Entropic Determinants of Massive Matrices

Diego Granziol and Stephen Roberts

***Abstract*—The ability of many powerful machine learning algorithms to deal with large data sets without compromise is often hampered by computationally expensive linear algebra tasks, of which calculating the log determinant is a canonical example. In this paper we demonstrate the optimality of Maximum Entropy methods in approximating such calculations. We prove the equivalence between mean value constraints and sample expectations in the big data limit, that Covariance matrix eigenvalue distributions can be completely defined by moment information and that the reduction of the self entropy of a maximum entropy proposal distribution, achieved by adding more moments reduces the KL divergence between the proposal and true eigenvalue distribution. We empirically verify our results on a variety of SparseSuite matrices and establish best practices.**

***Index Terms*—Maximum entropy methods, approximation methods, Matrix Theory, constrained optimization, noisy constraints, log determinants.**

I. Motivation

Scalability is one of the key challenges facing machine learning algorithms. In the era of large data sets, inference schemes are required to deliver optimal results within a constrained computational cost. Linear algebraic operations with high computational complexity pose a significant bottleneck to algorithmic scalability, and the log determinant of a matrix [1] falls firmly within this category of operations. The typical solution, involving Cholesky decomposition [2] 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 unfeasible for large matrices. We further find that, along with making multiple matrix copies, typical implementations of Cholesky decomposition require contiguous memory. Consequently, the difficulty in calculating this term greatly hinders widespread use of the learning models where it appears, which includes determinantal point processes [3], Gaussian processes [4], and graph problems [5].

II. Contributions of this Paper

Recent work combining Maximum Entropy algorithms with stochastic trace estimates of moments displayed state of the art performance on log determinant estimates with an $\mathcal{O}(n^2)$ computational time on both randomly generated and sparse matrices [6]. In this paper we address and answer many open pedagogical and practical concerns, such as:

1) Why should we characterize an eigenvalue probability distribution by its moments? To what extent do they embody relevant information?
2) What is the equivalence between sample averages and mean value constraints? When are they identical?
3) Can we characterize an eigenvalue distribution better with more moment constraints?
4) If a practitioner wants to use MaxEnt algorithms and stochastic trace estimates to generate an accurate log determinant estimate of a large matrix, how many samples and how many moments do they need to take?

Machine Learning Research Group, University of Oxford

III. Moments can fully describe probability distributions

For a probability measure μ having finite moments of all orders $\alpha_k = \int_{-\infty}^{\infty} x^k \mu(dx)$, if the power series $\sum_k \alpha_k/k!$ has a positive radius of convergence, that μ is the only probability measure with the moments $\alpha_1, \alpha_2, ...$ [7].

A. Application to Entropic Trace Estimation

Consider a random variable z with mean m and variance σ using the property of the expectations of quadratic forms, the expectation

$$\mathbb{E}[zz^t] = \sigma + mm^t = I, \tag{1}$$

where we have assumed that the variable is zero-mean and unit-variance. This allows us to calculate the trace of any matrix power A^m as

$$\begin{aligned} \mathrm{Tr}(A^m) &= \mathrm{Tr}(A^m I) = \mathrm{Tr}(A^m \mathbb{E}[zz^t]) \\ &= \mathbb{E}[\mathrm{Tr}(A^m zz^t)] = \mathbb{E}[z^t A^m z]. \end{aligned} \tag{2}$$

To explicitly prove that all covariance matrices have moments of finite order, we note from linear algebra that a $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. All normal matrices, of which real-symmetric (covariance) matrices are a subset, are diagonalizeable. Hence the eigenvectors of A span the space of the $\mathbb{R}^n$. We can thus write any vector as a linear combinations of the eigenvectors of A, $z = \sum_i \alpha_i |\phi_i\rangle$, where we have used dirac bra-ket notation to avoid confusion between the scalar α_i and the (ket) vector $|\phi_i\rangle$ the conjugate transpose (in this case just transpose as we are in the real space of ket is denoted as $z^t = \sum_i \alpha_i \langle\phi_i|$. Hence,

$$\begin{aligned} &\mathbb{E}(z^t A^n z) = \mathbb{E}\Big(\sum_j \alpha_j \langle\phi_j| A^n \sum_i \alpha_i |\phi_i\rangle \Big) \\ &= \mathbb{E}\Big(\sum_{i,j} \alpha_j \alpha_i \lambda^n \langle\phi_j|\phi_i\rangle \Big) = \mathbb{E}\Big(\sum_i |\alpha_i|^2 \lambda^n \Big) \\ &= \mathbb{E}\Big(\lambda_{\max}^n |\alpha_{\max}|^2 \Big[1 + \sum_{i \neq i_{\max}} \max \left| \frac{\alpha_i}{\alpha_{\max}} \right|^2 \left(\frac{\lambda_i}{\lambda_{\max}} \right)^n \Big] \Big) \\ &\xrightarrow{n \to \infty} \mathbb{E}\Big(\lambda_{\max}^n |\alpha_{\max}|^2 \Big) = \mathbb{E}\Big(|\alpha_{\max}|^2 \Big) \lambda_{\max}^n. \end{aligned} \tag{3}$$

in which we have used ortho-normality $\langle\phi_i|\phi_j\rangle = \delta_{i,j}$ along with the fact there are n distinct eigenvalues and that $\lambda_{i \neq i_{\max}} < \lambda_{\max}$. Although this is not strictly necessary, if

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

there were multiple degenerate maximum eigenvalues, the eigenvalue pre-factor in (3) would become:

$$\mathbb{E}\Big(|\alpha_{\max}|^2\Big) \xrightarrow{\text{k degenerate maxima}} \mathbb{E}\Big(\sum_i^k |\alpha_i|^2\Big). \tag{4}$$

Given that the n^{th} raw moment of the eigenvalue spectrum can be written as $\mathbb{E}(\lambda^n) = (1/n)Tr(A^n)$ thus

$$\begin{aligned}
&\frac{\mathbb{E}(\lambda^n)}{n!} = \frac{(1/n)\text{Tr}(A^n)}{n!} = \mathcal{C} \\
&\log \mathcal{C} = n \log \lambda_{\max} + \log \mathbb{E}(|\alpha_{\max}|^2) \\
&- (n+1)\log(n+1) + (n+1) \\
&\log \mathcal{C} \xrightarrow{n\to\infty} n(\log \lambda_{\max} - \log n + 1) \xrightarrow{\forall \lambda_{\max}} -\infty \\
&C \xrightarrow{n\to\infty} 0.
\end{aligned} \tag{5}$$

Here we use the fact that, for positive semi-definite matrices, all moments are positive and hence raw moments are equivalent to absolute power moments along with Stirling's approximation in the large n limit. This proves that it is possible to uniquely define a covariance matrix's eigenvalue probability distribution through its moment information.

This answers the question as to why beyond being computationally cheap $\mathcal{O}(n^2)$, it is worth sampling moments. They embody relevant information.

IV. Maximum Entropy

The method of maximum entropy (MaxEnt) [8] is a method which generates the least biased estimate of a proposal probability distribution, $q(x)$, given information in the form of functional expectations (also known as constraints). It is maximally non-committal in regards to missing information [9]. Mathematically we maximize the functional,

$$S = \int p(\vec{x}) \log p(\vec{x}) d\vec{x} - \sum_i \lambda_i \Big[\int p(\vec{x}) f_i(\vec{x}) d\vec{x} - \mu_i \Big], \tag{6}$$

with respect to $p(\vec{x})$, where $\langle f_i(\vec{x}) \rangle = \mu_i$ are the values of the imposed mean value constraints. For stochastic trace estimation, the functions are the power moments, $f_i = x^i$. The first term in Equation (6) is the Boltzmann-Shannon-Gibbs (BSG) entropy. This has been applied in a variety of disparate fields, from modelling crystal defects in lattice models in condensed matter physics [10] to inferring asset price movement distributions from option prices in finance [11], [12]. It can be used to derive statistical mechanics (without the a priori assumptions of ergodicity and metric transitivity [13]), non-relativistic quantum mechanics, Newton's laws and Bayes' rule [14], [15]. It can be proved under the axioms of consistency, uniqueness, coordinate invariance, subset and system independence, that for mean value constraints any self consistent inference scheme must either maximize the entropic functional (6), or any functional sharing its maximum [16], [8]. The Johnson and Shore axioms state that the entropy must have a unique maximum [16] and, given the convexity of the BSG entropy, it contains a unique maximum provided that the constraints are convex. This is satisfied for any polynomial in x and hence entropy maximization, given moment information, constitutes a self consistent inference scheme [8].

V. Functional Expectations

We provide a mathematical justification for the asymptotic equivalence between mean value constraints and sample expectations. From Chebyshev's inequality we have,

$$P(|X| \geq a) \leq \frac{1}{a^p}\mathbb{E}(|X|^p), \tag{7}$$

which, when applied to a set of independent random variables possessing a mean and variance, leads to:

$$\begin{aligned}
&P\left(\left|\frac{X_1 + X_2 + .. + X_N}{n} - \mu\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{\sum_{i=1}^n \text{Var}(X_i)}{n^2} \\
&\xrightarrow{i.i.d.} \frac{1}{\epsilon^2}\frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0.
\end{aligned} \tag{8}$$

Here we have used the Chebyshev inequality and the fact that the variance of a sum is the sum of the variances (for independent random variables), followed by the i.i.d. assumption and the asymptotic limit respectively. This is known as the weak law of large numbers, as the limit is outside the brackets. Note that this limit does not necessitate the variables to be identically distributed, nor does it preclude weak dependence[1]. Using extensions of the central limit theorem [17], [18] it can be shown that for at most weakly dependent random variables obeying weak conditions, this result also holds.

Considering each X_m to be a stochastic trace estimate of the power moment $\int p(x)x^m dx$, we see that in the large n or big data limit, we recover the true mean value constraint with probability 1.

A. Sufficiency of Statistics

All distributions derived from the method of maximum entropy are within the exponential family, i.e. they are of the form,

$$p(x|\theta) = A(x)\exp(\langle T(x), \theta\rangle - F(\theta)), \tag{9}$$

where $\theta \in \Theta \subset \mathbb{R}^d$. T and A are fixed functions that characterize the exponential family, $F(\theta)$ is a normalization factor with respect to some measure $\nu(x)$ and $A(x)$ is the carrier measure. For variables independently drawn from the probability measure, we have

$$\begin{aligned}
&p(x_1, x_2 ..... x_n|\theta) = \Pi_i p(x_i|\theta) \\
&= A(x)\exp(\langle \sum_i T(x_i), \theta\rangle - F(\theta)).
\end{aligned} \tag{10}$$

This distribution hence depends on the input data only through the sample statistic $\sum_i T(x_i)$, referred to as a sufficient statistic. Given the nature of our proposed inference scheme (MaxEnt) restricts us to the exponential family, it makes sense for us to compress the data with no loss of information. That no information is lost is implicit in the definition of a sufficient statistic and can be demonstrated using the data-processing inequality [19]. This insight is also discussed by Jaynes [20].

[1]For fully dependent variables $\text{Var}(\sum_i^n X_i) = n^2\text{Var}(X)$ and hence the limit is never reached.

VI. SELF ENTROPY AS A DIVERGENCE

Consider the KL divergence $\mathcal{D}_{kl}$, also known as the minimum discrimination information, or negative relative entropy [19] between a true eigenvalue distribution $p(x)$ and a proposal MaxEnt solution $q(x) = \exp(\sum_j \alpha_j x^i)$:

$$\mathcal{D}(P||Q) = \int p(x) \log p(x) dx - \int p(x) \log q(x) \tag{11}$$

note that the (self) entropy of the MaxEnt solution is given by

$$\begin{aligned} \mathcal{S}(Q) &= -\int q(x) \log q(x) dx \\ &= \sum_i \alpha_i \int x^i \exp\Big(-\sum_j \alpha_j x^j \Big) dx = \sum_i \alpha_i \langle x^i \rangle, \end{aligned} \tag{12}$$

where α denotes the Lagrange multipliers pertaining to the MaxEnt solution and $\langle x^j \rangle$ refers to the expectation of the j^{th} moment.

The first term in equation (11) is the negative entropy of the true unknown distribution $\mathcal{S}(p)$. We can thus rewrite equation (11) as:

$$\begin{aligned} &-\mathcal{S}(P) + \int p(x) \sum_i \alpha_i x^i = -\mathcal{S}(P) + \sum_i \alpha_i \langle x^i \rangle \\ &= -\mathcal{S}(P) + \mathcal{S}(Q). \end{aligned} \tag{13}$$

We have used the fact that the functional expectations of our MaxEnt distribution by construction (Equation (6)) match that of the underlying distribution.

Thus by minimizing $\mathcal{S}(Q)$, for which we have an analytic form, we manifestly reduce the KL divergence between our MaxEnt proposal $q(x)$ and our true eigenvalue distribution $p(x)$. It is further clear by the use of the information inequality [19] that the entropy of our proxy MaxEnt solution serves as an upper bound to that of the true solution, i.e.

$$\mathcal{D}_{kl}(P||Q) = \mathcal{S}(Q) - \mathcal{S}(P) \geq 0 \rightarrow \mathcal{S}(Q) \geq \mathcal{S}(P). \tag{14}$$

A. Consequences

In section III we proved that an eigenvalue distribution could be completely specified by its power moments. In the above section we show that that the entropy of the MaxEnt proposal distribution $q(x)$ is an upper bound to the entropy of the true data generating distribution $p(x)$ and prove that the reduction in self entropy $\mathcal{S}(q)$ is equivalent to reducing $\mathcal{D}_{kl}(p||q)$. In the next section we prove that adding information in the form of extra moment information necessarily reduces the self entropy $\mathcal{S}(q)$. This result generates an active procedure in which we can be principled in knowing how many functional expectations we need to take. We just sequentially calculate the self entropy of proposal MaxEnt distribution $q(x)$ using equation (12) and terminate the procedure at the point at which this decrease becomes negligible $-\Delta\mathcal{S} < \epsilon$.

VII. LAGRANGIAN DUALITY

Consider a generic optimization problem of the form,

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0,\ i = 1...m \\ &\text{subject to } h_i(x) = 0,\ i = 1...p \end{aligned} \tag{15}$$

where $x \in \mathbb{R}^n$ and the domain $\mathcal{D} = \bigcap_{i=0}^{m} f_i \cap \bigcap_{i=1}^{p} h_i$. We define the Lagrangian dual function as the infimum of the Lagrangian over the domain of x,

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} \Big(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \Big). \end{aligned} \tag{16}$$

As the dual is the pointwise infimum of a family of affine functions of (λ, ν), it is concave, irrespective of the convexity of f_0, f_i, h_i. [21]. It is easily verifiable due to the net negativity of the two summation terms in $g(\lambda, \nu)$ that the dual provides a lower bound on the optimal value p^* of the primal problem. This is known as weak duality. In the case of equality constraints this bound is tight.

For general inequality constraints the difference between the primal and dual optimal solution (duality gap) is not 0. However, for $f_0...f_m$ convex, Affine equality constraints and certain regularity conditions, we have a duality gap of 0, this is known as strong duality. An example of such a constraint qualification is Slater's condition, which states that there is an $x \in$ rel. int. $(\mathcal{D})$ which satisfies the constraints (where rel. int. refers to the relative interior).

A. Application to Maximum Entropy

We wish to maximise the entropic functional $\mathcal{S}(p) = -\int p(x) \log p(x) dx$ under certain moment constraints $\int p(x) x^m dx = \mu_m$. This can be written as,

$$\begin{aligned} &\text{minimize } f_0[p(x)] = \int p(x) \log p(x) dx \\ &\text{subject to } h_i[p(x)] = \int p(x) x^i dx - \mu_i = 0,\ i = 1...p. \end{aligned} \tag{17}$$

Given that the negative entropy is a convex objective and that the moment equality constraints are affine in the variable being optimised over $p(x)$ by strong duality we have an equivalence between the solution of the dual and that of the primal.

It is also clear that the domain defined as the intersection of the constraint sets can never increase upon the addition of an extra constraint. Hence,

$$\inf_{x \in \mathcal{D} = \bigcap_{i=0}^{m} f_i} L(x, \lambda, \nu) \leq \inf_{x \in \mathcal{D} = \bigcap_{i=0}^{m+1} f_i} L(x, \lambda, \nu) \tag{18}$$

and thus the entropy can only decrease when adding an extra constraint. Hence by adding more moment constraints, we always reduce the entropy and given equations (14) and (13) we necessarily reduce $\mathcal{D}_{kl}(p[x]||q[x])$, where $p[x], q[x]$ define the true eigenvalue and MaxEnt proposal distributions respectively.

VIII. Algorithm

We apply a numerically stable MaxEnt Algorithm (algorithm 1) [23], under the conditions that λ_i is strictly positive and the all power moments $0 \leq \lambda^k \leq 1$. We can satisfy these conditions by normalizing our positive definite matrix by the maximum of the Gershgorin intervals [24].

Algorithm 1 Optimising the Coefficients of the MaxEnt Distribution

Input: Moments $\{\mu_i\}$, Tolerance ϵ
Output: Coefficients $\{\alpha_i\}$
1: $\alpha_i \sim \mathcal{N}(0, 1)$
2: $i \leftarrow 0$
3: $p(\lambda) \leftarrow \exp(-1 - \sum_k \alpha_k \lambda^k)$
4: **while** error $< \epsilon$ **do**
5: $\quad \delta \leftarrow \log\left(\frac{\mu_i}{\int \lambda^i p(\lambda) d\lambda}\right)$
6: $\quad \alpha_i \leftarrow \alpha_i + \delta$
7: $\quad p(\lambda) \leftarrow p(\lambda|\alpha)$
8: $\quad$ error $\leftarrow \max |\int \lambda^i p(\lambda) d\lambda - \mu_i|$
9: $\quad i \leftarrow \text{mod}(i + 1, \text{length}(\mu))$

We follow the procedure from entropic trace estimation [6]. Firstly, the raw moments of the eigenvalues are estimated using stochastic trace estimation. These moments are then passed to the maximum entropy optimization of Algorithm 1 to produce an estimate of the distribution of eigenvalues, $p(\lambda)$. Consequently, $p(\lambda)$ is used to estimate the distribution's log geometric mean, $\int \log(\lambda) p(\lambda) d\lambda$. This term is multiplied by the matrix's dimensionality and if the matrix was normalized, the log of this normalization term is added. We lay out these steps 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 ϵ
Output: Log Determinant Approximation $\log |A|$
1: $B = A / \|A\|_2$
2: μ (moments)$\leftarrow$ StochasticTraceEstimation(B, k)
3: α (coefficients) $\leftarrow$ MaxEntOpt(μ, ϵ)
4: $p(\lambda) \leftarrow p(\lambda|\alpha)$
5: $\log |A| \leftarrow n \int \log(\lambda) p(\lambda) d\lambda + n \log(\|A\|_2)$

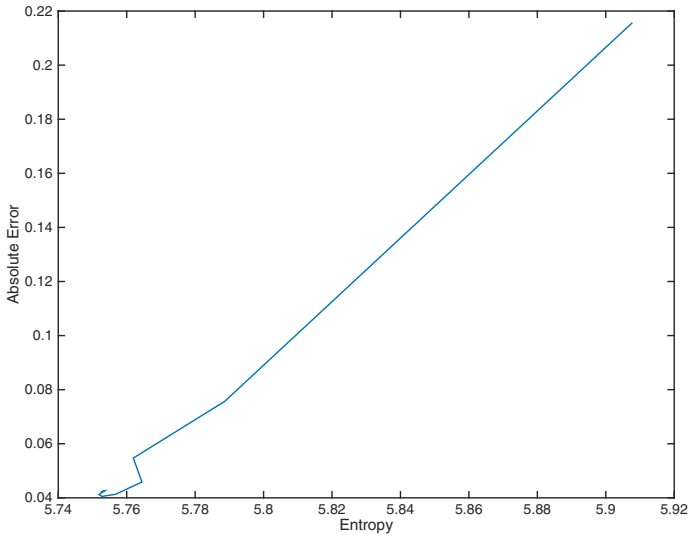

Fig. 1. Relative error against entropy for a single sample stochastic trace estimate from 2 to 10 moments for the Thermomech dataset.

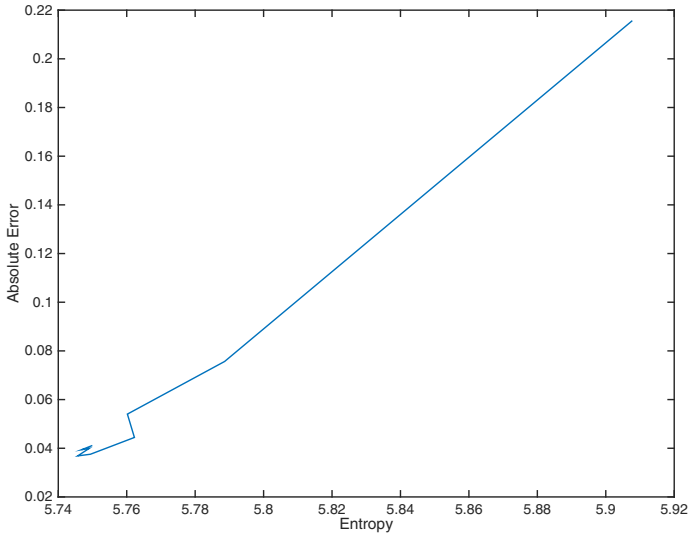

Fig. 2. Relative error against entropy for a 50 sample stochastic trace estimate from 2 to 10 moments for the Thermomech dataset.

A. Algorithmic details for Practitioners

Given that the MaxEnt approach of Algorithm 1 is numerical, we need to specify a gridding of the input space or choice of nodes. We find that a gridding between $0 \leq x \leq 1$ of $\Delta x = 0.001$ provides a good trade-off between speed and accuracy, with essentially the same results (measured by absolute error) as $\Delta x = 0.0001$. We find that even when we set the tolerance ϵ in algorithm 1 to 10^{-4} (smaller values than this significantly increase run time) the entropy, shown in Figure 4 and by the results of section VI the absolute error shown in Figure 3 does not seem to decrease beyond $m \approx 10$ moments. We note that ϵ represents a maximum error to which extent each moment constraint can be violated. Given that the eigenvalue power moments are strictly decreasing, this means the higher order moment constraints can be violated by a greater relative fraction, which when combined with the fact that the constraints are noisy and that the computational complexity of cycling through all the constraints with an approximate correction term (the basis for algorithm 1) increases with more moments (as does the error per constraint) the usefulness of using more moment information is limited. We hence don't recommend using more than 10 moments. It is also ill advised to use moment information which is smaller than your error tolerance.

IX. Stochastic trace estimates

A key component of the computational complexity of Entropic trace estimation [6], is the number of stochastic samples taken. Hence we ask the question, how does the quality of our inference, i.e the approximation to the log determinant depend on the number of samples taken?

To keep our results comparable and consistent, we keep with [6], [25] and consider Gaussian random unit vectors. We note that across a variety of sparse datasets, the number of samples taken neither largely effects the entropy of the proposal distribution (used to determine the number of moments required before attaining an optimal result) as is demonstrated by the indiscernability of Figures 1 and 2, where we compare the Entropy of the corresponding MaxEnt distribution for 1 of 50 stochastic trace samples. For Figure 3 the corresponding graph for 50 stochastic samples is also virtually identical. Given this behaviour we empirically investigate the extent to which reducing the number of stochastic samples affects the estimation of the log determinant.

A. Single sample result comparisons

We load five sparse (SuiteSparse) square PSD matrices, ranging from a maximum dimension of $999,999$ to a minimum of $81,200$ and run Cholesky using the Matlab 2014b 'Chol' function to calculate the log determinants on a 2.6 GHz Intel Core i7 16 GB 1600 MHz DDR3 notebook. This takes 4847 seconds. Using our MaxEnt Algorithm 1, with a gridding of 0.001 and 30 stochastic samples of 8 moments, we calculate the log determinants in 20 seconds. For a single sample, we calculate the log determinants in 16 seconds. The respective errors are shown in Table I. We note that even for relatively

Dataset	Dimension	Samples	Error
ecology2	999,999	30	0.0102
		1	0.0105
thermomech TC	102,158	30	0.0398
	102,158	1	0.0402
shallow water 1	81,920	30	0.0043
		1	0.0035
shallow water 2	81,920	30	0.0039
		1	0.0040
apache1	80,800	30	0.006571
		1	0.0101

TABLE I
Relative absolute error on SuiteSparse datasets, for 8 moments and 30/1 samples per moment

small matrices, $n \approx 80,000$, that the performance from reducing the number of samples is relatively unaffected. However, given that there is a slight decrease in performance and that 15 of the 20 and 16 seconds of compute time were spent on Alogorithm 1, determining the Maximum Entropy coefficients, we recommend reducing the number of samples only when it becomes a larger proportion of the overall cost.

We further test the performance impact, by evaluating the difference in MaxEnt estimate for the largest PSD matrices in the SuiteSparse data set, comparing a single sample to 30 samples. For Queen 4147 with a dimension of $4,147,110$ and $316,548,962$ non zero values, the difference in prediction from taking 1 sample instead of 30 is 0.0028% and the runtime is reduced from 173 to 60 seconds. We note that the standard Cholesky and LU functions in (e.g.) MATLAB are unable to handle matrices of that size, due to contiguous memory constraints, even on significantly more powerful machines than the one above. Table II shows results for a variety of large datasets. The reduction in samples from 30 to 1 reduces

Dataset	Dimension	Samples	Estimate	Time(s)	$\Delta\%$
Queen	4,147,110	30	-7.3951e+07	172.4	
Non 0's	316,548,962	1	-7.3953e+07	60.3	0.0028
Bump	2,911,419	30	-5.2282e+07	64.3	
Non 0's	127,729,899	1	-5.2297e+07	15.9	0.029
Serena	1,391,349	30	-1.5831e+07	34.4	
Non 0's	64,131,971	1	-1.5771e+07	8.867	0.38
Geo	1,437,960	30	-1.0186e+07	33.2	
Non 0's	60,236,322	1	-1.0203e+07	12.5	0.17
Hook	1,498,023	30	-4.6026e+06	32.3	
Non 0's	59,374,451	1	-4.6033e+06	11.8	0.015
StochF	1,465,137	30	-2.6807e+07	15.4	
Non 0's	21,005,389	1	-2.6812e+07	7.1	0.019
G3	1,585,478	30	-1.0263e+07	9.875	
Non 0's	7,660,826	1	-1.0262e+07	4.618	0.097

TABLE II
Results for the largest PSD SuiteSparse matrices, using 8 moments, with sample number either 30 or 1. Final column denotes percentage difference in estimate between using 30/1 sample(s).

the computational run-time by a factor of 3 and the difference in estimates, which is always less than 0.4% tends to increase in as the matrix dimension decreases. The exceptions, *StochF* and *G3*, are both significantly sparser than the others, which is why they run significantly faster and the MaxEnt calculation algorithm (which is independent of the number of samples taken) takes up a greater proportion of the total run-time and hence the reduction from taking less samples is less. We posit a potential link between sparsity and accuracy, but leave the investigation for future work.

The link between reduction in proposal self entropy and absolute error, is also unchanged as we reduce the number of samples, as can be seen by comparing Figures 1 and 2.

X. Conclusion

In this paper we formally establish the link between sample expectation and mean value constraint, proving asymptotic equivalence. We also prove that the eigenvalue distribution of a Covariance matrix can be uniquely determined by its moments. The combination of these two provides a solid foundation for using stochastic trace estimation sample estimates as mean value constraints for a Maximum Entropy estimation of a Covariance matrix eigenvalue density.

We further show how the inclusion of extra moment constraints, necessarily reduces the KL divergence $\mathbb{D}_{kl}(p||q)$ between the MaxEnt proposal $q(\lambda)$ and true eigenvalue spectrum $p(\lambda)$. We demonstrate empirically on SuiteSprase datasets how this reduction in $\mathbb{D}_{kl}(p||q)$ corresponds to increased estimation accuracy.

We investigate the effect of reducing the number of stochastic trace estimate samples empirically and experimentally demonstrate that the larger the matrix, the smaller the effect

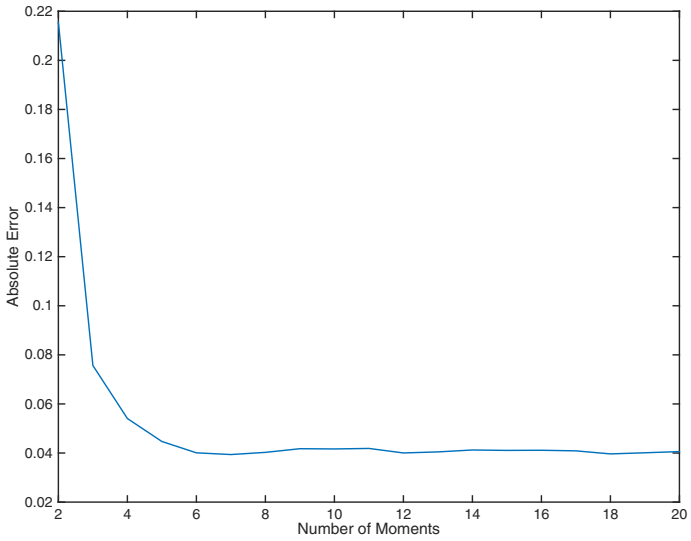

Fig. 3. Relative error against number of moments included for a single sample stochastic trace estimate for the Thermomech dataset. The corresponding graph for 50 samples is visually indistinguishable from this one.

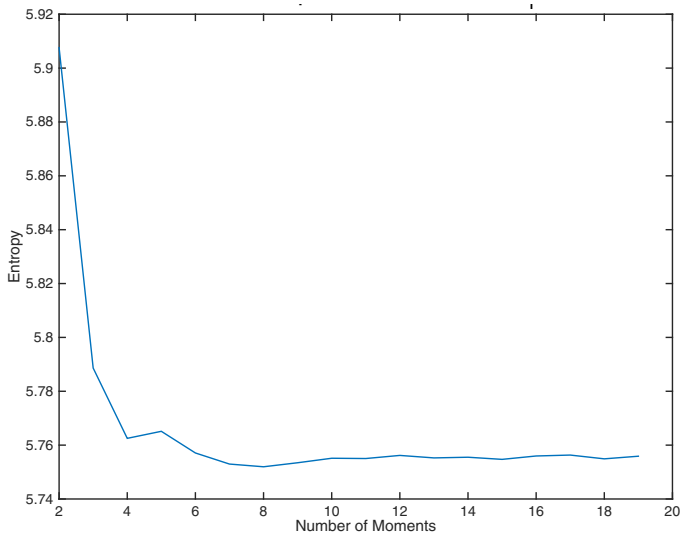

Fig. 4. Entropy against number of Moments for the Thermomech dataset

and the greater the computational benefit of reducing the number of samples.

We set up best practice guidelines, rooted in theory and experiment, for practitioners wishing to deal with large matrices. Our basic, non-optimized MaxEnt implementation is able to calculate determinants of 4 million by 4 million matrices on a laptop within a minute.

Acknowledgments

The authors would like to thank the Oxford-Man Institute and the Royal Academy of Engineering for their support, Thomas Gunter and Michael Osborne for illuminating discussions, Pawan Kumar for his input on Convex Analysis and Tim Davies for the upkeep or the SuiteSparse dataset.

References

[1] Zhaojun Bai and Gene H. Golub. Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices. Annals of Numerical Mathematics, 4:29–38, 1997.

[2] Gene H. Golub and Charles F. Van Loan. Matrix computations. The Johns Hopkins University Press, 3rd edition, October 1996.

[3] Odile Macchi. The Coincidence Approach to Stochastic point processes. Advances in Applied Probability, 7:83–122, 1975.

[4] Carl E. Rasmussen and Christopher Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.

[5] Martin J. Wainwright and Michael I. Jordan. Log-determinant relaxation for approximate inference in discrete markov random fields. IEEE Trans. Signal Processing, 54(6-1):2099–2109, 2006.

[6] Jack Fitzsimons, Diego Granziol, Kurt Cutajar, Michael Osborne, Maurizio Filippone, and Stephen Roberts. Entropic trace estimates for log determinants, 2017.

[7] Patrick Billingsley. Probability and measure. Wiley, 2012.

[8] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. Rev. Mod. Phys., 85:1115–1141, Jul 2013.

[9] E. T. Jaynes. Information theory and statistical mechanics. Phys. Rev., 106:620–630, May 1957.

[10] Adom Giffin, Carlo Cafaro, and Sean Alan Ali. Application of the maximum relative entropy method to the physics of ferromagnetic materials. Physica A: Statistical Mechanics and its Applications, 455:11 – 26, 2016.

[11] Cassio Neri and Lorenz Schneider. Maximum entropy distributions inferred from option portfolios on an asset. Finance and Stochastics, 16(2):293–318, 2012.

[12] Peter W Buchen and Michael Kelly. The maximum entropy distribution of an asset inferred from option prices. Journal of Financial and Quantitative Analysis, 31(01):143–159, 1996.

[13] Diego Granziol and Stephen Roberts. An information and field theoretic approach to the grand canonical ensemble, 2017.

[14] Diego González, Sergio Davis, and Gonzalo Gutiérrez. Newtonian dynamics from the principle of maximum caliber. Foundations of Physics, 44(9):923–931, 2014.

[15] A Caticha. Entropic inference and the foundations of physics (monograph commissioned by the 11th brazilian meeting on Bayesian statistics–ebeb-2012, 2012.

[16] John Shore and Rodney Johnson. 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.

[17] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, pages 583–602, Berkeley, Calif., 1972. University of California Press.

[18] A. N. Tikhomirov. On the convergence rate in the central limit theorem for weakly dependent random variables. Theory of Probability '1&' Its Applications, 25(4):790809, 1981.

[19] Thomas M Cover and Joy A Thomas. Elements of information theory. John Wiley & Sons, 2012.

[20] Edwin T Jaynes. On the rationale of maximum-entropy methods. Proceedings of the IEEE, 70(9):939–952, 1982.

[21] Stephen P. Boyd and Lieven Vandenberghe. Convex optimization. Cambridge University Press, 2009.

[22] Peter Walley. Statistical reasoning with imprecise probabilities. Chapman and Hall, 1991.

[23] K Bandyopadhyay, Arun K Bhattacharya, Parthapratim Biswas, and DA Drabold. Maximum entropy and the problem of moments: A stable algorithm. Physical Review E, 71(5):057701, 2005.

[24] Semyon Gershgorin. Uber die Abgrenzung der Eigenwerte einer Matrix. Izvestija Akademii Nauk SSSR, Serija Matematika, 7(3):749–754, 1931.

[25] Jack Fitzsimons, Kurt Cutajar, Michael Osborne, Stephen Roberts, and Maurizio Filippone. Bayesian inference of log determinants, 2017.