CARL VON OSSIETZKY universität OLDENBURG

IfM INSTITUT FÜR MATHEMATIK

Approximation of probability density functions by the Multilevel Monte Carlo Maximum Entropy method*

Claudio Bierig[1], Alexey Chernov[2]

IfM Preprint No. 2016-01
March 2016

Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
D-26111 Oldenburg, Germany

*This version is available at https://www.uni-oldenburg.de/alexey-chernov
Published in *J. Comput. Physics*, 314 (2016), 661–681, DOI: 10.1016/j.jcp.2016.03.027
[1] claudio.bierig@uni-oldenburg.de
[2] alexey.chernov@uni-oldenburg.de

Approximation of probability density functions by the Multilevel Monte Carlo Maximum Entropy method

Claudio Bierig, Alexey Chernov

Carl von Ossietzky University, 26111 Oldenburg, Germany

Abstract

We develop a complete convergence theory for the Maximum Entropy method based on moment matching for a sequence of approximate statistical moments estimated by the Multilevel Monte Carlo method. Under appropriate regularity assumptions on the target probability density function, the proposed method is superior to the Maximum Entropy method with moments estimated by the Monte Carlo method. New theoretical results are illustrated in numerical examples.

Keywords: Multilevel Monte Carlo method, Maximum Entropy method, Kullback-Leibler divergence, statistical moments, moment matching

2010 MSC: 65N30, 65C05, 65C30, 65K15

1. Introduction

The Multilevel Monte Carlo Method (MLMC) is a recently established technique for efficient computation of an observable's statistics by approximate sampling in the case when generation of samples of different accuracy is possible. The method is particularly advantageous for complex problems with low regularity, typically resulting in high memory and CPU time demands. The idea is based on the observation that coarse sample approximations can be used as control variates for more accurate sample approximations and thereby re-

Email addresses: claudio.bierig@uni-oldenburg.de (Claudio Bierig), alexey.chernov@uni-oldenburg.de (Alexey Chernov)

duce the variance of the Monte Carlo estimator. This family of methods has been introduced by M. Giles [1] for Itô stochastic differential equations arising in mathematical finance after similar ideas have been published in the earlier work by S. Heinrichs [2] on numerical quadrature. Since then MLMC has been extended to elliptic PDEs [3, 4], parabolic problems [5], conservation laws [6], variational inequalities [7, 8], multiscale PDEs [9], Kalman filtering [10] and other fields. The recent work [11] contains a recipe for an efficient evaluation of central statistical moments of arbitrary order. The aim of the present article is the further extension of the MLMC methodology for estimation of probability density functions.

Setting up probability density functions (PDF) on the basis of incomplete information on the observable is a prominent problem in statistics and information theory. One way to solve it is to recover the PDF from a truncated sequence of statistical moments (see the recent work by Giles et al. [12] for an alternative approach). This task (also known as solving the *truncated moment problem*) is by no means trivial and has been extensively studied in measure and probability theory [13, 14, 15, 16]. It is well known that depending on the prescribed moments, the truncated moment problem may have no solution or multiple (infinitely many) solutions. The latter is typically the case when the truncated sequence of moments is admissible, i.e. it corresponds to some PDF (ruling out the case of negative even-order monomial moments and similar incompatibilities). However, in the presence of significant statistical and approximation errors, the sequence of *estimated* moments may become inadmissible even when the sequence of *exact* moments is admissible.

Assuming that the truncated moment sequence is admissible, one needs a criterion to select a PDF which is the "most appropriate" among infinitely many solutions to the truncated moment problem. The strategy of selecting the least biased estimate brings us to the concept of the Maximum Entropy (ME) method [17]. The ME solution is the (nonnegative) maximiser of the Shannon entropy constraint at the prescribed moment values. Obviously, the error of this approximation depends on the number of statistical moments and the accuracy of

the *estimated* moments. Under appropriate assumptions the original constraint Maximum Entropy formulation is equivalent to the matching of moments with a density function whose logarithm is approximated by a polynomial. References [18, 19] contain a rigorous error analysis of this class of ME methods, [18] also combines it with the Monte Carlo approach. The purpose of this work is to combine the Maximum Entropy approach with the Multilevel Monte Carlo estimation of moments and develop a rigorous error analysis in terms of i) the number of statistical moments, ii) statistical error and iii) discretization error. We derive complexity estimates for the proposed approach, test its performance on a set of synthetic problems with known PDFs, and demonstrate its applicability in a more realistic context: on a problem modelling contact of an elastic membrane with a rough random obstacle.

The outline of the paper is as follows. After a brief introduction to the Multilevel Monte Carlo and the Maximum Entropy methods in Section 2 we give a complete *a priori* error analysis for the proposed method in Section 3. In particular, we consider three different approximation methods for the set of the statistical moments: the Monte Carlo method based on exact sampling, the Monte Carlo method based on approximate sampling, and the Multilevel Monte Carlo approach. The error estimates naturally depend on the number of statistical moments, the sample size and the level of accuracy for approximate samples. In Section 4 we identify the optimal relation between these parameters and derive error-versus-cost relations for the three aforementioned methods. In Section 5 we give a series of numerical experiments illustrating convergence of the suggested Maximum Entropy approximations and compare them with a variant of Kernel Density Estimators available from the literature.

In the following, $\ln(\cdot)$ stands for the natural logarithm. We use a convention that for two scalar quantities f and g the notation $f \lesssim g$ means that there exists a nonnegative constant C independent of the approximation parameters such that $f \leq Cg$. The notation $f \sim g$ is equivalent to $f \lesssim g$ and $g \lesssim f$.

2. Preliminaries

In this section we recall some preliminary information needed for the subsequent analysis, see e.g. [20] and the references therein for the general framework of the multilevel Monte Carlo method (we utilise the notations from [8, 11]), and [18, 19, 21] for the description of the Maximum Entropy method.

2.1. Multilevel Monte Carlo method

Suppose $(\Omega, \Sigma, \mathbb{P})$ is a probability space and X is a real-valued random variable which is not available for direct sampling. Instead, there exists an approximation X_ℓ to X, so that samples X_ℓ^i of X_ℓ can be generated. In this case the mean $\mathbb{E}[X]$ can be approximated by the sample average $E_M[X_\ell] := \frac{1}{M}\sum_{i=1}^M X_\ell^i$ of iid samples X_ℓ^i admitting the decomposition of the mean square error (MSE)

$$\|E_M[X_\ell] - \mathbb{E}[X]\|_{L^2}^2 = |\mathbb{E}[X - X_\ell]|^2 + \frac{1}{M}\operatorname{Var}[X_\ell] \tag{1}$$

where $\operatorname{Var}[X_\ell]$ is the variance of X_ℓ. The idea of the two-level Monte Carlo approach is to use samples from a coarser approximation $X_{\ell-1}$ to reduce the variance of the estimator. Indeed, for the two-level estimator it holds that

$$\begin{aligned}\|E_{M_\ell}[X_\ell - X_{\ell-1}] + E_{M_{\ell-1}}[X_{\ell-1}] - \mathbb{E}[X]\|_{L^2}^2 &= |\mathbb{E}[X - X_\ell]|^2 \\ &\quad + \frac{1}{M_\ell}\operatorname{Var}[X_\ell - X_{\ell-1}] + \frac{1}{M_{\ell-1}}\operatorname{Var}[X_{\ell-1}]\end{aligned}$$

where $E_{M_\ell}[X_\ell - X_{\ell-1}]$ and $E_{M_{\ell-1}}[X_{\ell-1}]$ are based on independent samples.

This situation occurs for example when X depends on a solution of an ODE or a PDE which is not available in closed form, but can be computed approximately, e.g. by the Finite Element Method or another numerical approximation method. In this setting the parameter ℓ plays the role of a discretization parameter. It is plausible that the samples of the fine approximation X_ℓ are better approximations to samples of X, but are typically more expensive to compute than samples of the coarse approximation $X_{\ell-1}$. The Multilevel Monte Carlo Method extends the two-level approach to multiple levels. In particular, the multilevel sample mean estimator is defined as

$$E^{\mathrm{ML}}[X] := \sum_{\ell=1}^{L} E_{M_\ell}[X_\ell - X_{\ell-1}], \quad X_0 := 0.$$

When $E_{M_\ell}[X_\ell - X_{\ell-1}]$ are independent for $\ell = 1, \ldots, L$, there holds the error representation

$$\|E^{\mathrm{ML}}[X] - \mathbb{E}[X]\|_{L^2}^2 = |\mathbb{E}[X - X_L]|^2 + \sum_{\ell=1}^{L} \frac{1}{M_\ell} \operatorname{Var}[X_\ell - X_{\ell-1}]. \tag{2}$$

If $X_\ell \to X$ in a suitable sense, the variance of the correction $\operatorname{Var}[X_\ell - X_{\ell-1}]$ becomes smaller for larger values of ℓ. Then the appropriate balancing of the summands relies on the rapidly decaying sequence of the number of samples $\{M_1, \ldots, M_L\}$. This allows us to generate fewer computationally intensive samples at the fine grids (i.e. when $1 \ll \ell \leq L$) at the cost of computing more samples of the coarse grid approximations ($1 \leq \ell \ll L$). This may lead to significant savings in the computational cost, while preserving the accuracy.

2.2. Maximum Entropy Method

Let X be a real-valued random variable in the probability space $(\Omega, \Sigma, \mathbb{P})$ with the associated probability density function $\rho \geq 0$. The Shannon entropy is defined as a functional

$$\mathbb{E}[-\ln(\rho(X))] = -\int_I \rho(x) \ln \rho(x)\, dx.$$

The integral can be taken over the support of the density ρ. The subsequent theory will be based on the Assumption 1 on $\operatorname{supp}(\rho)$ implying that I can be seen as a bounded interval on the real line so that $\operatorname{supp}(\rho) \subseteq I$. Suppose ρ is unknown and has to be recovered on the basis of finitely many generalized moments

$$\mu_k = \int_I \phi_k(x)\rho(x)\, dx, \qquad 0 \leq k \leq R \tag{3}$$

where $\{\phi_k\}_{k=0}^R$ are linearly independent functions. In this paper we concentrate on the case when $\{\phi_0, \ldots, \phi_R\}$ are algebraic polynomials forming a basis in the space of algebraic polynomials $\mathcal{P}_R$ of degree less than or equal to R. In this case ϕ_0 is a constant, and since ρ is a probably density function we require that $\phi_0 \equiv \mu_0$. According to Assumption 1 the density function ρ has bounded support, which guarantees in particular, that moments μ of arbitrary order k exist. If

the truncated sequence of moments is admissible (i.e. when it corresponds to some PDF), then the system of equations (3) may have multiple solutions.

A popular method to select the "best" candidate PDF satisfying (3) is to choose the one with the greatest Shannon entropy. Under the assumption that there exists a strictly positive PDF fulfilling (3), the unique maximizer is given by

$$\begin{cases} \rho_R(x) := \exp\left(\sum_{k=0}^{R} \lambda_k \phi_k(x)\right), & \lambda_k \in \mathbb{R}, \\ \mu_k = \int_I \phi_k(x)\rho_R(x)\,dx, & 0 \le k \le R. \end{cases} \tag{4}$$

For a rigorous derivation of this result see e.g. [18, Lemma 3] and [21]. Assuming that the statistical moments (3) are not known exactly and a finite sequence of perturbed moments $\tilde{\mu}_k \approx \mu_k$ (the perturbation may be caused by estimation of μ_k's and may contain sampling and discretization errors, etc.) is available instead, we define the perturbed Maximum Entropy solution $\tilde{\rho}_R$ as the solution to the following problem:

$$\begin{cases} \tilde{\rho}_R(x) := \exp\left(\sum_{k=0}^{R} \tilde{\lambda}_k \phi_k(x)\right), & \tilde{\lambda}_k \in \mathbb{R}, \\ \tilde{\mu}_k = \int_I \phi_k(x)\tilde{\rho}_R(x)\,dx, & 0 \le k \le R \end{cases} \tag{5}$$

where $\tilde{\lambda}_k \approx \lambda_k$ is the sequence of the corresponding perturbed coefficients. If ρ is sufficiently regular (we refer to the subsequent error analysis for the precise statements), it is plausible that the truncation error $\rho - \rho_R$ vanishes when $R \to \infty$ and the estimation error $\rho_R - \tilde{\rho}_R$ becomes small when $\tilde{\mu}_k \to \mu_k$ implying that the total error $\rho - \tilde{\rho}_R$ converges to zero in a suitable sense. It turns out that the Kullback-Leibler distance (or the relative entropy, KL-divergence)

$$D_{\mathrm{KL}}(\rho\|\eta) = \int_I \rho(x) \ln \frac{\rho(x)}{\eta(x)}\,dx$$

is a natural measure of distance between two probability densities [22, 23]. In particular, the following Pythagorean-like identity which separates contributions of the truncation and estimation error has been shown in [18, Lemma 3]

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R) = D_{\mathrm{KL}}(\rho\|\rho_R) + D_{\mathrm{KL}}(\rho_R\|\tilde{\rho}_R) \tag{6}$$

so that ρ_R is sometimes called the *information projection*.

It is well known that $D_{\mathrm{KL}}(\rho\|\eta)$ is non-negative and $D_{\mathrm{KL}}(\rho\|\eta) = 0$ if and only if $\rho = \eta$. Moreover, D_{KL} is an upper bound for all L^p norms of the difference $\rho - \eta$. The following result can be found in [19, Lemma 2.2, Proposition 2.3], see also [22, 23] for the proof of (7).

Lemma 1. *For two probability density functions $\rho, \eta \in L^1(I)$ it holds that*

$$\frac{1}{2}\|\rho - \eta\|_{L^1}^2 \leq D_{\mathrm{KL}}(\rho\|\eta). \tag{7}$$

If moreover $\rho, \eta \in L^\infty$, then it holds for $2 \leq p < \infty$ that

$$\|\rho - \eta\|_{L^p}^p \leq p(p-1)\left(\max(\|\rho\|_{L^\infty}, \|\eta\|_{L^\infty}\right)^{p-1} D_{\mathrm{KL}}(\rho\|\eta). \tag{8}$$

Observe that the KL-divergence is non-symmetric and $D_{\mathrm{KL}}(\rho\|\eta) < \infty$ implies that $\mathrm{supp}(\rho) \subseteq \mathrm{supp}(\eta)$. Another two-sided bound in terms of the weighted $L^2(\rho)$ norms of the log-densities is required for the subsequent analysis. We refer to [18, Lemma 1] for the proof of the following statement:

Lemma 2. *For two probability density functions ρ and η with $\ln(\rho/\eta) \in L^\infty(\mathcal{I})$ where $\mathcal{I} = \mathrm{supp}(\rho)$ it holds that*

$$D_{\mathrm{KL}}(\rho\|\eta) \geq \frac{1}{2} e^{-\|\ln(\rho/\eta)\|_{L^\infty(\mathcal{I})}} \int_{\mathcal{I}} \rho(x) \left(\ln \frac{\rho(x)}{\eta(x)}\right)^2 dx$$

and

$$D_{\mathrm{KL}}(\rho\|\eta) \leq \frac{1}{2} e^{\|\ln(\rho/\eta) - c\|_{L^\infty(\mathcal{I})}} \int_{\mathcal{I}} \rho(x) \left(\ln \frac{\rho(x)}{\eta(x)} - c\right)^2 dx$$

for any $c \in \mathbb{R}$.

Thanks to the error splitting (6) it is sufficient to estimate the truncation error and the estimation error with respect to the KL-divergence. The size of the truncation error $D_{\mathrm{KL}}(\rho\|\rho_R)$ is determined by the smoothness of the probability density ρ and is controlled by the number of moments R. In the forthcoming Theorem 1 we address two typical cases: when a) $\ln(\rho)$ has a finite Sobolev regularity and b) $\ln(\rho)$ is analytic on $\mathcal{I}$. The estimation error $D_{\mathrm{KL}}(\rho_R\|\tilde{\rho}_R)$ is determined by the perturbation of the statistical moments, see Theorem 2, and

is controlled by a number of "estimation parameters". Depending on the method of estimation (we consider the Monte Carlo method with exact samples, approximate single- and Multilevel Monte Carlo methods) the generic term "estimation parameters" may include the number of samples M_ℓ and/or deterministic discretization parameters N_ℓ. For various cases mentioned above we identify the optimal relation between parameter values and obtain upper bounds for the required computational cost. The estimates for the error-versus-cost relation are given in the statements of Theorems 3, 4 and 5 below.

3. A priori error analysis

The forthcoming error analysis relies on further assumptions on the smoothness of the log-density $\ln(\rho)$ which has to be recovered from computations. In particular, we shall analyse the best approximation of ρ by probability density functions η, where $\ln(\eta)$ is an algebraic polynomial, with the aid of Lemma 2. Since $\ln(\eta)$ is a polynomial, the requirement $\ln(\rho/\eta) \in L^\infty(\mathcal{I})$ is satisfied if the following assumption holds true.

Assumption 1. *The probability density ρ has a bounded and connected support $\mathcal{I} := \operatorname{supp}(\rho) \subset I$. Moreover, there exist constants $c_1, c_2 > 0$ such that*

$$c_1 < \rho(x) < c_2 \qquad \forall x \in \mathcal{I}.$$

By the translation and scaling argument, we assume that $I = [-1, 1]$ without loss of generality.

3.1. Truncation error

The following best approximation estimate for the truncation error is a consequence of Lemma 2 and basic properties of the KL-divergence.

Corollary 1. *Suppose ρ is such that Assumption 1 is satisfied and ρ_R is defined in* (4)*, so that ρ and ρ_R have identical first R generalized moments* (3)*. Then*

$$D_{\mathrm{KL}}(\rho\|\rho_R) \le \frac{1}{2} \inf_{\psi\in\mathcal{P}_R} \left(e^{\|\ln(\rho)-\psi\|_{L^\infty(\mathcal{I})}} \|\ln(\rho) - \psi\|^2_{L^2(\mathcal{I},\rho)} \right)$$

Proof. Let $\psi \in \mathcal{P}_R$ and define $\eta_R := e^\psi / \int e^\psi$ so that η_R is a probability density. By (6), non-negativity of the KL-divergence and Lemma 2 with $c = \ln(\int e^\psi)$

$$D_{\mathrm{KL}}(\rho\|\rho_R) \le D_{\mathrm{KL}}(\rho\|\eta_R) \le \frac{1}{2} e^{\|\ln(\rho)-\psi\|_{L^\infty(\mathcal{I})}} \int_{\mathcal{I}} \rho(x)\big(\ln(\rho(x)) - \psi(x)\big)^2 \, dx.$$

Taking the infimum with respect to all $\psi \in \mathcal{P}_R$ yields the assertion. □

Now we are ready to prove *a priori* convergence estimates for the truncation error dependent on the smoothness of the probability density ρ.

Theorem 1. *Suppose ρ is such that Assumption 1 is satisfied and ρ_R is defined in (4), in particular, ρ and ρ_R have identical first R generalized moments (3). Assume in addition that $\rho \in H^s(\mathcal{I})$ with $s \ge 1$. Then it holds that*

$$D_{\mathrm{KL}}(\rho||\rho_R) \le C(s) R^{-2s} \|\rho\|_{L^\infty(\mathcal{I})} \|\ln(\rho)\|^2_{H^s(\mathcal{I})} \tag{9}$$

with a constant $C(s) > 0$ which is independent of R and ρ. If moreover, ρ is analytic on $\mathcal{I}$ and $\ln(\rho)$ admits a unique analytic continuation $\ln(\rho(z))$ in $z \in \mathcal{E}_a \subset \mathbb{C}$, where $\mathcal{E}_a$ is the ellipse with focuses ± 1 and semiaxes' sum $a > 1$, it holds that

$$D_{\mathrm{KL}}(\rho||\rho_R) \le C R^{-1} a^{-2R} \|\rho\|_{L^\infty}, \tag{10}$$

where C depends only on a and $\sup_{z\in\mathcal{E}_a} |\ln(\rho(z))|$.

Proof. By the trace inequality we have that

$$\|\ln(\rho) - \psi\|_{L^\infty(\mathcal{I})} \le \sqrt{2} \|\ln(\rho) - \psi\|_{H^1(\mathcal{I})}$$

and moreover

$$\|\ln(\rho) - \psi\|^2_{L^2(\mathcal{I},\rho)} \le \|\rho\|_{L^\infty(\mathcal{I})} \|\ln(\rho) - \psi\|^2_{L^2(\mathcal{I})}$$

for any polynomial $\psi \in \mathcal{P}_R$. Next, we fix $\psi = \psi_R$ as the unique projection of $\ln(\rho)$ determined as the solution of the variational problem

$$\begin{aligned} \int_{\mathcal{I}} (\ln(\rho) - \psi)' v' = 0, \qquad & \forall v \in \mathcal{P}_R, \\ (\ln(\rho) - \psi)|_{\partial\mathcal{I}} = 0, & \end{aligned}$$

see [24, Theorem 3.14]. Then $\|\ln(\rho) - \psi_R\|_{H^1(\mathcal{I})} \to 0$ for $R \to \infty$ and

$$\|\ln(\rho) - \psi_R\|^2_{L^2(\mathcal{I})} \leq CR^{-2s} \|\ln(\rho)\|^2_{H^s(\mathcal{I})}$$

If $\ln(\rho)$ is analytic, the result follows for the same projection ψ_R from [24, (3.3.32)] and [25]. □

3.2. Estimation error

In this section we address existence of the Maximum Entropy solution $\tilde{\rho}_R$ for the set of perturbed moments $\tilde{\mu}_k \approx \mu_k$ and derive *a priori* upper bounds for the quantity $D_{\mathrm{KL}}(\rho_R \| \tilde{\rho}_R)$ – the estimation error part in the splitting (6).

The forthcoming Lemma 3 is an important stability result which establishes and quantifies continuous dependence of the perturbation in PDFs on the perturbation of the first R moments. Evidently, this stability bound depends on the type of the generalized moments, that is on the particular basis $\{\phi_0, \ldots, \phi_R\}$. In the remaining part of the paper we work with generalized moments (3) for the orthonormal basis on $[-1, 1]$, i.e. we choose $\phi_k(x) := P_k(x)$ where P_k are orthonormalized Legendre polynomials. Then (3) takes the form

$$\mu_k = \int_{-1}^{1} P_k(x)\rho(x)\, dx. \tag{11}$$

Recall from [26] that for any $\psi \in \mathcal{P}_R$

$$\|\psi\|_{L^\infty(-1,1)} \leq A_R \|\psi\|_{L^2(-1,1)} \quad \text{where} \quad A_R := \frac{R+1}{\sqrt{2}}. \tag{12}$$

Lemma 3. *Suppose $R \in \mathbb{N}$, $\{\mu_1, \ldots, \mu_R\}$ is the sequence of exact moments (3) of the target probability density ρ, and ρ_R is the corresponding Maximum Entropy probability density determined by (4). Let $\{\tilde{\mu}_1, \ldots, \tilde{\mu}_R\}$ be small perturbations of $\{\mu_1, \ldots, \mu_R\}$ in the sense that*

$$\left(\sum_{k=1}^{R} (\mu_k - \tilde{\mu}_k)^2 \right)^{1/2} \leq \frac{1}{2C_R A_R} \qquad \textit{with} \quad C_R := 2e^{1+\|\ln \rho_R\|_{L^\infty(-1,1)}} \tag{13}$$

and A_R defined in (12). Then the Maximum Entropy solution $\tilde{\rho}_R$ of (5) constraint at perturbed moments exists, is unique and is close to ρ_R in the sense that

$$D(\rho_R \| \tilde{\rho}_R) \leq C_R \sum_{k=1}^{R} (\mu_k - \tilde{\mu}_k)^2. \tag{14}$$

Proof. [18, Lemma 5]. □

Notice that C_R is uniformly bounded when $\log(\rho) \in H^s(I)$ for any $s > 1$. This can be shown involving the arguments from [18, Theorem 3].

The next observation is that when the sequence of perturbed moments $\tilde{\mu}_k$ is obtained by sampling, $\tilde{\mu}_k$'s are not deterministic but are, in fact, random variables. As a consequence, it may happen that (13) is not fulfilled with certain probability, so that the existence of $\tilde{\rho}_R$ is not guaranteed in this case. The following theorem is a probabilistic variant of Lemma 3, which gives an upper bound on the probability that $\tilde{\rho}_R$ may not exist. The proof follows [18] and is given here for the sake of completeness.

Theorem 2. *Suppose that the assumptions of Lemma 3 are satisfied and additionally, let $\tilde{\mu}_k$ be probabilistic estimators for μ_k satisfying* (13) *in mean, i.e.*

$$\mathbb{E}\Big[\sum_{k=1}^{R}(\mu_k - \tilde{\mu}_k)^2\Big] \leq \Phi(R,\rho) \leq \frac{1}{(2A_RC_R)^2}$$

for some function Φ. Then $\tilde{\rho}_R$ exists with probability at least $1 - p_*$ where

$$p_* = (2A_RC_R)^2\mathbb{E}\Big[\sum_{k=1}^{R}(\mu_k - \tilde{\mu}_k)^2\Big]. \tag{15}$$

Furthermore, for any $p \in [p_*, 1]$ the estimate

$$D_{\mathrm{KL}}(\rho_R \| \tilde{\rho}_R) \leq C_R p^{-1} \Phi(R,\rho) \tag{16}$$

holds true with probability at least $1 - p$.

Proof. The proof is a combination of Lemma 3 and the Markov inequality. For a random variable Z taking nonnegative values it holds that

$$\mathbb{P}(Z > c) = \int_{Z>c} 1 \mathrm{d}\mathbb{P} \leq \int_{Z>c} c^{-1} Z \,\mathrm{d}\mathbb{P} \leq c^{-1}\mathbb{E}[Z].$$

This estimate and (15) imply that

$$\mathbb{P}\Big(\sum_{k=1}^{R}(\mu_k - \tilde{\mu}_k)^2 > (2A_RC_R)^{-2}\Big) \leq (2A_RC_R)^2\mathbb{E}\Big[\sum_{k=1}^{R}(\mu_k - \tilde{\mu}_k)^2\Big] = p_*.$$

Thus, (13) is satisfied with probability at least $1 - p_*$ and $\tilde{\rho}_R$ exists with the same probability by Lemma 3. In the same way we obtain for $p \in [p_*, 1]$

$$\mathbb{P}\Big(\sum_{k=1}^{R} (\mu_k - \tilde{\mu}_k)^2 > p^{-1}\Phi(R,\rho) \Big) \leq p\Phi(R,\rho)^{-1}\mathbb{E}\Big[\sum_{k=1}^{R} (\mu_k - \tilde{\mu}_k)^2 \Big] \leq p.$$

Therefore, with probability at least $1 - p$ it holds that

$$\sum_{k=1}^{R} (\mu_k - \tilde{\mu}_k)^2 \leq p^{-1}\Phi(R,\rho)$$

and together with (14) this implies that $D_{\mathrm{KL}}(\rho_R \| \tilde{\rho}_R) \leq C_R p^{-1}\Phi(R,\rho)$ with the same probability, $1 - p$. □

4. Accuracy and cost of single- and multilevel Monte Carlo estimators

In this section we analyse convergence of the Maximum Entropy method for three types of estimators for the statistical moments:

- A Monte Carlo estimator for the case when exact samples of X can be generated (referred to as exact sampling);
- A Monte Carlo estimator for the case when only approximate samples $X_\ell \approx X$ can be generated (single level approximate sampling);
- A Multilevel Monte Carlo estimator for the case when approximate samples $X_\ell \approx X$ can be generated for $\ell = 1, \ldots, L$ (multilevel approximate sampling).

Throughout this section we require that Assumption 1 is satisfied and, in particular, that $\mathcal{I} := \mathrm{supp}(\rho) \subseteq [-1, 1] =: I$. Recall that P_k is the orthonormalized Legendre polynomial of degree k. Then it easy to see that the following auxiliary estimate holds.

Lemma 4. *Suppose Z is a random variable with values in $[-1, 1]$ and ρ_Z is its probability density function. Then it holds that*

$$\mathrm{Var}[P_k(Z)] \leq \|\rho_Z\|_{L^\infty(-1,1)}.$$

Proof. We have for any $k \in \mathbb{N}$

$$\mathrm{Var}[P_k(Z)] = \mathbb{E}[P_k^2(Z)] - \mathbb{E}[P_k(Z)]^2 \leq \int_{-1}^{1} P_k^2(z)\rho_Z(z)\,dz \leq \|\rho_Z\|_{L^\infty(-1,1)}$$

and hence the assertion. □

In this work we focus on the case when the cost of generation of all necessary samples significantly dominates the computational cost of solving the system of nonlinear equations (5). This situation typically occurs when generation of one sample requires an approximate solution of a differential model e.g. by the Finite Element Method or some other numerical method. Therefore the term "computational cost" refers to the computational cost required for generation of all necessary samples neglecting the cost of solving the system of nonlinear equations (5).

The minimal smoothness assumptions on the log-density throughout this section will be that $\log(\rho) \in H^s(I)$, $s > 1$. In this case the constants in (9), (10) and (16) are uniformly bounded.

Next, we address the accuracy and the cost of the aforementioned Maximum Entropy Monte Carlo estimators.

4.1. Monte Carlo estimators based on exact samples

In this section we assume that X is a random variable and exact samples of X can be generated so that the following assumption is satisfied.

Assumption 2. *Let X be a random variable in a complete probability space $(\Omega, \Sigma, \mathbb{P})$ having the probability density function ρ with $\mathcal{I} := \mathrm{supp}(\rho) \subseteq [-1,1] =: I$. We assume that iid samples X^i of X can be generated and the cost to generate one sample X^i is independent of i and is equal to $\mathcal{C}(X)$.*

Then exact moments μ_k from (11) can be approximated by sample means

$$\tilde{\mu}_k^{\mathrm{MC}} := E_M[P_k(X)]. \tag{17}$$

Notice that all approximate moments $\{\tilde{\mu}_1^{\mathrm{MC}}, \dots \tilde{\mu}_R^{\mathrm{MC}}\}$ can be evaluated for the same set of samples. Evaluation of the Legendre polynomials P_k is fast, therefore

we assume that the computational cost is independent of the number of involved statistical moments R. The following bound holds for this approximation.

Lemma 5. *Suppose that Assumptions 1 and 2 are satisfied and μ_k and $\tilde{\mu}_k^{\mathrm{MC}}$ are defined by (11) and (17) respectively. Then it holds that*

$$\mathbb{E}\Big[\sum_{k=1}^{R}\big(\mu_k-\tilde{\mu}_k^{\mathrm{MC}}\big)^2\Big]\leq\frac{R}{M}\|\rho\|_{L^\infty}.$$

Proof. Recall that $\mathbb{E}[E_M[P_k(X)]]=\mu_k$. Then the assertion follows from

$$\mathbb{E}\Big[\sum_{k=1}^{R}\big(\mu_k-\tilde{\mu}_k^{\mathrm{MC}}\big)^2\Big]=\sum_{k=1}^{R}\mathrm{Var}[E_M[P_k(X)]]\leq\frac{R}{M}\|\rho\|_{L^\infty}$$

where Lemma 4 has been used in the last step. □

Theorem 3. *Suppose that Assumptions 1 and 2 are satisfied and $\ln(\rho)\in H^s(I)$ for some $s>1$. Let $R\in\mathbb{N}$ and $\tilde{\rho}_R^{\mathrm{MC}}$ be the Maximum Entropy solution of problem (5) for a sequence of approximate moments $\{\tilde{\mu}_1^{\mathrm{MC}},\dots,\tilde{\mu}_R^{\mathrm{MC}}\}$ defined in (17). Then for any $\varepsilon>0$ and $p\in(0,1)$ it is possible to select the number of samples M and the number of moments R so that*

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{MC}})<\varepsilon \tag{18}$$

is satisfied with probability at least $1-p$ and the cost of evaluation of all samples required for the recovery of $\tilde{\rho}_R^{\mathrm{MC}}$ satisfies the following asymptotic relations: when $\ln(\rho)\in H^s(I)$ for some $s>1$, it holds that

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{MC}})\sim p^{-1}\varepsilon^{-1-\frac{1}{2s}}, \tag{19}$$

when, moreover, $\ln(\rho)$ is analytic, it holds that

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{MC}})\sim p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|. \tag{20}$$

Proof. When $\ln(\rho)\in H^s(I)$ for some $s>1$ we have by Theorems 1, 2 and Lemma 5 that

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{MC}})\lesssim R^{-2s}+\frac{R}{pM}$$

with probability at least $1-p$. The cost of evaluation of all required samples is proportional to M, thus $D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{MC}})$ is minimized for a fixed cost and satisfies (18) when

$$R^{-2s} \sim \frac{R}{pM} \sim \varepsilon,$$

or analogously $M \sim p^{-1}R^{2s+1}$ and $R \sim \varepsilon^{-\frac{1}{2s}}$. In this case the cost of evaluation of all samples scales as

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{MC}}) = M \cdot \mathcal{C}(X) \sim p^{-1}\varepsilon^{-\frac{2s+1}{2s}}$$

and (19) follows. When $\ln(\rho)$ is analytic, Theorems 1, 2 and Lemma 5 imply

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{MC}}) \lesssim R^{-1}a^{-2R} + \frac{R}{pM}. \tag{21}$$

for $a > 1$. Both summands in the right-hand side are comparable (and thus $D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R)$ is minimized for a fixed computational cost) when $M \sim p^{-1}Ra^{2R}$ and (18) holds for $R = -\frac{1}{2}\log_a(\varepsilon)$. In this case

$$\mathcal{C}(\tilde{\rho}_R) = M \cdot \mathcal{C}(X) \sim p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|$$

and the assertion follows. □

Remark 1. *Notice that both summands in* (21) *are asymptotically of the same order when* $M \sim p^{-1}R^2a^{2R}$ *(rather than* $M \sim p^{-1}Ra^{2R}$*). However this leads to the same asymptotic estimate for the computational cost. Indeed, in this case* (18) *holds when*

$$R^{-1}a^{-2R} = \varepsilon \qquad \Leftrightarrow \qquad R = c^{-1}W(\varepsilon^{-1}c) \quad \textit{for} \quad c = 2\ln(a)$$

where W *is the Lambert function. Thus*

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{MC}}) = M \cdot \mathcal{C}(X) \sim p^{-1}\varepsilon^{-1}R \sim p^{-1}\varepsilon^{-1}W(\varepsilon^{-1}) \sim p^{-1}\varepsilon^{-1}|\ln(\varepsilon)| \tag{22}$$

since for sufficiently small ε

$$W(\varepsilon^{-1}) \sim |\ln(\varepsilon)| - \ln(|\ln(\varepsilon)|) \sim |\ln(\varepsilon)|,$$

see [27], and the new estimate (22) *is equivalent to* (20)*. An analogous situation appears in the proofs of Theorem 4 and 5. In order to keep the presentation simple we do not further elaborate on this.*

4.2. Approximate single level Monte Carlo estimators

Now, let us assume that it is not possible to sample from X, but X_ℓ is available for sampling instead and X_ℓ is close to X in the sense of Assumption 3.

Assumption 3. *Let X be a random variable in a complete probability space $(\Omega, \Sigma, \mathbb{P})$ having the probability density function ρ with $\mathcal{I} := \mathrm{supp}(\rho) \subseteq [-1,1] =: I$. We assume that it is possible to generate iid samples of random variables $X_\ell \approx X$, where $\mathcal{C}_\ell$ is the cost required for generation of one sample of X_ℓ. Let $\{N_\ell\}_{\ell=1}^\infty$ be an exponentially increasing sequence satisfying $\bar{c} \geq N_\ell/N_{\ell-1} \geq c$ for some fixed $\bar{c} \geq c > 1$. Moreover, assume there exist constants $\beta, \gamma > 0$ and $\delta \geq 0$ such that the following asymptotic bounds hold*

$$1)\ \ \mathbb{E}\big[\big(P_k(X_\ell) - P_k(X)\big)^2\big] \lesssim k^\delta N_\ell^{-\beta}, \qquad 2)\ \ \mathcal{C}_\ell \lesssim N_\ell^\gamma.$$

As we shall show next, relation 1) in Assumption 3 can be replaced by a simpler relation (23). Then relation 1) follows with $\delta = 5$ in the general case, or with $\delta = 2$ if additional assumptions are satisfied. However, in the numerical experiments we observed that this estimate for δ may be too pessimistic (see Figure 8 in Section 5) and therefore prefer to put relation 1) as an additional assumption.

Proposition 1. *If Assumption 3 holds with relation 1) replaced by*

$$\mathbb{E}[(X_\ell - X)^2] \lesssim N_\ell^{-\beta}, \tag{23}$$

then relation 1) is satisfied too. Moreover, relation 1) is satisfied with $\delta = 2$ when $\mathcal{I} \subset (-1,1)$ and $\overline{\mathrm{Im}(X_\ell)} \subset (-1,1)$ for all ℓ, and with $\delta = 5$ otherwise.

Proof. By the mean value theorem there exists ξ in the interval (a,b) where $a = \inf_{x\in\mathbb{R}} \big\{x \in \mathcal{I} \cup \mathrm{Im}(X_\ell)\big\} > -1$ and $b = \sup_{x\in\mathbb{R}} \big\{x \in \mathcal{I} \cup \mathrm{Im}(X_\ell)\big\} < 1$ such that

$$\mathbb{E}[(P_k(X_\ell) - P_k(X))^2] = \mathbb{E}[P_k'(\xi)^2 (X_\ell - X)^2] \leq \|P_k'\|_{L^\infty}^2 \mathbb{E}[(X_\ell - X)^2].$$

Recalling the classical result [26, Theorem 7.32.2, p. 168,(4.21.7)] that

$$\|P_k'\|_{L^\infty(-1,1)} \lesssim k^{\frac{5}{2}}, \qquad \text{and} \qquad \|P_k'\|_{L^\infty(a,b)} \lesssim k, \quad -1 < a < b < 1$$

we obtain the assertion. □

Now we are in the position to prove the counterpart of Lemma 5 for the case of approximate estimation of moments where the exact moments from (11) are estimated by approximate (single level) sample means

$$\tilde{\mu}_k^{\mathrm{SL}} := E_M[P_k(X_L)]. \tag{24}$$

Lemma 6. *Suppose Assumptions 1 and 3 are satisfied and let μ_k and $\tilde{\mu}_k^{\mathrm{SL}}$ be defined by (11) and (24) respectively. Then it holds that*

$$\mathbb{E}\left[\sum_{k=1}^{R}\left(\mu_k - \tilde{\mu}_k^{\mathrm{SL}}\right)^2\right] \lesssim R^{\delta+1}N_L^{-\beta} + \frac{R}{M}.$$

Proof. Recall that $\mathbb{E}[\tilde{\mu}_k^{\mathrm{SL}}] = \mathbb{E}[P_k(X_L)]$ which is in general not equal to μ_k. Then, analogously to (1), we observe that

$$\begin{aligned}\mathbb{E}\left[\left(\mu_k - \tilde{\mu}_k^{\mathrm{SL}}\right)^2\right] &= \mathbb{E}\big[(P_k(X) - P_k(X_L))\big]^2 + \frac{1}{M}\operatorname{Var}[P_k(X_L)] \\ &\lesssim k^{\delta}N_L^{-\beta} + M^{-1}(k^{\delta}N_L^{-\beta} + \|\rho\|_{L^\infty(-1,1)})\end{aligned}$$

where Jensen's inequality, Assumption 3 and Lemma 4 were used in the last step. Then the assertion follows by the summation over $k = 1 \dots R$. □

Theorem 4. *Suppose Assumptions 1 and 3 are satisfied and let $\ln(\rho) \in H^s(I)$ for some $s > 1$. Let $\{\tilde{\mu}_1^{\mathrm{SL}}, \dots, \tilde{\mu}_R^{\mathrm{SL}}\}$ be a sequence of perturbed moments defined in (24) for some fixed values of the parameters R, M, L and suppose $\tilde{\rho}_R^{\mathrm{SL}}$ is the corresponding perturbed Maximum Entropy solution. Then for any $\varepsilon > 0$ and $p \in (0,1)$ the parameters R, M, L can be selected such that the bound*

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{SL}}) < \varepsilon \tag{25}$$

is satisfied with probability at least $1-p$ and the cost of evaluation of all samples required for the recovery of $\tilde{\rho}_R^{\mathrm{SL}}$ satisfies the following asymptotic relations: when $\ln(\rho) \in H^s(I)$, $s > 1$, it holds that

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{SL}}) \sim p^{-\frac{\beta+\gamma}{\beta}}\varepsilon^{-\frac{\beta+\gamma}{\beta}-\frac{1}{2s}\frac{\beta+\gamma+\gamma\delta}{\beta}}, \tag{26}$$

when, moreover, $\ln(\rho)$ is analytic, the computational cost scales as

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{SL}}) \sim p^{-\frac{\beta+\gamma}{\beta}}\varepsilon^{-\frac{\beta+\gamma}{\beta}}|\ln(\varepsilon)|^{\frac{\beta+\gamma+\gamma\delta}{\beta}}. \tag{27}$$

Proof. Recalling decomposition (6), Theorem 2 and Lemma 6 we get

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{SL}}) \lesssim D_{\mathrm{KL}}(\rho\|\rho_R) + \frac{1}{p}\left(R^{\delta+1}N_L^{-\beta} + \frac{R}{M}\right) \tag{28}$$

with probability at least $1-p$, whereas the computational cost satisfies $\mathcal{C}(\tilde{\rho}_R^{\mathrm{SL}}) = M \cdot \mathcal{C}_L \sim MN_L^{\gamma}$. For a fixed computation cost the expression in parentheses in (28) is minimized for

$$M \sim R^{-\delta}N_L^{\beta}.$$

In this case we can estimate (28) by

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{SL}}) \lesssim D_{\mathrm{KL}}(\rho\|\rho_R) + \frac{1}{p}R^{\delta+1}N_L^{-\beta}. \tag{29}$$

Theorem 1 allows to determine the optimal choice of the number of moments R depending on the smoothness of the log-density $\ln(\rho)$. In particular, assuming that $\ln(\rho) \in H^s(I)$ with $s > 1$ we find with (9) that (29) is minimised when

$$R^{-2s} \sim \frac{1}{p}R^{\delta+1}N_L^{-\beta} \sim \varepsilon,$$

or, equivalently, when $R \sim \varepsilon^{-\frac{1}{2s}}$ and $N_L \sim (p^{-1}R^{2s+\delta+1})^{\frac{1}{\beta}}$. In this case the computational cost is proportional to

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{SL}}) \sim MN_L^{\gamma} \sim R^{-\delta}N_L^{\beta+\gamma} \sim p^{-\frac{\beta+\gamma}{\beta}}\varepsilon^{-\frac{\beta+\gamma}{\beta}-\frac{1}{2s}\frac{\beta+\gamma+\gamma\delta}{\beta}}.$$

On the other hand, when $\ln(\rho)$ is analytic, we select $R = -\frac{1}{2}\log_a(\varepsilon)$ and $N_L \sim (p^{-1}R^{\delta+1}a^{2R})^{\frac{1}{\beta}}$. Then (25) is satisfied and the computational cost is proportional to

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{SL}}) \sim p^{-\frac{\beta+\gamma}{\beta}}\varepsilon^{-\frac{\beta+\gamma}{\beta}}|\ln(\varepsilon)|^{\frac{\beta+\gamma+\gamma\delta}{\beta}}$$

The proof is complete. □

4.3. Approximate multilevel Monte Carlo estimators

Finally, we analyse the Multilevel Monte Carlo estimator for the statistical moments. For this purpose we define

$$\tilde{\mu}_k^{\mathrm{ML}} := E^{\mathrm{ML}}[P_k(X)] = \sum_{\ell=1}^{L} E_{M_\ell}[P_k(X_\ell) - P_k(X_{\ell-1})], \quad X_0 := 0 \tag{30}$$

where $E_{M_\ell}[P_k(X_\ell) - P_k(X_{\ell-1})]$, $\ell = 1, \ldots, L$ are based on independent samples.

Lemma 7. *Suppose Assumption 1 and 3 are satisfied and let μ_k and $\tilde{\mu}_k^{\mathrm{ML}}$ be defined by* (11) *and* (30) *respectively. Then it holds that*

$$\mathbb{E}\Big[\sum_{k=1}^{R}\big(\mu_k-\tilde{\mu}_k^{\mathrm{ML}}\big)^2\Big]\lesssim R^{\delta+1}\bigg(N_L^{-\beta}+\sum_{\ell=1}^{L}N_\ell^{-\beta}M_\ell^{-1}\bigg).$$

Proof. The Lemma follows by the same arguments as in Lemma 6 and analogously to the standard Multilevel Monte Carlo estimate (2). Recall that $\mathbb{E}[\tilde{\mu}_k^{\mathrm{ML}}]=\mathbb{E}[P_k(X_L)]$, and thus by Jensen's inequality and Assumption 3

$$\begin{aligned}\mathbb{E}\big[(\mu_k-\tilde{\mu}_k^{\mathrm{ML}})^2\big]&=\mathbb{E}\big[P_k(X)-P_k(X_L)\big]^2\\&\quad+\sum_{\ell=2}^{L}\frac{1}{M_\ell}\operatorname{Var}\big[P_k(X_\ell)-P_k(X_{\ell-1})\big]+\frac{1}{M_1}\operatorname{Var}[P_k(X_1)]\\&\lesssim k^{\delta}\bigg(N_L^{-\beta}+\sum_{\ell=1}^{L}M_\ell^{-1}N_\ell^{-\beta}\bigg)+\frac{1}{M_1}\|\rho\|_{L^\infty(-1,1)}.\end{aligned}$$

The last summand can be absorbed by the sum in parentheses, thus the assertion follows by summation over $k=1\dots R$. □

Theorem 5. *Suppose Assumptions 1 and 3 are satisfied. Let $\{\tilde{\mu}_1^{\mathrm{ML}},\dots,\tilde{\mu}_R^{\mathrm{ML}}\}$ be a sequence of perturbed moments defined in* (30) *for some fixed values of the parameters R,L and $\{M_1,\dots,M_L\}$ and suppose $\tilde{\rho}_R^{\mathrm{ML}}$ is the corresponding perturbed Maximum Entropy solution. Then for any $\varepsilon>0$ and $p\in(0,1)$ the parameters R,L and $\{M_1,\dots,M_L\}$ can be selected such that the bound*

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R)<\varepsilon \tag{31}$$

holds with probability at least $1-p$ and the cost of evaluation of all samples required for the recovery of $\tilde{\rho}_R^{\mathrm{ML}}$ satisfies the following asymptotic relations: when $\ln(\rho)\in H^s(I)$ with $s>1$ it holds that

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}})\sim\begin{cases}p^{-1}\varepsilon^{-1-\frac{\delta+1}{2s}}, & \text{if } \beta>\gamma,\\ p^{-1}\varepsilon^{-1-\frac{\delta+1}{2s}}(|\ln(p)|+|\ln(\varepsilon)|)^2, & \text{if } \beta=\gamma,\\ p^{-\frac{\gamma}{\beta}}\varepsilon^{-\frac{\gamma}{\beta}-\frac{(\delta+1)\gamma}{2s\beta}}, & \text{if } \beta<\gamma,\end{cases} \tag{32}$$

when $\ln(\rho)$ *is analytic the computational cost is proportional to*

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}}) \sim \begin{cases} p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|^{\delta+1}, & \text{if } \beta > \gamma, \\ p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|^{\delta+1}(|\ln(p)| + |\ln(\varepsilon)|)^2, & \text{if } \beta = \gamma, \\ p^{-\frac{\gamma}{\beta}}\varepsilon^{-\frac{\gamma}{\beta}}|\ln(\varepsilon)|^{(\delta+1)\frac{\gamma}{\beta}}, & \text{if } \beta < \gamma. \end{cases} \tag{33}$$

Proof. Recalling decomposition (6), Theorem 2 and Lemma 7 we get

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{ML}}) \lesssim D_{\mathrm{KL}}(\rho\|\rho_R) + \frac{R^{\delta+1}}{p}\left(N_L^{-\beta} + \sum_{\ell=1}^{L} M_\ell^{-1} N_\ell^{-\beta}\right) \tag{34}$$

with probability at least $1-p$. We start by minimizing the expression in the parentheses w.r.t. $\{M_1, \dots, M_L\}$ so that the computational cost

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}}) = \sum_{\ell=1}^{L} M_\ell \mathcal{C}_\ell \sim \sum_{\ell=1}^{L} M_\ell N_\ell^\gamma$$

remains fixed. Solving this constraint minimization problem analytically we find the optimal sample size

$$M_\ell \sim N_\ell^{-\frac{\beta+\gamma}{2}} \begin{cases} N_L^\beta, & \text{if } \beta > \gamma, \\ L N_L^\beta, & \text{if } \beta = \gamma, \\ N_L^{\frac{\beta+\gamma}{2}}, & \text{if } \beta < \gamma, \end{cases}$$

which is the standard selection of the sample size in MLMC (cf. [4, 8]). This implies the computational cost

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}}) \sim \begin{cases} N_L^\beta, & \text{if } \beta > \gamma, \\ L^2 N_L^\beta, & \text{if } \beta = \gamma, \\ N_L^\gamma, & \text{if } \beta < \gamma. \end{cases}$$

From (34) it follows that

$$D_{\mathrm{KL}}(\rho\|\tilde{\rho}_R^{\mathrm{ML}}) \lesssim D_{\mathrm{KL}}(\rho\|\rho_R) + R^{\delta+1} p^{-1} N_L^{-\beta},$$

which is the same estimation as (29) and thus the error is minimized for the same R, N_L as in Theorem 4. Suppose that $\ln(\rho) \in H^s(I)$ for $s > 1$ then we have for $N_L \sim (p^{-1}R^{2s+\delta+1})^{\frac{1}{\beta}}$ and $R \sim \varepsilon^{-\frac{1}{2s}}$ the computational cost of order

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}}) \sim \begin{cases} p^{-1}\varepsilon^{-1-\frac{\delta+1}{2s}}, & \text{if } \beta > \gamma, \\ p^{-1}\varepsilon^{-1-\frac{\delta+1}{2s}}(|\ln(p)| + |\ln(\varepsilon)|)^2, & \text{if } \beta = \gamma, \\ p^{-\frac{\gamma}{\beta}}\varepsilon^{-\frac{\gamma}{\beta}-\frac{(\delta+1)\gamma}{2s\beta}}, & \text{if } \beta < \gamma. \end{cases}$$

For analytic $\ln(\rho)$ we choose $N_L \sim (p^{-1}R^{\delta+1}a^{2R})^{\frac{1}{\beta}}$ and $a^{-2R} \sim \varepsilon$. Hence the computational cost is proportional to

$$\mathcal{C}(\tilde{\rho}_R^{\mathrm{ML}}) \sim \begin{cases} p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|^{\delta+1}, & \text{if } \beta > \gamma, \\ p^{-1}\varepsilon^{-1}|\ln(\varepsilon)|^{\delta+1}(|\ln(p)| + |\ln(\varepsilon)|)^2, & \text{if } \beta = \gamma, \\ p^{-\frac{\gamma}{\beta}}\varepsilon^{-\frac{\gamma}{\beta}}|\ln(\varepsilon)|^{(\delta+1)\frac{\gamma}{\beta}}, & \text{if } \beta < \gamma \end{cases}$$

and the proof is complete. □

5. Numerical Experiments

In this section we present results of numerical experiments assessing theoretical convergence estimates from the previous sections. We describe a numerical algorithm for solution of the truncated moment problem and report results of the convergence studies.

5.1. Synthetic probability density functions

We start by setting up a series of synthetic problems involving random variables with known probability density functions of different regularity. In particular, we study approximations of density functions $\rho = \rho^1, \ldots, \rho^4$ of four different random variables, explicitly defined in Table 1 and visualized in Figure 1. Observe that ρ^1 is analytic, but not polynomial, ρ^2 is piecewise constant with one jump, ρ^3 and ρ^4 are piecewise linear and continuous with one kink. Notice that ρ^2 and ρ^4 violate the minimal smoothness assumption $\ln(\rho) \in H^s(-1,1)$ for some $s > 1$, so that the convergence theory from the previous sections is not applicable in these two cases.

For a given random variable X with the density function ρ we introduce an approximation X_ℓ at level ℓ with the density function ρ_ℓ as follows. Let $\mathcal{T}_\ell$ be a uniform mesh on the interval $I := [-1,1]$ consisting of $N_\ell := 2^{\ell+1}$ subintervals of the same length. Then let ρ_ℓ be the piecewise constant function on $\mathcal{T}_\ell$ determined by

$$\int_K (\rho(x) - \rho_\ell(x))\,dx = 0 \qquad \forall K \in \mathcal{T}_\ell.$$

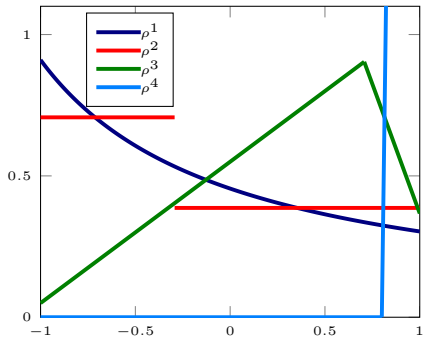

Figure 1: Graphs of the synthetic probability density functions $\rho^1, \ldots, \rho^4$.

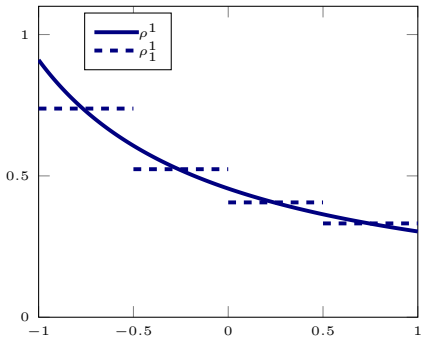

Figure 2: Density function ρ^1 and its approximation ρ^1_ℓ for $\ell = 1$.

Observe that ρ_ℓ is the L^2 projection of ρ. For example, $\rho^1(x)$ and the corresponding ρ^1_ℓ for $\ell = 1$ are visualized in Figure 2.

Let F be the cumulative distribution function of X. Then $X^i = F^{-1}(Z^i)$ is a sample of X where Z^i is a sample from the uniform distribution $Z \sim \mathcal{U}(0,1)$. Recall that the multilevel estimator (30) involves dependent pairs of samples X^i_ℓ and $X^i_{\ell-1}$ corresponding to two approximations of the sample X^i of different fidelity. To fulfil this dependence requirement we generate samples Z^i of the uniformly distributed random variable $Z \in \mathcal{U}(0,1)$ and declare $X^i_\ell := F^{-1}_\ell(Z^i)$ and $X^i_{\ell-1} := F^{-1}_{\ell-1}(Z^i)$ where F_ℓ is the cumulative distribution function of X_ℓ. The estimators (17), (24) and (30) for the Legendre moments are then evaluated directly by computing involved sums.

pdf	ρ^1	ρ^2	ρ^3	ρ^4
x_0	-1	$\sqrt{\frac{1}{2}} - 1$	$\sqrt{\frac{1}{2}}$	0.8
$\rho(x),\ x < x_0$		$\sqrt{\frac{1}{2}}$	$0.5x + 0.55$	0
$\rho(x),\ x \geq x_0$	$\frac{1}{\log(3)} \frac{1}{x+2}$	$\frac{1}{4-\sqrt{2}}$	$\approx -1.8314x + 2.1985$	$50x - 40$

Table 1: The definition of the density functions $\rho^1, \ldots, \rho^4$. In the second row x_0 denotes the location of a possible discontinuity of ρ^n or its derivative. The second and third row contain explicit expressions for ρ^n on both sides of x_0.

5.2. Solution algorithm

In this section we describe the numerical procedure of computing the Maximum Entropy solution ρ_R for a given set of moments $\mu_1, \ldots, \mu_R$ (strictly speaking the moments μ_k are estimated by sampling and therefore are inexact, but the skip the tilde-notation for the perturbed quantities until the end of this section to simplify the notations). For a fixed basis of global algebraic polynomials $\{\phi_0, \ldots, \phi_R\}$ this task reduces to finding unknown expansion coefficients $\lambda = (\lambda_0, \ldots, \lambda_R)^\top$ so that (4) is fulfilled. This leads to a system of nonlinear equations $F(\lambda) = 0$ with

$$F_j(\lambda) = \int_{-1}^{1} \phi_j(x)\rho[\lambda](x)\,dx - \mu_j, \qquad \rho[\lambda] := \rho_R = \exp\left(\sum_{k=0}^{R} \lambda_k \phi_k\right) \tag{35}$$

which can be solved by the Newton method. The Newton update step reads

$$\lambda^{(m+1)} = \lambda^{(m)} - J[\lambda^{(m)}]^{-1} F(\lambda^{(m)}), \qquad m = 1, 2, \ldots$$

with the Jacobi matrix

$$J[\lambda^{(m)}]_{jk} = \int_{-1}^{1} \phi_j(x)\phi_k(x)\rho[\lambda^{(m)}](x)\,dx.$$

The choice of the basis $\{\phi_k\}$ has, of course, no influence on the solution of the problem in the exact arithmetic, but it has a crucial impact on stability and convergence properties of the Newton algorithm. For example, for the monomial basis $\phi_k(x) = x^k$ the Jacobi matrix is a Hankel matrix with the condition number growing exponentially with the number of moments R, cf. [28]. On the other hand, the Jacobi matrix in the basis of Legendre polynomials $\phi_k(x) = P_k(x)$ is well-conditioned with linearly increasing condition number [29]. However, the constants in this estimate blow up when the minimum of the probability density function $\rho[\lambda]$ on $[-1, 1]$ is close to zero.

In our numerical experiments we use the basis of orthonormalized Legendre polynomials and define the initial value $\lambda^{(0)} = (\ln(\frac{1}{2}), 0, \ldots, 0)$ corresponding to the uniform distribution $\rho[\lambda^{(0)}] = \frac{1}{2}$. It is expected that the Newton algorithm will not perform well when the target probability density function ρ vanishes

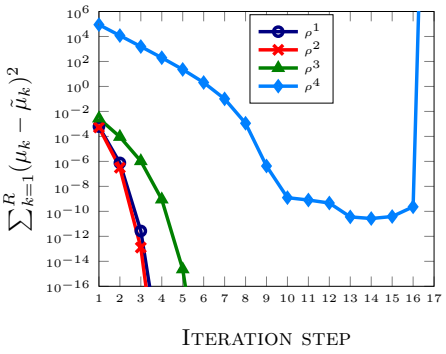

Figure 3: Convergence of the Newton method for $R = 100$ statistical moments.

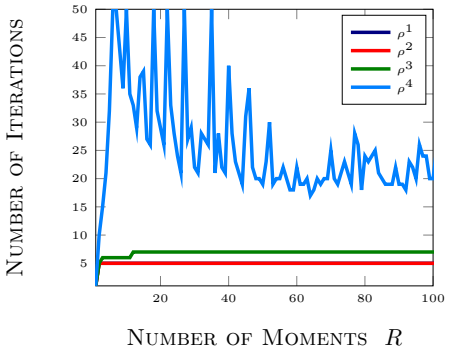

Figure 4: The number of Newton iterations needed until convergence for R different number of prescribed statistical moments.

or is close to zero on the interval $[-1, 1]$. When the Newton iteration does not achieve the prescribed level of accuracy, we select as an output the vector of coefficients $\lambda^{(m_*)}$ at the iteration m_* having the smallest residual, that is $m_* = \text{argmin}_m \sum_{k=0}^{R}(\mu_k - \mu_k^{(m)})^2$ where $\mu_k^{(m)} = \int_{-1}^{1} P_k(x)\rho[\lambda^{(m)}](x)\,dx$.

Now we discuss a series of numerical experiments where we use the exact first R moments $\mu_1, \ldots, \mu_R$ of the target density ρ (i.e. in this case the estimation error equals zero) in the nonlinear system (35) and compute the Maximum Entropy solution using the Newton method as described above. Figure 3 shows the convergence of the Newton algorithm for $R = 100$ moments for the four synthetic density functions $\rho = \rho^1, \ldots, \rho^4$. The plotted value is the squared Euclidean norm of the error between Legendre coefficients of the logarithms of ρ and $\rho[\lambda^{(m)}]$

$$\sum_{k=1}^{R}(\lambda_k - \lambda_k^{(m)})^2.$$

Observe that the Newton algorithm for the first three density functions achieves machine precision in a few (4-5) iterations. The fourth example starts with a slow convergence, but then the error saturates and the method diverges at the 17th iteration.

Figure 4 shows the number of iterations needed to converge to precision

10^{-15} dependent on the number of prescribed moments. The first three probability density functions only need a few iteration steps to converge almost independently of the number of moments. For ρ^4 the Newton method does not converge to the tolerance 10^{-15} when $R > 5$ and returns only an approximate solution.

5.3. Convergence of the Maximum Entropy method for synthetic density functions

In this section we discuss convergence of the Maximum Entropy method and compare it with a Kernel Density Estimator from [30]. Here the target distribution ρ is one of the synthetic distributions $\rho^1, \ldots, \rho^4$ defined in Section 5.1.

First we study the behaviour of the truncation error estimated in Theorem 1 which is independent of the type of the moment estimator. Recall that $\log(\rho^2), \log(\rho^4) \notin H^1(I)$ violating assumptions of Theorem 1. Nonetheless, we observe in Figure 5 that the slope of the convergence curves for ρ^1, ρ^2, ρ^3 is as predicted in Theorem 1. The approximations to ρ^4 converge at a reduced rate of about R^{-1}. This indicates that the minimal smoothness assumption $\ln(\rho) \in H^1(I)$, $s > 1$ may be relaxed. Notice that the convergence curve for ρ^4 is non-monotone. This effect may be due to inaccurate iterative approximation of ρ_R^4 by the Newton method.

Next, we validate the stability bound (14) in Lemma 3. For this we introduce an artificial noise in the moments $\tilde{\mu}_k = (1 + y\xi_k)\mu_k$, where ξ_k are iid normal random variables and the scaling parameter y taking different values for each data point. Figure 6 indicates that (14) is a sharp bound for ρ^1, ρ^2 and ρ^3 whereas the graph for ρ^4 shows a very different behaviour. It can be explained by the very large value of the constant C_R for the approximation ρ_R^4 (indeed, in this example we estimate $C_{20} > 10^{100}$) and thus it is not surprising that the stability relation (14) is not realized. On the other hand, this effect (and also the non-monotone convergence in Figure 5) can be explained by the poor accuracy of the Newton solution in the case $\rho = \rho^4$, as mentioned in Section 5.2.

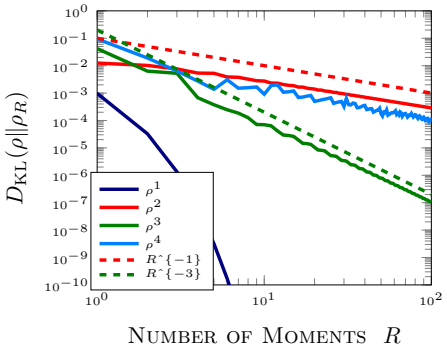

Figure 5: Truncation error versus the number of exact moments.

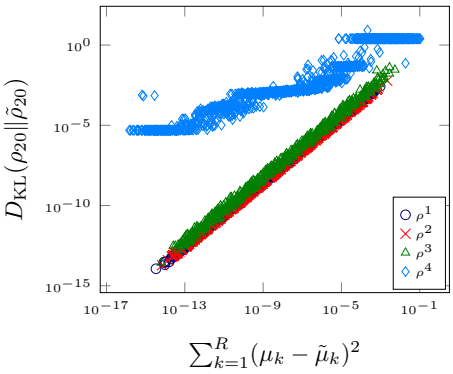

Figure 6: Estimation error versus the size of the moment perturbation with $R = 20$ perturbed moments.

Our next aim is to verify the bound 1) in Assumption 3. In Figure 7 we plot the values $\mathbb{E}[(X_\ell - X)^2]$ for varying ℓ. The slope of the convergence curves indicates that $\beta \approx 4$ in all four examples. In Figure 8 we show the values $\mathbb{E}[(P_k(X_\ell) - P_k(X))^2]$ for the fixed approximation level $\ell = 3$ and varying polynomial degree k. In all cases we observe that $\delta \approx 2$. Of course, in the described synthetic examples the generation cost for approximate samples is independent of ℓ. To make a meaningful test for the MLMC estimator (30) with synthetic density functions we make a convention that the cost to generate one sample of X_ℓ is proportional to N_ℓ^4, i.e. Assumption 3 is fulfilled with $\gamma = \beta$.

As seen from the proof of Theorem 5, the number of the statistical moments R has to be coupled to other discretization parameters, e.g. the sample size. To verify the statement of Theorem 5 we chose the optimal value R manually (a practical estimator would ideally choose R in an adaptive manner). In what follows we compare the Maximum Entropy method with the Kernel Density Estimators from [30] (their implementation makes use of an adaptive selection of the kernel width).

We start by verifying the statement of Theorem 3. Figures 9 and 10 show the error of the Maximum Entropy method and the Kernel Density Estimator

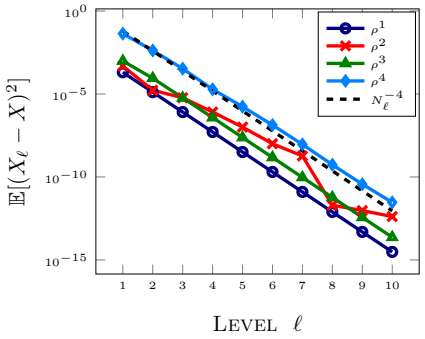

Figure 7: Numerical approximation of the parameter β in Assumption 3.

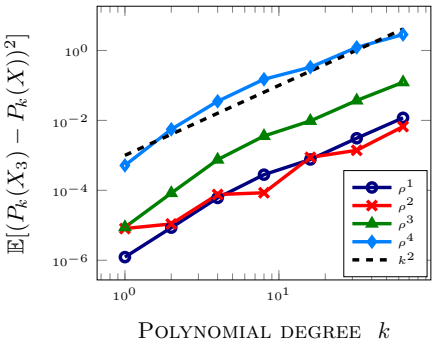

Figure 8: Numerical approximation of the parameter δ in Assumption 3.

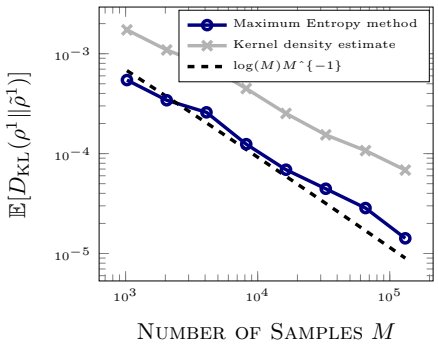

Figure 9: The mean of the KL-divergence for ρ^1 and its approximation by the ME and KDE with M exact samples.

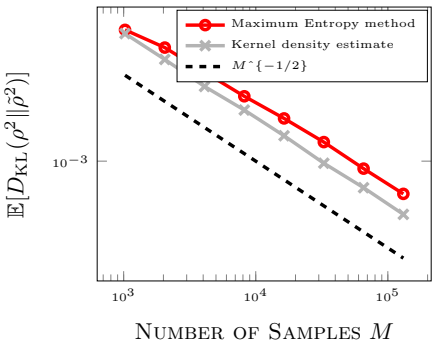

Figure 10: The mean of the KL-divergence for ρ^2 and its approximation by the ME and KDE with M exact samples.

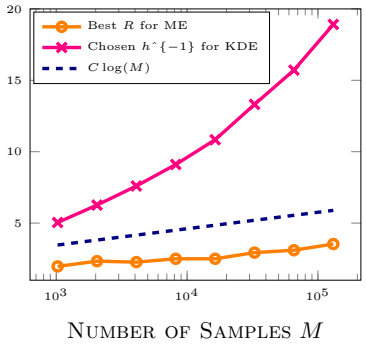

Figure 11: The average values of h^{-1} for the KDE and R for the ME for M exact samples of ρ^1.

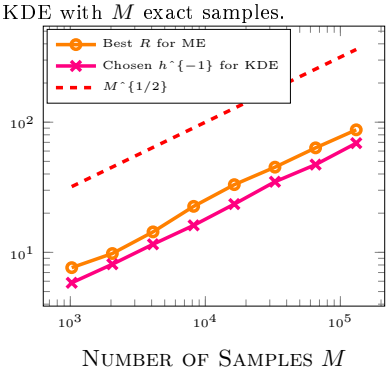

Figure 12: The average values of h^{-1} for the KDE and R for the ME for M exact samples of ρ^2.

from [30] dependent on the sample size for ρ^1 and ρ^2 (the data points represent averages of 30 independent simulations). The two other examples with $\rho = \rho^3$ and ρ^4 show similar behaviour and therefore are omitted. The dashed line indicates the rate of the Maximum Entropy method predicted by Theorem 3. Recall that the bound (18) only holds on a set of probability smaller than 1, therefore there might be runs with no ME solution. However, this problem has never occurred in the experiments for ρ^1 and ρ^2. This seems reasonable if we recall Figure 6. Comparing the Maximum Entropy method and the Kernel Density Estimate we observe in both cases a similar convergence behaviour. In Figures 11 and 12 we show the relations between R and h^{-1} (where h is the kernel width) and the number of moments M. Notice the very different behaviour of $R(M)$ and $h^{-1}(M)$ for the analytic density function ρ^1 and the step-function ρ^2. The behaviour $R = R(M)$ is in excellent agreement with the one suggested in the proof of Theorem 3 in both cases.

In Figures 13 and 14 we compare the MLMC-ME approach to the Kernel Density Estimator for ρ^1 and ρ^2 (again, each data point is an average over 30 independent runs). Each curve in the KDE method corresponds to a fixed level of approximation $\ell = 1, \ldots, 4$ and an increasing number of samples. This explains the saturation of the KDE convergence curves once the sampling error has achieved the magnitude of the fixed approximation error. The convergence of the MLMC-ME approach is in good agreement with Theorem 5 and is faster than that of the envelope of all KDE curves. By construction, it is naturally refinable in both the sample size and the level of sample approximation, and therefore does not saturate.

5.4. Application to contact with rough random obstacles

In this section we apply the Multilevel Monte Carlo Maximum Entropy method to a class of contact problems with rough random obstacles. We utilize the mathematical framework and notations from the recent articles [8, 11] and recall the most important ingredients for the sake of completeness.

Let $D = [-1, 1]^2$ and assume that the obstacle is parametrized by a con-

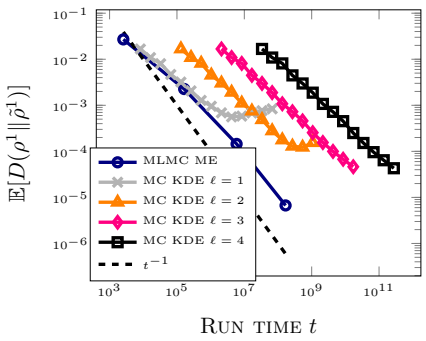

Figure 13: The mean KL-divergence of ρ^1 approximated by the MLMC-ME and KDE on different levels.

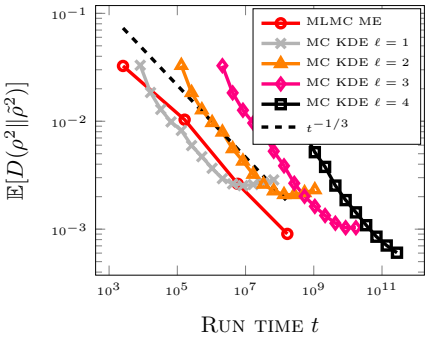

Figure 14: The mean KL-divergence of ρ^2 approximated by the MLMC-ME and KDE on different levels.

tinuous function ψ satisfying $\psi \leq 0$ on ∂D and $f \in L^2(D)$. The deterministic obstacle problem can be formulated as finding $u : D \to \mathbb{R}$ such that

$$\begin{cases} -\Delta u & \geq & f & \text{in } D, \\ u & \geq & \psi & \text{in } D, \\ (-\Delta u - f)(u - \psi) & = & 0 & \text{in } D, \\ u & = & 0 & \text{on } \partial D. \end{cases} \tag{36}$$

The weak formulation of (36) is a variational inequality of the first kind having a unique solution $u \in H_0^1(D)$ satisfying $u \geq \psi$ a.e. in D and depending continuously on the data ψ and f. We are interested in the probability distribution of the area of the coincidence set

$$X = |\Lambda|, \qquad \Lambda = \{x \in D \ : \ u(x) = \psi(x)\}$$

in the case when $\psi = \psi(x, \omega)$ is uncertain. Precisely, we assume that

$$\psi(x) = \sum_{q_0 \leq |q| \leq q_s} B_q(H) \cos(q \cdot x + \varphi_q), \tag{37}$$

where the oscillation amplitudes $B_q(H)$ obey the law

$$B_q(H) = \frac{\pi}{25}(2\pi \max(|q|, q_l))^{-(H+1)}, \qquad q_0 \leq |q| \leq q_s,$$
$$q_0 = 1, \quad q_l = 10, \quad q_s = 26$$

see [8, Sect. 8], [11, Sect. 6] and references therein for the details. The values of the *Hurst coefficient* H correspond to obstacles of different roughness [31] (notice that H is independent of the frequency q). In the subsequent numerical experiments we assume that H is uniformly distributed in $[0, 1]$ and the phase shifts φ_q in (37) are uniformly distributed in $[0, 2\pi]$.

Evidently, exact samples of X are out of reach and we shall work with approximate samples $X_\ell \approx X$ at the approximation level ℓ. The samples X_ℓ are obtained from the approximate Finite Element solutions u_ℓ computed on the Finite Element meshes $\mathcal{T}_\ell$ consisting of congruent triangles: the coarsest solution u_1 involves 13 degrees of freedom whereas the finest solution u_8 involves 130 561 degrees of freedom. Clearly, the sizes of the the exact coincidence set X and of its Finite Element approximations X_ℓ are enclosed in the interval $[0, 4]$. In the forthcoming numerical experiments the unknown probability density ρ is approximated on a smaller interval $[x_* - \Delta, x^* + \Delta] \subset [0, 4]$ where x_* and x^* are the smallest and the largest realizations of X_ℓ for all $\ell = 1, \ldots, L$ and $\Delta := (x^* - x_*)/10$.

Figure 15 shows graphs of the Maximum Entropy solutions for different number of statistical moments computed by the MLMC. Observe that a significant number of moments (50 to 60) is required to reach a reasonable approximation of the target density function which appears quite concentrated around the value $x_0 = 1.13$ and having a heavy tail towards smaller values of X (notice non-negligible spurious oscillations in that region). The parameters are selected so that the computational cost for every ME solution in Figure 15 is the same. The outcomes of this simulation can be compared directly with approximate density functions computed by KDE simulations at fixed levels $\ell = 4, \ldots, 7$, see Figure 16. The KDE density functions are quite rough in the left tail region and are quantitatively close to the ME density functions with 50–60 statistical moments. The computation of each KDE density requires the same computational cost which is, however, by factor 9 smaller than the computational cost for the ME density functions. Figure 17 shows the zoom into the tips of the KDE and ME distributions.

Next, we consider convergence of the truncation error $D_{\mathrm{KL}}(\rho||\rho_R)$. Since the target density function ρ is unknown, we use the overkill approximation $\rho \approx \tilde{\rho}_{60}^{\mathrm{ML}}$ instead which involves $L = 9$ levels of approximation and $M_L = 300$ samples on the finest grid. The convergence history for different values of R is shown in Figure 18 and indicates the convergence rate of order R^{-1}.

The behaviour of the estimation error $D_{\mathrm{KL}}(\rho_R\|\tilde{\rho}_R)$ can be seen in Figure 19 (each data point is an average over 30 independent runs). The approximations $\tilde{\rho}_R$ are computed with 10 samples on the finest level. Observe that for rising number of moments the slopes of the convergence curves rapidly become flat. To interpret this effect we show in Figure 20 the behaviour of the corresponding moment perturbation. Recall that due to Lemma 3 this is an upper bound for $D_{\mathrm{KL}}(\rho_R\|\tilde{\rho}_R)$. Here the slope of the convergence curves rapidly becomes parallel and decaying for all moments up to $R = 50$. This indicates that the stability constant C_R in (14) should grow significantly with increasing R. Indeed, we observe huge values of C_R in the course of the simulation.

6. Conclusions and discussion

In this work we have developed a complete convergence theory of the Multilevel Monte Carlo Maximum Entropy method and presented numerical examples showing that this approach provides a good alternative for numerical approximation of probability density function of the system output. If the target probability density function is strictly positive and smooth, this can lead to significant savings in computation time. On the other hand, when the target probability density function has low regularity, the savings are smaller, but the method still has a good error-versus-cost relation. The convergence of the method is not guaranteed if the probability density function is not strictly positive.

An important open question for practical computations, which lies outside the scope of this paper, is how to choose the number of statistical moments R in the Maximum Entropy method dependent on the discretization and sampling parameters in the case when the smoothness of the target distribution is not

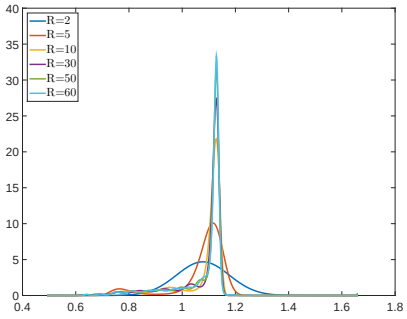

Figure 15: Solutions of the MLMC-ME for different number of moments R.

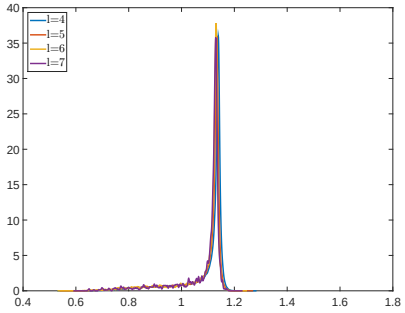

Figure 16: Solutions of the KDE with samples taken from different discretization levels ℓ.

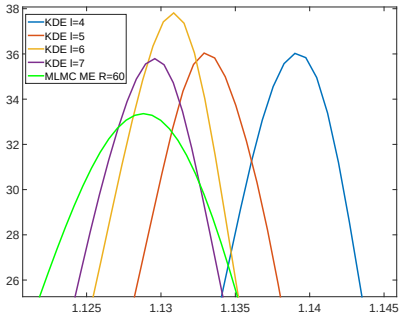

Figure 17: Zoom into the tips of the probability density functions in Figures 15 and 16.

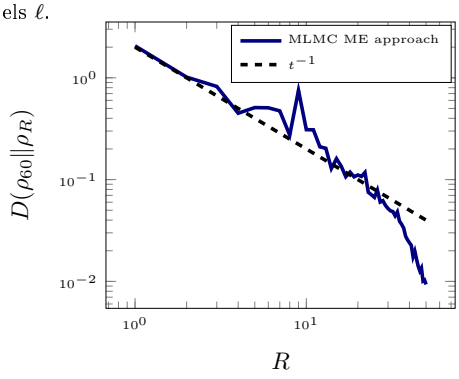

Figure 18: Convergence of the MLMC-ME solution for an increasing number of moments.

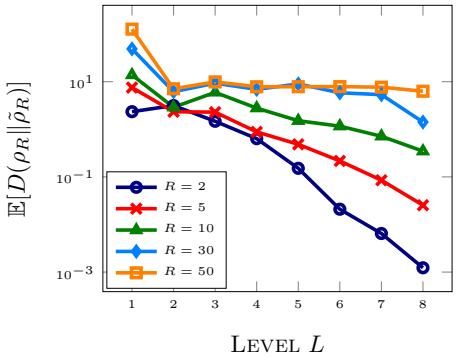

Figure 19: Expected KL-divergence of the ME solution for approximating R moments.

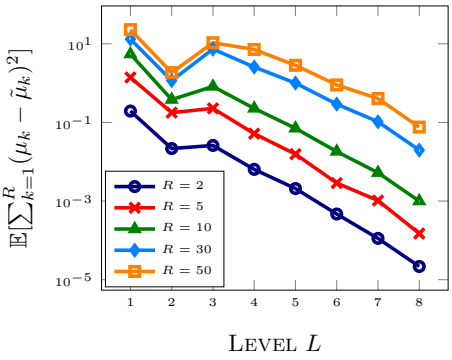

Figure 20: Mean square error of R approximated moments with the MLMC.

known. A possible adaptive strategy may start with low values of R and increase them in the course of the simulation when necessary.

References

[1] M. B. Giles, Multilevel Monte Carlo path simulation, Oper. Res. 56 (3) (2008) 607–617.

[2] S. Heinrich, Multilevel Monte Carlo methods, in: Large-Scale Scientific Computing, Third International Conference, LSSC 2001, Sozopol, Bulgaria, June 6–10, 2001, Vol. 2179, 2011, pp. 58–67.

[3] A. Barth, C. Schwab, N. Zollinger, Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients, Numer. Math. 119 (1) (2011) 123–161.

[4] K. A. Cliffe, M. B. Giles, R. Scheichl, A. L. Teckentrup, Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients, Comput. Vis. Sci. 14 (1) (2011) 3–15.

[5] M. B. Giles, C. Reisinger, Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance, SIAM J. Financial Math. 3 (1) (2012) 572–592.

[6] S. Mishra, C. Schwab, J. Šukys, Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions, J. Comput. Phys. 231 (8) (2012) 3365–3388.

[7] R. Kornhuber, C. Schwab, M.-W. Wolf, Multilevel Monte Carlo finite element methods for stochastic elliptic variational inequalities, SIAM J. Numer. Anal. 52 (3) (2014) 1243–1268.

[8] C. Bierig, A. Chernov, Convergence analysis of Multilevel Monte Carlo variance estimators and application for random obstacle problems, Numer. Math. 130 (4) (2015) 579–613.

[9] A. Abdulle, A. Barth, C. Schwab, Multilevel Monte Carlo methods for stochastic elliptic multiscale PDEs, Multiscale Model. Simul. 11 (4) (2013) 1033–1070.

[10] H. Hoel, K. J. Law, R. Tempone, Multilevel ensemble Kalman filtering, arXiv preprint arXiv:1502.06069.

[11] C. Bierig, A. Chernov, Estimation of arbitrary order central statistical moments by the Multilevel Monte Carlo methods, Stoch. Partial Differ. Equ. Anal. Comput. (in press), DOI: 10.1007/s40072-015-0063-9

[12] M. B. Giles, T. Nagapetyan, K. Ritter, Multilevel Monte Carlo approximation of distribution functions and densities, SIAM/ASA J. Uncertain. Quantif. 3 (1) (2015) 267–295.

[13] S. Karlin, L. S. Shapley, Geometry of moment spaces, Mem. Amer. Math. Soc. 1953 (12) (1953) 93.

[14] A. Tagliani, Numerical aspects of finite Hausdorff moment problem by maximum entropy approach, Appl. Math. Comput. 118 (2-3) (2001) 133–149.

[15] G. Athanassoulis, P. Gavriliadis, The truncated Hausdorff moment problem solved by using kernel density functions, Probabilistic Engineering Mechanics 17 (3) (2002) 273–291.

[16] R. M. Mnatsakanov, Hausdorff moment problem: reconstruction of distributions, Statist. Probab. Lett. 78 (12) (2008) 1612–1618.

[17] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. (2) 106 (1957) 620–630.

[18] A. R. Barron, C.-H. Sheu, Approximation of density functions by sequences of exponential families, Ann. Statist. 19 (3) (1991) 1347–1369.

[19] J. M. Borwein, A. S. Lewis, Convergence of best entropy estimates, SIAM J. Optim. 1 (2) (1991) 191–205.

[20] M. B. Giles, Multilevel Monte Carlo methods, Acta Numer. 24 (2015) 259–328.

[21] J. M. Borwein, A. S. Lewis, Duality relationships for entropy-like minimization problems, SIAM J. Control Optim. 29 (2) (1991) 325–338.

[22] I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, Studia Sci. Math. Hungar. 2 (1967) 299–318.

[23] S. Kullback, A lower bound for discrimination information in terms of variation, IEEE Trans. Inf. Theory 13 (1967) 126–127.

[24] C. Schwab, p- and hp-finite element methods, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 1998, theory and applications in solid and fluid mechanics.

[25] P. J. Davis, Interpolation and approximation, Dover Publications, Inc., New York, 1975, republication, with minor corrections, of the 1963 original, with a new preface and bibliography.

[26] G. Szegő, Orthogonal polynomials, 4th Edition, American Mathematical Society, Providence, R.I., 1975, american Mathematical Society, Colloquium Publications, Vol. XXIII.

[27] A. Hoorfar, M. Hassani, Inequalities on the Lambert W function and hyperpower function, JIPAM. J. Inequal. Pure Appl. Math. 9 (2) (2008), Article 51, 5.

[28] E. E. Tyrtyshnikov, How bad are Hankel matrices?, Numer. Math. 67 (2) (1994) 261–269.

[29] D. Fasino, Spectral properties of Hankel matrices and numerical solutions of finite moment problems, in: Proceedings of the International Conference on Orthogonality, Moment Problems and Continued Fractions (Delft, 1994), Vol. 65, 1995, pp. 145–155.

[30] Z. I. Botev, J. F. Grotowski, D. P. Kroese, Kernel density estimation via diffusion, Ann. Statist. 38 (5) (2010) 2916–2957.

[31] B. N. J. Persson, O. Albohr, U. Tartaglino, A. I. Volokitin, E. Tosatti, On the nature of surface roughness with application to contact mechanics, sealing, rubber friction and adhesion, J. Phys.: Condens. Matter 17 (2005) R1–R62.