Entropic Trace Estimates for Log Determinants

Jack Fitzsimons[1], Diego Granziol[1], Kurt Cutajar[2]
Michael Osborne[1], Maurizio Filippone[2], and Stephen Roberts[1]

[1] Department of Engineering, University of Oxford, UK
{jack,diego,mosb,sjrob}@robots.ox.ac.uk
[2] Department of Data Science, EURECOM, France
{kurt.cutajar,maurizio.filippone}@eurecom.fr

Abstract. The scalable calculation of matrix determinants has been a bottleneck to the widespread application of many machine learning methods such as determinantal point processes, Gaussian processes, generalised Markov random fields, graph models and many others. In this work, we estimate log determinants under the framework of maximum entropy, given information in the form of moment constraints from stochastic trace estimation. The estimates demonstrate a significant improvement on state-of-the-art alternative methods, as shown on a wide variety of matrices from the SparseSuite Matrix Collection. By taking the example of a general Markov random field, we also demonstrate how this approach can significantly accelerate inference in large-scale learning methods involving the log determinant.

1 Introduction

Scalability is a key concern for machine learning in the big data era, whereby inference schemes are expected to yield optimal results within a constrained computational budget. Underlying these algorithms, linear algebraic operations with high computational complexity pose a significant bottleneck to scalability, and the log determinant of a matrix [5] falls firmly within this category of operations. The canonical solution involving Cholesky decomposition [16] for a general $n \times n$ positive definite matrix, A, entails time complexity of $\mathcal{O}(n^3)$ and storage requirements of $\mathcal{O}(n^2)$, which is infeasible for large matrices. Consequently, this term greatly hinders widespread use of the learning models where it appears, which includes determinantal point processes [24], Gaussian processes [31], and graph problems [36].

The application of kernel machines to vector valued input data has gained considerable attention in recent years, enabling fast linear algebra techniques. Examples include Gaussian Markov random fields [32] and Kronecker-based algebra [33], while similar computational speed-ups may also be obtained for sparse matrices. Nonetheless, such structure can only be expected in selected applications, thus limiting the widespread use of such techniques.

In light of this computational constraint, several approximate inference schemes have been developed for estimating the log determinant of a matrix more efficiently. Generalised approximation schemes frequently build upon iterative stochastic trace estimation techniques [4]. This includes polynomial approximations such as Taylor and Chebyshev expansions [2, 20]. Recent developments shown to outperform the aforementioned approximations include estimating the trace using stochastic Lanczos quadrature [35], and a probabilistic numerics approach based on Gaussian process inference which incorporates bound information [12]. The latter technique is particularly significant as it introduces the possibility of quantifying the numerical uncertainty inherent to the approximation.

In this paper, we present an alternative probabilistic approximation of log determinants rooted in information theory, which exploits the relationship between stochastic trace estimation and the moments of a matrix's eigenspectrum. These estimates are used as moment constraints on the probability distribution of eigenvalues. This is achieved by maximising the entropy of the probability density $p(\lambda)$ given our moment constraints. In our inference scheme, we circumvent the issue inherent to the Gaussian process approach [12], whereby positive probability mass may occur in the region of negative densities. In contrast, our proposed entropic approach implicitly encodes the constraint that densities are necessarily positive. Given equivalent moment information, we achieve competitive results on matrices obtained from the SuiteSparse Matrix Collection [11] which consistently outperform competing approximations to the log-determinant [12, 7].

The most significant contributions of this work are listed below.[3]

1. We develop a novel approximation to the log determinant of a matrix which relies on the principle of maximum entropy enhanced with moment constraints derived from stochastic trace estimation.
2. We present the theory motivating the use of maximum entropy for solving this problem, along with insights on why we expect particularly significant improvements over competing techniques for large matrices.
3. We directly compare the performance of our entropic approach to other state-of-the-art approximations to the log-determinant. This evaluation covers real sparse matrices obtained from the SuiteSparse Matrix Collection [11].
4. Finally, to showcase how the proposed approach may be applied in a practical scenario, we incorporate our approximation within the computation of the log-likelihood term of a Gaussian Markov random field, where we obtain a significant increase in speed.

1.1 Related Work

The methodology presented in this work predominantly draws inspiration from the recently introduced probabilistic numerics approach to estimating the log determinant of a matrix [12]. In that work, the computation of the log determinant

[3] Code for algorithms proposed in this paper are available at https://github.com/OxfordML/EntropicTraceEstimation

is reinterpreted as a probabilistic estimation problem, whereby results obtained from budgeted computations are used to infer accurate estimates for the log determinant. In particular, within that proposed framework, the eigenvalues of a matrix A are modelled from noisy observations of $\mathrm{Tr}(A^k)$ obtained from stochastic trace estimation [4] using the Taylor approximation method. By modelling such noisy observations using a Gaussian process [31], Bayesian quadrature [29] can then be invoked for making predictions on the infinite series of the Taylor expansion, and in turn estimating the log determinant. Of particular interest is the uncertainty quantification inherent to this approach, which is a notable step forward in the direction of measuring the complete numerical uncertainty associated with approximating large-scale inference models. The estimates obtained using this Bayesian set-up may be further improved by considering known upper and lower bounds on the value of the log determinant [5]. In this paper, we provide an alternative to this approach by interpreting the observed moments as being constraints on the probability distribution of eigenvalues underlying the computation of the log determinant. As we shall explore, our novel entropic formulation makes better calibrated prior assumptions than the previous work, and consequently yields superior performance.

More traditional approaches to approximating the log determinant build upon iterative algorithms, and exploit the fact that the log determinant may be rewritten as the trace of the logarithm of the matrix. This features in both the Chebyshev expansion approximation [20], as well as the widely-used Taylor series approximation upon which the aforementioned probabilistic inference approaches are built. Recently, an approximation to the log determinant using stochastic Lanczos quadrature [35] has been shown to outperform the aforementioned polynomial approaches, while also providing probabilistic error bounds. Finally, given that the logarithm of a matrix often appears multiplied by a vector (for example the log likelihood term of a Gaussian process [31]), the spline approximation proposed in [10] may be used to accelerate computation.

2 Background

In this section, we shall formally introduce the concepts underlying the proposed maximum entropy approach to approximating the log determinant. We start by describing stochastic trace estimation and demonstrate how this can be applied to estimating the trace term of matrix powers. Subsequently, we illustrate how the latter terms correspond to the raw moments of the matrix's eigenspectrum, and show how the log determinant may be inferred from the distribution of eigenvalues constrained by such moments.

2.1 Stochastic Trace Estimation

Estimating the trace of implicit matrices is a central component of many approaches to approximate the log determinant of a matrix. Stochastic trace estimation [4] builds a Monte Carlo estimate of the trace of a matrix A by repeatedly multiplying it by *probing vectors* $\mathbf{z}$,

$$\mathrm{Tr}(A) \approx \frac{1}{m}\sum_{i=1}^{m} \mathbf{z}_i^T A \mathbf{z}_i,$$

such that the expectation of $\mathbf{z}_i\mathbf{z}_i^T$ is the identity, $\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^T] = I$. This can be readily verified using the expectation of $\mathrm{Tr}(\mathbf{z}_i^T A\mathbf{z}_i)$ by exploiting the cyclical property of the trace operation. As such, many choices of how to sample the probing vectors have emerged. Possibly the most naïve choice involves sampling from the columns of the identity matrix; however, due to poor expected sample variance this is not widely used in the literature. Sampling from vectors on the unit hyper-sphere, and correspondingly sampling normal random vectors (Gaussian estimator), significantly reduces the sample variance, but more random bits are required to generate each sample. A major progression for stochastic trace estimation was the introduction of Hutchinson's method [21], which sampled each element as a Bernoulli random variable requiring only a linear number of random bits, while also reducing the sample variance even further. A more recent approach involves sampling from sets of mutually unbiased bases (MUBs) [13], requiring only a logarithmic number of bits. Table 1 (adapted from [13]) provides a concise overview of the landscape of probing vectors.

Table 1. Comparison of single shot variance V, worst case single shot variance V^{worst} and number of random bits R required for commonly used trace estimators and the MUBs estimator. (* *required for floating point precision*)

	V	V^{worst}	R
Fixed basis estimator	$d\sum_{i=1}^{d} M_{ii}^2 - \mathrm{Tr}(A)^2$	$(d-1)\mathrm{Tr}(A)^2$	$\log_2(d)$
MUBs estimator	$\frac{d}{d+1}\mathrm{Tr}(A^2) - \frac{1}{d+1}\mathrm{Tr}(A)^2$	$\frac{d-1}{d+1}\mathrm{Tr}(A^2)$	$2\log_2(d)$
Hutchinson estimator	$2\left(\mathrm{Tr}(A^2) - \sum_{i=1}^{d} A_{ii}^2\right)$	$\frac{2(d-1)}{d}\mathrm{Tr}(A^2)$	d
Gaussian estimator	$2\mathrm{Tr}(A^2)$	$2\mathrm{Tr}(A^2)$	$\mathcal{O}(d)^*$

A notable application of stochastic trace estimation is the approximation of the trace term for matrix powers, $\mathrm{Tr}(A^k)$. Stochastic trace estimation enables vector-matrix multiplications to be propagated right to left, costing $\mathcal{O}(n^2)$, rather than the $\mathcal{O}(n^3)$ complexity required by matrix multiplication. This simple trick has been applied in several domains such as counting the number of triangles in graphs [3], string pattern matching [1] and of course estimating the log determinant of matrices, as discussed in this work.

2.2 Raw Moments of the Eigenspectrum

The relation between the raw moments of the eigenvalue distribution and the trace of matrix powers allows us to exploit stochastic trace estimation for estimating the log determinant. Raw moments are defined as the mean of the random variables raised to integer powers. Given that the function of a matrix is implicitly applied to its eigenvalues, in the case of matrix powers this corresponds to raising the eigenvalues to a given power. For example, the k^{th} raw moment of the distribution over the eigenvalues (a mixture of Dirac delta functions) is $\sum_{i=1}^{m} \lambda^k p(\lambda)$, where $p(\lambda)$ is the distribution of eigenvalues. The first few raw moments of the eigenvalues are trivial to compute. Denoting the k^{th} raw moment as $\mathbb{E}[\lambda^k]$, we have that $\mathbb{E}[\lambda^0] = 1$, $\mathbb{E}[\lambda^1] = \frac{1}{n}\text{Tr}(A)$ and $\mathbb{E}[\lambda^2] = \frac{1}{n}\sum_{i,j} A_{i,j}^2$. More generally, the k^{th} raw moment can be formulated as $\mathbb{E}[\lambda^k] = \frac{1}{n}\text{Tr}(A^k)$, which can be estimated using stochastic trace estimation. These identities can be easily derived using the definitions and well known identities of the trace term and Frobenius norm.

2.3 Approximating the Log Determinant

In view of the relation presented in the previous subsection, we can reformulate the log determinant of a matrix in terms of its eigenvalues using the following derivation:

$$\log\big(\text{Det}(A)\big) = \sum_{i=1}^{n} \log(\lambda_i) := n\mathbb{E}\left[\log(\lambda)\right] \approx n \int p(\lambda)\log(\lambda)\text{d}\lambda, \tag{1}$$

where the approximation is introduced due to our estimation of $p(\lambda)$, the probability distribution of eigenvalues. If we knew the true distribution of $p(\lambda)$ it would hold with equality.

Given that we can obtain information about the moments of $p(\lambda)$ through stochastic trace estimation, we can solve this integral by employing the principle of maximum entropy, while treating the estimated moments as constraints. While not explored in this work, it is worth noting that in the event of moment information combined with samples of eigenvalues, we would use the method of maximum relative entropy with data constraints, which is in turn a generalisation of Bayes' rule [9]. This can be applied, for example, in the quantum linear algebraic setting [28].

3 Estimating the Log Determinant using Maximum Entropy

The maximum entropy method (MaxEnt) [30] is a procedure for generating the most conservatively uncertain estimate of a probability distribution possible with the given information, which is particularly valued for being maximally noncommittal with regard to missing information [22]. In particular, to determine a probability density $p(\boldsymbol{x})$, this corresponds to maximising the functional

$$S = -\int p(\boldsymbol{x})\log p(\boldsymbol{x})\text{d}\boldsymbol{x} - \sum_i \alpha_i \left[\int p(\boldsymbol{x}) f_i(\boldsymbol{x})\text{d}\boldsymbol{x} - \mu_i\right] \tag{2}$$

with respect to $p(\boldsymbol{x})$, where $\mathbb{E}[f_i(\boldsymbol{x})] = \mu_i$ are given constraints on the probability density. In our case, each μ_i constraint denotes the stochastic trace estimate of the i^{th} raw moment of the matrix eigenvalues. The first term in the above equation is referred to as the Boltzmann-Shannon-Gibbs (BSG) entropy, which has been applied in multiple fields, ranging from condensed matter physics [15] to finance [27, 8]. Along with its path equivalent, maximum caliber [17], it has been successfully used to derive statistical mechanics [18], non-relativistic quantum mechanics, Newton's laws and Bayes' rule [17, 9]. Under the axioms of consistency, uniqueness, invariance under coordinate transformations, sub-set and system independence, it can be proved that for constraints in the form of expected values, drawing self-consistent inferences requires maximising the entropy [34, 30]. Crucial for our investigation are the functional forms $f_i(\boldsymbol{x})$ of constraints for which the method of maximum entropy is appropriate. The axioms of Johnson and Shore [34] assert that the entropy must have a unique maximum and that the BSG entropy is convex. The entropy hence has a unique maximum provided that the constraints are convex. This is satisfied for any polynomial in $\boldsymbol{x}$ and hence, maximising the entropy given moment constraints constitutes a self-consistent inference scheme [30].

3.1 Implementation

Our implementation of the system follows straight from stochastic trace estimation to estimate the raw moments of the eigenvalues, maximum entropy distribution given these moments and, finally, determining the log of the geometric mean of this distribution. The log geometric mean is an estimate of the log determinant divided by the dimensionality of $A \in \mathbb{R}^{n \times n}$. We explicitly step through the subtleties of the implementation in order to guide the reader through the full procedure.

By taking the partial derivatives of S from Equation (2), it is possible to show that the maximum entropy distribution given moment information is of the form

$$p(\lambda) = \exp(-1 + \sum_i \alpha_i \mu_i).$$

The goal is to find the set of α_i which match the raw moments of $p(\lambda)$ to the observed moments $\{\mu_i\}$. While this *may* be performed symbolically, this becomes intractable for larger number of moments, and our experience with current symbolic libraries [25, 37] is that they are not extendable beyond more than 3 moments. Instead, we turn our attention to numerical optimisation. Early approaches to optimising maximum entropy coefficients worked well for a small number of coefficients but became highly unstable as the number of observed moments grew [26]. However, building on these concepts, more stable approaches emerged [6]. Algorithm 1 outlines a stable approach to this optimisation under the conditions that λ_i is strictly positive and the moments lie between zero and one. We can satisfy these conditions by normalising our positive definite matrix by the maximum of the Gershgorin intervals [14].

Algorithm 1 Optimising the Coefficients of the MaxEnt Distribution

```
Input: Moments $\{\mu_i\}$, Tolerance $\epsilon$
Output: Coefficients $\{\alpha_i\}$
$\alpha_i \sim \mathcal{N}(0,1)$
$i \leftarrow 0$
$p(\lambda) \leftarrow \exp(-1 - \sum_k \alpha_k \lambda^k)$
while error $< \epsilon$ do
    $\delta \leftarrow \log\left(\frac{\mu_i}{\int \lambda^i p(\lambda) \mathrm{d}\lambda}\right)$
    $\alpha_i \leftarrow \alpha_i + \delta$
    $p(\lambda) \leftarrow p(\lambda|\alpha)$
    error $\leftarrow \max |\int \lambda^i p(\lambda)\mathrm{d}\lambda - \mu_i|$
    $i \leftarrow \mathrm{mod}(i+1, \mathrm{length}(\mu))$
```

Given Algorithm 1, the pipeline of our approach can be pieced together. First, the raw moments of the eigenvalues are estimated using stochastic trace estimation. These moments are then passed to the maximum entropy optimisation algorithm to produce an estimate of the distribution of eigenvalues, $p(\lambda)$. Finally, $p(\lambda)$ is used to estimate the log geometric mean of the distribution, $\int \log(\lambda)p(\lambda)d\lambda$. This term is multiplied by the dimensionality of the matrix and if the matrix is normalised, the log of this normalisation term is added again. These steps are laid out more concisely in Algorithm 2.

Algorithm 2 Entropic Trace Estimation for Log Determinants

```
Input: PD Symmetric Matrix $A$, Order of stochastic trace estimation $k$, Tolerance $\epsilon$
Output: Log Determinant Approximation $\log|A|$
$B = A/\|A\|_2$
$\mu$ (moments)$\leftarrow$ StochasticTraceEstimation$(B, k)$
$\alpha$ (coefficients) $\leftarrow$ MaxEntOpt$(\mu, \epsilon)$
$p(\lambda) \leftarrow p(\lambda|\alpha)$
$\log|A| \leftarrow n \int \log(\lambda)p(\lambda)\mathrm{d}\lambda + n\log(\|A\|_2)$
```

4 Insights for Large Matrices

The method of entropic trace estimation has the interesting property where we expect the relative error to decrease as the matrix size N increases. Colloquially, we can liken maximum entropy to a maximum likelihood over distributions, where this likelihood functional is raised to the number of eigenvalues in the matrix. This corresponds to the number of particles in the system, in traditional particle physics parlance. Given that there is a global maximum, as the number of eigenvalues increases the functional tends to a delta functional around the $p(\boldsymbol{x})$ of maximum entropy. This confirms that within the scope of our problem's continuous distribution over eigenvalues, whenever the number of eigenvalues (and correspondingly the dimensionality of the matrix) tends towards infinity, we expect the maximum entropy solution to converge to the true solution. This gives further credence to the suitability of our method when applied to large matrices. We substantiate this claim by delving into the fundamentals of maximum entropy for physical systems and extending the analogy to functionals over the space of densities. We show that in the limit of $N \to \infty$ the maximum entropy distribution dominates the space of solutions satisfying the constraints. We demonstrate the practical significance of this assertion by setting up an experiment using synthetically constructed random matrices, where this is in fact verified.

4.1 Law of Large Numbers for Maximum Entropy in Matrix Methods

In order to demonstrate our result, we consider the quantity W, which represents the number of ways in which our candidate probability distribution can recreate the observed moment information. In order to make this quantity finite we consider the discrete distribution characterised by machine precision ϵ. We show that $W = \exp(NS)$, where S is the entropy. Hence maximising the entropy is equivalent to maximising W, as N is fixed. In the continuous limit, we consider the ratio of two such terms $F_i = W_i / \sum_j W_j$, which is also finite. We consider this quantity F_i to represent the probability of a candidate solution i occurring, given the space of all possible solutions. We further show in the discrete and continuous space that for large N, the candidate distribution maximising S occurs with probability 1.

Consider the analogy of having a physical system made up of particles. The different ways, W, in which we can organise this system of N particles with T distinguishable groups each containing n_t particles, can be expressed as the combinatorial

$$W = \frac{N!}{\prod_{t=1}^{T} n_t!}, \tag{3}$$

where $\sum_t n_t = N$. If we consider the logarithm of the above term, we can invoke Stirling's approximation that $\log(n!) \approx n\log(n) - n$, which is exact in the limits $N \to \infty$ and $n_i \to \infty$. Using this relation, we obtain

$$\log W = N\bigg(- \sum_{t=1}^{T} \frac{n_t}{N} \log\left[\frac{n_t}{N}\right]\bigg) = NS, \tag{4}$$

where S is the Boltzmann-Shannon-Gibbs entropy and in the continous case we identify $p(t) = n_t/N$, where $p(t)$ represents the probability of being in group t. Hence, $W = \exp(NS)$.

The number of formulations, W_{maxent}, in which the maximum entropy realisation is more probable than any other realisation can be succinctly expressed as

$$\frac{W_{\text{maxent}}}{W_{\text{other}}} = \exp\bigl(N(S_{\text{maxent}} - S_{\text{other}})\bigr), \tag{5}$$

which exposes that in the limit of large N, the maximum entropy solution dominates other solutions satisfying the same constraints. More significantly, we can also show that it dominates the space of *all* solutions. Let $\sum_i W_i$ denote the total number of ways in which the system can be configured for all possible underlying densities satisfying the constraints.

If we consider the ratio between this term and the number of ways the maximum entropy distribution can be configured, we observe that

$$\frac{W_{\text{maxent}}}{\sum_i W_i} = \frac{\exp(NS(P_{\text{maxent}}))}{\sum_i \exp(NS_i)} = \frac{1}{\sum_i \exp\left(N(S(P_i(\boldsymbol{x})) - S(P_{\text{maxent}}(\boldsymbol{x})))\right)} \xrightarrow{N\to\infty} 1, \tag{6}$$

with $S\left(P(\boldsymbol{x})\right)$ denoting the entropy of the probability of a probability distribution $P(\boldsymbol{x})$ and where we have exploited the fact that one of the S_i is S_{max} and that $S_{i\neq j} < S_{\text{max}}$. More formally, we consider the probability mass about maxima in the functional describing all possible configurations, which is characterised via their entropy, S:

$$W_{\text{total}} = \int \exp(NS)[\mathcal{D}P] = \int \cdots \int_{-\infty}^{\infty} \exp(NS) \prod_x \mathrm{d}P(\boldsymbol{x}). \tag{7}$$

When $N \to \infty$, the maximum value of S accounts for the majority of the integral's mass. To see this consider the ratio of functional integrals,

$$\frac{W_{\text{maxent}}}{W_{\text{total}}} = \frac{\int \exp\left(NS(P_{\text{maxent}}(\boldsymbol{x}))\right) \mathrm{d}P_{\text{maxent}}(\boldsymbol{x})}{\int_{-\infty}^{\infty} \exp\left(NS(P(\boldsymbol{x}))\right) \prod_x \mathrm{d}P(\boldsymbol{x})}, \tag{8}$$

which tends to 1 as $N \to \infty$. The argument is the continuous version of that which is displayed in Equation (6), while the convergence to 1 as N approaches infinity follows directly from Laplace's method in the multivariate case, as well as the definition of a probability density and the functional integral.

Laplace's method gives a theoretical basis for the canonical distributions in statistical mechanics. Its equivalent in the complex space, the method of steepest descent, in Feynman's path integral formulation, shows that points in the vicinity of the extrema of the action functional (the classical mechanical solution), contribute maximally to the path integral. This is the corresponding result for the matrix eigenvalue distributions.

4.2 Validation on Synthetic Data

We generate random, diagonally dominant positive semi-definitive matrices, M, which are constructed as

$$M = \frac{A^\top A}{||A^\top A||_2} + I_N, \tag{9}$$

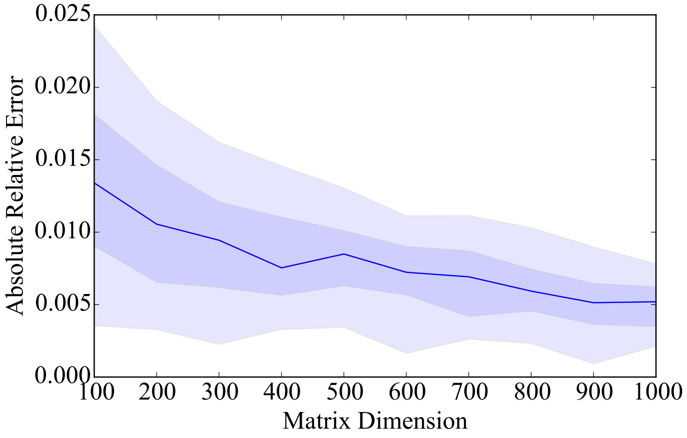

Fig. 1. The absolute relative error with respect to matrix dimensionality. Plotted is the median error, with the 30-70 and 10-90 percentile regions shaded in dark and light blue respectively.

where $A \in \mathbb{R}^{N \times N}$ is an $N \times N$ matrix filled with Gaussian random variables and I is the identity. In order to test the hypothesis that the maximum entropy solution dominates the space of possible solutions with increasing matrix size, we investigate the relative error of the log determinant L, for $100 \leq N \leq 1000$. As can be seen in Figure 1, there is a clear decrease in relative error for all plotted percentiles with increasing matrix size N.

5 Experiments

So far, we have supplemented the theoretic foundations of our proposal by devising experiments on synthetically constructed matrices. In this section, we extend our evaluation to include real matrices obtained from a variety of problem domains, and demonstrate how the results obtained using our approach consistently outperform competing state-of-the-art approximations. Moreover, in order to demonstrate the applicability of our method within a practical domain, we highlight the benefits of replacing the exact computation of the log determinant term appearing in the log likelihood of a Gaussian Markov random field with our maximum entropy approximation.

5.1 SparseSuite Matrix Collection

While the ultimate goal of this work is to accelerate inference of large-scale machine learning algorithms burdened by the computation of the log determinant, this is a general approach which can be applied to a wide variety of application domains. The SuiteSparse Matrix Collection [11] (commonly referred to as the

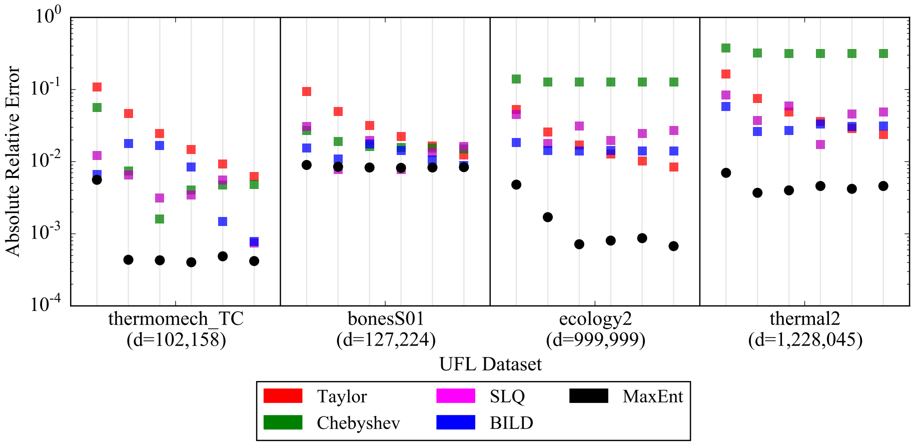

Fig. 2. Comparison of competing approaches over four UFL datasets. Results are also shown for increasing computational budgets, i.e. 5, 10, 15, 20, 25 and 30 moments respectively. Our method obtains substantially lower error rates across 3 out of 4 datasets, and still performs very well on 'bonesS01'.

set of UFL datasets) is a collection of sparse matrices obtained from various real problem domains. In this section, we shall consider a selection of these matrices as 'matrices in the wild' for comparing our proposed algorithm against established approaches. In this experiment we compare against Taylor [2] and Chebyshev [20] approximations, stochastic lanczos quadrature (SLQ) [35] and Bayesian inference of log determinants (BILD) [12]. In Figure 2, we report the absolute relative error of the approximated log determinant for each of the competing approaches over four different UFL datasets. Following [12], we assess the performance of each method for an increasing computational budget, in terms of matrix vector multiplications, which in this case corresponds to the number of moments considered. It can be immediately observed that our entropic approach vastly outperforms the competing techniques across all datasets, and for any given computational budget. The overall accuracy also appears to consistently improve when more moments are considered.

Complementing the previous experiment, Table 2 provides a further comparison on a range of other sample matrices which are large, yet whose determinants can be computed by standard machines in reasonable time (by virtue of being sparse). For this experiment, we consider 10 estimated moments using 30 probing vectors, and their results are reported for the aforementioned techniques. The results presented in Table 2 are the relative error of the log determinants *after* they have been normalised using Gershgorin intervals [14]. We note, however, that the methods improve at different rates as more raw moments are taken.

Table 2. Comparison of competing approximations to the log determinant over additional sparse UFL datasets. The technique yielding the lowest relative error is highlighted in bold, and our approach is consistently superior to the alternatives. Approximations are computed using 10 moments estimated with 30 probing vectors.

Dataset	Dimension	Taylor	Chebyshev	SLQ	BILD	MaxEnt
shallow_water1	81,920	**0.0023**	0.7255	0.0058	0.0163	0.0030
shallow_water2	81,920	0.5853	0.9846	0.9385	1.1054	**0.0051**
apache1	80,800	0.4335	0.0196	0.4200	0.1117	**0.0057**
finan512	74,752	0.1806	0.1158	0.0142	**0.0005**	0.0171
obstclae	40,000	0.0503	0.5269	0.0423	0.0733	**0.0026**
jnlbrng1	40,000	0.1084	0.2079	0.0465	0.0805	**0.0158**

5.2 Computation of GMRF Likelihoods

Gaussian Markov random fields (GMRFs) [32] specify spatial dependence between nodes of a graph with Markov properties, where each node denotes a random variable belonging to a multivariate joint Gaussian distribution defined over the graph. These models appear in a wide variety of applications, ranging from interpolation of spatio-temporal data to computer vision and information retrieval. While we refer the reader to [32] for a more comprehensive review of GMRFs, we highlight the fact that the model relies on a positive-definite precision matrix Q_θ parameterised by θ, which defines the relationship between connected nodes; given that not all nodes in the graph are connected, we can generally expect this matrix to be sparse. Nonetheless, parameter optimisation of a GMRF requires maximising the following equation:

$$\log p(\mathbf{x} \mid \theta) = \frac{1}{2}\log\bigl(\mathrm{Det}(Q_\theta)\bigr) - \frac{1}{2}\mathbf{x}^\top Q_\theta \mathbf{x} - \frac{n}{2}\log(2\pi),$$

where computing the log determinant poses a computational bottleneck, even where Q_θ is sparse. This arises because it is possible for the Cholesky decomposition of a sparse matrix with zeros outside a band of size k to be nonetheless dense *within* that bound. Thus, the Cholesky decomposition is still expensive to compute.

Following the experimental set-up and code provided in [19], in this experiment we evaluate how incorporating our approximation into the log likelihood term of a GMRF improves scalability when dealing with large matrices, while still maintaining precision. In particular, we construct lattices of increasing dimensionality and in each case measure the time taken to compute the log likelihood term using both approaches. The precision kernel is parameterised by κ and τ [23], and is explicitly linked to the spectral density of the Matérn covariance

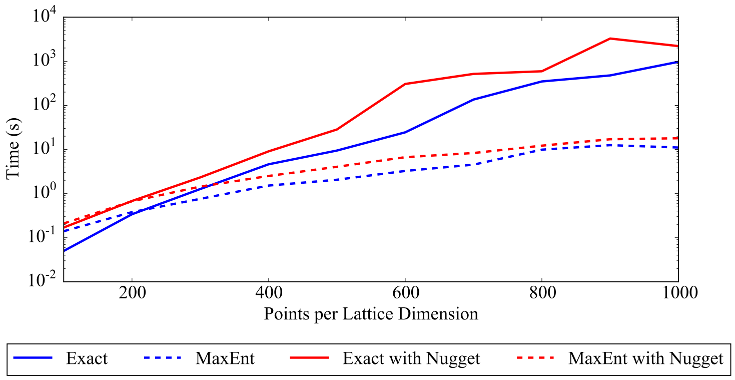

Fig. 3. Time in seconds for computing the log likelihood of a GMRF via Cholesky decomposition or using our proposed MaxEnt approach for estimating the log determinant term. Results are shown for GMRFs constructed on square lattices with increasing dimensionality, with and without a nugget term.

function for a given smoothness parameter. We repeat this evaluation for the case where a nugget term, which denotes the variance of the non-spatial error, in included in the constructed GMRF model. Note that for the maximum entropy approach we employ 30 sample vectors in the stochastic trace estimation procedure, and consider 10 moments. As illustrated in Figure 3, the computation of the log likelihood is orders of magnitude faster when computing the log determinant using our proposed maximum entropy approach. In line with our expectations, this speed-up is particularly significant for larger matrices. Similar improvements are observed when a nugget term is included. Note that we set $\kappa = 0.1$ and $\tau = 1$ for this experiment.

Needless to say, improvements in computation time mean little if the quality of inference degrades. Figure 4 illustrates the comparable quality of the log likelihood for various settings of κ and τ, and the results confirm that our method enables faster inference without compromising on performance.

6 Conclusion

Inspired by the probabilistic interpretation introduced in [12], in this work we have developed a novel approximation to the log determinant which is rooted in information theory. While lacking the uncertainty quantification inherent to the aforementioned technique, this formulation is appealing because it uses a comparatively less informative prior on the distribution of eigenvalues, and we have also demonstrated that the method is theoretically expected to yield superior approximations for matrices of very large dimensionality. This is especially significant given that the primary scope for undertaking this work was to accelerate the log determinant computation in large-scale inference problems. As illustrated

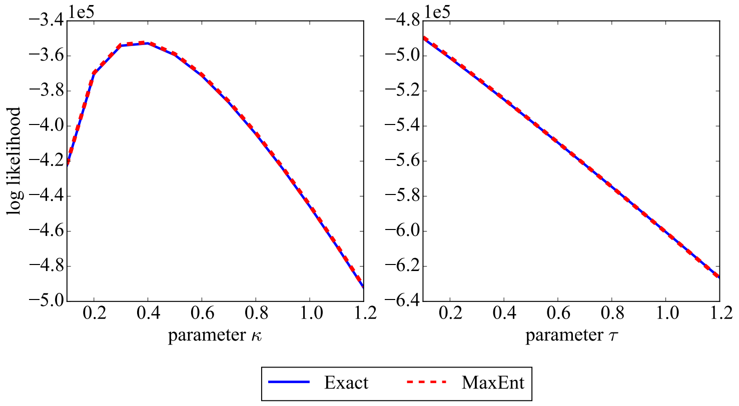

Fig. 4. The above plots indicate the difference of log likelihood between exact computation of the likelihood and the maximum entropy approach for a range of hyperparameters of the model. We note that the extrema of both exact and approximate inference align and it is difficult to distinguish the two lines.

in the experimental section, the proposed approach consistently outperforms all other state-of-the-art approximations by a sizeable margin.

Future work will include incorporating the empirical Monte Carlo variance of the stochastic trace estimates into the inference scheme, extending the method of maximum entropy to include noisy constraints, and explicitly evaluating the ratio of the functional integrals for large matrices to obtain uncertainty estimates similar to those in [12]. We hope that the combination of these advancements will allow for an apt active sampling procedure given pre-specified computational budgets.

References

1. Atallah, M.J., Grigorescu, E., Wu, Y.: A lower-variance randomized algorithm for approximate string matching. Information Processing Letters 113(18), 690–692 (2013)
2. Aune, E., Simpson, D.P., Eidsvik, J.: Parameter Estimation in High Dimensional Gaussian Distributions. Statistics and Computing 24(2), 247–263 (2014)
3. Avron, H.: Counting triangles in large graphs using randomized matrix trace estimation. In: Workshop on Large-scale Data Mining: Theory and Applications. vol. 10, pp. 10–9 (2010)
4. Avron, H., Toledo, S.: Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-definite Matrix. Journal of the ACM (JACM) 58(2), 8:1–8:34 (2011)
5. Bai, Z., Golub, G.H.: Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices. Annals of Numerical Mathematics 4, 29–38 (1997)

6. Bandyopadhyay, K., Bhattacharya, A.K., Biswas, P., Drabold, D.: Maximum entropy and the problem of moments: A stable algorithm. Physical Review E 71(5), 057701 (2005)
7. Boutsidis, C., Drineas, P., Kambadur, P., Zouzias, A.: A Randomized Algorithm for Approximating the Log Determinant of a Symmetric Positive Definite Matrix. CoRR abs/1503.00374 (2015)
8. Buchen, P.W., Kelly, M.: The Maximum Entropy Distribution of an Asset inferred from Option Prices. Journal of Financial and Quantitative Analysis 31(01), 143–159 (1996)
9. Caticha, A.: Entropic Inference and the Foundations of Physics (monograph commissioned by the 11th Brazilian Meeting on Bayesian Statistics–EBEB-2012 (2012)
10. Chen, J., Anitescu, M., Saad, Y.: Computing f(A)b via Least Squares Polynomial Approximations. SIAM Journal on Scientific Computing 33(1), 195–222 (2011)
11. Davis, T.A., Hu, Y.: The University of Florida Sparse Matrix Collection. ACM Transactions on Mathematical Software (TOMS) 38(1), 1 (2011)
12. Fitzsimons, J., Cutajar, K., Osborne, M., Roberts, S., Filippone, M.: Bayesian Inference of Log Determinants (2017)
13. Fitzsimons, J., Osborne, M., Roberts, S., Fitzsimons, J.: Improved stochastic trace estimation using mutually unbiased bases. arXiv preprint arXiv:1608.00117 (2016)
14. Gershgorin, S.: Uber die Abgrenzung der Eigenwerte einer Matrix. Izvestija Akademii Nauk SSSR, Serija Matematika 7(3), 749–754 (1931)
15. Giffin, A., Cafaro, C., Ali, S.A.: Application of the Maximum Relative Entropy method to the Physics of Ferromagnetic Materials. Physica A: Statistical Mechanics and its Applications 455, 11 – 26 (2016), http://www.sciencedirect.com/science/article/pii/S0378437116002478
16. Golub, G.H., Van Loan, C.F.: Matrix computations. The Johns Hopkins University Press, 3rd edn. (Oct 1996)
17. González, D., Davis, S., Gutiérrez, G.: Newtonian Dynamics from the Principle of Maximum Caliber. Foundations of Physics 44(9), 923–931 (2014)
18. Granziol, D., Roberts, S.: An Information and Field Theoretic approach to the Grand Canonical Ensemble (2017)
19. Guinness, J., Ipsen, I.C.F.: Efficient Computation of Gaussian Likelihoods for Stationary Markov Random Fields (2015)
20. Han, I., Malioutov, D., Shin, J.: Large-scale Log-Determinant computation through Stochastic Chebyshev Expansions. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (2015)
21. Hutchinson, M.: A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. Communications in Statistics - Simulation and Computation 19(2), 433–450 (1990)
22. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. 106, 620–630 (May 1957), http://link.aps.org/doi/10.1103/PhysRev.106.620
23. Lindgren, F., Rue, H., Lindström, J.: An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(4), 423–498 (2011)
24. Macchi, O.: The Coincidence Approach to Stochastic point processes. Advances in Applied Probability 7, 83–122 (1975)
25. Meurer, A., Smith, C.P., Paprocki, M., Čertík, O., Kirpichev, S.B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J.K., Singh, S., Rathnayake, T., Vig, S., Granger,

B.E., Muller, R.P., Bonazzi, F., Gupta, H., Vats, S., Johansson, F., Pedregosa, F., Curry, M.J., Terrel, A.R., Roučka, v., Saboo, A., Fernando, I., Kulal, S., Cimrman, R., Scopatz, A.: Sympy: symbolic computing in python. PeerJ Computer Science 3, e103 (Jan 2017), https://doi.org/10.7717/peerj-cs.103

26. Mohammad-Djafari, A.: A matlab program to calculate the maximum entropy distributions. In: Maximum Entropy and Bayesian Methods, pp. 221–233. Springer (1992)
27. Neri, C., Schneider, L.: Maximum Entropy Distributions inferred from Option Portfolios on an Asset. Finance and Stochastics 16(2), 293–318 (2012)
28. Nielsen, M.A., Chuang, I.: Quantum computation and quantum information (2002)
29. O'Hagan, A.: Bayes-Hermite Quadrature. Journal of Statistical Planning and Inference 29, 245–260 (1991)
30. Pressé, S., Ghosh, K., Lee, J., Dill, K.A.: Principles of Maximum Entropy and Maximum Caliber in Statistical Physics. Reviews of Modern Physics 85, 1115–1141 (Jul 2013), http://link.aps.org/doi/10.1103/RevModPhys.85.1115
31. Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning. MIT Press (2006)
32. Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications, Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall, London (2005)
33. Saatçi, Y.: Scalable Inference for Structured Gaussian Process Models. Ph.D. thesis, University of Cambridge (2011)
34. Shore, J., Johnson, R.: Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. IEEE Transactions on information theory 26(1), 26–37 (1980)
35. Ubaru, S., Chen, J., Saad, Y.: Fast Estimation of tr (f (a)) via Stochastic Lanczos Quadrature (2016)
36. Wainwright, M.J., Jordan, M.I.: Log-determinant relaxation for approximate inference in discrete markov random fields. IEEE Trans. Signal Processing 54(6-1), 2099–2109 (2006)
37. Wolfram Research Inc.: Mathematica, https://www.wolfram.com/mathematica/