Integrals over Gaussians under Linear Domain Constraints

Alexandra Gessner
University of Tuebingen and
MPI for Intelligent Systems
Tübingen, Germany
agessner@tue.mpg.de

Oindrila Kanjilal
University of Tuebingen and
Technical University of Munich
Germany
oindrila.kanjilal@tum.de

Philipp Hennig
University of Tuebingen and
MPI for Intelligent Systems
Tübingen, Germany
ph@tue.mpg.de

Abstract

Integrals of linearly constrained multivariate Gaussian densities are a frequent problem in machine learning and statistics, arising in tasks like generalized linear models and Bayesian optimization. Yet they are notoriously hard to compute, and to further complicate matters, the numerical values of such integrals may be very small. We present an efficient black-box algorithm that exploits geometry for the estimation of integrals over a small, truncated Gaussian volume, and to simulate therefrom. Our algorithm uses the Holmes-Diaconis-Ross (HDR) method combined with an analytic version of elliptical slice sampling (ESS). Adapted to the linear setting, ESS allows for *rejection-free* sampling, because intersections of ellipses and domain boundaries have closed-form solutions. The key idea of HDR is to decompose the integral into easier-to-compute conditional probabilities by using a sequence of nested domains. Remarkably, it allows for direct computation of the logarithm of the integral value and thus enables the computation of extremely small probability masses. We demonstrate the effectiveness of our tailored combination of HDR and ESS on high-dimensional integrals and on entropy search for Bayesian optimization.

1 INTRODUCTION

Multivariate Gaussian *densities* are omnipresent in statistics and machine learning. Yet, Gaussian *probabilities* are hard to compute—they require solving an integral over a constrained Gaussian volume—owing to the intractability of the multivariate version of the Gaussian cumulative distribution function (CDF). The probability mass that lies within a domain $\mathcal{L} \subset \mathbb{R}^D$ restricted by M linear constraints can be written as

$$Z = P(\mathbf{x} \in \mathcal{L}) = \int_{\mathbb{R}^D} \prod_{m=1}^{M} \Theta\left(\mathbf{a}_m^\intercal \mathbf{x} + b_m\right) \; d\,\mathcal{N}(\mathbf{x}; 0, \mathbf{1}), \tag{1}$$

with the Heaviside step function $\Theta(x) = 1$ if $x > 0$ and zero otherwise. We take the integration measure to be a standard normal without loss of generality, because any correlated multivariate Gaussian can be whitened by linearly transforming the integration variable.

Gaussian models with linear domain constraints occur in a myriad of applications that span all disciplines of applied statistics and include biostatistics (Thiébaut & Jacqmin-Gadda, 2004), medicine (Chen & Chang, 2007), environmental sciences (Wani et al., 2017), robotics and control (Fisac et al., 2018), machine learning (Su et al., 2016) and more. A common occurrence of this integral is in spatial statistics, such as Markov random fields (Bolin & Lindgren, 2015), the statistical modeling of spatial extreme events called max-stable processes (Huser & Davison, 2013; Genton et al., 2011), or in modeling uncertainty regions for latent Gaussian models. An example for the latter is to find regions that are likely to exceed a given reference level, e.g., pollution levels in geostatistics and environmental monitoring (Bolin & Lindgren, 2015), or in climatology (French & Sain, 2013). Another area where integrals like Eq. (1) are often encountered is in reliability analysis (Au & Beck, 2001a; Melchers & Beck, 2018; Andersen et al., 2018; Straub et al., 2020). A key problem there is to estimate the probability of a rare event to occur (e.g., a flood) or for a mechanical system to enter a failure mode.
In machine learning, there are many Bayesian models in which linearly constrained multivariate normal distributions play a role, such as Gaussian processes under linear constraints (López-Lopera et al., 2017;

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

López-Lopera et al., 2019; Agrell, 2019; Da Veiga & Marrel, 2012), inference in graphical models (Mulgrave & Ghosal, 2018), multi-class Gaussian process classification (Rasmussen & Williams, 2006), ordinal and probit regression (Lawrence et al., 2008; Ashford & Sowden, 1970), incomplete data classification (Liao et al., 2007), and Bayesian optimization (Hennig & Schuler, 2012; Wang et al., 2016), to name a few.

This practical relevance has fed a slow-burn research effort in the integration of truncated Gaussians over decades (Geweke, 1991; Genz, 1992; Joe, 1995; Vijverberg, 1997; Nomura, 2014). Gassmann et al. (2002) and Genz & Bretz (2009) provide comparisons and attest that the algorithm by Genz (1992) provides the best accuracy across a wide range of test problems, which has made it a default choice in the literature. Genz's method applies a sequence of transformations to transform the integration region to the unit cube $[0,1]^D$ and then solves the integral numerically using quasi-random integration points. Other methods focus on specialized settings such as bivariate or trivariate Gaussian probabilities (Genz, 2004; Hayter & Lin, 2013), or on orthant probabilities (Miwa et al., 2003; Craig, 2008; Nomura, 2016; Hayter & Lin, 2012). Yet, these methods are only feasible for at most a few tens of variables. Only recent advances have targeted higher-dimensional integrals: Azzimonti & Ginsbourger (2017) study high-dimensional orthant probabilities and Genton et al. (2018) consider the special case where the structure of the covariance matrix allows for hierarchical decomposition to reduce computational complexity. Phinikettos & Gandy (2011) employ a combination of four variance reduction techniques to solve such integrals with Monte Carlo methods. Botev (2016) constructs an exponential tilting of an importance sampling measure that builds on the method by Genz (1992) and reports effectiveness for $D \lesssim 100$. A different approach has been suggested by Cunningham et al. (2011): They use expectation propagation to approximate the constrained normal integrand of Eq. (1) by a moment-matched multivariate normal density. This allows for fast integration, at the detriment of guarantees. Indeed, the authors report cases in which EP is far off the ground truth integral.

Closely related to integration is *simulation* from linearly constrained Gaussians, yet these tasks have rarely been considered concurrently, except for Botev (2016) who proposes an accept-reject sampler alongside the integration scheme. Earlier attempts employ Gibbs sampling (Geweke, 1991), or other Monte Carlo techniques (Cong et al., 2017). Koch & Bopp (2019) recently introduced an algorithm for exact simulation from truncated Gaussians. Their method iteratively samples from transformed univariate truncated Gaussians that satisfy the box constraints.

In our work, we jointly address the sampling and the normalization problem for linearly constrained domains in a Gaussian space, making the following contributions:

- We present an adapted version of elliptical slice sampling (ESS) which we call LIN-ESS that allows for *rejection-free* sampling from the linearly constrained domain $\mathcal{L}$. Its effectiveness is not compromised even if the probability mass of $\mathcal{L}$ is very small (cf. Section 2.1).
- Based on the above LIN-ESS algorithm, we introduce an efficient integrator for truncated Gaussians. It relies on a sequence of nested domains to decompose the integral into multiple, easier-to-solve, conditional probabilities. The method is an adapted version of the Holmes-Diaconis-Ross algorithm (Diaconis & Holmes, 1995; Ross, 2012; Kroese et al., 2011) (cf. Section 2.2).
- With increasing dimension D, the integral value Z can take extremely small values. HDR with a LIN-ESS sampler allows to compute such integrals efficiently, and to even compute the logarithm of the integral.
- With LIN-ESS, sampling is sufficiently efficient to also compute *derivatives* of the probability with respect to the parameters of the Gaussian using expectations.

We provide a PYTHON implementation available at `https://github.com/alpiges/LinConGauss`.

2 METHODS

We first introduce an adapted version of elliptical slice sampling, LIN-ESS, which permits efficient sampling from a linearly constrained Gaussian domain of arbitrarily small mass once an initial sample within the domain is known. This routine is a special case of elliptical slice sampling that leverages the analytic tractability of intersections of ellipses and hyperplanes to speed up the ESS loop. LIN-ESS acts at the back-end of the integration method, which is introduced in Section 2.2.

For further consideration, it is convenient to write the linear constraints of Eq. (1) in vectorial form, $\mathbf{A}^\intercal \mathbf{x} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{D\times M}$, $\mathbf{x} \in \mathbb{R}^D$, and $\mathbf{b} \in \mathbb{R}^M$. The integration domain $\mathcal{L} \subset \mathbb{R}^D$ is given by the intersection of the region where all the M constraints exceed zero. For example, orthant probabilities of a correlated Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written in the form of Eq. (1) by using the transformation $\mathbf{x} = \mathbf{L}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{L}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$. Typically, we expect $M \geq D$, i.e., there are at least as many linear constraints as dimensions. This is because if $M < D$, there exists a transformation of $\mathbf{x}$ such that $D - M$ dimensions can be integrated out in closed form, and

an M-dimensional integral with M constraints remains. However, there are situations in which integrating out dimensions might be undesired. This is the case, e.g., when samples from the untransformed integrand are required.

2.1 Sampling from truncated Gaussians

Elliptical slice sampling (ESS) by Murray et al. (2010) is a Markov chain Monte Carlo (MCMC) algorithm to draw samples from a posterior when the prior is a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Given an initial location $\mathbf{x}_0 \in \mathbb{R}^D$, an auxiliary vector $\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is drawn to construct an ellipse $\mathbf{x}(\theta) = \mathbf{x}_0 \cos\theta + \boldsymbol{\nu} \sin\theta$ parameterized by the angle $\theta \in [0, 2\pi]$. In the general case, the algorithm proceeds similarly to regular slice sampling (Neal, 2003), but on the angular domain. A likelihood threshold is defined, and rejected proposals (in θ) with likelihood values below the threshold are used to adapt the bracket $[\theta_{\min}, \theta_{\max}]$ to sample from, until a proposal is accepted that serves as new $\mathbf{x}_0$ (see Murray et al. (2010) for details).

ESS is designed for generic likelihood functions. The special form of the likelihood in Eq. (1) can be leveraged to significantly simplify the ESS algorithm:

1. The selector $\ell(\mathbf{x}) := \prod_{m=1}^{M} \Theta[\mathbf{a}_m^\intercal \mathbf{x} + b_m]$ can take only the values 0 and 1. Hence there is no need for a likelihood threshold, the domain to sample from is always defined by $\ell(\mathbf{x}) = 1$ for $\mathbf{x}(\theta)$ on the ellipse.
2. The intersections between the ellipse and the linear constraints have closed-form solutions. The angular domain(s) to sample from can be constructed analytically, and LIN-ESS is thus rejection-free. The typical bisection search of slice sampling becomes a simple analytic expression.

With these simplifications of ESS, each sample from $\mathcal{L}$ requires exactly one auxiliary normal sample $\boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{1}) \in \mathbb{R}^D$ and a scalar uniform sample $u \sim \text{Uniform}[0, 1]$ to sample from the angular domain. Fig. 1 illustrates the process of drawing a sample from the domain of interest (blue shaded area) using our version of ESS. Given the two base vectors $\mathbf{x}_0 \in \mathcal{L}$ and $\boldsymbol{\nu}$, the ellipse is parameterized by its angle $\theta \in [0, 2\pi]$. The intersections between the ellipse and the domain boundaries $\mathbf{A}^\intercal \mathbf{x} + \mathbf{b} = \mathbf{0}$ can be expressed in closed form in terms of angles on the ellipse as solution to the set of equations $\mathbf{A}^\intercal(\mathbf{x}_0 \cos\theta + \boldsymbol{\nu} \sin\theta) + \mathbf{b} = \mathbf{0}$. For the m^{th} constraint, this equation typically has either zero or two solutions,

$$\theta_{m,1/2} = \pm \arccos\left(-\frac{b_m}{r}\right) + \arctan\left(\frac{\mathbf{a}_m^\intercal \boldsymbol{\nu}}{r + \mathbf{a}_m^\intercal \mathbf{x}_0}\right) \quad (2)$$

with $r = \sqrt{(\mathbf{a}_m^\intercal \mathbf{x}_0)^2 + (\mathbf{a}_m^\intercal \boldsymbol{\nu})^2}$. A single solution occurs in the case of a tangential intersection, which is unlikely.

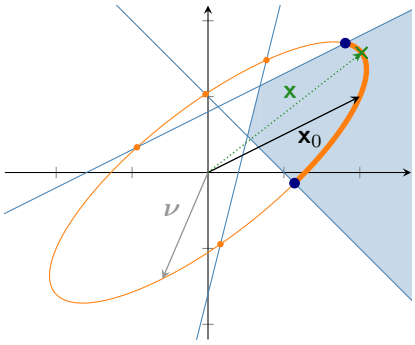

Figure 1: Sampling from a constrained normal space using ESS. $\mathbf{x}_0$ is a previous sample from the domain $\mathcal{L}$ and, with the auxiliary $\boldsymbol{\nu}$, defines the ellipse. From all intersections of the ellipse and zero lines (or hyperplanes in higher dimensions), the *active* intersections at the domain boundary are identified (•). These define the slice from which a uniform sample is drawn (×).

Not all intersection angles lie on the domain boundary and we need to identify those *active* intersections where $\ell(\mathbf{x}(\theta))$ switches on or off. To identify potentially multiple brackets, we sort the angles in increasing order and check for each of them if adding/subtracting a small $\Delta\theta$ causes a likelihood jump. If there is no jump, the angle is discarded, otherwise the sign of the jump is stored (whether from 0 to 1 or the reverse), in order to know the direction of the relevant domain on the slice. Pseudocode for LIN-ESS can be found in Algorithm 2 in the appendix.

The computational cost of drawing one sample on the ellipse is dominated by the M inner products that need to be computed for the intersections, hence the complexity is $\mathcal{O}(MD)$. This is comparable with standard ESS for which drawing from a multivariate normal distribution is $\mathcal{O}(D^2)$, but the suppressed constant can be much smaller because there is no need to evaluate a likelihood function in LIN-ESS. This version of ESS is a rejection-free sampling method to sample from a truncated Gaussian of arbitrarily small mass—except that it requires an initial point within the domain from where to launch the Markov chain. How to obtain such a sample will be discussed in Section 2.2.2.

2.2 Computing Gaussian probabilities

2.2.1 The Holmes-Diaconis-Ross algorithm

The Holmes-Diaconis-Ross algorithm (HDR) (Diaconis & Holmes, 1995; Ross, 2012; Kroese et al., 2011) is a specialized method for constructing an unbiased estimator for probabilities of the form $P(\mathbf{x} \in \mathcal{L})$ under an arbitrary prior measure $\mathbf{x} \sim p_0(\mathbf{x})$ and a domain $\mathcal{L} = \{\mathbf{x} \text{ s.t. } f(\mathbf{x}) \geq 0\}$ with a deterministic function

Algorithm 1 The Holmes-Diaconis-Ross algorithm applied to linearly constrained Gaussians

```
procedure HDR(A, b, {γ_1, ..., γ_T}, N)
    X ~ N(0, 1)                                   // N samples
    log Z = 0                                     // initialize log integral value
    for t = 1...T do
        L_t = {x : min_m(a_m^T x_n + b_m) + γ_t > 0}_{n=1}^N
                                                  // find samples inside current nesting
        log Z ← log Z + log(#(X ∈ L_t)) − log N
        choose x_0 ∈ L_t
        X ← LinESS(A, b + γ_t, N, x_0)
        // draw new samples from constrained domain
    end for
    return log Z
end procedure
```

$f : \mathbb{R}^D \mapsto \mathbb{R}$. If this domain has very low probability mass, $P(\mathcal{L})$ is expensive to compute with direct Monte Carlo because most samples are rejected. HDR mitigates this by using a sequence of T nested domains $\mathbb{R}^D = \mathcal{L}_0 \supset \mathcal{L}_1 \supset \mathcal{L}_2 \supset \ldots \supset \mathcal{L}_T = \mathcal{L}$, s.t. $\mathcal{L}_t = \bigcap_{i=1}^{t} \mathcal{L}_i$. The probability mass of the domain of interest can be decomposed into a product of conditional probabilities,

$$Z = P(\mathcal{L}) = P(\mathcal{L}_0) \prod_{t=1}^{T} P(\mathcal{L}_t | \mathcal{L}_{t-1}). \tag{3}$$

If each of the conditional probabilities $P(\mathcal{L}_{t+1}|\mathcal{L}_t)$ is closer to $1/2$, they all require quadratically fewer samples, reducing the overall cost despite the linear increase in indidivual sampling problems. Noting that $P(\mathcal{L}_0) = 1$ and introducing the shorthand $\rho_t = P(\mathcal{L}_t|\mathcal{L}_{t-1})$, Eq. (3) can be written in logarithmic form as $\log Z = \sum_{t=1}^{T} \log \rho_t$.

HDR does not deal with the construction of these nested domains—a method to obtain them is discussed in Section 2.2.2. For now, they are assumed to be given in terms of a decreasing sequence of positive scalar values $\{\gamma_1, \ldots, \gamma_T\}$, where $\gamma_T = 0$. Each shifted domain $\mathcal{L}_t$ can then be defined through its corresponding shift value γ_t. In the general setting, this is $\mathcal{L}_t = \{\mathbf{x} \text{ s.t. } f(\mathbf{x}) + \gamma_t \geq 0\}$; in our specific problem of linear constraints, $\mathbf{x} \in \mathcal{L}_t$ if $\ell_t(\mathbf{x}) = \prod_{m=1}^{M} \Theta(\mathbf{a}_m^\intercal \mathbf{x} + b_m + \gamma_t) = 1$. Any positive shift γ_t thus induces a domain $\mathcal{L}_t$ that contains all domains $\mathcal{L}_{t'}$ with $\gamma_{t'} < \gamma_t$, and that engulfs a larger volume than $\mathcal{L}_{t'}$. The T^{th} shift $\gamma_T = 0$ identifies $\mathcal{L}$ itself.

Given the shift sequence $\{\gamma_1, \ldots, \gamma_T\}$, the HDR algorithm proceeds as follows: Initially, N samples are drawn from $\mathcal{L}_0$, the integration measure, in our case a standard normal. $\mathcal{L}_0$ corresponds to $\gamma_0 = \infty$ which is ignored in the sequence. The conditional probability $\rho_1 = P(\mathcal{L}_1 \,|\, \mathcal{L}_0)$ is estimated as the fraction of samples from $\mathcal{L}_0$ that also fall into $\mathcal{L}_1$. To estimate the subsequent conditional probabilities ρ_t for $t > 1$ as the fraction of samples from $\mathcal{L}_{t-1}$ falling into $\mathcal{L}_t$, standard HDR uses an MCMC sampler to simulate from $\mathcal{L}_{t-1}$. If the sequence of nestings is chosen well and initial seeds in the domain $\mathcal{L}_{t-1}$ are known, these samplers achieve a high acceptance rate. This procedure is repeated until $t = T$. With the estimated conditional probabilities $\hat{\rho}_t$, the estimator for the probability mass is then

$$\log \hat{Z} = \sum_{t=1}^{T} \log \hat{\rho}_t. \tag{4}$$

In our adapted version of HDR, the LIN-ESS algorithm (cf. Section 2.1) comes into play, which achieves a 100% acceptance rate for simulating from the nested domains. In order to simulate rejection-free from $\mathcal{L}_t$, LIN-ESS requires an initial sample from the domain $\mathcal{L}_t$, which is obtained from the previous iteration of the algorithm. Every location sampled requires evaluating the linear constraints, hence the cost for each subset in HDR is $\mathcal{O}(NMD)$. Pseudocode for this algorithm is shown in Algorithm 1, where LinESS is a call to the LIN-ESS sampler (cf. Section 2.1 and Algorithm 2 in the appendix) that simulates from the linearly constrained domain.

2.2.2 Obtaining nested domains

As the final missing ingredient, the HDR algorithm requires a sequence of nested domains or level sets defined by positive shifts γ_t, $t = 1, \ldots, T$. In theory, the nested domains should ideally have conditional probabilities of $\rho_t = 1/2 \; \forall t$ (then each nesting improves the precision by one bit). Yet, in a more practical consideration, the computational overhead for constructing the nested domains should also be small. In practice, the shift sequence is often chosen in an ad hoc way, hoping that conditional probabilities are large enough to enable a decently accurate estimation via HDR (Kanjilal & Manohar, 2015). This is not straightforward and requires problem-specific knowledge.

We suggest to construct the nestings via subset simulation (Au & Beck, 2001a) which is very similar to HDR. It only differs in that the conditional probabilities ρ_t are fixed a priori to a value ρ, and then the shift values γ_t are computed such that a fraction ρ of the N samples drawn from $\mathcal{L}_{t-1}$ falls into the subsequent domain $\mathcal{L}_t$.
The construction of the nested domains is depicted in Fig. 2. To find the shifts, N samples are drawn from the integration measure initially (cf. Fig. 2, left). Then the first (and largest) shift γ_1 is determined such that a fraction ρ of the samples fall into the domain $\mathcal{L}_1$. This is achieved by computing for each sample by how

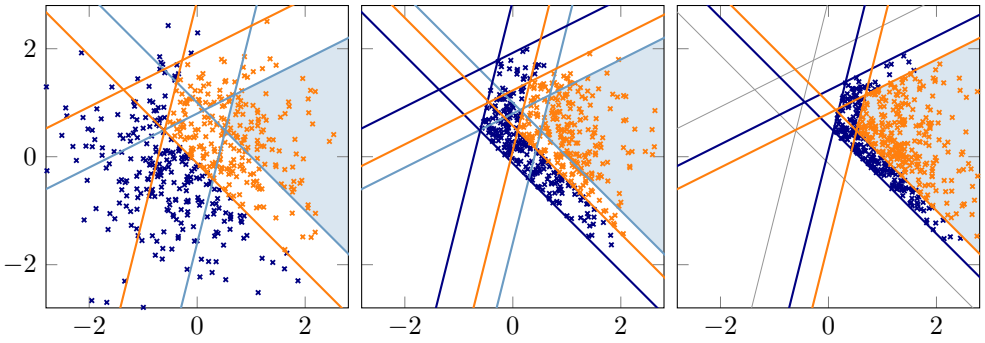

Figure 2: Finding the level sets in subset simulation for linear constraints. *Left:* Draw standard normal samples and find the shift γ_1 for which a fraction ρ of the samples lie inside the new domain (orange lines); *center:* Use LIN-ESS to draw samples from the subsequent domain defined by γ_1 (now in dark blue) and find γ_2 (orange lines) similarly; *right:* Proceed until the domain of interest (shaded area) is reached. Details in text.

much the linear constraints would need to be shifted to encompass the sample. For the subsequent shifts, N samples are simulated from the current domain $\mathcal{L}_t$, and the next shift γ_t is again set s.t. $\lfloor N\rho \rfloor$ samples fall into the next domain $\mathcal{L}_{t+1}$ (Fig. 2, center). This requires an initial sample from $\mathcal{L}_t$ to launch the LIN-ESS sampler, which is obtained from the samples gathered in the previous nesting $\mathcal{L}_{t-1}$ that also lie in $\mathcal{L}_t$, while all other samples are discarded to reduce dependencies. This nesting procedure is repeated until more than $\lfloor N\rho \rfloor$ samples fall into the domain of interest $\mathcal{L}$ (cf. Fig. 2, right). We set $\rho = 1/2$ to maximize the entropy of the binary distribution over whether samples fall in- or outside the next nested domain, yet in reliability analysis a common choice is $\rho = 0.1$ (Au & Beck, 2001b), which has the advantage of requiring less nestings (to the detriment of more samples). Pseudocode can be found in Algorithm 3 in the appendix.

In fact, subset simulation itself also permits the estimation of the integral Z, without appealing to HDR: Since the subsets are constructed such that the conditional probabilities take a predefined value, the estimator for the integral is $\hat{Z}_{\text{ss}} = \rho^{T-1}\rho_T$ where $\rho_T = P(\mathcal{L}_T|\mathcal{L}_{T-1}) \in [\rho, 1]$ is the conditional probability for the last domain. For $\rho = 1/2$ the number of nestings is roughly the negative binary logarithm of the integral estimator $T \approx -\log_2 \hat{Z}_{\text{ss}}$ (cf. Fig. 3). The main reason not to rely on subset simulation alone is that its estimator $\hat{Z}_{\text{ss}}$ is biased, because the samples are both used to construct the domains and to estimate Z. We thus use HDR for the integral estimation and subset simulation for the construction of the level sets.

Both subset simulation and HDR are instances of a wider class of so-called *multilevel splitting* methods which are related to *sequential Monte Carlo* (SMC) in that they are concerned with simulating from a sequence of probability distributions. SMC methods (aka. *particle filters*) were conceived for online inference in state space models, but can be extended to non-Markovian latent variable models (Naesseth et al., 2019). In this form, SMC methods have gained popularity for the estimation of rare events (Del Moral et al., 2006; Bect et al., 2017; Cérou et al., 2012).

2.2.3 Derivatives of Gaussian probabilities

Many applications (e.g. Bayesian optimization, see below) additionally require *derivatives* of the Gaussian probability w.r.t. to parameters λ of the integration measure or the linear constraints. The absence of such derivatives in classic quadrature sub-routines (such as from Genz (1992)) has thus sometimes been mentioned as an argument against them (e.g. Cunningham et al., 2011)). Our method allows to efficiently compute such derivatives, because it can produce samples. This leverages the classic result that derivatives of exponential families with respect to their parameters can be computed from expectations of the sufficient statistics. To do so, it is advantageous to rephrase Eq. (1) as the integral over a *correlated* Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ with axis-aligned constraints (or constraints that are independent of λ). The derivatives w.r.t. a parameter λ can then be expressed as an expected value,

$$\frac{\mathrm{d}\, Z}{\mathrm{d}\,\lambda} = \mathbb{E}\left[\frac{\mathrm{d}\log \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathrm{d}\,\lambda}\right], \tag{5}$$

where the expectation is taken with respect to the transformed integrand Eq. (1). Since LIN-ESS permits us to simulate from the integrand of Eq. (1), derivatives can be estimated via expectations. We demonstrate in Section 3.2 that this is a lot more efficient than finite differences, which requires Z to be estimated twice, and at considerably higher accuracy.

3 EXPERIMENTS

To shed light on the interplay of subset simulation, HDR, and LIN-ESS, we consider a 500-dimensional synthetic integration problem with a closed-form solution. Further 1000-d integrals can be found in Section B.1. We then turn to Bayesian optimization and demonstrate our algorithm's ability to estimate derivatives.

3.1 Synthetic experiments

As an initial integration problem we consider axis-aligned constraints in a 500-dimensional space. Since this task amounts to computing the mass of a shifted orthant under a standard normal distribution, it allows comparison to an exact analytic answer. The goal of this setup is two-fold: 1) to demonstrate that our method can compute small Gaussian probabilities to high accuracy, and 2) to explore configurations for the construction of nested domains using subset simulation. The domain is defined by $\ell(\mathbf{x}) = \prod_{d=1}^{D} \Theta(x_d + 1)$. The true mass of this domain is $3.07 \cdot 10^{-38} = 2^{-124.6}$. Estimating this integral naïvely by sampling from the Gaussian would require of the order of 10^{38} samples for one to fall into the domain of interest. With a standard library like `numpy.random.randn`, this would take about 10^{15} ages of the universe.

Subset simulation First, we compute the shift sequence $\{\gamma_1, \dots, \gamma_T\}$ using subset simulation for various numbers of samples N per subset and a fixed conditional probability of $\rho = 1/2$. Since the contributing factor of each nesting is $\rho = 1/2$, the integral estimate is roughly 2^{-T} for our choice of ρ (cf. Section 2.2.2). The relation between the number of subsets T and the estimated integral value $\hat{Z}_{\text{ss}}$ is visualized in Fig. 3. It shows the sequences of shift values for increasing sample sizes and the resulting integral estimate $\log_2 \hat{Z}_{\text{ss}}$. The T^{th} nesting has shift value $\gamma = 0$ and is the only subset with a conditional probability that deviates from the chosen value of ρ, yet T is a good indicator for the value of the negative binary logarithm of the estimated integral. Hence we use the same axis to display the number of subsets and $-\log_2 \hat{Z}_{\text{ss}}$. The plot highlights the bias of subset simulation: For small sample sizes, e.g. $N = 2, 4, 8$, the integral is severely underestimated. This bias is caused by the dependency of the subset construction method on the samples themselves: Since we are using a MCMC method for simulating from the current domain, samples are correlated and do not fall into the *true* next subset with probability exactly ρ. This is why we only accept every 10^{th} sample to diminish this effect when constructing the subsets. For the subsequent HDR simulation, we accepted every second sample from the ESS procedure.
We choose powers of 2 for the number of samples per

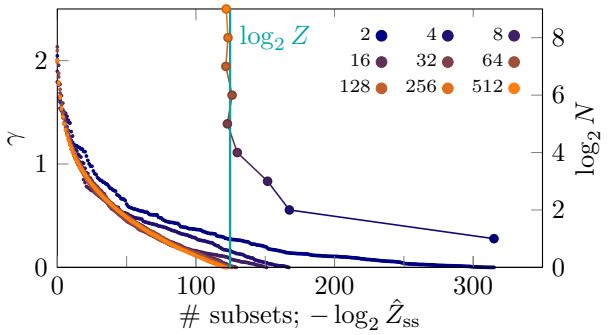

Figure 3: Shift values γ against number of subsets T for different sample size per nesting N (small dots). The connected dots show $-\log_2 \hat{Z}_{\text{ss}}$ vs. $\log_2 N$. The ground truth is indicated by the vertical line. This plot emphasizes the connection between T and $-\log_2 Z$ for $\rho = 1/2$ (see text for details).

subset and observe that as of 16 samples per subset, the subset sequence is good enough to be handed to HDR for more accurate and unbiased estimation. This low requirement of 16 samples per nesting also means that subset simulation is a low-cost preparation for HDR, and causes only minor computational overhead.

Holmes-Diaconis-Ross Fig. 4 shows the results achieved by HDR for the nine subset sequences obtained with 2^1 to 2^9 samples per subset and for different numbers of samples per nesting for HDR. The top left panel of Fig. 4 shows the binary logarithm of the HDR integral estimator. The bad performance for the subsets created with 2, 4, or 8 samples per nesting indicates that a good nesting sequence is essential for the effectiveness of HDR, but also that such a sequence can be found using only about 16 samples per subset (this is thus the number used for all subsequent experiments). The bottom left panel displays the relative error of the HDR estimator. It is to bear in mind that the relative error is $9/11$ if the estimator is one order of magnitude off, indicating that HDR achieves the right order of magnitude with a relatively low sample demand. The right panel of Fig. 4 shows the values for the conditional probabilities found by HDR, using 2^{11} samples per subdomain. If subset simulation were perfectly reliable, these should ideally be $\rho = 1/2$. The plot confirms that, with $N \geq 16$, all conditional probabilities found by HDR are far from 0 and 1, warranting the efficiency of HDR.

3.2 Bayesian optimization

Bayesian optimization is a sample-efficient approach to global optimization of expensive-to-evaluate black-box functions (see Shahriari et al. (2016) for a review). A surrogate over the objective function $f(\mathbf{x})$ serves to build a utility function and ultimately derive a pol-

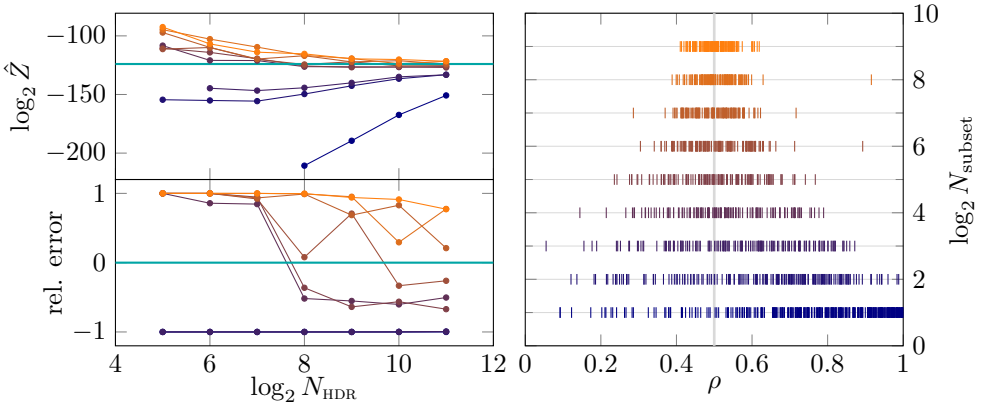

Figure 4: *Left:* HDR integral estimates for different subset sequences (same color coding as in Fig. 3) for 2^5 to 2^{11} samples, *top:* compared to the binary logarithm of the ground truth (horizontal line), and *bottom:* the relative error. *Right:* Conditional probabilities obtained by HDR for the same subset sequences, where $\rho = 1/2$ was chosen for the construction of the subsets (vertical line).

icy to determine the next query point. Information-based utilities are directly concerned with the posterior distribution over the minimizer, $p_{\min}(\mathbf{x} \mid \mathcal{D})$, where $\mathcal{D} = \{\mathbf{x}_n, f(\mathbf{x}_n)\}_{n=1}^N$ summarizes previous evaluations of f. Entropy search (Hennig & Schuler, 2012) seeks to evaluate the objective function at the location that bears the most information about the minimizer. The expression $p_{\min}(\mathbf{x} \mid \mathcal{D})$ is an infinite-dimensional integral itself, but for practical purposes, it can be discretized considering the distribution over so-called *representer points*. The probability of the i^{th} representer point to be the minimum can be approximated as

$$\hat{p}_{\min}(\mathbf{x}_i) = \int d\mathbf{f}\, \mathcal{N}(\mathbf{f}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{j \neq i} \Theta(f(\mathbf{x}_j) - f(\mathbf{x}_i)), \quad (6)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the posterior mean and covariance of the Gaussian process over f, respectively. Clearly, this is a linearly constrained Gaussian integral in the form of Eq. (1) which has to be solved for all N_R representer points. Eq. (6) is stated in matrix form in the appendix Section B.2. The original paper and implementation uses expectation propagation (EP) to approximate this integral.

Probability of minimum For our experiment, we consider the one-dimensional Forrester function (Forrester et al., 2007) with three initial evaluations. The top plot in Fig. 5 shows the ground truth distribution over the minimum obtained by Thompson sampling, i.e., drawing samples from the discretized posterior GP and recording their respective minimum, and the approximation over this distribution obtained by EP. It is apparent that EP fails to accurately represent $\hat{p}_{\min}$. For HDR, we consider four locations (indicated by the vertical lines) and show that while it takes longer to compute, the estimate obtained by HDR converges to the true solution (see bottom plot of Fig. 5). In the experiment we use 200 representer points—which is an unusually high number for a 1-d problem—to show that our method can deal with integrals of that dimension. Also note that we are reporting CPU time, which means that due to automatic parallelization in PYTHON the wall clock time is considerably lower.

Derivatives Entropy search requires derivatives of Eq. (6) to construct a first-order approximation of the predictive information gain from evaluating at a new location $\mathbf{x}_\star$. We can estimate derivatives using expectations (cf. Section 2.2.3 and B.2). Initially we choose 5 representer points to validate the approach of computing derivatives via moments against finite differences. The latter requires estimating $\hat{p}_{\min}$ at very high accuracy and has thus a high sample demand even in this low-dimensional setting, for which we employ both rejection sampling and HDR. The derivatives computed via moments from rejection sampling and LIN-ESS take 0.7% of the time required to get a similar accuracy with finite differences. Unsurprisingly, rejection sampling is faster in this case, with $\hat{p}_{\min}(\mathbf{x}_i) \approx 1/4$, i.e. only $\sim 3/4$ of the samples from the posterior over f need to be discarded to obtain independent draws that have their minimum at $\mathbf{x}_i$. LIN-ESS only outperforms rejection sampling at higher rejection rates common to higher-dimensional problems.
Therefore, we also consider 20 representer points, which corresponds to a 20-d linearly constrained space to sample from. In this setting, we consider a location of low probability, with $\hat{p}_{\min} = 1.6 \cdot 10^{-4}$, which renders an estimation via finite differences impossible and highly disfavors rejection sampling even for computing the moments. LIN-ESS, however, enables us to estimate the gradient of the normal distribution w.r.t. its mean and

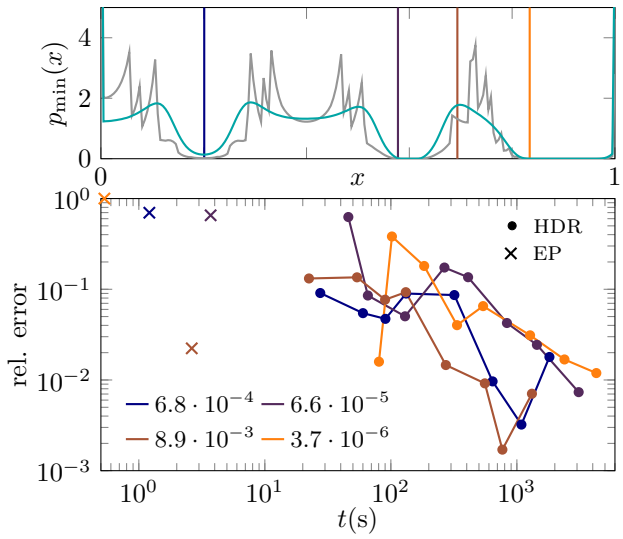

Figure 5: *Top:* Probability for x to be the minimum, estimated via Thompson sampling (blue), and EP (gray). Vertical lines indicate locations at which we run HDR. *Bottom:* Absolute relative error by EP and HDR against CPU time at the locations indicated above. Each HDR sequence shown uses 2^6 to 2^{13} samples per nesting. The smaller $\hat{p}_{\min}$, the longer takes the HDR run, since there are more subsets to traverse.

covariance matrix with a relative standard deviation on the 2-norm of the order of 10^{-2} using $5 \cdot 10^5$ samples and an average CPU time of 325 s for a problem that was previously unfeasible. A badly conditioned covariance matrix in Eq. (5) deteriorates runtime (which is already apparent in the considered case) since it requires estimating moments at very high accuracy to compensate for numerical errors.

3.3 Constrained samples

We emphasize that LIN-ESS allows to draw samples from linearly constrained Gaussians *without rejection*. In the Gaussian process setting, this permits to efficiently draw samples that are subject to linear restrictions (Agrell, 2019; López-Lopera et al., 2017; Da Veiga & Marrel, 2012). In particular, the time required for sampling is essentially independent of the probability mass of the domain of interest. This probability mass only affects the precomputation required to find an initial sample in the domain for LIN-ESS (cf. Section 2.2.2). Since this can be achieved with $\sim$16 samples per subset (cf. Section 3.1), this initial runtime is typically negligible compared to the actual sampling. Fig. 6 displays the posterior distribution of a GP conditioned on the location of the minimum from the Bayesian optimization context, estimated from LIN-ESS samples. This distribution is required in predictive entropy search

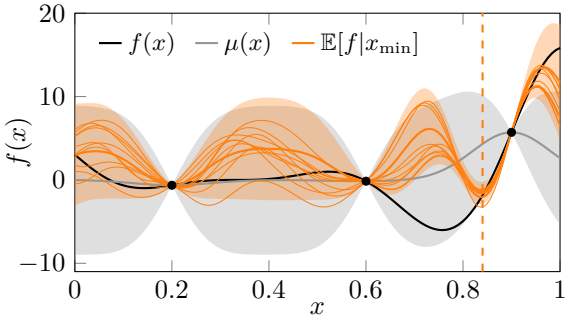

Figure 6: The Forrester function (black), the posterior GP given three evaluations (gray), and the posterior distribution over f conditioned on the minimum being located at where the vertical line indicates (orange), each with the 2σ confidence interval shaded. The latter has been obtained from drawing 10^5 samples using LIN-ESS, 10 of which are shown (thin orange lines).

(Hernández-Lobato et al., 2014)—a reformulation of the original entropy search—where it is approximated by imposing several related constraints (e.g., on the derivatives at the minimizer $\mathbf{x}_{\min}$). The probability for the given location to be the minimizer is $\lesssim 10^{-6}$, which renders direct sampling virtually impossible. The unaltered ESS algorithm fails on this problem due to the domain selector—a binary likelihood.

4 CONCLUSIONS

We have introduced a black-box algorithm that computes Gaussian probabilities (i.e. the integral over linearly constrained Gaussian densities) with high numerical precision, even if the integration domain is of high dimensionality and the probability to be computed is very small. This was achieved by adapting two separate pieces of existing prior art and carefully matching them to the problem domain: We designed a special version of elliptical slice sampling that takes explicit advantage of the linearly-constrained Gaussian setting, and used it as an internal step of the HDR algorithm. We showed that, because this algorithm can not just compute integrals but also produces samples from the nestings alongside, it also permits the evaluation of derivatives of the integral with respect to the parameters of the measure. One current limitation is that, because our algorithm was designed to be unbiased, it has comparably high computational cost (but also superior numerical precision) over alternatives like expectation propagation. This problem could be mitigated if one is willing to accept unbiasedness and thus reuse samples. Furthermore, both HDR and LIN-ESS are highly parallelizable (as opposed to EP) and thus offer margin for implementational improvement.

Acknowledgements

AG and PH gratefully acknowledge financial support by the European Research Council through ERC StG Action 757275 / PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. The work was carried out while OK was at the University of Tuebingen, funded by the German Research Foundation (Research Unit 1735). OK also acknowledges financial support through the Alexander von Humboldt Foundation. AG is grateful to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support.

References

Agrell, C. (2019). "Gaussian processes with linear operator inequality constraints". In: *Journal of Machine Learning Research* 20.135, pp. 1–36.

Andersen, L. N., P. J. Laub, & L. Rojas-Nandayapa (2018). "Efficient simulation for dependent rare events with applications to extremes". In: *Methodology and Computing in Applied Probability* 20.1, pp. 385–409.

Ashford, J. R. & R. R. Sowden (1970). "Multi-variate probit analysis". In: *Biometrics*, pp. 535–546.

Au, S.-K. & J. L. Beck (2001a). "Estimation of small failure probabilities in high dimensions by subset simulation". In: *Probabilistic Engineering Mechanics* 16.4, pp. 263–277.

Au, S.-K. & J. L. Beck (2001b). "First excursion probabilities for linear systems by very efficient importance sampling". In: *Probabilistic Engineering Mechanics* 16.3, pp. 193–207.

Azzimonti, D. & D. Ginsbourger (Aug. 2017). "Estimating orthant probabilities of high-dimensional Gaussian vectors with an application to set estimation". In: *Journal of Computational and Graphical Statistics* 27.2, pp. 255–267.

Bect, J., L. Li, & E. Vazquez (2017). "Bayesian subset simulation". In: *SIAM/ASA Journal on Uncertainty Quantification* 5.1, pp. 762–786.

Bolin, D. & F. Lindgren (Jan. 2015). "Excursion and contour uncertainty regions for latent Gaussian models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1, pp. 85–106.

Botev, Z. I. (Feb. 2016). "The normal law under linear restrictions: simulation and estimation via minimax tilting". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.1, pp. 125–148.

Cérou, F., P. Del Moral, T. Furon, & A. Guyader (2012). "Sequential Monte Carlo for rare event estimation". In: *Statistics and Computing* 22.3, pp. 795–908.

Chen, Y.-I. & Y.-M. Chang (2007). "Identification of the minimum effective dose for right-censored survival data". In: *Computational Statistics & Data Analysis* 51.6, pp. 3213–3222.

Cong, Y., B. Chen, & M. Zhou (Dec. 2017). "Fast simulation of hyperplane-truncated multivariate normal distributions". In: *Bayesian Anal.* 12.4, pp. 1017–1037.

Craig, P. (Feb. 2008). "A new reconstruction of multivariate normal orthant probabilities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 227–243.

Cunningham, J. P., P. Hennig, & S. Lacoste-Julien (Nov. 2011). "Gaussian probabilities and expectation propagation". In: *arXiv e-prints*, arXiv:1111.6832, arXiv:1111.6832.

Da Veiga, S. & A. Marrel (2012). "Gaussian process modeling with inequality constraints". en. In: *Annales de la Faculté des sciences de Toulouse : Mathématiques* Ser. 6, 21.3, pp. 529–555.

Del Moral, P., A. Doucet, & A. Jasra (2006). "Sequential Monte Carlo samplers". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.3, pp. 411–436.

Diaconis, P. & S. Holmes (1995). "Three examples of Monte-Carlo Markov chains: at the interface between statistical computing, computer science, and statistical mechanics". In: *Discrete Probability and Algorithms*. New York, NY: Springer New York, pp. 43–56.

Fisac, J. F. et al. (2018). "A general safety framework for learning-based control in uncertain robotic systems". In: *IEEE Transactions on Automatic Control*.

Forrester, A. I. J., A. Sóbester, & A. J. Keane (2007). "Multi-fidelity optimization via surrogate modelling". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 463.2088, pp. 3251–3269.

French, J. P. & S. R. Sain (Sept. 2013). "Spatio-temporal exceedance locations and confidence regions". In: *Ann. Appl. Stat.* 7.3, pp. 1421–1449.

Gassmann, H. I., I. Deák, & T. Szántai (2002). "Computing multivariate normal probabilities: a new look". In: *Journal of Computational and Graphical Statistics* 11.4, pp. 920–949.

Genton, M. G., D. E. Keyes, & G. Turkiyyah (2018). "Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities". In: *Journal of Computational and Graphical Statistics* 27.2, pp. 268–277.

Genton, M. G., Y. Ma, & H. Sang (2011). "On the likelihood function of Gaussian max-stable processes". In: *Biometrika*, pp. 481–488.

Genz, A. (1992). "Numerical computation of multivariate normal probabilities". In: *Journal of Computational and Graphical Statistics* 1.2, pp. 141–149.

Genz, A. (Aug. 2004). "Numerical computation of rectangular bivariate and trivariate normal and t probabilities". In: *Statistics and Computing* 14.3, pp. 251–260.

Genz, A. & F. Bretz (2009). *Computation of multivariate normal and t probabilities*. Vol. 195. Springer Science & Business Media.

Geweke, J. (1991). "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities". In: *Computing science and statistics: Proceedings of the 23rd symposium on the interface*. Fairfax, Virginia: Interface Foundation of North America, Inc, pp. 571–578.

Hayter, A. J. & Y. Lin (Sept. 2012). "The evaluation of two-sided orthant probabilities for a quadrivariate normal distribution". In: *Computational Statistics* 27.3, pp. 459–471.

Hayter, A. J. & Y. Lin (2013). "The evaluation of trivariate normal probabilities defined by linear inequalities". In: *Journal of Statistical Computation and Simulation* 83.4, pp. 668–676.

Hennig, P. & C. Schuler (June 2012). "Entropy search for information-efficient global optimization". In: *Journal of Machine Learning Research* 13, pp. 1809–1837.

Hernández-Lobato, J. M., M. W. Hoffman, & Z. Ghahramani (2014). "Predictive entropy search for efficient global optimization of black-box functions". In: *Advances in Neural Information Processing Systems*, pp. 918–926.

Huser, R. & A. C. Davison (Feb. 2013). "Composite likelihood estimation for the Brown–Resnick process". In: *Biometrika* 100.2, pp. 511–518.

Joe, H. (1995). "Approximations to multivariate normal rectangle probabilities based on conditional expectations". In: *Journal of the American Statistical Association* 90.431, pp. 957–964.

Kanjilal, O. & C. Manohar (2015). "Markov chain splitting methods in structural reliability integral estimation". In: *Probabilistic Engineering Mechanics* 40, pp. 42–51.

Koch, H. & G. P. Bopp (2019). *Fast and exact simulation of multivariate normal and wishart random variables with box constraints*.

Kroese, D. P., T. Taimre, & Z. I. Botev (2011). *Handbook of Monte Carlo methods*. Wiley New Jersey.

Lawrence, E., D. Bingham, C. Liu, & V. N. Nair (2008). "Bayesian inference for multivariate ordinal data using parameter expansion". In: *Technometrics* 50.2, pp. 182–191.

Liao, X., H. Li, & L. Carin (2007). "Quadratically gated mixture of experts for incomplete data classification". In: *Proceedings of the 24th International Conference on Machine learning*. ACM, pp. 553–560.

López-Lopera, A. F., F. Bachoc, N. Durrande, & O. Roustant (2017). *Finite-dimensional Gaussian approximation with linear inequality constraints*.

López-Lopera, A. F., S. John, & N. Durrande (2019). *Gaussian process modulated Cox processes under linear inequality constraints*.

Melchers, R. E. & A. T. Beck (2018). *Structural reliability analysis and prediction*. John Wiley & Sons.

Miwa, T., A. J. Hayter, & S. Kuriki (2003). "The evaluation of general non-centred orthant probabilities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 223–234.

Mulgrave, J. J. & S. Ghosal (2018). "Bayesian inference in nonparanormal graphical models". In: *Bayesian Anal.*

Murray, I., R. Adams, & D. MacKay (13–15 May 2010). "Elliptical slice sampling". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. Vol. 9. PMLR, pp. 541–548.

Naesseth, C. A., F. Lindsten, & T. B. Schön (2019). *Elements of sequential Monte Carlo*.

Neal, R. M. (June 2003). "Slice sampling". In: *Ann. Statist.* 31.3, pp. 705–767.

Nomura, N. (2014). "Computation of multivariate normal probabilities with polar coordinate systems". In: *Journal of Statistical Computation and Simulation* 84.3, pp. 491–512.

Nomura, N. (Jan. 2016). "Evaluation of Gaussian orthant probabilities based on orthogonal projections to subspaces". In: *Statistics and Computing* 26.1, pp. 187–197.

Phinikettos, I. & A. Gandy (2011). "Fast computation of high-dimensional multivariate normal probabilities". In: *Computational Statistics & Data Analysis* 55.4, pp. 1521–1529.

Rasmussen, C. E. & C. K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, p. 248.

Ross, S. (2012). *Simulation*. Knovel Library. Elsevier Science.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, & N. de Freitas (2016). "Taking the human out of the loop: a review of Bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175.

Straub, D., R. Schneider, E. Bismut, & H.-J. Kim (2020). "Reliability analysis of deteriorating structural systems". In: *Structural Safety* 82, p. 101877.

Su, Q., X. Liao, C. Chen, & L. Carin (20–22 Jun 2016). "Nonlinear statistical learning with truncated Gaussian graphical models". In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. PMLR, pp. 1948–1957.

Thiébaut, R. & H. Jacqmin-Gadda (2004). "Mixed models for longitudinal left-censored repeated measures". In: *Computer Methods and Programs in Biomedicine* 74.3, pp. 255–260.

Vijverberg, W. P. (1997). "Monte Carlo evaluation of multivariate normal probabilities". In: *Journal of Econometrics* 76.1, pp. 281–307.

Wang, J., S. C. Clark, E. Liu, & P. I. Frazier (2016). "Parallel Bayesian global optimization of expensive functions". In: *arXiv preprint arXiv:1602.05149*.

Wani, O., A. Scheidegger, J. P. Carbajal, J. Rieckermann, & F. Blumensaat (2017). "Parameter estimation of hydrologic models using a likelihood function for censored and binary observations". In: *Water Research* 121, pp. 290–301.