold constraints)

Ambient dimension: $K \times n + 1$. Number of constraints: n.

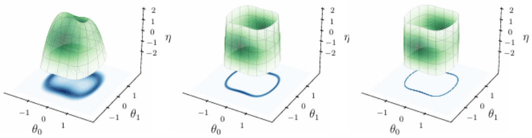
Context: submanifolds in Bayesian statistics

- Models with additive noise, with $\pi(\theta|y^{obs})$ strongly anisotropic



- State space augmentation: $D = d + n = \dim(\theta) + \dim(y^{obs})$[Au et al., 2023]

$$\mathcal{S} = \{(\theta,\ u_i) \in \Theta \times U | g(\theta, u_i) = y_i^{obs}, i = 1, \ldots, n\}.$$



Another view on the proposal

From x on $\mathbb{R}^D$ (ambient space), the proposal z on $\mathcal{T}_x$ can be obtained either

- by drawing $\nu \sim \text{Normal}(0, \Sigma)$ *for a fixed* Σ

  and computing $z = x + U_x \nu$

- by drawing $\xi \sim \text{Normal}(0, \Sigma_a)$

  with $\Sigma_a = \begin{pmatrix} \Sigma^\star & C \\ C' & \Sigma \end{pmatrix}$,

  and computing $z = x + P_x Q_x \xi$,

  with Q_x the Q matrix of the QR decomposition of $\nabla q(x)$
  $P_x = I_D - N_x N_x'$ orthogonal projector onto $\mathcal{T}_x$,
  N_x the first m columns of Q_x

Hug and Hop

Hug: propose moves nearly on same contour level. The state of the chain is (x, v). The velocity v is updated as:

$$v_{b+1} = v_b - 2\frac{v_b^\top g(x_b')}{\|g(x_b')\|^2} g(x_b'), \quad \text{with } g(x) = \nabla \log \pi(x)$$

.

The position becomes $x_{b+1} = x_b + \delta v_{b+1}$, $b = 1, \dots, B$, δ is the step size,

Hug and Hop

Hug: propose moves nearly on same contour level. The state of the chain is (x, v). The velocity v is updated as:

$$v_{b+1} = v_b - 2\frac{v_b^\top g(x_b')}{\|g(x_b')\|^2} g(x_b'), \quad \text{with } g(x) = \nabla \log \pi(x)$$

.

The position becomes $x_{b+1} = x_b + \delta v_{b+1}$, $b = 1, \ldots, B$, δ is the step size,

Hop: jumps between contours:

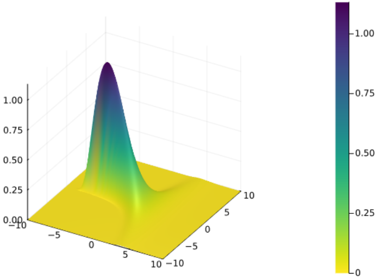
$$x \sim \mathcal{N}\left(x', \frac{\mu^2}{\|g(x')\|^2} I + \frac{\lambda^2 - \mu^2}{\|g(x')\|^4} g(x')g(x')^\top\right),$$

where, λ controls jumps along the gradient, and μ controls jumps perpendicular to the gradient.

Sampling on generic targets

Take $g(\theta) = \log(\pi(\theta))$, define $\mathcal{S} = \{\theta \in \Theta, u = g(\theta) \in \mathbb{R} | g(\theta) - u = 0\}$

Sample on the graph of the function:



- $d+1$ dimensional state space, $m = 1$ constraint
- adapting to the curvature of the target without computing second derivatives
- target distribution on $\mathcal{S}$

$$\pi_g(x) \propto \pi(x) G^*(x)^{-1/2}, G^*(x) = 1 + \|\nabla q(x)\|^2.$$