Elliptical Slice Sampling with Expectation Propagation

Francois Fagan
Department of IEOR
Columbia University
New York, NY 10027
ff2316@columbia.edu

Jalaj Bhandari
Department of IEOR
Columbia University
New York, NY 10027
jb3618@columbia.edu

John P. Cunningham
Department of Statistics
Columbia University
New York, NY 10027
jpc2181@columbia.edu

Abstract

Markov Chain Monte Carlo techniques remain the gold standard for approximate Bayesian inference, but their practical issues — including onerous runtime and sensitivity to tuning parameters — often lead researchers to use faster but typically less accurate deterministic approximations. Here we couple the fast but biased deterministic approximation offered by expectation propagation with elliptical slice sampling, a state-of-the-art MCMC method. We extend our hybrid deterministic-MCMC method to include recycled samples and analytical slices, and we rigorously prove the validity of each enhancement. Taken together, we show that these advances provide an order of magnitude gain in efficiency beyond existing state-of-the-art sampling techniques in Bayesian classification and multivariate gaussian quadrature problems.

1 INTRODUCTION

Exact posterior inference in Bayesian models is rarely tractable, a fact which has prompted vast amounts of research into efficient approximate inference techniques. Deterministic methods such as the Laplace approximation, Variational Bayes, and Expectation Propagation offer fast and analytical posterior approximations, but introduce potentially significant bias due to their restricted form which can not capture important characteristics of the true posterior. Markov Chain Monte Carlo (MCMC) methods represent the target posterior with samples, which while asymptotically exact, can be slow, require substantial tuning, and perform poorly when variables are highly correlated.

Conceptually, these two techniques can be combined to great benefit: if a deterministic approximation can cover the true posterior mass accurately, then a subsequent MCMC sampler should be much faster and be less susceptible to inefficiency due to correlation (as the deterministic approximation would have captured this correlation). To do so, however, is practically quite difficult. First, both the Laplace and Variational Bayesian approximations fit local mass of a posterior (in Variational Bayes this is sometimes called the *exclusive* property of optimizing the Kullback-Liebler divergence from the approximation to the true posterior [Minka, 2005]). While excellent in many situations, this property is inappropriate for initializing an MCMC sampler, since it will be very difficult for that sampler to explore other areas of posterior mass (e.g., other modes). Expectation Propagation (EP, [Minka, 2001]), on the other hand, is typically derived as an inclusive approximation that, at least approximately, attempts to match the global sufficient statistics of the true posterior (most often the first and second moments, producing a Gaussian approximation). Such a choice is superior for an MCMC sampler.

Secondly, we require a sensible choice of MCMC sampler so as to leverage a deterministic approximation like EP. Given an unnormalized target distribution $p^*(\mathbf{x})$, we can write:

$$p^*(\mathbf{x}) \;=\; \hat{p}(\mathbf{x})\frac{p^*(\mathbf{x})}{\hat{p}(\mathbf{x})} \;\equiv\; \hat{p}(\mathbf{x})\hat{\mathcal{L}}(\mathbf{x}),$$

which allows us to treat the true posterior as the product of an effective prior $\hat{p}$ and likelihood $\hat{\mathcal{L}}$. We then have freedom to choose $\hat{p}$, which we will set to be the deterministic (Gaussian) posterior approximation from EP. Amongst all MCMC methods, Elliptical Slice Sampling (ESS, [Murray et al., 2010]) handles the above reformulations seamlessly. ESS has become an important and generic method for posterior inference with models that have a strong Gaussian prior. It inherits the attractive properties of slice sampling generally [Neal, 2003], and notably lacks tuning parameters that are often highly bur-

densome in other state-of-the-art methods like Hamiltonian Monte Carlo (HMC; Neal [2011]). The critical observation is that, if EP provides a quality posterior approximation $\hat{p} \approx p^*$, the likelihood term $\hat{\mathcal{L}}$ will typically be weak, which puts ESS in the regime where it is most efficient.

What results is a new MCMC sampler that combines EP and ESS, is faster than state-of-the-art samplers like HMC, and is able to explore the parameter space efficiently even in the presence of strong dependency among variables. Specifically, our contributions include:

1. In Section 2, we propose *Expectation Propagation based Elliptical Slice Sampling (EPESS)* where we justify the use of EP as the "prior" for ESS.

2. In Section 3, we investigate a method to improve the overall run time of ESS by sampling multiple points each iteration. It reduces the average number of shrinkage steps giving it a computational advantage. We call it *Recycled ESS* and integrate it with EPESS to further increase its efficiency.

3. We extend our method to *Analytic Elliptical Slice Sampling* in Section 4. As the name suggests, we can analytically find the region corresponding to a slice and sample uniformly from it. In addition to decorrelating samples, it offers the computational advantage of avoiding expensive shrinkage steps. It is applicable to only a few target distributions and we illustrate it, in the context of EPESS, for linear Truncated Multivariate Gaussian (TMG) quadrature.

4. We offer empirical evaluation of EPESS (Section 5), which show an order of magnitude improvement over the state-of-the art MCMC methods for TMG and probit models.

2 EXPECTATION PROPAGATION AND ELLIPTICAL SLICE SAMPLING

In this section we introduce our combined EP and ESS sampling method. We begin with background of the two building blocks of this method, to place them in context of current literature.

2.1 ELLIPTICAL SLICE SAMPLING

There are many problems where dependency between latent variables is induced through a Gaussian prior, for example in Gaussian Processes. Elliptical Slice Sampling (ESS, [Murray et al., 2010]) is specifically designed for efficiently sampling from such distributions and is considered state-of-the-art on these problems. ESS considers posteriors of the form

$$p^*(\mathbf{x}) = \frac{1}{Z}\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma})\mathcal{L}(\mathbf{x}) \tag{1}$$

where $\mathcal{L}$ is a likelihood function, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a multivariate Gaussian prior and Z is the normalizing constant.

ESS is a variant of slice sampling [Neal, 2003] that takes advantage of the Gaussian prior to improve mixing time and eliminate parameter tuning. At the beginning of each iteration of ESS two random variables are sampled. The first is the slice height y which is uniformly distributed over $[0, \mathcal{L}(\mathbf{x})]$, where $\mathbf{x}$ is the current sample. The second variable $\boldsymbol{\nu}$ is sampled from the prior $\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma})$ and, together with the current sample $\mathbf{x}$, defines an ellipse:

$$\mathbf{x}'(\theta) = \mathbf{x}\cos(\theta) + \boldsymbol{\nu}\sin(\theta). \tag{2}$$

Next, a one-dimensional angle bracket $[\theta_{\min}, \theta_{\max}]$ of length 2π is proposed containing the point $\theta = 0$ (corresponding to the current point $\mathbf{x}$). The bracket is then shrunk toward $\theta = 0$ until a point is found within the bracket that satisfies $\mathcal{L}(\mathbf{x}'(\theta)) > y$. This point is accepted as the next point in the Markov chain.

ESS is known to work well when the prior aligns with the posterior and the likelihood is weak [Murray et al., 2010, Section 2.5]. However, when this is not the case then ESS can perform poorly, as we demonstrate below.

Figure 1 illustrates the problem when the prior and the posterior do not align: here we have a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior with an observed Bernoulli likelihood $\mathcal{L}(\mathbf{x}) = \mathbb{1}(\mathbf{x} \in A)$ for some rectangle A. The posterior is a truncated Gaussian within A. In this example we have placed A away from the origin, with the result that that most of the posterior density lies vertically on the left boundary of the box. Accordingly, a good sampler should be able to make large vertical moves to effectively explore the posterior mass.

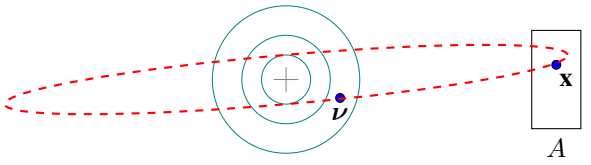

Figure 1: ESS ellipse shown in dashed red.

As the likelihood rectangle A moves further right, the posterior moves away from the prior. As a result most of the points proposed on the ellipse will not lie in A, so more shrinkage steps will be necessary until a point is accepted, leading to an inefficient algorithm. Moving A

further to the right also makes the ellipse more eccentric which prevents vertical movement, resulting in further inefficiency.

The other pathology afflicting ESS is that of strong likelihoods. This happens when $\mathcal{L}(\mathbf{x})$ is extremely large in regions of non-negligible posterior density. Once the sampler is in such a region, only with low probability will it be able to accept points proposed outside the region, hence it will get stuck. This will occur, for instance, when the prior underestimates the variance of the posterior and $\mathcal{L}(\mathbf{x})$ becomes large in the tails. We refer the reader to an extended explanation of this effect in [Nishihara et al., 2014]. Indeed, this motivates our choice of EP as a prior, since Variational Bayes and Laplace approximations are known to often underestimate posterior variance whereas EP does not [Minka, 2005].

We address both these problems by choosing an EP prior (Section 2.2) for ESS. How to incorporate EP into ESS is explained in Section 2.3.

2.2 EXPECTATION PROPAGATION

Expectation Propagation (EP) is a method for finding a Gaussian approximation q to a given distribution p^* by iteratively matching local moments and then updating the global approximation via a so-called 'tilted' distribution [Minka, 2001]. At termination the distribution q will optimize a global objective that approximates the Kullback-Liebler divergence $KL(p^*||q)$ [Wainwright and Jordan, 2008]. The resulting Gaussian approximation is an *inclusive* estimate of p^* that approximately matches its zeroth, first, and second moments.

Although EP has few theoretical guarantees [Dehaene and Barthelmé, 2015], it is known to be accurate for many models including truncated multivariate gaussian [Cunningham et al., 2011], probit and logistic regression [Nickisch and Rasmussen, 2008], log-Gaussian Cox processes [Ko and Seeger, 2015], and more [Minka, 2001]. It is also known to have superior performance compared to the Laplace approximation and Variational Bayes in terms of approximating marginal distributions accurately [Kuss and Rasmussen, 2005, Cseke and Heskes, 2011, Deisenroth and Mohamed, 2012].

2.3 ELLIPTICAL SLICE SAMPLING WITH EXPECTATION PROPAGATION

As outlined in Section 1, we incorporate a posterior approximation $\hat{p}$ as a proposal distribution for ESS. We do so by defining:

$$p^*(\mathbf{x}) \;=\; \hat{p}(\mathbf{x})\frac{p^*(\mathbf{x})}{\hat{p}(\mathbf{x})} \;=\; \hat{p}(\mathbf{x})\hat{\mathcal{L}}(\mathbf{x}) \qquad (3)$$

where p^* is the posterior distribution of interest from Equation (1), $\hat{p}$ is our new prior and $\hat{\mathcal{L}}$ is our new likelihood. As explained in Section 2.1, for ESS to work well, $\hat{p}$ should have two desirable properties: *(i)* It should approximate the posterior p^*. The most obvious candidates for $\hat{p}$ includes Laplace, Variational Bayes and EP approximations, *(ii)* It should ensure that the new likelihood $\hat{\mathcal{L}} = p^*/\hat{p}$ is weak, in the sense as described in Section 2.1. Using either Laplace or Variational Bayes may result in large values of $\hat{\mathcal{L}}$ in the tails due to variance underestimation, which could cause the sampler to get stuck. The more inclusive nature of the EP estimate, on the other hand, makes it a sensible choice to obtain a Gaussian posterior approximation $\hat{p}$.

To demonstrate the power of this approach we return to the problematic example given in Figure 1. Using the EP approximation we can shift our prior to align with the posterior density on the left side of the likelihood rectange A. The ellipses become short and vertical, allowing ESS to mix efficiently. This is illustrated in Figure 2. To demonstrate the difference in the sampling behavior between EPESS and ESS, Figure 2.3 plots 400 samples from both EPESS and ESS. EPESS is clearly superior and manages to explore the entire distribution whereas ESS moves consistently less.

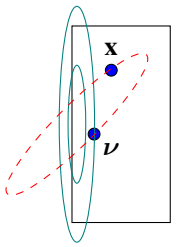

Figure 2: The EP approximation is in teal and an EPESS elliptical slice is in dashed red.

The idea of Equation (3) is not unique to this paper. Nishihara et al. [2014] use a similar construction where the Gaussian approximation is learned from samples. Although this has the advantage of not relying on EP to do moment matching, it requires parallelism and expensive moment calculations. EPESS will be simpler and more efficient when an accurate EP approximation is available. Braun and Bonfrer [2011] also have a similar method where they use the Laplace approximation, which as discussed, is a poor choice. We remark that using Power EP approximations is also a viable choice for a prior, a point that we will return to in Section 6.

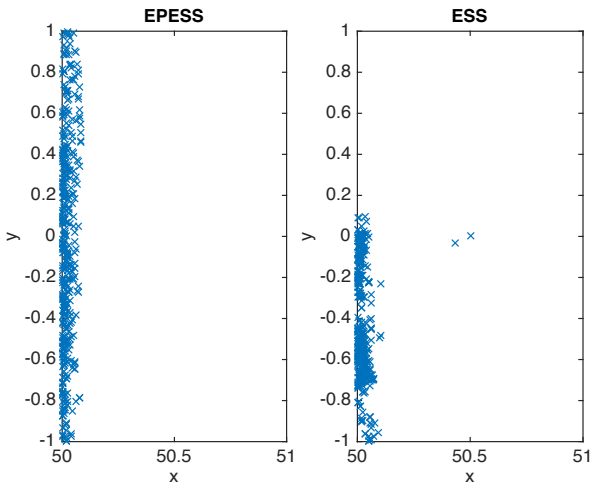

Figure 3: EPESS vs ESS: 400 samples of EPESS and ESS for a 2-d Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ truncated in a rectangular box $\{50 \leq x \leq 51, -1 \leq y \leq 1\}$. EPESS explores the parameter space effectively whereas ESS does not.

Algorithm 1: `Recycled_ESS`

Input : Log-likelihood function $(\log \mathcal{L})$, initial point $\mathbf{x}_1^{(1)} \in \mathbb{R}^d$, prior $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, number of iterations N, number of recycled points J

Output: Samples from Markov Chain $((\mathbf{x}_1^{(1)}, ..., \mathbf{x}_J^{(1)}), ..., (\mathbf{x}_1^{(N)}, ..., \mathbf{x}_J^{(N)}))$

for $i = 1$ **to** N **do**
  $u \sim$ Uniform $[0, 1]$
  $\log y \leftarrow \log \mathcal{L}(\mathbf{x}_1^{(i-1)}) + \log u$
  $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
  $\theta_{max} \sim$ Uniform $[0, 2\pi]$
  $\theta_{min} \leftarrow \theta_{max} - 2\pi$
  for $j = 1$ **to** J **do**
    $(\hat{\mathbf{x}}_j^{(i)}, \theta_{min}, \theta_{max}) \leftarrow$ `ESS_inner_loop`
      $(\log \mathcal{L}, \log y, \boldsymbol{\nu}, \mathbf{x}_1^{(i-1)}, \theta_{min}, \theta_{max})$
  end
  $(\mathbf{x}_1^{(i)}, ..., \mathbf{x}_J^{(i)}) \leftarrow$ `rand_perm`$(\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)})$
end
return $((\mathbf{x}_1^{(1)}, ..., \mathbf{x}_J^{(1)}), ..., (\mathbf{x}_1^{(N)}, ..., \mathbf{x}_J^{(N)}))$

3 RECYCLED ELLIPTICAL SLICE SAMPLING

In this section we show how to sample $J > 1$ points at every ESS iteration without a significant increase in computational complexity. This idea is inspired by the work of Nishimura and Dunson [2015] on HMC. In that work an HMC algorithm is devised which "recycles" the intermediate points as valid samples from the target distribution. We borrow the phrase "recycling" from them and call our method *Recycled Elliptical Slice Sampling*.

Recall that in every ESS iteration, we propose points along an ellipse within an angle bracket, which is iteratively shrunk, until a point is accepted. In Recycled ESS, we don't stop after accepting the first point but continue to propose points starting from the last angle bracket used. This procedure is continued until J points are accepted. One of the J points is randomly selected to propagate the Markov chain.

As we shrink the angle bracket $[\theta_{\min}, \theta_{\max}]$ towards $\theta = 0$ (corresponding to the current point), the probability of the next proposal point being accepted tends to increase. Hence the number of shrinkage steps required to accept latter points is typically smaller than that for first accepted point. Since the number of likelihood function evaluations is proportional to the number of shrinkage steps, Recycled ESS is able to sample more points with only a small increase in computational complexity, leading to improved run times per sample. This approach is formalized in Algorithm 1, where `ESS_inner_loop` function is the regular ESS inner loop and can be found in Figure 2 of [Murray et al., 2010].

It is clear from Algorithm 1 that we treat each sample $\mathbf{x}_j^{(i)}$ as an element in a large Markov chain with state space $(\mathbf{x}_1^{(i)}, ..., \mathbf{x}_J^{(i)})$. We prove in Theorem 3.2 that each element $\mathbf{x}_j^{(i)}$ has its stationary marginal distribution as p^*. In order to do so, we first show in Lemma 3.1 that the transition operator of accepting the j^{th} point is reversible.

Lemma 3.1. *Let T_j correspond to the transition operator from $\mathbf{x}_1^{(i-1)} \rightarrow \hat{\mathbf{x}}_j^{(i)}$. Then T_j is invariant to p^*.*

A detailed proof is given in the appendix. Theorem 3.2 easily follows:

Theorem 3.2. *Each element in the Recycled ESS Markov chain has marginal stationary distribution p^*.*

Proof. The sequence of points $\{\mathbf{x}_1^{(i)}\}$ follow a Markov Chain. At each step the transition operator is uniformly sampled from the set $\{T_j : j = 1, ..., J\}$, with each T_j being invariant to p^* (Lemma 3.1). Therefore we have that $\mathbf{x}_1^{(i)} \xrightarrow{dist.} \mathbf{x}^*$ where $\mathbf{x}^* \sim p^*$. Also, at any fixed iteration i, we have that all points in $\{\mathbf{x}_j^{(i)} : j = 1, ..., J\}$ are identically distributed. This follows from the random permutations:

$$\begin{aligned} p(\mathbf{x}_j^{(i)} | (\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)})) &= \text{Uniform}(\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)}) \\ &= p(\mathbf{x}_k^{(i)} | (\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)})). \end{aligned}$$

Integrating over $p(\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)})$ gives us that $p(\mathbf{x}_j^{(i)}) =$

$p(\mathbf{x}_1^{(i)})$ for any i, j. Since we have that $\mathbf{x}_1^{(i)} \xrightarrow{dist.} \mathbf{x}^*$, it follows that for all j: $\mathbf{x}_j^{(i)} \xrightarrow{dist.} \mathbf{x}^*$. □

The downside of Recycled ESS is that the latter accepted points (corresponding to $j \approx J$) are sampled from a very small angle bracket and so are highly correlated. On the other hand these points only require a small number of function evaluations. Overall the effect of recycling is a small increase in the effective number of samples, with a small increase in computational complexity. Whether or not this is beneficial is investigated empirically in Section 5.

4 ANALYTIC ELLIPTICAL SLICE SAMPLING

Consider the ellipse

$$\mathcal{E} = \{\mathbf{x}' : \mathbf{x}'(\theta) = \mathbf{x}\cos(\theta) + \boldsymbol{\nu}\sin(\theta)\}$$

as in an ESS iteration as defined by Equation (2). Let $\mathcal{S}(y; \mathcal{E})$ be the slice corresponding to the acceptable points in $\mathcal{E}$ for a given slice height y:

$$\mathcal{S}(y; \mathcal{E}) = \{\mathbf{x}' \in \mathcal{E} : \mathcal{L}(\mathbf{x}') > y\}.$$

If we can analytically characterize $\mathcal{S}(y; \mathcal{E})$ then we only need to sample a point uniformly from the slice to propagate the Markov Chain [Neal, 2003]. This has three advantages: *(i)* We eliminate expensive slice shrinkage steps which reduces the computational cost of our sampler; *(ii)* In standard slice sampling algorithms, shrinkage steps bias the next sample to be close to the current sample thereby introducing correlations. Since we uniformly sample over $\mathcal{S}(y; \mathcal{E})$, the resulting samples are less correlated as we are not biased towards the current point; *(iii)* We can easily incorporate the recycling idea here resulting in an extremely efficient algorithm, which we refer to as *Analytic Elliptical Slice Sampling*.

As in Recycled ESS, in Analytic ESS we sample $J > 1$ points from each ellipse $\mathcal{E}$. We first sample J different y values, which are evenly spaced in a Quasi Monte-Carlo way. Corresponding to each y value, we analytically solve for $\mathcal{S}(y; \mathcal{E})$ (which has only a small amortized computational cost). One point is then uniformly sampled from each slice $\mathcal{S}(y; \mathcal{E})$. The pseudocode for Analytic ESS is given in Algorithm 2 and in Theorem 4.1 we prove its validity.

Theorem 4.1. *Each element in the Analytic ESS Markov chain has marginal stationary distribution p^*.*

Proof. The proof follows exactly the same argument as in Theorem 3.2. □

Algorithm 2: `Analytic_Slice_Sampling`

Input : Likelihood $\hat{\mathcal{L}}$, prior $\hat{p}$, initial point $\mathbf{x}_1^{(0)}$, subroutine `Sample_Ellipse` to sample an ellipse, subroutine `Characterize_Slice` to analytically characterize $\mathcal{S}(\cdot; \mathcal{E})$, number of iterations N, number of slices per iteration J

Output: Samples from Markov Chain $((\mathbf{x}_1^{(1)}, ..., \mathbf{x}_J^{(1)}), ..., (\mathbf{x}_1^{(N)}, ..., \mathbf{x}_J^{(N)}))$

for $i = 1$ **to** N **do**
  $\mathcal{E} \leftarrow$ `Sample_Ellipse`$(\mathbf{x}_1^{(i-1)}, \hat{p})$
  $\mathcal{S}(\cdot; \mathcal{E}) \leftarrow$`Characterize_Slice`$(\mathcal{E})$
  $u \sim$ Uniform $[0, 1]$
  for $j = 1$ **to** J **do**
    $y \leftarrow (j - u)/J \cdot \hat{\mathcal{L}}(\mathbf{x}_1^{(i-1)})$
    $\mathbf{x}_j^{(i)} \leftarrow$ Uniform $\{\mathbf{x} : \mathbf{x} \in \mathcal{S}(y; \mathcal{E})\}$
  end
  $(\mathbf{x}_1^{(i)}, ..., \mathbf{x}_J^{(i)}) \leftarrow$ `rand_perm`$(\hat{\mathbf{x}}_1^{(i)}, ..., \hat{\mathbf{x}}_J^{(i)})$
end
return $((\mathbf{x}_1^{(1)}, ..., \mathbf{x}_J^{(1)}), ..., (\mathbf{x}_1^{(N)}, ..., \mathbf{x}_J^{(N)}))$

Unfortunately solving for $\mathcal{S}(y; \mathcal{E})$ in closed form is not possible in general, although it can be done for Truncated Multivariate Gaussian (TMG) quadrature as shown below.

4.1 ANALYTIC EPESS FOR TMG

The (linear) TMG distribution is defined as:

$$p^*(\mathbf{x}) = \frac{1}{Z}\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}) \prod_{j=1}^{M} \mathbb{1}(L_j^\top \mathbf{x} \geq 0).$$

Using the EP approximation $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and Equation (3), we can rewrite the density p^* as:

$$\begin{aligned}
p^*(\mathbf{x}) &\propto \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}) \prod_{j=1}^{m} \mathbb{1}(L_j^\top \mathbf{x} \geq 0) \\
&\propto \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I})}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \prod_{j=1}^{m} \mathbb{1}(L_j^\top \mathbf{x} \geq 0) \\
&= \mathcal{N}(\mathbf{z}; \mathbf{0}, \boldsymbol{\Sigma}) \frac{\mathcal{N}(\mathbf{z}; -\boldsymbol{\mu}, \mathbf{I})}{\mathcal{N}(\mathbf{z}; \mathbf{0}, \boldsymbol{\Sigma})} \prod_{j=1}^{m} \mathbb{1}(L_j^\top (\mathbf{z} + \mu) \geq 0) \\
&\equiv \mathcal{N}(\mathbf{z}; \mathbf{0}, \boldsymbol{\Sigma}) \hat{\mathcal{L}}(\mathbf{z}) \\
&\equiv \tilde{p}(\mathbf{z})
\end{aligned}$$

where $\mathbf{x} = \mathbf{z} + \boldsymbol{\mu}$ is a transformation with identity Jacobian. We are able to apply Analytic ESS to $\tilde{p}(\mathbf{z})$ and can then recover samples for $\mathbf{x}$ by reversing the transforma-

tion. First we analytically characterize the slice:

$$\begin{aligned}\mathcal{S}(y;\mathcal{E}) &= \{\theta \in [0,2\pi) : \mathcal{L}(\mathbf{z}'(\theta)) > y\} \\ &= \cap_{j=1}^{m}\{\theta \in [0,2\pi) : L_j^\top \mathbf{z}'(\theta) + L_j^\top \mu \geq 0\} \\ &\quad \cap \{\theta \in [0,2\pi) : \frac{\mathcal{N}(\mathbf{z}'(\theta); -\boldsymbol{\mu}, \mathbf{I})}{\mathcal{N}(\mathbf{z}'(\theta); \mathbf{0}, \boldsymbol{\Sigma})} > y\} \\ &\equiv \cap_{j=1}^{m}\Theta_j \cap \Theta_y,\end{aligned}$$

where $\mathbf{z}'(\theta) = \mathbf{z}\cos(\theta) + \boldsymbol{\nu}\sin(\theta)$. The region Θ_j is the part of the ellipse that lies in the halfspace defined by L_j. Since it is defined by a linear inequality of $\sin(\theta)$ and $\cos(\theta)$, it is easily characterized using basic trigonometry. The resulting region may be rewritten as $\Theta_j = [0, l_j] \cup [u_j, 2\pi]$ for some $l_j, u_j \in (0, 2\pi)$. Due to this nice structure, taking the intersection of all m regions can be computed in $\mathcal{O}(m)$. Region Θ_y can be simplified by taking logarithms on both sides of its inequality and reduces to the form:

$$a_0 + a_1\cos\theta + a_2\sin\theta + a_3\cos\theta\sin\theta + a_4\cos^2\theta > 0.$$

The roots of this inequality can be obtained by solving a quartic equation. This completes our analytic characterization of $\mathcal{S}(y;\mathcal{E})$.

The most expensive operations in Analytic ESS are generating $\mathcal{E}$ and characterizing $\mathcal{S}(y;\mathcal{E})$, as sampling from $\mathcal{S}(y;\mathcal{E})$ is relatively cheap. To see this, let d denote the dimension of $\mathbf{x}$ and m the number of linear truncations. To sample the ellipse $\mathcal{E}$ we must draw $\boldsymbol{\nu}$ from the Gaussian prior which costs $\mathcal{O}(d^2)$. Characterizing the slice $\mathcal{S}(y;\mathcal{E})$ involves m inner products of $\mathbf{x}$ and $\boldsymbol{\nu}$ with the linear truncations L_j and can be calculated in $\mathcal{O}(md)$. The total computational overhead is thus $\mathcal{O}(d^2 + md)$. Once this is done, sampling from $\mathcal{S}(y;\mathcal{E})$ is cheap. It involves sampling from an intersection of $\mathcal{O}(m)$ intervals which can be done in $\mathcal{O}(d + m)$ (here we have factored in the expense of storing each sample at a cost $\mathcal{O}(d)$).

Since the upfront cost of characterizing the slice $\mathcal{S}(y;\mathcal{E})$ is $\mathcal{O}(d^2 + md)$, we can sample $\mathcal{O}(d)$ points per ellipse $\mathcal{E}$ without significantly increasing the total computational complexity. This leads to a high effective sample size relative to the computational complexity.

The Exact-HMC algorithm for TMG [Pakman and Paninski, 2014] has an intimate relationship with Analytic ESS and inspired our analytic framework. We explain this connection in Section 5.2.

5 EXPERIMENTAL RESULTS

In this section we compare the empirical performance of the algorithms introduced in Sections 2.3–4 to other state-of-the-art MCMC methods. Comparisons are shown for the Probit regression and the TMG problem, both of which are often encountered in machine learning contexts, as well as the log-Gaussian Cox process. We quantify the mixing of MCMC samplers by comparing their effective number of samples. Effective sample size is estimated using the method as described in [Gelman et al., 2014] which is implemented in the MCMC Diagnostics Tool box for `Matlab` [Särkkä and Vehtari, 2014-02-34]. We compare the results in terms of effective sample size divided by the number of density function evaluations of p^*, which is the dominant computational expense of running the samplers.

5.1 PROBIT REGRESSION

Probit regression is one of the most common problems in machine learning and is often used as a benchmark for comparing Bayesian computation methods. A nice review of the state-of-the-art algorithms for probit can be found in [Chopin and Ridgway, 2015]. For our experiments, we choose 4 data sets of moderate size from UCI repository as listed in Table 1, with their dimension and number of datapoints. These are the Breast Cancer [Wolberg et al., 1995], Ionosphere [Sigillito, 1989], Sonar [Son] and Musk [AI Group at Arris Pharmaceutical Corporation, 1994] data sets. As is standard, each dataset is preprocessed to have zero mean and unit variance for each regressor, a unit intercept term has been included and the prior on each latent variable is $\mathcal{N}(0, 10)$.

Table 1: Datasets for probit: dimensions and number of data points.

DATASET	DIMENSION	DATA POINTS
Breast Cancer	31	569
Ionosphere	31	351
Sonar	61	97
Musk	165	419

We compare EPESS and Recycled EPESS (denoted by EPESS(J) where J is the number of recycled points per slice) against Metropolis-Hastings with an EP proposal (EPMH) and HMC using the No-U-Turn sampler as implemented in Stan [Carpenter et al., 2015]. EPMH is considered as state-of-the-art for Probit [Chopin and Ridgway, 2015] . The chains were initialized at the EP mean for all EPESS methods and the Stan implementation decides on its own initialization. We use the `R` package of Ridgway [2016] to find the EP approximation. Its CPU time is negligible compared to the time to run the samplers.

We run 100 chains with 20,000 samples per chain. For

EPMH we ran it until 20,000 unique samples (i.e., accepted proposed points) were collected to make it comparable with the other methods. The results are shown in the top three plots of Figure 4. EPESS outperforms HMC and EPMH by about a factor of 5 for effective sample size relative to number of function evaluations. As compared to EPMH, EPESS gives a slightly smaller effective number of samples but takes far fewer function evaluations. The number of function evaluations for Recycled EPESS is smaller than that of EPESS, however the effective number of samples are also proportionately small as the samples are highly correlated (this is expected: see Section 3). Overall Recycled EPESS does not improve the effective sample size relative to number of function evaluations above EPESS.

5.2 TRUNCATED MULTIVARIATE GAUSSIAN

The Truncated Multivariate Gaussian (TMG) is an important distribution which commonly arise in diverse models such as Probit/Tobit models, Neural models [Pillow et al., 2003], Bayesian bridge model in finance [Polson et al., 2014], True-skill model for competitions [Herbrich et al., 2006] and many others. There has been some recent work including [Lan and Shahbaba, 2015] focusing on sampling from TMG, but Exact-HMC algorithm [Pakman and Paninski, 2014] is considered to be state-of-the-art (see [Altmann et al., 2014] for a nice review). We treat it as the benchmark for comparisons in our experiments.

The equations of motion for Exact-HMC are the same as that of standard ESS,

$$\mathbf{x}'(t) = \mathbf{x}\cos(t) + \boldsymbol{\nu}\sin(t),$$

and so it suffers from the same problems as described in Section 2.1 and illustrated in Figure 1. The elliptical path enables *exact* calculation of where the HMC particle hits the truncations, hence the term *Exact*-HMC. These are the same calculations used to find the regions Θ_j in Analytic ESS, although being an HMC method, it does not have a slice height y with the corresponding region Θ_y. It also cannot incorporate an EP prior as this would destroy the elliptical path and render the calculations intractable. A tuning parameter T is required and for all our experiments we have fixed T to be $\pi/2$ as recommended by Pakman and Paninski [2014]. We only show comparative results for Analytic EPESS as it is faster than EPESS since it avoids slice shrinkages. We obtain an EP approximation for TMG using the method as described in [Cunningham et al., 2011] which is fast and scales well for high dimensions. It runs in negligible CPU time as compared to the running time of different sampling algorithms.

We run 100 chains with 20,000 samples per chain. The fourth plot in Figure 4 shows the results for a 2-d standard Gaussian where the truncated region is a rectangular box: $\{s \leq x \leq s+1, -1 \leq y \leq 1\}$. As we shift the box to the right, we see Analytic EPESS outperforming Exact-HMC by orders of magnitude. This trend carries over to higher dimensions as shown in the fifth plot in Figure 4. Results for $d = 500$ and $d = 1000$ have been omitted as Exact-HMC with $T = \pi/2$ takes prohibitively long to run.

5.3 LOG-GAUSSIAN COX PROCESS

We conducted experiments on a Log-Gaussian Cox Process (LGCP) applied to the coal mining disaster dataset as set up in the original ESS paper [Murray et al., 2010]. Although a convergent EP is available for the LGCP, it is not accurate with the EP mean substantially deviating from the true mean [Ko and Seeger, 2015]. Our experiments showed that EPESS fared no better than ESS on this problem, with the effective number of samples being about the same. This demonstrates the fact that EPESS will only perform well when EP is accurate.

6 DISCUSSION AND CONCLUSION

In this work we have shown how the ideas of ESS, EP and recycling can be combined to yield highly efficient MCMC methods. For both probit regression and Gaussian quadrature, performance exceeds state-of-the-art samplers by an order of magnitude. In the case of TMG, this can be multiple orders of magnitude.

We investigated two different types of recycling: sampling multiple points per slice (Recycled ESS), and sampling multiple points at different slice heights from the same ellipse (Analytic ESS). The benefit of Recycled ESS is questionable as it seems not to improve performance in probit, due to having highly correlated samples. It also introduces a tuning parameter which makes the algorithm more difficult to implement. Analytic EPESS for TMG does not have the above-mentioned issues of Recycled ESS. In this case recycling is of clear benefit as can be seen in the experimental results of Section 5.2. It is here where EPESS outperforms the state-of-the-art by the largest margin.

The example of the Log-Gaussian Cox process shows that EPESS will only offer an advantage over ESS when EP is accurate. This restricts the applicability of EPESS as a general method. Improving the accuracy of EP is a subject of active research and any developments made will be immediately be inherited by EPESS.

There are multiple directions of future work. Instead of

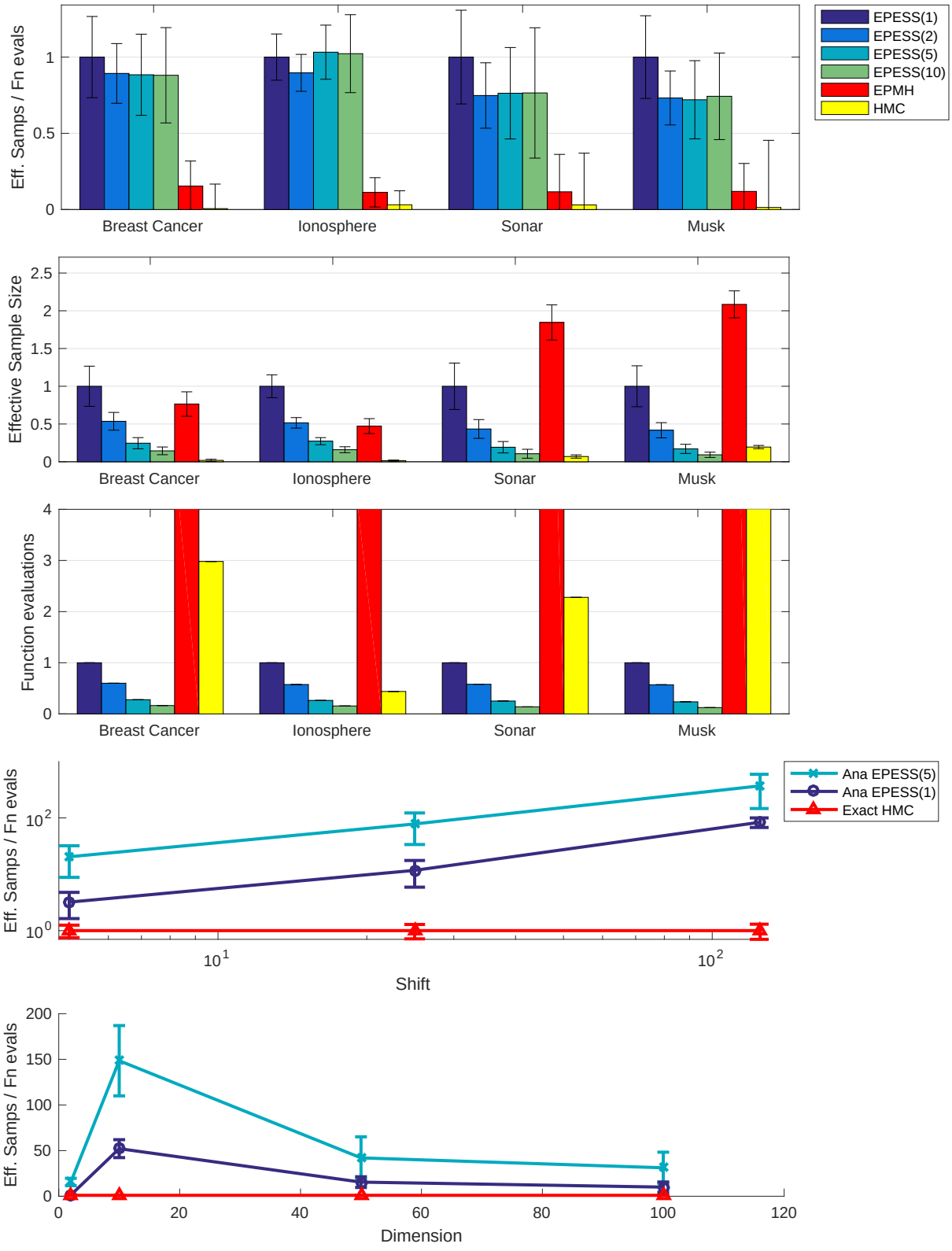

Figure 4: Plots of empirical results. In the top three plots all values are normalized so that EPESS(1) has value 1. In the bottom 2 plots all values are normalized so that Exact-HMC has value 1. The naming convention: EPESS(J) denotes Recycled EPESS with J points sampled per slice. EPESS(1) denotes EPESS without recycling. Ana EPESS(J) denotes Analytic EPESS with J threshold levels per iteration. Error bars in plot 3 for all algorithms are effectively zero.

choosing the prior that minimizes $\alpha = 1$ divergence, we could choose a prior corresponding to $\alpha > 1$ by replacing EP with Power-EP [Minka, 2004]. This might make the likelihood even weaker in EPESS and further improve performance.

Acknowledgements. We would like to thank Garud Iyengar for helpful discussions and Lichi Li for assisting in the code. FF is supported by a grant from Bloomberg. JPC is supported by Simons Foundation (SCGB#325171 and SCGB#325233), the Sloan Foundation, the McKnight Foundation, and the Grossman Center.

References

UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences.

AI Group at Arris Pharmaceutical Corporation. UCI Machine Learning Repository, 1994.

Yoann Altmann, Steve McLaughlin, and Nicolas Dobigeon. Sampling from a multivariate gaussian distribution truncated on a simplex: a review. In *Statistical Signal Processing (SSP), 2014 IEEE Workshop on*, pages 113–116. IEEE, 2014.

Michael Braun and Andre Bonfrer. Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Science*, 30(3):513–531, 2011.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: a probabilistic programming language. *Journal of Statistical Software*, 2015.

Nicolas Chopin and James Ridgway. Leave pima indians alone: binary regression as a benchmark for bayesian computation. *arXiv preprint arXiv:1506.08640*, 2015.

Botond Cseke and Tom Heskes. Approximate marginals in latent gaussian models. *The Journal of Machine Learning Research*, 12:417–454, 2011.

John P Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation. *arXiv preprint arXiv:1111.6832*, 2011.

Guillaume P Dehaene and Simon Barthelmé. Bounding errors of expectation-propagation. In *Advances in Neural Information Processing Systems*, pages 244–252, 2015.

Marc Deisenroth and Shakir Mohamed. Expectation propagation in gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2012.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2006.

Young Jun Ko and Matthias Seeger. Expectation propagation for rectified linear poisson regression. In *Proceedings of the Seventh Asian Conference on Machine Learning*, number EPFL-CONF-214372, 2015.

Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *The Journal of Machine Learning Research*, 6:1679–1704, 2005.

Shiwei Lan and Babak Shahbaba. Sampling constrained probability distributions using spherical augmentation. *arXiv preprint arXiv:1506.05936*, 2015.

Thomas Minka. Power EP. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

Thomas Minka. Divergence measures and message passing. Technical report, 2005.

Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Iain Murray, Ryan Prescott, Adams David, and J. C. Mackay. Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Radford Neal. Slice sampling. *Annals of Statistics*, 31: 705–741, 2003.

Radford M Neal. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(10), 2008.

Robert Nishihara, Iain Murray, and Ryan P. Adams. Parallel mcmc with generalized elliptical slice sampling. *J. Mach. Learn. Res.*, 15(1):2087–2112, January 2014. ISSN 1532-4435.

Akihiko Nishimura and David Dunson. Recycling intermediate steps to improve Hamiltonian Monte Carlo. *arXiv preprint arXiv:1511.06925*, 2015.

Ari Pakman and Liam Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23 (2):518–542, 2014.

Jonathan W Pillow, Liam Paninski, and Eero P Simoncelli. Maximum likelihood estimation of a stochastic integrate-and-fire neural model. In *NIPS*, pages 1311–1318. Citeseer, 2003.

Nicholas G Polson, James G Scott, and Jesse Windle. The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4): 713–733, 2014.

James Ridgway. *EPGLM: Gaussian Approximation of Bayesian Binary Regression Models*, 2016. R package version 1.1.1.

S Särkkä and A Vehtari. Mcmc diagnostics for matlab. *Helsinki University of Technology*, 2014-02-34.

Vince Sigillito. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 1989.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 1995.