MESSY Estimation: Maximum Entropy based Stochastic and Symbolic densitY Estimation

Mohsen Sadr, Tony Tohme, Kamal Youcef-Toumi, and Nicolas Hadjiconstantinou

Department of Mechanical Engineering, MIT, Cambridge, MA 02139, USA

Maximum entropy distribution function

Given a vector of N_m moments, $\boldsymbol{\mu}$, one can find the parent density $f_{\boldsymbol{X}}$ in a least bias sense, by minimizing the Shannon entropy functional

$$C[\mathcal{F}(\boldsymbol{x})] := \int \mathcal{F}(\boldsymbol{x}) \log(\mathcal{F}(\boldsymbol{x})) d\boldsymbol{x} + \sum_{i=1}^{N_m} \lambda_i \left(\int H_i(\boldsymbol{x}) \mathcal{F}(\boldsymbol{x}) d\boldsymbol{x} - \mu_i(\boldsymbol{x}) \right) .$$

The extremum of this functional gives the maximum entropy density function

$$\hat{f}(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\boldsymbol{\lambda} \cdot \boldsymbol{H}(\boldsymbol{x})\right), \qquad \text{where } \; Z = \int \exp\left(\boldsymbol{\lambda} \cdot \boldsymbol{H}(\boldsymbol{x})\right) d\boldsymbol{x}.$$

The Lagrange multipliers $\lambda_i, \;\; i = 1...N_m$, may be found using the unconstrained dual formulation $D(\boldsymbol{\lambda})$ with the gradient $\boldsymbol{g} = \nabla D(\boldsymbol{\lambda})$ and Hessian $\boldsymbol{H}(\boldsymbol{\lambda}) = \nabla^2 D(\boldsymbol{\lambda})$ leading to an iterative scheme

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - \boldsymbol{L}^{-1}(\boldsymbol{\lambda}) \boldsymbol{g}(\boldsymbol{\lambda}) ,$$
$$\text{where} \quad \boldsymbol{g} = \boldsymbol{\mu} - \frac{1}{Z} \int \boldsymbol{H} \exp\left(\boldsymbol{\lambda} \cdot \boldsymbol{H}\right) d\boldsymbol{x}$$
$$\text{and} \quad \boldsymbol{L} = -\frac{1}{Z} \int \boldsymbol{H} \otimes \boldsymbol{H} \exp\left(\boldsymbol{\lambda} \cdot \boldsymbol{H}\right) d\boldsymbol{x}.$$

Pros	Cons
✓ Least bias	✗ Ill-conditioned Hessian $\boldsymbol{L}$
✓ Convex optimization problem	✗ Requiring an accurate numerical integration method
✓ Matching moments	

Finding Lagrange multipliers via Gradient flow

Consider a Gradient flow that transitions from $f_{\boldsymbol{X}}$ to an ansatz $\hat{f}$

$$\frac{\partial f_{\boldsymbol{X}}}{\partial t} = \nabla_{\boldsymbol{x}} \left[\hat{f} \nabla_{\boldsymbol{x}} [f_{\boldsymbol{X}} / \hat{f}] \right]$$
$$= -\nabla_{\boldsymbol{x}} \cdot \left[\nabla_{\boldsymbol{x}} \left[\log(\hat{f}) \right] f_{\boldsymbol{X}} \right] + \nabla^2_{\boldsymbol{x}} \left[f_{\boldsymbol{X}} \right].$$

Using integration by parts, integrability of density and existence of its moments, we obtain an equation for the relaxation rate of moments as

$$\underbrace{\frac{d}{dt} \left[\int \boldsymbol{H} f_{\boldsymbol{X}} d\boldsymbol{x} \right]}_{\boldsymbol{g} :=} = \int \nabla_{\boldsymbol{x}} [\boldsymbol{H}] \cdot \nabla_{\boldsymbol{x}} [\log(\hat{f})] f_{\boldsymbol{X}} d\boldsymbol{x} + \int \nabla^2_{\boldsymbol{x}} [\boldsymbol{H}] f_{\boldsymbol{X}} d\boldsymbol{x} .$$

By substituting maximum entropy ansatz, we obtain the relaxation rates (or gradient) using the samples

$$\boldsymbol{g} = \underbrace{\sum_i \left\langle \nabla_{x_i} [\boldsymbol{H}(\boldsymbol{X}(t))] \otimes \nabla_{x_i} [\boldsymbol{H}(\boldsymbol{X}(t))] \right\rangle}_{\boldsymbol{L}^{\mathrm{ME}} :=} \boldsymbol{\lambda} + \sum_i \left\langle \nabla^2_{x_i} [\boldsymbol{H}(\boldsymbol{X}(t))] \right\rangle .$$

At the steady-state, $f_{\boldsymbol{X}} \to \hat{f}$, leading to $\boldsymbol{g} \to \boldsymbol{0}$. Lagrange multipliers can be computed directly as

$$\boldsymbol{\lambda} = -\left(\boldsymbol{L}^{\mathrm{ME}}\right)^{-1} \left(\sum_i \left\langle \nabla^2_{x_i} [\boldsymbol{H}(\boldsymbol{X}(t))] \right\rangle \right) .$$

Pros	Cons
✓ Least bias	✗ Ill-conditioned matrix $\boldsymbol{L}^{\mathrm{ME}}$
✓ No optimization problem	—
✓ Matching moments	—
✓ Integrating using samples	—

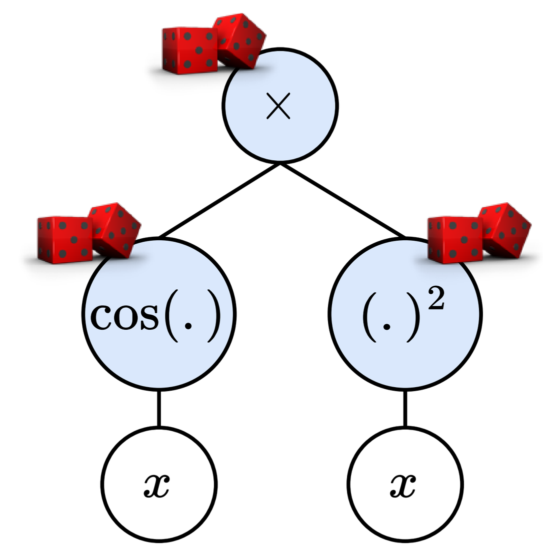
Examples of symbolic expressions for bi-modal problem

Method	Obtained density
MESSY-P	$\hat{f}(x) = 0.288 e^{-0.017x^{10}+0.106x^9-0.084x^8-0.659x^7+1.209x^6+1.179x^5-3.722x^4+0.075x^3+2.693x^2-0.612x}$
MESSY-S	$\hat{f}(x) = 0.993 e^{-1.85x^2-1.162x\cos(1.5x)+0.232x-0.652\cos(x)-0.424\cos(2x)-0.591\cos(3.5x)+0.47\cos(\cos(3.5x))}$

Table 1: Example of expressions obtained for the bi-modal problem using MESSY with polynomial (MESSY-P) and randomly created basis functions (MESSY-S).

Symbolic exploration for an optimal basis function

We perform a Monte Carlo and symbolic search in the space of smooth functions constructed using an expression tree to find a vector of basis functions $\boldsymbol{H}$ that guarantee small $\mathrm{cond}(\boldsymbol{L}^{\mathrm{ME}})$. Here, we also impose the necessary condition that the basis function with the highest growth rate is even.



Example : $x^2 \times \cos(x)$

Pros	Cons
✓ Least bias	✗ Additional cost of symbolic acc./rej.
✓ No optimization problem	—
✓ Matching moments	—
✓ Integration using samples	—
✓ Well-conditioned matrix $\boldsymbol{L}^{\mathrm{ME}}$	—

Results

We compare MESSY estimate using polynomials (MESSY-P) and randomly created basis functions (MESSY-S) to kernel density estimation and the maximum cross-entropy distribution function with Gaussian as the prior. As the test case, here we consider bi-modal distributions that are far from Gaussian.

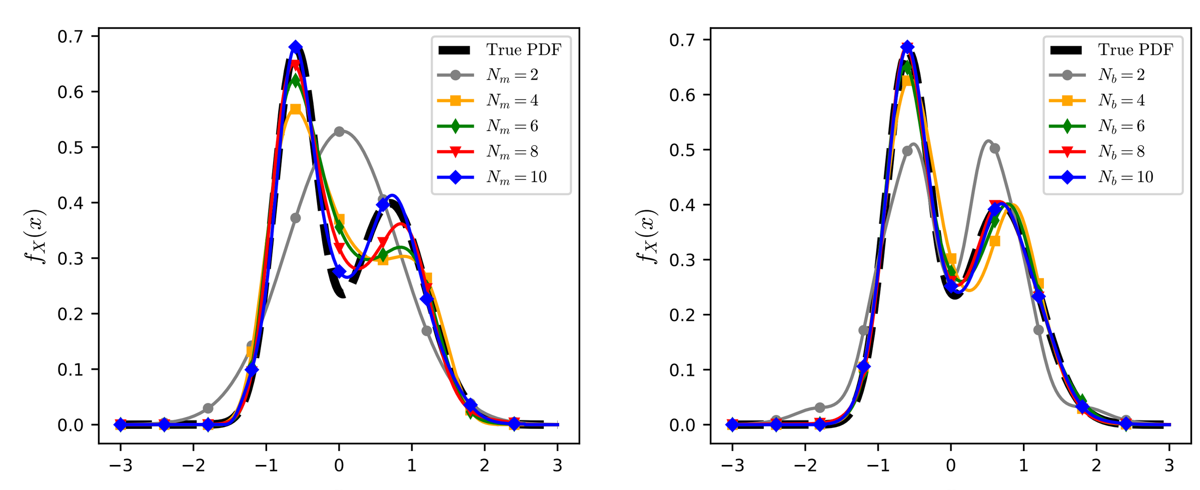


Figure 1: Convergence of MESSY estimation to target distribution function by (left) increasing the order of polynomial basis functions N_m or (right) increasing the number of random basis functions N_b with highest order $\mathcal{O}(x^2)$.

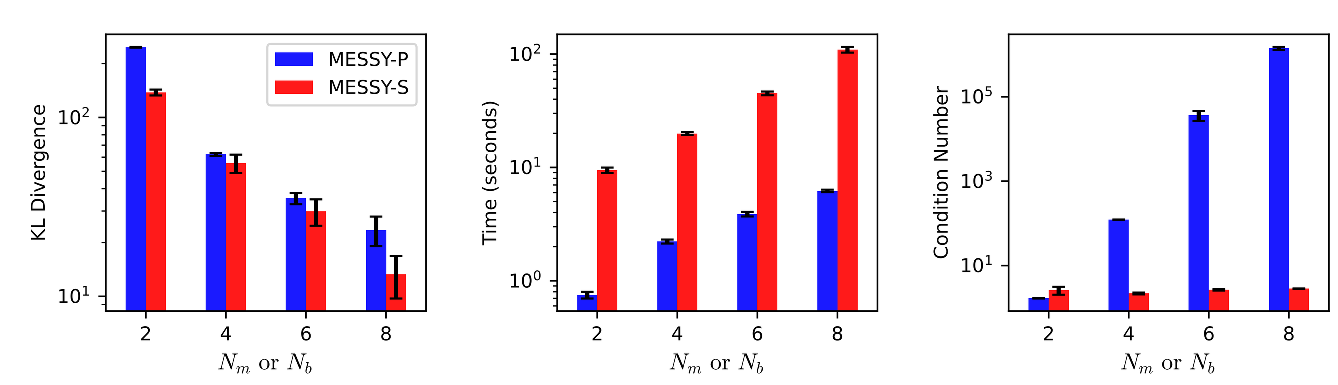


Figure 2: KL Divergence (left), execution time (middle) and condition number (right) against the degrees of freedom, i.e. the number of moments N_m for MESSY-P and the number of basis functions N_b with highest order $\mathcal{O}(x^2)$ for MESSY-S.

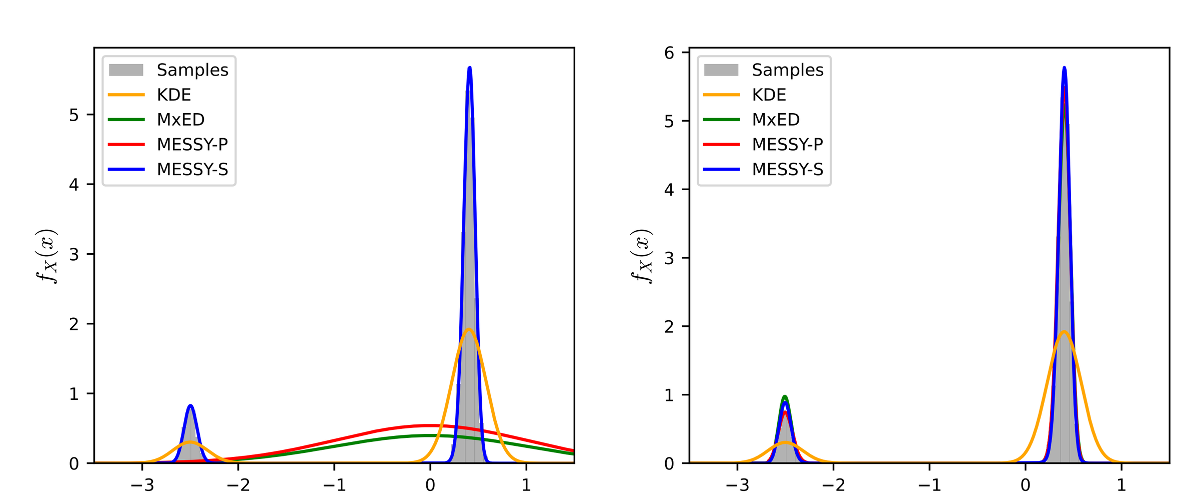


Figure 3: Estimating density for border case from samples using KDE, MxED, MESSY-P, and MESSY-S using basis functions with a growth rate of leading term up to $\mathcal{O}(x^2)$ (left) and $\mathcal{O}(x^4)$ (right).

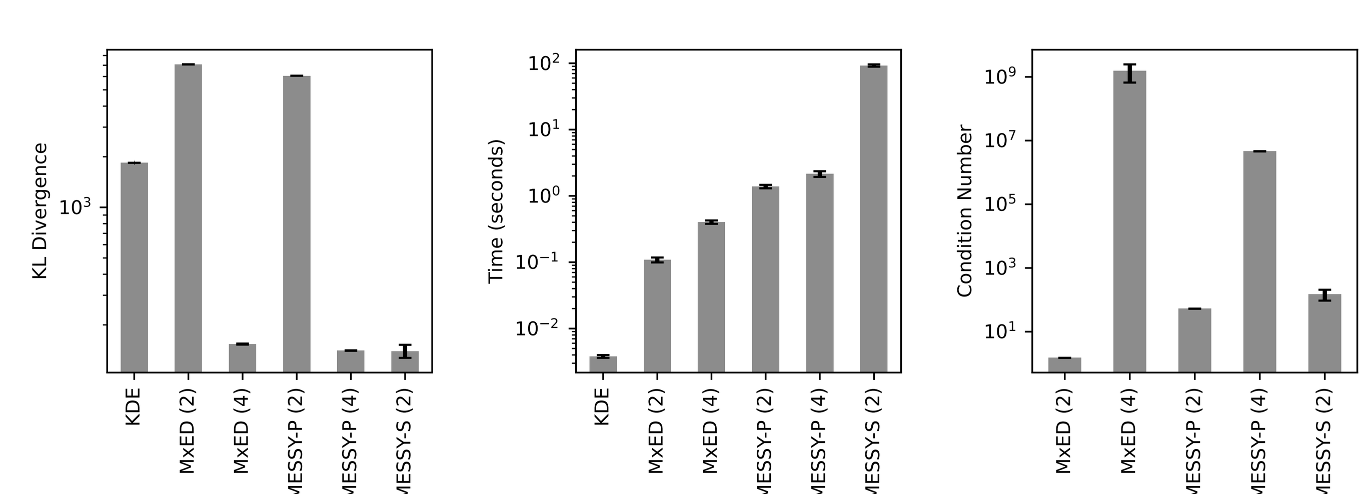


Figure 4: Comparing KL Divergence (left), execution time (middle), and condition number (right) for KDE, MxED, MESSY-P, and MESSY-S estimate of density in the limit of the realizability. Here, we consider matching moments up to $N_m = 2, 4$ for MxED and MESSY-P denoted by MxED (2), MxED (4), ..., while matching only up to $N_m = 2$ for MESSY-S.

Deutsche Forschungsgemeinschaft

Massachusetts Institute of Technology