Concave-Convex Adaptive Rejection Sampling

Dilan Görür*
Gatsby Computational Neuroscience Unit, University College London, UK
and
Yee Whye Teh
Gatsby Computational Neuroscience Unit, University College London, UK

January 20, 1010

Abstract

We describe a method for generating *independent* samples from univariate density functions using adaptive rejection sampling without the log-concavity requirement. The method makes use of the fact that many functions can be expressed as a sum of concave and convex functions. Using a concave-convex decomposition, we bound the log-density by separately bounding the concave and convex parts using piecewise linear functions. The upper bound can then be used as the proposal distribution in rejection sampling. We demonstrate the applicability of the concave-convex approach on a number of standard distributions and describe an application to the efficient construction of sequential Monte Carlo proposal distributions for inference over genealogical trees. Computer code for the proposed algorithms are available online.

Keywords: Random number generation, Monte Carlo sampling, Concave-convex decomposition, Log-concave densities

1 INTRODUCTION

Probabilistic graphical models have become popular tools for addressing many statistical and machine learning problems in recent years (Cowell et al., 1999; Jensen, 1996). This has been especially accelerated by the development of general-purpose approximate inference techniques based on Markov chain Monte Carlo (Gilks et al., 1996; Neal, 1993), sequential Monte Carlo (SMC) (Doucet et al., 2001) and variational approximations (Opper and Saad, 2001; Wainwright

*The authors gratefully acknowledge the Gatsby Charitable Foundation and PASCAL2 for funding. We thank two anonymous reviewers and the journal editors for their constructive comments.

and Jordan, 2008). Recent years have even seen the development of software toolkits automating the process of constructing inference algorithms so that users can concentrate their efforts on the design of probabilistic models rather than on developing inference algorithms (Spiegelhalter et al., 1999, 2004; Winn, 2004; Minka et al., 2008). Adaptive rejection sampling (ARS) (Gilks and Wild, 1992) and other approaches to generating from univariate densities are important components in the growing suite of Monte Carlo inference methods.

The basic adaptive rejection sampler assumes densities are log-concave and constructs piecewise exponential proposal distributions which are adaptively refined using past rejected samples. In this paper we propose a novel generalization of adaptive rejection sampling to distributions whose log-densities can be expressed as sums of concave and convex functions. These form a very large class of distributions; roughly, the only requirements are that the log-densities are differentiable with derivatives of bounded variation and that the tails decrease to $-\infty$ at least exponentially quickly. Further, many distributions of interest, including some multimodal distributions, have log-densities that trivially decompose into a sum of concave and convex components. We call our generalization concave-convex adaptive rejection sampling (CCARS).

The basic idea of CCARS is illustrated in Figure 1 and described in detail in Section 3. The logarithm of the density function is decomposed into a concave and a convex component. Upper bounds on both components are built using piecewise linear functions. These bounds are added to form an upper bound on the log-density. Exponentiating, we have a piecewise exponential upper bound on the density, which we use as a proposal distribution for rejection sampling. The bound is iteratively refined to improve the acceptance probability of future proposals.

In recent years a number of different generalizations of ARS has been proposed. Adaptive rejection Metropolis sampling (ARMS) (Gilks et al., 1995) and ARMS2 (Meyer et al., 2008) use adaptively refined proposals within a Metropolis-Hastings sampler. ARMS uses piecewise linear proposal distributions whereas ARMS2 uses second order polynomials to construct the proposal density. Both algorithms do not require log-concave densities, but they do not produce exact samples (since they are MCMC samplers). This is not a serious drawback if the sampler were used as a part of a more extensive MCMC sampler. However there are other occasions for which exact samples are desirable (e.g. if used as a component of a SMC algorithm). Hoermann (1995) extended the concept of log-concavity to T-concavity (Hoermann, 1995) and Evans and Swartz

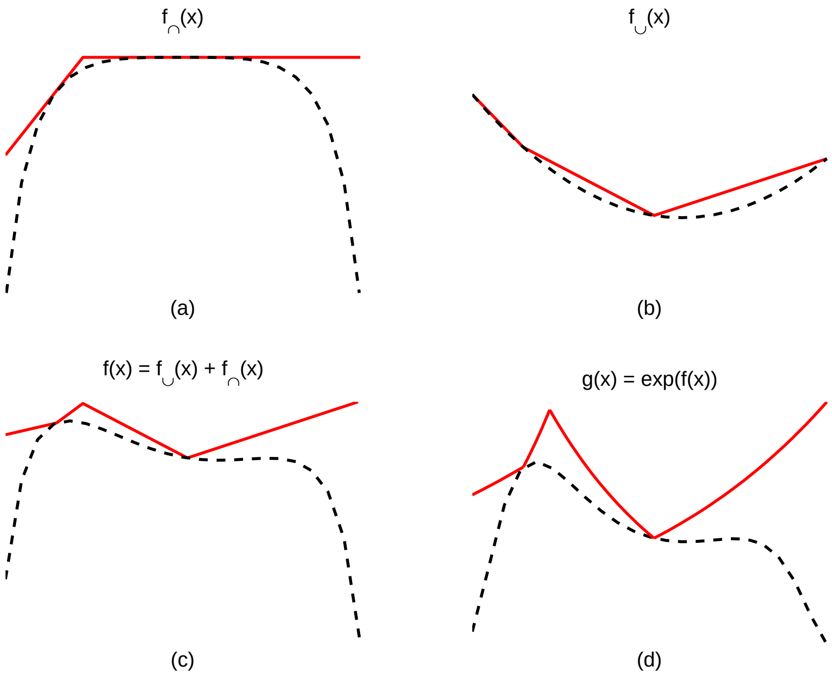

Figure 1: Bounding the log-density function using CCARS. The sum of the concave function **(a)** and the convex function **(b)** give the log-density function **(c)**. Piecewise linear bounds are constructed separately on the concave and convex functions and summed to form the bound on the log-density function. **(d)** shows the true density and the piecewise exponential bounds on the density. The functions are drawn with dashed black lines, and the bounds with solid red lines.

(1998) used this idea to develop a generalization of ARS (we refer to this as ES-ARS in the following) where the logarithmic transform is generalized to a monotonic transformation T. ES-ARS uses the inflection points of the transformed density to partition the domain into regions where the density is T-concave or T-convex, and adaptively constructs piecewise linear upper bounds separately on each partition to use as the proposal distribution. Though we restrict the description of CCARS to log-transforms of densities over a single interval, note that the same idea can be used with T-transforms and partitioning of the interval as well. We show experimentally in Section 4 that CCARS can be competitive with ES-ARS given the same information about inflection points. However, as opposed to ES-ARS, CCARS does not necessarily need knowledge of the inflection points, which may not always be easy to compute. In Section 3 we will further elaborate on the relationship between CCARS and ES-ARS.

We describe ARS in detail in Section 2 for completeness and to set the stage for the contributions of this paper in the following sections. We explain the CCARS algorithm and give details about how to decompose functions into concave and convex parts in Section 3. Section 4 presents example function decompositions and simulation results of CCARS on standard distributions. A distinguishing characteristic of CCARS is that it does not require knowledge of inflection points. The merit of this property is demonstrated in Section 5 where we describe use of CCARS in a SMC algorithm for inferring genealogical trees. We conclude in Section 6 with some discussion on the properties of CCARS and further extensions of the method.

2 ADAPTIVE REJECTION SAMPLING

Rejection sampling is a standard Monte Carlo technique for sampling from a distribution. Suppose the distribution has density $p(x)$, and assume that we have a proposal distribution with density $q(x)$ for which there exists a constant $c \geq 1$ such that $p(x) \leq cq(x)$ for all x, and from which it is easy to obtain samples. Rejection sampling proceeds as follows: obtain a sample $x \sim q(x)$; compute the acceptance probability $\alpha = p(x)/cq(x)$; and accept x with probability α, otherwise reject and repeat the procedure until some sample is accepted.

The intuition behind rejection sampling is straightforward. Generating a sample from $p(x)$ is equivalent to generating a sample from a uniform distribution under the curve $p(x)$. We obtain this sample by generating a uniform sample from under the curve $cq(x)$, and only accepting the sample if it by chance also falls under $p(x)$. This intuition also shows that the average acceptance probability is $1/c$, thus the expected number of samples required from $q(x)$ is c.

The choice of the proposal distribution is crucial to the efficiency of rejection sampling. When a sample is rejected, the computations performed to obtain the sample are discarded and thus wasted. Adaptive rejection sampling (ARS) (Gilks and Wild, 1992) is a sampling method for univariate densities which addresses this issue by making use of the rejected samples to improve the proposal distribution so that future proposals have higher acceptance probabilities.

ARS assumes that the density $p(x)$ is log-concave, that is, $f(x) = k + \log p(x)$ is a concave function, where k is an arbitrary constant making $f(x)$ tractable to compute. Since $f(x)$ is concave, it is upper bounded by its tangent lines: $f(x) \leq T^f_{x_0}(x)$ for all x_0 and x, where $T^f_{x_0}(x) = f(x_0) + (x - x_0)f'(x_0)$ is the tangent to $f(x)$ at abscissa x_0. ARS uses proposal distributions whose log-densities are constructed as the minimum of a finite set of tangents:

$$f(x) \leq g(x) = \min_{i=1\dots n} T^f_{x_i}(x) \tag{1}$$

$$q(x) \propto \exp(g(x)) \tag{2}$$

where $x_1, \dots, x_n$ are the abscissae of the tangent lines and $q(x)$ is the proposal density. Since the envelope function $g(x)$ is piecewise linear, $q(x)$ is a piecewise exponential density that can be efficiently sampled from. Say we propose a point $x' \sim q$, and accept x' with probability $\exp(f(x') - g(x'))$. That is, x' will be accepted with low probability if the sampling density is far from the true density at that point. If the sample is rejected, instead of simply discarding it, we add it to the set of abscissae so that the new $q(x)$ will more closely match $p(x)$ around x'.

It will generally be costly to evaluate the function $f(x)$. We can improve the computational efficiency of the procedure by using a lower bound $l(x)$ on $f(x)$ based on secant lines subtended by consecutive abscissae. This allows testing for acceptance of the proposed point x' without the need to evaluate $f(x')$ every time. We first sample $u \sim \text{Uniform}(0, 1)$. Then the proposal x' is accepted if $u \leq \exp(l(x') - g(x'))$. This is called the squeezing step. If x' is not accepted, then we apply the rejection step where we accept if $u \leq \exp(f(x') - g(x'))$. If the proposal x' is not accepted at the squeezing step, this implies that it is likely to be located in a part of the real line where the upper and lower bounds differ significantly. Therefore we add it to the set of abscissae even if it is accepted at the subsequent rejection step, that is, whenever we evaluate $f(x)$.

The idea of using both tangent and secant lines of $f(x)$ can be used in different ways to improve ARS. In ES-ARS, the assumption of $f(x)$ being log-concave is relaxed by partitioning the real line into intervals using its inflection points, so that each interval is either log-concave or log-convex (Evans and Swartz, 1998). On the log-convex intervals the upper bounds are constructed using secant lines instead. In the next section we decompose $f(x)$ into the sum of a log-concave and a log-convex part, and upper bound them individually using tangent and secant lines respectively. In both ES-ARS and our approach we can construct lower bounds as well by inverting the roles of the tangent and secant lines, though we will not detail this straightforward extension here.

3 CONCAVE-CONVEX ADAPTIVE REJECTION SAMPLING

In this section we propose a generalization to ARS that produces independent samples without requiring log-concavity. We assume that the log-density $f(x) = k + \log p(x)$ can be decomposed into a concave $f_\cap(x)$ and a convex $f_\cup(x)$ function: $f(x) = f_\cap(x) + f_\cup(x)$. As we will see in Section 3.1, many densities of interest satisfy this condition (Hartman, 1959). The approach we take is to bound $f_\cap(x)$ and $f_\cup(x)$ separately using piecewise linear functions, so that the sum of the bounds is itself piecewise linear and a bound on $f(x)$. A pictorial depiction of the bound is given in Figure 1. For simplicity we start by describing in detail the case where the support of $p(x)$ is a finite closed interval $[d_l, d_r]$ with both $f_\cap(x)$ and $f_\cup(x)$ continuous on $[d_l, d_r]$. We discuss the general case in Section 3.2. Finally, Section 3.3 uses the same ideas to approximate normalization constants.

As mentioned, concave-convex adaptive rejection sampling (CCARS) maintains separate piecewise linear upper bounds on $f_\cap(x)$ and $f_\cup(x)$. Sum of these are used to construct a proposal distribution for rejection sampling. If the proposed point is rejected, the bounds are refined to be tight at that point and the algorithm repeats until a proposal is accepted. CCARS is depicted in Figure 2, and given in pseudocode in Algorithm 1.

As in ARS, the upper bound on the concave part $f_\cap(x)$ is formed by a series of tangent lines at a set of n abscissae, say ordered $d_l < x_1 \cdots < x_n < d_r$. At each abscissa x_i we form the tangent line $T^{f_\cap}_{x_i}(x) = f_\cap(x_i) + (x - x_i)f'_\cap(x_i)$, and the upper bound on $f_\cap(x)$ is:

$$f_\cap(x) \leq g_\cap(x) = \min_{i=1\ldots n} T^{f_\cap}_{x_i}(x). \tag{3}$$

The consecutive tangent lines $T^{f_\cap}_{x_i}$, $T^{f_\cap}_{x_{i+1}}$ intersect at a point $y_i \in (x_i, x_{i+1})$:

$$y_i = \frac{f_\cap(x_{i+1}) - f'_\cap(x_{i+1})x_{i+1} - f_\cap(x_i) + f'_\cap(x_i)x_i}{f_\cap(x_i) - f_\cap(x_{i+1})}$$

and $g_\cap(x)$ is piecewise linear.

On the other hand, the upper bound on the convex part $f_\cup(x)$ is formed by a series of n secant lines subtended at the same set of points $\{x_1 \ldots x_n\}$ and the domain limits $x_0 = d_l,\ x_{n+1} = d_r$. For each consecutive pair $x_i < x_{i+1}$ the secant line

$$S^{f_\cup}_{x_i x_{i+1}}(x) = \frac{f_\cup(x_{i+1}) - f_\cup(x_i)}{x_{i+1} - x_i}(x - x_i) + f_\cup(x_i)$$

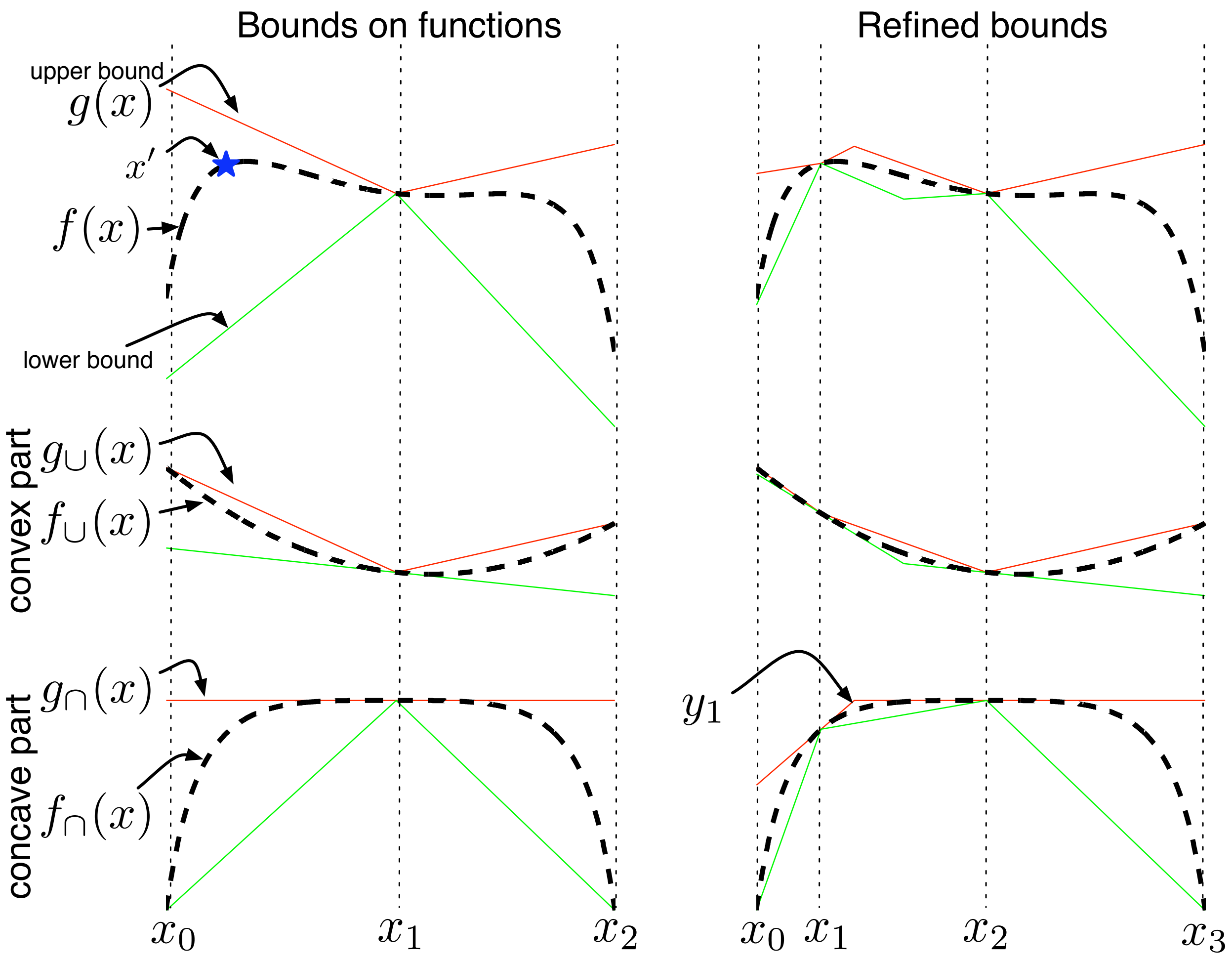

Figure 2: Concave-convex adaptive rejection sampling. **First column:** upper and lower bounds on functions $f(x)$, $f_{\cup}(x)$ and $f_{\cap}(x)$. We start with a single abscissa x_1. A point x' is proposed from the upper bound. **Second column:** refined bounds after the proposed point x' is rejected. x' is included in the set of abscissae and the abscissae are re-indexed. y_1 denotes the intersection of the two tangent lines.

is an upper bound on $f_{\cup}(x)$ on the interval $[x_i, x_{i+1}]$. Therefore the upper bound on $f_{\cup}(x)$ is:

$$f_{\cup}(x) \leq g_{\cup}(x) = \max_{i=0\ldots n-1} S^{f_{\cup}}_{x_i x_{i+1}}(x). \tag{4}$$

Finally the overall upper bound on the function $f(x)$ is just the sum of both upper bounds:

$$f(x) \leq g(x) = g_{\cap}(x) + g_{\cup}(x). \tag{5}$$

Note that $g(x)$ is a piecewise linear function with $2n$ segments, therefore the proposal distribution is a piecewise exponential distribution with $2n$ segments, $q(x) \propto \exp(g(x))$.

Algorithm 1 Concave-Convex Adaptive Rejection Sampling

```
inputs: functions $f_\cap$, $f_\cup$, domain $(d_l, d_r)$, numsamples
initialize: abscissae
search for a point $x_0$ on the left tail of $f_\cap + f_\cup$, add $x_0$ as left abscissa.
search for a point $x_1$ on the right tail of $f_\cap + f_\cup$, add $x_1$ as right abscissa.
initialize: bounds $g_\cap$ and $g_\cup$, numaccept $= 0$.
while numaccept $<$ numsamples do {generate samples}
  sample $x' \sim q(x) \propto \exp(g_\cap(x) + g_\cup(x))$.
  sample $u \sim \text{Uniform}(0, 1)$.
  if $u < \exp(l_\cap(x') + l_\cup(x') - g_\cap(x') - g_\cup(x'))$ then {squeezing step}
    accept the sample $x'$.
    numaccept := numaccept +1.
  else if $u < \exp(f_\cap(x') + f_\cup(x') - g_\cap(x') - g_\cup(x'))$ then {rejection step}
    accept the sample $x'$.
    numaccept := numaccept +1.
    include $x'$ in the set of abscissae.
    update the bounds.
  else
    reject sample $x'$.
    include $x'$ in the set of abscissae.
    update the bounds.
  end if
end while
```

As in Gilks and Wild (1992) we can also construct a lower bound $l(x)$ of $f(x)$ so that costly function evaluations can be avoided whenever possible using squeezing steps. This lower bound can be constructed by reversing the operations on the concave and convex functions: we lower bound $f_\cap(x)$ using its secant lines, and lower bound $f_\cup(x)$ using its tangent lines. This reversal is perfectly symmetrical and the same procedure described above can be used. We use the same set of points $\{x_0, \ldots, x_n\}$ to construct the lower bound, however note that the intersection points of the tangents are going to be different from those of the upper bound since they will depend on the derivative of $f_\cup$ rather than that of $f_\cap$.

The data structure maintained by CCARS consists of the n abscissae, the $n-1$ intersections of consecutive tangent lines of $f_\cap$, the $n-1$ intersections of the tangent lines of $f_\cup$, and the values of $g_\cap$, $g_\cup$, $l_\cap$ and $l_\cup$ evaluated at these points and the domain limits. Given the bounds, the algorithm is the same as ARS. At each iteration a sample $x' \sim q(x)$ is drawn from the proposal distribution and a value u from Uniform$(0, 1)$. The proposed point is accepted in the squeezing step if $u < \exp(l(x) - g(x))$. If the proposal fails at the squeezing step, it is accepted at the rejection step if $u < \exp(f(x) - g(x))$. If not accepted at the squeezing step, x' is added to the list of points to refine the upper and lower bounds. Pseudocode for CCARS is given in Algorithm 1.

3.1 Concave-Convex Decompositions

In this section we will briefly consider the existence of concave-convex decompositions before describing two different constructions for such decompositions. The following proposition gives a condition for existence which is satisfied by most log-densities of practical interest. In the next section we will describe conditions under which there are decompositions leading to valid CCARS procedures.

Proposition 1. *If $f(x)$ is differentiable with a derivative of bounded variation on every finite closed interval in its domain, then there exists a concave function $f_\cap(x)$ and a convex function $f_\cup(x)$ such that $f(x) = f_\cap(x) + f_\cup(x)$.*

Proof. This is straightforward and a slightly stronger proposition already noted by Hartman (1959). Since the derivative $f'(x)$ has bounded variation on every finite closed interval in the domain, it has a Jordan decomposition $f'(x) = h_\downarrow(x) + h_\uparrow(x)$ into a non-decreasing function $h_\uparrow(x)$ and a non-increasing function $h_\downarrow(x)$ (Hazewinkel, 1998). Integrating, we get $f(x) = k + \int_{d_l}^{x} h_\downarrow(x) + h_\uparrow(x)\,\mathrm{d}x$ where $f_\cap(x) = \int_{d_l}^{x} h_\downarrow(x)\,\mathrm{d}x$ is concave and $f_\cup(x) = \int_{d_l}^{x} h_\uparrow(x)\,\mathrm{d}x$ is convex. □

Note that the concave-convex decomposition of a function is not unique: one can always add a convex function to $f_\cup(x)$ and subtract the same from $f_\cap(x)$. Some decompositions may be easier to construct algebraically and some may result in more efficient sampling. Below we describe a procedure for obtaining the minimal concave-convex decomposition of a function (Hartman, 1959) and another algebraically simpler one that exploits additive structure (if there is one).

Minimal concave-convex decomposition using inflection points

A concave-convex decomposition $f(x) = f_\cap(x) + f_\cup(x)$ is minimal if there is no concave function (save affine ones) that can be subtracted from $f_\cap(x)$ and added to $f_\cup(x)$ while maintaining concavity and convexity of $f_\cap(x)$ and $f_\cup(x)$ respectively. Intuitively, both $f_\cap(x)$ and $f_\cup(x)$ have minimal amounts of variation, so we expect the bounds on $f(x)$ to be tighter and the resulting CCARS algorithm more efficient. This will be validated in the experiments in Section 4.

The construction in the proof of Proposition 1 is in fact a minimal concave-convex decomposition since the Jordan decomposition of $f'(x)$ involves minimally non-decreasing and non-increasing functions $h_\uparrow(x)$ and $h_\downarrow(x)$ respectively. This yields the following proposition:

Proposition 2. *Let $f(x)$ be differentiable with a derivative of bounded variation on every finite closed interval in its domain. The inflection points partition the domain of $f(x)$ into intervals over which $f(x)$ is either concave or convex. Denote by $D_\cap$ the union of the concave intervals, and likewise by $D_\cup$ for the convex intervals. If $f(x)$ has countably many inflection points $\{\xi_1, \ldots, \xi_M\}$, then it has a minimal decomposition given by*

$$f_\cap(x) = f(x)\,\mathbb{I}(x \in D_\cap) + \sum_{j=1}^{M} T^f_{\xi_j}(x)\,\big[\mathbb{I}\big(x > \xi_j \wedge (\xi_j^- \in D_\cap)\big) \vee \mathbb{I}\big(x < \xi_j \wedge (\xi_j^+ \in D_\cap)\big)\big], \tag{6}$$

$$f_\cup(x) = f(x)\,\mathbb{I}(x \in D_\cup) - \sum_{j=1}^{M} T^f_{\xi_j}(x)\,\big[\mathbb{I}\big(x > \xi_j \wedge (\xi_j^- \in D_\cap)\big) \vee \mathbb{I}\big(x < \xi_j \wedge (\xi_j^+ \in D_\cap)\big)\big]. \tag{7}$$

where $\mathbb{I}(A)$ is the indicator function taking value 1 when A is true and 0 otherwise, $T^f_\xi(x)$ is the tangent line of $f(x)$ at ξ, and $\xi_j^\pm$ means $\xi_j \pm \epsilon$ for an infinitesimal $\epsilon > 0$.

The first term in eq. (6) captures the concave parts of $f(x)$ but is only concave on each interval in $D_\cap$. The tangent lines are added to make $f_\cap(x)$ concave over the whole domain. Similarly for $f_\cup(x)$. The above equations allow for countably infinitely many inflection points M, however note that algorithmically the minimal concave-convex decomposition can be constructed only if $f(x)$ has a finite number of inflection points.

Using the minimal decomposition for CCARS has the same flavor as ES-ARS in that knowledge of the inflection points is necessary to partition the function domain. In ES-ARS each interval is treated independently thus allowing more flexibility. For example, different transform functions other than the logarithm can be applied on different intervals, and the decompositions need not be continuous at inflection points. On the other hand CCARS with the minimal decomposition potentially requires smaller numbers of abscissae since only one decomposition is used over the whole domain. We will compare the two algorithms experimentally in Section 4.

The construction of the minimal decomposition is generally applicable since most practically encountered functions have a finite number of inflection points. However the computation of the inflection points may require tedious algebra or expensive numerical solutions, neither of which is desirable. The next construction for concave-convex decompositions apply when $f(x)$ has additive structure, and it can be substantially cheaper to obtain. However it can be less efficient for CCARS (though possibly still cheaper than solving for the inflection points).

Concave-convex decomposition exploiting the additive structure

The CCARS algorithm is most naturally applied when the log density consists of a collection of additive components, $f(x) = \sum_{i=1}^{m} f_i(x)$. If each $f_i(x)$ is either concave or convex we can simply partition the sum accordingly to get a concave-convex decomposition of $f(x)$. When this is not the case, we can instead apply the minimal decomposition to each $f_i(x)$ separately and sum up the decompositions to form one for $f(x)$. This can be easier than finding the minimal decomposition of the whole function $f(x)$ since each $f_i(x)$ is typically much simpler.

Note that in principle CCARS is not limited to using the log transform and any transform function that can be used for ES-ARS can be used for CCARS as well. However the log transform will generally be the natural choice for CCARS for popular classes of models such as exponential family models or generalized linear models (Nelder and Wedderburn, 1972) which naturally have log additive structures.

3.2 Conditions for Applicability of CCARS

The first condition for CCARS to be applicable is that the log density $f(x)$ should be decomposable into concave $f_\cap(x)$ and convex functions $f_\cup(x)$. This condition is satisfied by most densities of practical interest, as discussed in the previous section. Secondly, CCARS uses piecewise exponential bounds on the density as the proposal distribution which is obtained by constructing piecewise linear upper bounds on the concave-convex decomposition of the log density. Therefore, $f(x)$ should be such that it can be upper bounded using piecewise linear functions with a finite number of segments (we call this a FPL bound in the following) and the area under the piecewise exponential bound on the density should be finite. The construction of the algorithm as described thus far along with Proposition 2 show that this condition is satisfied for differentiable log densities with a derivative of bounded variation defined over a finite closed interval. For functions that are not continuous at domain limits and for infinite domains, we need to check the tail behaviour. The proposition below sets out the conditions for the applicability of CCARS.

Proposition 3. *Let $f(x)$ be a log density defined over an interval with left and right limits d_l and d_r. Suppose $f(x)$ is differentiable with a derivative of bounded variation on every finite closed interval in the domain, and has finitely many inflection points. If one of the following holds on the right limit:*

1. *$d_r < \infty$ and $\lim_{x \to d_r} f(x) < \infty$, or*

2. *$d_r = \infty$ and $f(x)$ has a right tail decreasing to $-\infty$ at least linearly fast,*

and similarly for the left limit, then $f(x)$ has a minimal decomposition $f(x) = f_\cap(x) + f_\cup(x)$ where both $f_\cap(x)$ and $f_\cup(x)$ have FPL bounds. Furher, it follows that $f(x)$ has a FPL bound, and the exponentiated bound on the density has finite area.

Proof. Proposition 2 shows that $f(x)$ has a minimal decomposition given by eqs. (6) and (7). The usual approach of ARS shows that concave functions, thus $f_\cap(x)$, have FPL upper bounds. Next we show that $f_\cup(x)$ has a FPL upper bound on its right tail as well (similar arguments hold for the left tail). Since $f(x)$ has a finite number of inflection points, it is either concave or convex on the right limit (ξ, d_r), where ξ is the rightmost inflection point (use $\xi = d_l$ if $f(x)$ does not have any inflection points). If it is concave, eq. (7) shows that $f_\cup(x)$ is linear on (ξ, d_r) so it will have a FPL upper bound. Next suppose $f(x)$ is convex on (ξ, d_r) so that $f_\cup(x)$ is not linear in this range. Consider the two conditions in the proposition. In case 1, $\lim_{x \to d_r} f_\cup(x)$ must be finite since it is convex so is lower bounded by the tangent line $T_\xi^{f_\cup}(x)$ subtended at ξ, so the CCARS procedure as described previously will produce a FPL upper bound on $f_\cup(x)$. In case 2, $\lim_{x \to \infty} f'_\cup(x)$ exists and has some finite value f'_0, since $f'_\cup(x)$ is monotone increasing but $f_\cup(x)$ is upper bounded by a line on (ξ, ∞). We can construct a FPL bound on $f_\cup(x)$ using a set of secant lines subtended at a set of points $\{x_1, \ldots, x_n\}$ as before, except that on $[x_n, \infty)$ we use the upper bound $f_\cup(x_n) + (x - x_n) f'_0$ instead. The sum of the FPL upper bounds on $f_\cap(x)$ and $f_\cup(x)$ gives a FPL upper bound on $f(x)$. For a FPL upper bound on $f(x)$ with finite area after exponentiation, use any CCARS bound (with the above alteration) with $x_n > \xi$ (and x_1 less than the leftmost inflection point to handle the left limit). □

Note that the above bound construction on $f_\cup(x)$ is valid for the minimal decomposition. If a specific choice of $f_\cup(x)$ does not have a FPL upper bound a different decomposition should be sought. Alternatively, we can partition the domain into tail and non-tail regions and construct

Algorithm 2 Concave Convex Function Approximation

```
inputs: $f_\cap$, $f_\cup$, domain $(a, b)$, error threshold.
initialize: abscissae and upper and lower bounds $g_\cap$, $g_\cup$, $l_\cap$ and $l_\cup$ as in Algorithm 1.
initialize: calculate the areas under the bounds in each segment, $\{A_i^g, A_i^l\}$.
while $(\sum_i A_i^l)/(\sum_i A_i^g) <$ threshold do {refine bounds}
    find $i$ with largest difference in bound areas between abscissae $x_{i-1}$ and $x_i$.
    if $i =$ a log concave tail segment then
        sample $x' \sim q(x) \propto \exp(g_\cap(x) + g_\cup(x))$ on tail segment.
    else
        $x' = \text{argmax}_{x \in (x_{i-1}, x_i)} g(x) - l(x)$.
    end if
    include $x'$ in the set of abscissae.
    update the bounds.
end while
```

bounds separately in the different regions, using different function decompositions. In general, we can incorporate ideas from Evans and Swartz (1998) to partition the domain and apply different concave-convex decompositions on the transformed density (possibly using functions other than the logarithm). CCARS reduces to ES-ARS the domain is partitioned at every inflection point and the transformed density $f(x)$ is assigned to either $f_\cap(x)$ or $f_\cup(x)$ depending on whether it is concave or convex respectively.

3.3 Approximating Density Functions

The upper and lower bounds constructed by CCARS provide piecewise exponential envelopes approximating the true unnormalized density, and their easily evaluated integrals are approximations of the integral of the unnormalized density. This can be exploited to obtain a piecewise linear fit to the density function for instance to use as the proposal distribution within a sequential Monte Carlo algorithm.

The upper and lower bounds $l(x) \leq f(x) \leq g(x)$ provide bounds on the integral of $f(x)$,

$$\int_a^b \exp(l(x))\mathrm{d}x \leq \int_a^b \exp(f(x))\mathrm{d}x \leq \int_a^b \exp(g(x))\mathrm{d}x.$$

Thus, the area under the upper (lower) bounding piecewise exponential function gives an upper (lower) bound on the area under the unnormalized function $\exp(f(x))$ and the ratio of the areas under the upper and lower bounds can be used as a measure of the approximation error.

Algorithm 1 described above is optimized for requiring as few function evaluations as possible for generating samples from the distribution. Since in function approximation we are interested in optimally placing the abscissae rather than generating random samples, we can avoid sampling and update the bounds by deterministically choosing the abscissae. We start by initializing the abscissae and the upper and lower bounds $g(x)$, $l(x)$, and calculating the area under both bounds. At each iteration, we find the consecutive pair of abscissae with maximum area between $g(x)$ and $l(x)$ and add to the set of abscissae the point with largest discrepancy between $g(x)$ and $l(x)$ in this interval. The modification of CCARS for approximating functions is summarized in Algorithm 2.

4 EXPERIMENTS

In this section we demonstrate the performance of CCARS on a number of non-log-concave density functions for which there is no standard specialized sampling method. For all the log-density functions, we provide concave-convex decompositions obtained by exploiting the natural additive structure as well as noting the inflection points for obtaining the minimal decomposition described in Section 3.1. We compare CCARS using the two decompositions given and ES-ARS with the log transform, using the number of abscissae used by each algorithm as the measure of efficiency. We start with a description of the three classes of density functions and the corresponding concave-convex decompositions.

The **generalized inverse Gaussian** (GIG) distribution, introduced by Barndorff-Nielsen and Halgreen (1977), is ubiquitously used across many statistical domains, especially in financial data analysis (Eberlein and von Hammerstein, 2004). It is an example of an infinitely divisible distribution and this property allows the construction of nonparametric Bayesian models based on it (Barndorff-Nielsen, 1998; Jørgensen, 1982). The GIG density function on $x > 0$ is

$$p(x) = \frac{(a/b)^{\lambda/2}}{2K_\lambda(\sqrt{ab})} x^{\lambda-1} \exp\left\{-\frac{1}{2}(ax + bx^{-1})\right\},$$

where $K_\lambda(\cdot)$ is the modified Bessel function of the third kind and $a, b > 0$ and λ are parameters. Sampling from this distribution is not trivial, the most commonly used method being that of Dagpunar (1989). The unnormalized log-density is $f(x) = (\lambda - 1)\log(x) - \frac{1}{2}(ax + bx^{-1})$, which is log-concave when $\lambda \geq 1$ and a sum of concave and convex terms when $\lambda < 1$:

$$f_\cap(x) = -\frac{1}{2}(ax + bx^{-1}), \quad f_\cup(x) = (\lambda - 1)\log(x).$$

This decomposition constitutes an example of the case where the left tail of $f_\cup(x)$ cannot be bounded using a finite piecewise linear bounds. Therefore, we partition the domain into two regions $(0, \epsilon)$ and $[\epsilon, \infty)$ and use a different decomposition for the left tail. In detail, the log-densty is concave in $(0, \frac{b}{1-\lambda})$ and we can simply use $f_\cup(x) = 0$, $f_\cap(x) = f(x)$ in $(0, \epsilon)$, $\epsilon < b/(1 - \lambda)$.

To get the minimal decomposition, we note that $f(x)$ has one inflection point $\xi = \frac{b}{1-\lambda}$ for $\lambda < 1$. It is concave in $(0, \xi)$ and convex in (ξ, ∞), with $f'(x) \to -a/2$ as $x \to \infty$. We get a minimal decomposition by using the procedure outlined in Section 3.1, eq. (6) and (7).

Makeham's distribution is used to model the mortality rates at adult ages (Makeham, 1860). The density is

$$p(x) = (a + bc^x)\exp\left\{-ax - \frac{b}{\ln(c)}(c^x - 1)\right\},$$

where $x > 0$ and the parameters are $b > 0$, $c > 1$, $a > -b$. To the best of our knowledge, no specialized method for efficiently sampling from this distribution exists (Scollnik, 1995). When $a > 0$ the logarithm of the first term is convex, while that for the last term is concave. The natural concave-convex decomposition for the log-density $f(x) = \log p(x)$ is thus:

$$f_\cap(x) = -ax - \frac{b}{\ln(c)}(c^x - 1), \quad f_\cup(x) = \log(a + bc^x).$$

This distribution also has a single inflection point which is at $\xi = \log \frac{\sqrt{a \log c} - a}{b} \big/ \log c$ with the function being convex in $(0, \xi)$ and concave in (ξ, ∞). Evans and Swartz (1998) considered sampling from this distribution with parameter values $a = 0.01$, $b = 0.01$, and $c = \mathrm{e}$. In this case the log-density has an inflection point at $\xi = 2.197$.

The **polynomial-normal** distribution (Evans and Swartz, 1998) constitutes an interesting decomposition example. The unnormalized density function is given by

$$p(x) = \exp\left(-\frac{x^2}{2}\right)\prod_{k=1}^{m}((x - a_k)^2 + b_k^2), \tag{8}$$

where a_k and $b_k > 0$ are the parameters, and the product term defines a general nonnegative polynomial. Unlike the previous examples, the log-density function $f(x) = \log p(x)$ is not simply a sum of concave and convex terms. It is a sum of a quadratic term and logarithms of quadratic

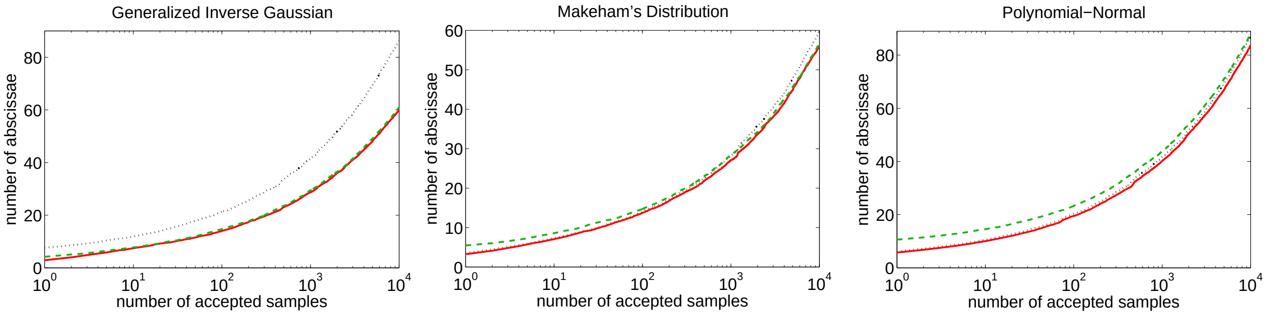

Figure 3: Numbers of abscissae as a function of the numbers of accepted samples generated by CCARS on the additive decomposition (dashed, black) and the minimal decomposition (solid, red), and ES-ARS using the log transform (dashed, green) averaged over 100 runs. The numbers of abscissae grow slowly for all methods. CCARS using the minimal decomposition is more efficient than the other two. Initially ES-ARS utilizes more abscissae but the gap disappears after a number of samples are accepted.

terms and computing its inflection points may be rather expensive for large m. It is however straightforward to calculate the inflection points of the logarithm of each quadratic term separately, construct a concave-convex decomposition for each term, and exploit the additive structure to obtain an overall concave-convex decomposition for $f(x)$. In detail, $f_k(x) = \log((x - a_k)^2 + b_k^2)$ has inflection points at $a_k \pm b_k$; it is convex in $\big(a_k - b_k, a_k + b_k\big)$ and concave outside. Thus we can decompose it as:

$$f_\cap(x) = -\frac{x^2}{2} + \sum_{k=1}^{m} f_k(x)\mathbb{I}(x \notin (a_k - b_k, a_k + b_k)) + T^{f_k}_{a_k - b_k}(x)\mathbb{I}(x > a_k - b_k) + T^{f_k}_{a_k + b_k}(x)\mathbb{I}(x < a_k + b_k),$$

$$f_\cup(x) = \sum_{k=1}^{m} f_k(x)\mathbb{I}(x \in (a_k - b_k, a_k + b_k)) - T^{f_k}_{a_k - b_k}(x)\mathbb{I}(x > a_k - b_k) - T^{f_k}_{a_k + b_k}(x)\mathbb{I}(x < a_k + b_k). \qquad (9)$$

Evans and Swartz (1998) considered this distribution for $m = 2$, $(a_1, b_1) = (1, .5)$ and $(a_2, b_2) = (-3, .5)$. For this case there are four inflection points, $\boldsymbol{\xi} = \{-3.398,\ -2.605,\ 0.605,\ 1.398\}$ and a minimal decomposition can be obtained using these. Note that computing inflection points, hence using ES-ARS and the minimal decomposition for CCARS, becomes harder for large m whereas the decomposition in eq. (9) is straightforward for any m.

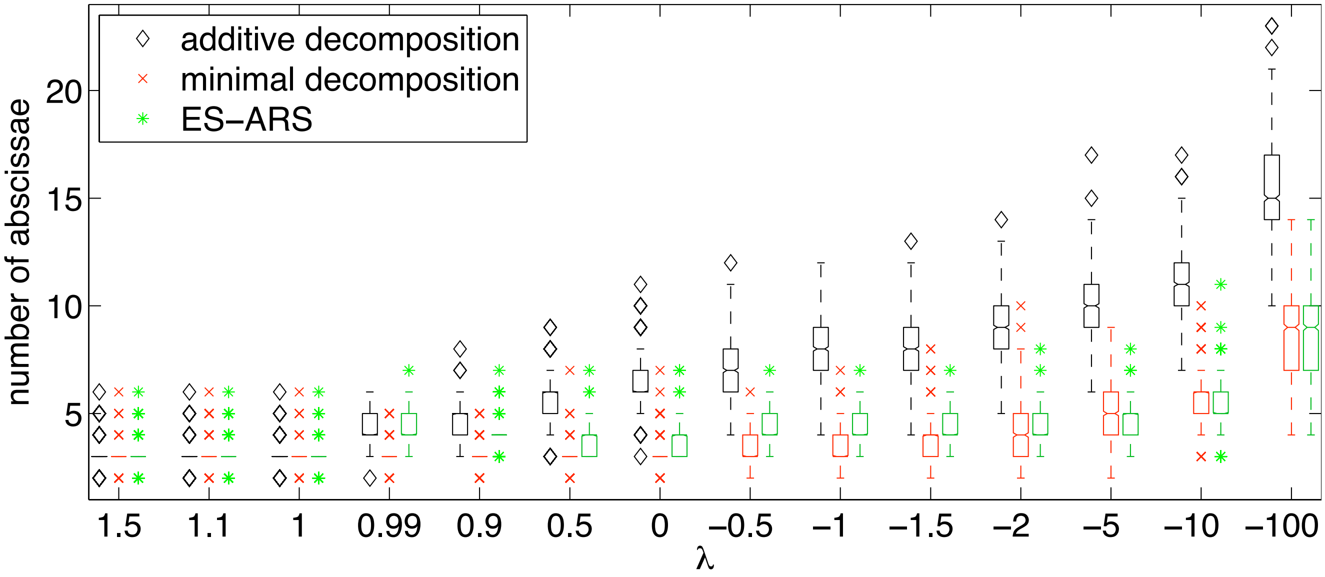

Figure 4: The numbers of abscissae used for generating a single sample from a GIG distribution with $a = b = 1$ and different values of λ, over 1000 runs. For each value of λ the left box-plot is CCARS with the natural decomposition, middle is CCARS with minimal decomposition, and right is ES-ARS. When $\lambda \geq 1$ the density function is log-concave and all methods reduce to ARS, thus have identical performances. CCARS with the natural but non-minimal decomposition consistently utilize more abscissae as expected, though the numbers are still reasonable. CCARS with the minimal decomposition and ES-ARS are comparable.

4.1 Generating Multiple Samples

For our first experiment, we compared CCARS with both decompositions against ES-ARS on the task of generating multiple samples from a given density. We experimented with one density from each family. For the GIG distribution we used parameter values $a = 1, b = 1$ and $\lambda = -1$. We used the parameter settings in (Evans and Swartz, 1998) for Makeham's distribution ($a = 0.01$, $b = 0.01$, $c = e$) and for the polynomial-normal distribution ($m = 2$, $(a_1, b_1) = (1, .5)$, $(a_2, b_2) = (-3, .5)$). Figure 3 shows the number of abscissae used as a function of the number of accepted samples generated by each algorithm, averaged over 100 runs. As expected, we observe that CCARS using the minimal decompositions generally utilize less abscissae than the natural but non-minimal decompositions. The performance of CCARS using the minimal decomposition is comparable to ES-ARS for these densities which also hold for other parameter settings that we tried.

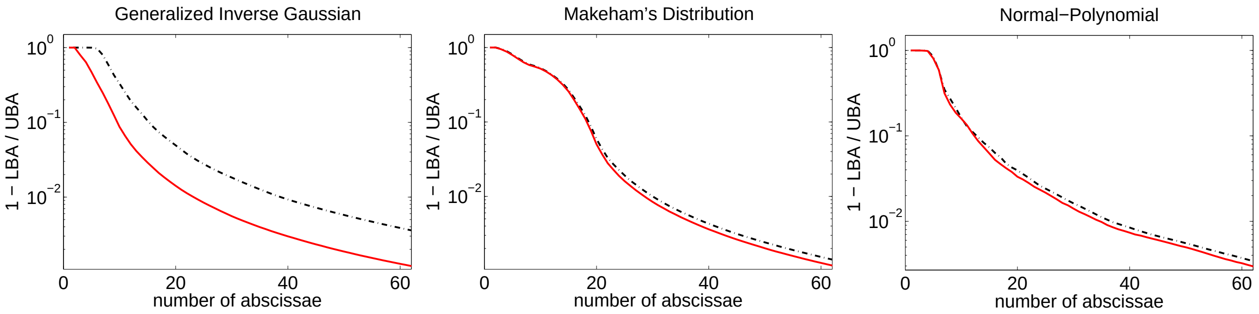

Figure 5: Accuracies of the integral approximations as a function of the number of abscissae. Dashed black curves are for the additive decompositions while the solid red lines are for minimal decompositions. Integral estimates quickly get more accurate as more abscissae are added.

4.2 Generating a Single Sample

Generally when these ARS methods are used within Gibbs sampling, one only needs a single sample from the conditional distribution for each Gibbs iteration. Therefore it is important to assess the cost of generating a single sample from a density. We consider the efficiencies of CCARS using both decompositions and of ES-ARS in this situation. We experimented with the GIG distribution with various parameter settings. We found that the comparative results are not sensitive to values of a and b thus report results only for $a = b = 1$ and various values of λ. Figure 4 shows the numbers of abscissae for the three methods and for different λ values over 1000 runs. The numbers of abscissae used increase with decreasing λ for all methods since the convexity increases. CCARS with the minimal decomposition performs similar to ES-ARS while both are faster than CCARS with the natural decomposition. It is interesting to note that the minimum number of abscissae used by ES-ARS is more than that used by CCARS with the minimal decomposition for most parameter settings.

4.3 Function Approximation

Figure 5 shows the performance of CCARS for approximating the unnormalized density functions using different function decompositions. We experimented with the three densities used for the experiment on generating multiple samples in Section 4.1. We see that only a small number of abscissae are needed to estimate the normalization constants accurately and the minimal decompositions are more efficient as expected.

5 PASSAGE TIMES AND GENEALOGICAL TREES

In this section we describe an application of CCARS to the inference of passage times in Markov processes. We first describe our problem formulation and then briefly describe how our solution to this passage time problem is used as part of a sequential Monte Carlo sampler for the coalescent (Kingman, 1982a,b).

5.1 Passage Times of Independent Markov Processes

Problems of determining passage (or waiting) times of diffusion and other stochastic processes arise from a variety of domains including statistics (Darling and Siegert, 1953; Durbin, 1971), probability (Ciesielski and Taylor, 1962; Kou and Wang, 2003), biology (Ewens, 1979), and engineering (Blake and Lindsey, 1973). Our problem differs from standard formulations in three ways: each Markov process is simple and discrete, but we deal with multiple Markov processes, and we deal with partial observations of the processes.

Let D be a positive integer and $X_1(\cdot), \ldots, X_D(\cdot)$ be D independent continuous time Markov processes. Each Markov process $X_d(\cdot)$ takes values in a finite state space $\{1, \ldots, K\}$ and has transition rate matrix $\Phi_d = \lambda_d(\phi_d \mathbf{1}^\top - I)$, where $\lambda_d > 0$ is the rate at which transitions occur, ϕ_d is the equilibrium distribution probability vector, $\mathbf{1}$ is a vector of 1's and I is the identity matrix. Each $X_d(\cdot)$ is a simple Markov process: transitions occur according to a Poisson process with uniform rate λ_d and at each transition the next state is independent of the previous state and has distribution ϕ_d.

Consider the problem of inferring the passage time $t = r - l$ separating observations at times l and r. For each $s \in \{l, r\}$, $d \in \{1, \ldots, D\}$ and $k \in \{1, \ldots, K\}$ let observation y_d^s be conditionally independent of the rest of the Markov chains given $X_d(s)$ and let M_{dk}^s be the likelihood of $X_d(s) = k$ given y_d^s. The likelihood for t given all observations is:

$$\mathcal{L}(t) = \prod_{d=1}^{D} \left\{ e^{-\lambda_d t} \left(\sum_{k=1}^{K} \phi_{dk} M_{dk}^l M_{dk}^r \right) + \left(1 - e^{-\lambda_d t}\right) \left(\sum_{k=1}^{K} \phi_{dk} M_{dk}^l \right) \left(\sum_{k=1}^{K} \phi_{dk} M_{dk}^r \right) \right\}. \qquad (10)$$

The first term is the probability of the observations and there being no transitions in X_d between times l and r. The second term is the probability of there being a transition and of the observations given that there was a transition. If $\sum_{k=1}^{K} \phi_{dk} M^l_{dk} M^r_{dk} > (\sum_{k=1}^{K} \phi_{dk} M^l_{dk})(\sum_{k=1}^{K} \phi_{dk} M^r_{dk})$ then the term in curly braces is convex; otherwise it is concave. Thus $\log \mathcal{L}(t)$ has a simple concave-convex decomposition. If the prior on t has a concave-convex decomposition as well, the resulting posterior will also have a concave-convex decomposition. Thus CCARS can be used to efficiently obtain samples from the posterior, and the function approximation can be used to obtain tight bounds on the posterior density.

5.2 Inference on Kingman's coalescent

In **?**, we used the above technique to develop a novel sequential Monte Carlo (SMC) sampler for the coalescent, a standard model of the genealogical structure of populations (Kingman, 1982a,b; Ewens, 2004). Given an observed set of genotypes, our aim is to estimate the posterior distribution over the possible genealogies, for the purpose of inferring parameters about the population.

Our SMC sampler operates assuming multiple independent sites, parent-independent mutation, and without recombination and selection. Under these assumptions, the genealogy of n genotypes has a binary tree structure with the n genotypes on the leaves of the tree. Our SMC sampler mirrors closely the generative process of the coalescent, constructing the genealogical tree by starting with the n observed genotypes each in its own (trivial) genealogy, and iteratively merging (coalescing) genealogies. In each iteration, a coalescence time for each pair of genealogies is proposed, and the pair with the shortest coalescence time is picked to be merged. The algorithm stops after $n - 1$ iterations when all genotypes have been coalesced into one genealogy. Being a SMC algorithm, multiple such runs (particles) are used, and resampling steps are taken to ensure that the particles are representative of the posterior distribution.

The passage time technique described above is used to construct the proposal distribution of the coalescence time of each pair of genealogies. With the assumption of independent sites and parent-independent mutation, the mutation process over genotypes is precisely the Markov process described above. The likelihood function of the coalescence time is of the form given in eq. (10), with the passage time being the total time from the proposed coalescence time to the times of the most recent common ancestors of both genealogies, while the partial observations are of the genotypes in each genealogy (with ancestral genotypes marginalized out). We use bounds fit using Algorithm 2 as the proposal densities. Figure 6 shows the density function and the proposal

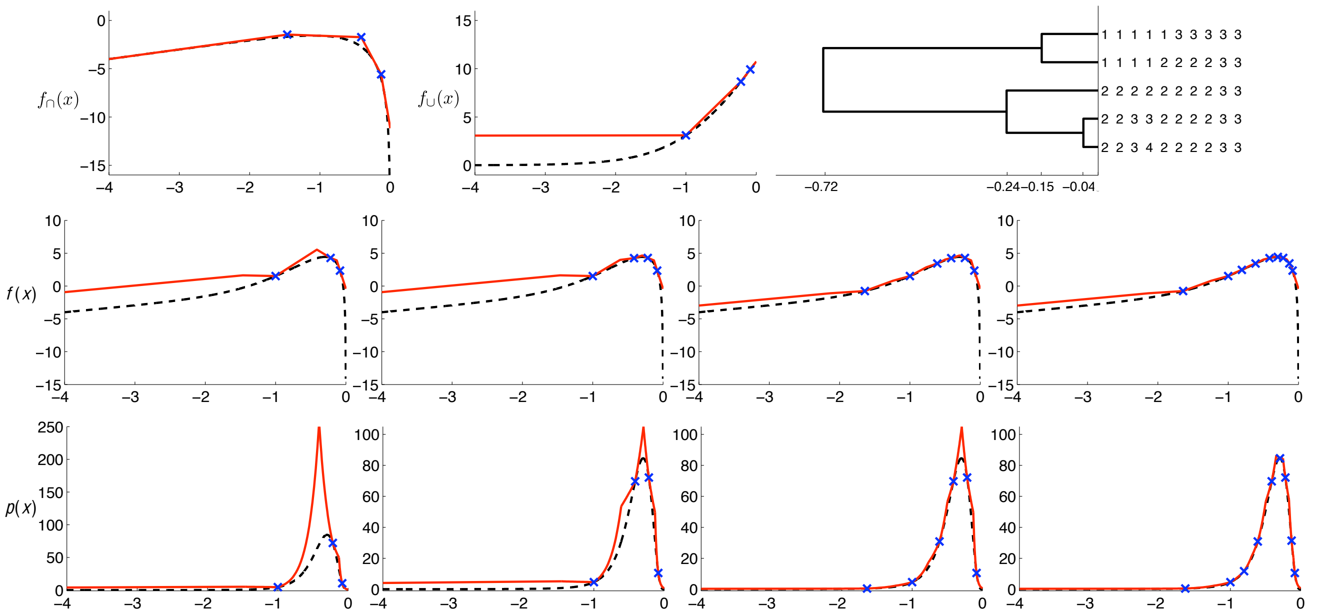

Figure 6: Sampling coalescence times using CCARS. For all plots, x-axis is time, running from past to present. The domain of the density function is truncated to $[-300, 0]$ for simplicity. Most coalescence times are greater than -10 therefore -300 is a safe bound and only one abscissa used at -297 was enough to appropriately bound the tail. **Top:** The concave and convex components of the log-density function (dashed, black) of coalescence times for a pair of subtrees and piecewise linear upper bounds on the functions (solid, red) using four abscissae (×) and an example tree structure inferred by sequential Monte Carlo sampling. **Middle:** Progressive fit of the CCARS bounds (solid, red) on the log-density function (dashed, black) and the abscissae (×) used to construct the bounds. **Bottom:** The density function and the piecewise exponential upper bounds. The abscissae are sequentially added to have a better fit. Note that the abscissae are concentrated around the high density region with high curvature.

density (i.e. the upper bound on the density) for a pair of genealogies. The tree in Figure 6 is a sample produced by the algorithm.

Our SMC sampler is related to a number of algorithms from the population genetics literature. The algorithm most similar to ours is by Slatkin (2002). Both Slatkin's and ours represent the coalescent events and the coalescence times, while they integrate out the ancestral genotypes. Where the algorithms differ is in their proposal distributions. Ours use a tight approximation to the posterior distributions of the coalescence times, while Slatkin's use their priors ignoring the genotypic observations. CCARS was essential in making our algorithm efficient, since the posterior distributions are not of standard form.Stephens and Donnelly (2000); Griffiths and Tavaré (1998); Chen et al. (2005) proposed sequential importance samplers with a similar flavour to our algorithm. The major difference is that while theirs marginalize out coalescence times, we marginalize out ancestral genotypes instead. As a result of this difference, their samplers can only efficiently handle single sites rather than multiple sites, since the space of ancestral genotypes becomes too large with multiple sites. We report detailed comparisons of our algorithm against these prior works in a separate manuscript under preparation.

6 DISCUSSION

We have proposed a generalization of adaptive rejection sampling to the case where the log density can be expressed as a sum of concave and convex functions. The generalization is based on the idea that both concave and convex functions can be upper bounded by piecewise linear functions, so that the sum of the piecewise linear functions is a piecewise linear upper bound on the log density itself. We experimentally verified that CCARS works efficiently on a number of well-known distributions, and showed that it is an indispensable component of a recently proposed SMC inference algorithm for genealogical trees under the coalescent.

The fact that the concave-convex decomposition is not unique has both advantages and disadvantages. An advantage is that we can have some function decompositions which may be straightforward, requiring minimal algebra, like the trivial decomposition of additive terms. A disadvantage is that we need to carefully choose between alternative decompositions as the redundancy in the decomposition may reduce efficiency of the sampler. Although the same function is being sampled from, a naïve decomposition may require more abscissae. We therefore suggest watching out for redundancy in the decomposition and when feasible, using a "effectual" concave-convex decomposition—one where both components are as close to linear as possible since the envelope functions are piecewise linear. The minimal concave-convex decomposition can considered to be effectual since in each interval defined by consecutive inflection points either $f_\cap$ or $f_\cup$ will be linear, and a single abscissa in that interval will make the bound exact on this term in that interval. Therefore CCARS will in general be more efficient on this decomposition since we can have a tighter envelope using the same number of abscissae. This can be seen as a compromise between the amount of calculations necessary to obtain the inflection points of the function versus the reduced efficiency of the sampler.

In the examples, we have employed concave-convex decompositions that fall out naturally from the functional forms of the log densities, as well as minimal decompositions that require knowledge of inflection points. We found that the minimal decompositions produce more efficient samplers, as expected. Another issue with an effect on efficiency, especially for single sample generation, is the initial abscissae positions. If available, information about the modes of the distribution or earlier runs of the algorithm can be used for better initialization. For example, if using CCARS as part of MCMC sampling, the abscissae information of previous iterations can be used, since the conditional distributions typically do not change drastically within a few MCMC iterations.

CCARS can be generalized by incorporating a number of ideas used to generalize the original adaptive rejection sampling. For log-concave densities Gilks (1992) proposed an alternative upper bound that is looser but does not require derivative evaluations. This can be applied to CCARS as well to obtain upper bounds on $f_\cap(x)$ (the upper bounds on $f_\cup(x)$ already do not require derivative information). We can also incorporate the ideas in ES-ARS (Evans and Swartz, 1998) to CCARS. Firstly, we can generalize the log transform to an arbitrary monotonic T transform. Secondly, we can partition the domain and apply a separate concave-convex decomposition (possibly with a different T transform) in each interval. Using interval partitions based on inflection points and the log transform across all intervals, we have found in our experiments (Section 4.2) that this approach has similar efficiency as CCARS with the minimal decomposition. The advantage of CCARS is that it can be applied in situations where inflection point information is unavailable or difficult to obtain.

Our algorithm is based on the fact that functions can be expressed as a sum of concave and convex functions, in other words a difference of convex functions (DC) (Hartman, 1959). A closely related area of research in the optimization literature is DC programming, where the objective function to be minimized is expressed as a DC function(Horst and Thoai, 1999; An and Tao, 2005). The idea of concave-convex decompositions have also been explored in the approximate inference context by Yuille and Rangarajan (2003), which can be seen as a special case of DC programming. We believe that concave-convex decompositions of functions are natural in other problems as well and exploiting such structure can lead to efficient solutions for such problems.

7 SUPPLEMENTAL MATERIALS

MATLAB code for CCARS algorithm: MATLAB code to generate samples from a distribution given the concave-convex decomposition of its log density and to fit bounds to the function for approximating its integral. (tar file)

References

*** (2009), "An Efficient Sequential Monte Carlo Algorithm for Coalescent Clustering," in *available to reviewers on AllenTrack*.

An, L. T. H. and Tao, P. D. (2005), "The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems," *Annals of Operations Research*, 133, 23–46.

Barndorff-Nielsen, O. (1998), "Processes of Normal Inverse Gaussian Type," *Finance And Stochastics*, 2, 41–68.

Barndorff-Nielsen, O. and Halgreen, C. (1977), "Infinite Divisibility of the Hyperbolic and Generalized Inverse Gaussian Distributions," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandete Gebiete*, 38, 309–311.

Blake, I. and Lindsey, W. (1973), "Level-crossing Problems for Random Processes," *IEEE Transactions on Information Theory*, 19, 295–315.

Chen, Y., Xie, J., and Liu, J. S. (2005), "Stopping-time Resampling for Sequential Monte Carlo Methods," *Journal of the Royal Statistical Society B*, 67, 199–217.

Ciesielski, Z. and Taylor, S. J. (1962), “First Passage times and Sojourn Times for Brownian Motion in Space and the Exact Hausdorff Measure of the Sample Path,” *Transactions of the American Mathematical Society*, 103, 434–450.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, Springer-Verlag.

Dagpunar, J. (1989), “An Easily Implemented Generalised Inverse Gaussian Generator,” *Communications in Statistics - Simulation and Computation*, 18, 703–710.

Darling, D. A. and Siegert, A. J. F. (1953), “The First Passage Problem for a Continuous Markov Process,” *Annals of Mathematical Statistics*, 24, 624–639.

Devroye, L. (1986), *Non-uniform Random Variate Generation*, Springer, New York.

Doucet, A., de Freitas, N., and Gordon, N. J. (2001), *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, New York: Springer-Verlag.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), “Hybrid Monte Carlo,” *Physics Letters*, 195(2), 216–222.

Durbin, J. (1971), “Boundary-crossing Probabilities for the Brownian Motion and Poisson Processes and Techniques for Computing the Power of the Kolmogorov-Smirnov Test,” *Journal of Applied Probability*, 8, 431–453.

Eberlein, E. and von Hammerstein, E. A. (2004), “Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes,” in *Seminar on Stochastic Analysis, Random Fields and Applications IV*, eds. Dalang, R., Dozzi, M., and Russo, F., vol. 58 of *Progress in Probability.*

Evans, M. and Swartz, T. (1998), “Random Variate Generation Using Concavity Properties of Transformed Densities,” *Journal of Computational and Graphical Statistics*, 7, 514–528.

Ewens, W. J. (1979), *Mathematical Population Genetics*, Springer-Verlag.

— (2004), *Mathematical Population Genetics I. Theoretical Foundations*, Springer.

Gilks, W. R. (1992), "Derivative-free adaptive rejection sampling for Gibbs sampling." in *Bayesian Statistics*, eds. Bernardo, J., Berger, J., Dawid, A. P., and Smith, A. F. M., Oxford University Press, vol. 4.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics*, 44, 455–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

Gilks, W. R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling," *Applied Statistics*, 41, 337–348.

Griffiths, R. C. and Tavaré, S. (1998), "The Age of a Mutation in a General Coalescent Tree," *Stochastic Models*, 14, 273–295.

Hartman, P. (1959), "On Functions Representable as a Difference of Convex Functions," *Pacific Journal of Mathematics*, 9, 707–713.

Hazewinkel, M. (ed.) (1998), *Encyclopedia of Mathematics*, Kluwer Academic Publ.

Hoermann, W. (1995), "A rejection technique for sampling from T-concave distributions," *ACM Transactions on Mathematical Software*, 21, 182–193.

Horst, R. and Thoai, N. V. (1999), "DC Programming: Overview," *Journal of Optimization Theory and Applications*, 103, 1–43.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*, UCL Press, London.

Jørgensen, B. (1982), *Statistical Properties of the Generalized Inverse Gaussian Distribution*, vol. 9 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

Kingman, J. F. C. (1982a), "The Coalescent," *Stochastic Processes and their Applications*, 13, 235–248.

— (1982b), "On the genealogy of large populations," *Journal of Applied Probability*, 19, 27–43.

Kou, S. G. and Wang, H. (2003), "First Passage Times of a Jump Diffusion Process," *Advances in Applied Probability*, 35, 504–531.

Makeham, W. M. (1860), "On the Law of Mortality and the Construction of Annuity Tables," *J. Inst. Actuaries and Assur. Mag.*, 8, 301310.

Meyer, R., Cai, B., and Perron, F. (2008), "Adaptive Rejection Metropolis sampling using Lagrange Interpolation Polynomials of Degree 2," *Computational Statistics and Data Analysis*, 52, 3408–3423.

Minka, T., Winn, J., Guiver, J., and Kannan, A. (2008), "Infer.NET," http://research.microsoft.com/mlp/ml/Infer/Infer.htm.

Neal, R. M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Tech. rep., Department of Statistics, University of Toronto.

— (2003), "Slice Sampling," *The Annals of Statistics*, 31, 705–767.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, 135, 370–384.

Opper, M. and Saad, D. (eds.) (2001), *Advanced Mean Field Methods : Theory and Practice*, The MIT Press.

Scollnik, D. P. M. (1995), "Simulating random variates from Makeham's distribution and from others with exact or nearly log-concave densities," *TRANSACTIONS OF SOCIETY OF ACTUARIES*, 47.

Slatkin, M. (2002), "A Vectorized Method of Importance Sampling with Applications to Models of Mutation and Migration," *Theoretical Population Biology*, 62, 339–348.

Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks., W. R. (1999, 2004), "BUGS: Bayesian inference using Gibbs sampling," http://www.mrc-bsu.cam.ac.uk/bugs/.

Stephens, M. and Donnelly, P. (2000), "Inference in Molecular Population Genetics," *Journal of the Royal Statistical Society*, 62, 605–655.

Wainwright, M. J. and Jordan, M. I. (2008), “Graphical Models, Exponential Families, and Variational Inference,” *Foundations and Trends in Machine Learning*, 1, 1–305.

Winn, J. (2004), “VIBES: Variational Inference in Bayesian networks,” http://vibes.sourceforge.net/.

Yuille, A. L. and Rangarajan, A. (2003), “The Concave-Convex Procedure,” *Neural Computation*, 15, 915–936.