Couplings for MCMC on submanifolds

Elena Bortolato (Padova), Pierre E. Jacob (ESSEC), Robin J. Ryder (Dauphine)

Journée AppliBUGS

June 13, 2023

Overview

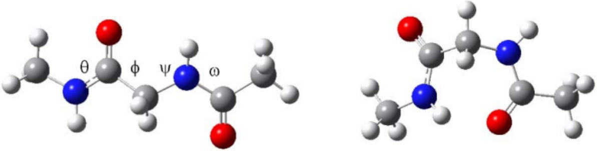
1. Submanifolds in statistics
2. MCMC on submanifolds
3. Coupling on submanifolds

Submanifolds in statistics - statistical mechanics

Andersen [1983] *Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations*

- Consider a system of 3-dimensional particles with configuration q evolving according to the motion equation and energy $V(q)$.
- Compute $\mathbb{E}_\mu[f(q)]$ with respect to the Gibbs measure $\mu(q) \propto \exp(-\frac{1}{\beta}V(q))$

  under some specific constraints on the angles ϕ ψ.



Glycine molecule, from Hartmann [2008]

$$\mathcal{S} = \{q \in Q | \psi = \psi_0, \phi = \phi_0\}$$

Submanifolds in statistics - ABC

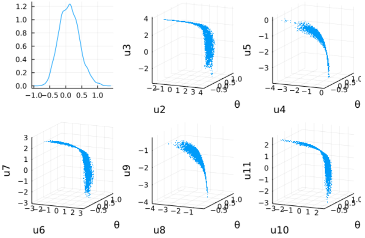
- Models with intractable likelihood function: Approximate Bayesian Computation
  Define $y^* = g(u, \theta^*)$, $\theta^* \sim \pi(\theta)$, $u \sim p(u|\theta)$, then

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

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

$$\mathcal{S} = \{(\theta,\ u_i) \in \Theta \times U | g(\theta, u_i) = y_i, \forall i\}.$$

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



[Graham and Storkey, 2017]

Submanifolds

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

is the zero level set of a smooth function $q : \mathbb{R}^D \to \mathbb{R}^m$.

- D: dimension of the ambient space
- Assume: $x \mapsto q_j(x)$ is C^∞ for all $1 \leq j \leq m$
- $\nabla q(x)$: $D \times m$ Jacobian matrix
- Assume: $\text{rank}(\nabla q(x)) = m$ for all $x \in \mathcal{S}$

$\mathcal{S}$ is of dimension $d := D - m$.

Probability distributions on submanifolds

Interest: probability distributions with density

$$\pi(x) = \frac{f(x)\mathbb{1}\{x \in \mathcal{S}\}}{Z}, \quad f(x) \geq 0, \quad Z = \int f(x)\sigma^d_{\mathcal{S}}(dx),$$

where $\sigma^d_{\mathcal{S}}$ is the Hausdorff/surface measure on $\mathcal{S}$.

MCMC algorithms can be used for drawing values from π starting from a initial position on $\mathcal{S}$ [Lelievre et al., 2012], [Zappa et al., 2018].

Open questions

- Diagnostics of convergence/choosing tuning parameters
- Compare the performance of different MCMC algorithms
- Parallelize computation

we aim to address them using *couplings*.

Coupled Markov chains

Design kernels for running couples of chains (X_t) and (Y_t), such that $\text{law}(X_t) = \text{law}(Y_t)$, and $X_t = Y_{t-L}$ for all $t \geq \tau$, almost surely ($L \in \mathbb{N}$ lag between chains, $\tau \in \mathbb{N}$ random meeting time).

With independent copies of τ, we can obtain Monte Carlo estimates of:

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

Our contribution:
design couplings of MCMC algorithms for distributions on submanifolds
e.g. Zappa et al. [2018].

Overview

1. Submanifolds in statistics
2. MCMC on submanifolds
3. Coupling on submanifolds

Metropolis–Rosenbluth–Teller–Hastings

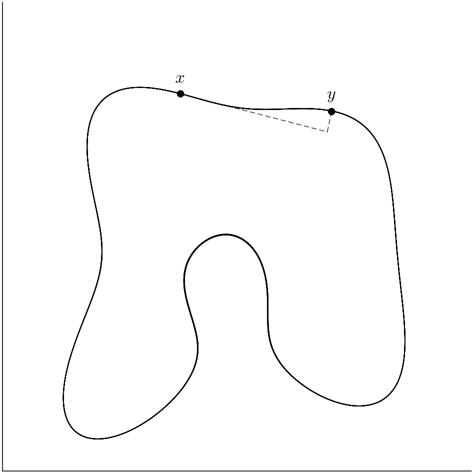
One step of random walk
(standard in $\mathbb{R}^D$)

1. Start from x.
2. Draw ϵ and propose $y = x + \epsilon$.
3. Accept/reject.

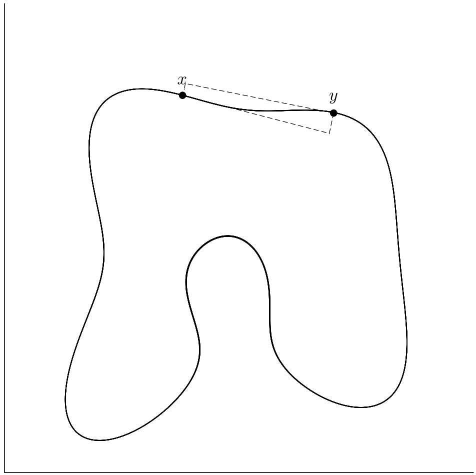
... and random walk on the tangent space [Zappa et al., 2018]:
$\mathcal{T}_x = \{x^* \in \mathbb{R}^D | \nabla q(x)^\top (x^* - x) = 0\}$.

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.

...in a picture



...in a picture



Projections

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

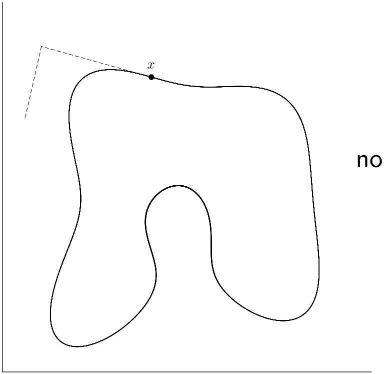
$$q(x + U_x\nu + \nabla q(x)\alpha)$$

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

There might not be a solution. Even if there is a solution, the number of iterations is limited and Newton's method can fail.

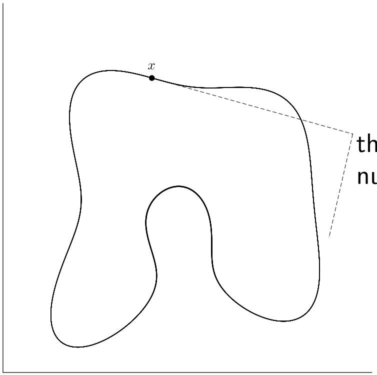
In these cases the chain remains at its current state.

Failure of projections



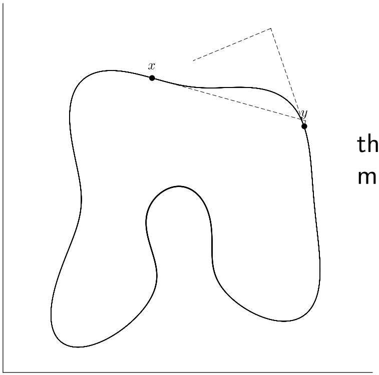
no solution

Failure of projections



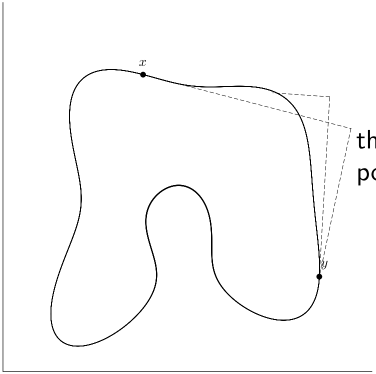
the projection fails with a prefixed maximum number of iterations

Failure of projections



the reverse projection fails with a prefixed maximum number of iterations

Failure of projections



the reverse projection delivers a different point

A closer look at the proposal

Denote by $G_x : \mathcal{S} \to \mathbb{R}^d$,

$$G_x(y) =: U_x^\top (y - x) = \nu. \tag{1}$$

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

The proposal distribution $q(x, dy)$ can be written as

$$q(x, dy) = r(x)\delta_x(dy) + (1 - r(x))|\det DG_x(y)|p_\nu(\nu)\sigma_{\mathcal{S}}(dy). \tag{2}$$

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

Acceptance probability

Acceptance ratio evaluated only when the projection steps succeed

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

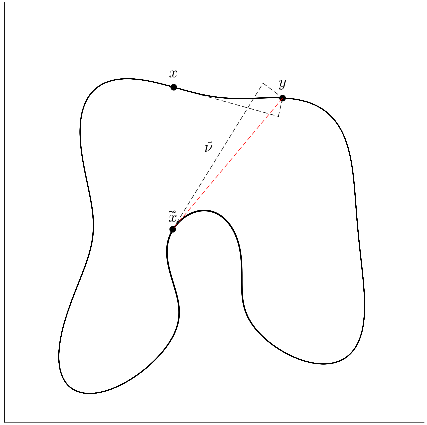
where f is target density, and $\nu' = G_y(x)$.

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

Overview

1. Submanifolds in statistics
2. MCMC on submanifolds
3. Coupling on submanifolds
   - Designing maximal couplings
   - Improving convergence

Proposing the same point



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

Coupling of probability distributions with point masses

The distributions to couple is

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

Simplify: $k(x, dy) = |\det DG_x(y)|p_\nu(\nu)\sigma_{\mathcal{S}}(dy)$.

We would like to couple two transition kernels of the form:

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

such that the two chains can meet: $(Y, \tilde{Y})$ with $Y \sim q(x, dy)$ and $\tilde{Y} \sim q(\tilde{x}, dy)$ can be such that $Y = \tilde{Y}$.

Coupling of probability distributions with point masses

- Draw $Y \sim q(x, dy)$, draw $W \sim \text{Uniform}(0,1)$.
- If $Y \neq x$ and $W \leq k(\tilde{x}, Y)/k(x, Y)$, return (Y, Y) (identical states).
- 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 $(Y, \tilde{Y})$.

Very similar setting to [Wang et al., 2021].

Coupling of random walk proposals on submanifolds

Remarkably we don't need to evaluate $r(x)$.

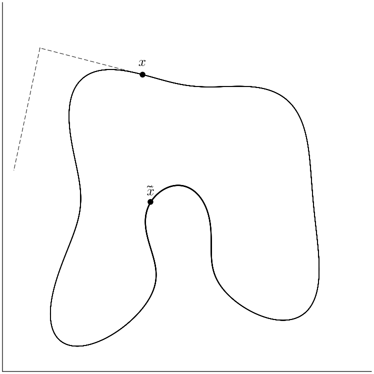
The algorithm requires evaluating ratios of the form

$$\frac{k(\tilde{x}, y)}{k(x, y)} = \frac{|\det DG_{\tilde{x}}(y)| p_\nu(G_{\tilde{x}}(y))}{|\det DG_x(y)| p_\nu(G_x(y))},$$

this time the determinants do not cancel out

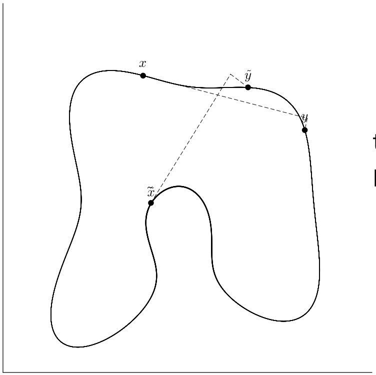
computational cost changes.

Two chains failing to meet



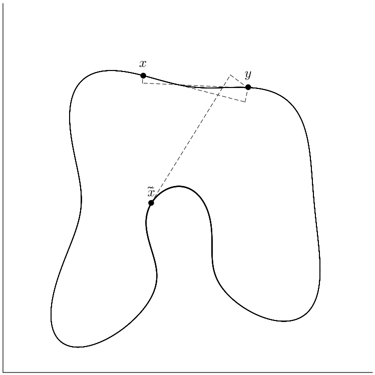
projection from one of the chains fails

Two chains failing to meet



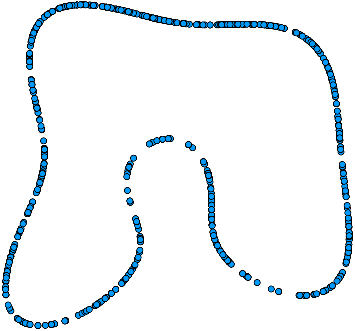
the secondary chain delivers a different point

Two chains failing to meet

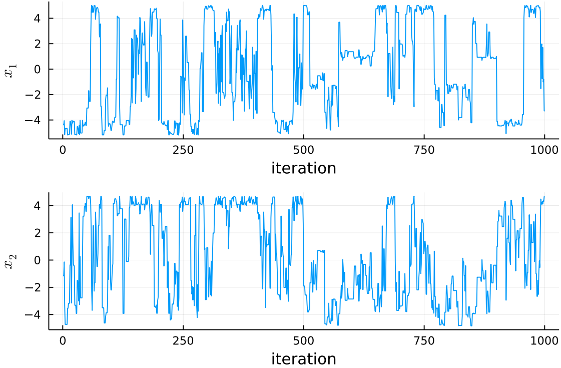


reverse projection fails

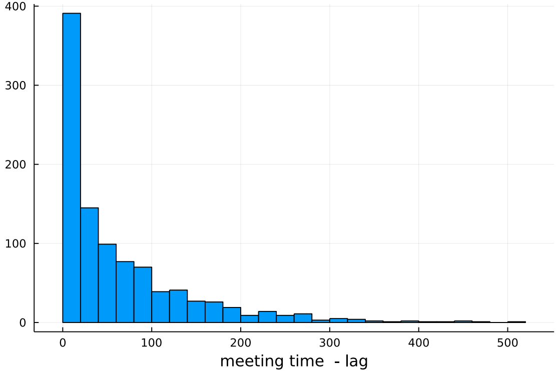
Example: 1000 iterations of RW



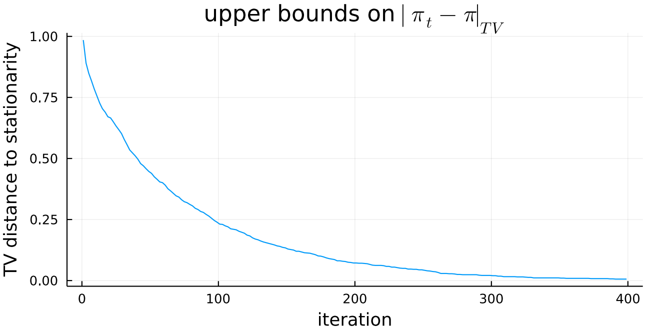
Mixing?


x_1
4
2
0
−2
−4
0
250
500
750
1000
iteration
x_2
4
2
0
−2
−4
0
250
500
750
1000
iteration

Meeting times



Upper bounds on TV to stationarity



Meeting times depend on

- properties of the base-algorithm
  (proposal standard deviation, number of iterations in Newton's method, initial distribution...)
- efficiency of the coupling strategy
  (bound improves to some degree when lag L increases...)

With the previous coupling. . .

- If chains are distant, they evolve independently and rarely meet.
- The problem is exacerbated in high dimensions.

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

Ambient proposal for defining reflection couplings

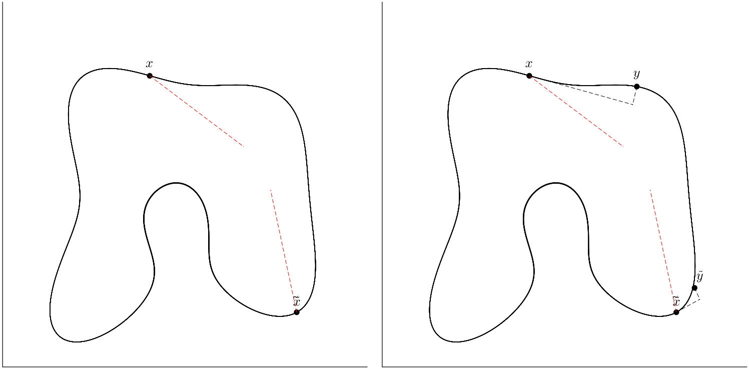
On unconstrained space: reflections between the chains induce contractions

From x and $\tilde{x}$ on $\mathbb{R}^D$, y and $\tilde{y}$ can be drawn from Normal(x, Σ) and Normal$(\tilde{x}, \Sigma)$

- Draw a perturbation u, $u \sim$ Normal$(0, I_D)$.
- Define $\tilde{u}$ in the opposite direction with respect to u.
- Rescale u, $\tilde{u}$ with original covariance matrix Σ and add them to x, $\tilde{x}$.

On manifolds: projections on $\mathcal{T}_x$ after reflections

...an intuition



We expect that reflection strategy helps in obtaining meeting times faster

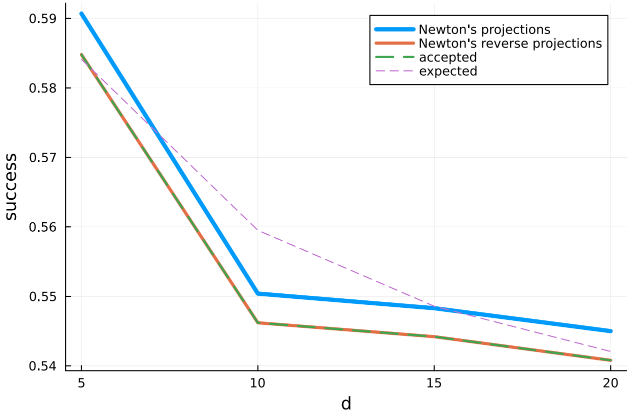
Scaling properties: choice of proposal

Consider the Uniform distribution on a sequence of Hyperspheres,

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

- Choose a proposal standard deviation to ensure comparable acceptance probability across all dimensions:
- if $\|\nu\|_2^2 = \sum_{i=1}^d \nu_i^2 \leq 1$, there are solutions for orthogonal projections.
- if $\nu \sim \text{Normal}(0, I_d/d)$ then $\|\nu\|_2^2 \sim \chi_d^2/d$
- $\mathbb{E}\|\nu\|_2^2 = 1$ and $\mathbb{P}(\|\nu\|_2^2 > 1) \leq 0.5$ (equal as $d \to \infty$).

Proportion of successful proposals



Proportion of successful proposals in different dimensions, computed on chains of length 10^4.

Scaling properties: maximal coupling vs reflections

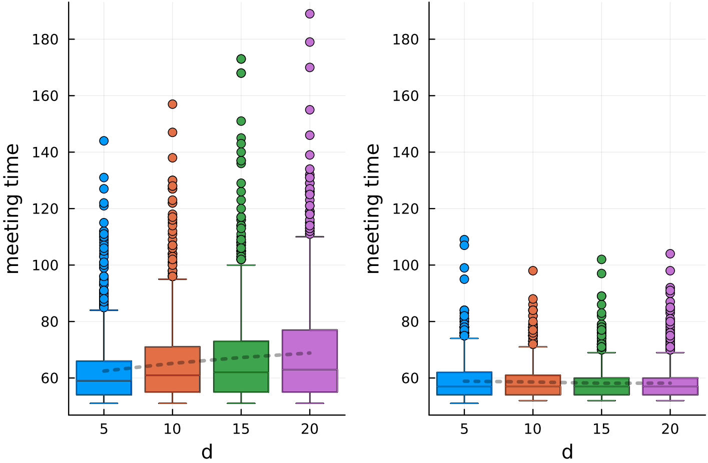
- 1000 parallel chains for each d
- initialization from opposed points $[1, 0, \ldots, 0]$ and $[-1, 0, \ldots, 0]$
- lag $L = 50$

Two strategies:

M Maximal coupling only

M+R Maximal coupling + reflections if $\|x - \tilde{x}\|_2^2 > 1/\sqrt{d} = \sigma$

- **M** Average meeting times increase linearly with the dimension of the space *(left)*
- **M+R** Average meeting times are constant *(right)*



Benefits of couplings of MCMC algorithms on submanifolds

- Diagnosing convergence, other measures of performance require asymptotic reasoning.

- Parallelizing computation, algorithms are computationally involving and long runs are hard.

Thanks for your attention!

Soon on arXiv: *Couplings of MCMC algorithms on submanifolds*, B. E., Jacob, P.E., Ryder, R.J.

Some references I

H. C. Andersen. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *Journal of computational Physics*, 52 (1):24–34, 1983.

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.

P. Diaconis, S. Holmes, M. Shahshahani, et al. Sampling from a manifold. *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, 10:102–125, 2013.

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

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.

Some references II

C. Hartmann. An ergodic sampling scheme for constrained hamiltonian systems with applications to molecular dynamics. *Journal of Statistical Physics*, 130:687–711, 2008.

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.

T. Lelievre, M. Rousset, and G. Stoltz. Langevin dynamics with constraints and computation of free energy differences. *Mathematics of computation*, 81(280):2071–2125, 2012.

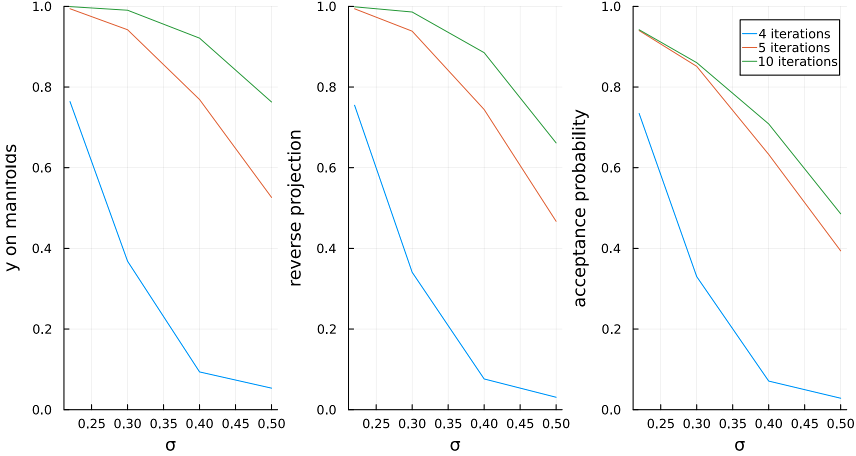
G. Wang, J. O'Leary, and P. Jacob. Maximal couplings of the Metropolis–Hastings algorithm. *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233, 2021.

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.

Comparing algorithms: Goodness of Fit Example

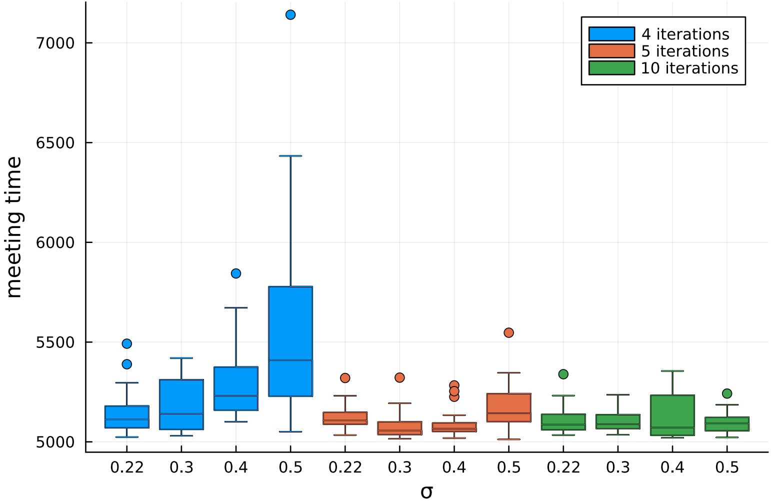
- In the field of testing, sampling over constrained spaces helps in improving the power of tests.
- Constraining here means conditioning on sufficient statistics for the model under the null hypothesis.
- We focus on a goodness-of-fit test to the Gamma distribution, conditioning on the sum and product $(S(x), P(x))$.
- Goal 1: study the impact of number of iterations in Newton's method on convergence properties.
- Goal 2: compare the Random walk of Zappa et al. [2018] *ZHG* to a MCMC algorithm proposed in Diaconis et al. [2013] *DHS*

Studying the impact of Newton's iterations, fixed $D = 20$ $m = 2$

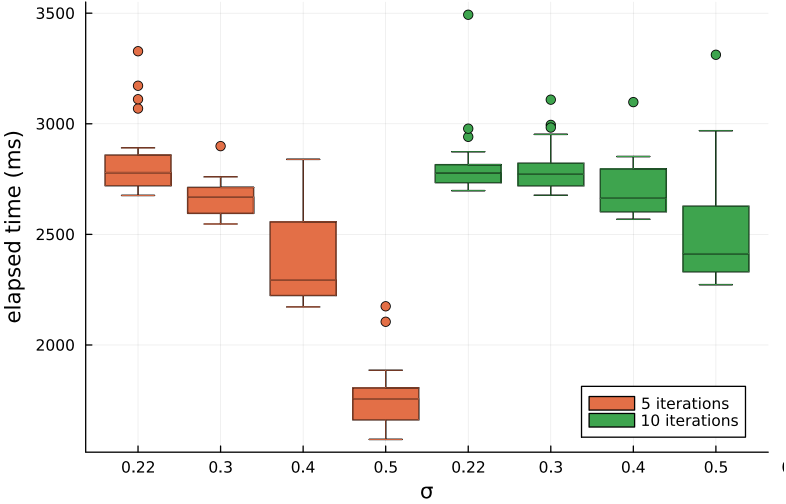


Fraction of successful proposals, reverse projections, and acceptance rate

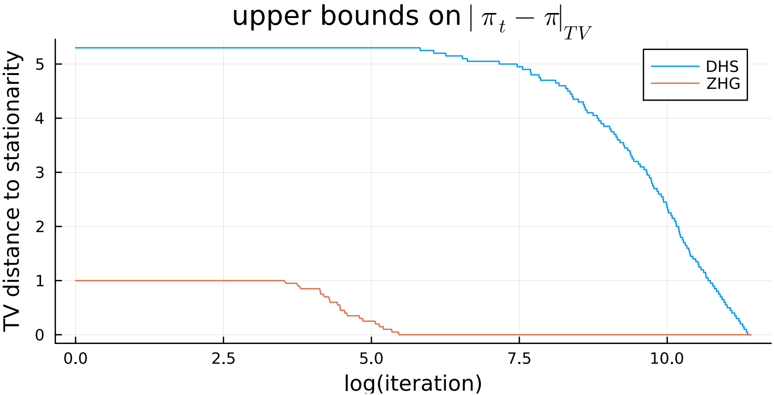
Zappa's RW with Coupled Chains



Zappa's RW with Coupled Chains (fair comparison)



Comparison between algorithms



Comparison of upper bounds on the distance from stationarity of tuned algorithms.