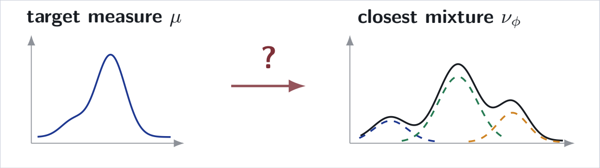
Mixtures Closest to a Given Measure:
A Semidefinite Programming Approach

Srećko Ðurašinović   Jean B. Lasserre   Victor Magron

July 7th, 2026, Seoul, South Korea

ICML International Conference On Machine Learning
NANYANG TECHNOLOGICAL UNIVERSITY
cnrs @CREATE Singapore
LAAS CNRS

Main Question



Question. Among all admissible mixtures from a parametric family, which one is **closest** to a target measure μ?

Distance:
2-Wasserstein (W_2) or Total Variation (TV)

The target μ is usually known from samples μ^N

$\rightsquigarrow$ **Parametric family**

$$\{\mu_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in S_{\boldsymbol{\theta}}}$$

$\rightsquigarrow$ **Mixing measure**

$$\phi \in \mathcal{P}(S_{\boldsymbol{\theta}})$$

$\rightsquigarrow$ **Induced mixture**

$$\nu_\phi(B) = \int_{S_{\boldsymbol{\theta}}} \mu_{\boldsymbol{\theta}}(B)\, d\phi(\boldsymbol{\theta}),\ B \in \mathcal{B}(\mathbb{R}^n)$$

Formal Framework

$$\tau^{\text{dist}}_{\varepsilon,R} = \inf_{\nu_\phi,\phi} \left\{ \text{dist}(\mu,\nu_\phi) + \varepsilon \int_{S_{\boldsymbol{\theta}}} R(\boldsymbol{\theta})\, d\phi(\boldsymbol{\theta}) \right\}, \quad \text{dist} \in \{W_2, \text{TV}\}$$

Regularization parameters

$$\varepsilon > 0, \quad R \in \mathbb{R}[\boldsymbol{\theta}]$$

Parameter space

$$S_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \ r_j(\boldsymbol{\theta}) \geq 0,\ j \in \{1, \ldots, \ell\}\}$$

$\rightsquigarrow$ Example: Multivariate Gaussian components

$$S_{\boldsymbol{m},\boldsymbol{\Sigma}} = \{(\boldsymbol{m}, \boldsymbol{\Sigma}) \mid b\boldsymbol{I} \succeq \boldsymbol{\Sigma} \succeq a\boldsymbol{I},\ \overline{\boldsymbol{m}} \geq \boldsymbol{m} \geq \underline{\boldsymbol{m}}\}$$

What is NOT FIXED beforehand:

- **mixture order K**
- **component means/variances**
- **mixing measure ϕ need not be finitely supported**
- **the family $\{\mu_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ need not be Gaussian**

Parametric Families with (overlooked) Crucial Property

For every moment of order $\boldsymbol{\alpha}$, there exists $p_{\boldsymbol{\alpha}} \in \mathbb{R}[\boldsymbol{\theta}]$ s.t.:

$$\boxed{\int_{\mathbb{R}^n} \boldsymbol{x}^{\boldsymbol{\alpha}} \, d\mu_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\alpha}}(\boldsymbol{\theta})}$$

Given a sequence $\boldsymbol{\phi} := (\phi_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}}$, define the Riesz linear functional $L_{\boldsymbol{\phi}} : \mathbb{R}[\boldsymbol{\theta}] \to \mathbb{R}$:

$$L_{\boldsymbol{\phi}} : p \mapsto \sum_{\boldsymbol{\gamma}} p_{\boldsymbol{\gamma}} \phi_{\boldsymbol{\gamma}}$$

$\rightsquigarrow$ Then, as a major consequence,

$$\boxed{\int_{\mathbb{R}^n} \boldsymbol{x}^{\boldsymbol{\alpha}} \, d\nu_{\phi}(\boldsymbol{x}) = \int_{S_{\boldsymbol{\theta}}} p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \, d\phi(\boldsymbol{\theta}) = L_{\boldsymbol{\phi}}(p_{\boldsymbol{\alpha}})}$$

Gaussian $(n = 1)$

For $\boldsymbol{\theta} = (m, \sigma)$,

$$p_3(m, \sigma) = m^3 + 3m\sigma^2$$
$$L_{\boldsymbol{\phi}}(p_3) = \phi_{(3,0)} + 3\phi_{(1,2)}$$

Poisson $(n = 1)$

For $\boldsymbol{\theta} = \lambda$,

$$p_3(\lambda) = \lambda^3 + 3\lambda^2 + \lambda$$
$$L_{\boldsymbol{\phi}}(p_3) = \phi_3 + 3\phi_2 + \phi_1$$

The unknown mixture moments are linear in the unknown moments ϕ

Moment-SOS Pipeline

Generic Hierarchy of SDP relaxations

$$\forall d \in \mathbb{N}, \quad \tau_{d,\varepsilon,R}^{\mathrm{dist}} = \min_{\mathbf{y}_d^{\mathrm{dist}}} \left\langle \mathbf{c}_{d,\varepsilon,R}^{\mathrm{dist}}, \mathbf{y}_d^{\mathrm{dist}} \right\rangle \quad \text{s.t.} \quad \mathbf{A}_d^{\mathrm{dist}} \mathbf{y}_d^{\mathrm{dist}} = \mathbf{b}_d^{\mathrm{dist}}, \qquad \mathbf{M}_d^{\mathrm{dist}}\left(\mathbf{y}_d^{\mathrm{dist}}\right) \succeq \mathbf{0}$$

Affine constraints
Moment matching

Moment positivity
Sequences come from measures

Support localization
Enforce $\phi \in \mathcal{P}(S_{\boldsymbol{\theta}})$

$\rightsquigarrow$ **By construction,**

$$\tau_{d,\varepsilon,R}^{\mathrm{dist}} \leq \tau_{d+1,\varepsilon,R}^{\mathrm{dist}} \leq \cdots \leq \tau_{\varepsilon,R}^{\mathrm{dist}}$$

Distance-Specific Routes

2-Wasserstein (W_2)

Decision vector: $\boldsymbol{y}_d^{W_2} = (\boldsymbol{\lambda}, \boldsymbol{\phi})$

$\rightsquigarrow$ $\boldsymbol{\lambda}$ is a **coupling** between the target and the induced mixture, as in classical optimal transport

$\downarrow$

$$\forall \boldsymbol{\alpha}: \quad \lambda_{\boldsymbol{\alpha},\mathbf{0}} = \mu_{\boldsymbol{\alpha}}, \qquad \lambda_{\mathbf{0},\boldsymbol{\alpha}} = L_{\boldsymbol{\phi}}(p_{\boldsymbol{\alpha}})$$

Total Variation (TV)

Decision vector: $\boldsymbol{y}_d^{\mathrm{TV}} = (\boldsymbol{\psi}_+, \boldsymbol{\psi}_-, \boldsymbol{\phi})$

$\rightsquigarrow$ the signed measure $\mu - \nu_\phi$ admits a unique **Hahn–Jordan decomposition** (ψ_+, ψ_-) satisfying $\|\mu - \nu_\phi\|_{\mathrm{TV}} = \psi_+(\mathbb{R}^n) + \psi_-(\mathbb{R}^n)$

$\downarrow$

$$\forall \boldsymbol{\alpha}: \quad \psi_{+,\boldsymbol{\alpha}} - \psi_{-,\boldsymbol{\alpha}} = \mu_{\boldsymbol{\alpha}} - L_{\boldsymbol{\phi}}(p_{\boldsymbol{\alpha}})$$

The distance choice **only changes** the auxiliary variables and linear equations

The moment-SOS machinery **stays the same**

Main Results

Globally convergent hierarchy

$$\tau^{\mathrm{dist}}_{d,\varepsilon,R} \nearrow \tau^{\mathrm{dist}}_{\varepsilon,R} \qquad (d \to +\infty)$$

- each SDP is a convex relaxation
- **fixed** d: guaranteed lower bound
- **increasing** d: accuracy vs complexity trade-off

Finite recovery under flatness 🏆

$$\operatorname{rank} \mathbf{M}_d(\phi^{*(d)}) = \operatorname{rank} \mathbf{M}_{d-d_{\min}}(\phi^{*(d)}) = \widehat{K}$$

- the relaxation is exact at order d
- $\widehat{K}$ is the recovered mixture order
- $\widehat{K}$-atomic ϕ^* gives component parameters $\left\{\widehat{\boldsymbol{\theta}}_i\right\}_{i=1,\dots,\widehat{K}}$ and associated weights

Moment-SOS (a.k.a. Lasserre's) hierarchy $\rightsquigarrow$ Global optimality certificate

In Practice: SDP-Based Initialization Algorithm

Input empirical moments $\mu_{\boldsymbol{\alpha}}^{N}$, for $|\boldsymbol{\alpha}| \leq 2d$

1 solve the W_2 or TV moment-SOS relaxation at order d

2 test **flatness** of the optimal pseudo-moment matrix $\mathbf{M}_d(\boldsymbol{\phi}^{*(d)})$

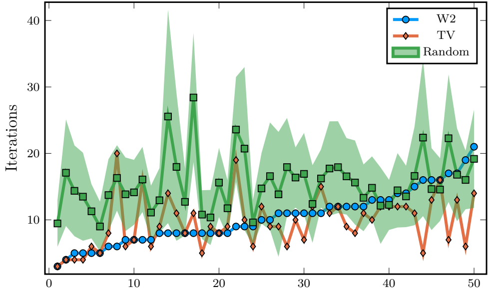
- if **exact**: complete cluster configuration is obtained
- if **approximate**: initialize k-means or Expectation-Maximization (EM)

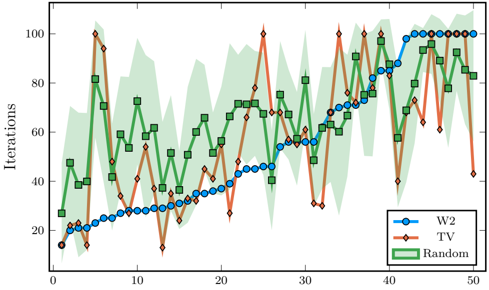
3 else: increase the relaxation order d and repeat

$\rightsquigarrow$ When flatness is approximate:

$$\left\{\widehat{\boldsymbol{\theta}}_i\right\}_{i=1}^{\widehat{K}} \implies \text{initial mixture estimates} \implies \text{globally informed initialization}$$

Synthetic Mixtures





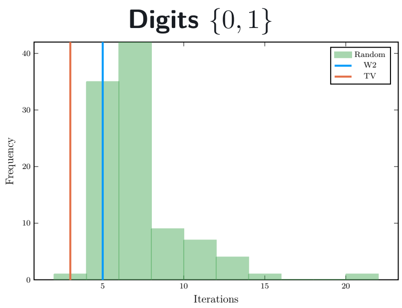
$n = 2$, $K = 5$, $N = 1000$, 50 non-spherical mixtures; $d = 4$

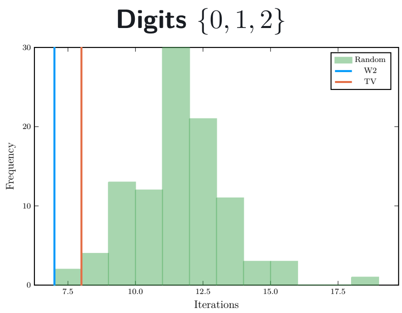
Average reduction relative to random starts:

31.8%	**36.5%**	**18.6%**	**13.9%**
W_2 on k-means	TV on k-means	W_2 on EM	TV on EM

The global layer supplies better structure before local refinement

MNIST Data





$n = 784 \to 2$ PCA components $\to$ Gaussian family; $d = 4$

Lowered number of iterations with the same k-means objective

Comparison

Classical route
decisions are fixed early

- choose number of components K
- ↓
- choose initial parameters
- ↓
- run EM or k-means
- ↓
- return **one local** mixture $\rightsquigarrow$ **sensitive to initialization**

This paper's route
global information first

- solve a moment-SOS relaxation
- ↓
- **exact** flatness: $\rightsquigarrow$ extract atoms or global seeds
- ↓
- **approximate** flatness: initialize EM or k-means $\rightsquigarrow$ **globally informed refinement**

Conclusion: Moment-SOS Hierarchy for Machine Learning

① Given $\mathrm{dist} \in \{W_2, \mathrm{TV}\}$ and $\{\mu_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in S_{\boldsymbol{\theta}}}$, solve

$$\inf_{\phi \in \mathcal{P}(S_{\boldsymbol{\theta}})} \mathrm{dist}\big(\mu, \nu_\phi\big)$$

A **global** mixture approximation problem with **minimal assumptions**:

$\rightsquigarrow$ interesting in its own right

② **Core restriction**: Use parametric families with the polynomial moments property

$$\forall \boldsymbol{\alpha},\ \exists p_{\boldsymbol{\alpha}} \in \mathbb{R}[\boldsymbol{\theta}] : \int_{\mathbb{R}^n} \boldsymbol{x}^{\boldsymbol{\alpha}}\, d\mu_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$$

③ **Consequence**: Efficient numerical scheme

$\rightsquigarrow$ Moment-SOS hierarchy

④ **By-product**

$\rightsquigarrow$ Complete mixture recovery

or

$\rightsquigarrow$ A globally informed initialization layer for clustering

Thank you!

HALL A, Poster Session 2

#**4400**