The minimum cross-entropy method: A general algorithm for one-dimensional problems

J.C. Cuchí *(Universitat de Lleida, Spain)*

J.C. Angulo *(Instituto Carlos I, Universidad de Granada, Spain)*

A. Zarzo *(Universidad Politécnica de Madrid* and *Instituto Carlos I, Spain)*

MaxEnt, July 2006, CNRS, Paris, France.

Aim of this talk is two-fold:

- On one side, to present a **general** (but rather simple) **algorithm** based on standard optimization methods **to obtain the MinxEnt solutions**. It can be applied to "general" densities: discrete and continuous, domains: bounded and unbounded, and constraints, being able to manage mixed constraints problems.

Aim of this talk is two-fold:

- On one side, to present a **general** (but rather simple) **algorithm** based on standard optimization methods **to obtain the MinxEnt solutions**. It can be applied to "general" densities: discrete and continuous, domains: bounded and unbounded, and constraints, being able to manage mixed constraints problems.

- On the other, **to illustrate** by means of two very different examples **the way the algorithm works**, **showing the** (rather well known) **good and accurate behavior of the** *Minimum cross-entropy method* **(MinxEnt method) when applied to one-dimensional problems.**

The Problem and The MinxEnt solution

Generalized (one-dimensional) reduced "expectation-value" problem:

To construct approximations to a **pdf**, $f : D \subseteq R \rightarrow R^+$, from a finite set of expectation values ($i = 1, 2, \ldots, n$):

$$\int_{D \subseteq R} f(x)dx = \mu_0 \, , \quad \langle k_i(x) \rangle_{[f]} := \int_{D \subseteq R} k_i(x) f(x) dx = \mu_i \, ,$$

"Generalized" because not only expectation values of type

$$k_i(x) = x^i$$

are considered, but also

$$k_i(x) = e^{-jp_i x} \quad \textbf{or} \quad k_i(x) = j_0(p_i x)$$

and others.

Moreover **"Mixed constraints"** are also allowed.

The Problem and The MinxEnt solution

Generalized (one-dimensional) reduced "expectation-value" problem:

To construct approximations to a **pdf**, $f : D \subseteq R \rightarrow R^+$, from a finite set of expectation values ($i = 1, 2, \ldots, n$):

$$\int_{D \subseteq R} f(x)dx = \mu_0 \, , \quad \langle k_i(x) \rangle_{[f]} := \int_{D \subseteq R} k_i(x) f(x) dx = \mu_i \, ,$$

The MinxEnt solution:

It is obtained by minimizing the relative entropy functional:

$$\mathcal{H}[f : f_0] = \int_{D \subseteq R} f(x) \ln \left[\frac{f(x)}{f_0(x)} \right] \, dx \, .$$

$f_0(x) \equiv$ prior information on $f(x)$, such that $\int_D f_0(x)dx = \mu_0$.

The Problem and The MinxEnt solution

The MinxEnt solution:

In terms of Lagrange multipliers λ_0 and $\mathbf{L} := (\lambda_1, \ldots, \lambda_n)$, the solution $f_n^{me}(x)$ (if it exists) is:

$$f_n^{me}(x) = \frac{\mu_0\, f_0(x)}{\mathcal{Z}(\mathbf{L})}\, \exp\left\{ -\sum_{i=1}^{n} \lambda_i\, k_i(x) \right\},$$

Partition function:

$$\mathcal{Z}(\mathbf{L}) := e^{-\lambda_0} = \int_D f_0(x)\, \exp\left\{ -\sum_{i=1}^{n} \lambda_i\, k_i(x) \right\} dx.$$

The Lagrange multipliers are solutions of the non-linear system:

$$\int_D k_i(x)\, f_n^{me}(x)\, dx = \mu_i\,, \quad i = 1, \ldots, n.$$

The Problem and The MinxEnt solution

The MinxEnt solution:

Existence, convergence and some other interesting properties of the MinxEnt solution have been widely studied in the literature. In this context and being non exhaustive, it is worth-mentioning the work of Csiszàr 1975, Einbu 1977, Johnson and Shore 1979–1981, Borwein and Lewis 1993 or Tagliani 2003, among others (see also J.C. Cuchí PhD Thesis (2005, in spanish), where a detailed summary of that properties has been recently done).

The Problem and The MinxEnt solution

The MinxEnt solution:

Dual Problem: The Lagrange multipliers are obtained by minimizing the relative entropy (with a minus) of the MinxEnt solution

$$\Gamma(\mathbf{L}) \;:=\; -\mathcal{H}[f_n^{me} : f_0] = -\lambda_0 + \sum_{i=1}^{n} \lambda_i \, \mu_i$$

$$= \; \mu_0 \, \ln \, \mathcal{Z}(\mathbf{L}) + \sum_{i=1}^{n} \lambda_i \, \mu_i \; - \; \mu_0 \, \ln \, \mu_0,$$

which is a convex function of them.

The Algorithm

A number of methods to deal with this optimization problem can be found in the literature. Among others and being non-exhaustive:

- Darroch and Ratcliff 1972.
- Mead and Papanicolau 1984.
- Turek 1988.
- Borwein and Huang 1995.
- Drabold et al. 2005.
- ...

Two main difficulties:

- These algorithms are not ready to work with Mixed constraints.
- When the number of moments increase, there appears some numerical instabilities.

The Algorithm

A typical algorithm of unidirectional descending.

Starting from one initial feasible point $\mathbf{L}^{(0)}$, a descending direction $\mathbf{d}^{(k)}$ is chosen on each iteration by solving the system

$$\mathbf{H}^{(k)} \cdot \mathbf{d}^{(k)} = -\nabla\Gamma(\mathbf{L}^{(\mathbf{k})}) \, .$$

We have employed Newton's algorithm and the BFGS or Broyden's algorithm (a quasi-Newton algorithm of rank 2).

Then a decision is taken on how much one should advance on the direction $\mathbf{d}^{(k)}$ using line-search with backtracking.

The Algorithm

Unbounded domains, e.g $[0, +\infty)$.

Our algorithm works in the following way

- First, it solves the problem for a set of finite intervals $[0, a)$, with increasing values of a, until the multiplier corresponding to the highest expectation value, $\langle x^n \rangle$, is positive and remains reasonably unchanged.
- Then, the solution for the highest value of a is used as a feasible initial value for the Newton's algorithm with a specific integration subroutine in unbounded intervals.

Difficulties.

- In some cases the Lagrange multipliers can have alternating signs and big absolute values (e.g. when the interval is not bounded, or the moment sequence increases fast enough).
- The Hessian matrix is ill conditioned.

To get round of this (at least partially) we have used Tchebyshev polynomials for rewriting the constraints in terms of them.

In most of the applications, it turns out that this strategy solve the multipliers problem **at the price to have big values for partition functions**. Moreover, using quad-precision, in most of the applications we have worked in detail, use of Tchebyshev polynomials also avoid the ill conditioning Hessian problem.

Example 1: Electron-pair density

Electron-pair density for the Helium atom: $h(\mathbf{u})$.

- $h(\mathbf{u})$ is the probability density associated to the inter-electronic vector $\mathbf{u} = \mathbf{r}_1 - \mathbf{r}_2$.
- Basic quantity in the study of the e^-– e^- correlation problem in many-electron systems.
- $h(\mathbf{u})d\mathbf{u}$ gives the probability of finding a pair of electrons with $\mathbf{r}_1 - \mathbf{r}_2$ between $\mathbf{u}$ and $\mathbf{u} + d\mathbf{u}$.
- In many cases, it is enough to consider its spherically averaged counterpart

$$h(u) := \frac{1}{4\pi} \int h(\mathbf{u})\, d\Omega_{\mathbf{u}} \quad \text{or} \quad H(u) := 4\pi u^2\, h(u)\,, \quad u \in [0, +\infty)\,.$$

Example 1: Electron-pair density

Electron-pair density for the Helium atom: $h(\mathbf{u})$.

- $h(\mathbf{u})$ is the probability density associated to the inter-electronic vector $\mathbf{u} = \mathbf{r}_1 - \mathbf{r}_2$.
- Basic quantity in the study of the e^-– e^- correlation problem in many-electron systems.
- $h(\mathbf{u})d\mathbf{u}$ gives the probability of finding a pair of electrons with $\mathbf{r}_1 - \mathbf{r}_2$ between $\mathbf{u}$ and $\mathbf{u} + d\mathbf{u}$.
- In many cases, it is enough to consider its spherically averaged counterpart

$$h(u) := \frac{1}{4\pi} \int h(\mathbf{u})\, d\Omega_{\mathbf{u}} \quad \text{or} \quad H(u) := 4\pi u^2\, h(u)\,, \quad u \in [0, +\infty)\,.$$

Why these problem ?

Example 1: Electron-pair density

Electron-pair density for the Helium atom: $h(\mathbf{u})$.

1. Using the Hylleras-type atomic wave-functions (Koga 1993) it is possible to compute accurately not only the density $H(u) = 4\pi u^2\, h(u)$, but its expectation values

$$\langle u^n \rangle = \int_0^{+\infty} u^n H(u) du = 4\pi \int_0^{+\infty} u^{n+2} h(u) du\,,\; n = -2, -1, 0, 1, \ldots,$$

and also its Hankel transform (related to the total scattering intensity)

$$K(k) = \int_0^{+\infty} H(u)\, j_0(ku)\, du \;=\; \int_0^{+\infty} 4\pi u\, \frac{\sin ku}{k}\, h(u)\, du, \quad k \in R^+\,,$$

where $j_0(ku) := \sin ku/(ku)$ is the spherical Bessel function of order zero.

2. The *overlap a priori function* (Koga 1984):

$$h_{ov}(u) = \left(\frac{\alpha}{(1+\gamma u^2)^2} + \frac{\alpha(3-\gamma u^2)}{(1+\gamma u^2)^3} + \frac{\beta(1-\gamma u^2)}{(1+\gamma u^2)^4} \right)\,.$$

Results on $H(u)$ for Helium atom

NOTATION

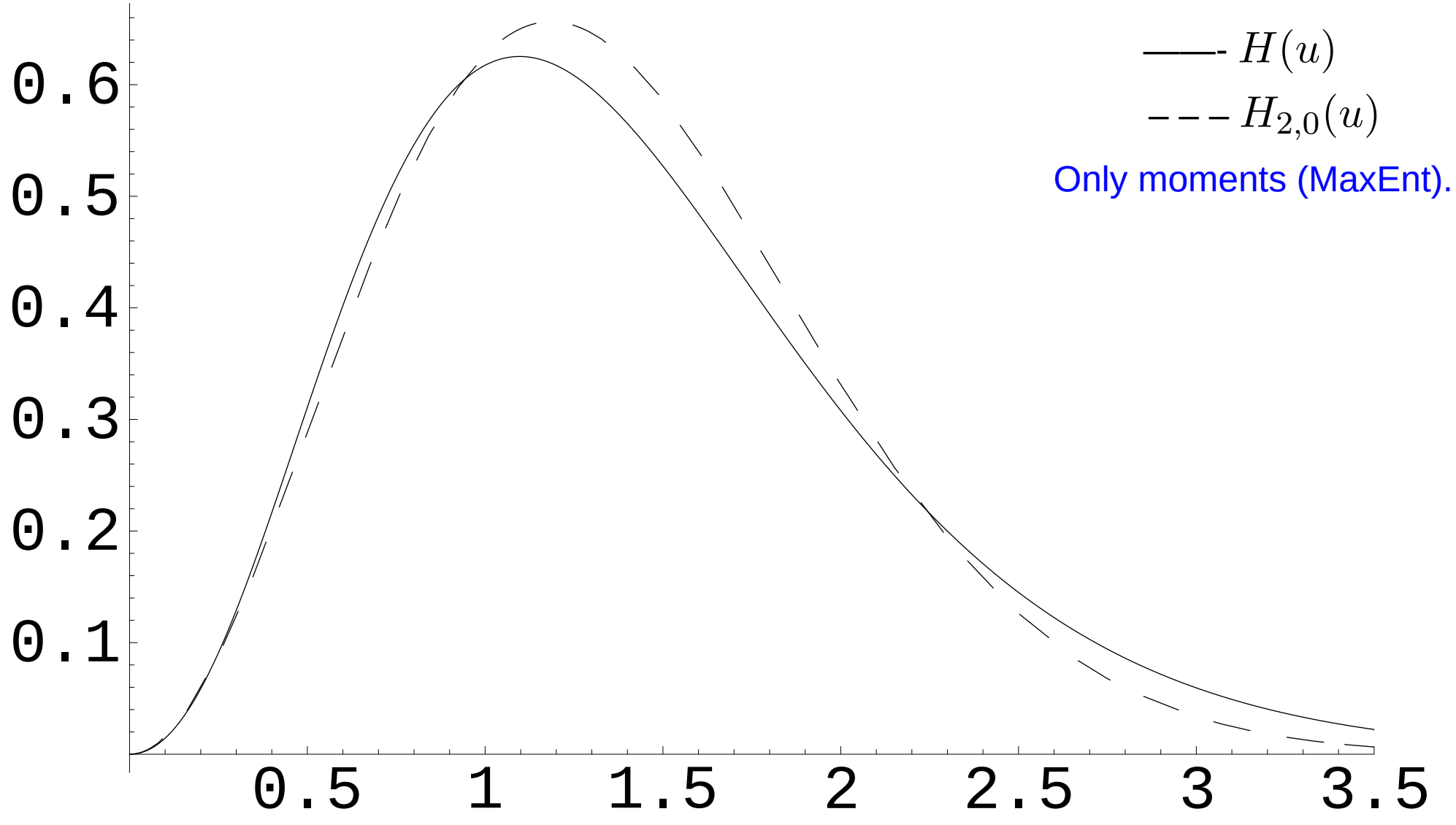
$H_{n,m}(u) := 4\pi\, u^2\, h_{n,m}(u)$ with

$$h_{n,m}(u) = \frac{\langle u^{-2}\rangle\, h_0(u)}{\mathcal{Z}(\mathbf{L})}\, \exp\left\{ -\sum_{i=1}^{n} \lambda_i u^i \; - 4\pi\, u \sum_{j=1}^{m} \lambda_{n+j}\, \frac{\sin(k_j u)}{k_j} \right\}$$

where $h_0(u)$ is the prior density ($h_0(u) = 1$ if no prior information is considered).

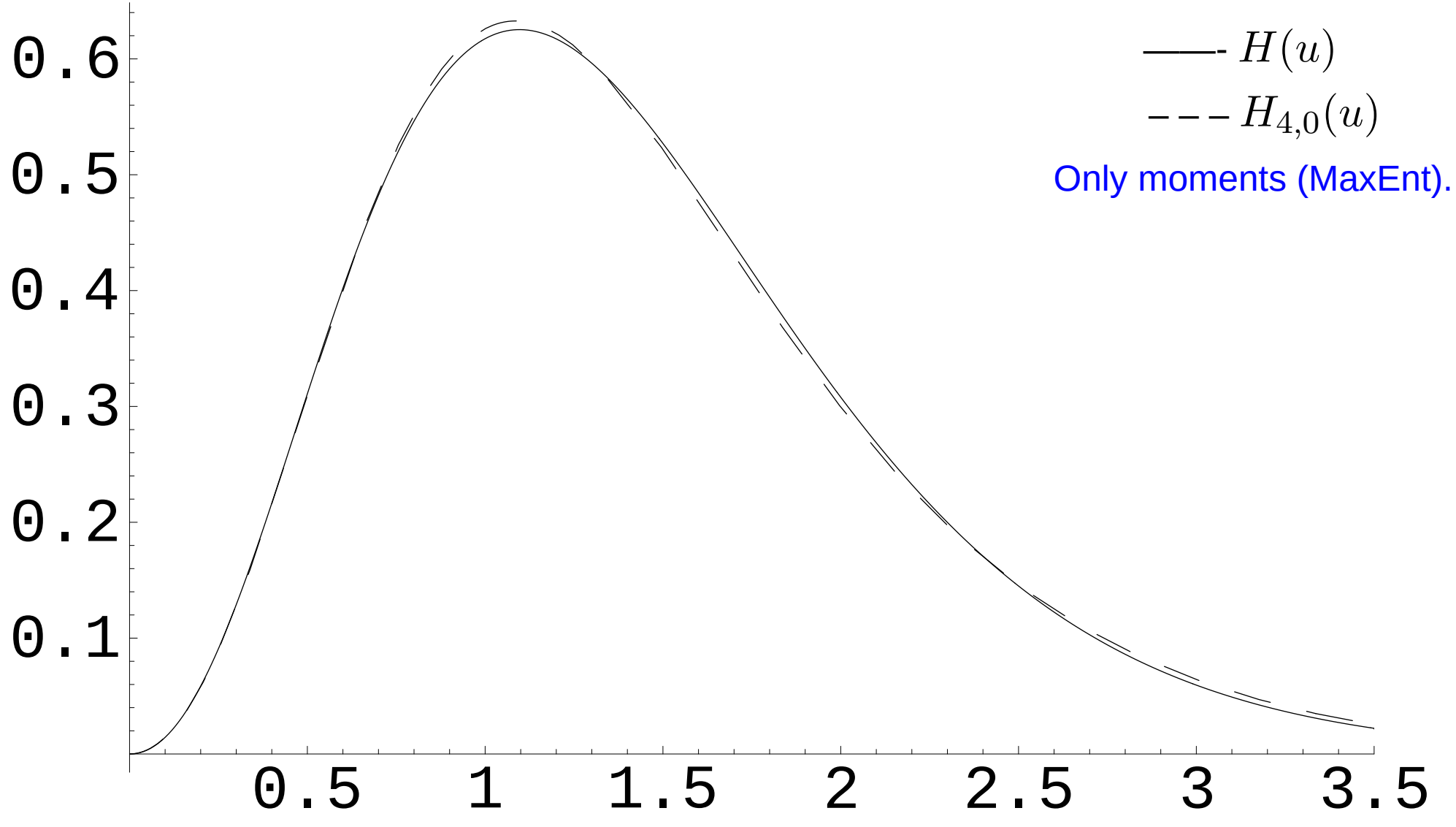
So, $H_{n,m}(x)$ is the MinxEnt (or MaxEnt) solution built up using n moments (plus the normalization) and m values of the Hankel transform.

Results on $H(u)$ for Helium atom



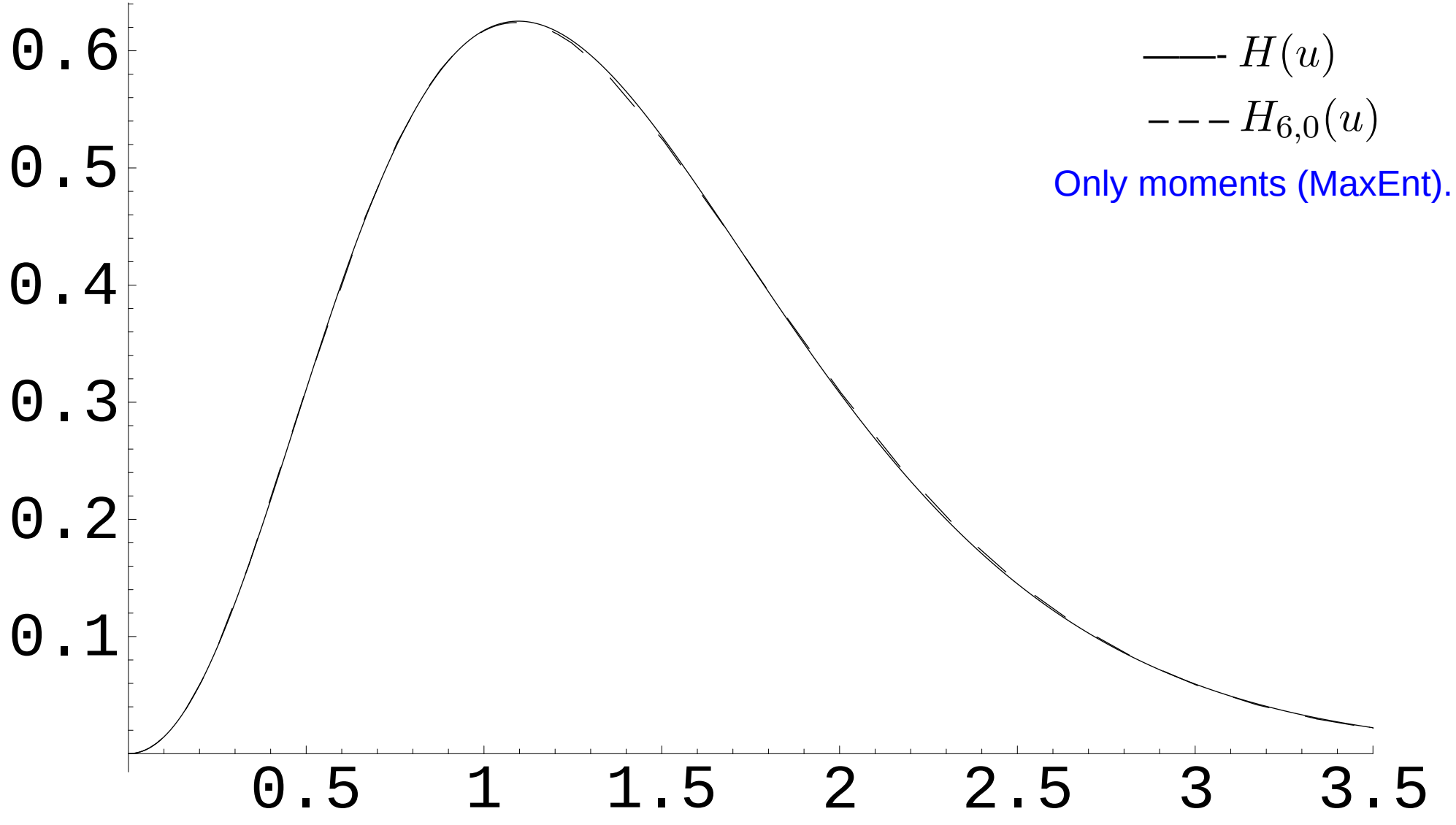
Only moments (MaxEnt).

Results on $H(u)$ for Helium atom



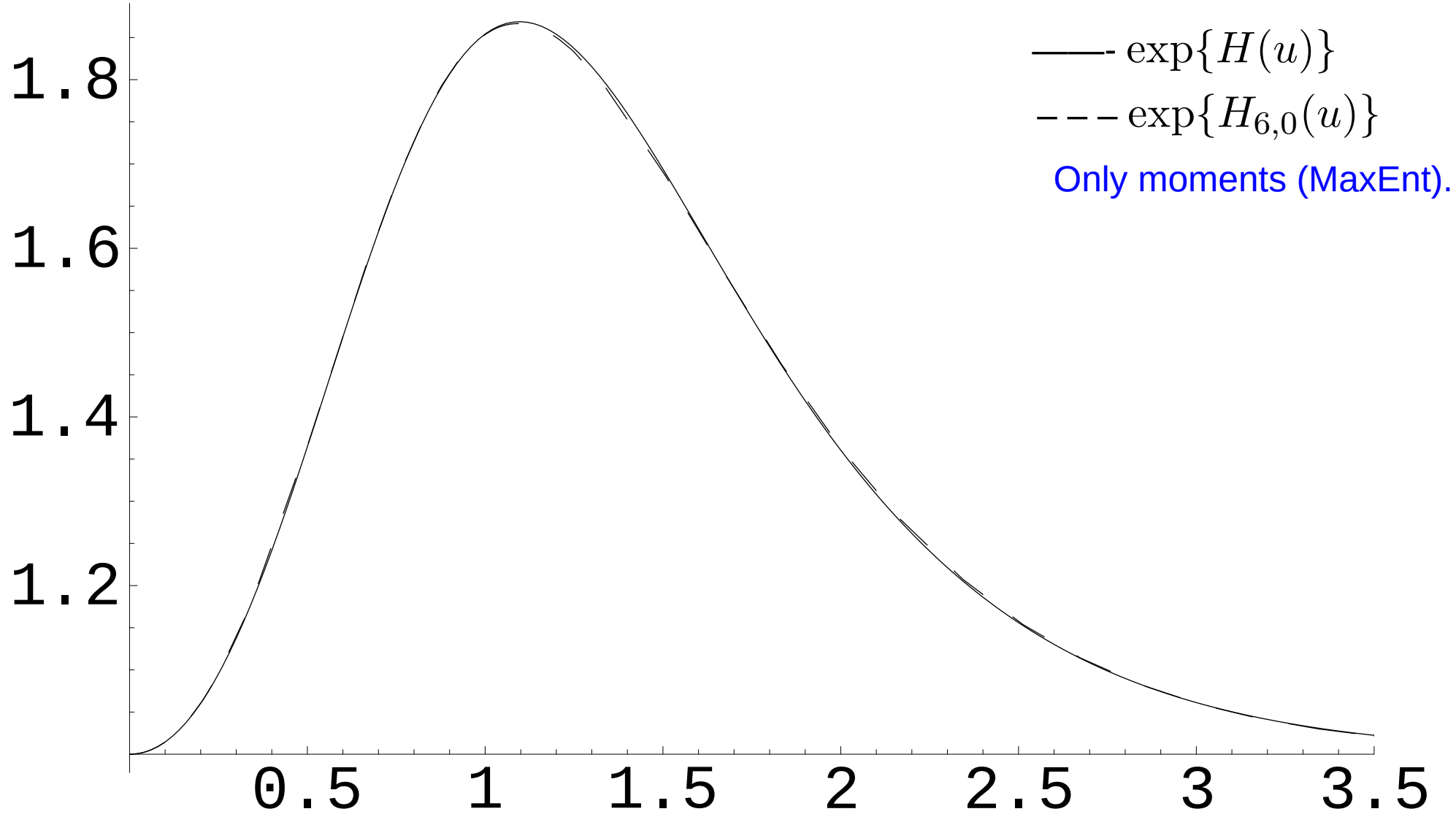
Only moments (MaxEnt).

Results on $H(u)$ for Helium atom



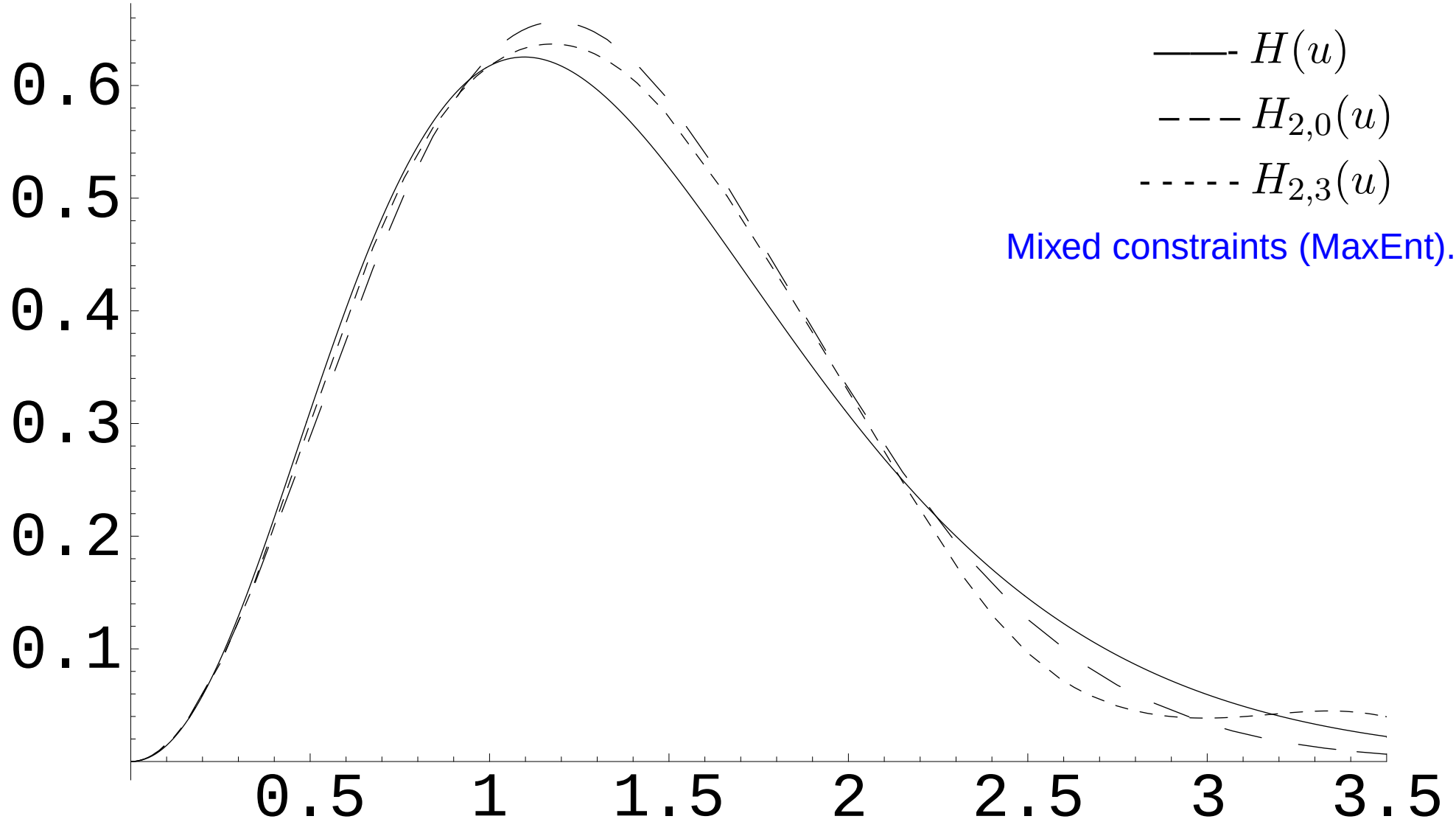
Only moments (MaxEnt).

Results on $H(u)$ for Helium atom



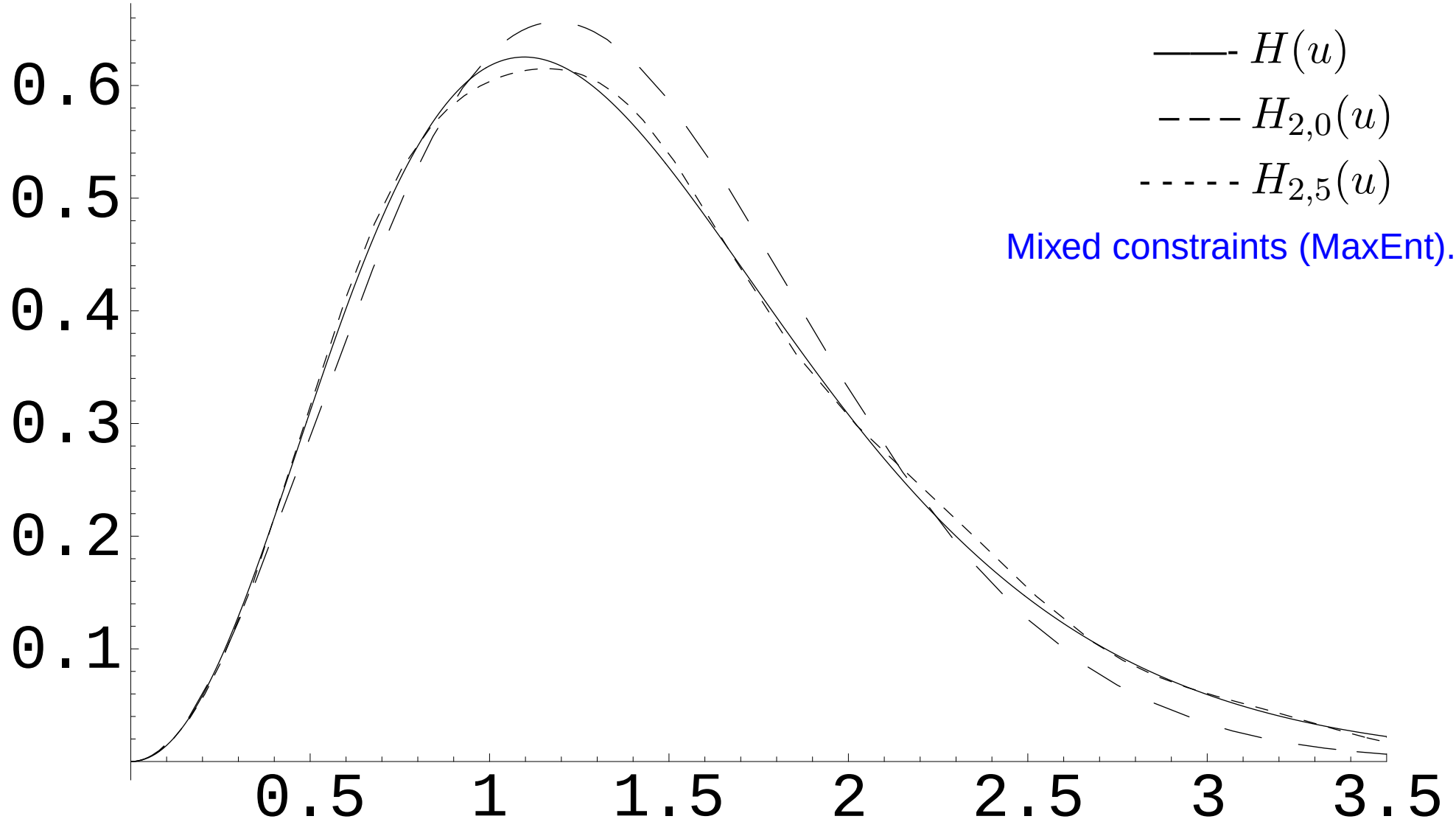
Only moments (MaxEnt).

Results on $H(u)$ for Helium atom



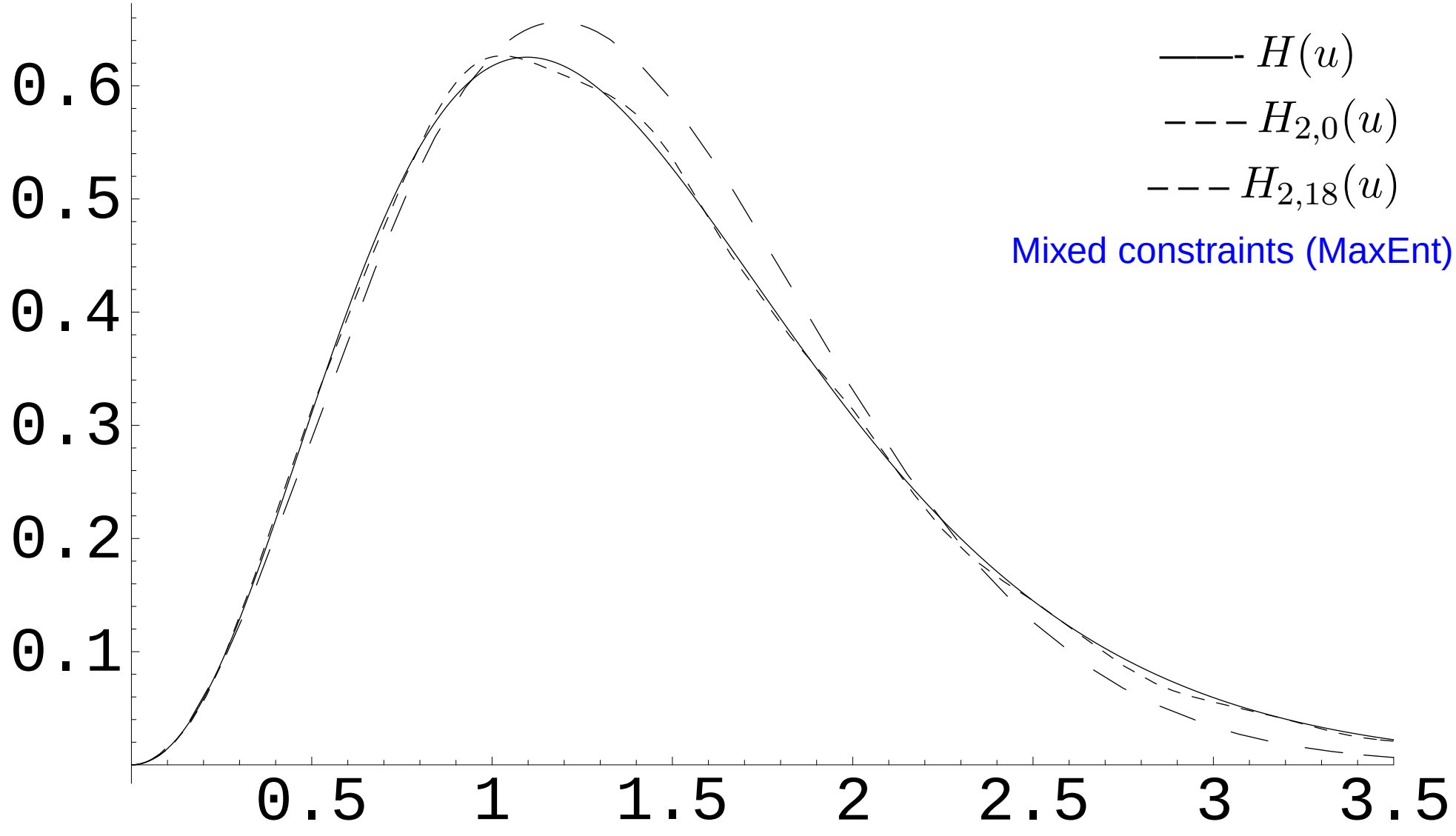
Mixed constraints (MaxEnt).

Results on $H(u)$ for Helium atom



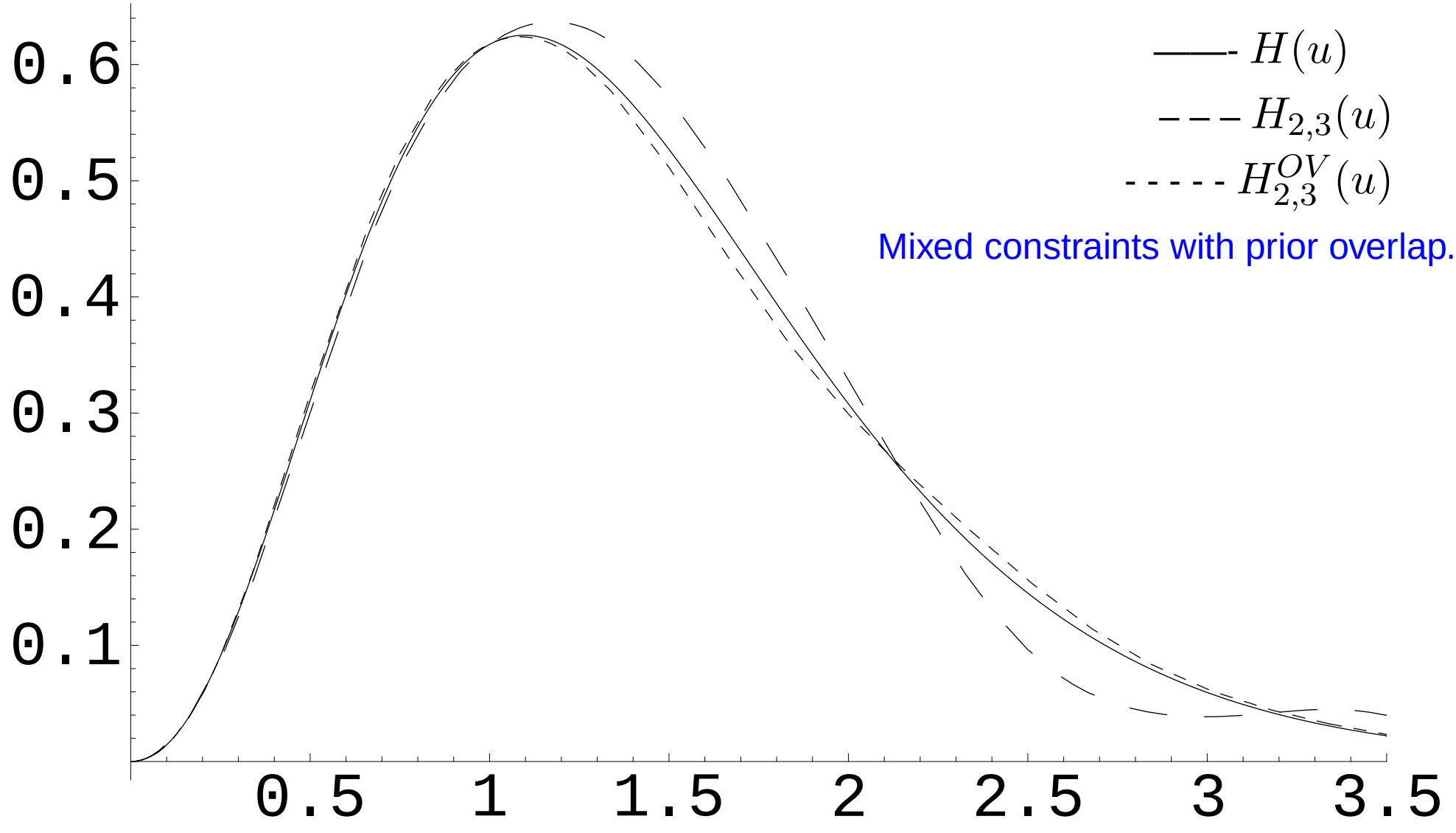
Mixed constraints (MaxEnt).

Results on $H(u)$ for Helium atom



Mixed constraints (MaxEnt).

Results on $H(u)$ for Helium atom



Mixed constraints with prior overlap.

Ex. 2: Spectrum of Jacobi matrices

A Jacobi matrix is a real, tridiagonal and symmetric matrix:

$$J_n := \begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & 0 \\ & b_2 & a_3 & \ddots & \\ & & \ddots & \ddots & b_{n-1} \\ 0 & & & b_{n-1} & a_n \end{pmatrix}$$

The characteristic polynomials $P_n(x) := det(xI_n - J_n)$ $(n = 1, 2, \ldots)$ satisfy a three-term recurrence relation:

$$P_{k+1}(x) = (x - a_{k+1})P_k(x) - b_k^2 P_{k-1}(x)\,, \quad k = 1, 2, \ldots, n-1\,,$$

with initial conditions $P_0(x) = 1$ and $P_1(x) = x - a_1$.

Ex. 2: Spectrum of Jacobi matrices

The spectrum of J_n is fully characterized by the zero distribution of $P_n(x)$ defined by

$$\rho_n(x) := \frac{1}{n}\sum_{j=1}^{n}\delta(x-x_j) \text{ with moments } \mu_r^{(n)} := \frac{1}{n}\sum_{j=1}^{n}x_j^r\,,$$

where $\delta(x-x_j)$ stands for the Dirac delta at the point x_j and $x_1 < \ldots < x_n$ are the real and simple zeros of $P_n(x)$.

The moments $\mu_r^{(n)}$ $(r = 0, 1, 2, \ldots)$ can be recurrently computed (Zarzo et al. 1988), so the MaxEnt method can be used to approximate $\rho_n(x)$.

To illustrate this we have chosen the well known Hermite polynomials, $H_n(x)$, because from the differential equation that they satisfy (Zarzo et al. 2002):

- All the zeros of $H_n(x)$ belongs to the interval $(-\sqrt{2n+1}, +\sqrt{2n+1})$.
- WKB approximation to the corresponding $\rho_n(x)$:

$$\rho_{wkb}^{(n)}(x) := \frac{2}{\pi}\,\frac{\sqrt{2n+1-x^2}}{2n+1}\,, \quad x \in (-\sqrt{2n+1}, +\sqrt{2n+1})\,.$$

Hermite Polynomial of degree 200.

- Moments of the zero distribution of $H_n(x)$: $\mu_{2j-1} = 0$ $(j = 1, 2, \ldots)$ and

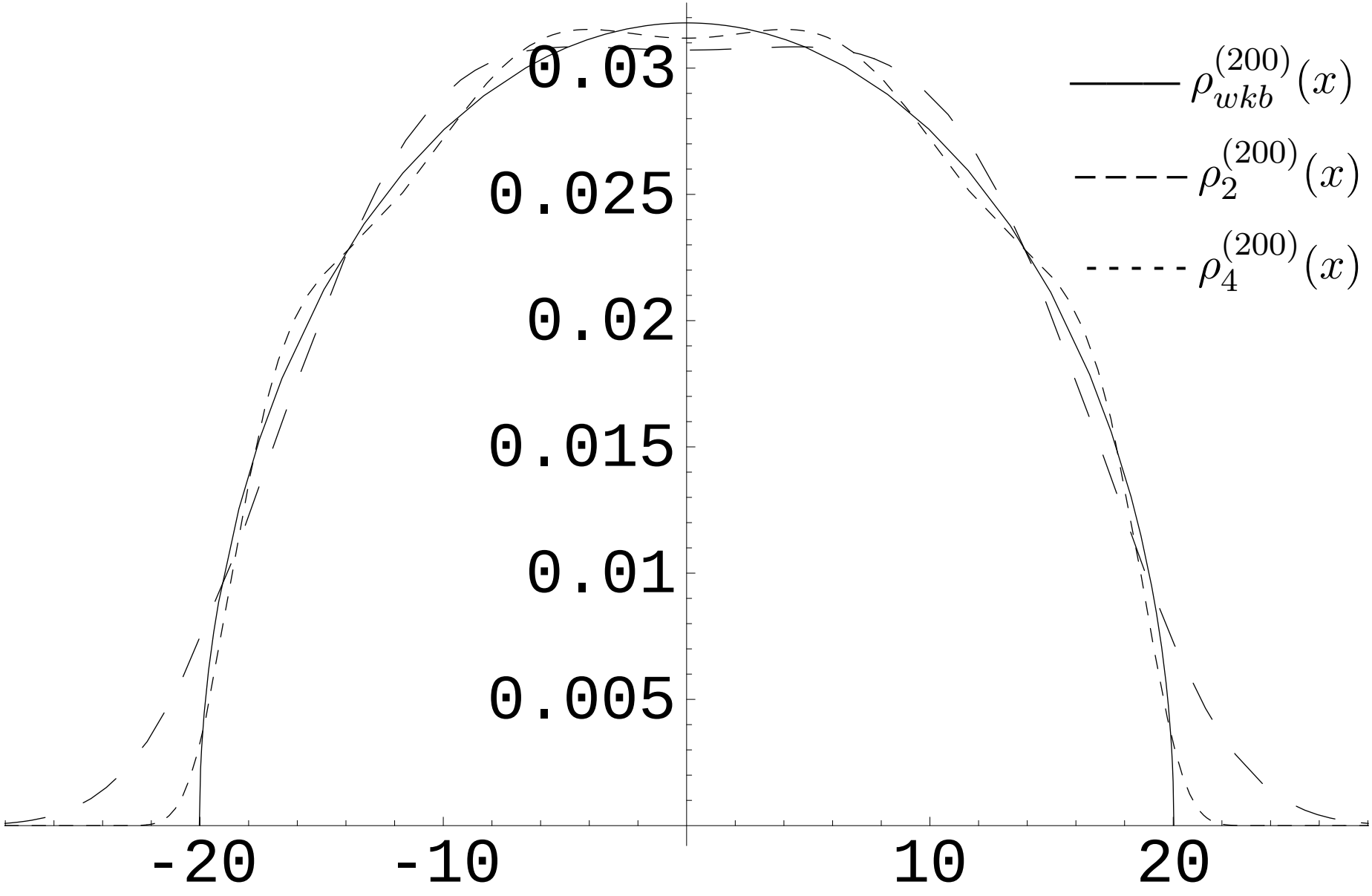
$$\begin{aligned}
\mu_0^{(n)} &= 1\,, \qquad \mu_2^{(n)} = \frac{n-1}{2} \\
\mu_4^{(n)} &= \frac{n^2}{2} - \frac{5n}{4} + \frac{3}{4}\,, \qquad \mu_6^{(n)} = \frac{5n^3}{8} - \frac{11n^2}{4} + 4n - \frac{15}{8} \\
\mu_8^{(n)} &= \frac{7n^4}{8} - \frac{93n^3}{16} + \frac{117n^2}{8} - \frac{65n}{4} + \frac{105}{16}\,, \ \ldots
\end{aligned}$$

- The MaxEnt solutions will be denoted by:

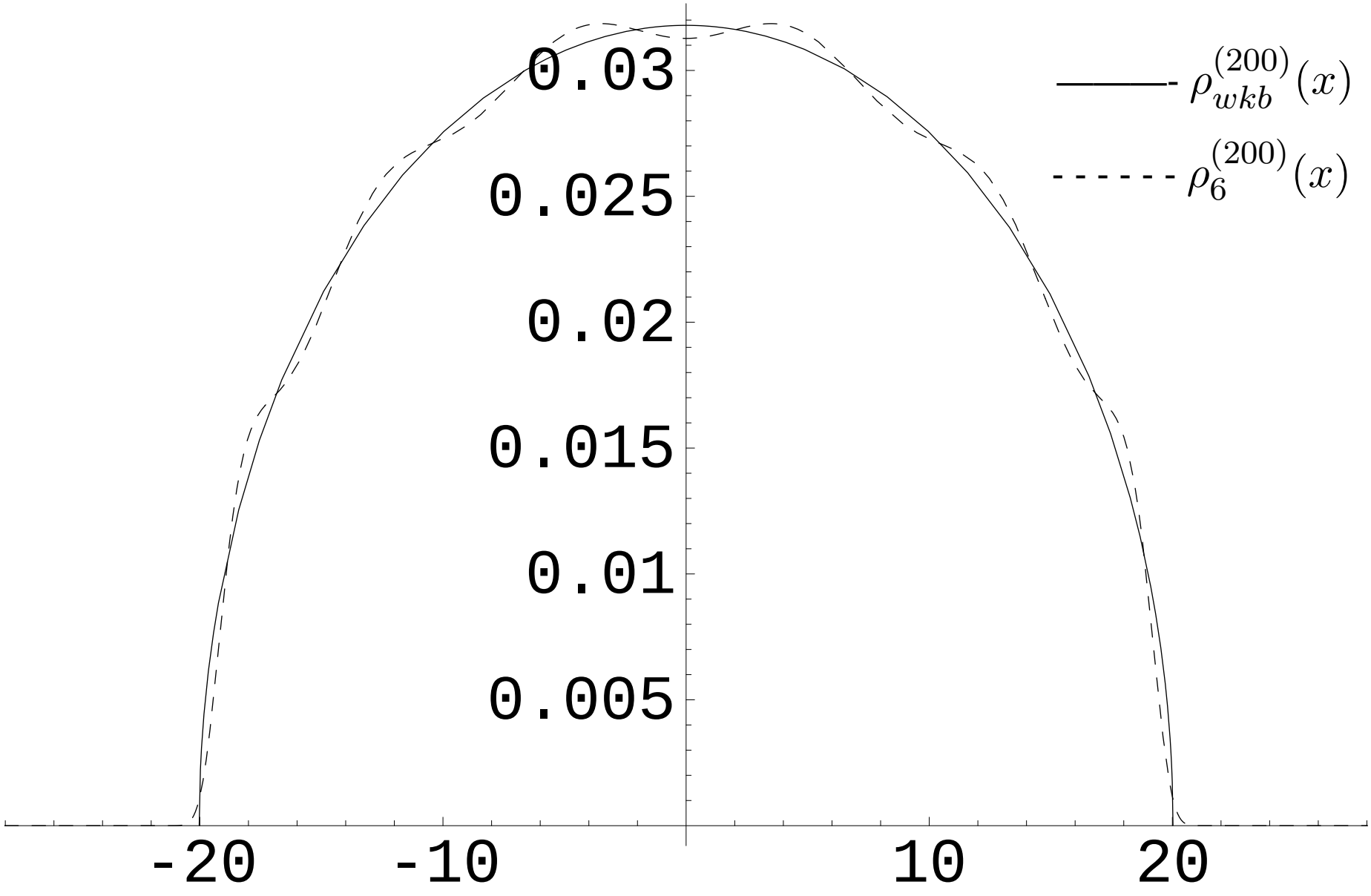
$$\rho_r^{(n)}(x) := \frac{1}{\mathcal{Z}(\mathbf{L})} \exp\left\{ -\sum_{i=1}^{r} \lambda_i \,, x^i \right\} .$$

where n is the degree of the polynomial and r is the number of moments used (excluding the normalization, which is always considered).

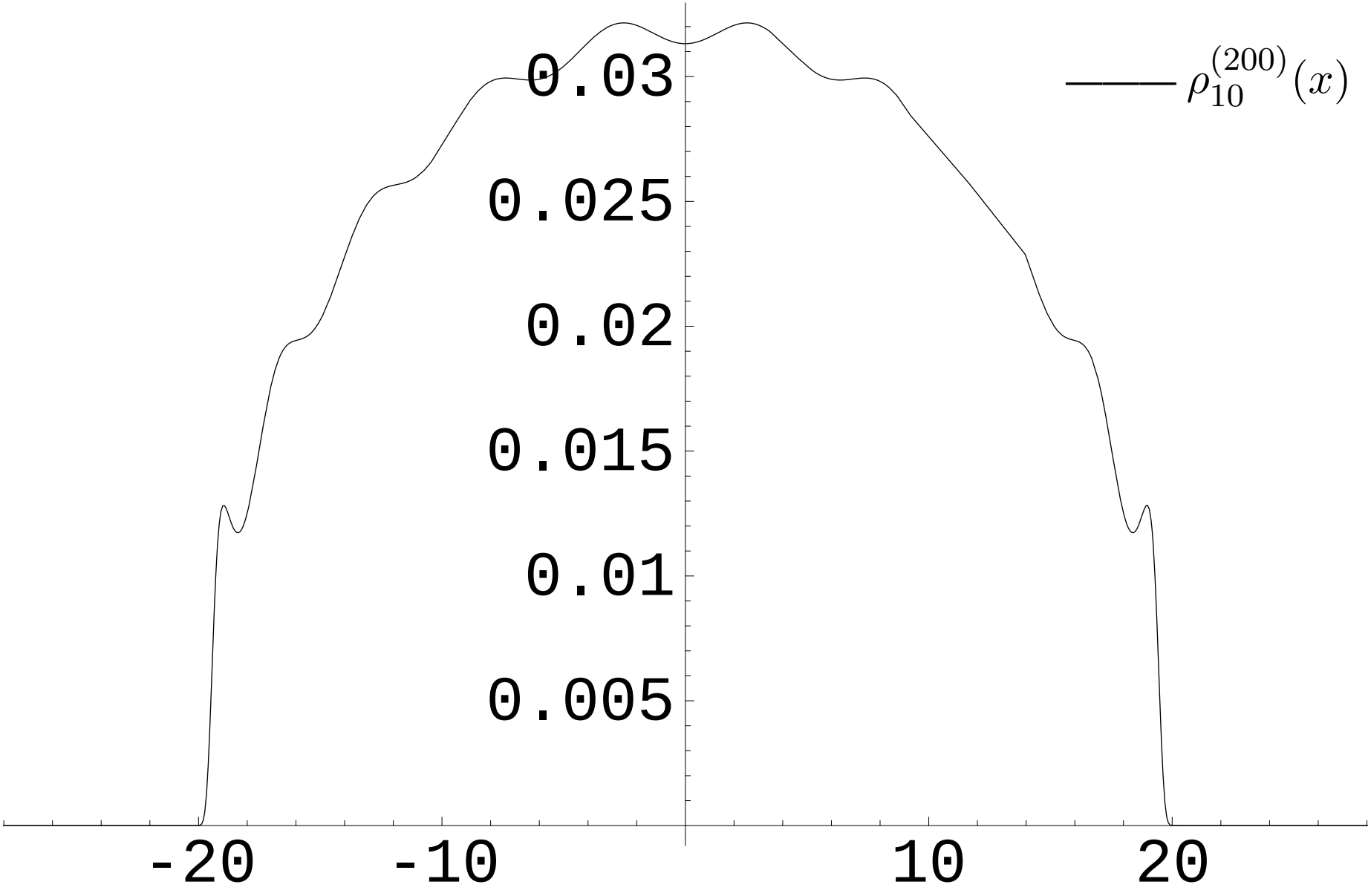
Hermite Polynomial of degree 200.



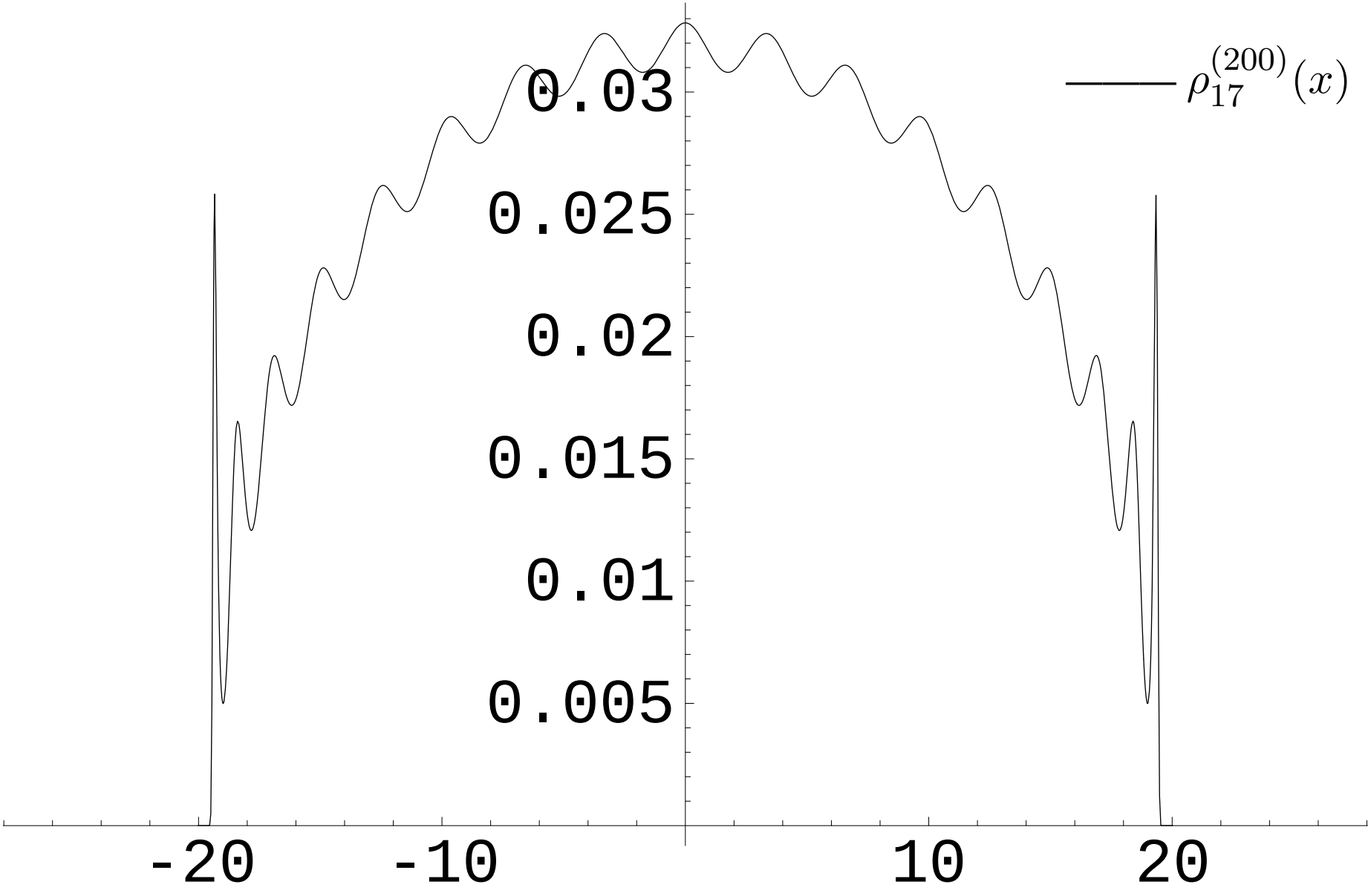
Hermite Polynomial of degree 200.



Hermite Polynomial of degree 200.



Hermite Polynomial of degree 200.



“Pseudo-MaxEnt" solutions

Hermite polynomial of degree 5: $H_5(x)$.

The ten first moments of the zero distribution of $H_5(x)$ fully characterize this distribution, in such a way that it is the unique one having those moments.

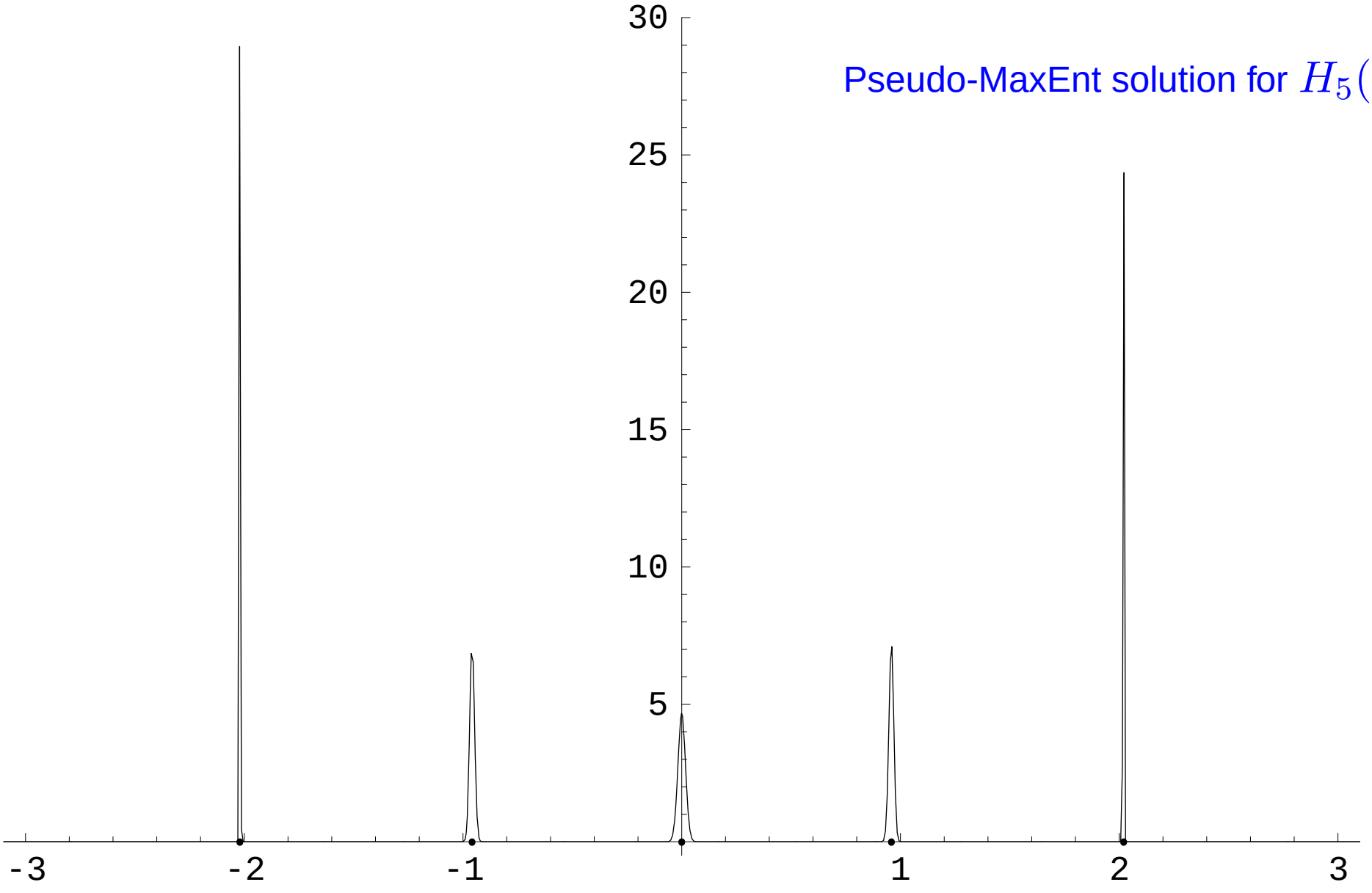
Hence, MaxEnt solution $\rho_{10}^{(5)}(x)$ does not exists

(in general, $\rho_{2n}^{(n)}$ neither do so)

and the algorithm gives no solution.

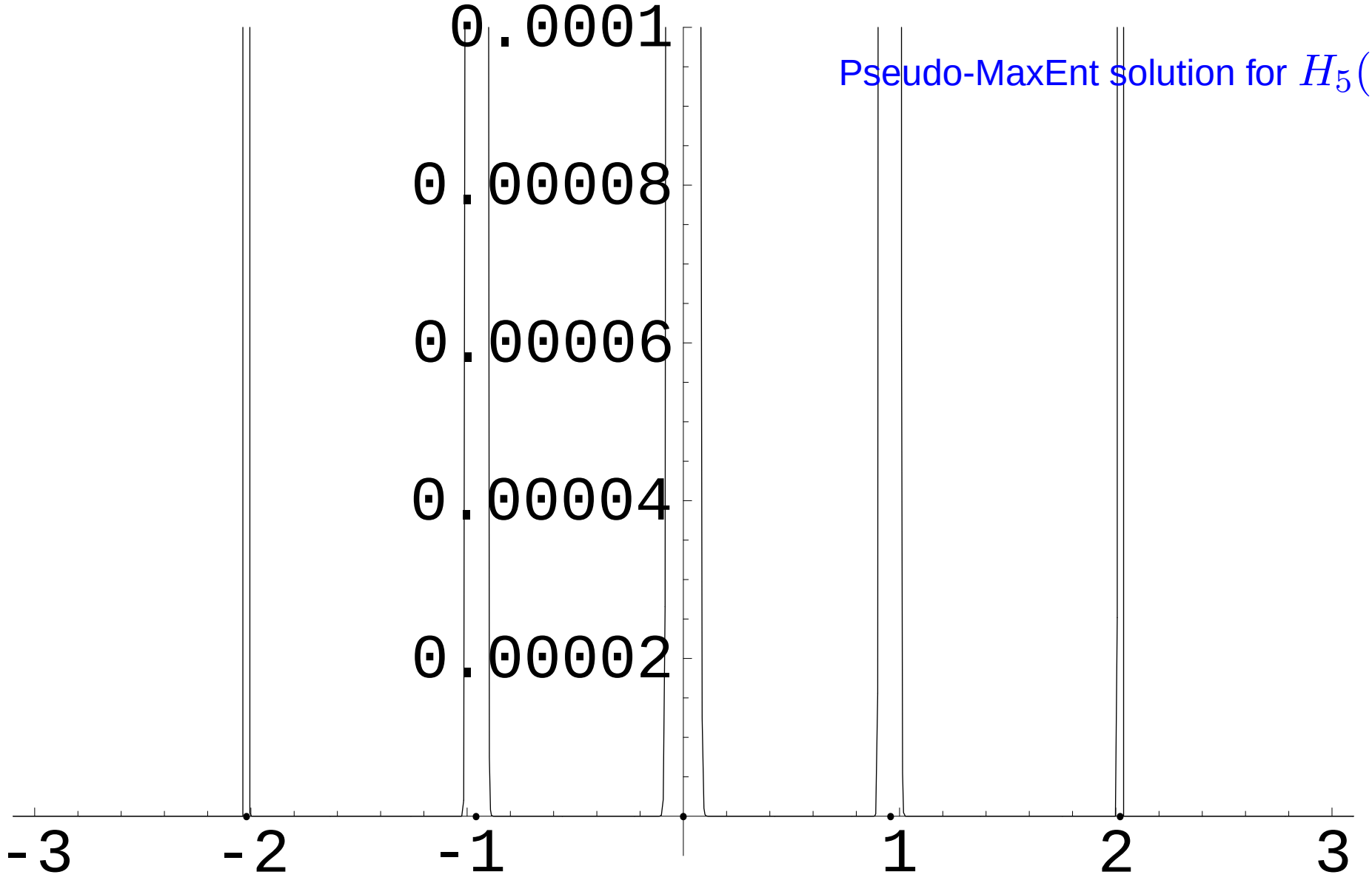
However, on running it, one can find several points in which the norm of the gradient is small ($\approx 10^{-6}$). We have called **“Pseudo-MaxEnt solutions”** to the solutions corresponding to such values of the gradient.

“Pseudo-MaxEnt" solutions



Pseudo-MaxEnt solution for $H_5(x)$.

“Pseudo-MaxEnt" solutions



Pseudo-MaxEnt solution for $H_5(x)$.


Why the MinxEnt method works so nicely ?
THANKS