Generalized maximum entropy estimation

Tobias Sutter,1,* David Sutter,2,† Peyman Mohajerin Esfahani,3,‡ and John Lygeros1,*

1Automatic Control Laboratory, ETH Zurich, Switzerland
2Institute for Theoretical Physics, ETH Zurich, Switzerland
3Delft Center for Systems and Control, TU Delft, The Netherlands

We consider the problem of estimating a probability distribution that maximizes the entropy while satisfying a finite number of moment constraints, possibly corrupted by noise. Based on duality of convex programming, we present a novel approximation scheme using a smoothed fast gradient method that is equipped with explicit bounds on the approximation error. We further demonstrate how the presented scheme can be used for approximating the chemical master equation through the zero-information moment closure method.

1. INTRODUCTION

Consider a one-dimensional moment problem formulated as follows: Given a set $\mathbb{K} \subset \mathbb{R}$ and a sequence $(y_i)_{i \in \mathbb{N}} \subset \mathbb{R}$ of moments, does there exist a measure $\mu$ supported on $\mathbb{K}$ such that

$$y_i = \int_{\mathbb{R}} x^i \mu(dx) \quad \text{for all } i \in \mathbb{N} \quad (1)$$

For $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = [a, b]$ with $-\infty < a < b < \infty$ the above moment problem is known as the Hamburger moment problem and Hausdorff moment problem, respectively. If the moment sequence is finite, the problem is called a truncated moment problem. In both full and truncated cases, a measure $\mu$ that satisfies (1), is called a representing measure of the sequence $(y_i)_{i \in \mathbb{N}}$. If a representing measure is unique, it is said to be determined by its moments. From the Stone-Weierstrass theorem it follows directly that every non-truncated representing measure with compact support is determined by its moments. In the Hamburger moment problem, given a representing measure $\mu$ for a moment sequence $(y_i)_{i \in \mathbb{N}}$, a sufficient condition for $\mu$ being determined by its moments is the so-called Carleman condition, i.e., $\sum_{i=1}^{\infty} y_i^{-1/2i} = \infty$. Roughly speaking this says that the moments should not grow too fast, see [1] for further details. For the Hamburger and the Hausdorff moment problem, there are necessary and sufficient conditions for the existence of a representing measure for a given moment sequence $(y_i)_{i \in \mathbb{N}}$ in both the full as well as the truncated setting, see [22, Theorems 3.2, 3.3, 3.4].

In this article, we investigate the problem of estimating an unknown probability distribution given a finite number of (possibly noisy) observed moments. Given that the observed moments are consistent (i.e., there exists a probability distribution that satisfies all the moment constraints), the problem is underdetermined and has infinitely many solutions. This therefore raises the question of which solution to choose. A natural choice would be to pick the one with the highest entropy, called the MaxEnt distribution. The main reason why the MaxEnt distribution is a natural choice is due to a concentration phenomenon described by Jaynes [19]:

“If the information incorporated into the maximum-entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally.”

See [17, 19] for a rigorous statement. This maximum entropy estimation problem, subject to moment constraints, also known as the principle of maximum entropy, is applicable to large classes
of problems in natural and social sciences — in particular in economics, see [16] for a comprehensive survey. Also it has important applications in approximation methods to dynamical objects, such as in systems biology, where MaxEnt distributions are key objects in the so-called moment closure method to approximate the chemical master equation [37], or more recently in the context of approximating dynamic programming [25] where MaxEnt distributions act as a regularizer, leading to computationally more efficient optimization programs.

Their operational significance motivates the study of numerical methods to compute MaxEnt distributions, which are the solutions of an infinite-dimensional convex optimization problem and as such computationally intractable in general. Since it was shown that the MaxEnt distribution maximizing the entropy subject to a finite number of moment constraints (if it exists) belongs to the exponential family of distributions [7], its computation can be reduced to solving a system of nonlinear equations, whose dimension is equal to the number of moment constraints [24]. Furthermore, the system of nonlinear equations involves evaluating integrals over $\mathbb{K}$ that are computationally difficult in general. Even if the support set $\mathbb{K}$ is finite, finding the MaxEnt distribution is not straightforward, since solving a system of nonlinear equations can be computationally demanding.

In [22, Section 12.3] it is shown that the maximum entropy subject to moment constraints can be approximated by using duality of convex programming. The problem can be reduced to an unconstrained finite-dimensional convex optimization problem and an approximation hierarchy of its gradient and Hessian in terms of two single semidefinite programs involving two linear matrix inequalities is presented. The desired accuracy is controlled by the size of the linear matrix inequalities constraints. The method seems to be powerful in practice, however a rate of convergence has not been proven. Furthermore, it is not clear how the method extends to the case of uncertain moment constraints. In a finite dimensional setting, [11] presents a treatment of the maximum entropy principle with generalized regularization measures, that as a special case contain the setting presented here. However, convergence rates of algorithms presented are not known and again it is not clear how the method extends to the case of uncertain moment constraints.

In this article, we present a new approximation scheme to minimize the relative entropy subject to noisy moment constraints. This is a generalization of the introduced maximum entropy problem and extends the principle of maximum entropy to the so-called principle of minimum discriminating information [20]. We show that its dual problem exhibits a particular favourable structure that allows us to apply Nesterov’s smoothing method [28] and hence tackle the presented problem using a fast gradient method obtaining process convergence properties, unlike [22].

In many applications, it is important to efficiently compute the MaxEnt distribution. For example, the zero-information moment closure method [37], as well as a recently developed method to approximate the channel capacity of a large class of memoryless channels [39] deal with iterative algorithms that require the numerical computation of the MaxEnt distribution in each iteration step.

**Structure.** The layout of this paper is as follows: In Section 2 we formally introduce the problem setting. Our results on an approximation scheme in a continuous setting are reported in Section 3. In Section 4, we show how these results simplify in the finite-dimensional case. Section 5 discusses the gradient approximation that is the dominant step of the proposed approximation method from a computational perspective. The theoretical results are applied in Section 6 to the zero-information moment closure method. We conclude in Section 7 with a summary of our work and comment on possible subjects of further research.

**Notation.** The logarithm with basis 2 and e is denoted by $\log(\cdot)$ and $\ln(\cdot)$, respectively. We define the standard $n$–simplex as $\Delta_n := \{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i = 1\}$. For a probability mass function $p \in \Delta_n$ we denote its entropy by $H(p) := \sum_{i=1}^n -p_i \log p_i$. Let $B(y, r) := \{x \in \mathbb{R}^n : \|x - y\|_2 \leq r\}$ denote the ball with radius $r$ centered at $y$. Throughout this article, measurability
always refers to Borel measurability. For a probability density \( p \) supported on a measurable set \( B \subset \mathbb{R} \) we denote the differential entropy by \( h(p) := -\int_B p(x) \log p(x) \, dx \). For \( A \subset \mathbb{R} \) and \( 1 \leq p \leq \infty \), let \( L^p(A) \) denote the space of \( L^p \)-functions on the measure space \((A, \mathcal{B}(A), dx)\), where \( \mathcal{B}(A) \) denotes the Borel \( \sigma \)-algebra and \( dx \) the Lebesgue measure. Let \( X \) be a compact metric space, equipped with its Borel \( \sigma \)-field \( \mathcal{B}(\cdot) \). The space of all probability measures on \((X, \mathcal{B}(X))\) will be denoted by \( \mathcal{P}(X) \). The relative entropy (or Kullback-Leibler divergence) between any two probability measures \( \mu, \nu \in \mathcal{P}(X) \) is defined by

\[
D(\mu \| \nu) := \begin{cases} 
\int_X \log \left( \frac{d\mu}{d\nu} \right) \, d\mu, & \text{if } \mu \ll \nu \\
+\infty, & \text{otherwise},
\end{cases}
\]

where \( \ll \) denotes absolute continuity of measures, and \( \frac{d\mu}{d\nu} \) is the Radon-Nikodym derivative. The relative entropy is non-negative, and is equal to zero if and only if \( \mu \equiv \nu \). Let \( X \) be restricted to a compact metric space and let us consider the pair of vector spaces \((\mathcal{M}(X), \mathcal{B}(X))\) where \( \mathcal{M}(X) \) denotes the space of finite signed measures on \( \mathcal{B}(X) \) and \( \mathcal{B}(X) \) is the Banach space of bounded measurable functions on \( X \) with respect to the sup-norm and consider the bilinear form

\[
\langle \mu, f \rangle := \int_X f(x) \mu(dx).
\]

This induces the total variation norm as the dual norm on \( \mathcal{M}(X) \), since by [18, p.2]

\[
\|\mu\|_* = \sup_{\|f\|_{\infty} \leq 1} \langle \mu, f \rangle = \|\mu\|_{TV},
\]

making \( \mathcal{M}(X) \) a Banach space. In the light of [18, p. 206] this is a dual pair of Banach spaces; we refer to [3, Section 3] for the details of the definition of dual pairs.

## 2. PROBLEM STATEMENT

Let \( \mathbb{K} \subset \mathbb{R} \) be compact and consider the scenario where a probability measure \( \mu \in \mathcal{P}(\mathbb{K}) \) is unknown and only observed via the following measurement model

\[
y_i = \langle \mu, x^i \rangle + u_i, \quad u_i \in \mathcal{U}_i \quad \text{for } i = 1, \ldots, M,
\]

where \( u_i \) represents the uncertainty of the obtained data point \( y_i \) and \( \mathcal{U}_i \subset \mathbb{R} \) is compact, convex and \( 0 \in \mathcal{U}_i \) for all \( i = 1, \ldots, M \). Given the data \( \{y_i\}_{i=1}^M \subset \mathbb{R} \), the goal is to estimate a probability measure \( \mu \) that is consistent with the measurement model \((2)\). This problem (given that \( M \) is finite) is underdetermined and has infinitely many solutions. Among all possible solutions for (2), we aim to find the solution that maximizes the entropy. Define the set \( T := \times_{i=1}^M \{y_i - u : u \in \mathcal{U}_i\} \subset \mathbb{R}^M \) and the linear operator \( A : \mathcal{M}(\mathbb{K}) \to \mathbb{R}^M \) by

\[
(A\mu)_i := \langle \mu, x^i \rangle = \int_{\mathbb{K}} x^i \mu(dx) \quad \text{for all } i = 1, \ldots, M.
\]

The operator norm is defined as \( \|A\| := \sup_{\|\mu\|_{TV} = 1, \|y\|_2 = 1} \langle A\mu, y \rangle \). Note that due to the compactness of \( \mathbb{K} \) the operator norm is bounded, see Lemma 3.3 for a formal statement. The adjoint operator to \( A \) is given by \( A^* : \mathbb{R}^M \to \mathcal{B}(\mathbb{K}) \), where \( A^* z(x) := \sum_{i=1}^M z_i x^i \); note that the domain and image spaces of the adjoint operator are well defined as \((\mathcal{B}(\mathbb{K}), \mathcal{M}(\mathbb{K}))\) is a topological dual pairs and the operator \( A \) is bounded [18, Proposition 12.2.5].

Given a reference measure \( \nu \in \mathcal{P}(\mathbb{K}) \), the problem of minimizing the relative entropy subject to moment constraints \((2)\) can be formally described by

\[
J^* = \min_{\mu \in \mathcal{P}(\mathbb{K})} \{ D(\mu \| \nu) : A\mu \in T \}.
\]
**Proposition 2.1** (Existence & uniqueness of (3)). The optimization problem (3) attains an optimal feasible solution that is unique.

**Proof.** The variational representation of the relative entropy [6, Corollary 4.15] states that the mapping $\mu \mapsto D(\mu\|\nu)$ is the Fenchel-Legendre dual of the mapping $X \mapsto \log \mathbb{E}e^X$, where $X$ is a random variable with law $\nu$. As a basic property of the Legendre-Fenchel transform, the mapping $\mu \mapsto D(\mu\|\nu)$ therefore is lower-semicontinuous [23]. Note also that the space of probability measures on $\mathbb{K}$ is compact [2, Theorem 15.11]. Moreover, since the linear operator $\mathcal{A}$ is bounded, it is continuous. As a result, the feasible set of problem (3) is compact and hence the optimization problem attains an optimal solution. Finally, the strict convexity of the relative entropy [7] ensures uniqueness of the optimizer.

Note that if $\mathcal{U}_i = \{0\}$ for all $i = 1, \ldots, M$, i.e., there is no uncertainty in the measurement model (2), Proposition 2.1 reduces to a known result [7]. Consider the special case where the reference measure $\nu$ is the uniform measure on $\mathbb{K}$ and let $p$ denote the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ (whose existence can be assumed without loss of generality). Since $\mathcal{A}$ is weakly continuous and the differential entropy is known to be weakly lower semi-continuous [6], we can restrict attention to a (weakly) dense subset of the feasible set and hence assume without loss of generality that $p \in L^1(\mathbb{K})$. Problem (3) then reduces to

$$\max_{p \in L^1(\mathbb{K})} \left\{ \langle h(p) : \mathcal{A}p(x)dx \in T \rangle, \langle dx, p \rangle = 1 \right\}. \tag{4}$$

Problem (4) is a generalized maximum entropy estimation problem that, in case $\mathcal{U}_i = \{0\}$ for all $i = 1, \ldots, M$, simplifies to the standard entropy maximization problem subject to $M$ moment constraints. In this article, we present a new approach to solve (3) that is based on its dual formulation. It turns out that the dual problem of (3) has a particular structure that allows us to apply Nesterov’s smoothing method [28]. Furthermore, we will show how an $\varepsilon$-optimal primal solution can be reconstructed. This is done by solving the dual problem and comparing with the existing approach in this context, it additionally requires a second smoothing step that is motivated by [8]. The problem of entropy maximization subject to uncertain moment constraints (4) can be seen as a special case of (3).

### 3. RELATIVE ENTROPY MINIMIZATION

We start by recalling that an unconstrained minimization of the relative entropy with an additional linear term in the cost admits a closed form solution. Let $c \in \mathbb{B}(\mathbb{K}), \nu \in \mathcal{P}(\mathbb{K})$ and consider the optimization problem

$$\min_{\mu \in \mathcal{P}(\mathbb{K})} \left\{ D(\mu\|\nu) - \langle \mu, c \rangle \right\}. \tag{5}$$

**Lemma 3.1** (Gibbs distribution). The unique optimizer to problem (5) is given by the Gibbs distribution, i.e.,

$$\mu^*(dx) = \frac{2^c(x)\nu(dx)}{\int_{\mathbb{K}} 2^c(x)\nu(dx)} \quad \text{for } x \in \mathbb{K},$$

which leads to the optimal value of $-\log \int_{\mathbb{K}} 2^c(x)\nu(dx)$.

**Proof.** The result is standard and follows from [7] or alternatively by [39, Lemma 3.10]. □
Let \( \mathbb{R}^M \ni z \mapsto \sigma_T(z) := \max_{x \in T} \langle x, z \rangle \in \mathbb{R} \) denote the support function of \( T \), which is continuous since \( T \) is compact [32, Corollary 13.2.2]. The primal-dual pair of problem (3) can be stated as

\[
\begin{align*}
(\text{primal program}) & : \quad J^* = \min_{\mu \in \mathcal{P}(\mathbb{K})} \left\{ D(\mu \| \nu) + \sup_{z \in \mathbb{R}^M} \left\{ \langle \mathcal{A}_\mu, z \rangle - \sigma_T(z) \right\} \right\} \\
(\text{dual program}) & : \quad J^*_D = \sup_{z \in \mathbb{R}^M} \left\{ -\sigma_T(z) + \min_{\nu \in \mathcal{P}(\mathbb{K})} \left\{ D(\mu \| \nu) + \langle \mathcal{A}_\mu, z \rangle \right\} \right\},
\end{align*}
\]

where the dual function is given by

\[
F(z) = -\sigma_T(z) + \min_{\mu \in \mathcal{P}(\mathbb{K})} \left\{ D(\mu \| \nu) + \langle \mathcal{A}_\mu, z \rangle \right\}.
\]

Note that the primal program (6) is an infinite-dimensional convex optimization problem. The key idea of our analysis is driven by Lemma 3.1 indicating that the dual function, that involves a minimization running over an infinite-dimensional space, is analytically available. As such, the dual problem becomes an unconstrained finite-dimensional convex optimization problem, which is amenable to first-order methods.

**Lemma 3.2** (Zero duality gap). There is no duality gap between the primal program (6) and its dual (7), i.e., \( J^* = J^*_D \). Moreover, if there exists \( \hat{\mu} \in \mathcal{P}(\mathbb{K}) \) such that \( \mathcal{A}\hat{\mu} \in \text{int}(T) \), then the set of optimal dual variables in (7) is compact.

**Proof.** Recall that the relative entropy is known to be lower semicontinuous and convex in the first argument, which can be seen as a direct consequence of the duality relation for the relative entropy [6, Corollary 4.15]. Hence, the desired zero duality gap follows by Sion’s minimax theorem [35, Theorem 4.2]. The compactness of the set of dual optimizers is due to [4, Proposition 5.3.1]. \( \square \)

Because the dual function (8) turns out to be non-smooth, in the absence of any additional structure, the efficiency estimate of a black-box first-order method is of order \( O(1/\varepsilon^2) \), where \( \varepsilon \) is the desired absolute additive accuracy of the approximate solution in function value [27]. We show, however, that the generalized entropy maximization problem (6) has a certain structure that allows us to deploy the recent developments in [28] for approximating non-smooth problems by smooth ones, leading to an efficiency estimate of order \( O(1/\varepsilon) \). This, together with the low complexity of each iteration step in the approximation scheme, offers a numerical method that has an attractive computational complexity. In the spirit of [8, 28], we introduce a smoothing parameter \( \eta := (\eta_1, \eta_2) \in \mathbb{R}^2_{>0} \) and consider a smooth approximation of the dual function

\[
F_\eta(z) := -\max_{x \in T} \left\{ \langle x, z \rangle - \frac{\eta_1}{2} \| x \|^2_2 \right\} + \min_{\mu \in \mathcal{P}(\mathbb{K})} \left\{ D(\mu \| \nu) + \langle \mathcal{A}_\mu, z \rangle - \frac{\eta_2}{2} \| z \|^2_2 \right\},
\]

with respective optimizers denoted by \( x^*_z \) and \( \mu^*_z \). It is straightforward to see that the optimizer \( x^*_z \) is given by

\[
x^*_z = \arg \min_{x \in T} \| x - \eta_1^{-1} z \|^2_2 = \pi_T \left( \eta_1^{-1} z \right).
\]

Hence, the complexity of computing \( x^*_z \) is determined by the projection operator onto \( T \); for simple enough cases (e.g., 2-norm balls, hypercubes) the solution is analytically available, while for more general cases (e.g., simplex, 1-norm balls) it can be computed at relatively low computational effort, see [30, Section 5.4] for a comprehensive survey. The optimizer \( \mu^*_z \) according to Lemma 3.1 is given by

\[
\mu^*_z(B) = \frac{\int_B 2^{-A^*(z) \mathcal{L}}(dx)}{\int_{\mathbb{K}} 2^{-A^*(z) \mathcal{L}}(dx)}, \quad \text{for all } B \in \mathcal{B}(\mathbb{K}).
\]
Lemma 3.3 (Lipschitz gradient). The dual function $F_\eta$ defined in (9) is $\eta_2$-strongly concave and differentiable. Its gradient $\nabla F_\eta(z) = -x^*_z + A\mu^*_z - \eta_2 z$ is Lipschitz continuous with Lipschitz constant $\frac{1}{m} + \left(\sum_{i=1}^M B^i\right)^2 + \eta_2$ and $B := \max\{|x| : x \in \mathbb{K}\}$.

Proof. The proof follows along the lines of [28, Theorem 1] and in particular by recalling that the relative entropy (in the first argument) is strongly convex with convexity parameter one and Pinsker’s inequality, that says that for any $\mu \in \mathcal{P}(\mathbb{K})$ we have $D(\mu \| \nu) \geq \frac{1}{2} \|\mu - \nu\|_{TV}$. Moreover, we use the bound

$$\|A\| = \sup_{\lambda \in \mathbb{R}^M, \mu \in \mathcal{P}(\mathbb{K})} \{\langle A\mu, \lambda \rangle : \|\lambda\|_2 = 1, \|\mu\|_{TV} = 1\}$$

$$\leq \sup_{\lambda \in \mathbb{R}^M, \mu \in \mathcal{P}(\mathbb{K})} \{\|A\mu\|_2 \|\lambda\|_2 : \|\lambda\|_2 = 1, \|\mu\|_{TV} = 1\}$$

$$\leq \sup_{\mu \in \mathcal{P}(\mathbb{K})} \{\|A\mu\|_1 : \|\mu\|_{TV} = 1\}$$

$$= \sup_{\mu \in \mathcal{P}(\mathbb{K})} \left\{\sum_{i=1}^M \int_{\mathbb{K}} x^i \mu(dx) : \|\mu\|_{TV} = 1\right\}$$

$$\leq \sum_{i=1}^M B^i,$$

where (10) is due to the Cauchy-Schwarz inequality. \(\square\)

Note that $F_\eta$ is $\eta_2$-strongly concave and according to Lemma 3.3 its gradient is Lipschitz continuous with constant $L(\eta) := \frac{1}{m} + \|A\|^2 + \eta_2$. We finally consider the approximate dual program given by

$$J^*_\eta(z) = \sup_{z \in \mathbb{R}^M} F_\eta(z).$$

(11)

It turns out that (11) belongs to a favorable class of smooth and strongly convex optimization problems that can be solved by a fast gradient method given in Algorithm 1 (see [27]) with an efficiency estimate of the order $O(1/\sqrt{T})$.

| Algorithm 1: Optimal scheme for smooth & strongly convex optimization |
|-------------------------|-------------------------|
| Choose $w_0 = y_0 \in \mathbb{R}^M$ and $\eta \in \mathbb{R}^2_{>0}$ |
| For $k \geq 0$ do* |
| **Step 1:** Set $y_{k+1} = w_k + \frac{1}{L(\eta)} \nabla F_\eta(w_k)$ |
| **Step 2:** Compute $w_{k+1} = y_{k+1} + \frac{\sqrt{L(\eta)} - \sqrt{\eta_2}}{\sqrt{L(\eta)} + \sqrt{\eta_2}} (y_{k+1} - y_k)$ |

[The stopping criterion is explained in Remark 3.7]

Under an additional regularity assumption, solving the smoothed dual problem (11) provides an estimate of the primal and dual variables of the original non-smooth problems (6) and (7), respectively, as summarized in the next theorem (Theorem 3.5). The main computational difficulty of the presented method lies in the gradient evaluation $\nabla F_\eta$. We refer to Section 5, for a detailed discussion on this subject.

Assumption 3.4 (Slater point). There exits a strictly feasible solution to (3), i.e., $\mu_0 \in \mathcal{P}(\mathbb{K})$ such that $A\mu_0 \in T$ and $\delta := \min_{\eta \in T^c} \|A\mu_0 - \eta\|_2 > 0$. 


Note that finding a Slater point $\mu_0$ such that Assumption 3.4 holds, in general can be difficult. In Remark 3.8 we present a constructive way of finding such an interior point. Given Assumption 3.4, for $\varepsilon > 0$ define

$$C := D(\mu_0\|\nu), \quad D := \frac{1}{2} \max_{x \in T} \|x\|_2, \quad \eta_1(\varepsilon) := \frac{\varepsilon}{4D}, \quad \eta_2(\varepsilon) := \frac{\varepsilon \delta^2}{2C^2}$$

$$N_1(\varepsilon) := 2 \left( \sqrt{\frac{8DC^2}{\varepsilon^2 \delta^2} + \frac{2\|A\|^2 C^2}{\varepsilon \delta^2}} + 1 \right) \ln \left( \frac{10(\varepsilon + 2C)}{\varepsilon} \right)$$

$$N_2(\varepsilon) := 2 \left( \sqrt{\frac{8DC^2}{\varepsilon^2 \delta^2} + \frac{2\|A\|^2 C^2}{\varepsilon \delta^2}} + 1 \right) \ln \left( \frac{C}{\varepsilon \delta(2 - \sqrt{3})} \right) \left( 4 \frac{4D}{\varepsilon} + \|A\|^2 + \frac{\varepsilon \delta^2}{2C^2} \right) \left( C + \frac{\varepsilon}{2} \right).$$

**Theorem 3.5** (Almost linear convergence rate). Given Assumption 3.4 and the definitions (12), let $\varepsilon > 0$ and $N(\varepsilon) := \left[ \max\{N_1(\varepsilon), N_2(\varepsilon)\} \right]$. Then, $N(\varepsilon)$ iterations of Algorithm 1 produce approximate solutions to the problems (7) and (6) given by

$$\hat{z}_{k,n} := y_k \quad \text{and} \quad \hat{u}_{k,n}(B) := \frac{\int_B 2^{-A^*\hat{z}_{k,n}(x)}\nu(dx)}{\int_\mathcal{X} 2^{-A^*\hat{z}_{k,n}(x)}\nu(dx)}, \quad \text{for all } B \in \mathcal{B}(\mathcal{X}),$$

which satisfy

- **dual $\varepsilon$-optimality:** $0 \leq J^* - F(\hat{z}_{k(\varepsilon)}) \leq \varepsilon$ (14a)
- **primal $\varepsilon$-optimality:** $|D(\hat{u}_{k(\varepsilon)}\|\nu) - J^*| \leq 2(1 + 2\sqrt{3})\varepsilon$ (14b)
- **primal $\varepsilon$-feasibility:** $d(A\hat{u}_{k(\varepsilon)}, T) \leq \frac{2\varepsilon \delta}{C}$, (14c)

where $d(\cdot, T)$ denotes the distance to the set $T$, i.e., $d(x, T) := \min_{y \in T} \|x - y\|_2$.

In some applications, Assumption 3.4 does not hold, as for example in the classical case where $\mathcal{U}_i = \{0\}$ for all $i = 1, \ldots, M$, i.e., there is no uncertainty in the measurement model (2). Moreover, in other cases satisfying Assumption 3.4 using the construction described in Remark 3.8 might be computationally expensive. Interestingly, Algorithm 1 can be run irrespective of whether Assumption 3.4 holds or not, i.e., for any choice of $C$ and $\delta$. While explicit error bounds of Theorem 3.5 as well as the a-posteriori error bound discussed below do not hold anymore, the asymptotic convergence is not affected.

**Proof.** Using Assumption 3.4, note that the constant defined as

$$L := \frac{D(\mu_0\|\nu) - \min_{\mu \in \mathcal{P}(\mathcal{X})} D(\mu\|\nu)}{\min_{y \in T^c} \|A\mu_0 - y\|_2} = \frac{C}{\delta}$$

can be shown to be an upper bound for the optimal dual multiplier [26, Lemma 1], i.e., $\|z^*\|_2 \leq L$. The dual function can be bounded from above by $C$, since weak duality ensures $F(z) \leq J^* \leq D(\mu_0\|\nu) = C$ for all $z \in \mathbb{R}^M$. Moreover, if we recall the preparatory Lemmas 3.2 and 3.3, we are finally in the setting such that the presented error bounds can be derived from [8].

**Theorem 3.5** directly implies that we need at most $O\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon} \right)$ iterations of Algorithm 1 to achieve $\varepsilon$-optimality of primal and dual solutions as well as $\varepsilon$-feasible primal variable. Note that Theorem 3.5 provides an explicit bound on the so-called a-priori errors, together with approximate optimizer of the primal (6) and dual (7) problem. The latter allows us to derive an a-posteriori error depending on the approximate optimizers, which is often significantly smaller than the a-priori error.
Corollary 3.6 (Posterior error estimation). Given Assumption 3.4, let \( z^* \) denote the dual optimizer to (7). The approximate primal and dual variables \( \tilde{\mu} \) and \( \tilde{z} \) given by (13), satisfy the following a-posteriori error bound

\[
F(\tilde{z}) \leq J^* \leq D(\tilde{\mu} \| \nu) + \frac{C}{\delta} d(\hat{\mu}, T),
\]

where \( d(\cdot, T) \) denotes the distance to the set \( T \), i.e., \( d(x, T) := \inf_{y \in T} \|x - y\|_2 \).

Proof. The two key ingredients of the proof are Theorem 3.5 and the Lipschitz continuity of the so-called perturbation function of convex programming. We introduce the perturbed program as

\[
J^*(\varepsilon) = \min_{\mu \in \mathcal{P}(\mathbb{K})} \{ D(\mu \| \nu) : d(A\mu, T) \leq \varepsilon \}
\]

\[
= \min_{\mu \in \mathcal{P}(\mathbb{K})} D(\mu \| \nu) + \sup_{\lambda \geq 0} \inf_{y \in T} \lambda \|A\mu - y\| - \lambda \varepsilon
\]

\[
= \sup_{\lambda \geq 0} \inf_{y \in T} -\lambda \varepsilon + \inf_{\mu \in \mathcal{P}(\mathbb{K})} \sup_{\|z\|_2 \leq \lambda} \langle A\mu - y, z \rangle + D(\mu \| \nu)
\]

\[
\geq -\|z^*\|_2 \varepsilon + \inf_{\mu \in \mathcal{P}(\mathbb{K})} \sup_{y \in T} \langle A\mu - y, z^* \rangle + D(\mu \| \nu)
\]

\[
= -\|z^*\|_2 \varepsilon + J^*.
\]

Equation (15a) uses the strong duality property that follows by the existence of a Slater point that is due to the definition of the set \( T \), see Section 2. Step (15b) follows by Sion’s minimax theorem [35, Theorem 4.2]. Hence, we have shown that the perturbation function is Lipschitz continuous with constant \( \|z^*\|_2 \). Finally, recalling \( \|z^*\|_2 \leq \frac{C}{\delta} \), established in the proof of Theorem 3.5 completes the proof. \( \Box \)

Remark 3.7 (Stopping criterion of Algorithm 1). There are two alternatives for defining a stopping criterion for Algorithm 1. Choose desired accuracy \( \varepsilon > 0 \).

(i) a-priori stopping criterion: Theorem 3.5 provides the required number of iterations \( N(\varepsilon) \) to ensure an \( \varepsilon \)-close solution.

(ii) a-posteriori stopping criterion: Choose the smoothing parameter \( \eta \) as in (12). Fix a (small) number of iterations \( \ell \) that are run using Algorithm 1. Compute the a-posteriori error \( D(\tilde{\mu} \| \nu) + \frac{C}{\delta} d(\hat{\mu}, T) - F(\tilde{z}) \) according to Corollary 3.6 and if it is smaller than \( \text{If} \varepsilon \) terminate the algorithm. Otherwise continue with another \( \ell \) iterations.

Remark 3.8 (Slater point computation). To compute the respective constants in Assumption 3.4, we need to construct a strictly feasible point for (3). For this purpose, we consider a polynomial density of degree \( r \) defined as \( p_r(\alpha, x) := \sum_{i=0}^{r-1} \alpha_i x^i \). For notational simplicity we assume that the support set is the unit interval \( (\mathbb{K} = [0, 1]) \), such that the moments induced by the polynomial density are given by

\[
\langle p_r(\alpha, x), x^i \rangle = \int_0^1 \sum_{j=0}^{r-1} \alpha_j x^{j+i} dx = \sum_{j=0}^{r-1} \frac{\alpha_j}{j + i + 1},
\]

for \( i = 0, \ldots, M \). Consider \( \beta \in \mathbb{R}^{M+1} \), where \( \beta_1 = 1 \) and \( \beta_i = y_{i-1} \) for \( i = 2, \ldots, M + 1 \). Hence, the feasibility requirement of (3) can be expressed as the linear constraint \( A\alpha = \beta \), where
A \in \mathbb{R}^{(M+1)\times r}, \quad \alpha \in \mathbb{R}^r, \quad \beta \in \mathbb{R}^{M+1} \quad \text{and} \quad A_{i,j} = \frac{1}{i+j-1} \quad \text{and finding a strictly feasible solution reduces to the following feasibility problem}

\[
\begin{align*}
\max_{\alpha \in \mathbb{R}^r} & \quad \text{const} \\
\text{s.t.} & \quad A\alpha = \beta \\
& \quad p_r(\alpha, x) \geq 0 \quad \forall x \in [0, 1],
\end{align*}
\tag{16}
\]

where \( p_r \) is a polynomial function in \( x \) of degree \( r \) with coefficients \( \alpha \). The second constraint of the program (16) (i.e., \( p_r(\alpha, x) \geq 0 \forall x \in [0, 1] \))\(^1\) can be equivalently reformulated as linear matrix inequalities of dimension \([\frac{r}{2}]\), using a standard result in polynomial optimization, see [22, Chapter 2] for details. We note that for small degree \( r \), the set of feasible solutions to problem (16) may be empty, however, by choosing \( r \) large enough and assuming that the moments can be induced by a continuous density, problem (16) becomes feasible. Moreover, if \( 0 \in \text{int}(T) \) the Slater point leads to a \( \delta > 0 \) in Assumption 3.4.

**Example 3.9 (Density estimation).** We are given the first 3 moments of an unknown probability measure defined on \( K = [0, 1] \) as\(^2\)

\[
y := \left( \frac{1 - \ln 2}{\ln 2}, \frac{\ln 4 - 1}{\ln 4}, \frac{5 - \ln 64}{\ln 64} \right) \approx (0.44, \ 0.28, \ 0.20).
\]

The uncertainty set in the measurement model (2) is assumed to be \( \mathcal{U}_i = [-u, u] \) for all \( i = 1, \ldots, 3 \). A Slater point is constructed using the method described in Remark 3.8, where \( r = 5 \) is enough for the problem (16) to be feasible, leading to the constant \( C = 0.0288 \). The Slater point is depicted in Figure 1 and its differential entropy can be numerically computed as \(-0.0288\).

We consider two simulations for two different uncertainty sets (namely, \( u = 0.01 \) and \( u = 0.005 \)). The underlying maximum entropy problem (4) is solved using Algorithm 1. The respective features of the a-priori guarantees by Theorem 3.5 as well as the a-posteriori guarantees (upper and lower bounds) by Corollary 3.6 are reported in Table I. Recall that \( \hat{\mu}_k(\varepsilon) \) denotes the approximate primal variable after \( k \)-iterations of Algorithm 1 as defined in Theorem 3.5 and that \( d(\mathcal{A}\hat{\mu}_k(\varepsilon), T) \) (resp. \( \frac{2\varepsilon\delta}{C} \)) represent the a-posteriori (resp. a-priori) feasibility guarantees. It can be seen in Table I that increasing the uncertainty set \( \mathcal{U} \) leads to a higher entropy, where the uniform density clearly has the highest entropy. This is also intuitively expected since enlarging the uncertainty set is equivalent to relaxing the moment constraints in the respective maximum entropy problem. The corresponding densities are graphically visualized in Figure 1.

**Table I.** Some specific simulation points of Example 3.9.

| \( \mathcal{U} \) | \( \mathcal{U} = [-0.01, 0.01] \) | \( \mathcal{U} = [-0.005, 0.005] \) |
| --- | --- | --- |
| a-priori error \( \varepsilon \) | 0.01 | 0.001 | 0.01 | 0.001 |
| \( J_{\text{UB}} \) | 0.00184 | 0.00194 | 0.00194 | 0.00194 |
| \( J_{\text{LB}} \) | 0.00184 | 0.00194 | 0.00194 | 0.00194 |
| iterations \( k(\varepsilon) \) | 232 | 1241 | 12170 | 157865 |
| \( d(\mathcal{A}\hat{\mu}_k(\varepsilon), T) \) | 0.0008 | 0.00036 | 0.00065 | 0.00069 |
| \( \frac{2\varepsilon\delta}{C} \) | 0.69 | 0.069 | 0.0069 | 0.00069 |

\(^{1}\) In a multi-dimensional setting one has to consider a tightening (i.e., \( p_r(\alpha, x) > 0 \forall x \in [0, 1]^n \)).

\(^{2}\) The considered moments are actually induced by the probability density \( f(x) := (\ln 2 (1 + x))^{-1} \). We, however, do not use this information at any point of this example.
FIG. 1: Maximum entropy densities obtained by Algorithm 1 for two different uncertainty sets. As a reference, the Slater point density, that was computed as described in Remark 3.8 is depicted in red.

4. FINITE-DIMENSIONAL CASE

We consider the finite-dimensional case where \( K = \{1, \ldots, N\} \) and hence we optimize in (3) over the probability simplex \( P(K) = \Delta_N \). One substantial simplification, when restricting to the finite-dimensional setting, is that the Shannon entropy is non-negative and bounded from above (by \( \log N \)). Therefore, we can substantially weaken Assumption 3.4 to the following assumption.

Assumption 4.1.

(i) There exists \( \delta > 0 \) such that \( B(0, \delta) \subset \{ x - A\mu : \mu \in \Delta_N, x \in T \} \).

(ii) The reference measure \( \nu \in \Delta_N \) has full support, i.e., \( \min_{1 \leq i \leq N} \nu_i > 0 \).

Consider the the definitions given in (12) with \( C := \max_{1 \leq i \leq N} \log \frac{1}{\nu_i} \), then the following finite-dimensional equivalent to Theorem 3.5 holds.

Corollary 4.2 (a-priori error). Given Assumption 4.1, \( C := \max_{1 \leq i \leq N} \log \frac{1}{\nu_i} \) and the definitions (12), let \( \varepsilon > 0 \) and \( N(\varepsilon) := \lfloor \{ N_1(\varepsilon), N_2(\varepsilon) \} \rfloor \). Then, \( N(\varepsilon) \) iterations of Algorithm 1 produce the approximate solutions to the problems (7) and (6), given by

\[
\hat{z}_{k(\varepsilon)} := y_{k(\varepsilon)} \quad \text{and} \quad \hat{\mu}_{k(\varepsilon)}(B) := \frac{\sum_{i \in B} 2^{-(A^*\hat{z}_{k(\varepsilon)} )_{i} \nu_i}}{\sum_{i=1}^{N} 2^{-(A^*\hat{z}_{k(\varepsilon)} )_{i} \nu_i}} \quad \text{for all } B \subset \{1, 2, \ldots, N\}, \tag{17}
\]

which satisfy

\[
dual \varepsilon\text{-optimality:} \quad 0 \leq F(\hat{z}_{k(\varepsilon)}) - J^* \leq \varepsilon \tag{18a}
\]

\[
primal \varepsilon\text{-optimality:} \quad |D(\hat{\mu}_{k(\varepsilon)}\|\nu) - J^*| \leq 2(1 + 2\sqrt{3})\varepsilon \tag{18b}
\]

\[
primal \varepsilon\text{-feasibility:} \quad d(A\hat{\mu}_{k(\varepsilon)}, T) \leq \frac{2\varepsilon\delta}{C}, \tag{18c}
\]

where \( d(\cdot, T) \) denotes the distance to the set \( T \), i.e., \( d(x, T) := \min_{y \in T} \|x - y\|_2 \).
Proof. Under Assumption 4.1 the dual optimal solutions in (7) are bounded by
\[ \|z^*\| \leq \frac{1}{r} \max_{1 \leq i \leq N} \log \frac{1}{\nu_i}. \] (19)
This bound on the dual optimizer follows along the lines of [8, Theorem 6.1]. The presented error bounds can then be derived along the lines of Theorem 3.5. \qed 

In addition to the explicit error bound provided by Corollary 4.2, the a-posteriori upper and lower bounds presented in Corollary 3.6 directly apply to the finite-dimensional setting as well.

5. GRADIENT APPROXIMATION

The computationally demanding element for Algorithm 1 is the evaluation of the gradient \( \nabla F_{\eta}(\cdot) \) given in Lemma 3.3. In particular, Theorem 3.5 and Corollary 4.2 assume that this gradient is known exactly. While this is not restrictive if, for example, \( K \) is a finite set, in general, \( \nabla F_{\eta}(\cdot) \) involves an integration that can only be computed approximately. In particular if we consider a multi-dimensional setting (i.e., \( K \subset \mathbb{R}^d \)), the evaluation of the gradient \( \nabla F_{\eta}(\cdot) \) represents a multi-dimensional integration problem. This gives rise to the question of how the fast gradient method (and also Theorem 3.5) behaves in a case of inexact first-order information. We refer the interested readers to [9] for further details in this regard.

In this section we discuss two numerical methods to approximate this gradient. Note that in Lemma 3.3, given that \( T \) is simple enough the optimizer \( x^*_z \) is analytically available, so what remains is to compute \( A\mu^*_z \), that according to Lemma 3.1 is given by
\[ (A\mu^*_z)_i = \frac{\int_{K} x^i 2^{-A^*z(x)} \nu(dx)}{\int_{K} 2^{-A^*z(x)} \nu(dx)} \quad \text{for all } i = 1, \ldots, M. \] (20)

Semidefinite programming. Due to the specific structure of the considered problem, (20) represents an integration of exponentials of polynomials for which an efficient approximation in terms of two single semidefinite programs involving two linear matrix inequalities has been derived, where the desired accuracy is controlled by the size of the linear matrix inequalities constraints, see [5, 22] for a comprehensive study.

Quasi-Monte Carlo. The most popular methods for integration problems of the from (20) are Monte Carlo (MC) schemes, see [31] for a comprehensive summary. The main advantage of MC methods is that the root-mean-square error of the approximation converges to 0 with a rate of \( O(N^{-1/2}) \) that is independent of the dimension, where \( N \) are the number of samples used. In practise, this convergence often is too slow. Under mild assumptions on the integrand, the MC methods can be significantly improved with a more recent technique known as Quasi-Monte Carlo (QMC) methods. QMC methods can reach a convergence rate arbitrarily close to \( O(N^{-1}) \) with a constant not depending on the dimension of the problem. We would like to refer the reader to [10, 21, 29, 36, 40] for a detailed discussion about the theory of QMC methods.

Remark 5.1 (Computational stability). The evaluation of the gradient in Lemma 3.3 involves the term \( A\mu^*_z \), where \( \mu^*_z \) is the optimizer of the second term in (9). By invoking Lemma 3.1 and the definition of the operator \( A \), the gradient evaluation reduces to
\[ (A\mu^*_z)_i = \frac{\int_{K} x^i 2^{-\sum_{j=1}^{M} z_j x^j} dx}{\int_{K} 2^{-\sum_{j=1}^{M} z_j x^j} dx} \quad \text{for } i = 1, \ldots, M. \] (21)
Note that a straightforward computation of the gradient via \((\ref{eq:21})\) is numerically difficult. To alleviate this difficulty, we follow the suggestion of \cite[p. 148]{28} which we briefly elaborate here. Consider the functions \(f(z, x) := -\sum_{j=1}^{M} z_j x_j\), \(\bar{f}(z) := \max_{x \in K} f(z, x)\) and \(g(z, x) := f(z, x) - \bar{f}(z)\). Note, that \(g(z, x) \geq 0\) for all \((z, x) \in \mathbb{R}^M \times \mathbb{R}\). One can show that
\[
(A\mu^*_z)_i = \frac{\int_{K} 2g(z, x) \frac{\partial}{\partial z_i} g(z, x) \, dx}{\int_{K} 2g(z, x) \, dx} + \frac{\partial}{\partial z_i} \bar{f}(z) \quad \text{for} \quad i = 1, \ldots, M,
\]
which can be computed with significantly smaller numerical error than \((\ref{eq:21})\) as the numerical exponent are always negative, leading to values always being smaller than 1.

6. ZERO-INFORMATION MOMENT CLOSURE METHOD

In chemistry, physics, systems biology and related fields, stochastic chemical reactions are described by the chemical master equation (CME), that is a special case of the Chapman-Kolmogorov equation as applied to Markov processes \cite{42, 43}. These equations are usually infinite-dimensional and analytical solutions are generally impossible. Hence, effort has been directed toward developing of a variety of numerical schemes for efficient approximation of the CME, such as stochastic simulation techniques (SSA) \cite{13}. In practical cases, one is often interested in the first few moments of the number of molecules involved in the chemical reactions. This motivated the development of approximation methods to those low-order moments without having to solve the underlying infinite-dimensional CME. One such approximation method is the so-called moment closure method \cite{12}, that briefly described works as follows: First the CME is recast in terms of moments as a linear ODE of the form
\[
\frac{d}{dt} \mu(t) = A\mu(t) + B\zeta(t),
\]
where \(\mu(t)\) denotes the moments up to order \(M\) at time \(t\) and \(\zeta(t)\) is an infinite vector describing the contains moments of order \(M + 1\) or higher. In general \(\zeta\) can be an infinite vector, but for most of the standard chemical reactions considered in, e.g., systems biology it turns out that only a finite number of higher order moments affect the evolution of the first \(M\) moments. Indeed, if the chemical system involves only the so-called zeroth and first order reactions the vector \(\zeta\) has dimension zero (reduces to a constant affine term), whereas if the system also involves second order reactions then \(\zeta\) also contains some moments of order \(M + 1\) only. It is widely speculated that reactions up to second order are sufficient to realistically model most systems of interest in chemistry and biology \cite{14, 15}. The matrix \(A\) and the linear operator \(B\) (that may potentially be infinite-dimensional) can be found analytically from the CME. The ODE \((\ref{eq:22})\), however, is intractable due to its higher order moments dependence. The approximation step is introduced by a so-called closure function
\[
\zeta = \varphi(\mu),
\]
where the higher-order moments are approximated as a function of the lower-order moments, see \cite{33, 34}. A closure function that has recently attracted interest is known as the zero-information closure function (of order \(M\)) \cite{37}, and is given by
\[
(\varphi(\mu))_i = \langle p^*_\mu, x^{M+1} \rangle, \quad \text{for} \quad i = 1, 2, \ldots,
\]
where \(p^*_\mu\) denotes the maximizer to the problem \((\ref{eq:4})\), where \(T = \times_{i=1}^{M} \{\mu, x^i - u : u \in \mathcal{U}_i\}\), where \(\mathcal{U}_i = [-\kappa, \kappa] \subset \mathbb{R}\) for all \(i\) and for a given \(\kappa > 0\), that acts as a regularizer, in the sense
of Assumption 3.4. This approximation reduces the infinite-dimensional ODE (22) to a finite-dimensional ODE

$$\frac{d}{dt} \mu(t) = A\mu(t) + B\varphi(\mu(t)).$$

(24)

To numerically solve (24) it is crucial to have an efficient evaluation of the closure function \( \varphi \). In the zero-information closure scheme this is given by to the entropy maximization problem (4) and as such can be addressed using Algorithm 1.

To illustrate this point, we consider a reversible dimerisation reaction where two monomers (\( \mathcal{M} \)) combine in a second-order monomolecular reaction to form a dimer (\( \mathcal{D} \)); the reverse reaction is first order and involves the decomposition of the dimer into the two monomers. This gives rise to the chemical reaction system

$$2\mathcal{M} \underset{k_1}{\overset{k_2}{\rightleftharpoons}} \mathcal{D},$$

(25)

with reaction rate constants \( k_1, k_2 > 0 \). Note that the system as described has a single degree of freedom since \( \mathcal{M} = 2\mathcal{D}_0 - \mathcal{D} + \mathcal{M}_0 \), Where \( \mathcal{M} \) denotes the count of the monomers, \( \mathcal{D} \) the count of dimers, and \( \mathcal{M}_0 \) and \( \mathcal{D}_0 \) the corresponding initial conditions. Therefore, the matrices can be reduced to include only the moments of one component as a simplification and as such the zero-information closure function (23) consists of solving a one-dimensional entropy maximization problem such as given by (4), where the support are the natural numbers (upper bounded by \( \mathcal{M}_0 + \mathcal{D}_0 \) and hence compact). For illustrative purposes, let us look at a second order closure scheme, where the corresponding moment vectors are defined as \( \mu = (1, \langle \mathcal{M} \rangle, \langle \mathcal{M}^2 \rangle)^\top \in \mathbb{R}^3 \) and \( \zeta = \langle \mathcal{M}^3 \rangle \in \mathbb{R} \) and the corresponding matrices are given by

$$A = \begin{pmatrix} 0 & 0 & 0 \\ k_2 S_0 & 2k_1 - k_2 & -2k_1 \\ 2k_2 S_0 & 2k_2(S_0 - 1) - 4k_1 & 8k_1 - 2k_2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ -4k_1 \end{pmatrix},$$

where \( S_0 = \mathcal{M}_0 + 2\mathcal{D}_0 \). The simulation results, Figure 2, show the time trajectory for the average and the second moment of the number of \( \mathcal{M} \) molecules in the reversible dimerization model (25), as calculated for the zero information closure (24) using Algorithm 1, for a second-order closure as well as a third-order closure. To solve the ODE (24) we use an explicit Runge-Kutta (4,5) formula (ode45) built into MATLAB. The results are compared to the average of \( 10^6 \) SSA [43] trajectories. It can be seen how increasing the order of the closure method improves the approximation accuracy.

7. CONCLUSION AND FUTURE WORK

We presented an approximation scheme to a generalization of the classical problem of estimating a density via a maximum entropy criterion, given some moment constraints. The key idea used is to apply smoothing techniques to the non-smooth dual function of the entropy maximization problem, that enables us to solve the dual problem efficiently with fast gradient methods. Due to the favourable structure of the considered entropy maximization problem, we provide explicit error bounds on the approximation error as well as a-posteriori error estimates.

The proposed method requires one to evaluate the gradient (20) in every iteration step, which, as highlighted in Section 5, in the infinite-dimensional setting involves an integral. As such the method used to compute those integrals has to be included to the complexity of the proposed algorithm and, in higher dimensions, may become is the dominant factor. Therefore, it would
be interesting to investigate this integration step in more detail. Two approaches, one based on semidefinite programming and another invoking Quasi-Monte Carlo integration techniques, are briefly sketched. Another potential direction, would be to test the proposed numerical method in the context of approximating the channel capacity of a large class of memoryless channels [39], as mentioned in the introduction.

Finally it should be mentioned that the approximation scheme proposed in this article can be further generalized to quantum mechanical entropies. In this setup probability mass functions are replaced by density matrices (i.e., positive semidefinite matrices, whose trace is equal to one). The von Neumann entropy of such a density matrix $\rho$ is defined by $H(\rho) := -\text{tr}(\rho \log \rho)$, which reduces to the (Shannon) entropy in case the density matrix $\rho$ is diagonal. Also the relative entropy can be generalized to the quantum setup [41] and general treatment of our approximation scheme,
its analysis can be lifted to the this (strictly) more general framework. As demonstrated in [38], (quantum) entropy maximization problems can be used to efficiently approximate the classical capacity of quantum channels.

ACKNOWLEDGMENTS

The authors thank Andreas Milias-Argeitis for helpful discussions regarding the moment closure example. TS and JL acknowledge "SystemsX" under the grant "SignalX". DS acknowledges support by the Swiss National Science Foundation (SNSF) via the National Centre of Competence in Research "QSIT" and by the Air Force Office of Scientific Research (AFOSR) via grant FA9550-16-1-0245. PME was supported by the Swiss National Science Foundation under grant "P2EZP2 165264".

[1] N. I. Akhiezer. *The classical moment problem and some related questions in analysis*. Translated by N. Kemmer. Hafner Publishing Co., New York, 1965.
[2] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer, 2007. DOI: 10.1007/3-540-29587-9.
[3] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, 1987.
[4] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific optimization and computation series. Athena Scientific, 2009.
[5] D. Bertsimas, X. V. Doan, and J. Lasserre. Approximating integrals of multivariate exponentials: a moment approach. *Oper. Res. Lett.*, 36(2):205–210, 2008. DOI: 10.1016/j.orl.2007.07.002.
[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.
[7] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975. DOI: 10.1214/aop/1176996454.
[8] O. Devolder, F. Glineur, and Y. Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization*, 22(2):702–727, 2012. DOI: 10.1137/110826102.
[9] O. Devolder, F. Glineur, Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, pages 1–39, 2013. DOI: 10.1007/s10107-013-0677-5.
[10] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The Quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013. DOI: 10.1017/S0962492913000044.
[11] M. Dudík, S. J. Phillips, and R. E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.
[12] C. S. Gillespie. Moment-closure approximations for mass-action models. *IET Systems Biology*, 3(1):52–58, 2009. DOI: 10.1049/iet-syb:20070031.
[13] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Computational Phys.*, 22(4):403–434, 1976. DOI: 10.1016/0021-9991(76)90041-3.
[14] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007. DOI: 10.1146/annurev.physchem.58.032806.104637. PMID: 17037977.
[15] D. T. Gillespie, A. Hellander, and L. R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013. DOI: 10.1063/1.4801941.
[16] A. Golan. Information and entropy econometrics a review and synthesis. *Foundations and Trends in Econometrics*, 2:1–145, 2008. DOI: 10.1561/0800000004.
[17] P. Grünwald. Entropy concentration and the empirical coding game. *Statistica Neerlandica*, 62(3):374–392, 2008. DOI: 10.1111/j.1467-9574.2008.00401.x.
[18] O. Hernández-Lerma and J. Lasserre. *Further topics on discrete-time Markov control processes*. Applications of Mathematics Series. Springer, 1999. DOI: 10.1007/978-1-4612-0561-6.

[19] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, Cambridge, 2003. DOI: 10.1017/CBO9780511790423.

[20] S. Kullback. *Information theory and statistics*. John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London, 1959.

[21] F. Y. Kuo and I. H. Sloan. Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320–1328, 2005.

[22] J. B. Lasserre. *Moments, Positive Polynomials and Their Applications*. Imperial College Press optimization series. Imperial College Press, 2009. DOI: 10.1142/9781848164468_0001.

[23] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York-London-Sydney, 1969.

[24] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984. DOI: 10.1063/1.526446.

[25] P. Mohajerin Esfahani, T. Sutter, D. Kuhn, and J. Lygeros. From Infinite to Finite Programs: Explicit Error Bounds with Applications to Approximate Dynamic Programming. *ArXive e-prints*, 2017. https://arxiv.org/pdf/1701.06379.pdf.

[26] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2008.

[27] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004. DOI: 10.1007/978-1-4419-8853-9.

[28] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. DOI: 10.1007/s10107-004-0552-5.

[29] H. Niederreiter. Quasi-Monte Carlo methods. *Encyclopedia of quantitative finance*, 2010.

[30] S. Richter. Computational complexity certification of gradient methods for real-time model predictive control. *PhD thesis, ETH Zurich*, 2012. available at http://dx.doi.org/10.3929/ethz-a-007587480.

[31] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2 edition, 2004.

[32] R. T. Rockafellar. *Convergence analysis*. Princeton Landmarks in Mathematics and Physics Series. Princeton University Press, 1997.

[33] A. Singh and J. a. P. Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, 69(6):1909–1925, 2007. DOI: 10.1007/s11538-007-9198-9.

[34] A. Singh and J. P. Hespanha. Lognormal moment closures for biochemical reactions. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 2063–2068, 2006. DOI: 10.1109/CDC.2006.376994.

[35] M. Thély. Stochastic programming with Quasi-Monte Carlo methods. Master’s thesis, ETH Zurich, Switzerland, 2017.

[36] H. Umegaki. Conditional expectation in an operator algebra. *Kodai Mathematical Seminar Reports*, 14:59–85, 1962. DOI: 10.2996/kmj/1138844604.

[37] N. G. van Kampen. *Stochastic processes in physics and chemistry*, volume 888 of *Lecture Notes in Mathematics*. North-Holland Publishing Co., Amsterdam-New York, 1981.

[38] D. J. Wilkinson. *Stochastic modelling for systems biology*. Chapman & Hall/CRC Mathematical and Computational Biology Series. Chapman & Hall/CRC, Boca Raton, FL, 2006.