HMC on Manifolds

August 29, 2018

Outline

Outline

High Curvature

- Simple HMC performs badly when the target distribution contains certain regions with high curvature
- Consider the following example to be the distribution of our interest

$$\pi(\boldsymbol{q}) = \prod_{i=1} N(q_i|0, e^{q_0}) N(q_0|0, 9) \tag{1}$$

- We assume that we only know the unnormalized distribution of the target

Funnel Distribution

- Therefore, our unnormalized distribution is in form of

$$exp(-\frac{\sum_i q_i^2}{2exp(q_0)})exp(-\frac{q_0^2}{18}) \tag{2}$$

- The potential function in HMC is defined as negative log of the unnormalized distribution. It takes the form

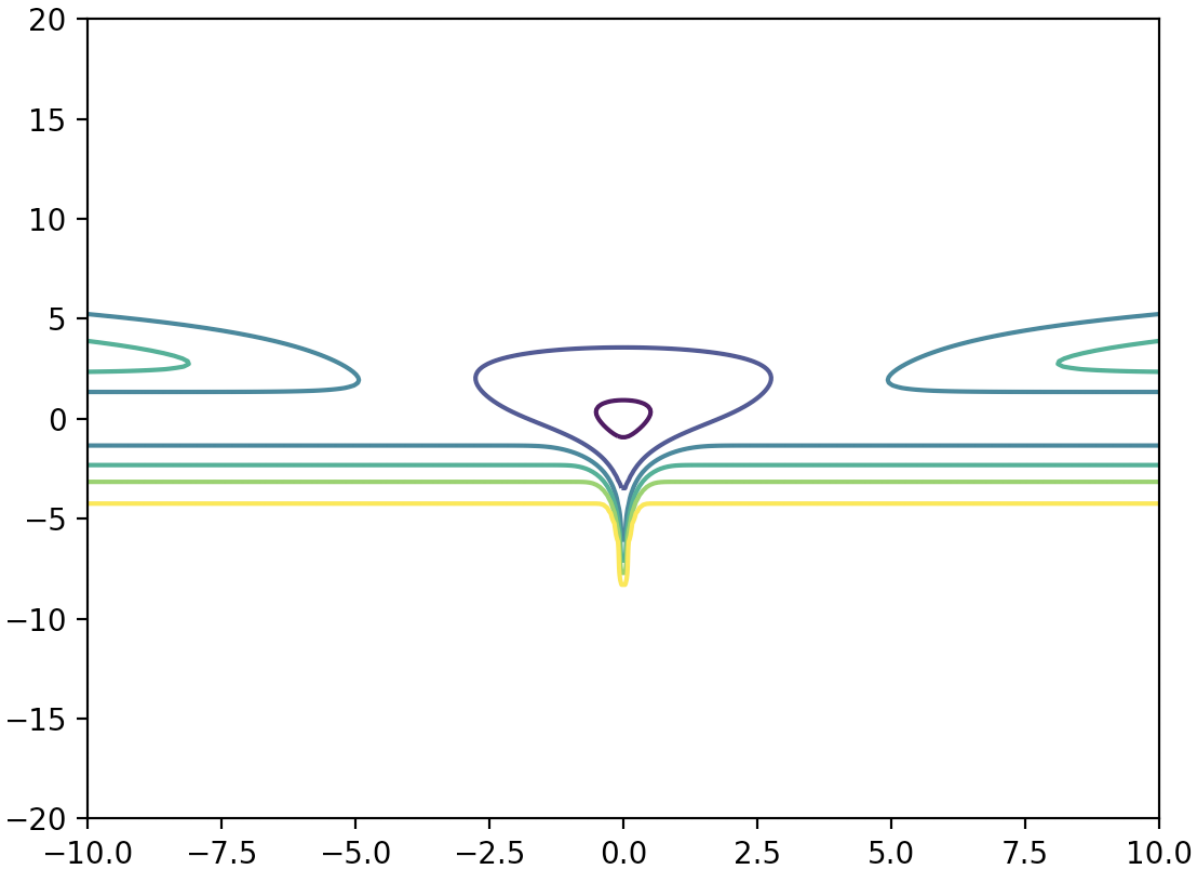
$$\frac{\sum_i q_i^2}{2exp(q_0)} + \frac{D * q_0^2}{18} \tag{3}$$

Funnel Distribution

- The Funnel Distribution is firstly introduced by Radford Neal couple years ago to illustrate the challenges encountered in MCMC
- MCMC, Gibbs Sampler, and Slicing Sampling all having difficult time of approximate such distribution.
- Note that, Funnel Distribution is actually very simple but still can cause lots of trouble to many sophisticated distribution.

Funnel Distribution

- The following is the level set of our target distribution. The Y-axis is q_0 and X-axis is q_i for anyone of the i



Funnel Distribution

The contour illustrates two noticeable features

- The graph has extrememly high curvature around near the neighborhood of (0,-5)
- The graph is very smooth with low curvature everywhere else.
- Unfortunately, Hamiltonian MC, jsut as all other MCMC algorithm, performs badly for this example.

Funnel Distribution

Note, we already know that the mean for q_0 and q_i $(i = 1, .., D)$ to be zero.Let our dimension $D = 10$.

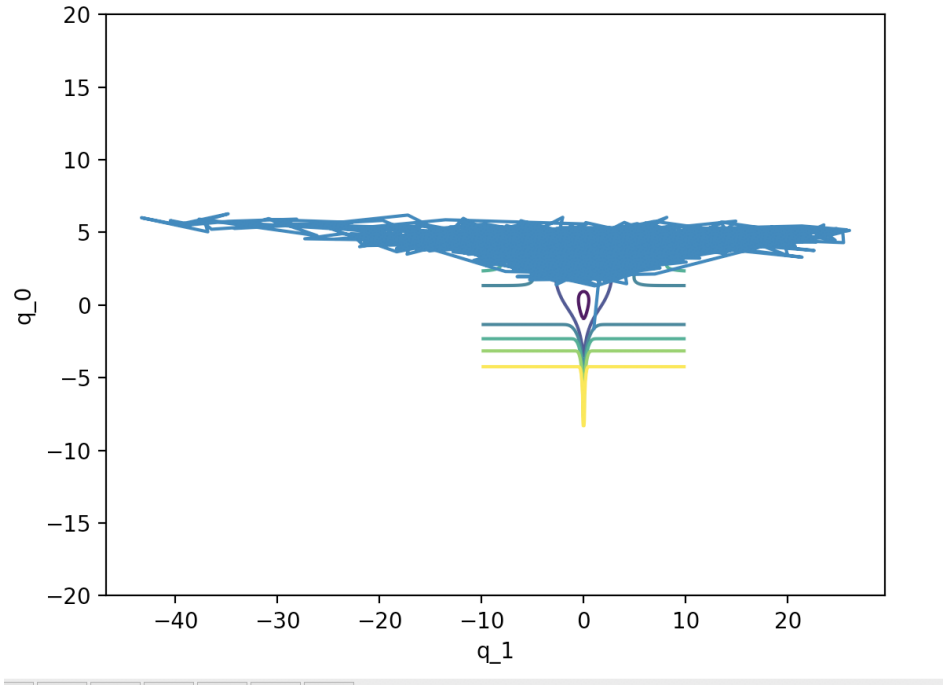


Figure: $\epsilon = 0.3$

Note that due to the large stepsize ϵ, we unable to explore the high curvature area and leads the q_0 has biased value.

Funnel Distribution

We see that the estimated value of q_0 is inaccurate

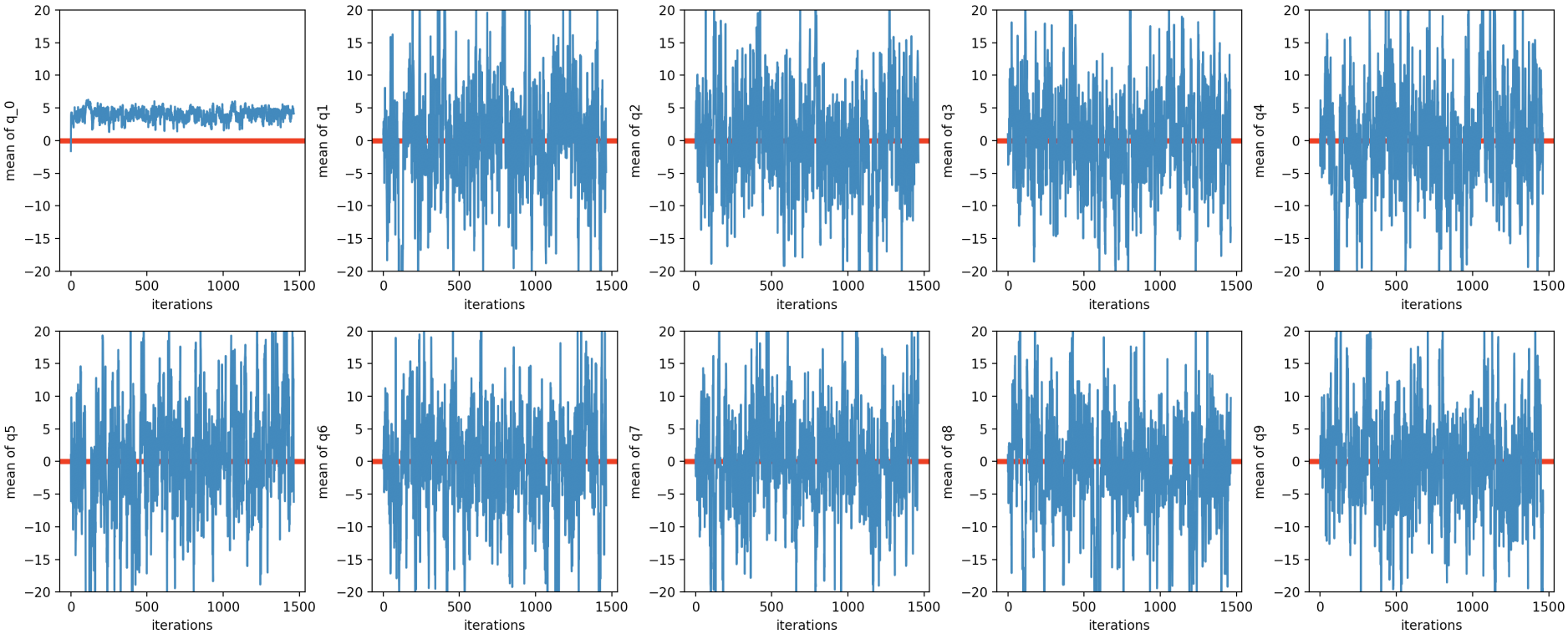


Figure: $\epsilon = 0.3$

Funnel Distribution

However, Due to Monte Carlo Convergence THM, we are guaranteed that our estimator will converge to true value. What will happen is that, when we run the algorithm for so long, once a while, it will step into the high curvature region and stay there very long time to collect samples to correct the estimated mean.

Funnel Distribution

Here is what happen when we stuck in high curvature. We having negative estimated mean for q_0

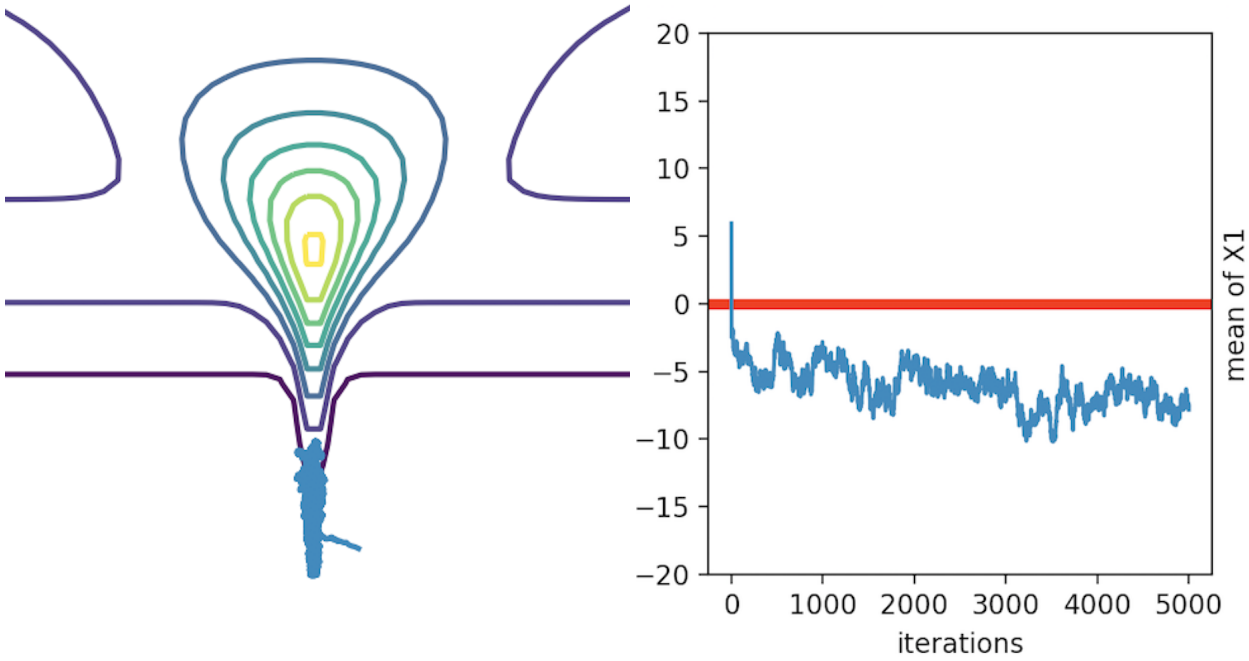


Figure:

Outline

Adaptive HMC

- The Major failure of HMC here is the badly tuned hyperparameter ϵ.
- One globally fixed ϵ will not work in practice.
- ϵ should be able to adjust itself to maneuver through high curvature region and capable of escaping from it as well.
  (See Aofei's work on Adaptive HMC tuning ϵ)

Outline

Manifolds

- Another approach is to adaptively adjust our Mass Matrix.
- Simple HMC use a global Mass Matrix through all iterations.
- This global matrix is used as co-variance matrix to sample momentum variable p. This doesn't capture the complicated characteristics of local regions in many distributions of interests.

Manifolds

- In the simple case, we usually treat our momentum variable and state variable lies in Cartesian coordinate system (orthonormal coordinate system).
- However, in the complicated situation, our dual space might be in curvilinear coordinated system.
- We treat the dual space as a manifolds where each points has its own local metric.

Manifolds

- Let x be a point on the manifolds. A tangent vector on the point x is the velocity vector for some curve lies on the manifolds and pass through point x
- Define the tangent space at point x on manifolds M, $T_{x_0}M$ to be the collection of all tangent vector. For distance between two points on the manifolds now can be defined on the tangent plane.

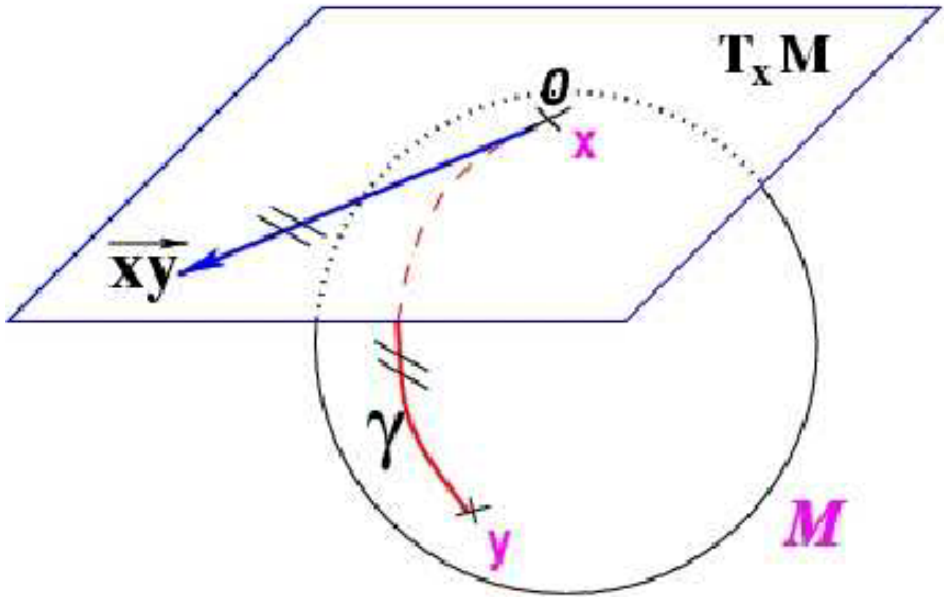


Figure: Manifolds

Manifolds

- We require the the tangent space endowed with an inner product via metric tensor. $G : T_{x_0}M \times T_{x_0}M \to R$

Since G here needed to represent the distance of two points. It requires additional properties.

- Symmetric (from x to y should be equal distance as from y to x)
- Bilinear ($G(x + y, z) = G(x, z) + G(y, z)$)
- Positive Definite (We require distance to be positive for two different points.)

G is called Riemannian Metric

HMC on Manifolds

- Instead of having momentum variable p to be independent of q. we sample p through a conditonal Gaussian with covariance matrix that is a Riemannian Metric

$$\pi(p|q) \sim N(0, G(q))$$

The kinetic function becomes

$$K(q,p) = \frac{1}{2} p^T G(q)^{-1} p + \frac{1}{2} log|G(q)|$$

HMC on Manifolds

- The joint distribution of momemtum and state variable becomes

$$H(q, p) = U(q) + K(q, p)$$

- Note that Simple HMC leapfrog no longer will be working here because now, kinetic function involves variable q, implies that $\partial_q H$ requires differentiate through $K(q, p)$. And this needed to be taken into consideration when doing leapfrog

Generalized Leapfrog

$$p^{n+1/2} \leftarrow p^n - \frac{\epsilon}{2}\partial_q H(q^n, p^{n+1/2}) \tag{4}$$

$$q^{n+1} \leftarrow q^n + \frac{\epsilon}{2}[\partial_p H(q^n, p^{n+1/2}) + \partial_p H(q^{n+1}, p^{n+1/2})] \tag{5}$$

$$p^{n+1} \leftarrow p^{n+1/2} - \frac{\epsilon}{2}\partial_q H(q^{n+1}, p^{n+1/2}) \tag{6}$$

Fixed Point Iteration Method

- Note, equation (4) and equation (5) are defined implicitly.
- Use fixed point iteration. In more general case, we are solving situation like $g(x) = x$
- Randomly choose a point x_0, consider a recursive process

$$x_{n+1} = g(x_n)$$

- Theoretically, if $g(x) - x$ is a continuous function and $\{x_n\}$ converges, then it converges to the solution of $g(x) = x$
- Hence, equation (4) and (5) will be solved through such recursive method. (the number of iteration is a hyperparameters we choose beforehand)

Experiments

It capable of moving big step and also capable of moving into the small high curvature region.

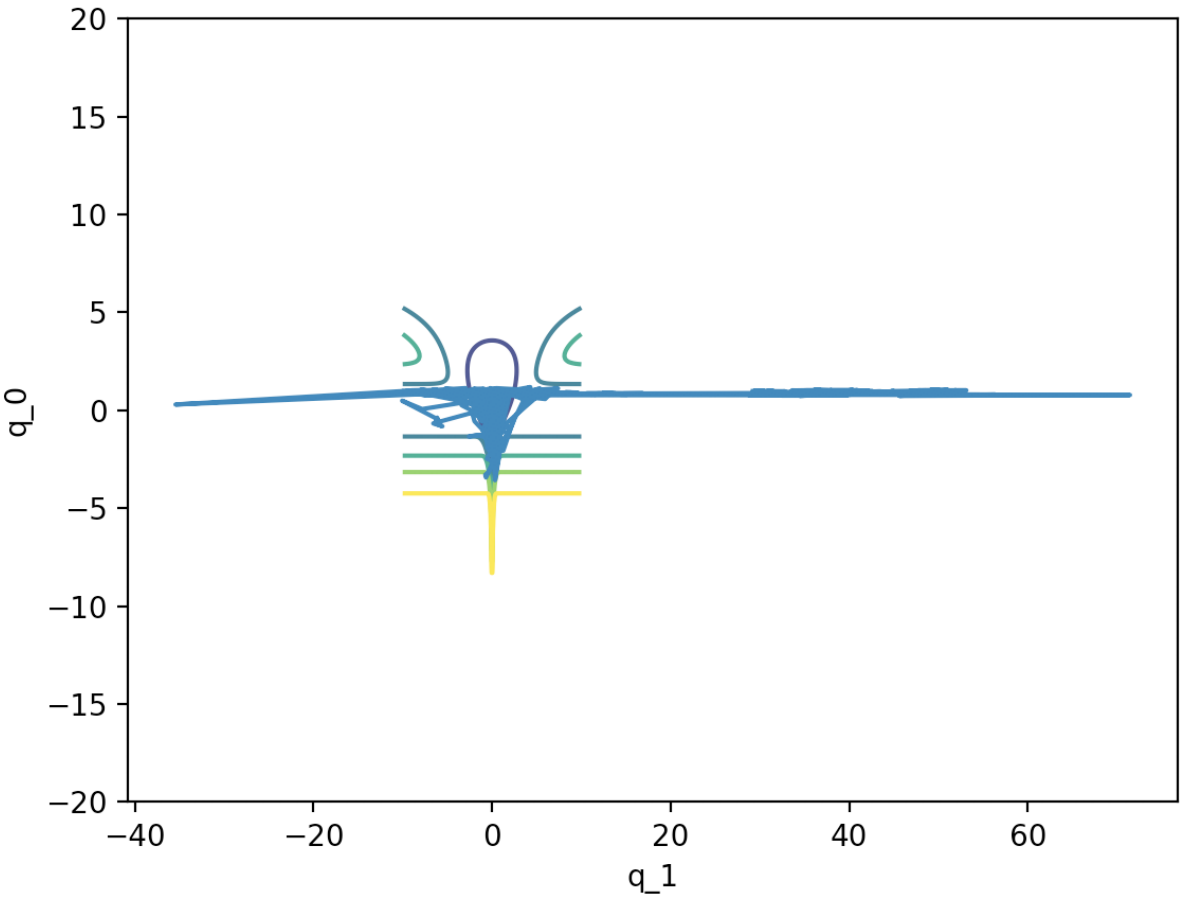
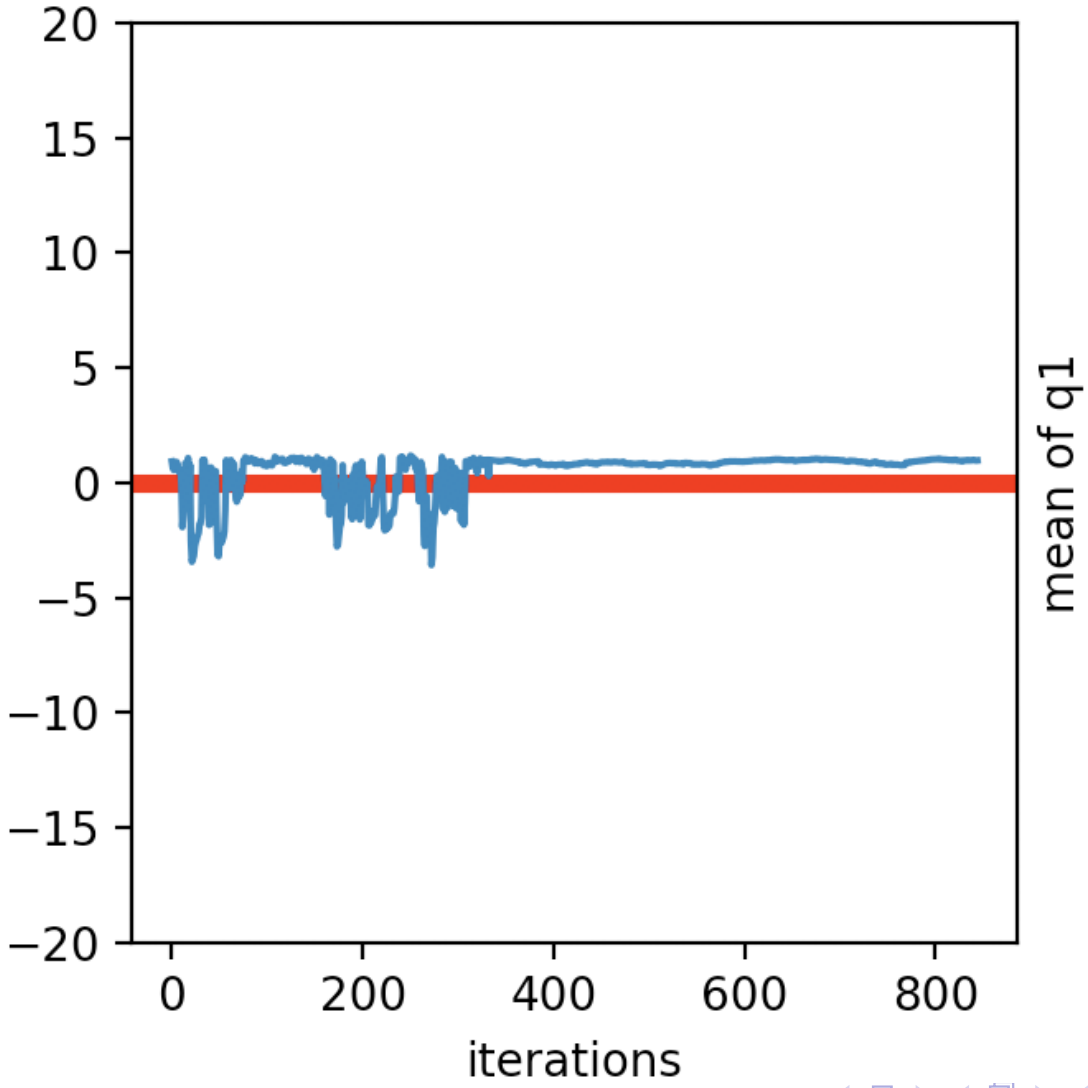


Figure: Contour, $\epsilon = 0.3$

Experiments

It outperforms the simple HMC



Future Work

- In the paper "Riemann Manifold Langevin and Hamiltonian Monte Carlo"; author doesn't dicuss the choice of Riennmanian Metric.
- If the Riennmanian Metric G chosen to be identity matrix, it reduces the algorithm back to Simple HMC.
- The performance highly depends on how to choose a good Metric.
- In the original paper, author carefully design metric depends on what kind of problem they trying to solve.

Future Work

- In my experiment, I notice that we need to take the local curvature into the consideration. I choose Hessian matrix of the joint distribution. As it's also symmetric.
- However, unless our potential distribution happens to be strictly convex, we can't guarantee that Hessian matrix is positive definite.
- In my experiment, we let $G = H^T H + \sigma I$ where H is Hessian matrix. σI is added ensure that our G is also invertible.
- Although, the result is outperforms the Simple HMC, such approach shouldn't be capable of generalized.

Future Work

- Design a good generalized metric that is independent of specific problem environment.
- It should also be computational efficient as we need to compute it's inverse and determinants.

Future Work

- I notice that some Bayesian Model use Fisher Information matrix. In our situation, it will be

$$G(q) = -\mathbb{E}_\theta(\partial_{ij}^2 logp(q|\theta))$$

- Use Fisher Information Matrix might be naive in some situation.
- The reason is that in Bayesian model, we usually define a prior distribution. The equation above can be approximated by sampling θ from our prior.
- But not all target distribution have such nice condition (Hence, unless the target is posterior distribution, we usually don't have ability to sample θ)
- Fisher Information can still be zero matrix. (it only guarantees positive semidefinite)

For Further Reading I

Mark Girolami; Ben Calderhead; Siu A. Chin

Riemann Manifold Langevin and Hamiltonian Monte Carlo

Semantic Scholar

Radford, Neal

The Short-Cut Metropolis Method