On the Entropy and Divergence of Gaussian Mixture Models

Diego Granziol and Stephen Roberts

***Abstract*—For the widely applied and practical Gaussian mixture model (GMM), there exists no closed form expression for the differential entropy. We derive a novel upper bound for the differential entropy of a GMM using the method of maximum entropy and propose an algorithm for its calculation. We further show that the Kullback Leibler (KL) divergence between two GMM's can be trivially estimated using the parameters of its .**

***Index Terms*—Gaussian mixtire models, differential entropy, KL divergence, Maximum entropy methods.**

I. Introduction

Differential entropy, defined as $\int p(x)\log p(x)dx$ [1] extends the notion of classical entropy [2] to continuous measures. In the case of bounded measures, or for those for which it makes sense to propose a uniform prior [3], it has a natural interpretation as uncertainty over the density $p(x)$, reaching a maximum for the uniform and a minimum for the Dirac delta distribution. It is used for parameter estimation [4] and for the computation of mutual information (MI).

$$\text{MI} = \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{1}$$

This is an ubiquitous term, meausing the Kullback-Leiber divergence between a joint density and its marginals. It is used in in a variety of applications, such as communication channel capacity [1], sensor management [5] and image restoration [6]. The Kullback-Leiber divergence [1], also known as the relative entropy, or minimum discrimination information, between two densities $f(x)$ and $g(x)$ is defined as

$$\mathcal{D}_{kl}(f||g) = \int f(x)\log\frac{f(x)}{g(x)}dx \tag{2}$$

and satisfies the properties of

1) Self Similarity: $\mathcal{D}_{kl}(f||f) = 0$
2) Self Identification: $\mathcal{D}_{kl}(f||g) = 0$ iff $f = g$
3) Positivity: $\mathcal{D}_{kl}(f||g) \geq 0 \, \forall f, g$

It is widely applied as a similarity measure for speech [7], image, text [8] and HMMs [9]. Futhermore, its Hessian defines the Fisher-Rao metric, which allows us to define a unique Riemmanian manifold, on which the points corresponds to probability measures and from which we can calculate informational difference between measurements [10].

Machine Learning Research Group, University of Oxford

II. Entropy, KL and Gaussian Mixture Models

For Gaussian distributions both the differential entropy and KL divergence have closed form solutions. However for highly skewed or multimodal distributions which describe real world data, the Gaussian approximation may be in-appropriate. The Gaussian mixture (where $\sum w_i = 1$)

$$\text{GMM} = \sum_i^N w_i \mathcal{N}(\vec{x}; \vec{\mu_i}, \vec{\Sigma_i}) \tag{3}$$

is a widely adopted and practical universal function approximator [11], used to deal with the short comings of the single Gaussian. Unfortunately for the GMM, due to the summation term in the logarithm, no analytic expression of the differential entropy or the KL divergence between two GMM's exists.

III. Related work

In [12] the authors propose an analytical approximation to the differential entropy, by using the GMM moments and truncating the logarithm expansion after R terms

$$\begin{aligned} H(\vec{x}) \approx &- \sum_i^N \int_{\mathbb{R}^n} w_i \mathcal{N}(\vec{x}; \vec{\mu_i}, \vec{\Sigma_i}) \\ &\times \left(\sum_{k=0}^R \frac{1}{k!} \left((x-\mu_k)\odot\nabla \right)^k \log g(x) \Big|_{x=\mu_i} \right) \end{aligned} \tag{4}$$

Where $\odot$ stands for the matrix contraction operator. They further derive an upper bound to the GMM differential entropy

$$H_{u1} = \sum_i^N -w_i \log w_i + \frac{1}{2}\log((2\pi e)^N |\Sigma_i|) \tag{5}$$

In [13] they propose and compare a variety of approximations to the computationally demanding gold standard of Monte Carlo simulation. They propose to replace the full GMM density $p(x)$ with either a single or product of Gaussian(s) in order to calculate the KL divergence. They further propose and finally conclude in favour of their analytical variational approximation

$$\mathcal{D}_{var}(f||g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-\mathcal{D}(f_a||f_{a'})}}{\sum_b w_b e^{-\mathcal{D}(f_a||g_b)}} \tag{6}$$

where π_a represents the weight of the a'th Gaussian component in mixture f and f_a corresponds to that density. check this

IV. Contributions of this Paper

In this paper, we use the method of maximum entropy and the moments of the GMM to derive a novel upper bound on the GMMs differential entropy and a novel approximation to the KL divergence between two GMMs. The structure of the paper is as follows

- We prove that the power moments of a GMM completely define its characteristic function and hence its density. This establishes the validity of using the GMMs power moments to estimate a more tractable surrogate density.
- We use the distribution of maximum entropy given the GMM's moments as constraints to upper bound the differential entropy of the GMM
- We prove, using the non-negativity of the multivariate mutual information [1] that the entropy of the multivariate distribution of maximum entropy is a lower bound to the sum of entropies of the univariate (marginal) distributions of maximum entropy
- This allows us to avoid computing computationally costly multi-dimensional numerical integrations, reducing an n^d numerical integration to $n \times 1^d$ numerical integrations.
- We propose an algorithm written in Python for the computation of an upper bound of the differential entropy of a GMM, make the code available and compare it to various other approximations in the literature
- We further show how the parameters of the distribution of maximum entropy allow for an approximate calculation of the KL divergence between two GMMs and compare it to other approximations in the literature

V. Maximum Entropy

The method of maximum entropy (MaxEnt) [14] 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 [15]. Mathematically we maximize the functional,

$$S = \int q(\vec{x}) \log q(\vec{x}) d\vec{x} - \sum_i \lambda_i \left[\int p(\vec{x}) f_i(\vec{x}) d\vec{x} - \mu_i \right], \quad (7)$$

with respect to $q(\vec{x})$, 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 (7) 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 [16] to inferring asset price movement distributions from option prices in finance [17], [18]. It can be used to derive statistical mechanics (without the a priori assumptions of ergodicity and metric transitivity [19]), non-relativistic quantum mechanics, Newton's laws and Bayes' rule [20], [3]. 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 (7), or any functional sharing its maximum [21], [14]. The Johnson and Shore axioms state that the entropy must have a unique maximum [21] 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 [14].

[1] There are many variants of the multivariate mutual information, some of which can be negative, we propose a natural definition of what we call the dependence information, which is non-negative

VI. Are moments sufficient to fully describe the problem?

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 $\alpha_1, \alpha_2, ...$ [22].
Informally, a Gaussian has finite orders of all moments, hence any finite combination of Gaussians must necessarily possess finite moments of all orders and the above condition is satisfied.
For the case of a one dimensional Gaussian with a location parameter μ_i and standard deviation σ_i it can be seen that the $2k$'th moment

$$G_{2k} = \sum_{2p=0}^{2k} \binom{2k}{2p} \mu_i^{2(k-p)} \beta_i^{-p} \quad (8)$$

As all odd power moments of the Gaussian are 0. Hence for the Gaussian mixture model we have

$$GMM_{2k} = \sum_{i=1}^{N} w_i \sum_{2p=0}^{2k} \binom{2k}{2p} \mu_i^{2(k-p)} \beta_i^{-p} \quad (9)$$

We note that since $\sum_i w_i = 1$ and $w_i \geq 0$ that $w_i \leq 1$. Furthermore $\mu_i \leq \mu_{upper} > 1$ and $\beta_i \geq \beta_{lower} < 1 \ \forall i$. We can bound the above expression by further seeing that $\sum_{2p=0}^{2k} \binom{2k}{2p} < (k+1) \frac{(2k)!}{(k!)^2}$. Hence

$$GMM_{2k} < N(k+1) \frac{(2k)!}{(k!)^2 (\mu_{max}^{2k} \beta_i^{-k})} \quad (10)$$

Which we can see is smaller than $(2k)!$ in the $k \to \infty$ limit by taking logarithms

$$\frac{\log N}{2k} + \frac{\log(k+1)}{2k} + \log mu_{max} + \frac{|\log \beta_i|}{2} \leq \log k \quad (11)$$

VII. Proof of MaxEnt proposal as an upper bound

Consider the KL divergence $\mathcal{D}_{kl}$, [1] between the true Gaussian mixture density $p(x)$ and a proposal MaxEnt solution $q(x) = \exp(\sum_j \alpha_j x^i)$:

$$\mathcal{D}(P||Q) = \int p(x) \log p(x) dx - \int p(x) \log q(x) \quad (12)$$

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

$$\begin{aligned} H(Q) &= -\int q(x) \log q(x) dx \\ &= \sum_i \alpha_i \int x^i \exp\left(-\sum_j \alpha_j x^j \right) dx = \sum_i \alpha_i \langle x^i \rangle, \end{aligned} \quad (13)$$

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 (12) is the (unknown) negative entropy of the GMM $\mathcal{S}(p)$. We can thus rewrite equation (12) as:

$$\begin{aligned} &- H(P) + \int p(x) \sum_i \alpha_i x^i = -H(P) + \sum_i \alpha_i \langle x^i \rangle \\ &= -H(P) + H(Q). \end{aligned} \tag{14}$$

Where we have used the fact that the functional expectations of our MaxEnt distribution by construction (Equation (7)) match that of the underlying distribution. Using the information inequality [1] it is clear, that the entropy of our proxy MaxEnt solution serves as an upper bound to that of the true solution, i.e.

$$\mathcal{D}_{kl}(P||Q) = H(Q) - H(P) \geq 0 \rightarrow H(Q) \geq H(P). \tag{15}$$

VIII. Multidimensional Entropy

For an n dimensional variable, we can define a variant of the mutual information as

$$I_d(x_1.....x_n) \equiv \int p(x_1.....x_n) \log \frac{p(x_1.....x_n)}{\prod_i^n p(x_i)} \prod_i^n dx_i \tag{16}$$

Which we denote as the dependence information. This can be split into a sum of entropic terms

$$I_d(x_1.....x_n) = -H(x_1...x_n) + \sum_i^N H(x_i) \tag{17}$$

and is thus seen to measure the difference in entropy between the joint and the sum of the entropies of the marginals, justifying its name.

It can be easily shown using the convexity of the negative log function and Jensens inequality ($\mathbb{E}[f(x)] \geq f[\mathbb{E}(x)]$) that

$$\begin{aligned} &- \int p(x_1..x_n) \log \frac{\prod_i^n p(x_i)}{p(x_1...x_n)} \prod_i^n dx_i \geq -\log \int \prod_i^n p(x_i) dx_i \\ &\therefore I_d(x_1.....x_n) \geq 0 \\ &\therefore \sum_i^N H(x_i) \geq H(x_1...x_n) \end{aligned} \tag{18}$$

Note that this is different to the interaction information [23] or conditional mutual information, which are defined (for the case of 3 variables) as

$$I(x_1, x_2, x_3) = -\sum_i H(x_i) + \sum_{i \neq j} H(x_i, x_j) - H(x_1, x_2, x_3)) \tag{19}$$

With a differing sign convention for odd variables. This type of multivariate mutual information can be negative.

It is thus clear that we can reduce the problem of multi-variate GMM differential entropy estimation, to that of a sum of univariate GMM differential entropy estimation, for which we can apply numerical techniques such as quaderature and sampling in a computational efficient manner to solve the problem of Maximum entropy. Formally for an n dimensional GMM

$$H(x_1...x_n) \leq \sum_i^n H(x_i) \leq \sum_i^n \mathcal{S}(x_i) \tag{20}$$

Where $H(x)$ denotes the entropy,of the joint and the marginals respectively and $\mathcal{S}$ denotes the surrogate density of maximum entropy. In the case of the GMM

$$GMM(\vec{x}) = \sum_i \frac{w_i}{\sqrt{(2\pi)^k \Sigma_i}} \exp\left((\vec{x} - \vec{\mu_i})^t \Sigma_i^{-1} (\vec{x} - \vec{\mu_i}) \right) \tag{21}$$

we have an analytic expression for the marginal densities $p(x_i)$, where we simply drop the terms from the mean vector and covariance matrix, as can be shown by the definition of the multivariate normal and linear algebra and hence

$$GMM(x) = \sum_i \frac{w_i}{\sqrt{(2\pi)\sigma_i}} \exp\left(\frac{(x - \mu_i)^2}{2\sigma_i^2} \right) \tag{22}$$

IX. KL divergence between two GMMs

Consider two GMMs with true underlying density $p_1(x)$ and $p_2(x)$ and their Maximum Entropy surrogates $q_1(x)$ and $q_2(x)$, Then it is easy to see that

$$\mathcal{D}_{kl}(p_1||p_2) \approx \int q_1(x) \log \frac{q_1(x)}{q_2(x)} dx \tag{23}$$

Writing $q_1(x) = \exp(\sum_i \alpha_i x^i)$ and $q_2(x) = \exp(\sum_j \lambda_j x^j)$

$$\mathcal{D}_{kl}(p_1||p_2) \approx \mathcal{S}(p_1) - \sum_j \lambda_j \mathbb{E}_{p_2}(x^j) \tag{24}$$

Extending this notion to higher dimensions, we consider the surrogate to be the product of its maximally entropic marginals and hence we arrive at for an n^d GMM where we take 40 moments of its marginals

$$\mathcal{D}_{kl}(p_1(\vec{x})||p_2(\vec{x})) = -\sum_i^n \mathcal{S}(x_i) + \sum_i^n \sum_j^m \lambda_{i,j} \mathbb{E}_{p_2}(x_i^j) \tag{25}$$

X. Lagrangian Duality

Consider a generic optimization problem of the form,

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0,\ i = 1...m \\ &\text{subject to } h_i(x) = 0,\ i = 1...p \end{aligned} \tag{26}$$

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,

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right). \end{aligned} \tag{27}$$

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. [24]. 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,

$$\begin{aligned} &\text{minimize } \ f_0[p(x)] = \int p(x)\log p(x)dx \\ &\text{subject to } \ h_i[p(x)] = \int p(x)x^i dx - \mu_i = 0,\ i = 1...p. \end{aligned} \tag{28}$$

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,

$$\inf_{x\in\mathcal{D}=\bigcap_{i=0}^{m} f_i} L(x,\lambda,\nu) \leq \inf_{x\in\mathcal{D}=\bigcap_{i=0}^{m+1} f_i} L(x,\lambda,\nu) \tag{29}$$

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 (15) and (14) we necessarily reduce $\mathcal{D}_{kl}(p[x]||q[x])$, where $p[x], q[x]$ define the true eigenvalue and MaxEnt proposal distributions respectively.

XI. Algorithm

We apply a numerically stable MaxEnt Algorithm (algorithm 1) [25], 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 [26].

We follow the procedure from entropic trace estimation [27]. 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 1 Optimising the Coefficients of the MaxEnt Distribution

Input: Moments $\{\mu_i\}$, Tolerance ϵ
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**
  $\delta \leftarrow \log\left(\frac{\mu_i}{\int \lambda^i p(\lambda)d\lambda}\right)$
  $\alpha_i \leftarrow \alpha_i + \delta$
  $p(\lambda) \leftarrow p(\lambda|\alpha)$
  error $\leftarrow \max|\int \lambda^i p(\lambda)d\lambda - \mu_i|$
  $i \leftarrow \text{mod}(i+1, \text{length}(\mu))$

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 $\log|A|$

$B = A/\|A\|_2$
μ (moments)$\leftarrow$ StochasticTraceEstimation(B,k)
α (coefficients) $\leftarrow$ MaxEntOpt(μ,ϵ)
$p(\lambda) \leftarrow p(\lambda|\alpha)$
$\log|A| \leftarrow n\int \log(\lambda)p(\lambda)d\lambda + n\log(\|A\|_2)$

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 VII 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.

XII. Stochastic trace estimates

A key component of the computational complexity of Entropic trace estimation [27], 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?

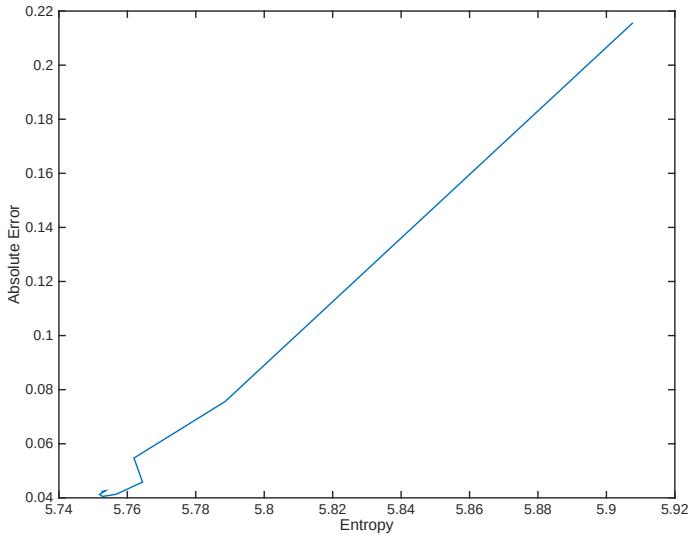

Fig. 1. Relative error against entropy for a single sample stochastic trace estimate from 2 to 10 moments for the Thermomech dataset.

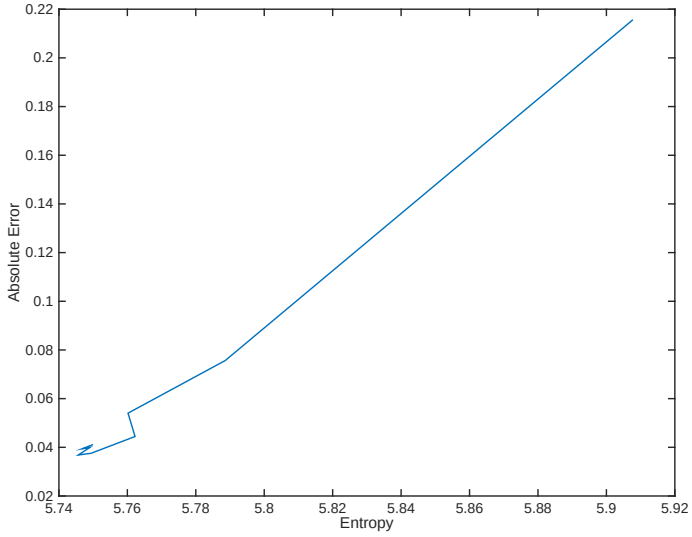

Fig. 2. Relative error against entropy for a 50 sample stochastic trace estimate from 2 to 10 moments for the Thermomech dataset.

To keep our results comparable and consistent, we keep with [27], [28] 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 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.

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).

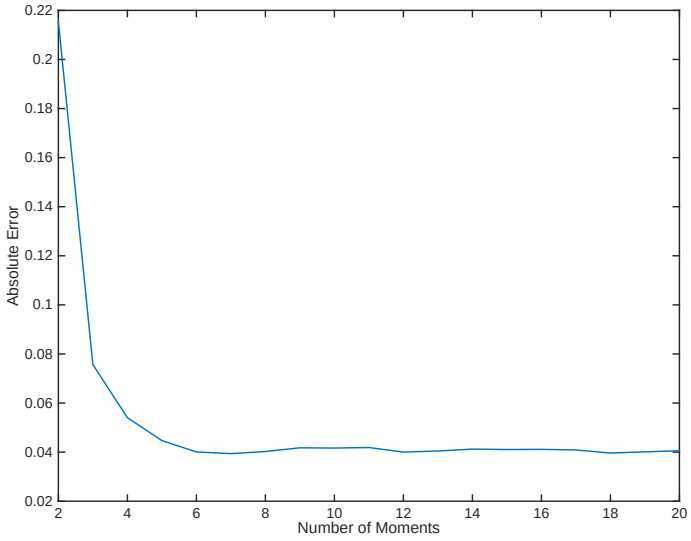

Fig. 3. Relative error against number of moments included for a single sample stochastic trace estimate for the Thermomech dataset. The corresponding graph for 50 samples is visually indistinguishable from this one.

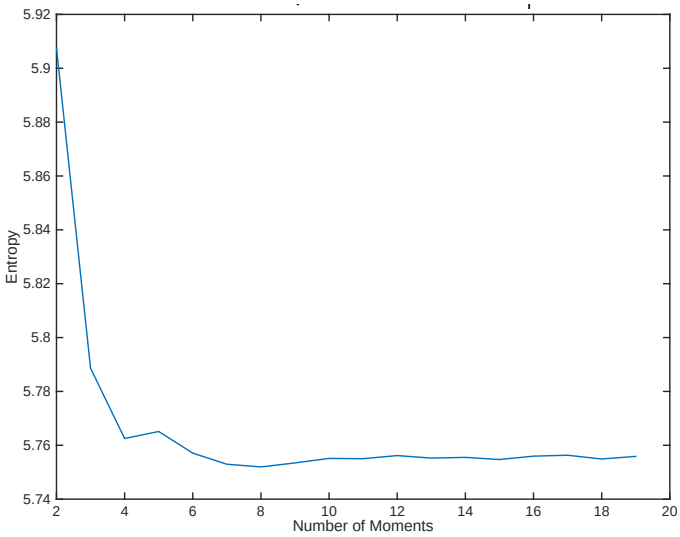

Fig. 4. Entropy against number of Moments for the Thermomech dataset

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

[2] Claude E Shannon. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1):3–55, 2001.

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

[4] RCH Cheng and NAK Amin. Estimating parameters in continuous univariate distributions with a shifted origin. Journal of the Royal Statistical Society. Series B (Methodological), pages 394–403, 1983.

[5] J Manyika and H Durrant-Whyte. Information as a basis for management and control in decentralized fusion architectures. In IEEE Conference on Decision and Control (CDC), 1992.

[6] Paul Viola and William M Wells III. Alignment by maximization of mutual information. International journal of computer vision, 24(2):137–154, 1997.

[7] Peder A Olsen and Satya Dharanipragada. An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models. In INTERSPEECH, 2003.

[8] Qiang Huo and Wei Li. A dtw-based dissimilarity measure for left-to-right hidden markov models and its application to word confusability analysis. In Ninth International Conference on Spoken Language Processing, 2006.

[9] Jorge Silva and Shrikanth Narayanan. Average divergence distance as a statistical discrimination measure for hidden markov models. IEEE Transactions on Audio, Speech, and Language Processing, 14(3):890–906, 2006.

[10] Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry, volume 191. American Mathematical Soc., 2007.

[11] Vladimir Maz'ya and Gunther Schmidt. On approximate approximations using gaussian kernels. IMA Journal of Numerical Analysis, 16(1):13–29, 1996.

[12] Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for gaussian mixture random vectors. In Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on, pages 181–188. IEEE, 2008.

[13] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 4, pages IV–317. IEEE, 2007.

[14] 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.

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

[16] 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.

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

[18] 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.

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

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

[21] 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.

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

[23] William McGill. Multivariate information transmission. Transactions of the IRE Professional Group on Information Theory, 4(4):93–111, 1954.

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

[25] 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.

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

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

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