Laplacian Hamiltonian Monte Carlo

Yizhe Zhang(✉), Changyou Chen, Ricardo Henao, and Lawrence Carin

Department of Electrical and Computer Engineering,
Duke University, Durham, NC, USA
{yz196,cc448,rhenao,lcarin}@duke.edu

Abstract. We proposed a Hamiltonian Monte Carlo (HMC) method with Laplace kinetic energy, and demonstrate the connection between slice sampling and proposed HMC method in one-dimensional cases. Based on this connection, one can perform slice sampling using a numerical integrator in an HMC fashion. We provide theoretical analysis on the performance of such sampler in several univariate cases. Furthermore, the proposed approach extends the standard HMC by enabling sampling from discrete distributions. We compared our method with standard HMC on both synthetic and real data, and discuss its limitations and potential improvements.

1 Introduction

One pivotal question in modern statistical computation is to efficiently sample from an unnormalized probability density function, where the normalization constant (partition function) is intractable. Towards this end, many Markov Chain Monte Carlo (MCMC) [22] methods have been developed. One of the most influential algorithms is Metropolis-Hastings (MH) [15]. Despite its great success, the *random walk* nature often delivers inefficient mixing of the Markov chain [22]. An inappropriate setting of transition kernel would result in either low acceptance ratio or slow moves. Such situation is exaggerated in high dimensional cases, where the samples from the chain can be highly correlated. As a consequence, the effective sample size is usually relatively small. A number of adaptations have been proposed to mitigate these issues [12,20], however, achievable improvements are limited if attempting maintaining the Markov property and reversibility of the chain [1,10,18].

To mitigate the random walk behavior in MH, several approaches have been proposed, such as Hamiltonian Monte Carlo (HMC) [9,20]. HMC augments a target distribution with auxiliary momentum variables, and uses gradient information to propose distant samples, while maintaining ergodic property and detailed balance. The ability of long-range movement with a high acceptance ratio significantly improves mixing performance. However, HMC is sensitive to parameter settings and can only sample continuous distributions. Towards solving these issues, methods were proposed to use adaptive leap-frog steps [13], or automatic stepsize [16], and to relax the discrete distributions sampling tasks to continuous distributions [21,26]. The improvement can be further boosted by leveraging

P. Frasconi et al. (Eds.): ECML PKDD 2016, Part I, LNAI 9851, pp. 98–114, 2016.
DOI: 10.1007/978-3-319-46128-1_7

geometric manifold information [10], by considering better numerical integrators [6], or by relaxing the detailed balance constraint [24].

A different direction towards improving sampling performance is the slice sampler [19]. The slice sampler is related to HMC in the sense that both use auxiliary variables for efficient moves. These moves can be automatically adapted to match the relative scale of the local region being sampled [19]. The sampling procedure alternates between uniformly drawing samples from the target distribution and uniformly drawing the slice variables. Unlike HMC, slice sampling does not require local gradient information. Instead, the primary effort is to locate slice intervals, where the unnormalized density values are greater than the slicing variable. This is typically hard to compute directly, thus requires local search [19]. Further, it is generally less feasible in high-dimensional parameter spaces, because the slice interval is difficult to approximate. For example, using hyper-rectangle estimation may result in high rejection rates [19]. Elliptical slice sampling [17] alleviate this issue by slicing on a high dimensional elliptical curve parameterized by a single scalar. However it assumes the latent variable to be Gaussian distributed.

In this paper, we leverage the Hamiltonian-Jacobi equation from classical physics [11] to unveil a deeper connection between HMC with modified kinetics and standard slice sampling in one-dimensional cases. We propose an equivalent slice sampler, which exploits gradient information without evaluating the slice interval. We formally show that, in several univariate scenarios where theoretical analysis is tractable, the proposed sampler yields lower autocorrelation compared with standard HMC, thus potentially yielding higher effective sample sizes. Finally, we discuss the scenario where our method is most desirable and validate it with synthetic and real-world experiments.

2 Preliminaries

Hamiltonian Monte Carlo. Consider sampling from a probability density function $p(\boldsymbol{x}) \propto \exp[-E(\boldsymbol{x})]$, where $\boldsymbol{x} \in \mathbb{R}^d$ and $E(\boldsymbol{x})$ is the potential energy. One can augment the density with an auxiliary momentum random variable $\boldsymbol{p} \in \mathbb{R}^d$. By Assumption, $\boldsymbol{p}$ is independent of $\boldsymbol{x}$, and has a marginal Gaussian distribution with zero-mean and covariance matrix $\boldsymbol{M}$. The joint distribution $p(\boldsymbol{x}, \boldsymbol{p})$ is defined as $p(\boldsymbol{x}, \boldsymbol{p}) \propto \exp[-H(\boldsymbol{x}, \boldsymbol{p})] = \exp[-E(\boldsymbol{x}) - K(\boldsymbol{p})]$, where $H(\boldsymbol{x}, \boldsymbol{p})$ is the total energy or *Hamiltonian*, and $K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}$ is the kinetic energy. Hamiltonian Monte Carlo leverages Hamiltonian dynamics to propose new samples for $\boldsymbol{x}$, driven by the following ordinary differential equations (ODE):

$$\frac{d\boldsymbol{x}}{dt} = \nabla_p K(\boldsymbol{p}), \qquad \frac{d\boldsymbol{p}}{dt} = -\nabla_x E(\boldsymbol{x}). \tag{1}$$

The Hamiltonian is preserved under perfect simulation, *i.e*, it is constant over t. However, closed-form dynamic updates are typically infeasible. As a result, one typically employs numerical integrators, *e.g.*, the leap-frog [20], to approximate the Hamiltonian flow. If the integrator is symplectic, by Liouville's theorem, the corresponding sampler is invariant to the target distribution [20].

Slice sampling. Slice sampling [19] was originally proposed as an approach to overcome the need of manually selecting the proposal scale (or stepsize) in the Metropolis-Hastings algorithm. Slice sampling leverages the fact that sampling the unnormalized target distribution $f(x)$ can be perceived as sampling a joint distribution. Therefore, sampling from the points under the unnormalized density curve is the same as sampling from the target distribution. The iterative procedure consists the following *slicing* and *sampling* steps:

$$\begin{aligned} &\text{Slicing :} && p(y_t|x_t) = \frac{1}{f(x_t)}\,, && s.t.\ 0 < y_t < f(x_t) \\ &\text{Sampling :} && q(x_{t+1}|y_t) = \frac{1}{Z_2(y_t)}\,, && s.t.\ f(x_t) > y_t\,, \end{aligned} \tag{2}$$

where y is the augmented slicing variable. $f(x) \triangleq e^{-E(x)}$ is the unnormalized density and $Z_2(y) = \int_{f(x)>y} 1dx$ is the measure of regions that have functional values greater than the slice variable y. The density function is given by

$$p(x, y) = \begin{cases} \frac{1}{Z_1}\,, & 0 < y < f(x) \\ 0\,, & \text{otherwise} \end{cases}\,,$$

where $Z_1 = \int f(x)dx$ is the normalizing constant. The marginal distribution for x exactly recovers the target distribution $f(x)/Z_1$. The evaluation of slice interval $x : f(x) > y$ is typically non-trivial, where iterative procedures to adaptively capture the boundaries of such slice interval are used [19].

3 Canonical Transformation

In this section we use the *canonical transformation* and the *Hamilton-Jacobi equation* (HJE) [11] to reveal a connection between HMC with Laplace kinetics and slice sampling. Without loss of generality to the multivariate cases, for simplicity, here we detail our derivations for the univariate case.

Suppose the kinetic energy function $K(p)$ in HMC can be defined as an arbitrary function of p, as long as the $K(p)$ is convex and symmetric *w.r.t.* p. We consider two particular kinetics forms. The standard HMC uses quadratic kinetics $K(p) = p^2/m$, where m is the *mass parameter* and the marginal distribution of p is proportional to $e^{-K(p)}$, thus is Gaussian distributed with variance m.

We employ the canonical transformation from classical physics to transform the original HMC system (H, x, p, t) in (1), into a new system space, termed as canonical space [11]: (H', x', p', t). The transformation $(H, x, p, t) \to (H', x', p', t)$ satisfies the *Hamilton's principle* [11]:

$$\lambda(p \cdot \dot{x} - H) = p' \cdot \dot{x}' - H' + \frac{\delta G}{\delta t}\,, \tag{3}$$

where $\lambda \in \mathbb{R}$ is a constant, $\dot{x} \triangleq dx/dt$, δ denotes functional derivative and G is a user-defined generating function [25]. Such a generating function can be of several types; here we use a type-2 generating function defined as

$$G \triangleq -x' \cdot p' + S(x, p', t)\,.$$

The explicit form of $S(x, p', t)$ is defined below. By substituting G into (3), one can establish the following equations:

$$p = \frac{\partial S}{\partial x}\,, \qquad x' = \frac{\partial S}{\partial p'}\,, \qquad H'(x', p') = H(x, p) + \frac{\partial S}{\partial t}\,. \tag{4}$$

In the HJE, we let the new Hamiltonian H' to be zero, *i.e.*,

$$H(x, p) + \frac{\partial S}{\partial t} = H'(x', p') = 0\,. \tag{5}$$

The Hamilton-Jacobi equation states that after this transformation, the motion of particles collapse into a point in the new space, *i.e.*, (x', p') are constant over time [25].

Consider setting the *Hamilton's principal function* as $S(x, p', t) = W(x) - p't$, where $W(x)$ is an unknown function of x that needs to be solved. Thereby, (5) becomes

$$H(x, p) + \frac{\partial S}{\partial t} = H(x, p) - p' = 0\,. \tag{6}$$

The implication from (6) is that $p' = H$; *i.e.*, the *generalized momentum* in the new phase space, (x', p'), represents the total Hamiltonian in the original space. We consider the standard Gaussian kinetic function $K(p) = |p|^2/m$. From (4) and (5), we can solve the functional equation in (6) to obtain,

$$W(x) = \int_{x_{min}}^{x(t)} f(z)dz + C\,, \tag{7}$$

where $f(z) = H - E(z)$ if $z \in \mathbb{X} \triangleq \{x : H - E(x) \geq 0\}$, and 0 otherwise, and $x_{min} = \min\{x : x \in \mathbb{X}\}$. From (4), (6) and (7),

$$x' = \frac{\partial S}{\partial p'} = \frac{\partial W}{\partial H} - t = \frac{1}{2}\int_{x_{min}}^{x(t)} f(z)^{-1/2}dz - t\,. \tag{8}$$

Note that x' is a constant. In (8), $\int_{x_{min}}^{x(t)} f(z)^{-1/2}dz \in [0, \int_{\mathbb{X}}[H - E(z)]^{-1/2}dz]$. Our objective is to mimic the Hamiltonian dynamics evolving with a random evolution time, t. If we assume a closed contour, the Hamiltonian dynamics has period $T \triangleq \int_{\mathbb{X}}[H - E(z)]^{-1/2}dz$. To sample a new point $x(t)$ on the contour, one can first sample the time, t, constrained to a single period of movement, *i.e*,

$$t \sim \text{uniform}\left(-x', -x' + \int_{\mathbb{X}}[H - E(z)]^{-1/2}dz\right)\,. \tag{9}$$

where x' can be understood as the "initial" timestamp of x. With a sampled time t from (9), one could solve the Eq. (8) for $x^* \triangleq x(t)$, *i.e.*, the value of x at time t.

However, the integral in (8) is not always tractable. Note that the integral in (8) can be interpreted as (up to normalization) a cumulative density function (CDF) of x. As a result, one can circumvent uniformly sampling t from (9), by directly sampling x^* from the following density function

$$p(x^*|H) \propto \; [H - E(x^*)]^{-1/2}, \quad s.t., \;\; H - E(x^*) \geq 0\,. \tag{10}$$

Note that p^* is not of interest because it is discard after each dynamic update.

This transformation provides the basic setup to reveal the equivalence between the slice sampler and HMC, which is discussed in Sect. 4.

4 Laplacian HMC

Let $\mathcal{L}(\cdot; m)$ denote the Laplace distribution with scale parameter m, the probability density function is given by

$$\mathcal{L}(p; m) \propto \exp(-|p|/m)$$

We denote the non-standard HMC with Laplace distribution for the momentum variable as Laplacian HMC (L-HMC). Suppose we assume the momentum variable have Laplace kinetics, i.e. employing an ℓ_1 norm for the kinetic function, similar to the derivation in (10), we have

$$p(x^*|H) \propto \; 1\,, \quad s.t., \;\; H - E(x^*) \geq 0\,. \tag{11}$$

In light of the above observation, we propose to perform standard HMC and L-HMC with the procedure described in Algorithm 1.

Algorithm 1. HMC/L-HMC in canonical space.

```
Input: Sample size $T$, energies $E(x)$ and $K(p; m)$.
Output: Sample results, $\{x_0, \dots, x_T\}$.
Initialization: Choose initial sample point, $x_0$.
for $t \in \{1, \dots, T\}$ do
    Sample $p_t \sim \mathcal{N}(p; m)$ (standard HMC) or $\mathcal{L}(p; m)$ (L-HMC).
    Compute Hamiltonian: $H_t = E(x_t) + K(p_t)$.
    Compute $\mathbb{X} \triangleq \{x : x \in \mathbb{R}; E(x) \leq H_t\}$.
    Sample $q(x_{t+1}|H_t) \propto [H_t - E(x_{t+1})]^{-1/2}$(standard HMC) or $q(x_{t+1}|H_t) \propto 1$ (L-HMC), with $x_{t+1} \in \mathbb{X}$.
end for
```

Denote $y_t = e^{-H_t}$, the conditional updates for the L-HMC sampling procedure in Algorithm 1 share the same formulas as standard slice sampling in (2)

Note that the mass parameter m (scale parameter of the Laplace distribution) is cancelled out.

Accordingly, the equivalent non-standard slice sampling that corresponds to standard HMC can be written as

$$p(y_t|x_t) = \frac{1}{f(x_t)}[\log f(x_t) - \log y_t]^{-\frac{1}{2}}\,, s.t.\ \ 0 < y_t < f(x_t) \tag{12}$$

$$q(x_{t+1}|y_t) = \frac{1}{Z_2(y_t)}[\log f(x_{t+1}) - \log y_t]^{-\frac{1}{2}}\ .s.t.\ \ f(x_{t+1}) > y_t \tag{13}$$

We denote this slice sampler as HMC-SS (the slice sampler corresponding to standard HMC). This iterative procedure yields an invariant joint distribution

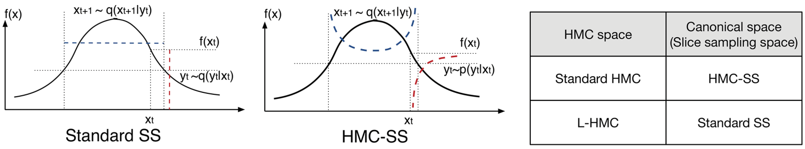
$$p(x,y) = \begin{cases} \frac{1}{\sqrt{\pi}Z_1}[\log f(x) - \log y]^{\frac{1}{2}}\,,\ 0 < y < f(x) \\ 0\,,\ \ \text{otherwise} \end{cases}\,,$$

leaving the marginal distribution for x as the desired target distribution, while the marginal distribution of y is given by

$$p(y) = Z_2(y)/(\sqrt{\pi}Z_1)\,. \tag{14}$$

The equivalent slice sampler for standard HMC and L-HMC is illustrated in Fig. 1. For HMC-SS, the conditional distribution of $q(x_{t+1}|y_t)$ is skewed, so that points that are close to the boundary of the slice interval are more likely to be drawn. In addition, from (12) the conditional draw of slice variable y_t given x_t tends to take values close to $f(x_t)$.

Intuitively, in contrast with the standard slice sampling, the auxiliary variable y_t in HMC-SS tend to stay close with $f(x_t)$, rendering x_{t+1} to be close to x_t. Thus the standard slice sampler with a larger a is expected to be more efficient. Based on the connection between HMC-SS and HMC, as well as standard SS with L-HMC, this seems suggest L-HMC is more efficient that standard HMC. We elaborate more about the mixing performance in Sect. 6.



HMC space	Canonical space (Slice sampling space)
Standard HMC	HMC-SS
L-HMC	Standard SS

Fig. 1. Standard slice sampling (Left). The equivalent slice sampler of standard HMC, HMC-SS (Middle). Mapping between HMC space and canonical space (Right). $y_t|x_t$ is sampled from (12) (red line) and $x_{t+1}|y_t$ from (13) (blue line). L-HMC is essentially the same but with $y_t|x_t$ and $x_{t+1}|y_t$ sampled from uniform distributions. (Color figure online)

5 Performing L-HMC with Numerical Integrators

Section 4 shows that performing L-HMC in the canonical space can be viewed as performing standard slice sampling. In practice, however, analytically solving the slice interval, $\mathbb{X}$, is typically infeasible. By leveraging the connection between L-HMC and standard slice sampling, one can perform the standard slice sampling in the original space using a numerical integrator, as done in standard HMC. Here we consider the second order Störmer-Verlet integration [20]. The updates for L-HMC[1] are thus given as the following leap-frog steps

$$\mathbf{p}_{t+1/2} = \mathbf{p}_t - \tfrac{1}{2}\varepsilon\nabla E(\mathbf{x}_t)\,, \tag{15}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon \mathrm{sign}(\mathbf{p})/m\,, \tag{16}$$

$$\mathbf{p}_{t+1} = \mathbf{p}_{t+1/2} - \tfrac{1}{2}\varepsilon\nabla E(\mathbf{x}_{t+1})\,, \tag{17}$$

Note that the mass matrix in our specification is $\mathbf{M} = m\mathbf{I}$. Here we use a random step size, ε, drawn from a uniform distribution with user-defined width, as suggested in [20]. Note that this specification is necessary for L-HMC to avoid moving on a fixed grid determined by ε.

Reflection. Another practical issue that comes with the fact that each contour in the phase space (x, p) has at least 2^D stiff (non-differentiable) points due to the non-differentiable kinetic function $K(\boldsymbol{p})$. The stiff points occur whenever the contour intersect with hyperplanes $p_d = 0$, for $d \in \{1 \cdots D\}$; D denotes the total dimension.

The naive leap-frog approach of L-HMC in (15)–(17) would lead to high integration errors, comparing with standard HMC, especially when the dimensionality is high. To alleviate this issue, we take a "reflection" action when encountering these stiff points, which shares some similarities with the "bouncing ball" strategy mentioned by [20]. Specifically, in (15) and (17), whenever the d-th component of momentum $p^{(d)}$ changes sign, we set $x_{t+1}^{(d)}$ and $p_{t+1}^{(d)}$ back to $x_t^{(d)}$ and $p_t^{(d)}$, and flip $p_t^{(d)} = -p_t^{(d)}$. A caveat of such a simple remedy lies in the fact that it may not guarantee the conservation of volume in phase space, thus may not leave the distribution invariant. Also, one will probably face "stickiness" in the high dimensional case [20]. This is because when negating the momentum in certain dimension(s), the next sample $\boldsymbol{x}_t$ may stay at the previous position, for those dimension(s). In high dimensions, this problem becomes more prominent since the chance of "reflection" for each update is considerably higher, yielding the whole sampler to perform less efficiently. Besides, this reflection strategy may render the sampler to be less sensitive on tail region of the target distribution. We note that this strategy may violate the invariance property. We hope to remark that the reflection is a first-remedy to ameliorate numerical difficulties. Nevertheless, this approach preserves the total Hamiltonian, and performs well in practice for low-dimensional cases.

[1] In the following, we denote L-HMC as the one in the original space, except otherwise explicitly stated.

Sample with constrained domain. As mentioned by [20], one could split the total Hamiltonian, to approach sampling from a bounded domain. An imaginary infinite potential energy can be imposed on regions that violate the constraints, which will give such points zero probability. Whenever the new proposed sample exceeds the constraint, we bounce the sample back. For example, when sampling from a truncated distribution with constraint $x^{(d)} > m$, if at time t the proposed $x_t^{(d)} < m$, the value $2m - x_t^{(d)}$ would be used instead, while the corresponding moment, $p_t^{(d)}$, changes sign.

Partial momentum refreshment. Using fewer number of leap-frog steps would reduce the computational cost of L-HMC, however rendering the algorithm less likely to adequately explore the contour and move to a distant point. [20] described a strategy to partially update the momentum variable, as an approach to further suppress the random-walk behavior when only a small number of leap-frog steps are taken, in which the distribution of the momentum would still be invariant. For the double-exponential kinetic energy form, one could consider a similar strategy to partially refresh the momentum. For the univariate case, without loss of generality to high dimensions, the update for momentum is given by

$$\tilde{p} = \min(p/\alpha, \eta/(1-\alpha))\mathrm{sign}(p)\,, \tag{18}$$

where $\alpha \in (0, 1)$ is a tuning parameter and η is an exponential random variable with mean $1/m$. It can be shown that $\tilde{p}$ has the same distribution as p. When α is close to 1, the generated new momentum $\tilde{p}$ would be similar to p. When α is close to zero, the absolute value of the new momentum becomes independent of its previous value. Similar to partial refreshment in standard HMC [20], one iteration applying the modification in (18) consists of three steps: (1) Updating momentum using (18), (2) performing a leap-frog discretization and Metropolis step, and (3) negating the momentum. In practice, the value of α has to be manually selected to achieve good performance.

Sampling discrete distributions. Sampling from discrete distributions such as Poisson, multinomial, Bernoulli, *etc.*, is generally infeasible for standard HMC, primarily due to the lack of gradient information. Recently proposed techniques tackle the discrete case by transforming it into sampling from a continuous distribution [21,26]. We show here that one can directly sample from a discrete distribution with L-HMC.

Notice from Eq. (16) that the update of $\boldsymbol{x}$ for each leap-frog discretization step depend only on the sign of the momentum variable $\boldsymbol{p}$. Based on this observation, one can sample a discrete distribution exactly, in an HMC manner. Consider a scenario, where a multivariate distribution with D dimensions is defined on an infinite grid with equidistant step m. Equation (16) allows the Hamiltonian dynamics to move in such a way, that each update in $\boldsymbol{x}$ moves with multiples of m, so as to stay on the grid. Meanwhile, the gradients in (15) and (17) are substituted with the difference vector $\triangle E(\boldsymbol{x})$, where its d-th component is $\triangle^{(d)} E(\boldsymbol{x}_{t-1/2}) \triangleq E(\boldsymbol{x}_t) - E(x_{t-1}^{(d)}, \boldsymbol{x}_t^{(-d)})$, $x^{(d)}$ denotes the d-th component of

$\boldsymbol{x}$ and $\boldsymbol{x}_t^{(-d)}$ denotes the remaining $D-1$ components. The iterative updates become

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon} \circ \mathrm{sign}(\boldsymbol{p})/m \,, \; \boldsymbol{p}_t = \boldsymbol{p}_{t-1} - \boldsymbol{\varepsilon} \circ \triangle E(\boldsymbol{x}_{t-1/2}) \,,$$

where the stepsize $\boldsymbol{\varepsilon}$ is constrained to $\mathbb{Z}^D$ and $\circ$ is the element-wise product. The reason that this strategy can not be applied to standard HMC is because in L-HMC, each increment $x_{t+1} - x_t$ is a constant that does not depend on the absolute value of momentum p, while in standard HMC, different value of p will yield different increment $x_{t+1} - x_t$. As a result, the sampler may not move on a uniform grid. In practice, one could sample $\boldsymbol{\varepsilon} \in \mathbb{R}^D$ and round it to the closest integer vector. It can be shown that the Hamiltonian is preserved under such procedure in univariate cases. For multivariate cases, the difference vector can be normalized to enforce the conservation of Hamiltonian, *i.e.* $\sum \triangle E(\boldsymbol{x}_{t-1/2}) = E(\boldsymbol{x}_t) - E(\boldsymbol{x}_{t-1})$. Note that when the Hamiltonian is preserved, the Metropolis-Hasting step can be omitted. As in the continuous scenario, the momentum is negated whenever it would change sign in the next iteration. This specification works well in practice for our tested cases when the dimensionality is low ($D < 5$), however, we remark that this specification would violate the volume preservation and is not the principled way to perform high-dimensional discrete sampling (when the dimensionality increase, the error between $E(\boldsymbol{x}_{t+1}) - E(\boldsymbol{x}_t)$ and $\triangle E(\boldsymbol{x}_t$ would inevitably become larger). How to perform a high dimensional discrete sampling remain as a interesting topic for future investigation. If $E(\boldsymbol{x})$ has well-defined gradient information over the real domain that covers the grid, one can relax the calculation to the continuous space, where the gradient $\nabla E(\boldsymbol{x})$ is computed, instead of D evaluations of the potential energy, $E(\boldsymbol{x})$.

Adaptive search. The fact that updating $\boldsymbol{x}$ does not explicitly involve $\boldsymbol{p}$ may have additional implications. Following [23], this observation enables applying adaptive search for appropriate scale of stepsize, $\boldsymbol{\varepsilon}$, based on the sufficient statistics from previous samples. For example, one could set the relative scale of the stepsize for each coordinate to match the diagonal elements from the empirical covariance matrix. Note that this strategy is particularly suitable to be applied to L-HMC, due to the fact that the update of the dynamics in L-HMC is moving exactly in the direction of the stepsize, $\boldsymbol{\varepsilon}$. This strategy would be expected to perform better than choosing a common stepsize for each dimension, when the landscape has different scales for each dimension. The convergence of adaptive parameters requires establishing regularity conditions [12]. Though it works well in many cases, it is known that this strategy results in a chain that is no longer Markovian, thus it will not always leave the target distribution invariant [23]. Besides, when the distribution has more than one mode, applying this method may render the sampler prone to get trapped into one of the modes.

6 Efficiency Analysis

We note that most of previous work of analyzing the mixing performance of HMC is based on empirical studies. Little work has been done on theoretical analysis [10,20]. Interestingly, we can leverage the implicit connection between HMC and slice sampling, to briefly touch on the analysis of the mixing performance for HMC and L-HMC from examining their corresponding slice samplers. We use the autocorrelation function and effective sample size to monitor mixing performance. We consider sampling from a univariate distribution $p(x) \propto e^{-E(x)}$ for the analysis. The one-time-lag autocorrelation for HMC and L-HMC, $\rho(1)$, is given by

$$\begin{aligned} \rho(1) = & \ (\mathbb{E}[x_t x_{t+1}] - \mathbb{E}[x]^2)/\mathrm{Var}(x)\,. && (19) \\ = & \ (\ \mathbb{E}_{p(y_t)}[\mathbb{E}_{q(x_{t+1}|y_t)}[x_{t+1}]]^2 - \mathbb{E}[x]^2)/\mathrm{Var}(x) && (20) \end{aligned}$$

From (12) and stationary assumption, for standard HMC

$$q(x_t|y_t) \propto p(y_t|x_t)p(x_t) \propto [\log f(x_t) - \log y_t]^{-1/2}, s.t. \ \ f(x_t) > y_t$$

For L-HMC, $q(x_t|y_t) \propto 1, s.t. \ \ f(x_t) > y_t$

Given the potential energy form $E(x)$, $\rho(1)$ can be computed from (14), (20) and (2). The h-time-lag autocorrelation function can be obtained as

$$\rho(h) = \ (\mathbb{E}_{p(x)}[\mathbb{E}_{\kappa_h(x'|x)}[x'x]] - \mathbb{E}[x]^2)/\mathrm{Var}(x)\,,$$

where, $\kappa_h(x_{t+h}|x_t)$ represents the *h-order transition kernel*, and can be calculated recursively as

$$\begin{aligned} \kappa_1(x_{t+1}|x_t) &= \int q(x_{t+1}|y_t)p(y_t|x_t)dy_t\,, \\ \kappa_h(x_{t+h}|x_t) &= \int \kappa_{h-1}(x'|x_t)\kappa_1(x_{t+h}|x')dx'\,. \end{aligned}$$

Finally, the resulting Effective Sample Size (ESS) [5] is given by ESS $= N/(1+2\times \sum_{h=1}^{\infty} \rho(h))$. Analyzing the efficiency of L-HMC for the general case is difficult, however, we can specify a special case where the ESS can be explicitly calculated.

We consider a simple case to assess the efficiency of standard HMC and L-HMC. We aim to sample from a univariate exponential distribution, $\mathrm{Exp}(x;\theta)$, with energy function, $E(x) = \theta x$, for $x > 0$. From the above analysis, for standard HMC

$$\rho(1) = \frac{2}{3}\,, \quad \rho(h) = (\frac{2}{3})^h\,, \quad \mathrm{ESS} = \frac{N}{5}\,,$$

For L-HMC, we have

$$\rho(1) = \frac{1}{2}\,, \quad \rho(h) = (\frac{1}{2})^h\,, \quad \mathrm{ESS} = \frac{N}{3}\,,$$

We observe that the ESS becomes larger with L-HMC. As a result, under these conditions, and many other univariate cases discussed in the experiments, L-HMC has a theoretical advantage of the mixing rate in stationary period over standard HMC. This observation is consistent with the intuition discussed in Sect. 4.

7 Experiments

7.1 Synthetic Toy Examples

We conduct several experiments to validate the theoretical results, as well as the performance of standard HMC and L-HMC.

Synthetic 1D problems. We first perform our experiments on several univariate distributions, where evaluation of theoretical mixing performance is possible. Our primary objective for this simulation study is to validate that the theoretical results are consistent with the empirical results. Each density is given by $p(x) = \frac{1}{Z_1}\exp(-E(x))$, $s.t\ x \geq 0$ and

- Exponential distribution: $\mathrm{Exp}(x;\theta)$, where $E(x) = \theta x$.
- Truncated Gaussian: $\mathcal{N}_+(x;0,\theta)$, where $E(x) = \theta x^2$.

We truncate the Gaussian distribution to the positive side, because for a symmetric distribution the theoretical autocorrelation is always 0, thus rendering the comparison less interesting. Note that for each case, as long as the parameter $\theta > 0$, the performance of the sampler does not depend on θ.

We perform standard HMC and L-HMC, as well as "analytic" slice sampling[2] when available. We collected 30,000 Monte Carlo samples, with 10,000 burn-in samples. The leap-frog steps are set to 100 for each experiment. The mass parameter m and stepsize ε are selected manually to achieve around 0.9 acceptance ratio. We observed that applying the partial momentum refreshment can provided additional help, especially when taking fewer leap-frog steps. However, the improvements are not significant when the number of leap-frog steps is adequate for the tested cases.

As shown in Table 1, in the tested cases, theoretical autocorrelations and ESS match well with empirical performance of standard HMC, L-HMC and analytic slice samplers. In every case, L-HMC obtained better empirical results, which is consistent with our theoretical analysis.

Sampling from a discrete distribution. To demonstrate that the L-HMC can perform sampling of distributions with discrete support, we consider sampling from a univariate Poisson distribution, $\mathcal{P}(\lambda)$, with fixed rate parameter λ (we use $\lambda = 10$ in our experiment). The potential energy is given by

[2] Analytic slice sampling is achieved by analytically solving the slice interval and computing the expectation in (20), and is only available for exponential and positive-truncated Gaussian cases.

Table 1. 1D theoretical (Th.) and empirical $\rho(1)$ and ESS. SS denotes the analytical slice sampler corresponding to standard HMC or L-HMC.

	Th. $\rho(1)$	Th. ESS	SS $\rho(1)$	SS ESS	(L-)HMC $\rho(1)$	(L-)HMC ESS
standard HMC (Exp)	0.6667	6000	0.6620	6204	0.6711	6069
L-HMC (Exp)	0.5	10000	0.4868	10227	0.5218	9773
standard HMC ($\mathcal{N}_+$)	0.4787	10576	0.4736	10705	0.4802	10510
L-HMC ($\mathcal{N}_+$)	0.3120	15732	0.3040	15457	0.3061	15595

$E(x) = -x \log \lambda + \log x!$. We apply the update scheme described in Sect. 5, and run 10,000 iterations with 3,000 burn-in samples. The number of iterative dynamic updates, stepsize, and mass parameter m were set to 15, 2, and 1, respectively. Results are shown in Fig. 2. The empirical results match well with the probability mass function of $\mathcal{P}(\lambda)$ with $\lambda = 10$. The acceptance ratio is always one, as during the iterative process, the Hamiltonian is exactly conserved. As a consequence, the Metropolis step can be omitted. The empirical $\rho(1)$ and ESS are 0.024 and $9,984$, respectively.

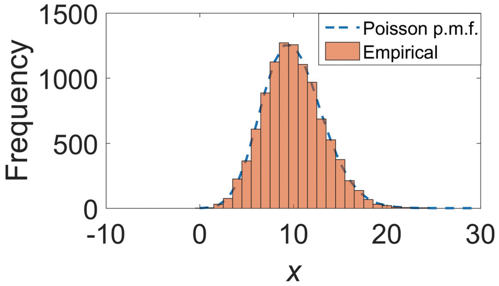


Fig. 2. Histogram of samples for a Poisson distribution, $x \sim \mathcal{P}(\lambda)$ with $\lambda = 10$.

We also apply our methods to sample from a bivariate Poisson distribution [14]. The bivariate Poisson with random covariates (z_1, z_2) can be constructed as $z_1 = y_1 + y_3, z_2 = y_1 + y_2$, where (y_1, y_2, y_3) are three independent Poisson variables with mean parameters $(\lambda_1, \lambda_2, \lambda_3)$. The probability function can be written as

$$\Pr(z_1 = k_1, z_2 = k_2) = \exp(-\lambda_1 - \lambda_2 - \lambda_3)\frac{\lambda_1^{k_1}}{k_1!}\frac{\lambda_2^{k_2}}{k_2!}\sum_{k=0}^{k_1 \wedge k_2}\binom{k_1}{k}\binom{k_2}{k}k!(\frac{\lambda_3}{\lambda_1\lambda_2})^k,$$

We set the ground truth model parameters to $(\lambda_1, \lambda_2, \lambda_3) = (1, 2, 3)$. The dynamic update step, stepsize and mass parameter m are set to be 10, 1 and 1, respectively. When performing the discrete sampling, we normalized the difference vector to enforce the total Hamiltonian to be conserved. We collect 10,000 Monte Carlo samples after 3,000 burn-in samples. The sampled distribution is shown in Fig. 3. The theoretical Pearson correlation for the target bivariate Poisson distribution is given by $\frac{\lambda_3}{\sqrt{\lambda_1+\lambda_3}\sqrt{\lambda_2+\lambda_3}} = 0.6708$. We observed that the empirical Pearson correlation is 0.6983, which matches well with the theoretical value. We also observed that when the dimensionality increases, the discrepancy between target distribution and empirical estimated distribution becomes larger. For this reason, we suggest to consider our method only for low dimensional sampling tasks. How to use HMC to sample from high-dimensional distributions is left for interesting future work.

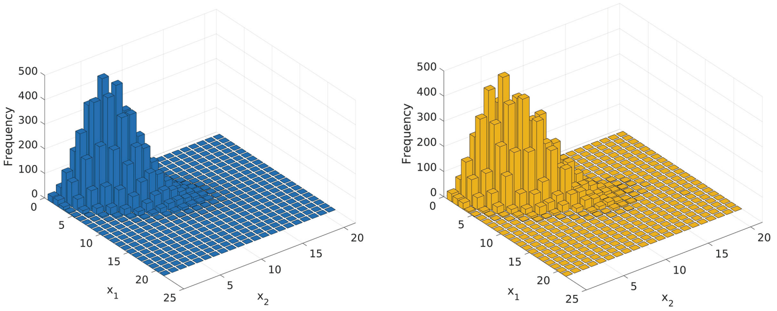


Fig. 3. Histogram of samples for bivariate Poisson distribution parameterized by $(\lambda_1, \lambda_2, \lambda_3) = (1, 2, 3)$. Left: theoretical sample frequency for target distribution. Right: samples from discrete L-HMC.

High-dimensional synthetic problems. We test the performance of standard HMC and L-HMC when sampling a high-dimensional Gaussian distribution. We consider a 100-dimensional Gaussian distribution with zero-mean and diagonal covariance matrix, with its diagonal elements uniformly drawn from $(0, 10]$. We ran 5,000 MC iterations, after 2,500 burn-in samples. For both standard HMC and L-HMC, we use 5 different leap-frog stepsizes, ε_t, $t = \{1, \ldots, 5\}$, where $\varepsilon_{t+1} = 0.8\varepsilon_t$. This scheme allows us to find the elbow points where performance is optimal. The ε_1 and m for standard HMC and L-HMC are set to $(0.025, 2)$ and $(0.015, 1)$, respectively. The sampler was initialized at MLE (estimated by gradient descent) to accelerate burn-in period.

We also compared with the adaptive scheme described in Sect. 5, where the stepsize is automatically tuned at each 500 interactions during the burn-in rounds using an empirically estimated covariance. The adaptation is stopped after burn-in, as suggested by [22]. Both L-HMC and adaptive L-HMC achieved median effective sample size near to the full sample size, and obtained a lower discrepancy between the empirically estimated covariance and the ground truth than standard HMC, see Fig. 4 (left). Employing the adaptive scheme improved the median ESS, probably due to the fact that the stepsize learned from the samples can automatically match the scale of each dimension, Fig. 4 (right).

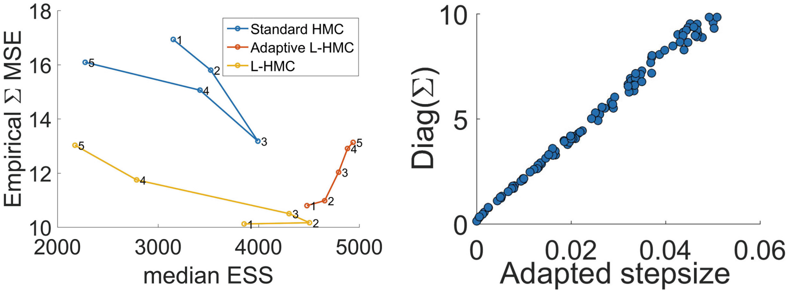


Fig. 4. Standard HMC and L-HMC performance on a 100-dimensional simulated Gaussian distribution. Left: Mean Squared Error (MSE) of estimated Σ vs. median ESS. Labels denote the stepsize index. Right: Elements of diag(Σ) vs. the adapted stepsize after 2,500 burn-in rounds.

7.2 Real Data Analysis

We perform an empirical comparison on two real-world probabilistic modeling tasks: Bayesian Logistic Regression (BLR) and Latent Dirichlet Allocation (LDA).

Bayesian logistic regression. We evaluated the mixing performance of standard HMC and L-HMC on 5 Bayesian logistic regression datasets from the UCI repository [2]. For data $\boldsymbol{X} \in \mathbb{R}^{d\times N}$, response variable $\boldsymbol{t} \in \{0,1\}^N$ and target parameters $\boldsymbol{\beta} \in \mathbb{R}^d$, suppose a Gaussian prior is imposed $\mathcal{N}(\mathbf{0}, \alpha\boldsymbol{I})$ (where $\alpha > 0$) on $\boldsymbol{\beta}$, the log posterior is given by [10],

$$\mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\boldsymbol{X}\boldsymbol{t} - \sum_{n=1}^{N}\log(1+\exp(\boldsymbol{\beta}^T\boldsymbol{X}^T_{n,\cdot})) - \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2\alpha}$$

Feature dimensions range from 7 to 15 and the number of data instances are between 250 and $1,000$. All datasets are normalized to have zero mean and unit variance. The sampler was initialized at gradient estimated MLE as in above experiments.

The mass matrix for kinetic function is defined as $\boldsymbol{M} = m \times \boldsymbol{I}$, where m is mass parameter. Gaussian priors $\mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$ were imposed on the regression coefficients. The leap-frog steps were set to be uniformly drawn from $[1, 100]$, as suggested by [20]. We manually select the stepsize and mass parameter m, so that the acceptance ratios fall in $[0.6, 0.9]$ [3]. On each dataset, the running time for each method is roughly identical, due to the fact that each method took approximately the same number of leap-frog steps. All experiments are based on 5,000 samples, with 1,000 burn-in samples.

Since the MCMC methods that we compared are asymptotically exact to the true posterior, the sample-based estimator is guaranteed to converge to the true expectation over the posterior. ESS indicates the variance of sample based estimator, thus is a good metric for comparison. For this reason, following [6,10,21],

Table 2. The minimum effective sample size, as well as the AUROC (in parenthesis) for each method. Dimensionality of each dataset is indicated in parenthesis after the name of each dataset.

Dataset (D)	Australian (15)	German (25)	Heart (14)	Pima (8)	Ripley (7)
Standard HMC	3124 (0.92)	3447 (0.78)	3524 (0.92)	3434 (0.90)	3317 (0.99)
L-HMC	**4308** (0.93)	**4353** (0.79)	**4591** (0.93)	**4664** (0.88)	**4226** (0.99)

Table 3. MNIST results. $D = 101, N = 12,214$. Total sample size is 4,000. AR denotes acceptance ratio.

	ESS min	Median	Max	Time (s)	AR
Standard HMC	2812	3441	3807	287.8	0.978
L-HMC	**3198**	**3808**	**4000**	291.0	0.968

we primarily compare on each method in terms of minimum ESS. We also evaluate the average predictive AUROC based on 10 fold cross-validation, the results showed no significant differences between standard HMC and L-HMC. The results are summarized in Table 2. L-HMC outperforms standard HMC in all datasets.

To further assess the scalability to high-dimensional problems, we also conduct an experiment on the MNIST dataset restricted to digits 7 and 9. We use 12,214 training instances, where the first 100 components from PCA were employed as regression features [6]. We ran $4,000$ MC iterations with $1,000$ burn-in samples, the results are shown in Table 3. L-HMC scales well, and achieved better mixing performance than standard HMC, while taking roughly the same running time. The acceptance ratio of L-HMC decreased by 0.01 *w.r.t.* standard HMC, presumably because the contours for L-HMC are slightly stiffer than those for standard HMC.

Topic modeling. We also evaluate our methods with LDA [4]. LDA models a document as a mixture of multinomial distributions over a vocabulary of size V. The multinomial distributions are parametrized by $\phi_k \in \Delta^V$ for $k = 1, \ldots, K$, where Δ^V denotes the V-dimensional simplex. Each ϕ_k is associated with a symmetric Dirichlet prior with parameter β. Specifically, the generative process for a document is as follows:

- For each topic k, sample a topic-word distribution: $\phi_k | \beta \sim \text{Dirichlet}(\beta)$.
- For each document d, sample a topic distribution: $\theta_d | \alpha \sim \text{Dirichlet}(\alpha)$.
    - For each word i, sample a topic indicator: $z_{di} | \theta_d \sim \text{Discrete}(\theta_d)$.
    - Sample an observed word: $w_{di} | \phi_{z_{di}} \sim \text{Discrete}(\phi_{z_{di} i})$.

To apply the L-HMC and standard HMC, following [8], we re-parametrize ϕ_k with $\tilde{\phi}_k$ as $\phi_{ki} = e^{\tilde{\phi}_{ki}} / (\sum_j e^{\tilde{\phi}_{kj}})$. Similar to [8], a semi-collapsed LDA formulation is used for sampling, where the distribution over topics for each document is integrated out. We use the ICML dataset [7] for the experiment, which

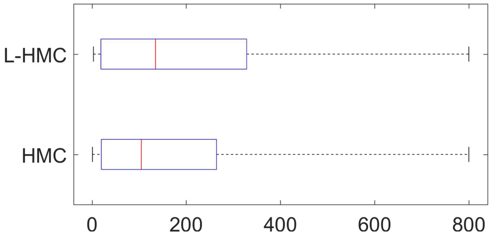


Fig. 5. Empirical distribution of coordinate-wise effective sample size of standard HMC and L-HMC, over 57,540 dimensions.

contains 765 documents corresponding to abstracts of ICML proceedings from 2007 to 2011. After stopword removal, we obtain a vocabulary size of 1,918 and total words of about 44K. We used 80 % of the documents for training and the remaining 20 % for testing. The number of topics is set to 30, resulting in 57,540 parameters. We use a symmetric Dirichlet prior (*i.e.*, all of the elements of parameter vector $\boldsymbol{\beta}$ have the same value) with parameter $\beta = 0.1$. All experiments are based on 800 MCMC samples with 200 burn-in rounds. We set the stepsizes to be 2.0 for both L-HMC and standard HMC, to obtain acceptance ratios around 0.68. For each iteration we set 20 leap-frog steps. L-HMC has best mixing performance as seen in Fig. 5, and the perplexity is comparable with standard HMC. The perplexities for L-HMC and standard HMC is 958 and 963, respectively.

8 Conclusion

We demonstrated the equivalency between the slice sampler and HMC with a Laplace kinetic energy. This enables us to perform the leap-frog numerical integrator for standard slice sampling in high-dimensional space. We further demonstrated that the resulting sampler can be applied to sampling from discrete distributions, *e.g.*, Poisson. Our method can be seen as a drop-in replacement for scenarios where standard HMC applies, and thus it has many potential extensions. However, our method has its limitations. For high dimensional problems, the numerical issues associated with the sampler are less negligible, and requires carefully selecting the sampler parameters. Future directions include (1) employing more sophisticated numerical methods to reduce the numerical error of our L-HMC approach (2) formal study of the ESS of the proposed L-HMC compared to standard HMC, and (3) exploiting geometric information [10] in the leap-frog updates to further improve the sampling efficiency.

Acknowledgments. The research reported here was supported in part by ARO, DARPA, DOE, NGA and ONR.

References

1. Andrieu, C., Thoms, J.: A tutorial on adaptive mcmc. Stat. Comput. **18**, 4 (2008)
2. Bache, K., Lichman, M.: UCI machine learning repository (2013)
3. Betancourt, M., Byrne, S., Girolami, M.: Optimizing the integrator step size for Hamiltonian Monte Carlo. ArXiv (2014)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3** (2003)
5. Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton (2011)
6. Chao, W.-L., Solomon, J., Michels, D., Sha, F.: Exponential integration for Hamiltonian Monte Carlo. In: ICML (2015)
7. Chen, C., Rao, V., Buntine, W., Whye Teh, Y.: Dependent normalized random measures. In: ICML (2013)
8. Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R.D., Neven, H.: Bayesian sampling using stochastic gradient thermostats. In: NIPS (2014)
9. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**, 2 (1987)
10. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. Roy. Stat. Soc. Ser. B (Stat. Method.) **73**, 2 (2011)
11. Goldstein, H.: Classical Mechanics. Pearson Education India, New Delhi (1965)
12. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli (2001)
13. Homan, M.D., Gelman, A.: The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res. **15**, 1 (2014)
14. Karlis, D., Meligkotsidou, L.: Multivariate poisson regression with covariance structure. Stat. Comput. **15**, 4 (2005)
15. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 6 (1953)
16. Mohamed, S., De Freitas, N., et al.: Adaptive hamiltonian and riemann manifold monte carlo samplers. Arxiv (2013)
17. Murray, I., Adams, R.P., MacKay, D.J.: Elliptical slice sampling. ArXiv (2009)
18. Neal, R.M.: Probabilistic inference using markov chain monte carlo methods. Technical report CRG-TR-93-1 (1993)
19. Neal, R.M.: Slice sampling. Ann. Stat. **31**, 705–767 (2003)
20. Neal, R.M.: MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo 2 (2011)
21. Pakman, A., Paninski, L.: Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. In: NIPS (2013)
22. Robert, C., Casella, G.: Monte Carlo statistical methods. Springer Science & Business Media, New York (2004)
23. Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. J. Comput. Graph. Stat. **18**, 2 (2009)
24. Sohl-Dickstein, J., Mudigonda, M., DeWeese, M.R.: Hamiltonian monte carlo without detailed balance. ArXiv (2014)
25. Taylor, J.R.: Classical Mechanics. University Science Books, Colorado (2005)
26. Zhang, Y., Ghahramani, Z., Storkey, A.J., Sutton, C.A.: Continuous relaxations for discrete Hamiltonian Monte Carlo. In: NIPS (2012)