Maximum Entropy Estimation via Gauss-LP Quadratures ⋆

Maxime Thély * **Tobias Sutter** *
Peyman Mohajerin Esfahani ** **John Lygeros** *

* *Automatic Control Laboratory, ETH Zürich (e-mail: mthely@student.ethz.ch, {sutter, lygeros}@control.ee.ethz.ch)*
** *Delft Center for Systems and Control, TU Delft, The Netherlands (e-mail: P.MohajerinEsfahani@tudelft.nl)*

Abstract: We present an approximation method to a class of parametric integration problems that naturally appear when solving the dual of the maximum entropy estimation problem. Our method builds up on a recent generalization of Gauss quadratures via an infinite-dimensional linear program, and utilizes a convex clustering algorithm to compute an approximate solution which requires reduced computational effort. It shows to be particularly appealing when looking at problems with unusual domains and in a multi-dimensional setting. As a proof of concept we apply our method to an example problem on the unit disc.

Keywords: Entropy maximization, convex clustering, linear programming, importance sampling

1. INTRODUCTION

Consider the problem where given a finite number of moments generated by an unknown probability density, we wish to estimate the unknown density. Obviously, this problem is under-determined and will have infinitely many solutions. To obtain a unique solution one can introduce a concave objective to be maximized. A natural choice for this is the information entropy, leading to the *MaxEnt density*, see Section 4 for a formal treatment of the problem.

The MaxEnt approach to density estimation and some of its remarkable properties has first been established by Jaynes in his seminal work Jaynes (1957). Since then it has found many important applications in various areas of physics, engineering, and systems biology in particular; see e.g. the discussions in Mead and Papanicolaou (1984); Sutter et al. (2015); Smadbeck and Kaznessis (2013). Its operational significance motivates the quest for efficient numerical methods to compute the MaxEnt density. The latter is the solution to an infinite-dimensional convex optimization problem that is as such intractable in general. It was shown in Csiszár (1975) that the MaxEnt distribution subject to a finite number of moment constraints, if it exists, belongs to the family of exponentials of polynomials. Its computation can thus be reduced to solving a system of nonlinear equations of dimension equal to the number of moment constraints. However, solving this system of nonlinear equations involves evaluating multi-dimensional integrals, a computationally difficult task in general. One way to overcome this issue is to approximate these by a quadrature rule, see for example Ormoneit and White (1999); Abramov et al. (2010). An alternative approach to approximate the MaxEnt density is presented in (Lasserre, 2010, Section 12.3), where by using duality of convex programming the problem is reduced to an unconstrained finite-dimensional convex optimization problem. An approximation hierarchy of the objective's gradient and Hessian in terms of two single semidefinite programs involving two linear matrix inequalities (LMI) is presented, where the desired accuracy is controlled by the size of the LMI constraints.

In this paper, we introduce a framework to numerically approximate a class of parametric integration problems using recent advances generalizing Gauss quadratures via an infinite-dimensional linear program (LP) Ryu and Boyd (2015). The difficulty resides in approximating the infinite-LP, as solving these problems exactly is computationally hard in general. Ryu and Boyd (2015) propose discretising the infinite-LP to a finite LP, that can be readily solved by standard solvers. We augment this technique by running a convex clustering algorithm which is built on the solution of a convex optimization problem, see Lashkari and Golland (2008). Embedded in an information theoretic context, the algorithm systematically filters out the most important discretization points, making it comparable to importance sampling Robert and Casella (2004). The reduction of the discretization set through the identification of the important exemplars is motivated by high-dimensional problems, where standard discretization techniques suffer from exponential growth. To complete the picture, we show that the dual program of the mentioned entropy maximization problem falls into the class of parametric integration problems addressed by the developed approximation method.

Section 2 formally presents the class of parametric integration problems that we aim to approximate. The ap-

⋆ Research was supported by the Swiss National Science Foundation under grant ”P2EZP2_165264” and by the European Commission under the project SPEEDD.

proximation method to these problems is introduced in Section 3. We show in Section 4 how the maximum entropy estimation problem can be addressed via the studied class of parametric integration problems. To illustrate the proposed methodology, the theoretical results are applied to a two-dimensional maximum entropy estimation problem, in Section 5. We conclude in Section 6 with a summary of our work and comment on possible topics of further research.

2. PROBLEM STATEMENT

Let $\eta \in \mathbb{R}^r$ be a known vector, μ a Borel measure defined on $\Omega \subset \mathbb{R}^d$ and consider the optimization problem

$$\min_{\lambda \in \mathbb{R}^r} \left\{ -\eta^\top \lambda + \log \int_\Omega f(\lambda, x)\, \mathrm{d}\mu(x) \right\}, \tag{1}$$

where λ is the r-dimensional decision variable and $f : \mathbb{R}^r \times \mathbb{R}^d \to \mathbb{R}$ a nonnegative function that is assumed to be convex and twice continuously differentiable in λ for fixed x. This type of objectives turns out to be of importance in a number of applications, among them the *maximum entropy estimation problem*, that aims to estimate a probability density supported on Ω only by knowledge of its first r moments; indeed in Section 4 we show that (1) is the dual of this problem. Note that if we set $\eta = 0$ and recall that the logarithm is a monotonic function, problem (1) reduces to

$$\min_{\lambda \in \mathbb{R}^r} \left\{ \int_\Omega f(\lambda, x)\, \mathrm{d}\mu(x) \right\}, \tag{2}$$

i.e., the minimization of an expected value cost in the presence of uncertainty distributed according to μ, also a problem of major importance, see the comprehensive monograph Shapiro et al. (2014) and the references therein.

To solve (1), we propose to replace the integral by a finite weighted sum of function evaluations

$$\int_\Omega f(\lambda, x)\, \mathrm{d}\mu(x) \approx \sum_{j=1}^m w_j f(\lambda, x_j). \tag{3}$$

As the sum inherits the structural properties of f in λ, (3) can be tackled by invoking methods from smooth convex optimization Nesterov (2004).

The cornerstone of the proposed approximation consists of finding the locations of the nodes x_j and their respective weights w_j. In this study, based on some recent work on extensions of Gauss quadrature Ryu and Boyd (2015), the nodes x_j and their weights w_j can be characterized as the solution to a semi-infinite LP. The key step of the presented approach is to use a convex clustering algorithm described in Lashkari and Golland (2008) to identify the k nodes which are the most important from an information theoretic perspective; we shall elaborate and provide greater details about this in Section 3.2.

3. METHODOLOGY

This section articulates in two parts. Based on Ryu and Boyd (2015), we first set the scene that allows us to find the nodes and weights of the Gauss-LP quadratures. By necessity we don't aim for a comprehensive treatment, but concentrate on the main points of central importance to our work. We then present the ideas behind the clustering algorithm used in this work.

3.1 Gauss-LP Quadratures

We start by briefly recalling traditional Gauss quadrature in one dimension before exposing an approach for a generalisation to multidimensional (potentially non-standard) domains.

A quadrature approximates an integral on the real interval $[-1, 1]$ by a finite sum of weighted function evaluations

$$\int_{-1}^{1} f(x)\, \mathrm{d}x \approx \sum_{j=1}^m w_j f(x_j), \tag{4}$$

where $x_j \in [-1, 1], j = 1, \ldots, m$ are called quadrature nodes and $w_j \geq 0$ their corresponding weights. Note that without loss of generality, the domain can be assumed to be equal to $[-1, 1]$ by making an appropriate change of integration variable. A quadrature is said to be of order n if it exactly computes the integral of polynomials up to degree $n - 1$, that is

$$\int_{-1}^{1} x^s\, \mathrm{d}x = \sum_{j=1}^m w_j x_j^s, \quad \text{for } s = 0, \ldots, n-1. \tag{5}$$

For an a priori fixed n, the whole difficulty resides in finding the appropriate number m, locations and weights w_j of the nodes x_j. Intuitively, the higher n is set, the more precise the resulting quadrature will be, but the number of required function evaluations m will increase as a consequence.

The *Gauss quadrature* is a unique set of nodes and weights such that $m = \frac{n}{2}$. For the standard interval $[-1, 1]$, locations of the Gauss nodes and values of their weights as a function of n can be looked up in a table, see for example Cheney and Kincaid (1980).

Unfortunately, classical Gauss quadrature does not easily extend to multiple dimensions and to non-standard domains (e.g., polytopes).

Ryu and Boyd (2015) proposed a way to extend the method, based on the fact that any quadrature can be interpreted as a measure μ with finite support, i.e.,

$$\begin{aligned} \int_{-1}^{1} f(x)\, \mathrm{d}x \approx \int_{-1}^{1} f(x)\, \mathrm{d}\mu = \int_{-1}^{1} f(x) (\sum_{j=1}^m w_j \delta_{x_j}) \\ = \sum_{j=1}^m w_j f(x_j), \end{aligned}$$

where δ_{x_j} denotes the Dirac measure. This way, the quadrature problem reduces to searching for a nonnegative Borel measure on $[-1, 1]$ that satisfies (5). Ryu and Boyd (2015) propose to pick the measure that will minimize its sensitivity to the polynomial of next degree, yielding the optimization problem

$$\begin{cases} \min\limits_{\mu} \int_{-1}^{1} x^n\, \mathrm{d}\mu \\ \text{s.t.} \quad \int_{-1}^{1} x^s\, \mathrm{d}\mu = \int_{-1}^{1} x^s dx, \text{ for } s = 0, \ldots, n-1 \\ \qquad \mu \geq 0, \end{cases} \tag{6}$$

where the constraints represent (5), and the last constraint ensures non-negativity of the weights.

Theorem 1. ((Ryu and Boyd, 2015, Theorem 1)). The linear program (6) admits a unique solution $\mu^\star = \sum_{j=1}^{n/2} w_j \delta_{x_j}$, where $w_1, \ldots, w_{n/2}$ and $x_1, \ldots, x_{n/2}$ are the weights and nodes of the Gauss quadrature.

Polynomials in (5) can be replaced by any linearly independent set of *test functions* $p_0, \ldots, p_{n-1}$, for example a sinusoidal basis. The objective can also be generalised; we denote this generalised integrand by Φ and, following Ryu and Boyd (2015), refer to it as the *sensitivity*. We can now rewrite (6) in a more general setting, yielding

$$\begin{cases} \min\limits_{\mu} & \int_\Omega \phi \, \mathrm{d}\mu \\ \text{s.t.} & \int_\Omega p_s \, \mathrm{d}\mu = \int_\Omega p_s \, \mathrm{d}x, \text{ for } s = 0, \ldots, n-1 \\ & \mu \geq 0. \end{cases} \tag{7}$$

We call a solution to (7) a *Gauss-LP quadrature*.

Theorem 2. ((Ryu and Boyd, 2015, Theorem 2)). The LP (7) admits a solution $\mu^\star$ supported on at most n points.

Problem (7) is infinite-dimensional and in general difficult to solve. It is therefore approximated by drawing a large set of points $\chi_M = \{x_1, \ldots, x_M\}$, sampled from a uniform density on Ω. The search for the decision variable is then restricted to those having finite support on χ_M, that is $\mu = \sum_{i=1}^M w_i \delta_{x_i}$, where δ denotes the Dirac measure. With this simplification, the problem (7) reduces to a finite-dimensional LP with decision variables $w_i, i = 1, \ldots, M$, which can be readily solved. By carefully choosing ϕ, one can guarantee that any basic feasible solution of the LP has at most n nonzero coefficients, which yields the corresponding support set $\chi_n = \{x_i \in \chi_M | w_i > 0\} \subseteq \chi_M$ with cardinality n.

3.2 Convex Clustering

To further reduce the cardinality of the samples set χ_n we propose to use the convex clustering method of Lashkari and Golland (2008). This will result in less function evaluations in (3) and therefore speed up the method. Consider the elements of χ_n as n realisations of a random variable with unknown distribution defined on $\mathbb{R}^d$. Lashkari and Golland (2008) extend classical mixture model clustering, considering all elements of χ_n as cluster-center candidates. Hence they look for the mixture distribution that maximizes the log-likelihood

$$l(q) = \frac{1}{n} \sum_{i=1}^{n} \log \sum_{j=1}^{n} q_j f_j(x_i), \tag{8}$$

where f_j is an exponential family member with expectation value $x_j \in \chi_n$ and $x_i \in \chi_n, i = 1, \ldots, n$. The convex combination of distributions that maximizes (8) is the distribution that most likely generates the data set χ_n. It is shown in Banerjee et al. (2005) that there is a bijection between exponential families and Bregman divergences [1]. Equation (8) can thus be reformulated as the optimization problem

$$\begin{cases} \max\limits_{q \in \mathbb{R}^n_{\geq 0}} & l(q) = \frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^n q_j e^{-\beta d(x_i, x_j)} \right) \\ \text{s.t.} & \sum_{i=1}^n q_j = 1, \end{cases} \tag{9}$$

where $d(x_i, x_j)$ is some Bregman divergence, and $x_i, x_j \in \chi_n$. The Euclidean distance is one example of a Bregman divergence and yields the normal distribution. What will result from (9) is a vector q with entry q_j positive if and only if x_j is a cluster-center. The parameter β controls the width of the clusters and therefore directly influences the sparsity of q, as shown in Figure 1 and Figure 2. Ultimately, we define $\chi_k = \{x_j \in \chi_n | q_j > 0\}$ as our final nodes set, composed only of the cluster-centers and of cardinality $k \leq n$.

This approach offers several advantages. Equation (9) is a convex optimization problem, making it possible to deploy algorithms that will converge to the global optimum. Moreover, there is no initial guess required as it would be the case in other clustering approaches such as *k-means*. Finally, the clustering approach of Lashkari and Golland (2008) has an information theoretic interpretation. Maximizing the log-likelihood function (9) is equivalent to a certain instance of the so-called rate-distortion problem, which considers an instance of lossy data compression, see Berger (1971) for a detailed treatment. Interestingly, this connection not only provides a rigorous quantification that higher number of clusters is the inherent cost of attaining a high objective function in (9), it also naturally suggests numerical algorithms derived for the rate-distortion problem to solve the optimization problem (9), such as the celebrated Blahut-Arimoto algorithm Blahut (1972); Arimoto (1972).

In view of the likelihood criteria, the proposed clustering approach can also be cast as a process to filter out important samples from a given dataset. Importance sampling is a well-known technique and has been extensively studied in the literature, see for instance Richard and Zhang (2007) in the context of high-dimensional integration and Kotecha and Djuric (2003) for building Gaussian particle filters to model uncertainty propagation in a dynamical environment. The idea of using an optimization-based process to filter out the important samples has recently received attention, see the comprehensive survey Dick et al. (2013) and the references therein. We believe that the proposed approach in this study falls into this category.

During the clustering process, we put the weights aside to focus on the locations of the nodes. We are thus left with unchanged weights w_j corresponding to the nodes $x_j \in \chi_k$ with $q_j > 0$, designated as cluster-centers. This quadrature does not fulfil the constraints of (7) anymore, and the weights need to be adjusted. Since the clustering reduced the cardinality of the nodes set, we now need to satisfy n constraints with the $k < n$ variables $w_1, \ldots, w_k$, which is impossible in general. We therefore seek to minimize the maximal violation of the constraints, that is

$$\min_{w \in \mathbb{R}^k} \max_{s \in \{0, \ldots, n-1\}} \left| \sum_{j=1}^{k} w_j p_s(x_j) - \int_\Omega p_s \, \mathrm{d}\Omega \right|. \tag{10}$$

This can be solved with a standard solver, and we eventually find our final nodes set $x_1, \ldots, x_k$ and respective weights $w_1, \ldots, w_k$.

4. MAXIMUM ENTROPY ESTIMATION

Assume we are given a family of moments, summarized by η, that are induced by an unknown probability density supported on a given set $\Omega \subset \mathbb{R}^d$. Given that the number of observed moments is finite, the problem of finding a density matching these moments is underdetermined and will have infinitely many solutions. To select among these candidate solutions one typically introduces a concave objective to be maximized by the unknown density. Here,

[1] See Banerjee et al. (2005) for a precise definition.

we introduce the multi-indexes $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbf{N}^d$ and select the Shannon entropy as the objective, defined for a given density p as $h(p) = \int_\Omega p(x) \log(p(x))\, \mathrm{d}x$, leading to the optimization problem

$$\begin{cases} \max\limits_{p \in \mathrm{L}_1(\Omega)} & h(p) \\ \text{s.t.} & \langle x^\alpha, p \rangle = \eta_\alpha, \quad \text{for all } |\alpha| \le r \\ & p \ge 0, \end{cases} \tag{11}$$

where we assume that $\eta_(0, \dots, 0) = 1$, which ensures that the solution is a valid probability density. Note that $|\alpha| = \sum_{\ell=1}^d \alpha_\ell$ and $x^\alpha = \prod_{\ell=1}^d \alpha_\ell$. We call a solution to (11) the *MaxEnt* density. By using (Sutter et al., 2015, Lemma 3.10) (that follows from Csiszár (1975)) and Sion's minimax theorem Sion (1958), one can show that the dual program to (11) is given by

$$\min_\lambda \left\{ -\sum_{|\alpha| \le r} \eta_\alpha \lambda_\alpha + \log \int_\Omega \exp\left(\sum_{|\alpha| \le r} x^\alpha \lambda_\alpha \right) \mathrm{d}x \right\}, \tag{12}$$

and that strong duality holds. Note that the vector of all monomials x^α of degree less than or equal to r has dimension $s(r) := \binom{r+d}{r}$. Let us enumerate the multi-indexes as $\alpha^1, \dots, \alpha^{s(r)}$ and denote $\lambda_i := \lambda_{\alpha^i}$ and as such consider $\lambda \in \mathbb{R}^{s(r)}$ as the decision variable in the dual program. Similarly let us denote $\eta_i := \eta_{\alpha^i}$. Moreover, let us denote the dual objective function by $F(\lambda)$, such that (12) reads as $\min_{\lambda \in \mathbb{R}^{s(r)}} F(\lambda)$. Note that the dual program (12) has exactly the structure of problem (1).

We aim to solve the dual program (12) with the Newton method

$$\lambda^{(k+1)} = \lambda^{(k)} - H(\lambda^{(k)})^{-1}(g(\lambda^{(k)}) - \eta). \tag{13}$$

The Hessian $H(\lambda^{(k)}) \in \mathbb{R}^{s(r) \times s(r)}$ and the gradient $g(\lambda^{(k)}) \in \mathbb{R}^{s(r)}$ are then approximated by

$$\begin{aligned} g(\lambda)_i &= \frac{\sum_{j=1}^m w_j x_j^{\alpha_i} \exp\left(\sum_{|\alpha| \le r} x_j^\alpha \lambda_\alpha\right)}{\sum_{j=1}^m w_j \exp\left(\sum_{|\alpha| \le r} x_j^\alpha \lambda_\alpha\right)} \\ H(\lambda)_{iq} &= g(\lambda)_{i+q} - g(\lambda)_i g(\lambda)_q, \end{aligned} \tag{14}$$

for $i, q = 1, \dots, s(r)$, where the nodes x_j and the respective weights w_j are found using the methodology presented in Section 3.

Remark 3. (Computational stability). The evaluation of the gradient and Hessian required in the Newton method (13) involves the term (14). Note that a straightforward computation of the gradient and Hessian via (14) is numerically difficult. In the light of (Nesterov, 2005, p. 148), we present a numerically stable technique for computing the term (14).

We consider the functions $\mathbb{R}^{s(r)} \ni \lambda \mapsto f_j(\lambda) = \sum_{|\alpha| \le n} x_j^\alpha \lambda_\alpha \in \mathbb{R}$, $\bar{f}(\lambda) := \max_{j=1,\dots,m} f_j(\lambda)$ and $\mathbb{R}^{s(r)} \ni \lambda \mapsto \varphi_j(\lambda) = f_j(\lambda) - \bar{f}(\lambda) \in \mathbb{R}$, such that all components of $\varphi_j(\lambda)$ are non-positive. One can show that the term (14) is equivalent to

$$g(\lambda)_i = \frac{\sum_{j=1}^m w_j e^{\varphi_j(\lambda)} \frac{\partial}{\partial \lambda_i} \varphi_j(\lambda)}{\sum_{j=1}^m w_j e^{\varphi_j(\lambda)}} + \frac{\partial}{\partial \lambda_i} \bar{f}(\lambda)$$

which can be computed with a small numerical error.

5. SIMULATION RESULTS

Three important features of the technique are illustrated through an example problem on the unit disc. First, we solve Problem (7), then present the clustering algorithm, in particular how β alone controls the amount of clusters k, and finally see how the resulting Gauss-LP quadratures (with k clusters) perform on the maximum entropy estimation problem.

Consider problem (7) where $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \le 1\}$. We follow the procedure described at the end of Section 3.1, extract $M = 20000$ samples x_i uniformly on Ω and thus define $\chi_M = \{x_1, \dots, x_M\}$. Further, we choose the two-dimensional test functions $p_{ij} = x^i y^j$ for $i + j < 7$, leading to $n = \frac{7(7+1)}{2} = 28$. Heuristically, we know that the sensitivity function $\phi = x^7 + y^7$ yields a sparse w. The resulting problem is an LP with M variables and n constraints:

$$\begin{cases} \min\limits_{w \in \mathbb{R}^M_{\ge 0}} & \sum_{\ell=1}^M w_\ell (x_\ell^7 + y_\ell^7) \\ \text{s.t.} & \sum_{\ell=1}^M w_\ell x_\ell^i y_\ell^j = \int_\Omega x^i y^j \, \mathrm{d}\Omega, \ \text{for } i + j < 7 \\ & w_\ell \ge 0, \end{cases}$$

which can be solved with a standard LP solver (e.g., `MOSEK`), yielding the upper left repartition of nodes in Figure 1. As expected from Theorem 2, the solution is sparse with $\mathtt{nnz}(w) = 28$.

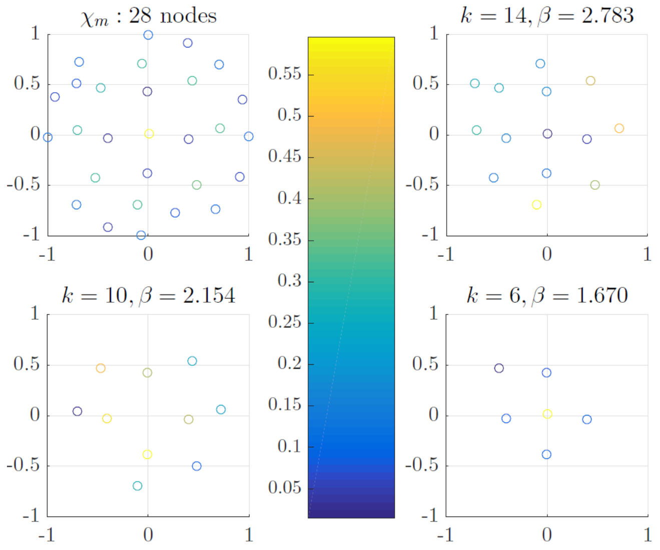


Fig. 1. Results of the clustering algorithm for different values of β. The colorbar represents the scale of the weights. For clarity only nodes with nonzero weights are depicted. The upper left graph shows the unclustered set χ_m as a reference.

In a second step, we run the clustering algorithm introduced in Section 3.2 for several values of β, and report the results in Figure 1. We observe that the sparsity of q, represented by k, grows monotonically with β, as mentioned in Section 3.2, see Figure 2.

We then regularize the weights according to Section **??**. Finally, we tackle the maximum entropy estimation problem on the unit disc as special case of (P). Let $r = 2$ and thus $s(r) = 6$, and let the vector of moments be given by the moments of the uniform density on the unit disc.

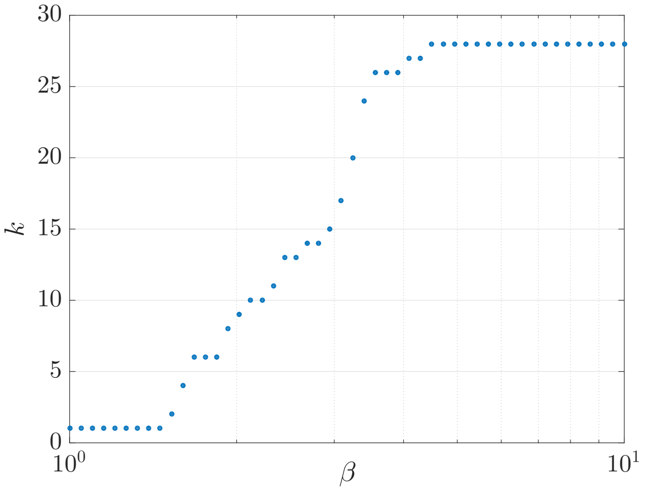


Fig. 2. The number of clusters k grows monotonically in the cluster-width controlling parameter β.

$$\eta_{ij} = \int_\Omega \mathcal{U}(\Omega) x^i y^j \,\mathrm{d}x\,\mathrm{d}y = \frac{1}{\pi}\int_\Omega x^i y^j \,\mathrm{d}x\,\mathrm{d}y.$$

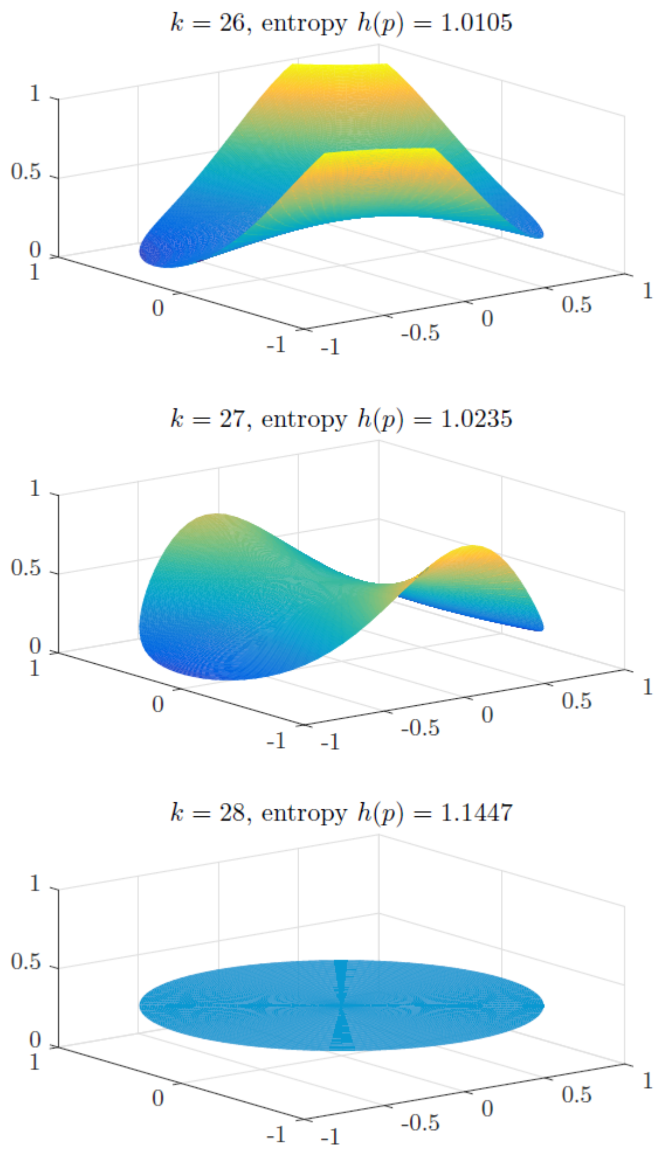
i	0	0	0	1	1	2
j	0	1	2	0	1	0
η_{ij}	1	0	1/4	0	0	1/4

The uniform density is known to be the maximum entropy distribution $h(\mathcal{U}(\Omega)) = -\int_\Omega \frac{1}{\pi}\log(\frac{1}{\pi})\,\mathrm{d}\Omega = \log\pi \simeq 1.1447$. Figure 3 shows how our approximations perform for $n = 28$ constraints and $k \leq n$. We see that the scheme developed in this work recovers the uniform density perfectly for $k = n$, that is no clustering. As soon as $k < n$, we see that the computational savings offered by the clustering algorithm come with a loss of accuracy in the entropy. We also see that the approximated maximum entropy distribution for $k < n$ has a very different shape from the uniform density. This is due to the fact that, by approximating integrals with finite sums, we are no longer optimizing the entropy, but a surrogate, leading to different solutions. In fact, this phenomenon is not specific to this particular example, it has been observed while conducting tests on different domains Ω and moments vectors η. Nevertheless the approximated entropy is indeed approaching the desired maximum entropy.

6. CONCLUSION

We presented a simple approximation scheme to a class of parametric integration problems we showed to appear when one wants to solve the dual of the maximum entropy estimation problem. Starting from a recent generalization of Gauss quadratures, we augmented the method by running a convex clustering algorithm in order to bring out the most important nodes of the quadrature. We thus paved the way towards approximate solutions with reduced computational cost. The method is particularly appealing when looking at problems with unusual domains and in a multi-dimensional setting.

We showed how the scheme performs in a two-dimensional context. The extension of the Gauss quadrature yields exact solution of the maximum entropy problem, see Figure 3. For a fixed exactness degree (set by n being the number of initial Gauss nodes), we ran the clustering algorithm for different cluster-sizes, resulting in improving estimates of the entropy as k (the number of clusters) goes to n.


$k = 26$, entropy $h(p) = 1.0105$

$k = 27$, entropy $h(p) = 1.0235$

$k = 28$, entropy $h(p) = 1.1447$

Fig. 3. Approximation of the uniform density on the unit disc for different values of $k \leq n$, where $n = 28$. The entropy $h(p)$ is growing as k goes to n, and is equal to $h(\mathcal{U}(\Omega))$ for $k = n$.

For future work, we opt for a rigorous framework to quantify the approximation error introduced by the proposed clustering approach as well as to better understand the sparsity phenomenon related to the choice of the β parameter. Moreover, we plan to apply the presented methodology to particular high-dimensional examples in the context of systems biology, where the MaxEnt densities are the key objects in the so-called *moment closure method* to approximate the chemical master equation Smadbeck and Kaznessis (2013).

An important research direction to investigate further would also be the complexity and runtime comparison between the presented approach and exisiting numerical schemes for the maximum entropy estimation problem, such as Lasserre's SDE approach (Lasserre, 2010, Section 12.3). Infinite-dimensional linear programs of the type (7) naturally appear in the context of the so-called static, convex-analytic formulation of Markov decision processes. The approximation of these linear programs by means of a finite program is the core of a methodology known as approximate dynamic programming, see Bertsekas and Tsitsiklis (1996); Hernández-Lerma and Lasserre (1999). Recent developments in the performance of the first order convex optimization methods Mohajerin Esfahani et al. (2017), as well as randomized optimization Mohajerin Esfahani et al. (2015), are among approaches that we aim to use as a tool to investigate this connection in our subsequent works.

REFERENCES

Abramov, R.V. et al. (2010). The multidimensional maximum entropy moment problem: A review of numerical methods. *Communications in Mathematical Sciences*, 8(2), 377–392.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. 18(1), 14–20. doi:10.1109/TIT.1972.1054753.

Banerjee, A., Merugu, S., Dhillon, I.S., and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct), 1705–1749.

Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall electrical engineering series. Prentice-Hall. URL http://books.google.ch/books?id=-HV1QgAACAAJ.

Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Blahut, R.E. (1972). Computation of channel capacity and rate-distortion functions. 18(4), 460–473. doi:10.1109/TIT.1972.1054855.

Cheney, W. and Kincaid, D. (1980). *Numerical mathematics and computing*. Brooks/Cole Publishing Co., Monterey, Calif. Contemporary Undergraduate Mathematics Series.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3, 146–158.

Dick, J., Kuo, F.Y., and Sloan, I.H. (2013). High-dimensional integration: The quasi-monte carlo way. *Acta Numerica*, 22, 133–288. doi:10.1017/S0962492913000044. URL http://journals.cambridge.org/article_S0962492913000044.

Hernández-Lerma, O. and Lasserre, J. (1999). *Further topics on discrete-time Markov control processes*. Applications of Mathematics Series. Springer.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, 106, 620–630. doi:10.1103/PhysRev.106.620. URL http://link.aps.org/doi/10.1103/PhysRev.106.620.

Kotecha, J.H. and Djuric, P.M. (2003). Gaussian particle filtering. *IEEE Transactions on signal processing*, 51(10), 2592–2601.

Lashkari, D. and Golland, P. (2008). Convex clustering with exemplar-based models. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis (eds.), *Advances in Neural Information Processing Systems 20*, 825–832. Curran Associates, Inc.

Lasserre, J.B. (2010). *Moments, Positive Polynomials and Their Applications*, volume 1 of *Imperial College Press Optimization Series*. Imperial College Press, London.

Mead, L.R. and Papanicolaou, N. (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8).

Mohajerin Esfahani, P., Sutter, T., Kuhn, D., and Lygeros, J. (2017). From Infinite to Finite Programs: Explicit Error Bounds with Applications to Approximate Dynamic Programming. *ArXiv e-prints*.

Mohajerin Esfahani, P., Sutter, T., and Lygeros, J. (2015). Performance bounds for the scenario approach and an extension to a class of non-convex programs. *IEEE Transactions on Automatic Control*, 60(1), 46–58. doi:10.1109/TAC.2014.2330702. URL http://dx.doi.org/10.1109/TAC.2014.2330702.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 127–152. doi:10.1007/s10107-004-0552-5. URL http://dx.doi.org/10.1007/s10107-004-0552-5.

Ormoneit, D. and White, H. (1999). An efficient algorithm to compute maximum entropy densities. *Econometric Reviews*, 18(2), 127–140.

Richard, J.F. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *J. Econometrics*, 141(2), 1385–1411. doi:10.1016/j.jeconom.2007.02.007. URL http://dx.doi.org/10.1016/j.jeconom.2007.02.007.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2 edition.

Ryu, E.K. and Boyd, S.P. (2015). Extensions of gauss quadrature via linear programming. *Foundations of Computational Mathematics*, 15(4), 953–971. doi:10.1007/s10208-014-9197-9. URL http://dx.doi.org/10.1007/s10208-014-9197-9.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on Stochastic Programming*. SIAM, second edition.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, 8, 171–176.

Smadbeck, P. and Kaznessis, Y.N. (2013). A closure scheme for chemical master equations. *Proceedings of the National Academy of Sciences*, 110(35), 14261–14265. doi:10.1073/pnas.1306481110. URL http://www.pnas.org/content/110/35/14261.abstract.

Sutter, T., Sutter, D., Mohajerin Esfahani, P., and Lygeros, J. (2015). Efficient approximation of channel capacities. *IEEE Transactions on Information Theory*, 61(4), 1649–1666. doi:10.1109/TIT.2015.2401002.