Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo

Michael Betancourt

Abstract. When properly tuned, Hamiltonian Monte Carlo scales to some of the most challenging high-dimensional problems at the frontiers of applied statistics, but when that tuning is suboptimal the performance leaves much to be desired. In this paper I show how suboptimal choices of one critical degree of freedom, the cotangent disintegration, manifest in readily observed diagnostics that facilitate the robust application of the algorithm.

Key words and phrases: Markov Chain Monte Carlo, Hamiltonian Monte Carlo, Microcanonical Disintegration, Diagnostics.

Once a statistical model has been specified as a probability distribution, applied statistics reduces to the evaluation of expectations with respect to the that target distribution. Consequently, the fundamental computational challenge in these statistics is the accurate and efficient estimation of these expectations.

Given its general applicability, Markov chain Monte Carlo (Robert and Casella, 1999; Brooks et al., 2011) has become one of the most of the most popular frameworks for developing practical estimation algorithms, as evident from decades of theoretical analysis and empirical success. In particular, Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 2011; Betancourt et al., 2014) pushes Markov chain Monte Carlo deep into the frontiers of applied statistics by exploiting the geometry inherent to many probability distributions. Implementing Hamiltonian Monte Carlo in practice, however, is frustrated by algorithmic degrees of freedom that present a delicate tuning problem which can not only impede the scalable performance of the algorithm but also introduce biases in the estimation.

In this paper I consider the choice of a *cotangent disintegration* that arises in any Hamiltonian Monte Carlo algorithm. Because the performance of the resulting implementation is highly sensitive to the interaction of the cotangent disintegration with the given target distribution, a careful choice is critical for robust performance.

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK (e-mail: betanalpha@gmail.com).

After first reviewing the general construction of Hamiltonian Monte Carlo, I show how the consequences of a given cotangent disintegration manifest in the performance of a single stage of the algorithm. I then analyze this stage to define not only a implicit criteria for the optimal disintegration relative to a given target distribution, but also an explicit diagnostics to identify a suboptimal cotangent disintegration in practice. Finally I demonstrate the utility of these diagnostics in various examples.

1. CONSTRUCTING HAMILTONIAN MONTE CARLO

In this paper I will let π be the target probability distribution over the D-dimensional sample space Q and some appropriate σ-algebra. To simplify the notation I will assume that π admits a density $\pi(q)$ with respect to some reference measure, $\mathrm{d}q$, although Hamiltonian Monte Carlo does not require this. For the more general construction of Hamiltonian Monte Carlo see Betancourt et al. (2014).

Here I will very briefly review Markov chain Monte Carlo and then Hamiltonian Monte Carlo, both in general and in its most common implementation.

1.1 Markov Chain Monte Carlo

Markov chain Monte Carlo builds expectation estimates by finding and exploring the neighborhoods on which the target probability distribution concentrates. The exploration itself is generated by repeatedly sampling from a Markov transition, given by the density $\mathcal{T}(q \mid q')$, to give a sequence of points, $\{q_0, \ldots, q_N\}$, known as a Markov chain. If the transition preserves the target distribution,

$$\pi(q) = \int_Q \mathcal{T}(q \mid q') \, \pi(q') \, \mathrm{d}q',$$

then the resulting Markov chain will eventually explore the entire target distribution and we can use the history of the Markov chain to construct consistent Markov chain Monte Carlo estimators of the desired expectations,

$$\lim_{N \to \infty} \hat{f}_N \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N} f(q_n) = \mathbb{E}_{\pi}[f] \, .$$

The performance of these Markov chain Monte Carlo estimators depends on how effectively the Markov transition guides the Markov chain along the neighborhoods of high probability. If the exploration is slow then the estimators will become computationally inefficient, and if the exploration is incomplete then the estimators will become biased. In order to scale Markov chain Monte Carlo to the high-dimensional and complex distributions of practical interest, we need a Markov transition that exploits the properties of the target distribution to make informed jumps through neighborhoods of high probability while avoiding neighborhoods of low probability entirely.

1.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo achieves such informed transitions by harnessing the differential geometry of the target distribution with auxiliary *momenta* parameters. The algorithm begins by first attaching to each point, in the sample space, q, a copy of $\mathbb{R}^D$ called a *momenta fiber*. Collecting these fibers together yields the $2D$-dimensional *cotangent bundle*, T^*Q, with a natural projection that collapses each fiber to the base point at which it was attached,

$$\begin{aligned} \varpi : T^*Q &\to Q \\ (q, p) &\mapsto q. \end{aligned}$$

We next lift our target probability distribution into a joint distribution on the cotangent bundle with the choice of a conditional probability distribution over the fibers known as a *cotangent disintegration*. Denoting the target distribution

$$\pi \propto \exp(-V(q)) \, \mathrm{d}q,$$

with $V(q)$ denoted the *potential energy*, and the cotangent disintegration as

$$\xi_q \propto \exp[-K(q,p)] \, \mathrm{d}p,$$

with $K(q,p)$ denoted the *kinetic energy*, then the joint distribution is defined as

$$\begin{aligned} \pi_H &= \xi_q \cdot \pi \\ &\propto \exp(-\left(K(q,p) + V(q)\right)) \, \mathrm{d}q \, \mathrm{d}p \\ &\propto \exp(-H(q,p)) \, \mathrm{d}q \, \mathrm{d}p, \end{aligned}$$

with $H(q,p)$ denoted the *Hamiltonian*.

When combined with the natural fiber structure of the cotangent bundle, this Hamiltonian immediately defines an infinite family of deterministic maps,

$$\begin{aligned} \phi^H_t : (q,p) &\to (q,p), \forall t \in \mathbb{R} \\ \phi^H_t \circ \phi^H_s &= \phi^H_{s+t}, \end{aligned}$$

called a *Hamiltonian flow*. By construction, the Hamiltonian flow traces through the neighborhoods where the joint distribution concentrations, while its projection, $\varpi \circ \phi^H_t$, traces through the neighborhoods where the target distribution concentrates, exactly as desired.

Hence we can build a powerful Markov transition in three stages. From the initial point, q, we first lift from the sample space onto the cotangent bundle by sampling a random momenta from the cotangent disintegration, $p \sim \xi_q$, apply the Hamiltonian flow for some time to generate exploration, $(q,p) \mapsto \phi^H_t(q,p)$, and then project back down to the target sample space, $\varpi : (q,p) \mapsto q$. In practice there are various strategies for choosing the integration time, as well as numercially approximating the Hamiltonian flow and correcting for the resulting error (Betancourt, 2016), but in general any Hamiltonian Markov transition will proceed with a lift, a flow, and a projection (Figure 1).

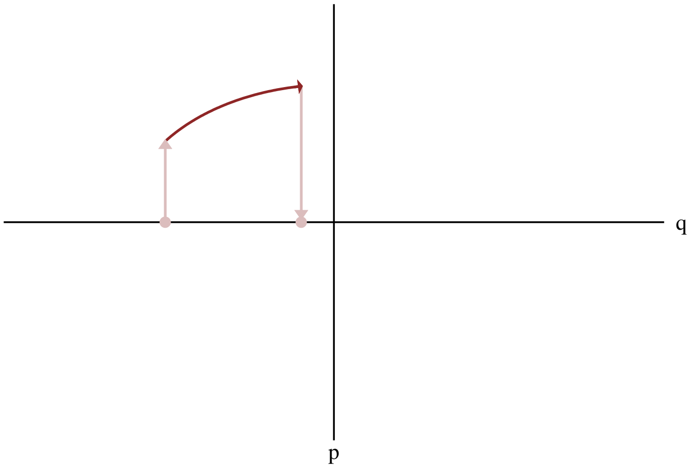


Fig 1. *Every Hamiltonian Markov transition is comprised of a random lift from the target sample space onto the cotangent bundle (light red), a deterministic Hamiltonian flow on the cotangent bundle (dark red), and a projection back down to the target space (light red).*

1.3 Gaussian-Euclidean Cotangent Disintegrations

An explicit choice of cotangent disintegration is facilitated when the sample space has a metric structure. In particular, if the sample space is equipped with a Riemannian metric, g, then we can define an entire family of *Riemannian cotangent disintegrations* with the kinetic energies

$$K(q,p) = A \cdot f\left(g_q^{-1}(p,p)\right) + \frac{1}{2} \log |g_q| + \text{const},$$

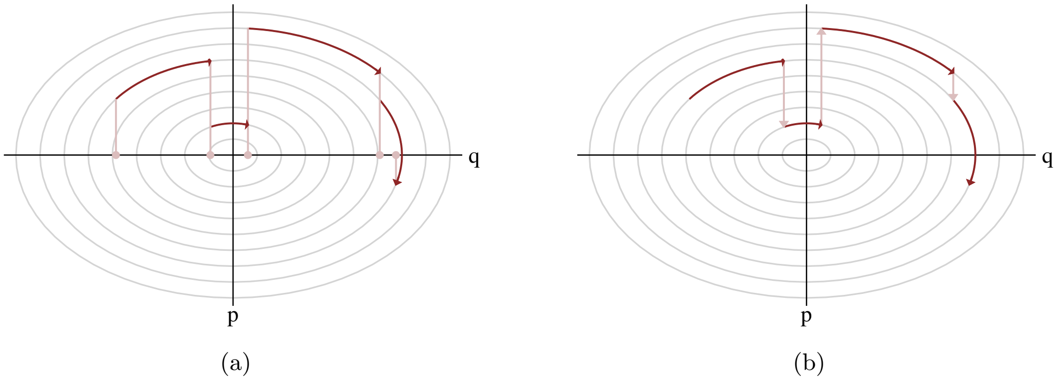
for some constant A and function $f : \mathbb{R} \to \mathbb{R}$. Riemannian disintegrations also define two helpful scalar functions: the *effective potential energy*,

$$\check{V}(q) = V(q) + \frac{1}{2} \log |g_q| + \text{const}.$$

and the *effective kinetic energy*,

$$\check{K}(q,p) = A \cdot f\left(g_q^{-1}(p,p)\right).$$

In practice most implementations of Hamiltonian Monte Carlo assume that any metric structure is *Euclidean*, where the metric g is constant across the sample space and then sometimes denoted as a *mass matrix*. Additionally, these implementations usually take $A = \frac{1}{2}$ and $f = \mathbb{I}$, in which case the cotangent disintegration defines a Gaussian distribution over each momenta fiber. Hence in common practice we typically consider only *Gaussian-Euclidean cotangent disintegrations*.



(a)

(b)

Fig 2. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with the Hamiltonian flow before projecting back down to the target sample space. (b) If we consider the projection and random lift steps as a single momentum resampling operation, then the Hamiltonian Markov chain alternates between deterministic flows along these level sets (dark red) and a random walk across the level sets (light red).*

2. THE MICROCANONICAL DISINTEGRATION

Although any choice of cotangent disintegration will yield a Hamiltonian flow that coherently explores the neighborhoods where the target distribution concentrates, not every choice will yield a flow that is as computationally efficient as others. How the interaction between a particular disintegration and the target distribution manifests in performance may at first seem abstruse, but it becomes straightforward to characterize if we examining these Hamiltonian systems from a more natural perspective.

One of the distinctive properties of Hamiltonian flow is that it preserves the Hamiltonian itself, which implies that each Hamiltonian trajectory is confined to a *level set* of the Hamiltonian,

$$H^{-1}(E) = \{(q,p) \in T^{*}Q \mid H(q,p) = E\} .$$

A Markov transition, then, first jumps to a random level set and then explores that level set with the Hamiltonian flow before projecting back to the sample space (Figure 2a). If we compose the projection and random lift stages together into a single *momentum resampling* operation, then the entire Hamiltonian Markov chain naturally decouples into exploration along each level set driven by the Hamiltonian flow, and exploration across level sets driven by the momentum resampling (Figure 2b).

Consequently a much more natural way to analyze Hamiltonian Monte Carlo is not through positions and momenta but rather level sets and the *energies* labeling each level set. The *microcanonical disintegration* formalizes this intuition by decomposing the joint distribution into a conditional *microcanonical distribution* over level sets, $\pi_{H^{-1}(E)}$, and a *marginal energy distribution*, π_E,

$$\pi_H = \pi_{H^{-1}(E)} \cdot \pi_E.$$

A Hamiltonian system always admits a microcanonical disintegration, although there are some technical subtleties (Betancourt et al., 2014).

From this perspective, the Hamiltonian flow generates exploration of the microcanonical distributions while the exploration of the marginal energy distribution is determined solely by the momentum resampling. Because the cotangent disintegration affects the geometry of the level sets, it also effects the efficacy of the Hamiltonian flow, but this can largely be overcome with an appropriate choice of integration times (Betancourt, 2016). The exploration of the marginal energy distribution, however, is determined solely by the momentum resampling which itself depends on only the interaction between the cotangent disintegration and the target distribution.

3. DIAGNOSING SUBOPTIMAL COTANGENT DISINTEGRATIONS

To quantify the efficacy of the momentum resampling consider $\pi_{E|q}$, the distribution of energies, E, induced by a momentum resampling at position q. The closer this distribution is to the marginal energy distribution for any q, the faster the random walk will explore energies and the smaller the autocorrelations we be in the overall Hamiltonian Markov chain (Figure 3a). Conversely, the more this distribution deviates from the marginal energy distribution the less effectively the random walk will explore and the larger the autocorrelations will be in the overall chain (Figure 3b).

Consequently, the compatibility of the momentum mresampling-induced distributions and the marginal energy distribution defines an implicit criterion for selecting an optimal cotangent disintegration for a given target distribution. There are many ways, however, of quantifying this compatibility in theory and in practice, and hence many ways of defining optimality criteria and resulting diagnostics.

3.1 General Criteria

Optimal performance is achieved only when the momentum resampling-induced energy distributions are uniformly equal to the marginal energy distribution,

$$\log \frac{\mathrm{d}\pi_{E|q}}{\mathrm{d}\pi_E} = 0, \, \forall q \in Q.$$

Consequently we could quantify the compatibility of the two distributions with the expectation

$$\mathbb{E}_\pi \left[\log \frac{\mathrm{d}\pi_{E|q}}{\mathrm{d}\pi_E} \right],$$

which would vanish only when the cotangent disintegration was optimal. Because we don't have closed-forms for the densities, however, this would be infeasible to even estimate in any nontrivial problem.

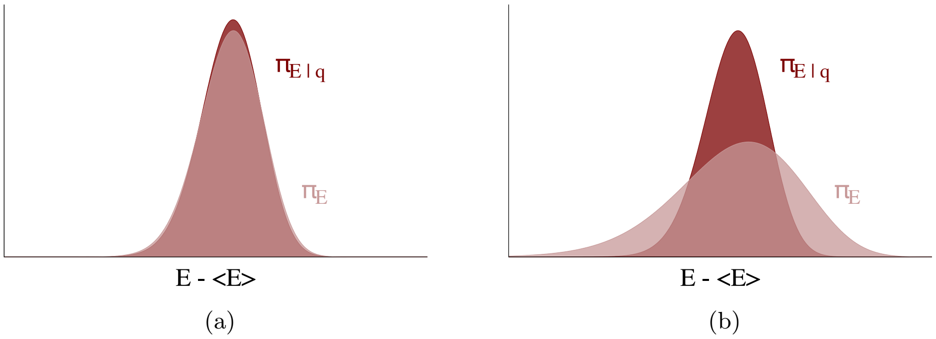


FIG 3. *The momentum resampling in a Hamiltonian Markov transition induces a change of energies and allows a Hamiltonian Markov chain to randomly walk across level sets. (a) When the energy distribution induced by momentum resampling at any point q, $\pi_{E|q}$ is similar to the marginal energy distribution, π_E, this random walk will rapidly explore all relevant energies and the resulting Hamiltonian Markov chain will enjoy small autocorrelations. (b) On the other hand, when the distributions deviate for any q, for example with the marginal energy distribution has heavier tails, then the exploration will be slow and the autocorrelations of the chain large.*

In practice we want a criterion that is readily estimating using the Hamiltonian Markov chain itself. One theoretically appealing choice is the Bayesian fraction of missing information (Rubin, 2004),

$$\mathrm{BFMI} = \frac{\mathbb{E}_{\pi}\left[\mathrm{Var}_{\pi_{E|q}}[E \mid q] \right]}{\mathrm{Var}_{\pi_E}[E]},$$

which quantifies how insufficient the energy variation induced by the momentum resampling is: in the worst case BFMI → 0 and the momentum resampling induces very slow exploration across the level sets, while in the best case BFMI → 1 and the momentum resampling effectively generates exact draws from the marginal energy distribution.

By construction,

$$\mathrm{Var}_{\pi_{E|q}}[E \mid q] = \mathrm{Var}_{\pi_{E|q}}[\Delta E \mid q]\,,$$

where ΔE is the change in energy induced by the momentum resampling. Because the momentum resampling does not change the position, q, this can also be interpreted as the change in kinetic energy, $\Delta E = \Delta K$, which depends only on the choice of cotangent disintegration, as expected. Using this latter form we can then readily estimate the Bayesian fraction of missing information using the history of energies in the Hamiltonian Markov chain,

$$\mathrm{BFMI} \approx \widehat{\mathrm{BFMI}} \equiv \frac{\sum_{n=1}^{N} \left(E_n - E_{n-1} \right)^2}{\sum_{n=0}^{N} \left(E_n - \bar{E} \right)^2}.$$

In this form the Bayesian fraction of missing information is similar to the lag-1 autocorrelation of the energies, suggesting that the effective sample size per transition of the energies, $\mathrm{ESS}/\mathrm{T}(E)$ might also be a useful quantification, with $\mathrm{ESS}/\mathrm{T}(E) \to 0$ indicating a suboptimal cotangent disintegration and $\mathrm{ESS}/\mathrm{T}(E) \to 1$ indicating an optimal one. This measure also has a strong intuitive appeal – up to the usual regularity conditions the effective sample size quantifies the rate of convergence of the marginal random walk over energies, and hence it directly quantifies the efficacy of the exploration induced by the momentum resampling.

Finally, we can also use the the change of energies induced by a momentum resampling to construct visual criteria. Averaging the momentum resampling-induced energy distribution over positions gives a marginal distribution over energy variations,

$$\pi_{\Delta E}(\Delta E) \equiv \int \pi_{E|q}(\Delta E \mid q)\, \pi(q)\, \mathrm{d}q,$$

whose density is readily estimated by histogramming the $\{\Delta E_n\}$ from the Hamiltonian Markov chain. We can also estimate the marginal energy density by histogramming the $\{E_n\}$, and then compare the variation of the two histograms by eye.

The Bayesian fraction of missing information, effective sample size per transition, and histograms can all be estimated directly from the history of the Hamiltonian Markov chain, but none of them define a criteria that can be explicitly inverted to identify an optimal disintegration. Hence in practice they best serve as diagnostics for distinguishing suboptimal disintegrations.

Additionally we must take care when applying these diagnostics as they make the strong assumption that the Hamiltonian Markov chain sufficiently explores the joint distribution. In order to improve the robustness of these diagnostics, in practice it is best to use them with multiple Markov chains and monitor additional diagnostics such as divergences (Betancourt, Byrne and Girolami, 2014; Betancourt and Girolami, 2015) and the Gelman-Rubin statistic (Gelman and Rubin, 1992).

3.2 Gaussian-Euclidean Criteria

Gaussian-Euclidean cotangent disintegrations are particularly useful as they admit some results that can simplify the general criteria introduced above.

Consider, for example, the random lift onto the cotangent bundle. For any Gaussian-Euclidean cotangent disintegration, in fact for any Gaussian-Riemannian cotangent disintegration, the effective kinetic energy introduced by the randomly sampled momentum is distributed according to a scaled-χ^2 distribution independent of position,

$$\breve{K} \sim \chi^2(D, 1/2)\,,$$

where

$$\chi^2(x \mid k, \sigma) = \frac{(2\sigma)^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2\sigma}}.$$

In general the projection can shed an arbitrarily large amount of effective kinetic energy, but in equilibrium we'd expect to loose as much energy as we gain, hence the change in energies should be distributed as the difference between two $\chi^2\big(D, \frac{1}{2}\big)$ variates describing the initial and final effective kinetic energies. As the number of dimensions, D, increases this distribution rapidly converges to a Gaussian distribution with zero mean and D variance, so to a good approximation we have

$$\Delta E \sim \mathcal{N}(0, D)\,,$$

for all positions, q.

In this case the numerator in the Bayesian fraction of missing information becomes

$$\mathbb{E}_{\pi}\Big[\mathrm{Var}_{\pi_{E|q}}[E \mid q]\Big] = \mathbb{E}_{\pi}[D] = D,$$

and we can quantify the efficacy of the cotangent disintegration simply by comparing the variance of the marginal energy distribution to the dimensionality of the target sample space, D.

4. EXAMPLES

In order to demonstrate the utility of these diagnostics, in this section we'll consider a series of pedagogic examples, starting with identically and independently distributed Gaussian and Cauchy distributions and then a typical hierarchical model.

All Markov chains were generated with CmdStan (Stan Development Team, 2016), using the No-U-Turn sampler (Hoffman and Gelman, 2014) to dynamically adapt the integration time and a Gaussian-Euclidean cotangent disintegration with a diagonal Euclidean metric adapted to the covariance of the target distribution. Unless otherwise specified all other settings were default. The exact version of CmdStan can be found at `https://github.com/stan-dev/cmdstan/commit/ad6177357d4d228e129eefa60c9f399b36e9ac19`, and all Stan programs and configurations can be found in the Appendix.

4.1 Gaussian Target

Let's first consider a 100-dimensional identically and independently distributed Gaussian distribution, $q_n \sim \mathcal{N}(0, 1)$. Given a Gaussian-Euclidean cotangent disintegration the marginal energy distribution reduces to a scaled χ^2 distribution,

$$E \sim \chi^2\left(2D, \frac{1}{2}\right),$$

which converges to a $\mathcal{N}(D, D)$ with increasing dimension. This perfectly matches the expected energy variation, as evident in both numerical diagnostics (Figure 4) and visual diagnostics (Figure 5).

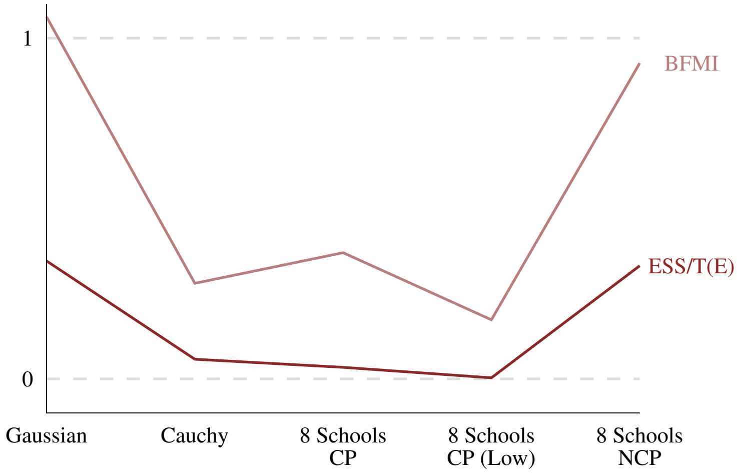


Fig 4. *Both the Bayesian fraction of mission information, BFMI, and effective sample size per transition for the energy ESS/T(E), quantify the compatibility of a cotangent disintegration with a given target distribution. Here a Gaussian-Euclidean cotangent disintegration works well for both a Gaussian and non-centered eight schools target, but is less effective for the heavier-tailed Cauchy and centered eight schools targets.*

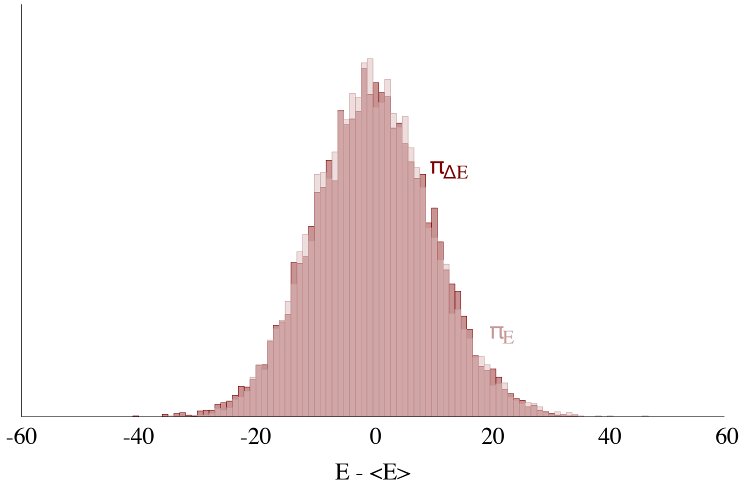


Fig 5. *A Gaussian-Euclidean cotangent disintegration is well-suited to a Gaussian target distribution – at each iteration the momentum resampling is able to jump the Hamiltonian Markov chain to any relevant level set.*

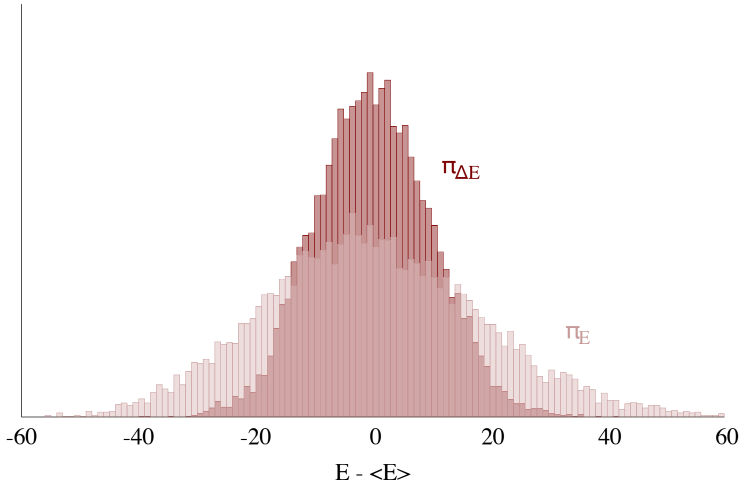


FIG 6. *The heavy tails of the Cauchy distribution induce a heavy-tailed marginal energy distribution which limits the efficacy of a Hamiltonian Markov chain utilizing the more lightly-tailed energy variation induced by a Gaussian-Euclidean cotangent disintegration.*

4.2 Cauchy Target

For a less compatible pairing consider instead a 100-dimensional identically and independently distributed Cauchy distribution, $q_n \sim \mathcal{C}(0, 1)$. The heavy tails of the Cauchy distribution induce a marginal energy distribution with heavier tails than the momentum resampling-induced energy variation. Consequently each transition is limited to only those level sets in close proximity to the initial level set, resulting in slower exploration and decreased performance (Figures 4, 6). Despite the suboptimality of this disintegration, however, the Hamiltonian Markov chain is able to explore all relevant energies within only a few transitions and ends up performing surprisingly well given the reputation of the Cauchy distribution.

4.3 Hierarchical Target

Finally, let's consider the eight schools posterior distribution, a relatively simple Bayesian hierarchical model that demonstrates both the utility of hierarchical modeling as well many of the computational difficulties inherent to these models (Rubin, 1981; Gelman et al., 2014). Here the test taking performance of eight schools is modeled with individual, centered Gaussian distributions,

$$y_n \sim \mathcal{N}\big(\theta_n, \sigma_n^2\big) \,,$$

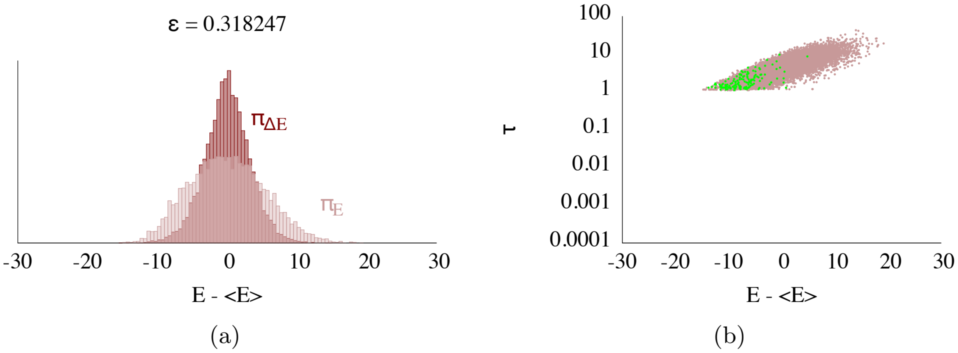


FIG 7. *(a) The empirically-derived marginal energy distribution of the centered eight schools model appears to have mildly-heavy tails, (b) but this is misleading due to the incomplete exploration of the tails as indicated by the divergent transitions in the Hamiltonian Markov chain shown in green. With so many divergences, the numerical and visual diagnostics are suspect.*

where the θ_n are modeled hierarchically,

$$\begin{aligned} \theta_n &\sim \mathcal{N}\big(\mu, \tau^2\big) \\ \mu &\sim \mathcal{N}\big(0, 10^2\big) \\ \tau &\sim \text{Half-}\mathcal{C}(0, 10)\,, \end{aligned}$$

and the $\{y_n, \sigma_n\}$ are given as data.

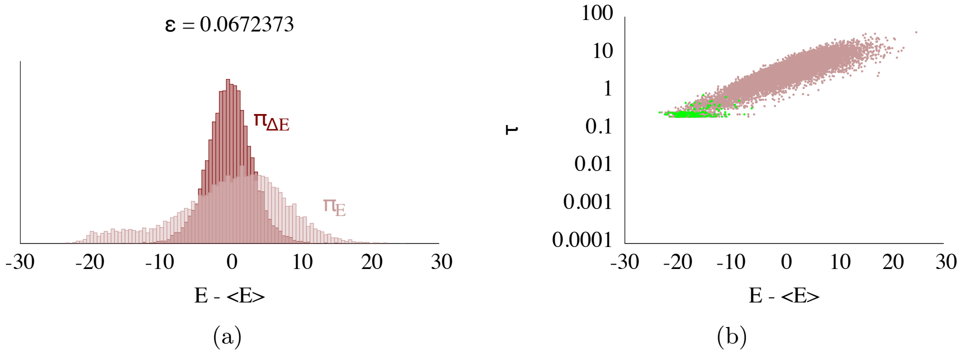
In the typical centered-parameterization (Papaspiliopoulos, Roberts and Sköld, 2007) the marginal energy distribution seems to exhibits only mildly-heavy tails (Figure 7a), but these empirical results are misleading. The problem is that the Hamiltonian Markov chain is not able to fully explore the tails of the target distribution, as exhibited by the large number of divergences at small τ, (Figure 7b).

Forcing the step size of the numerical integrator to a smaller value improves the exploration of the tails (Figure 8b) and better reveals the true heaviness of the marginal energy distribution (Figure 8a), although the exploration is still incomplete.

In order to completely explore the tails we have to utilize a non-centered parameterization which explores the hierarchical effects only indirectly,

$$\begin{aligned} y_n &\sim \mathcal{N}\Big(\mu + \tau \cdot \tilde{\theta}_n, \sigma_n^2\Big)\,, \\ \tilde{\theta}_n &\sim \mathcal{N}(0, 1) \\ \mu &\sim \mathcal{N}\big(0, 10^2\big) \\ \tau &\sim \text{Half-}\mathcal{C}(0, 10)\,. \end{aligned}$$

Not only does this implementation of the model not suffer from the heavy tails and patho-



(a)

(b)

Fig 8. *(a) Forcing a smaller step size in the Hamiltonian Markov transition admits further exploration of the centered eight schools model and better exhibits the true heaviness of the marginal energy distribution. (b) The persistent divergent transitions (green), however, indicate that the exploration is still incomplete. Regardless of the partial exploration, the Gaussian-Euclidean disintegration is suboptimal for this implementation of the model.*

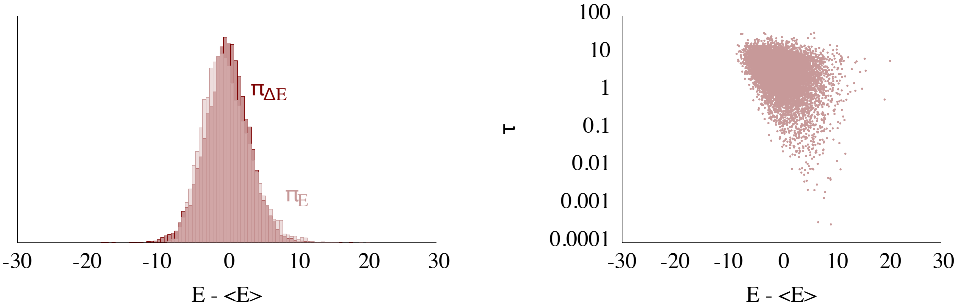
logical curvature of the centered implementation, the Gaussian-Euclidean cotangent disintegration is a nearly optimal pairing (Figures 4, 9)

5. DISCUSSION

As with any Markov chain Monte Carlo algorithm, the performance of Hamiltonian Monte Carlo is limited by its ability to sufficiently explore the target distribution. Livingstone et al. (2016), for example, demonstrates that both neighborhoods of strong curvature and heavy tails limit the exploration of a Hamiltonian Markov chain, ultimately obstructing geometric ergodicity and the central limit theorems needed to ensure robust Markov chain Monte Carlo estimation.

What is unique to Hamiltonian Monte Carlo, however, is its natural ability to diagnose these pathologies. Neighborhoods of strong curvature, for example, can be identified with the divergent transitions they provoke. Moreover, heavy tails manifest both in expansive level sets and heavy-tailed marginal energy distributions. When using dynamic integration time algorithms, the former manifests as long integration times which are readily reported to users, and we have seen in this paper that heavy-tailed marginal energy distributions are straightforward to report both numerically and visually. These intrinsic diagnostics make Hamiltonian Monte Carlo extremely robust, even in the challenging problems at the frontiers of applied statistics.

How to address the pathologies identified by these diagnostics is another question. For example, more heavily-tailed cotangent disintegrations, such as Laplace or even Cauchy disintegrations, may be useful. Generalizing from Euclidean to fully Riemannian disintegrations, for example with the SoftAbs metric (Betancourt, 2013), offers another potential strategy. Within existing tools like Stan, however, perhaps the best way to deal with



(a) (b)

Fig 9. *A non-centered implementation of the eight schools model is not only free of the pathologies that limit the exploration of the centered implementation but also is a nearly optimal pairing with the Gaussian-Euclidean cotangent disintegration.*

any identified pathologies is with alternative implementations, such as the non-centered parameterization utilized in the eight schools example.

6. ACKNOWLEDGEMENTS

I am grateful to Gareth Roberts for enlightening discussions. This work was supported under EPSRC grant EP/J016934/1.

A. STAN PROGRAMS

In this section I collect the Stan programs and configurations used in the examples.

A.1 Gaussian

Configuration:

```
./gauss sample num_samples=10000 random seed=2983157687
```

Stan Program:

```
parameters {
  real x[100];
}
model {
  x ~ normal(0, 1);
}
```

A.2 Cauchy

Configuration:

```
./cauchy sample num_samples=10000 random seed=2983158736
```

Stan Program:

```
parameters {
  real x[100];
}
model {
  x ~ cauchy(0, 1);
}
```

A.3 Centered Eight Schools

Nominal Configuration:

```
./eight_schools_cp sample num_samples=10000
data file=eight_schools.data.R random seed=483892929
```

Small Stepsize Configuration:

```
./eight_schools_cp sample num_samples=10000 adapt delta=0.99
data file=eight_schools.data.R random seed=483892929
```

Stan Program:

```
data {
  int<lower=0> J;
  real y[J];
  real<lower=0> sigma[J];
}
parameters {
  real mu;
  real<lower=0> tau;
  real theta[J];
}
model {
  mu ~ normal(0, 10);
  tau ~ cauchy(0, 10);
  theta ~ normal(mu, tau);
  y ~ normal(theta, sigma);
}
```

A.4 Noncentered Eight Schools

Configuration:

```
./eight_schools_ncp sample num_samples=10000
data file=eight_schools.data.R random seed=483892929
```

Stan Program:

```
data {
  int<lower=0> J;
  real y[J];
  real<lower=0> sigma[J];
}
parameters {
  real mu;
  real<lower=0> tau;
  real theta_tilde[J];
}
transformed parameters {
  real theta[J];
  for (j in 1:J)
    theta[j] = mu + tau * theta_tilde[j];
}
model {
  mu ~ normal(0, 10);
  tau ~ cauchy(0, 10);
  theta_tilde ~ normal(0, 1);
  y ~ normal(theta, sigma);
}
```

REFERENCES

Betancourt, M. (2013). A General Metric for Riemannian Hamiltonian Monte Carlo. In *First International Conference on the Geometric Science of Information* (F. Nielsen and F. Barbaresco, eds.). *Lecture Notes in Computer Science* **8085**. Springer.

Betancourt, M. (2016). Identifying the Optimal Integration Time in Hamiltonian Monte Carlo.

Betancourt, M., Byrne, S. and Girolami, M. (2014). Optimizing The Integrator Step Size for Hamiltonian Monte Carlo. *ArXiv e-prints* **1410.5110**.

Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In *Current Trends in Bayesian Methodology with Applications* (U. S. Dipak K. Dey and A. Loganathan, eds.) Chapman & Hall/CRC Press.

Betancourt, M., Byrne, S., Livingstone, S. and Girolami, M. (2014). The Geometric Foundations of Hamiltonian Monte Carlo. *ArXiv e-prints* **1410.5110**.

Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, New York.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B* **195** 216 - 222.

Gelman, A. and Rubin, D. B. (1992). Inference From Iterative Simulation Using Multiple Sequences. *Statistical science* 457–472.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014). *Bayesian Data Analysis*, third ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.

Livingstone, S., Betancourt, M., Byrne, S. and Girolami, M. (2016). On the Geometric Ergodicity of Hamiltonian Monte Carlo.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.) CRC Press, New York.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science* 59–73.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer New York.

Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. *Journal of Educational and Behavioral Statistics* **6** 377–401.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys. Wiley Classics Library*. Wiley-Interscience [John Wiley &amp; Sons], Hoboken, NJ.

Stan Development Team (2016). CmdStan: The command-line interface to Stan, Version 2.9.0. http://mc-stan.org/cmdstan.html.