Sampling-Based Approaches for Multimarginal Optimal Transport Problems with Coulomb Cost

Yukuan Hu

State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

Joint work with **Mengyu Li** (RUC), **Xin Liu** (AMSS), and **Cheng Meng** (RUC)

Applied Math Ph.D. Seminar
Fudan University
June 8, 2023

Outline

Outline

Optimal Transport Problem

Kantorovich formulation [Monge 1781; Kantorovich 1942; Villani 2021]

$$\begin{aligned} \min_{\Pi \in \mathcal{P}((\mathbb{R}^d)^2)} \quad & \int_{(\mathbb{R}^d)^2} c(\mathbf{r}_1, \mathbf{r}_2)\, \mathrm{d}\Pi(\mathbf{r}_1, \mathbf{r}_2), \\ \text{s. t.} \quad & \int_{\mathbb{R}^d} \mathrm{d}\Pi(\mathbf{r}, \mathbf{r}_2) = p(\mathbf{r}), \quad \int_{\mathbb{R}^d} \mathrm{d}\Pi(\mathbf{r}_1, \mathbf{r}) = q(\mathbf{r}). \end{aligned} \tag{OT}$$

- $d \in \mathbb{N}$: dimension of ambient space.
- Π: 2-point probability measure over $(\mathbb{R}^d)^2$.
- $\mathcal{P}((\mathbb{R}^d)^2)$: 2-point probability measure space over $(\mathbb{R}^d)^2$.
- $c : (\mathbb{R}^d)^2 \to \mathbb{R}$: cost function, usually with the form

$$c(\mathbf{r}_1, \mathbf{r}_2) = \|\mathbf{r}_1 - \mathbf{r}_2\|_a^a, \quad \text{where } a \in (0, +\infty].$$

- $p, q \in L^1(\mathbb{R}^d)$: normalized marginal probability densities.

Optimal Transport Problem (Cont.)

Applications [Peyré-Cuturi 2019]

- Machine learning. [Kusner et al. 2015; Arjovsky-Chintala-Bottou 2017; Meng et al. 2019; ...]
- Computer vision. [Solomon et al. 2015; Xu-Sun-Liu 2019; ...]
- Biomedical research. [Gramfort-Peyré-Cuturi 2015; Ma et al. 2019; ...]
- ...

Extensions to multimarginal settings and applications

$c(\mathbf{r}_1, \mathbf{r}_2) \rightsquigarrow c(\mathbf{r}_1, \ldots, \mathbf{r}_N)$, $\Pi(\mathbf{r}_1, \mathbf{r}_2) \rightsquigarrow \Pi(\mathbf{r}_1, \ldots, \mathbf{r}_N)$, $\{p, q\} \rightsquigarrow \{p_1, \ldots, p_N\}$.

- Incompressible fluids. [Brenier 1990/1993/1999/2008; ...]
- Economics. [Carlier-Ekeland 2010; Chiappori-McCann-Nesheim 2010; ...]
- Wasserstein barycenters. [Gangbo-Święch 1998; Carlier-Oberman-Oudet 2015; ...]
- ...

All the above involve costs that increase as $\|\mathbf{r}_i - \mathbf{r}_j\|$ increases.

Multimarginal OT Problem with Coulomb Cost

$$\min_{\Pi} \int_{(\mathbb{R}^d)^N} c_{\text{Coulomb}}(\mathbf{r}_1, \ldots, \mathbf{r}_N)\, \mathrm{d}\Pi(\mathbf{r}_1, \ldots, \mathbf{r}_N), \qquad \textbf{(MMOT)}$$
$$\text{s. t.} \quad \mathscr{S}_i(\Pi) = \rho,\ i = 1, \ldots, N;\ \Pi \in \mathcal{P}\big((\mathbb{R}^d)^N\big).$$

- $N \in \mathbb{N}$: #marginals.
- Π: N-point probability measure over $(\mathbb{R}^d)^N$.
- $\mathcal{P}\big((\mathbb{R}^d)^N\big)$: N-point probability measure space over $(\mathbb{R}^d)^N$.
- $c_{\text{Coulomb}} : (\mathbb{R}^d)^N \to \mathbb{R}$: repulsive Coulomb cost function.

$$c_{\text{Coulomb}}(\mathbf{r}_1, \ldots, \mathbf{r}_N) := \sum_{i=1}^{N} \sum_{j>i} \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|}, \quad \forall\, \mathbf{r}_1, \ldots, \mathbf{r}_N \in \mathbb{R}^d.$$

- $\rho \in L^1(\mathbb{R}^d)$: normalized marginal probability density.
- $\mathscr{S}_i(\Pi) = \rho$ $(i = 1, \ldots, N)$: marginal constraints.

$$[\mathscr{S}_i(\Pi)](\mathbf{r}) := \int_{(\mathbb{R}^d)^{i-1} \times (\mathbb{R}^d)^{N-i}} \mathrm{d}\Pi(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1}, \mathbf{r}, \mathbf{r}_{i+1}, \ldots, \mathbf{r}_N), \quad \forall\, \mathbf{r} \in \mathbb{R}^d.$$

Multimarginal OT Problem with Coulomb Cost (Cont.)

Application: strongly-correlated quantum systems.

[Seidl-Perdew-Levy 1999; Gori-Giorgi-Seidl-Vignale 2009; Cohen-Mori-Sánchez-Yang 2012; Malet-Gori-Giorgi 2012; Zhou-Bahmann-Ernzerhof 2015; Friesecke-Gerolin-Gori-Giorgi 2022; ...]

- Theory: strong-interaction limit of density functional theory.
- Terminologies: $d \in \{1, 2, 3\}$–system dimension, N–#electrons in the system, ρ–normalized single-particle density.

$$\int_{\mathcal{A}_1 \times \cdots \times \mathcal{A}_N} \mathrm{d}\Pi = \text{Prob}\left\{\text{the } i\text{th electron is located in } \mathcal{A}_i \subseteq \mathbb{R}^d,\ i = 1, \ldots, N\right\}.$$

Tasks: to compute accurate approximations to

- **the optimal value**;

  also known as strictly-correlated-electrons (SCE) energy.

- **the optimal dual potentials** associated with the marginal constraints.

  [Chen-Friesecke-Mendl 2014; Di Marino-Gerolin 2020]

  also known as Kantorovich potentials $\xrightarrow[\text{average}]{\text{shift}}$ ***SCE potential***.

Discretized Form and Difficulties

Linear program after discretization

$$\min_{\mathbf{X}\geq 0} \quad \langle \mathbf{C}, \mathbf{X} \rangle := \sum_{i_1,\ldots,i_N=1}^{K} x_{i_1\cdots i_N} c_{i_1\cdots i_N},$$

$$\text{s. t.} \quad \sum_{j\neq n} \sum_{i_j=1}^{K} x_{i_1\cdots i_N} = \varrho_{i_n},\ i_n = 1,\ldots,K,\ n = 1,\ldots,N.$$

- $K \in \mathbb{N}$: #finite elements.
- $\mathbf{X} \in \mathbb{R}^{K^N}$: discretization of Π; $\mathbf{C} \in \mathbb{R}^{K^N}$: discretization of c_{Coulomb}.
- $\boldsymbol{\varrho} := [\varrho_1, \ldots, \varrho_K]^\top \in \mathbb{R}^K_{++}$: discretization of ρ (defined later).

Difficulties

- **Few results** about the structure of the optimal solutions. [Pass 2015]
- Repulsive cost, $N > 2 \Rightarrow$ **uselessness** of the analytic tools for OT & combinatorial algorithms for network flow problems. [Villani 2008; Lin et al. 2022]
- Coupling in cost $\Rightarrow$ **curse of dimensionality**.
  $K = 10^4, N = 6, K^N = 10^{24} \xrightarrow{\text{double precision}} 6776.3\ \text{ZB} \approx 7 \times 10^{15}\ \text{GB}.$

Monge-Like Ansatz and Reformulation

Monge-like ansatz on the search space [Chen et al. 2014; H.-Chen-Liu 2023]

$$\begin{cases} \mathrm{d}\Pi(\mathbf{r}_1,\ldots,\mathbf{r}_N) = \rho(\mathbf{r}_1)\gamma_2(\mathbf{r}_1,\mathbf{r}_2)\cdots\gamma_N(\mathbf{r}_1,\mathbf{r}_N)\,\mathrm{d}\mathbf{r}_1\cdots\,\mathrm{d}\mathbf{r}_N, \\ \displaystyle\int_{\mathbb{R}^d}\gamma_n(\mathbf{r}_1,\mathbf{r}_n)\,\mathrm{d}\mathbf{r}_n = 1, \quad \int_{\mathbb{R}^d}\rho(\mathbf{r}_1)\gamma_n(\mathbf{r}_1,\mathbf{r}_n)\,\mathrm{d}\mathbf{r}_1 = \rho(\mathbf{r}_n),\ \gamma_n(\mathbf{r}_1,\mathbf{r}_n)\geq 0,\ \forall\, n. \end{cases}$$

- $\gamma_n \in L^1((\mathbb{R}^d)^2)$: *coupling* between the first and nth electron positions.
- Substantial physical information.
  [Seidl 1999; Seidl et al. 1999; Seidl-Perdew-Kurth 2000; Seidl-Gori-Giorgi-Savin 2007; ...]
- Spectacular dimension reduction: $\Pi \to \{\gamma_n\}_{n=2}^N$.

Reformulation of the MMOT under the Monge-like ansatz

$$\begin{aligned} \min_{\{\gamma_n\}_{n=2}^N} \quad & \sum_{n=2}^N \int_{(\mathbb{R}^d)^2} \frac{\rho(\mathbf{r})\gamma_n(\mathbf{r},\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}' + \sum_{n=2}^N\sum_{m>n}\int_{(\mathbb{R}^d)^3}\frac{\rho(\mathbf{r})\gamma_m(\mathbf{r},\mathbf{r}')\gamma_n(\mathbf{r},\mathbf{r}'')}{\|\mathbf{r}'-\mathbf{r}''\|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\,\mathrm{d}\mathbf{r}'', \\ \text{s. t.} \quad & \int_{\mathbb{R}^d}\gamma_n(\mathbf{r},\mathbf{r}')\,\mathrm{d}\mathbf{r}' = 1, \quad \int_{\mathbb{R}^d}\rho(\mathbf{r})\gamma_n(\mathbf{r},\mathbf{r}')\,\mathrm{d}\mathbf{r} = \rho(\mathbf{r}'),\ \gamma_n(\mathbf{r},\mathbf{r}')\geq 0,\ n=2,\ldots,N. \end{aligned}$$

Discretized Form of the Monge-Like Reformulation

$$\begin{aligned} \min_{Y} \quad & f(Y) := \sum_{n=2}^{N} \langle Y_n, \Lambda C \rangle + \sum_{n=2}^{N} \sum_{m>n} \langle Y_m, \Lambda Y_n C \rangle \,, \\ \text{s. t.} \quad & Y_n \mathbf{1} = \mathbf{1},\ Y_n^\top \boldsymbol{\varrho} = \boldsymbol{\varrho},\ \mathrm{Tr}(Y_n) = 0,\ Y_n \geq 0,\ n = 2, \ldots, N, \\ & \langle Y_m, Y_n \rangle = 0,\ m, n = 2, \ldots, N : m \neq n. \end{aligned} \qquad \textbf{(MPGCC)}$$

- $\{e_k\}_{k=1}^K$: non-overlapping finite elements mesh over a bounded domain Ω.
- $\boldsymbol{\varrho} := [\varrho_1, \ldots, \varrho_K]^\top \in \mathbb{R}^K_{++}$ with $\varrho_k := \int_{e_k} \rho$; $\Lambda := \mathrm{Diag}(\boldsymbol{\varrho}) \in \mathbb{R}^{K \times K}$.
- $\mathbf{1} \in \mathbb{R}^K$: all-ones vector.
- $C := (c_{ij})$, $Y_n := (y_{n,ij})_{ij} \in \mathbb{R}^{K \times K}$, $n = 2, \ldots, N$, where

$$c_{ij} := \begin{cases} \dfrac{1}{\|\mathbf{a}_i - \mathbf{a}_j\|}, & \text{if } i \neq j, \\ 0, & \text{otherwise,} \end{cases} \qquad y_{n,ij} := \frac{1}{|e_i|} \int_{e_i} \int_{e_j} \gamma_n(\mathbf{r}_1, \mathbf{r}_n)\, \mathrm{d}\mathbf{r}_n\, \mathrm{d}\mathbf{r}_1,$$

  $\{\mathbf{a}_k\}_{k=1}^K$ are the barycenters of $\{e_k\}_{k=1}^K$.
- $Y := (Y_n)_{n=2}^N \in \left(\mathbb{R}^{K \times K}\right)^{N-1}$.

$\Rightarrow$ Mathematical program with generalized complementarity constraints (MPGCC).

Karush-Kuhn-Tucker conditions may not be necessary for local minimizers.

ℓ_1 Penalty Problem

$$\begin{aligned} \min_{Y} \quad & f_\beta(Y) := f(Y) + \beta \left(\sum_{n=2}^{N} \mathrm{Tr}(Y_n) + \sum_{n=2}^{N} \sum_{m>n} \langle Y_m, Y_n \rangle \right), \\ \text{s. t.} \quad & Y_n \mathbf{1} = \mathbf{1},\ Y_n^\top \boldsymbol{\varrho} = \boldsymbol{\varrho},\ Y_n \geq 0,\ n = 2, \ldots, N. \end{aligned} \qquad \textbf{(NQP)}$$

- $\beta > 0$: penalty parameter.
- Exactness of the ℓ_1 penalty function. [H.-Liu 2023]
- Nonconvex quadratic program (NQP) $\Rightarrow$ **NP-hardness**. [Pardalos-Vavasis 1991]

Final goal: to compute approximate global solutions for (**NQP**).

$$\text{Global solver} \xRightarrow[\text{ingredient}]{\text{important}} \text{local solver} \xRightarrow[\text{structure}]{\text{block}} \textbf{splitting methods}.$$

Limitations of the Existing Splitting Methods

Block conditional gradient method (BCG)

[Lacoste-Julien et al. 2013; Beck-Pauwels-Sabach 2015; ...]

- Subproblem:

$$\begin{aligned} \min_{Y_n} \quad & \left\langle \nabla_n f_\beta(Y_{<n}^{(k+1)}, Y_{\geq n}^{(k)}), Y_n - Y_n^{(k)} \right\rangle, \\ \text{s. t.} \quad & Y_n \mathbf{1} = \mathbf{1},\ Y_n^\top \boldsymbol{\varrho} = \boldsymbol{\varrho},\ Y_n \geq 0. \end{aligned}$$

  Essentially, an OT problem with $\nabla_n f_\beta(Y_{<n}^{(k+1)}, Y_{\geq n}^{(k)})$ as the cost matrix.

- When K is large, $\{Y^{(k)}\}$ are stored as **sparse matrices** (theory & practice).

Drawbacks

- Sparse-dense matrix multiplications $\Rightarrow$ **low scalability**.

$$\nabla_n f_\beta(Y) = \Lambda C + \beta I + \Lambda \left(\sum_{m \neq n} Y_m \right) C + \beta \left(\sum_{m \neq n} Y_m \right).$$

- **Cubic computational complexity** using linear programming methods.

Limitations of the Existing Splitting Methods (Cont.)

Proximal alternating linearized minimization method (PALM)

[Xu-Yin 2013; Bolte-Sabach-Teboulle 2014; H.-Liu 2023; ...]

- Subproblem:

$$\begin{aligned} \min_{Y_n} \quad & \left\langle \nabla_n f_\beta(Y_{<n}^{(k+1)}, Y_{\geq n}^{(k)}), Y_n - Y_n^{(k)} \right\rangle + \frac{\mu_n^{(k)}}{2} \|Y_n - Y_n^{(k)}\|_{\mathrm{F}}^2, \\ \text{s. t.} \quad & Y_n \mathbf{1} = \mathbf{1},\ Y_n^\top \boldsymbol{\varrho} = \boldsymbol{\varrho},\ Y_n \geq 0, \end{aligned}$$

  which is equivalent to projecting

$$Y_n^{(k)} - \frac{1}{\mu_n^{(k)}} \nabla_n f_\beta(Y_{<n}^{(k+1)}, Y_{\geq n}^{(k)})$$

  onto the feasible region.

- When K is large, $\{Y^{(k)}\}$ are stored as **sparse matrices** (theory & practice).

Drawback

- Sparse-dense matrix multiplications $\Rightarrow$ **low scalability**.

Route and Contributions

Route: (**MMOT**) $\xrightarrow[\text{discretization}]{\text{Monge-like ansatz}}$ (**MPGCC**) $\xrightarrow{\ell_1 \text{ penalty}}$ (**NQP**).

Contributions

(i) Designing novel splitting methods with **better scalabilities**.

entropy regularization / Kullback-Leibler divergence + entrywise sampling.

(ii) Analyzing the theoretical properties for the proposed methods.

best achieved violation $\to 0$ *as* $K \to +\infty$.

(iii) Combining the proposed methods with a grid refinements-based framework for approximately globally solving (**NQP**).

numerical observation: a local solver + random initialization "nearly" suffices for global resolution.

(iv) Conducting numerical simulations on several typical physical systems.

higher dimensionality, finer meshes.

results conforming to theoretical predictions and physical intuitions.

first visualization *of the mappings between electron positions for* ***3D systems***.

Outline

Optimization with Transportation Constraints

Let $X_n := \Lambda Y_n \in \mathbb{R}^{K\times K}$ $(n = 2, \dots, N)$.

$$\begin{aligned} \min_{X} \quad & g_\beta(X) := f_\beta(\Lambda^{-1}X_2, \dots, \Lambda^{-1}X_N), \\ \text{s. t.} \quad & X_n \in \mathcal{S} := \{T \in \mathbb{R}^{K\times K} : T\mathbf{1} = T^\top \mathbf{1} = \boldsymbol{\varrho},\ T \geq 0\},\ \forall\, n. \end{aligned} \tag{P}$$

- $X := (X_2, \dots, X_N) \in (\mathbb{R}^{K\times K})^{N-1}$.

In our context, more beneficial for numerical resolution.

Subproblems in the BCG-Like Methods

$$\min_{X_n} \left\langle C_n^{(k)}, X_n - X_n^{(k)} \right\rangle, \quad \text{s. t. } X_n \mathbf{1} = \boldsymbol{\varrho},\ X_n^\top \mathbf{1} = \boldsymbol{\varrho},\ X_n \geq 0.$$

- $C_n^{(k)} := \nabla_n g_\beta \left(X_{<n}^{(k+1)}, X_{\geq n}^{(k)} \right) \in \mathbb{R}^{K \times K}$.
- Essentially an OT problem, with $C_n^{(k)}$ as the cost matrix.
- Computing $C_n^{(k)}$ involves sparse-dense matrix multiplications.
- Cubic computational complexity using linear programming methods.

How to reduce the computational complexity for solving an OT problem?

Add extra entropy regularization.

Entropy Regularized Subproblems

$$\min_{X_n} \ \left\langle C_n^{(k)}, X_n - X_n^{(k)} \right\rangle + \lambda_n^{(k)} h(X_n), \quad \text{s. t. } X_n \mathbf{1} = \boldsymbol{\varrho},\ X_n^\top \mathbf{1} = \boldsymbol{\varrho}.$$

- $\lambda_n^{(k)} > 0$: regularization parameter.
- $h(T) := \sum_{ij} t_{ij}(\log t_{ij} - 1)$: negative entropy of any $T \in \mathbb{R}_+^{K \times K}$.
  (statistical) thermodynamics, quantum statistical mechanics, information theory.
- Early usage in transportation field. [Wilson 1969]
- Wide adoption in machine learning community nowadays. [Flamary et al. 2021]

Entropy Regularized Subproblems

$$\min_{X_n} \ \left\langle C_n^{(k)}, X_n - X_n^{(k)} \right\rangle + \lambda_n^{(k)} h(X_n), \quad \text{s. t. } X_n \mathbf{1} = \boldsymbol{\varrho},\ X_n^\top \mathbf{1} = \boldsymbol{\varrho}.$$

- $\lambda_n^{(k)} > 0$: regularization parameter.
- $h(T) := \sum_{ij} t_{ij}(\log t_{ij} - 1)$: negative entropy of any $T \in \mathbb{R}_+^{K \times K}$.
  (statistical) thermodynamics, quantum statistical mechanics, information theory.
- Early usage in transportation field. [Wilson 1969]
- Wide adoption in machine learning community nowadays. [Flamary et al. 2021]

Good things about entropy regularization

- Strongly convex problem $\Rightarrow$ differentiability of the optimal value w.r.t. $\boldsymbol{\varrho}$.
- Upper bound on the optimal value difference (related to $\boldsymbol{\varrho}$ and $\lambda_n^{(k)}$).
  [Blondel-Seguy-Rolet 2018; Genevay et al. 2019; Kerdoncuff-Emonet-Sebban 2021]
- Convergence of the optimal value and solution as $\lambda_n^{(k)} \to 0$.
  [Cominetti-San Martín 1994]
- **Highly scalable** iterative schemes for numerical resolution. [Cuturi 2013]
- **Multiplicative formula** for the optimal solution (given an optimal dual one).

Dual Form

$$\min_{\mathbf{u}_n,\mathbf{v}_n} \; q(\mathbf{u}_n,\mathbf{v}_n;\lambda_n^{(k)},\Psi_n^{(k)}) := \lambda_n^{(k)} \exp\left(\frac{\mathbf{u}_n}{\lambda_n^{(k)}}\right)^\top \Psi_n^{(k)} \exp\left(\frac{\mathbf{v}_n}{\lambda_n^{(k)}}\right) - (\mathbf{u}_n+\mathbf{v}_n)^\top \boldsymbol{\varrho},$$

where $\mathbf{u}_n,\, \mathbf{v}_n \in \mathbb{R}^K$ are the dual variables, "$\exp(\cdot)$" is entrywise exponential,

$$\Psi_n^{(k)} := \exp\left(-C_n^{(k)}/\lambda_n^{(k)}\right) \in \mathbb{R}^{K\times K}$$

is called the (Gibbs) kernel matrix.

Relation between the optimal primal and dual solutions

$$X_n^{(k+1,\star)} := \mathrm{Diag}\left(\exp\left(\frac{\mathbf{u}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right) \Psi_n^{(k)} \mathrm{Diag}\left(\exp\left(\frac{\mathbf{v}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right). \qquad (1)$$

- $X_n^{(k+1,\star)} \in \mathbb{R}^{K\times K}$: optimal primal solution.
- $(\mathbf{u}_n^{(k+1,\star)}, \mathbf{v}_n^{(k+1,\star)}) \in (\mathbb{R}^K)^2$: optimal dual solution.

Dual BCD Method for Entropy Regularized Subproblems

$$\begin{aligned}
\mathbf{u}_n^{(k,j+1)} &:= \lambda_n^{(k)} \log\left(\boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)} \exp\left(\mathbf{v}_n^{(k,j)}/\lambda_n^{(k)}\right)\right)\right), \\
\mathbf{v}_n^{(k,j+1)} &:= \lambda_n^{(k)} \log\left(\boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)\top} \exp\left(\mathbf{u}_n^{(k,j+1)}/\lambda_n^{(k)}\right)\right)\right).
\end{aligned}$$

- "$\log(\cdot)$": entrywise logarithm; "$\oslash$": entrywise division.

Dual BCD Method for Entropy Regularized Subproblems

$$\begin{aligned}\mathbf{u}_n^{(k,j+1)} &:= \lambda_n^{(k)} \log\left(\boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)} \exp\left(\mathbf{v}_n^{(k,j)}/\lambda_n^{(k)}\right)\right)\right),\\ \mathbf{v}_n^{(k,j+1)} &:= \lambda_n^{(k)} \log\left(\boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)\top} \exp\left(\mathbf{u}_n^{(k,j+1)}/\lambda_n^{(k)}\right)\right)\right).\end{aligned}$$

- "$\log(\cdot)$": entrywise logarithm; "$\oslash$": entrywise division.

Letting $\check{\mathbf{u}}_n^{(k,j)} := \exp\left(\mathbf{u}_n^{(k,j)}/\lambda_n^{(k)}\right)$, $\check{\mathbf{v}}_n^{(k,j)} := \exp\left(\mathbf{v}_n^{(k,j)}/\lambda_n^{(k)}\right) \in \mathbb{R}^K$.

$$\check{\mathbf{u}}_n^{(k,j+1)} := \boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)} \check{\mathbf{v}}_n^{(k,j)}\right),\ \check{\mathbf{v}}_n^{(k,j+1)} := \boldsymbol{\varrho} \oslash \left(\Psi_n^{(k)\top} \check{\mathbf{u}}_n^{(k,j+1)}\right).$$

- Matrix-vector multiplications: $\mathcal{O}(K^2)$ complexity and **high scalability**.
- R-linear convergence rate. [Luo-Tseng 1992/1993]
- Warm start $\Rightarrow$ **further acceleration**.
- Other names: iterative proportional fitting procedure, RAS methods, Sinkhorn's algorithm, ... [Deming-Stephan 1940; Sinkhorn 1964; Bacharach 1965]

Entropy Regularized Alternating Linearized Minimization Method

Algorithm 1: Entropy regularized alternating linearized minimization method (**ERALM**).

Input: C, $X_n^{(0)} \in \mathbb{R}^{K\times K}$ $(n = 2,\dots,N)$, $\boldsymbol{\varrho}$, $\tilde{\mathbf{v}}_n^{(0)} \in \mathbb{R}^K$ $(n = 2,\dots,N)$, $\alpha^{(0)} \in [0,1]$, $\beta > 0$, $k_{\max} \in \mathbb{N}$.

Set $k := 0$;
while *certain conditions not satisfied* **and** $k < k_{\max}$ **do**
  for $n = 2,\dots,N$ **do**
    Choose regularization parameter $\lambda_n^{(k)} > 0$;
    Compute $C_n^{(k)}$ and construct $\Psi_n^{(k)} \in \mathbb{R}^{K\times K}$;
    Starting from $\tilde{\mathbf{v}}_n^{(k)}$, solve the following subproblem

$$\min_{\tilde{\mathbf{u}}_n,\tilde{\mathbf{v}}_n} \; q(\tilde{\mathbf{u}}_n,\tilde{\mathbf{v}}_n;\lambda_n^{(k)},\Psi_n^{(k)})$$

    using BCD to obtain $\tilde{\mathbf{u}}_n^{(k+1)}$ and $\tilde{\mathbf{v}}_n^{(k+1)} \in \mathbb{R}^K$;
    Update $\tilde{X}_n^{(k+1)} \in \mathbb{R}^{K\times K}$ as in (1) with $\tilde{\mathbf{u}}_n^{(k+1)}$ and $\tilde{\mathbf{v}}_n^{(k+1)}$;
    Update $X_n^{(k+1)} := (1-\alpha^{(k)})X_n^{(k)} + \alpha^{(k)}\tilde{X}_n^{(k+1)} \in \mathbb{R}^{K\times K}$;
  end
  Choose step size $\alpha^{(k+1)} \in (0,1]$;
  Set $k := k+1$;
end

Output: Approximate solution $X^{(k)} \in (\mathbb{R}^{K\times K})^{N-1}$.

How to avoid the full matrix multiplications in $C_n^{(k)}$?

Optimal solution of the subproblem

$$\tilde{X}_n^{(k+1,\star)} := \mathrm{Diag}\left(\exp\left(\frac{\tilde{\mathbf{u}}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right)\Psi_n^{(k)}\mathrm{Diag}\left(\exp\left(\frac{\tilde{\mathbf{v}}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right) \in \mathbb{R}^{K\times K},$$

where $(\tilde{\mathbf{u}}_n^{(k+1,\star)}, \tilde{\mathbf{v}}_n^{(k+1,\star)}) \in (\mathbb{R}^K)^2$ is an optimal dual solution.

Observation: $\Psi_{n,ij}^{(k)} = 0 \Rightarrow \tilde{x}_{n,ij}^{(k+1,\star)} = 0.$

Optimal solution of the subproblem

$$\tilde{X}_n^{(k+1,\star)} := \mathrm{Diag}\left(\exp\left(\frac{\tilde{\mathbf{u}}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right)\Psi_n^{(k)}\mathrm{Diag}\left(\exp\left(\frac{\tilde{\mathbf{v}}_n^{(k+1,\star)}}{\lambda_n^{(k)}}\right)\right) \in \mathbb{R}^{K\times K},$$

where $(\tilde{\mathbf{u}}_n^{(k+1,\star)}, \tilde{\mathbf{v}}_n^{(k+1,\star)}) \in (\mathbb{R}^K)^2$ is an optimal dual solution.

Observation: $\Psi_{n,ij}^{(k)} = 0 \Rightarrow \tilde{x}_{n,ij}^{(k+1,\star)} = 0.$

Idea: sparse optimal solution to (P). [Hosseini-Steinerberger 2022; H.-Liu 2022]
$\Rightarrow$ **only a small portion** of $\Psi_n^{(k)}$ (and thus $C_n^{(k)}$) is needed.

Tools: **matrix sparsification** by **importance sampling**. [Liu 1996/2004; Owen 2013]
$\Rightarrow$ **good guesses** on the sparsity pattern.

Importance Sampling

- Optimal sampling probabilities: $p_{n,ij}^{(k,\star)} \propto \tilde{x}_{n,ij}^{(k+1,\star)}$ $(i,j=1,\dots,K)$.
- Alternative: $p_{n,ij}^{(k)\prime} \propto x_{n,ij}^{(k)}$ $(i,j=1,\dots,K)$.
  reasonable when $X^{(k)}$ is close to an optimal solution.
- Shrinkage strategy: interpolate between $p_{n,ij}^{(k)\prime}$ and p_{ij}''.

$$p_{n,ij}^{(k)} := \gamma p_{n,ij}^{(k)\prime} + (1-\gamma) p_{ij}'' = \gamma \frac{x_{n,ij}^{(k)}}{\sum_{i,j} x_{n,ij}^{(k)}} + (1-\gamma) \frac{\sqrt{\varrho_i \varrho_j}}{\sum_{i',j'} \sqrt{\varrho_{i'} \varrho_{j'}}}, \quad (2)$$

where $\gamma \in [0,1]$ is the interpolation factor. [Ma-Mahoney-Yu 2014; Yu et al. 2022]

when using equi-mass discretization, $\{p_{ij}''\}$ reduces to uniform distribution.

Construction of Sparse Kernel Matrices

Let $\mathcal{I}_n^{(k)}$ be the sampled set of indices. The **sparse approximation** for $\Psi_n^{(k)}$:

$$\hat{\Psi}_{n,ij}^{(k)} := \begin{cases} \Psi_{n,ij}^{(k)} / \left(\left| \mathcal{I}_n^{(k)} \right| \cdot p_{n,ij}^{(k)} \right), & \text{if } (i,j) \in \mathcal{I}_n^{(k)}, \\ 0, & \text{otherwise,} \end{cases} \qquad i,j = 1,\ldots,K, \quad (3)$$

where the adjustment factors $\left| \mathcal{I}_n^{(k)} \right| \cdot p_{n,ij}^{(k)}$ $(i,j = 1,\ldots,K)$ ensure the unbiasedness of the random approximation.

Sampling-Based ERALM

Algorithm 2: Sampling-based ERALM (**S-ERALM**).

Input: C, $X_n^{(0)} \in \mathbb{R}^{K\times K}$ $(n = 2,\dots,N)$, $\boldsymbol{\varrho}$, $\tilde{\mathbf{v}}_n^{(0)} \in \mathbb{R}^K$ $(n = 2,\dots,N)$, γ, $\alpha^{(0)} \in [0,1]$, $\beta > 0$, s, $k_{\max} \in \mathbb{N}$.

Set $k := 0$;

while *certain conditions not satisfied* **and** $k < k_{\max}$ **do**

  for $n = 2,\dots,N$ **do**

    Choose regularization parameter $\lambda_n^{(k)} > 0$;

    Randomly pick a subset $\mathcal{I}_n^{(k)} \subseteq \{(i,j) : i,j = 1,\dots,K\}$ according to the probability distribution $P_n^{(k)} = (p_{n,ij}^{(k)}) \in \mathbb{R}^{K\times K}$ in (2) such that $\left|\mathcal{I}_n^{(k)}\right| = s$;

    Construct sparse approximate kernel matrix $\hat{\Psi}_n^{(k)} \in \mathbb{R}^{K\times K}$ as in (3);

    Starting from $\tilde{\mathbf{v}}_n^{(k)}$, solve the following subproblem

$$\min_{\tilde{\mathbf{u}}_n,\tilde{\mathbf{v}}_n} \; q(\tilde{\mathbf{u}}_n,\tilde{\mathbf{v}}_n;\lambda_n^{(k)},\hat{\Psi}_n^{(k)})$$

    using BCD to obtain $\tilde{\mathbf{u}}_n^{(k+1)}$ and $\tilde{\mathbf{v}}_n^{(k+1)} \in \mathbb{R}^K$;

    Update $\tilde{X}_n^{(k+1)} \in \mathbb{R}^{K\times K}$ as in (1) with $\tilde{\mathbf{u}}_n^{(k+1)}$ and $\tilde{\mathbf{v}}_n^{(k+1)}$;

    Update $X_n^{(k+1)} := (1-\alpha^{(k)})X_n^{(k)} + \alpha^{(k)}\tilde{X}_n^{(k+1)} \in \mathbb{R}^{K\times K}$;

  end

  Choose step size $\alpha^{(k+1)} \in (0,1]$;

  Set $k := k+1$;

end

Output: Approximate solution $X^{(k)} \in (\mathbb{R}^{K\times K})^{N-1}$.

Subproblems in the PALM-Like Methods

$$\min_{X_n} \ \left\langle C_n^{(k)}, X_n - X_n^{(k)} \right\rangle + \frac{\mu_n^{(k)}}{2} \|X_n - X_n^{(k)}\|_{\mathrm{F}}^2, \quad \text{s. t. } X_n \in \mathcal{S}.$$

- Equivalent to projecting $X_n^{(k)} - C_n^{(k)}/\mu_n^{(k)}$ onto $\mathcal{S}$;
- Computing $C_n^{(k)}$ involves sparse-dense matrix multiplications.

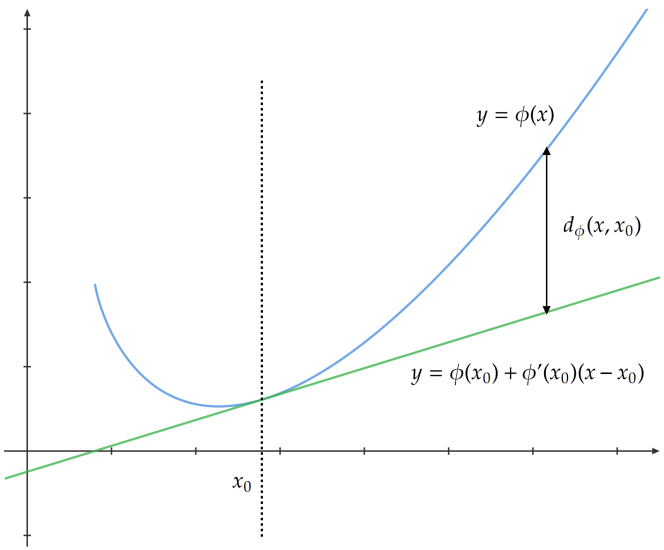
How to reduce the complexity for solving proximal subproblems?

Use Kullback-Leibler (KL) divergence.

Definition 1 (Bregman distance [Bregman 1967])

Given a proper closed strictly convex function $\phi : \mathbb{R}^{K\times K} \to (-\infty, +\infty]$, finite at $T, T' \in \mathbb{R}^{K\times K}$ and differentiable at T', the Bregman distance between T and T' associated with the kernel function ϕ is

$$d_\phi(T, T') := \phi(T) - \phi(T') - \left\langle \nabla\phi(T'), T - T' \right\rangle.$$



- $d_\phi(T, T') \geq 0$ and "=" holds iff $T = T'$.
- $\phi(\cdot) = \|\cdot\|_{\mathrm{F}}^2/2 \Rightarrow d_\phi(T, T') = \|T - T'\|_{\mathrm{F}}^2/2$.

Bregman Distance and KL Divergence

Definition 1 (Bregman distance [Bregman 1967])

Given a proper closed strictly convex function $\phi : \mathbb{R}^{K\times K} \to (-\infty, +\infty]$, finite at $T, T' \in \mathbb{R}^{K\times K}$ and differentiable at T', the Bregman distance between T and T' associated with the kernel function ϕ is

$$d_\phi(T, T') := \phi(T) - \phi(T') - \left\langle \nabla\phi(T'), T - T' \right\rangle.$$

Definition 2 (KL divergence [Kullback-Leibler 1951])

Given any $T \in \mathbb{R}_+^{K\times K}$ and $T' \in \mathbb{R}_{++}^{K\times K}$, the KL divergence between T and T' is

$$\mathrm{KL}(T; T') := \sum_{i,j} [t_{ij}(\log t_{ij} - \log t'_{ij}) - (t_{ij} - t'_{ij})].$$

$$\mathrm{KL}(T; T') = d_h(T, T').$$

KL Divergence-Based Subproblems

$$\begin{aligned} \min_{X_n} \quad & \left\langle C_n^{(k)}, X_n - X_n^{(k)} \right\rangle + \mu_n^{(k)} \mathrm{KL}(X_n; X_n^{(k)}), \\ \text{s. t.} \quad & X_n \mathbf{1} = \boldsymbol{\varrho},\ X_n^\top \mathbf{1} = \boldsymbol{\varrho}. \end{aligned}$$

- $\mu_n^{(k)} > 0$: proximal parameter.
- Related works:
  - Bregman proximal point algorithm.
    [Xie et al. 2020; Yang-Toh, 2022; Chu et al. 2023; ...]
  - Bregman PALM with Lipschitz smooth kernel functions.
    [Hua-Yamashita 2016; Li et al. 2019; Ahookhosh et al. 2021; ...]

The highly scalable dual BCD schemes & multiplicative formula STILL apply.

Dual Form and Dual BCD

Dual form

$$\min_{\mathbf{u}_n,\mathbf{v}_n} \; q(\mathbf{u}_n,\mathbf{v}_n;\mu_n^{(k)},\Phi_n^{(k)}) = \mu_n^{(k)} \exp\left(\frac{\mathbf{u}_n}{\mu_n^{(k)}}\right)^\top \Phi_n^{(k)} \exp\left(\frac{\mathbf{v}_n}{\mu_n^{(k)}}\right) - (\mathbf{u}_n+\mathbf{v}_n)^\top \boldsymbol{\varrho},$$

where $\mathbf{u}_n,\, \mathbf{v}_n \in \mathbb{R}^K$ are the dual variables,

$$\Phi_n^{(k)} := \exp\left(-C_n^{(k)}/\mu_n^{(k)}\right) \odot X_n^{(k)} \in \mathbb{R}^{K\times K}$$

is the kernel matrix.

Dual BCD: $\check{\mathbf{u}}_n^{(k,j)} := \exp\left(\mathbf{u}_n^{(k,j)}/\mu_n^{(k)}\right)$, $\check{\mathbf{v}}_n^{(k,j)} := \exp\left(\mathbf{v}_n^{(k,j)}/\mu_n^{(k)}\right) \in \mathbb{R}^K$.

$$\check{\mathbf{u}}_n^{(k,j+1)} := \boldsymbol{\varrho} \oslash \left(\Phi_n^{(k)} \check{\mathbf{v}}_n^{(k,j)}\right), \;\; \check{\mathbf{v}}_n^{(k,j+1)} := \boldsymbol{\varrho} \oslash \left(\Phi_n^{(k)\top} \check{\mathbf{u}}_n^{(k,j+1)}\right).$$

KL Divergence-Based Alternating Linearized Minimization

Algorithm 3: KL divergence-based alternating linearized minimization (**KLALM**).

Input: C, $X_n^{(0)} \in \mathbb{R}^{K\times K}$ $(n = 2, \ldots, N)$, $\boldsymbol{\varrho}$, $\mathbf{v}_n^{(0)} \in \mathbb{R}^K$ $(n = 2, \ldots, N)$, $\beta > 0$, $k_{\max} \in \mathbb{N}$.

Set $k := 0$;

while *certain conditions not satisfied* **and** $k < k_{\max}$ **do**

  for $n = 2, \ldots, N$ **do**

    Choose proximal parameter $\mu_n^{(k)} > 0$;

    Compute $C_n^{(k)}$ and construct $\Phi_n^{(k)} \in \mathbb{R}^{K\times K}$;

    Starting from $\mathbf{v}_n^{(k)}$, solve the following subproblem

$$\min_{\mathbf{u}_n, \mathbf{v}_n} \; q(\mathbf{u}_n, \mathbf{v}_n; \mu_n^{(k)}, \Phi_n^{(k)})$$

    using BCD to obtain $\mathbf{u}_n^{(k+1)}$ and $\mathbf{v}_n^{(k+1)} \in \mathbb{R}^K$;

    Update $X_n^{(k+1)} \in \mathbb{R}^{K\times K}$ as in (1) with $\mathbf{u}_n^{(k+1)}$ and $\mathbf{v}_n^{(k+1)}$;

  end

  Set $k := k + 1$;

end

Output: Approximate solution $X^{(k)} \in (\mathbb{R}^{K\times K})^{N-1}$.

Optimal solution of the subproblem

$$X_n^{(k+1,\star)} := \mathrm{Diag}\left(\exp\left(\frac{\mathbf{u}_n^{(k+1,\star)}}{\mu_n^{(k)}}\right)\right)\Phi_n^{(k)}\mathrm{Diag}\left(\exp\left(\frac{\mathbf{v}_n^{(k+1,\star)}}{\mu_n^{(k)}}\right)\right) \in \mathbb{R}^{K\times K},$$

where $(\mathbf{u}_n^{(k+1,\star)}, \mathbf{v}_n^{(k+1,\star)}) \in (\mathbb{R}^K)^2$ is an optimal dual solution.

Observation: $x_{n,ij}^{(k)} = 0 \Rightarrow \Phi_{n,ij}^{(k)} = 0 \Rightarrow x_{n,ij}^{(k+1,\star)} = 0.$

$$\Phi_n^{(k)} = \exp\left(-C_n^{(k)}/\mu_n^{(k)}\right) \odot X_n^{(k)}.$$

Idea: matrix sparsification by importance sampling **only in a critical iteration**.

reasonable when $X^{(k)}$ is close to an optimal solution.

Sampling-Based KLALM

Algorithm 4: Sampling-based KLALM (**S-KLALM**).

Input: C, $X_n^{(0)} \in \mathbb{R}^{K\times K}$ $(n = 2,\dots,N)$, $\boldsymbol{\varrho}$, $\mathbf{v}_n^{(0)} \in \mathbb{R}^K$ $(n = 2,\dots,N)$, $\gamma \in [0,1]$, $\beta > 0$, s, $\hat{k}$, $k_{\max} \in \mathbb{N}$.

Set $k := 0$ and $\mathcal{I}_n := \{(i,j) : i,j = 1,\dots,K\}$ $(n = 2,\dots,N)$;

while *certain conditions not satisfied* **and** $k < k_{\max}$ **do**

  for $n = 2,\dots,N$ **do**

    Choose proximal parameter $\mu_n^{(k)} > 0$;

    if $k = \hat{k}$ **then**

      Randomly pick a subset $\mathcal{I}_n^{(k)} \subseteq \{(i,j) : i,j = 1,\dots,K\}$ according to the probability distribution $P_n^{(k)} = (p_{n,ij}^{(k)}) \in \mathbb{R}^{K\times K}$ in (2) such that $\left|\mathcal{I}_n^{(k)}\right| = s$;

    end

    if $k < \hat{k}$ **then**

      Let $\hat{\Phi}_n^{(k)} := \Phi_n^{(k)} \in \mathbb{R}^{K\times K}$;

    else

      Construct sparse approximate kernel matrix $\hat{\Phi}_n^{(k)} \in \mathbb{R}^{K\times K}$ as in (3) with $\mathcal{I}_n^{(\hat{k})}$ and $P_n^{(\hat{k})}$;

    end

    Starting from $\mathbf{v}_n^{(k)}$, solve the following subproblem

$$\min_{\mathbf{u}_n,\mathbf{v}_n} \; q(\mathbf{u}_n,\mathbf{v}_n;\mu_n^{(k)},\hat{\Phi}_n^{(k)})$$

    using BCD to obtain $\mathbf{u}_n^{(k+1)}$ and $\mathbf{v}_n^{(k+1)} \in \mathbb{R}^K$;

    Update $X_n^{(k+1)} \in \mathbb{R}^{K\times K}$ as in (1) with $\mathbf{u}_n^{(k+1)}$ and $\mathbf{v}_n^{(k+1)}$;

  end

  Set $k := k + 1$;

end

Output: Approximate solution $X^{(k)} \in (\mathbb{R}^{K\times K})^{N-1}$.

Computational Complexities

Table 1: A comparison of computational complexities.

Ingredients in one iteration	**ERALM**	**KLALM**
Kernel matrices	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$
Subiterations	$j_{\max} \times \mathcal{O}(K^2)$	$j_{\max} \times \mathcal{O}(K^2)$
Total	$k_{\max}\big(\mathcal{O}(K^3) + j_{\max} \times \mathcal{O}(K^2)\big)$	$k_{\max}\big(\mathcal{O}(K^3) + j_{\max} \times \mathcal{O}(K^2)\big)$
Ingredients in one iteration	**S-ERALM**	**S-KLALM**
Sampling	$\mathcal{O}(K^2)$	$\mathcal{O}(K^2)$ (only when $k = \hat{k}$)
Kernel matrices	$\mathcal{O}(s^2/K)$	$\mathcal{O}(s^2/K)$
Subiterations	$j_{\max} \times \mathcal{O}(s)$	$j_{\max} \times \mathcal{O}(s)$
Total	$k_{\max}\big(\mathcal{O}(K^2) + \mathcal{O}(s^2/K) + j_{\max} \times \mathcal{O}(s)\big)$	$k_{\max}\big(\mathcal{O}(s^2/K) + j_{\max} \times \mathcal{O}(s)\big) + \mathcal{O}(K^2)$

- $k_{\max}, j_{\max} \in \mathbb{N}$: the maximum iteration number and subiteration number.
- Assumption: the sampled entries are uniformly distributed in each row and column.
- Warm start for BCD + "$s = K^\tau$" ($\tau \in (1, 2)$) $\Rightarrow$ best complexity per iteration:

$$\mathcal{O}(K^{2\tau-1}) \textbf{ by S-KLALM}.$$

Outline

Lemma 1

Let $L := \|C\|_2/(\min_k \varrho_k) + \beta/(\min_k \varrho_k)^2$. *Then, for* $n = 2, \dots, N$, $\nabla_n g_\beta$ *is* L*-Lipschitz continuous, i.e., for any* X, $X' \in (\mathbb{R}^{K\times K})^{N-1}$,

$$\|\nabla_n g_\beta(X) - \nabla_n g_\beta(X')\|_{\mathrm{F}} \le L\|X - X'\|_{\mathrm{F}}.$$

Assumption 1

There exists $q > 0$ *such that* $\|\boldsymbol{\varrho}\|_\infty \le q \cdot \min_k \varrho_k$ *and* $\|C\|_2 \le qK$ *for any* K.

$\Rightarrow L = \mathcal{O}(K^2)$, $\|\boldsymbol{\varrho}\|_\infty = \mathcal{O}(1/K)$, $-h(\boldsymbol{\varrho}\boldsymbol{\varrho}^\top) = \mathcal{O}(\log K)$.

Residual Functions

Define the residual functions $R_n : (\mathbb{R}^{K\times K})^{N-1} \to \mathbb{R}$ $(n = 2, \ldots, N)$ as

$$R_n(X) := g_\beta(X) - \min_{T\in\mathcal{S}} g_\beta(X_{<n}, T, X_{>n}), \quad \forall\, X \in (\mathbb{R}^{K\times K})^{N-1}.$$

Let $R := \sum_{n=2}^{N} R_n$.

Lemma 2

For any $X \in \mathcal{S}^{N-1}$, $R(X) \geq 0$ and $X \in \mathcal{S}^{N-1}$ is a KKT point of (P) if and only if $R(X) = 0$.

$R(X)$ can characterize the stationarity violation at X.

Theorem 1

Let $\{X^{(k)}\}$ *be the sequence generated by* ERALM *where the subproblems are exactly solved. Suppose that*

$$\alpha^{(k)} \equiv \alpha := \frac{1}{2(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty}\sqrt{\frac{g_\beta(X^{(0)}) - g_\beta^\star}{KL \cdot k_{\max}}}$$

for any $k \geq 0$*, where* $g_\beta^\star \in \mathbb{R}$ *is the optimal value of* (P)*. Then*

$$0 \leq \min_{k=1}^{k_{\max}} R(X^{(k)}) \leq 4(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty \sqrt{\frac{KL(g_\beta(X^{(0)}) - g_\beta^\star)}{k_{\max}}} - (N-1)\bar{\lambda}h(\boldsymbol{\varrho}\boldsymbol{\varrho}^\top),$$

where $\bar{\lambda} := \max_{k=1}^{k_{\max}} \max_{n=2}^{N} \lambda_n^{(k)}$.

Corollary 1

Let $\{X^{(k)}\}$ *be the sequence generated by* ERALM *where the subproblems are exactly solved. Suppose that Assumption 1 holds and* $X^{(0)}$ *is chosen such that* $g_\beta(X^{(0)}) \leq M$*, where* M *is a constant irrelevant to* K*. Also assume that*

$$\lambda_n^{(k)} \equiv \lambda = o(1/\log K),\ k_{\max} = \mathcal{O}(K^{1+\eta}),$$

$$\alpha^{(k)} \equiv \alpha := \frac{1}{2(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty}\sqrt{\frac{g_\beta(X^{(0)}) - g_\beta^\star}{KL \cdot k_{\max}}},$$

for any $k \geq 0$ *and* $n = 2, \ldots, N$*, where* $\eta > 0$*. Then* $\min_{k=1}^{k_{\max}} R(X^{(k)}) \to 0$ *as* $K \to +\infty$*.*

Assumption 2

(i) *There exist constants* $t \in (1/2, 1]$, c_1, c_2, $\hat{c}_2 > 0$ *such that, for any* $k \geq 0$, $\|\Psi_n^{(k)}\|_2 \geq K^t/c_1$ *and the spectral condition numbers of* $\Psi_n^{(k)}$, $\hat{\Psi}_n^{(k)}$ *are bounded by* c_2, $\hat{c}_2$, *respectively.*

(ii) *The interpolation factor* $\gamma < 1$ *and there exists* $\varepsilon > 0$ *such that*

$$s \geq \frac{8K^{1-2t}\log^4(2K)}{(1-\gamma)\underline{\varrho}\cdot\log^4(1+\varepsilon)},$$

where $\underline{\varrho} := \min_{i,j}\sqrt{\varrho_i\varrho_j}/\sum_{i',j'}\sqrt{\varrho_{i'}\varrho_{j'}}$.

- Under Assumption 1, $\underline{\varrho} = \mathcal{O}(1/K^2)$ and $s \geq \mathcal{O}(K^{3-2t}\log^4(K))$.
- Similar assumptions are also used in [Li et al. 2022].

Theorem 2

Let $\{X^{(k)}\}$ be the sequence generated by S-ERALM *where the subproblems are feasible and exactly solved. Suppose that Assumption 2 holds. Also assume that*

$$\alpha^{(k)} \equiv \alpha := \frac{1}{2(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty}\sqrt{\frac{g_\beta(X^{(0)}) - g_\beta^\star}{KL \cdot k_{\max}}}$$

for any $k \geq 0$, where $g_\beta^\star \in \mathbb{R}$ is the optimal value of (P). Then, for any $\theta > 0$ and $K > 76$, with probability no less than $\left(1 - 2\exp(-16\theta^2 \log^4(K)/\varepsilon^4)\right)^{k_{\max}}$, there holds

$$\begin{aligned} 0 \leq \min_{k=1}^{k_{\max}} R(X^{(k)}) &\leq 4(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty \sqrt{\frac{KL(g_\beta(X^{(0)}) - g_\beta^\star)}{k_{\max}}} \\ &\quad - 2(N-1)\bar{\lambda}h(\boldsymbol{\varrho}\boldsymbol{\varrho}^\top) + (N-1)\bar{\lambda}\frac{\hat{c}_2 c_3}{\log^2(2K) - c_3} \\ &\quad + (N-1)\bar{\lambda}\sqrt{K}\|\boldsymbol{\varrho}\|_\infty \sqrt{s}\log\frac{1}{(1-\gamma)\underline{\varrho} \cdot s} \end{aligned}$$

where $\bar{\lambda} := \max_{k=1}^{k_{\max}} \max_{n=2}^{N} \lambda_n^{(k)}$, $c_3 = c_1(1+\varepsilon+\theta)\log^2(1+\varepsilon)$.

Corollary 2

Let $\{X^{(k)}\}$ *be the sequence generated by* S-ERALM *where the subproblems are feasible and exactly solved. Suppose that Assumptions 1 and 2 hold with* t, c_1, c_2, c_2', ε *and* $X^{(0)}$ *is chosen such that* $g_\beta(X^{(0)}) \leq M$, *where* t, c_1, c_2, c_2', ε, M *are constants irrelevant to* K. *Also assume that*

$$\lambda_n^{(k)} \equiv \lambda := o(1/(K^{1-t}\log^3(K))),\ s = \mathcal{O}(K^{3-2t}\log^4(K)),$$

$$\alpha^{(k)} \equiv \alpha := \frac{1}{2(N-1)^{3/4}\|\boldsymbol{\varrho}\|_\infty}\sqrt{\frac{g_\beta(X^{(0)}) - g_\beta^\star}{KL\cdot k_{\max}}},\ k_{\max} = \mathcal{O}(K^{1+\eta}),$$

for any $k \geq 0$ *and* $n = 2, \ldots, N$, *where* $\eta > 0$. *Then* $\min_{k=1}^{k_{\max}} R(X^{(k)}) \to 0$ *as* $K \to +\infty$ *with probability going to* 1.

Outline

Implementation Details

Optimization model

- Equi-mass or equi-size discretization (fixed later).
- Penalty parameter $\beta = 1$. [H. et al. 2023]
- Truncation on $\boldsymbol{\varrho}$: discard the entries smaller than 0.1% of $\|\boldsymbol{\varrho}\|_\infty$. [Dvurechensky-Gasnikov-Kroshnin 2018]

Outer loop

- Step sizes $\alpha^{(k)} = 1/(k+1)$ (ERALM and S-ERALM).
- Interpolation factor $\gamma = 0.99$ (S-ERALM and S-KLALM).
- Sample size $s = \lfloor K^{1.5} \rfloor$ (S-ERALM and S-KLALM).
- Critical iteration number $\hat{k} = 0$ (S-KLALM).
- Regularization/proximal parameters

$$\lambda_n^{(k)} = \|\tilde{\mathbf{v}}_n^{(k)}\|_\infty/(20\log(K)), \quad \mu_n^{(k)} = \|\mathbf{v}_n^{(k)}\|_\infty/(20\log(K)).$$

Implementation Details (Cont.)

Outer loop (cont.)

- Stopping rules: one of the following three holds.
  - $\Delta^{(k)} := \sum_{n=2}^{N} \|\Lambda^{-1}(X_n^{(k)} - X_n^{(k-1)})\|/(N-1) \leq tol$ (fixed later).
  - $k \geq k_{\max}$ (fixed later).
  - CPU time in seconds $\leq T_{\max}$ (fixed later).

Inner loop

- Warm start with the previous dual iterates.
- Stopping rules: $\|X_n^{(k,j)}\mathbf{1} - \boldsymbol{\varrho}\|_\infty \leq 10^{-6}$ or $j \geq j_{\max} = 20$.

Running environment

- CPU: Intel Xeon Gold 6242R CPU @ 3.10GHz.
- RAM: 510GB.
- Operating system: Ubuntu 20.04.
- Software: MATLAB R2019b.

Metrics of Interest

(i) Converged objective value (obj).

(ii) Approximate SCE potential $\mathbf{v}$. [Chen et al. 2014; H. et al. 2023]

Taking ERALM *for example,* $\mathbf{v} := \tilde{\mathbf{v}} - \min_{k=1}^{K}\{\tilde{v}_k\} \cdot \mathbf{1} \in \mathbb{R}^K$*, where*

$$\tilde{\mathbf{v}} := \frac{1}{N-1}\sum_{n=2}^{N} \tilde{\mathbf{v}}_n \in \mathbb{R}^K$$

and $\{\tilde{\mathbf{v}}_n\}_{n=2}^{N}$ *are the dual solutions given by dual BCD.*

(iii) Errors of objective value (err_obj) and SCE potential (err_sce).[1)]

$$\text{err\_obj} := \left|\frac{f - f^\star}{f^\star}\right|, \quad \text{err\_sce} := \frac{\|\mathbf{v} - \mathbf{v}^\star\|_1}{K},$$

where f *and* $f^\star \in \mathbb{R}$ *denote respectively the converged and optimal objective values of* (P)*,* $\mathbf{v}^\star \in \mathbb{R}^K$ *refers to the vector made up by the values of the SCE potential at barycenters.*

(iv) CPU time in seconds (T).

[1)] Available only if there admit explicit constructions of the optimal solutions and the SCE potentials, e.g., in 1D settings [Colombo-De Pascale-Di Marino 2015; H.-Liu 2022]

Model Systems

Table 2: Information on the model systems.

System No.	Unnormalized single-particle densities ρ	Domains Ω	#electrons N
	1D systems		
1	$\cos(\pi x) + 1$	$[-1, 1]$	3
2	$2e^{-6(x+0.5)^2} + 1.5e^{-4(x-0.5)^2}$	$[-1.5, 1.5]$	3
3	$e^{-x^2}/\sqrt{\pi}$	$[-2, 2]$	7
4	$e^{-4(x+2)^2} + e^{-4(x+1.5)^2} + e^{-4(x+1)^2} + e^{-4(x+0.5)^2}$ $+e^{-4(x-2/3)^2} + e^{-4(x-4/3)^2} + e^{-4(x-2)^2}$	$[-3, 3]$	7
	2D systems		
5	$e^{-3(x^2+(y-0.96)^2)} + e^{-3((x-1.032)^2+(y+0.84)^2)}$ $+e^{-3((x+1.032)^2+(y+0.84)^2)}$	$[-2.5, 2.5]^2$	3
6	$2e^{-3(x^2+(y-1.2)^2)} + e^{-3((x-1.29)^2+(y+1.05)^2)}$ $+e^{-3((x+1.29)^2+(y+1.05)^2)}$	$[-3, 3]^2$	4
	3D systems		
7	$e^{-3((x+1)^2+(y+1)^2+(z+1)^2)} + e^{-3((x-1)^2+(y-1)^2+(z+1)^2)}$ $+e^{-3((x+1)^2+(y-1)^2+(z-1)^2)}$	$[-2, 2]^3$	3
8[2)]	$3e^{-4((x+1)^2+y^2+z^2)} + e^{-4((x-1)^2+y^2+z^2)}$	$[-2, 2] \times [-1, 1]^2$	4

[2)] System 8 can describe a dissociating lithium hydride (LiH) [Filatov 2015].

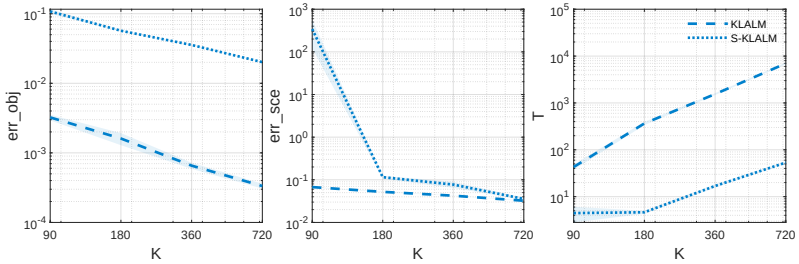
A Numerical Observation

Settings

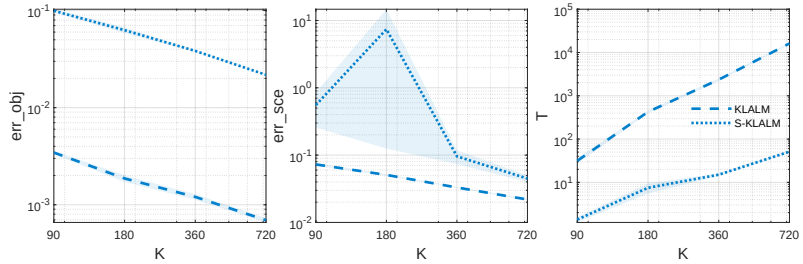
- Systems 1-4, equi-mass discretization $\Rightarrow$ closed-form optimal solutions.
  - Systems 1 and 2: $K = 90, 180, 360, 720$ ($K_0 := 90$).
  - Systems 3 and 4: $K = 140, 280, 560, 1120$ ($K_0 := 140$).
- For each K, invoke KLALM, S-KLALM with 10 trials from random initial points.
- Stopping rules[3] : $tol = 10^{-3} \cdot \sqrt{2}^{\log_2(K/K_0)}$, $k_{\max} = +\infty$, $T_{\max} = 10^5$.

[3] We increase tol with K because the diameter of the feasible region in (**NQP**) grows as $\mathcal{O}(\sqrt{K})$.

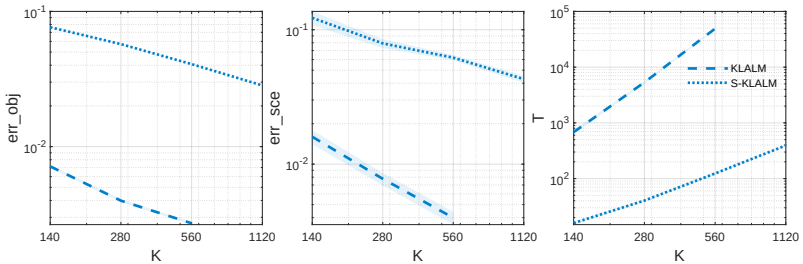
A Numerical Observation (Cont.)



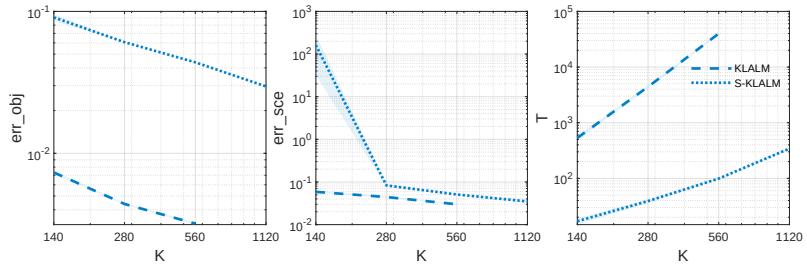
(a) System 1 ($N = 3$)



(b) System 2 ($N = 3$)



(c) System 3 ($N = 7$)



(d) System 4 ($N = 7$)

Fig. 1. $\mathrm{err\_obj}$, $\mathrm{err\_sce}$, and T of KLALM, S-KLALM on the 1D systems with different K.

(i) A local solver + a moderate K + random initialization $\Rightarrow$ relatively low $\mathrm{err\_obj}$.

(ii) Average $\mathrm{err\_obj}$, $\mathrm{err\_sce}$ and their standard errors $\downarrow$ as $K \uparrow$.

(iii) Sampling-based algorithm is more advantageous on T yet with worse accuracy.

A Numerical Observation (Cont.)

(i) A local solver + a moderate K + random initialization $\Rightarrow$ relatively low $\mathrm{err\_obj}$.

(ii) Average $\mathrm{err\_obj}$, $\mathrm{err\_sce}$ and their standard errors $\downarrow$ as $K \uparrow$.

(iii) Sampling-based algorithm is more advantageous on T yet with worse accuracy.

Numerical implications

- As $K \uparrow$ in (P), $\{$the objective values at the stationary points$\} \rightarrow$ the optimal value.
  $\Rightarrow$ "a local solver + random initialization" is **"*nearly*" enough** for globally solving (P).
- It is better to **combine** the merits of the algorithms with and without sampling.
  $\Rightarrow$ a **grid refinements-based framework** for large-scale settings. [H. et al. 2023]

A Grid Refinements-Based Framework for Global Optimization

Similar to the one in [H. et al. 2023].

Algorithm 5: Grid refinements (GR)-based framework.

Input: Discretization oracle, refinement oracle, local solver, initial number of finite elements $K^{(0)} \in \mathbb{N}$.

Set $\ell := 0$;
while *certain conditions not satisfied* **do**
  if $\ell = 0$ **then**
    Discretization: discretize the Monge-like formulation into (P) with $K^{(0)}$ finite elements $\{e_k^{(0)}\}_{k=1}^{K^{(0)}} \subseteq \mathbb{R}^d$;
    Construct a random initial point $X^{(0,0)} \in \left(\mathbb{R}^{K^{(0)} \times K^{(0)}}\right)^{N-1}$;
  else
    Grid refinement: refine the last mesh to obtain $\{e_k^{(\ell)}\}_{k=1}^{K^{(\ell)}} \subseteq \mathbb{R}^d$ with $K^{(\ell)} \in \mathbb{N} : K^{(\ell)} \geq K^{(\ell-1)}$;
    Grid refinements-based initialization: construct an initial point $X^{(\ell,0)} \in \left(\mathbb{R}^{K^{(\ell)} \times K^{(\ell)}}\right)^{N-1}$ based upon $X^{(\ell-1,\star)}$ as well as the relation between $\{e_k^{(\ell)}\}_{k=1}^{K^{(\ell)}}$ and $\{e_k^{(\ell-1)}\}_{k=1}^{K^{(\ell-1)}}$;
  end
  Local solution: start the local solver from $X^{(\ell,0)}$ for (P) and obtain $X^{(\ell,\star)}$;
  Set $\ell := \ell + 1$;
end

Output: Approximate solution $X^{(\ell,\star)}$.

Empirically, $K^{(0)} \sim \mathcal{O}(10^d)$.

Illustration of the GR-Based Initialization

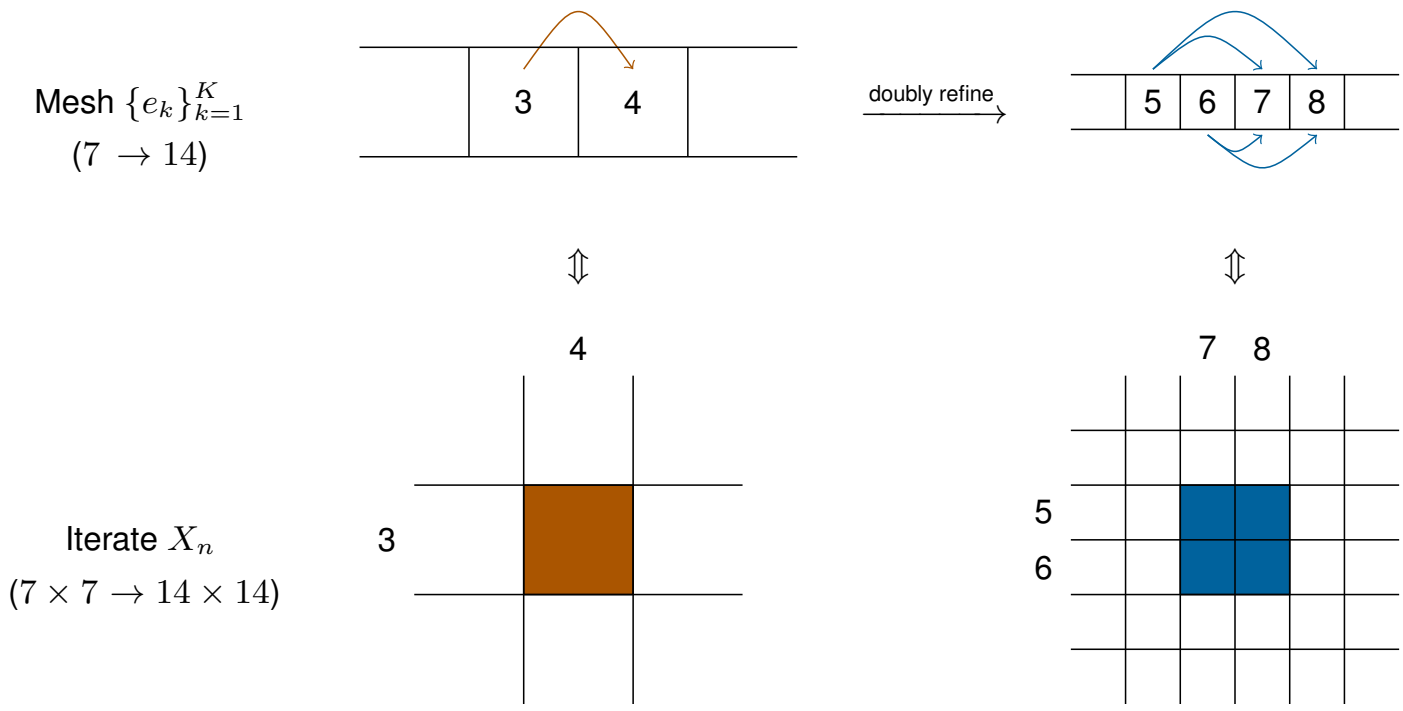


Fig. 2. GR-based initialization in 1D context (from $K^{(\ell)} = 7$ to $K^{(\ell+1)} = 14$).

- Standing at the OT backgrounds. [H. et al. 2023]
- Providing both **warm start and sampling** for S-ERALM, S-KLALM.

Effect of the GR-Based Framework

(S-)KLALM-GR = (S-)KLALM + GR-based framework ($\ell = 0$: KLALM; $\ell > 0$: S-KLALM).

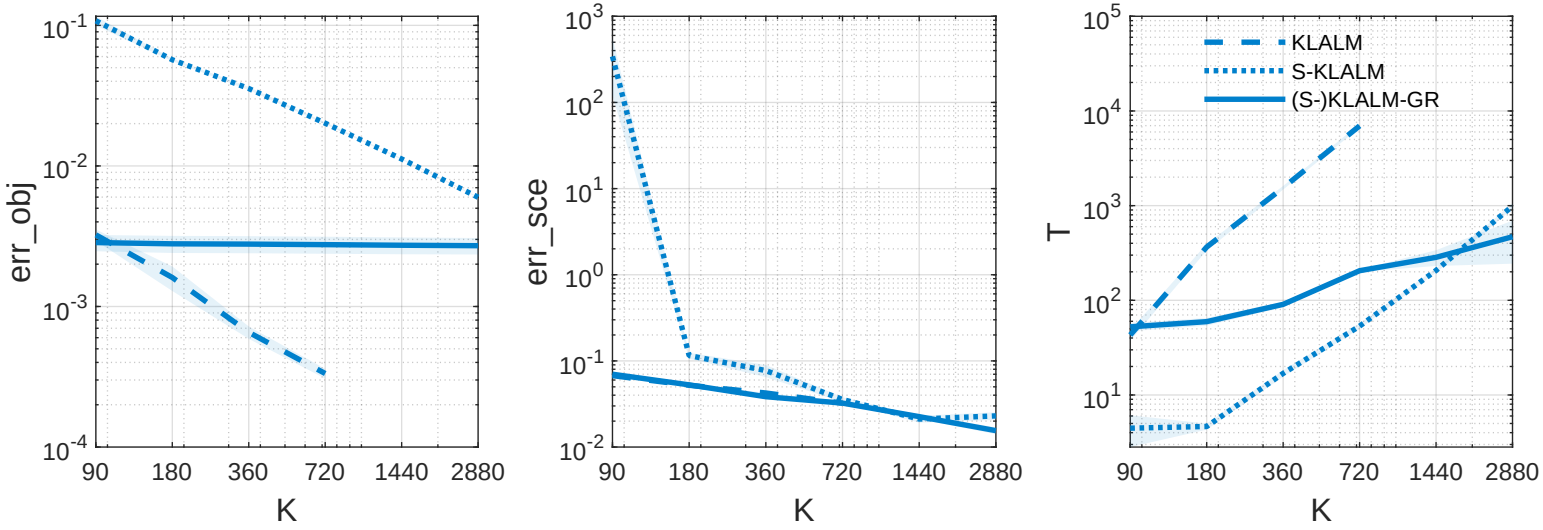


Fig. 3. $\mathrm{err\_obj}$, $\mathrm{err\_sce}$, and T of KLALM, S-KLALM, and (S-)KLALM-GR on system 1 with different K.

GR-based framework is well suited for obtaining solutions of **relatively high accuracy within short CPU time in large-scale contexts**.

Comparison Among Algorithms

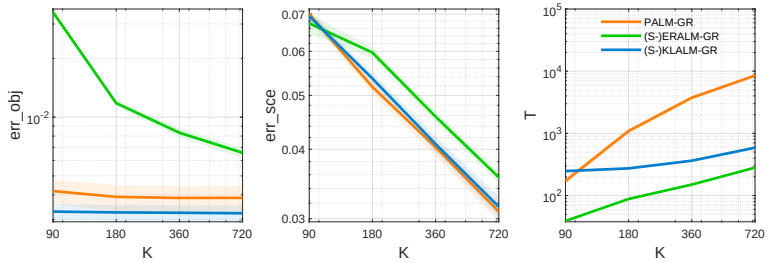
Algorithms

- **PALM-GR** = PALM + GR-based framework.
  the implementation of PALM follows [H. et al. 2023].
- **(S-)ERALM-GR** = (S-)ERALM + GR-based framework.
  $\ell = 0$: ERALM; $\ell > 0$: S-ERALM.
- **(S-)KLALM-GR** = (S-)KLALM + GR-based framework.
  $\ell = 0$: KLALM; $\ell > 0$: S-KLALM.

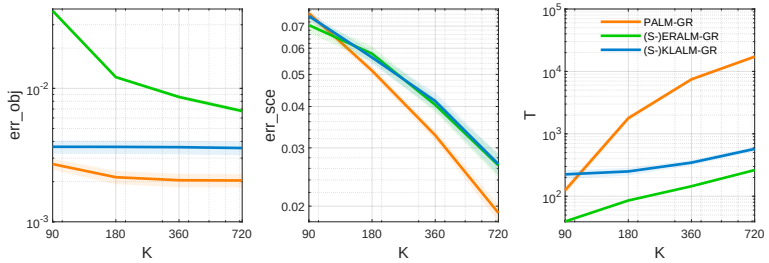
Settings

- Systems 1-4: equi-mass discretization ($\ell = 0$) / refinements ($\ell > 0$).
  - Systems 1 and 2: $K^{(0)} = 90 \xrightarrow[\text{grid refinements}]{\text{3 times}} K^{(3)} = 720.$
  - Systems 3 and 4: $K^{(0)} = 140 \xrightarrow[\text{grid refinements}]{\text{3 times}} K^{(3)} = 1120.$
- For each system, invoke the three algorithms with 10 trials.
- Stopping rules: $tol = 10^{-3} \cdot \sqrt{2}^{\log_2(K/K^{(0)})}$, $k_{\max} = 10^4$, $T_{\max} = 10^5$.

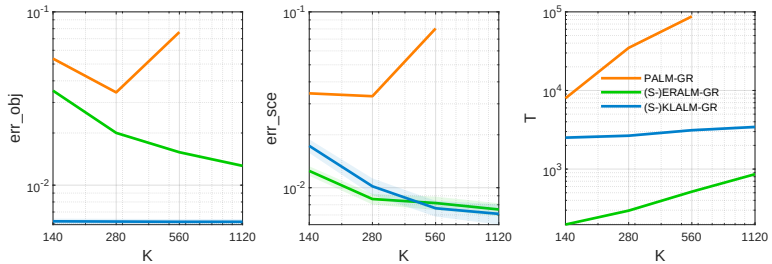
Comparison Among Algorithms (Cont.)



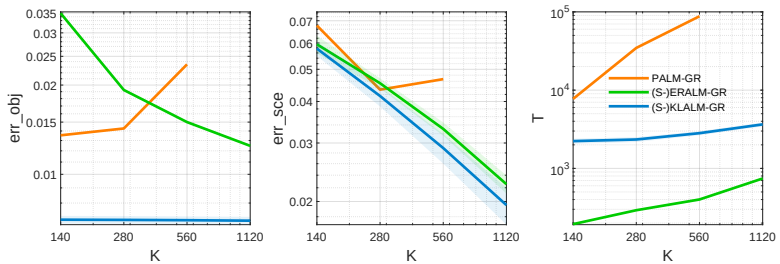
(a) System 1 ($N = 3$)



(b) System 2 ($N = 3$)



(c) System 3 ($N = 7$)



(d) System 4 ($N = 7$)

Fig. 4. $\mathrm{err\_obj}$, $\mathrm{err\_sce}$, and T of PALM-GR, (S-)ERALM-GR, and (S-)KLALM-GR on the 1D systems with different K.

(S-)KLALM-GR stands out with

- comparable (or even better) $\mathrm{err\_obj}$ and $\mathrm{err\_sce}$ to (than) PALM-GR;
- comparable $\mathrm{err\_sce}$ and T to (S-)ERALM-GR.

Settings

- Systems 5-8: equi-size discretization ($\ell = 0$) / uniform refinements ($\ell > 0$).
  - Systems 5 and 6 (2D): $K^{(0)} = 900 \xrightarrow[\text{grid refinements}]{\text{3 times}} K^{(3)} = 57600.$
  - System 7 (3D): $K^{(0)} = 1728 \xrightarrow[\text{grid refinements}]{\text{2 times}} K^{(2)} = 110592.$
  - System 8 (3D): $K^{(0)} = 1000 \xrightarrow[\text{grid refinements}]{\text{2 times}} K^{(2)} = 64000.$
- Stopping rules: $k_{\max} = 10^4$, $T_{\max} = +\infty$,

$$tol = \begin{cases} 5 \times 10^{-3}, & \ell = 0, \\ 10^{-2} \times (\sqrt{2^d})^{\log_2(K/K^{(0)})}, & \ell > 0. \end{cases}$$

- Mappings between electron positions $\{\mathcal{T}_n\}_{n=2}^N$.

$$\mathcal{T}_n(\mathbf{a}_i) := \sum_{1 \le j \le K} x_{n,ij} \mathbf{a}_j / \varrho_i, \ i = 1, \ldots, K, \ n = 2, \ldots, N.$$

Simulations on 2D/3D Systems (Cont.)

Table 3: obj given by (S-)KLALM-GR when simulating the 2D/3D systems. K_{trunc} refers to #entries in $\boldsymbol{\varrho}$ that are larger than $0.999\|\boldsymbol{\varrho}\|_\infty$.

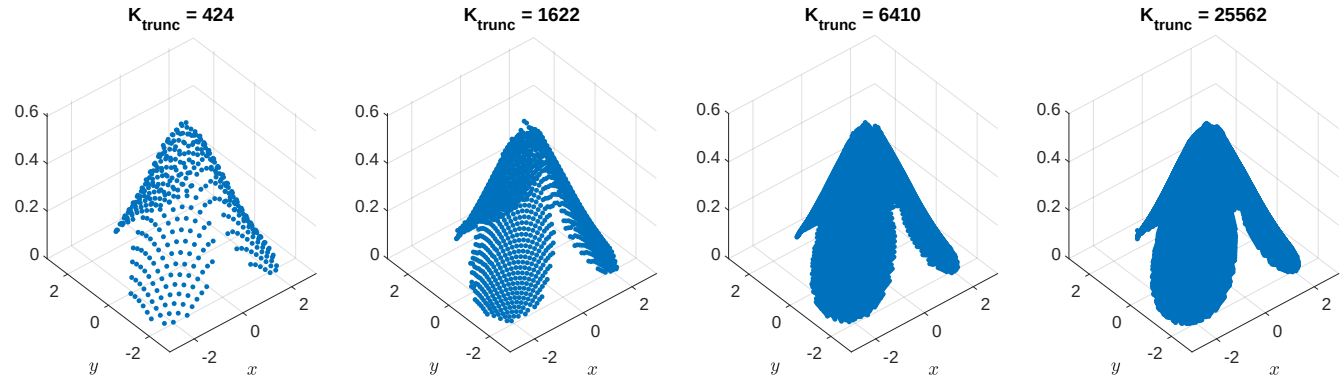
Step	System 5			System 6		
	K	K_{trunc}	obj	K	K_{trunc}	obj
0	900	424	1.1339	900	408	3.0690
1	3600	1622	1.1337	3600	1534	3.0690
2	14400	6410	1.1335	14400	6068	3.0677
3	57600	25562	1.1334	57600	24176	3.0667

(a) 2D systems

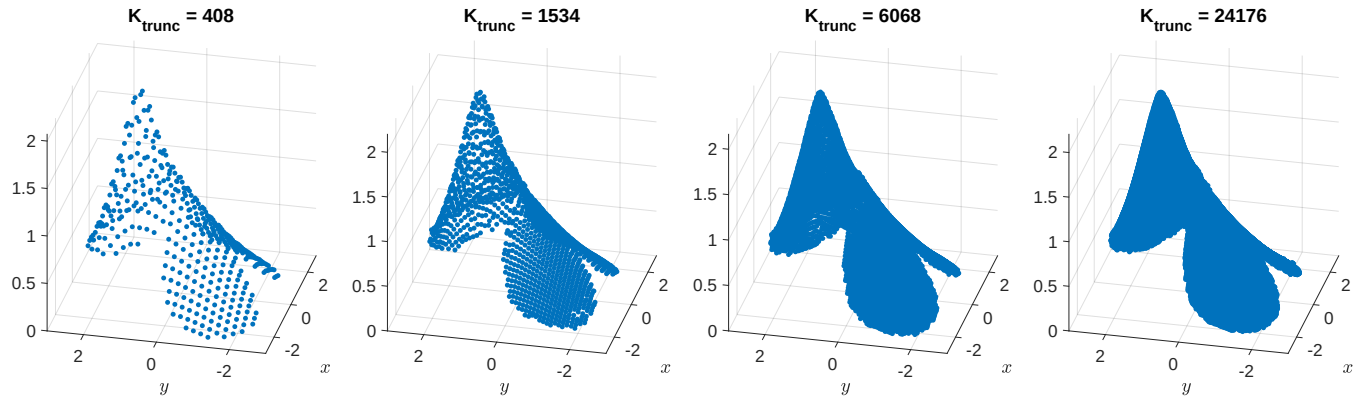
Step	System 7			System 8		
	K	K_{trunc}	obj	K	K_{trunc}	obj
0	1728	780	1.0202	1000	720	4.6193
1	13824	5628	1.0209	8000	5272	4.6716
2	110592	42936	1.0209	64000	40764	4.6833

(b) 3D systems

Simulations on 2D/3D Systems (Cont.)



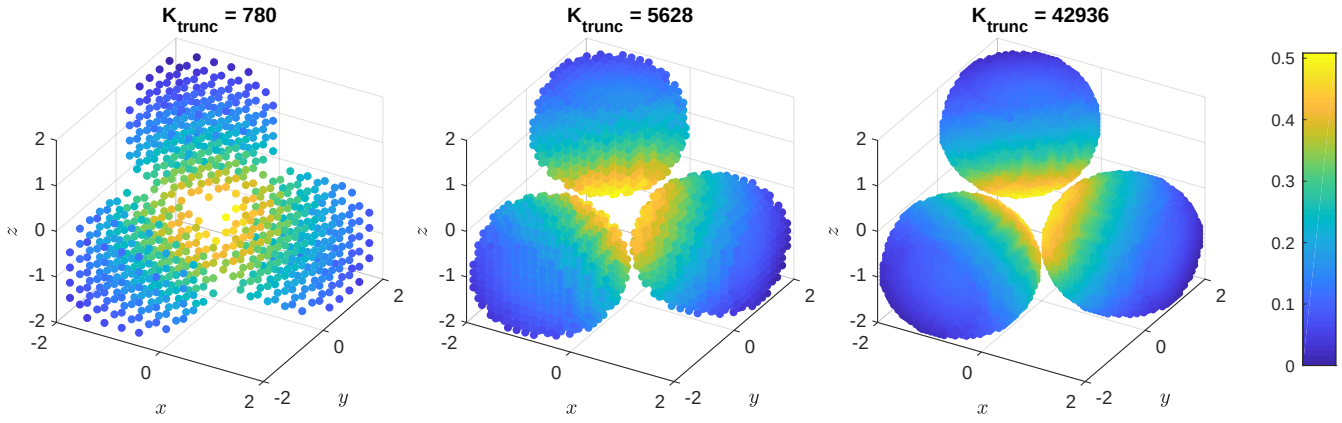
(a) System 5



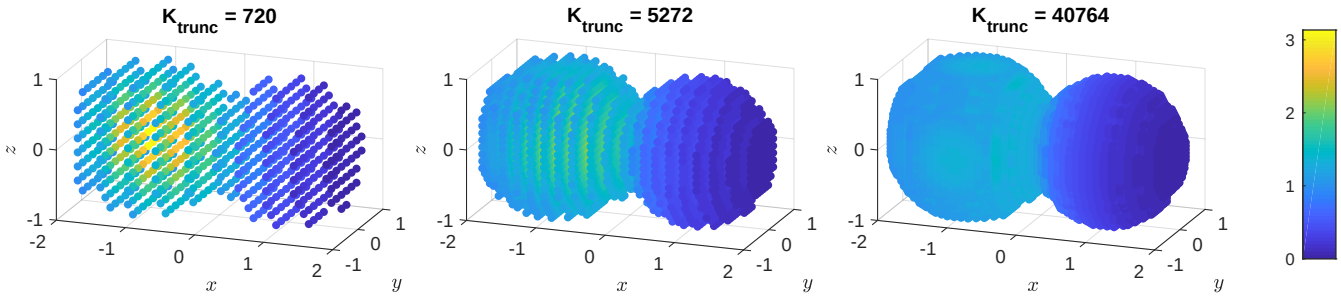
(b) System 6

Fig. 5. Approximate SCE potentials given by (S-)KLALM-GR on the 2D systems.

Simulations on 2D/3D Systems (Cont.)



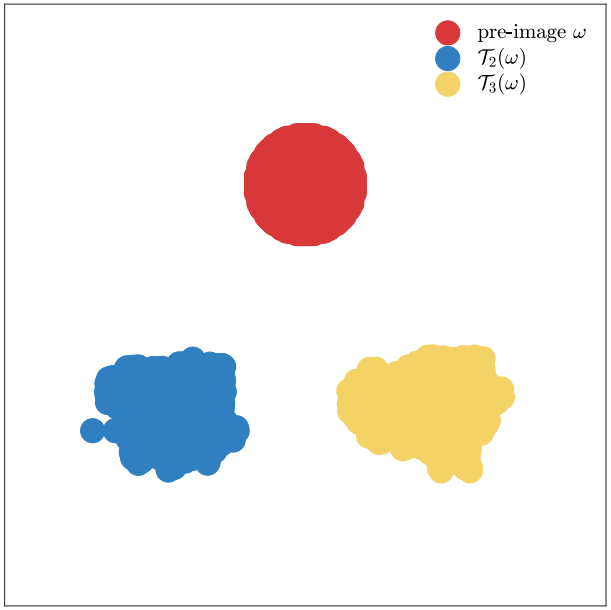
(a) System 7



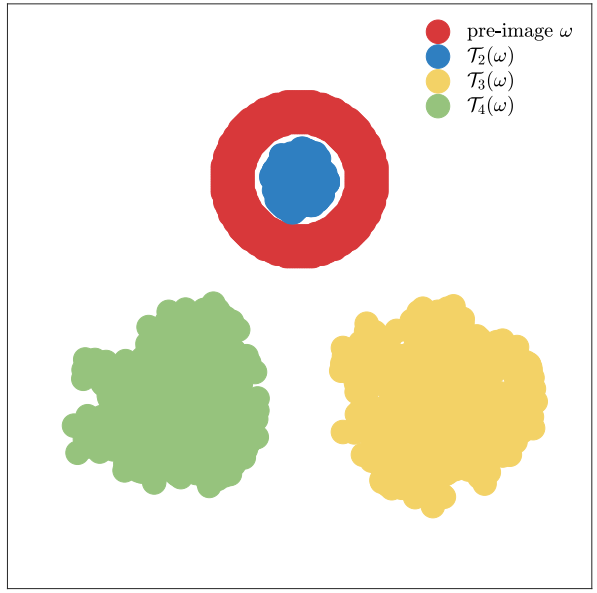
(b) System 8

Fig. 6. Approximate SCE potentials given by (S-)KLALM-GR on the 3D systems.

Simulations on 2D/3D Systems (Cont.)



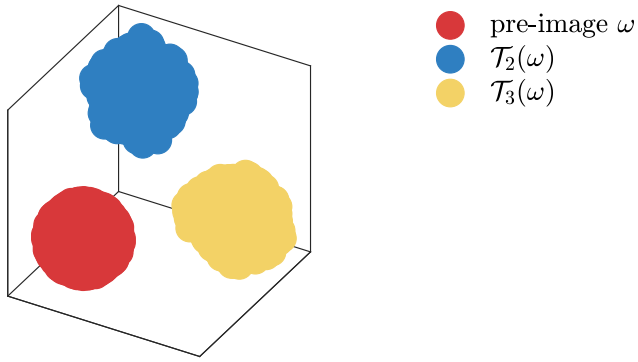
(a) System 5



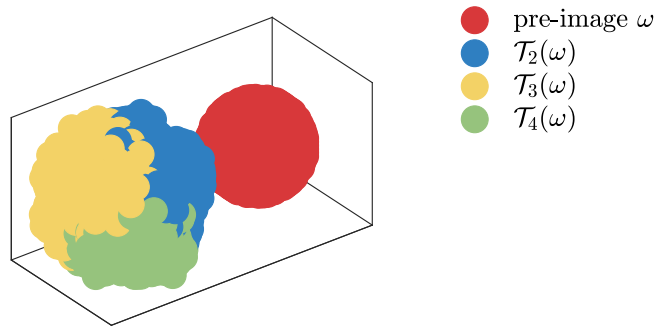
(b) System 6

Fig. 7. Mappings between electron positions given by (S-)KLALM-GR on the 2D systems.

Simulations on 2D/3D Systems (Cont.)



(a) System 7



(b) System 8

Fig. 8. Mappings between electron positions given by (S-)KLALM-GR on the 3D systems.

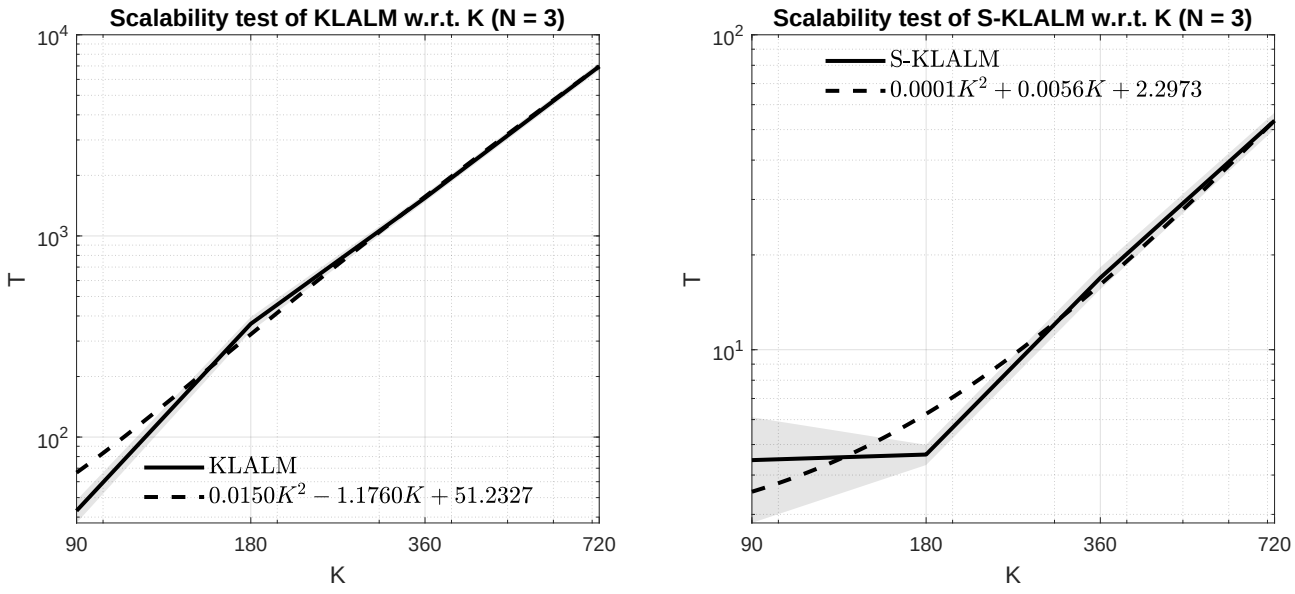
First visualization in 3D contexts.

Scalability Tests

Settings

- $\rho(x) \propto \cos(\pi x) + 1$, equi-mass discretization.
- Algorithms: KLALM and S-KLALM.
- **Scalability w.r.t. K:**
  - $K = 90, 180, 360, 720, 1440, 2880$, fix $N = 3$.
  - Stopping rules: $tol = 10^{-3} \times \sqrt{2}^{\log_2(K/90)}$, $k_{\max} = 10^4$, $T_{\max} = +\infty$.
- **Scalability w.r.t. N:**
  - $N = 3, 6, 12, 24, 48$, fix $K = 144$.
  - Stopping rules: $tol = 10^{-3}$, $k_{\max} = 10^4$, $T_{\max} = +\infty$.
- Each K or N corresponds to 10 random trials.

Scalability Tests w.r.t. K

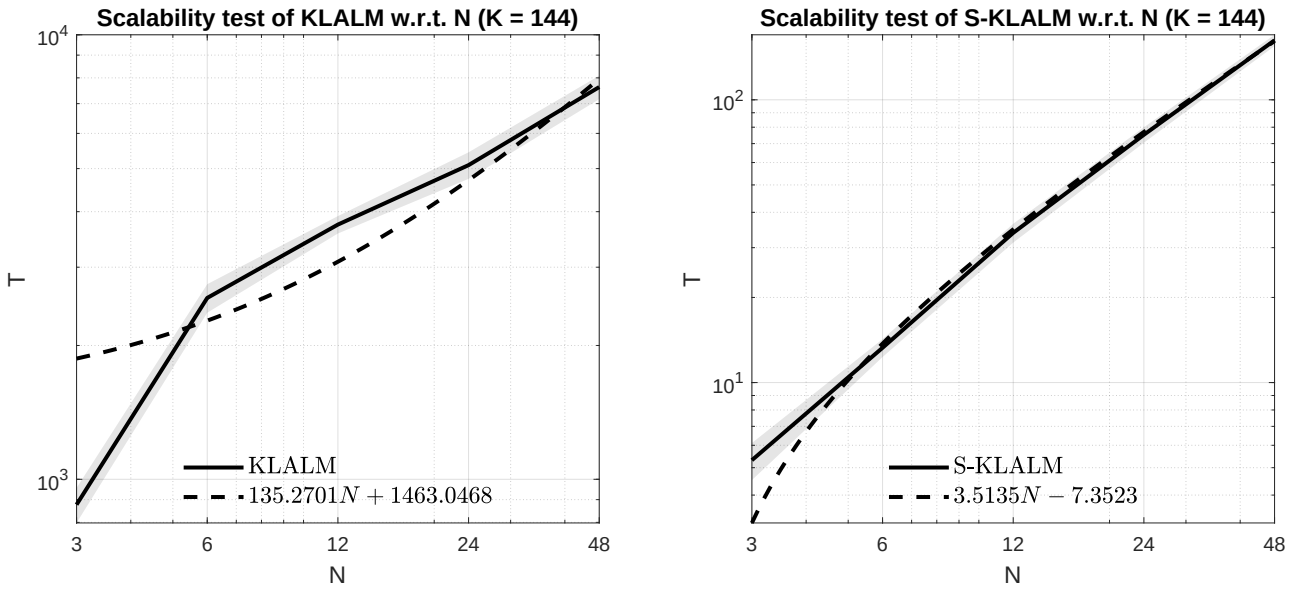


(a) Scalability of KLALM w.r.t. K (b) Scalability of S-KLALM w.r.t. K

Fig. 9. Scalability tests of KLALM and S-KLALM w.r.t. K.

Scalability w.r.t. $K \sim \mathcal{O}(K^2)$.

Scalability Tests w.r.t. N



(a) Scalability of KLALM w.r.t. N (b) Scalability of S-KLALM w.r.t. N

Fig. 10. Scalability tests of KLALM and S-KLALM w.r.t. N.

Scalability w.r.t. $N \sim \mathcal{O}(N)$.

Outline

Conclusions

(i) Sampling-based splitting methods for the MMOT: S-ERALM and S-KLALM.
  - Highly scalable schemes for subproblems.
  - No full matrix multiplications.

(ii) Convergence and asymptotic properties for ERALM and S-ERALM.

(iii) Numerical simulations on several model 1D/2D/3D systems.
  - Combination with the GR-based framework.
  - Much better scalabilities of S-ERALM and S-KLALM than PALM.
  - First visualization of the mappings between electron positions in 3D contexts.

Future Work

- Convergence and asymptotic properties for KLALM and S-KLALM.
- Solution landscape of (**NQP**) and its variations w.r.t. K.
- Relation between the dual solutions of (**NQP**) and the SCE potential.
- Support identification for further acceleration.

References

- **Y. Hu**, H. Chen, and X. Liu. A global optimization approach for multimarginal optimal transport problems with Coulomb cost. ***SIAM Journal on Scientific Computing***, 2023, accepted.
- **Y. Hu** and X. Liu. The exactness of the ℓ_1 penalty function for a class of mathematical programs with generalized complementarity constraints. ***Fundamental Research***, 2023, doi:10.1016/j.fmre.2023.04.006.
- **Y. Hu** and X. Liu. The convergence property of infeasible inexact proximal alternating linearized minimization. ***Science China Mathematics***, 2023, doi:10.1007/s11425-022-2074-7.
- X. Liu. Optimization models and approaches for strongly correlated electrons systems (in Chinese). ***Mathematica Numerica Sinica***, 2023, 45(2): 141–159.

Thanks for your attentions!

Email: ykhu@lsec.cc.ac.cn