Refractive Sampling

Abstract

Hamiltonian Monte Carlo (HMC) is an efficient, black-box Markov Chain Monte Carlo algorithm for sampling from densities where a gradient is available. However, the HMC proposal parameters – the stepsize ε and the number of "leapfrog" steps L – can be difficult to tune, and there have been many recent advances in alleviating the issues associated with tuning HMC. In this work we claim that the main difficulty in tuning HMC is its sensitivity to the gradient magnitudes of the target distribution, which may be highly variable across the parameter space. We propose *refractive sampling* as an alternative to HMC that makes use of only the gradient direction, and not its magnitude. We apply this sampler to three models and show that refractive sampling with minimal parameter tuning is generally able to find better modes than HMC.

1. Introduction

Markov Chain Monte Carlo (MCMC) is an effective tool for performing Bayesian inference. Frequently, a practitioner must design and implement a MCMC algorithm that is suited to his or her problem, which can be time consuming and error-prone. Tools that allow the general application of MCMC to wide varieties of models are thus attractive. State-of-the-art black-box samplers such as slice sampling (Neal, 2003) and Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011) that require only a logpdf (and perhaps its gradient) as input are hence popular tools for Bayesian modelling. Even so, there are some drawbacks: slice sampling does not generalize well to problems in high dimensions, and HMC has some associated hyperparameters which can be difficult to tune. In this work we propose refractive sampling, a new black-box sampler that makes effective use of gradient information while remaining easy to tune in complex settings.

HMC is a discrete approximation to a continuous Markov chain in which a gradient update is followed for a number of "leapfrog" steps L, and then accepted or rejected. The stepsize parameter ε controls the error of the approximation in the proposal and influences the acceptance probability. HMC with $L = 1$ is also known as the Metropolis Adjusted Langevin Algorithm (MALA), where the (proposed) updates follow:

$$x^{(t+1)} \leftarrow x^{(t)} + \frac{\varepsilon^2}{2}\nabla_x f(x^{(t)}) + \varepsilon\mathcal{N}(0,1)$$

where $f(x)$ is the log density of the target distribution. Note that ε controls both the ratio of noise to gradient in the update, as well as the size of the update.

Typically, there is a narrow range of ε that will produce reasonable acceptance rates while producing large steps. If ε is not chosen carefully, HMC will exhibit either low acceptance rates or very small updates per iteration. This problem becomes evident when the gradients become large, particularly when the curvature of the logpdf is also extreme. When the gradient is large, ε must be small, and thus the steps performed in regions with smaller gradients will be small as well. Techniques such as introducing preconditioning matrices, updates in alternate geometries (Girolami & Calderhead, 2011), allowing an intelligent or automatically tuned number of steps (Hoffman & Gelman, 2011; Wang et al., 2013), and automatically tuning ε (Robbins & Monro, 1951; Hoffman & Gelman, 2011; Wang et al., 2013) are thus popular for their ability to improve acceptance rates and allow larger steps. Even so, these extensions still rely on an underlying sampler that can be sensitive to markedly fluctuating gradients.

Additionally, in many applications updates are performed in a Gibbs sampling style, where different parameter sets are updated in turn, often because each set requires different hyperparameter settings, or one set has closed form updates. This complicates algorithms using automatic tuning, though the benefits may outweigh the costs as automatic tuning easily allows each parameter set to have its own hyperparameter settings. In cases where one set of parameters has more degrees of freedom or is more flexible than another, the more flexible parameters can update *too quickly* and take the chain into a mode that underfits or overfits the data. For example, if one were sampling the means and covariances of a Gaussian Mixture Model using a black-box

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

sampler, allowing the covariances to update too quickly can result in chains where a few large components (poorly) explain all the data. Because HMC has a narrow range of hyperparameter settings that allow for efficient sampling, it can be difficult to tune multiple HMC algorithms so that some update more slowly than others while still giving reasonably efficient sampling.

The main issue is that the choice of ε is sensitive to the magnitudes of the gradients found in the log-posterior. MCMC algorithms that instead use only the gradient direction, and not its magnitude, may be able to escape these problems.

We take reflective slice sampling and the notion of refraction as points of inspiration. In reflective slice sampling, updates "bounce" around a multivariate slice by reflecting off of the boundaries of the slice. The reflections occur off of the normal to the boundary, which is defined by the local gradient. When a ray of light refracts when travelling from a medium of lower to higher refractive index, it refracts "inwards" relative to the normal of the surface. If the surface is defined as a contour of a posterior, then this refraction will correspond to rotating a ray towards the gradient by some amount. Refraction is invertible and so valid MCMC moves can be constructed from the refraction transformation.

2. Reflective Slice Sampling

Reflective slice sampling (Neal, 2003) is a multivariate extension of slice sampling in which uniform draws are taken from a slice by moving in some dirction p and reflecting at the boundaries of the slice. Given a logpdf $f(x)$ and its gradient $g(x)$, we begin at state x_0, sample $p \sim N(0, I)$, and perform the following updates for a predetermined number of steps m:

$$\begin{aligned} x^{(1/2)} &= x + wp \\ p' &= \begin{cases} p & \text{if } f(x^{(1/2)}) > f(x_0) \\ p - 2g\frac{p^T g(x)}{||g(x)||^2} & \text{otherwise} \end{cases} \\ x' &= x + wp' \end{aligned} \tag{1}$$

These moves leave the joint distribution of x and p invariant as the Jacobian determinant of the joint transformation is one:

$$\left|\begin{array}{cc} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial p} \\ \frac{\partial p'}{\partial x} & \frac{\partial p'}{\partial p} \end{array}\right| = \left|\begin{array}{cc} I + w\frac{\partial p'}{\partial x} & w\frac{\partial p'}{\partial p} \\ \frac{\partial p'}{\partial x} & \frac{\partial p'}{\partial p} \end{array}\right| = \left|\frac{\partial p'}{\partial p}\right| = 1$$

where we have used the following determinant identity for block matrices: $\left|\begin{array}{cc} A & B \\ C & D \end{array}\right| = |D| \cdot |A - BD^{-1}C|$ for invertible D.

3. Refractive Sampling

We wish to construct a MCMC proposal scheme like the one above, but that makes stronger use of the normalized gradient than reflective slice sampling. We would also like it to be easy to tune and not sensitive to the peculiarities of the target distribution – therefore we still desire a proposal that preserves the norm of p rather than allowing the size of steps to grow or shrink with each step taken. One obvious choice is then to add the normalized gradient u to p, and then normalize p to keep the norm preserved. It is certainly possible to define proposals according to this add-then-normalize process, however, a few complications arise. For one, the reverse move does not take the form of a simple convex combination of p and u – one must solve for the p' that would have resulted in p from adding u and normalizing.

Refraction is similar to this process, and is in fact identical if we take $q = w_1 p + w_2 u$ and $p' = \frac{q}{||q||}$ with weights chosen so that $r_1 = w_1||p||$ and $r_2 = ||w_1 p + w_2 u||$ are constants (r_1 and r_2 are the refractive index parameters used in refractive sampling). Thus refractive sampling can be thought of as the add-then-normalize process described above, except that u is given a higher weight as p becomes more orthogonal to u.

Refraction occurs according to Snell's Law as follows: if a ray of light p travelling in a medium with index of refraction r_1 passes through a boundary to another medium with index of refraction r_2, then the ray refracts to a ray p'. If the boundary has surface unit normal u (defined so that $p^T u > 0$), then the angle of incidence θ_1 and angle of refraction θ_2 are determined by

$$\cos\theta_1 = \frac{p^T u}{||p||} \tag{2}$$

$$\cos\theta_2 = \left[1 - \frac{r_1^2}{r_2^2}\left(1 - \cos^2\theta_1\right)\right]^{\frac{1}{2}} \tag{3}$$

The refracted ray p' can be constructed as:

$$p' = \frac{r_1}{r_2}p - ||p||\left[\frac{r_1}{r_2}\cos\theta_1 - \cos\theta_2\right]u \tag{4}$$

This transformation preserves the norm: $||p'|| = ||p||$. If $r_2 > r_1$ then p is rotated towards u, otherwise it is rotated away.

Refractive sampling makes use of the same $x' = x + wp'$ scheme above for updating x, but uses an alternative update for p'. It also breaks from reflective slice sampling in that it is not a slice sampler, rather it is a gradient-informed Metropolis Hastings sampler. We still take $p \sim \mathcal{N}(0, I)$ or some other symmetric distribution.

Note that in reflective slice sampling, the fact that the Jacobian is equal to $\left|\frac{\partial p'}{\partial p}\right|$ doesn't depend on the actual form of p', as long as its Jacobian is nonzero. Thus we are free to pick any form of p' we wish under this scheme as long as we correct for transformations that have nonunit Jacobian in a Metropolis Hastings accept/reject step. To this end we update p according to refraction, where we are refracting into a higher index of refraction from a lower one if $p^T g(x) < 0$, and vice versa if $p^T g(x) > 0$. That is, the gradient will always be pointing into the side with higher index of refraction, so that moves into the half-plane in which the gradient points will rotate towards the gradient, and moves out of the half-plane will rotate away from the negative gradient.

The full transformation for (x, p) is defined as follows:

$$
\begin{aligned}
(u, r_1, r_2) &= \begin{cases} (\frac{g(x)}{||g(x)||}, 1, r) & \text{if } p^T g(x) > 0 \\ (-\frac{g(x)}{||g(x)||}, r, 1) & \text{otherwise} \end{cases} \\
\cos\theta_1 &= \frac{p^T u}{||p||} \\
\cos\theta_2 &= \left[1 - \frac{r_1^2}{r_2^2}\left(1 - \cos^2\theta_1\right)\right]^{\frac{1}{2}} \\
p' &= \frac{r_1}{r_2} p - ||p|| \left[\frac{r_1}{r_2}\cos\theta_1 - \cos\theta_2\right] u \\
x' &= x + w p'
\end{aligned}
\tag{5}
$$

where r is a parameter of the procedure defining the ratio between the indices of refraction r_1 and r_2, and u is the normalized gradient, but sign-flipped so that $p^T u > 0$. The Jacobian of this transformation is

$$
\begin{aligned}
&\left|\frac{\partial p'}{\partial p}\right| = \\
&\det\left(\frac{r_1}{r_2} I + \cos\theta_2 \left[1 - \left(\frac{r_1 \cos\theta_1}{r_2 \cos\theta_2}\right)^2\right] \frac{p u^T}{||p||} - \right. \\
&\left. \frac{r_1}{r_2}\left[1 - \frac{r_1 \cos\theta_1}{r_2 \cos\theta_2}\right] u u^T\right) = \\
&\left(\frac{r_1}{r_2}\right)^{d-1} \frac{\cos\theta_1}{\cos\theta_2}
\end{aligned}
\tag{6}
$$

where d is the dimension of p. This transformation is illustrated in Figure 1, and example trajectories are given in Figure 2.

Note that $\cos^2\theta_2$ can be negative; this occurs when moving from a medium of higher index of refraction to lower, and the angle of incidence is too shallow. In this case, the reverse move is impossible, and so we instead reflect[1].

[1]Incidentally, this is what occurs in nature as "total internal reflection"

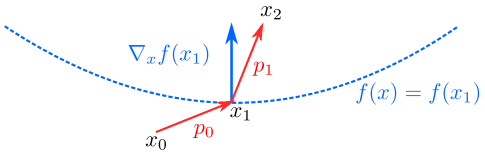

Figure 1. One step of a refractive sampling proposal. From state (x_0, p_0) we arrive to x_1. p_0 is then refracted through a surface with normal $\nabla_x f(x_1)$ to a produce a p_1 that has been rotated in towards the gradient. Going in reverse with $p = -p_1$ would result in $p' = -p_0$.

Thus, we can perform these updates for m steps[2], and then decide to accept or reject. Let $y = (x, p)$, $T(y)$ be the refraction transformation (5), and for $m = 1$, we need to choose acceptance probabilities α to satisfy detailed balance:

$$
\pi(y)\alpha(y \to T(y)) = \pi(T(y))\left|\frac{\partial T(y)}{\partial y}\right|\alpha(T(y) \to y) \tag{7}
$$

Where π is the target measure. Taking $y' = T(y)$, we have

$$
\alpha(y \to y') = \min\left[1, \frac{\pi(x')}{\pi(x)}\left(\frac{r_1}{r_2}\right)^{d-1}\frac{\cos\theta_1}{\cos\theta_2}\right] \tag{8}
$$

For $m > 1$ we simply have a product over multiple Jacobians.

To be clear, it is worth noting that the notion of a medium with a static index of refraction doesn't apply here: the indices of refraction used to refract p are determined entirely by the local gradient and its inner product with p. We are not attaching indices of refraction to various regions of parameter space, and using such a scaffold to propose updates. Rather, the index of refraction associated with one region may be different from iteration to iteration. See Algorithm 1 for pseudocode.

4. Experiments

In preliminary experiments comparing Refractive Sampling, HMC, NUTS, and reflective slice sampling on a Gaussian target distribution, we found that reflective slice sampling was about ten times slower per effective sample than the other algorithms; thus we focus here on the first three algorithms.

4.1. Bimodal Distribution

We begin with a simple demonstration on a two dimensional bimodal target distribution. We define the target

[2]Technically, we should flip p' after refraction and before the accept/reject step to maintain reversibility, but we omit this detail in the following analysis because we have chosen $\pi(p)$ to be symmetric.

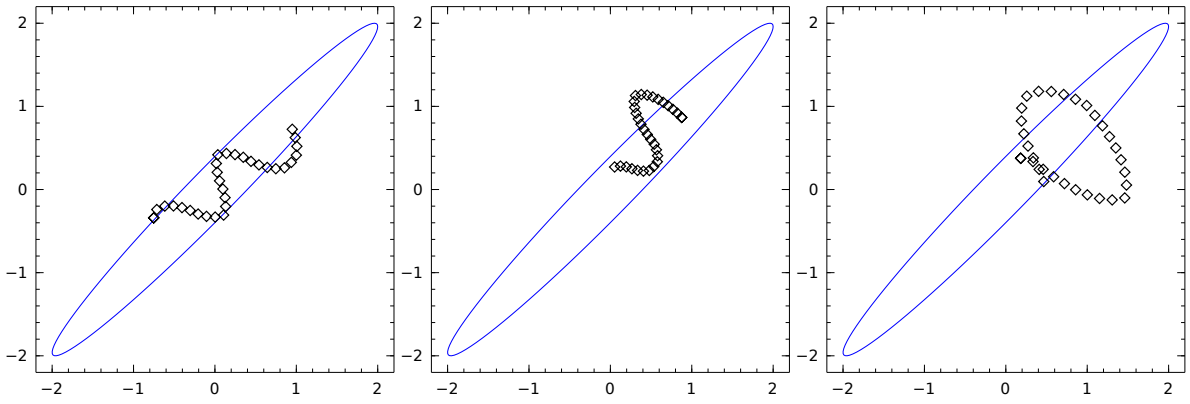

Figure 2. Example trajectories of refractive sampling whose final states were accepted. When the sampler repeatedly encounters gradients pointing in the opposite direction to p, it changes course until the angle of incidence to the gradient tangent plane is too shallow and p is reflected, giving a serpentine behavior reminiscent of HMC.

Algorithm 1 Pseudocode for Refractive Sampling

Input: x_0, $f(x)$, $g(x)$, m, w, r
$x \leftarrow x_0$
$p \sim \mathcal{N}(0, I)$
$\alpha \leftarrow 1$
for $i = 1 : m + 1$ **do**
  if $p^T g(x) > 0$ **then**
    $u \leftarrow \frac{g(x)}{||g(x)||}$
    $(r_1, r_2) \leftarrow (1, r)$
  else
    $u \leftarrow -\frac{g(x)}{||g(x)||}$
    $(r_1, r_2) \leftarrow (r, 1)$
  end if
  $\cos\theta_1 \leftarrow \frac{p^T u}{||p||}$
  $\cos^2\theta_2 \leftarrow 1 - \frac{r_1^2}{r_2^2}\left(1 - \cos^2\theta_1\right)$
  if $\cos^2\theta_2 < 0$ **then**
    $p \leftarrow p - 2(p^T u)u$
  else
    $p \leftarrow \frac{r_1}{r_2}p - ||p||\left[\frac{r_1}{r_2}\cos\theta_1 - \cos\theta_2\right]u$
    $\alpha \leftarrow \left(\frac{r_1}{r_2}\right)^{d-1}\frac{\cos\theta_1}{\cos\theta_2}\alpha$
  end if
  if $i \leq m$ **then**
    $x \leftarrow x + wp$
  end if
end for
$\alpha \leftarrow \frac{f(x)}{f(x_0)}\alpha$
$z \sim$ Uniform()
if $z < \alpha$ **then**
  return x
else
  return x_0
end if

distribution as an equal mixture of two Gaussians, with $\mu_1 = [1, 1]^T$ and $\mu_2 = [-1, -1]^T$. We take both covariance parameters to be Σ, with unit diagonal entries, and with the offdiagonal entries $\Sigma_{12} = \Sigma_{21}$ equal to some constant. As we increase Σ_{12}, the target distribution becomes a pair of parallel elliptical Gaussians. In order to measure how well a sampler mixes between the two modes, we count the number of times the sampler crosses the line $x_1 = -x_2$, a larger number of crossings indicating better mixing between the two modes. For Refractive Sampling we set $w = 0.5$, $m = 4$, and $r = 1.3$, and for HMC we set $\varepsilon = 0.5$ and $L = 4$. We performed four trials for each setting of Σ_{12}, reporting the mean and standard deviation estimates of the number of crossings and acceptance rates. As seen in Table 1, HMC and NUTS cross more often than Refractive sampling for small values of Σ_{12}, largely due to the higher acceptance rates. However, as Σ_{12} increases, the number of crossings for the HMC based algorithms degrades quickly, so that Refractive sampling crosses ten times more often when $\Sigma_{12} = 0.8$.

This demonstrates the fundamental difference between Refractive Sampling and HMC-related samplers: as the target distribution becomes more peaked, the gradient becomes more intense, and HMC has a more difficult time leaving the mode it is currently exploring. Refractive Sampling, on the other hand, is able to jump between peaked modes more freely.

4.2. Convergence Comparison

We compare refractive sampling to HMC and NUTS (Hoffman & Gelman, 2011) on how well they converge. For all HMC experiments, we chose ε so that the acceptance rate is larger than 0.5 but strictly less than 1.0. We used this ε as the starting value for the ε initialization scheme in NUTS, all other tuning hyperparameters

Table 1. Bimodal Mixing Example

Σ_{12}	0.0		0.5		0.8	
	Num. Cross	Accept Rate	Num. Cross	Accept Rate	Num. Cross	Accept Rate
Refractive	1002.5 ± 54.3	0.449 ± 0.003	760.5 ± 49.4	0.405 ± 0.006	527.0 ± 17.9	0.354 ± 0.003
HMC	2308.0 ± 48.1	0.977 ± 0.002	1163.5 ± 7.7	0.972 ± 0.003	64.3 ± 7.4	0.880 ± 0.004
NUTS	2229.0 ± 54.4	0.788 ± 0.008	804.5 ± 12.5	0.808 ± 0.008	44.5 ± 4.1	0.774 ± 0.003

for NUTS are as in (Hoffman & Gelman, 2011). We set $r = 1.3$ for all refractive sampling experiments, and set w so to give an acceptance rate larger than 0.05. We run each sampler for a burn-in period which we set as half the total number of iterations.

In the following experiments we report the log-posterior probability of the final states reached by the MCMC chains. These probabilities are representative of the probabilities found throughout the chain.

4.2.1. Gaussian Mixture Model

Gaussian Mixture Models (GMM), despite their apparent simplicity, can be difficult models for black box samplers. The covariance parameters are more flexible than the means, which manifests as modes in the posterior where a few components with large covariances explain the majority of the data, and the remaining clusters explain few points, if any. Specifically, we define our Bayesian GMM as:

$$
\begin{aligned}
w &\sim \text{Dirichlet}(\alpha) \\
\mu_k &\sim \mathcal{N}(0, \Sigma_\mu) \\
\Sigma_k &\sim \text{InvWishart}(\Sigma_0, \nu) \\
z_i &\sim \text{Categorical}(w) \\
X_i &\sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})
\end{aligned}
$$

We set $\alpha = \mathbf{2}$, $\nu = D + 2$, and $\Sigma_\mu = \Sigma_0 = I$. It is sensible to update the covariances and the means in turn, as well as to update each component's respective parameters in turn; we perform inference in this Gibbs style for all algorithms. Algorithms that have autotuning carry their tuned parameters between Gibbs iterations.

We compare refractive sampling to HMC and NUTS on inference of the means and variances[3] of a GMM on synthetic data. The data are generated from 3 Gaussians with dimension $D = 2$. All algorithms began with the same initial states for a given trial. After tuning, we set $w_\mu = 0.05$, $w_\Sigma = 0.005$, $m = 5$ for refractive sampling, and $\varepsilon_\mu = 0.02$, $\varepsilon_\Sigma = 0.005$, and $L = 5$ for HMC. We ran all samplers for 10000 iterations, with NUTS taking about twice as long as refractive sampling and HMC. Typically, all samplers found the same mode; see Figure 3. The log-posteriors of the final states for all algorithms and trials are summarized in Figure 4. In this experiment, log-posterior values fluctuated by about 5 points from iteration to iteration for all algorithms – refractive sampling and NUTS perform similarly, while HMC has not yet converged.

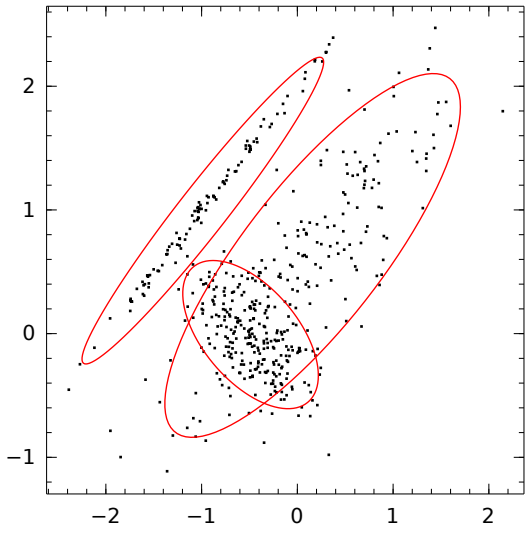

Figure 3. Typical refractive sampling MCMC state on the GMM synthetic data.

4.2.2. Binary Feature Linear Gaussian

We applied our sampler to a Binary Feature Linear Gaussian (BFLG) model[4]. The model with data $X \in \mathbb{R}^{N \times D}$ and parameters $Z \in \mathbb{Z}_1^{N \times K}$, $W \in \mathbb{R}^{K \times D}$ is specified as follows:

$$
\begin{aligned}
W_{k,d} &\sim \mathcal{N}(0, \sigma_W) \\
\beta_k &\sim \text{Beta}(a, b) \\
Z_{i,k} &\sim \text{Bernoulli}(\beta_k) \\
X_{i,d} &\sim \mathcal{N}((ZW)_{i,d}, \sigma_X)
\end{aligned}
$$

We apply this model to a subset of 500 of the "threes" digits of the MNIST (LeCun & Cortes, 1998) dataset. We first perform PCA on the 28×28 pixel images to give a $D = 80$

[3] For the mixing weights we used refractive sampling for all algorithms.

[4] this is a finite version of the linear-Gaussian model explored in (Griffiths & Ghahramani, 2005) and (Teh et al., 2007)

representation that captures 90% of the data variance. As sampling in this setting requires alternately sampling continuous and discrete variables[5], we expect there to be many modes in the posterior. We set $a = b = 2$, $\sigma_W = 10$, $\sigma_X = 1$, $K = 10$, and we integrate out β. We ran each algorithm for eight independent runs of 5000 iterations each.

For refractive sampling we set $w = 0.01$ and $m = 8$, and for HMC $\varepsilon = 0.15$ and $L = 8$, giving an acceptance rate of 0.9. All algorithms ran in similar times as the computational bottleneck is sampling Z. The results are summarized in Figure 4. As can be seen here, refractive sampling outperforms HMC and NUTS.

4.2.3. Bayesian Softmax Regression

Finally, we compare all algorithms on Bayesian Softmax Regression (also known as multinomial logistic regression). The model for data $X \in \mathbb{R}^{N\times D}$, $Y \in \mathbb{Z}_C^N$, and $C > 1$ is:

$$\begin{aligned} \beta_{c,j} &\sim \mathcal{N}(0, \sigma_\beta) \\ Y_i &\sim \text{Categorical}(P(Y_i|\beta, X_i)) \\ P(Y_i = c|\beta, X_i) &= \frac{\exp(\beta_c^T X_i)}{\sum_{c'=1}^{C} \exp(\beta_{c'}^T X_i)} \end{aligned}$$

where we set the parameters of the pivot class $\beta_C = \mathbf{0}$ as they are superfluous degrees of freedom. In this model there are dependencies between the β_cs, particularly in modelling the data of the pivot class which has no parameters of its own. Thus we expect that the gradient to only be a good local approximation as there is significant curvature in the posterior logpdf. We apply the model to the St. Jude Leukemia dataset (Yeoh et al., 2002), a data set with $N = 327$ and $D > 10000$. We preprocess the data using PCA to give $D = 140$, retaining about 90% of the data variance. The data are gene expression levels from 6 different diagnostic classes of leukemia, with a 7^{th} class denoting cases that were not assigned a diagnostic label. We treat each designation as its own class, including the 7^th "unlabeled" class, which we set as the pivot class.

We take $\sigma_\beta = 10$; $w = 0.01$, $m = 2$ for refractive sampling and $\varepsilon = 0.001$, and $L = 2$ for HMC, giving an acceptance rate of 0.55. The results are summarized in 4. In this case, refractive sampling outperforms both HMC based samplers.

4.3. Sample Efficiency

There is an inherent trade off between a sampler's ability to explore a mode quickly and its ability to escape that mode; a sampler that spends too much time attempting to find alternative modes will have inferior sample efficiency. As such, we should not expect Refractive Sampling to outperform HMC or NUTS on metrics such as Effective Sample Size (ESS) for unimodal posteriors. However, computing ESS in the above experiments would not have been informative; not all samplers had converged to the same mode or had converged at all.

Thus we compare sample efficiencies on Bayesian Logistic Regression applied to three benchmark datasets[6]. We mean-centered and whitened all datasets for evaluation. We compute ESS as estimated in (Hoffman & Gelman, 2011). In (Hoffman & Gelman, 2011), a 50,000 iteration run of NUTS is used to estimate the posterior mean and variance for each parameter, and for each algorithm being evaluated, the ESS for estimators of the mean and central second moment for each paramter is estimated, and the minimum is reported. We tuned Refractive Sampling and HMC manually, trying m and L in $\{1, 2, 4, 8\}$ with various stepsizes in order to maximize ESS per second in preliminary runs, still taking $r = 1.3$. We ran each algorithm for 10,000 iterations, discarding the first 5,000 as burn-in.

We repeated each evaluation for 8 trials and report the mean and standard deviations of the ESS and ESS per second in Table 2. On these problems Refractive Sampling is roughly 3-7 times less sample efficient than HMC and NUTS. This is not a prohibitively large difference. Perhaps the easiest way to maximize ESS while using Refractive Sampling is to mix it with another sampler such as HMC or NUTS.

5. Additional Remarks

There are possible improvements and variants of refractive sampling that have yet to be explored. One obvious direction is that any vector-valued function may be substituted for the gradient for use in the refraction transformation. Stochastic approximations of the gradient can be used while still providing a valid Markov Chain, and other schemes choosing directions other than that of steepest ascent may be fruitful. As presented, refractive sampling makes no use of the magnitude of the gradient. One possible improvement would be to allow the refractive index ratio r to depend on the gradient magnitude – when the gradient is larger it may be sensible to increase r. Many of the ideas in Riemannian geometry variants of HMC and MALA could be applied to refractive sampling, as well as the automated/incorporated choice of the number of steps and stepsizes in (Hoffman & Gelman, 2011) and (Wang et al., 2013).

[5] Technically, W can be integrated out analytically. We instead sample it for the purposes of this experiment.

[6] German Credit, Pima Indians, and Statlog Heart datasets, all available at the UCI Machine Learning Repository (Bache & Lichman, 2013)

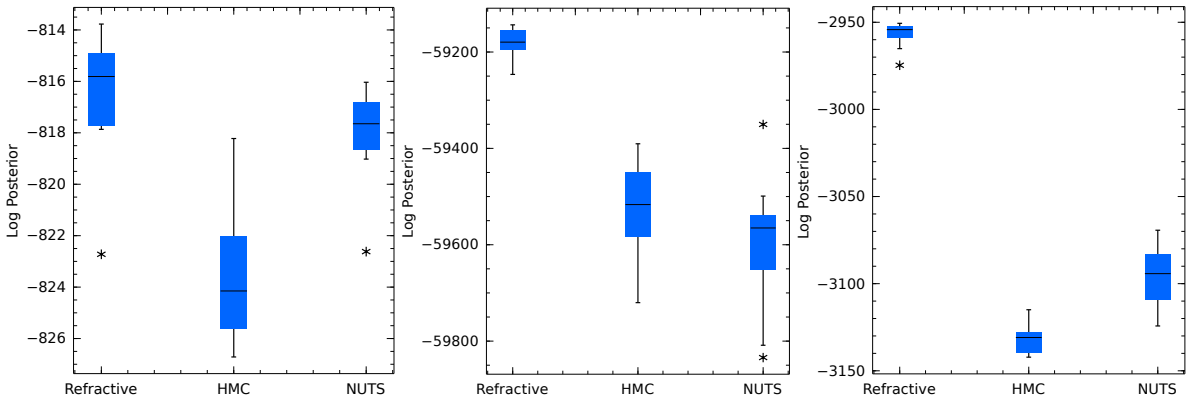

Figure 4. Boxplots depicting the posterior probabilities of the final states from 8 independent trials for each algorithm for the Gaussian Mixture Model (left), Binary Feature Linear Gaussian (middle), and Softmax Regression (right)

Table 2. Sample Efficiency

	German Credit		Pima		Heart	
	ESS	ESS/sec.	ESS	ESS/sec.	ESS	ESS/sec.
Refractive	175.8 ± 51.4	4.7 ± 2.0	445.3 ± 44.0	31.4 ± 2.0	92.6 ± 38.7	10.2 ± 4.5
HMC	1140.8 ± 167.8	34.1 ± 8.1	1603.4 ± 155.6	116.7 ± 13.6	359.4 ± 93.9	42.5 ± 10.9
NUTS	623.5 ± 85.8	13.6 ± 4.3	1474.4 ± 207.8	99.6 ± 14.0	912.8 ± 244.0	61.0 ± 16.5

6. Conclusions

We have introduced refractive sampling, a Metropolis Hastings sampler which uses the normalized gradient to guide its proposals. It constructs proposals based on basic physical processes and is thus easy to implement. Refractive sampling enjoys many of the benefits of other gradient-based samplers without the sensitivity to large fluctuations in gradients. Setting the refractive index ratio $r = 1.3$ was shown to work well in a large variety of problems. We have demonstrated in several experiments that refractive sampling is indeed easier to tune as simply requiring acceptance rates larger than 0.05 consistently gave better mode finding behavior than HMC tuned to have acceptance rates larger than 0.5. Refractive sampling also frequently outperformed NUTS in this manner, whose hyperparameters are all automatically tuned.

References

Bache, Kevin and Lichman, Moshe. UCI Machine Learning Repository, 2013. URL http://archive.ics.uci.edu/ml.

Duane, S, Kennedy, AD, Pendleton, BJ, and Roweth, D. Hybrid monte carlo. *Physics letters B*, 1987.

Girolami, M and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011.

Griffiths, T and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. 2005.

Hoffman, MD and Gelman, A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1111.4246*, 2011. URL http://arxiv.org/abs/1111.4246.

LeCun, Y and Cortes, C. The MNIST database of handwritten digits, 1998.

Neal, R. MCMC for Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, 2011.

Neal, Radford. Slice sampling. *Annals of statistics*, 2003.

Robbins, H and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

Teh, YW, Görür, D, and Ghahramani, Z. Stick-breaking construction for the Indian buffet process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

Wang, Ziyu, Mohamed, Shakir, and de Freitas, Nando. Adaptive Hamiltonian and Riemann Manifold Monte Carlo Samplers. *arXiv preprint arXiv:1302.6182*, pp. 10, February 2013. URL http://arxiv.org/abs/1302.6182.

Yeoh, E J, Ross, M E, Shurtleff, S A, Williams, W K, Patel, D, Mahfouz, R, Behm, F G, Raimondi, S C, Relling,

M V, Patel, A, and Others. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell*, 1(2):133–143, 2002.