The Mixing method: low-rank coordinate descent for diag SDPs

Carnegie Mellon University

Po-Wei Wang [1], Wei-Cheng Chang [2], and J. Zico Kolter [3]

[1]MLD, CMU, [2]LTI, CMU, [3]CSD, CMU & Bosch Center for AI

MACHINE LEARNING DEPARTMENT

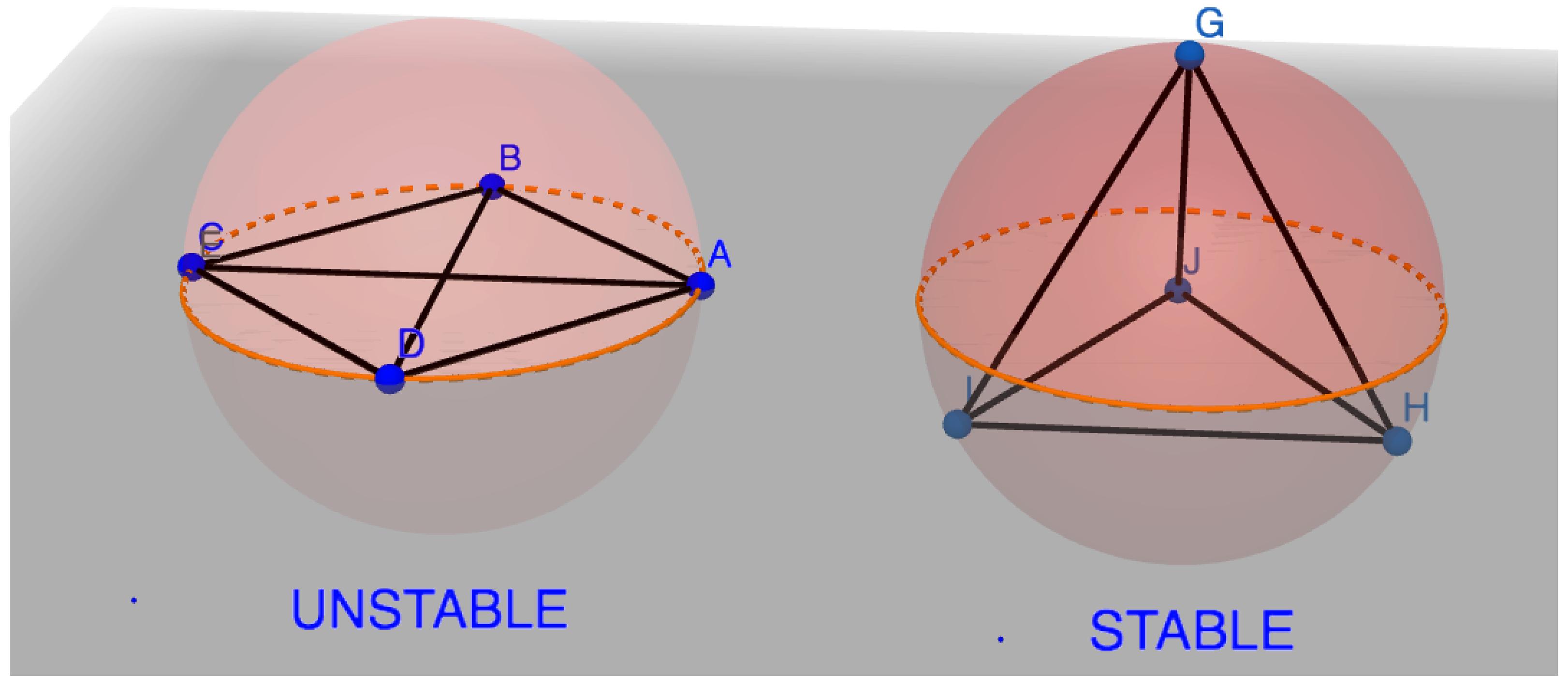
Summary

The paper presents **the first assumptionless proof** that a low-rank SDP method converges to a global optimum for constrained SDPs despite its non-convexity. The experiments suggest the proposed method is **10∼100x faster** than the state-of-the-art methods.

Problem

$$\underset{V\in\mathbb{R}^{k\times n}}{\text{maximize}} \sum_{ij} c_{ij}\|v_i - v_j\|^2 \text{ s.t. } \|v_i\| = 1,\ \forall i = 1\dots n.$$

Intuitively, maximize the sum of weighted distances between vertices. The problem is non-convex, and zero gradient (critical pt) $\neq$ optimum.



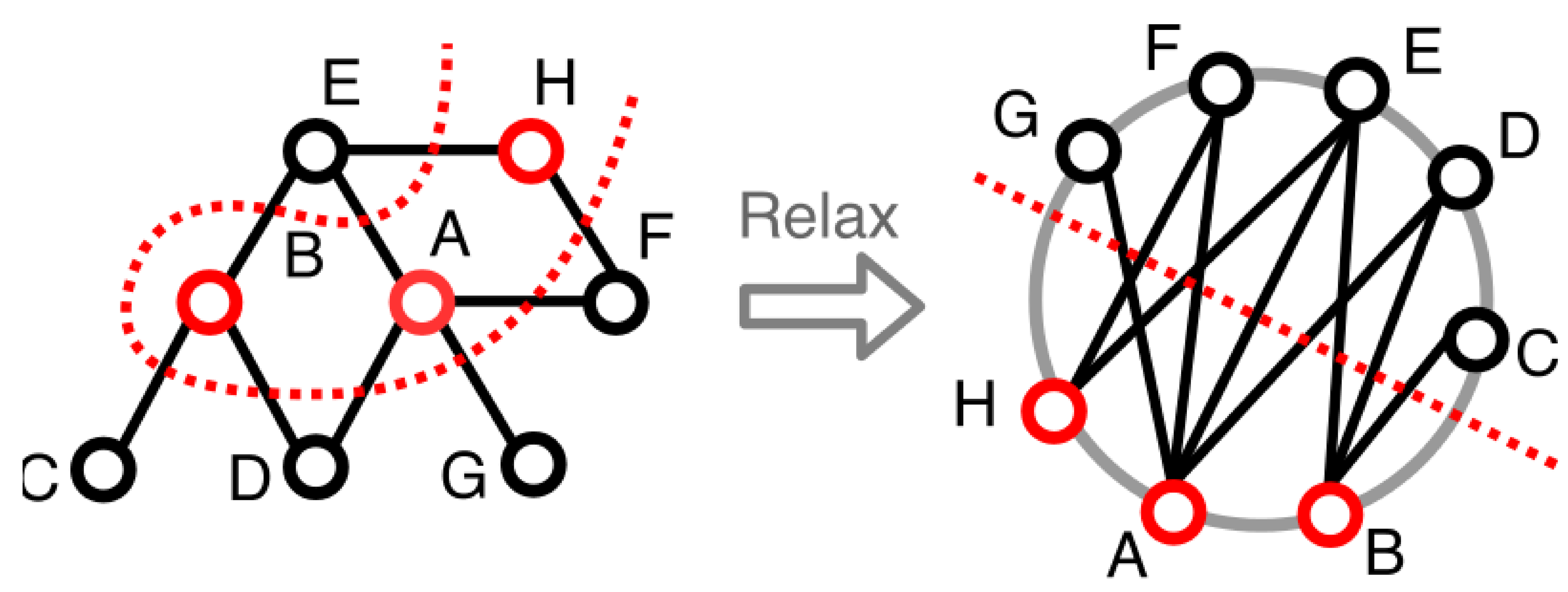
Goal: Can we optimize on the manifold and still reach global optimum?

Application

- **Diag constrained SDP (low-rank $X = V^TV \iff X \succeq 0$)**

$$\underset{V\in\mathbb{R}^{k\times n}}{\text{maximize}} \sum_{ij} c_{ij}\|v_i - v_j\|^2 \text{ s.t. } \|v_i\| = 1 \iff \underset{X\succeq 0}{\text{minimize}} \langle C, X\rangle \text{ s.t. } X_{ii} = 1$$

- **MAXCUT (Goesman-Williamson SDP relaxation)**



- **MAXSAT (Minimize convexified loss)**

The Mixing methods

Main ideas:

- **Randomize init** $\Rightarrow$ unlikely to reach unstable criticals.
- **BCD**: Move one vertices v_i at a time $\Rightarrow$ closed-form solution.

$$\underset{v_i\in\mathbb{R}^k}{\text{minimize}}\ g_i^T v_i, \text{ s.t. } \|v_i\| = 1 \ \Rightarrow\ v_i = -g_i/\|g_i\|.$$

The Mixing method = Mix and normalize neighbors $\forall v_i$.

- Initialize v_i randomly on a unit sphere (e.g. normalized uniform).
- **While** not yet converge **:**
- **For** $i = 1\dots n$ **:**
- $v_i := -g_i/\|g_i\|$, where $g_i = \sum_{j=1}^n c_{ij}v_j$;

Adding step size (technical):

- Instead of exact BCD, use $v_i := (v_i - \theta g_i)/\|v_i - \theta g_i\|$.
- **Avoid degeneracy** ($\|g_i\| = 0$), which never happens in practice.

Convergence analysis:

- Low-rank = low-memory complexity ($V \in \mathbb{R}^{k\times n}$ v.s. $X \in \mathbb{R}^{n\times n}$).
- Converge to global opt: observed in exps, **open prob for 17 yrs**.
- **Difficulties**:
  - ∗ Non-convex (spherical manifold dom), rotational equivalence.
  - ∗ Random initialization required.
  - ∗ Singularity of Jacobian and Hessian.

Lyapunov instability and stable manifolds

Instability: the operator has **expansive** direction $\forall$ non-opt criticals.

- Eigenvalues of Jacobian on manifold contain those in Euclidean

$$\text{Eigvals}(A \otimes I_k \operatorname{diag}(I - v_iv_i^T)B \otimes I_k) \supseteq \text{Eigvals}(AB) \text{ for } k > \sqrt{2n}.$$

- Jacobian of Gauss-Seidel is unstable when not PSD ($|\lambda_i| > 1$).
- All non-optimal criticals corresponds to a G.-S. on non-PSD system.
- Thus, the Mixing methods are unstable on non-optimal criticals.

Center-stable manifold thm: Existence of **invariant manifolds**.

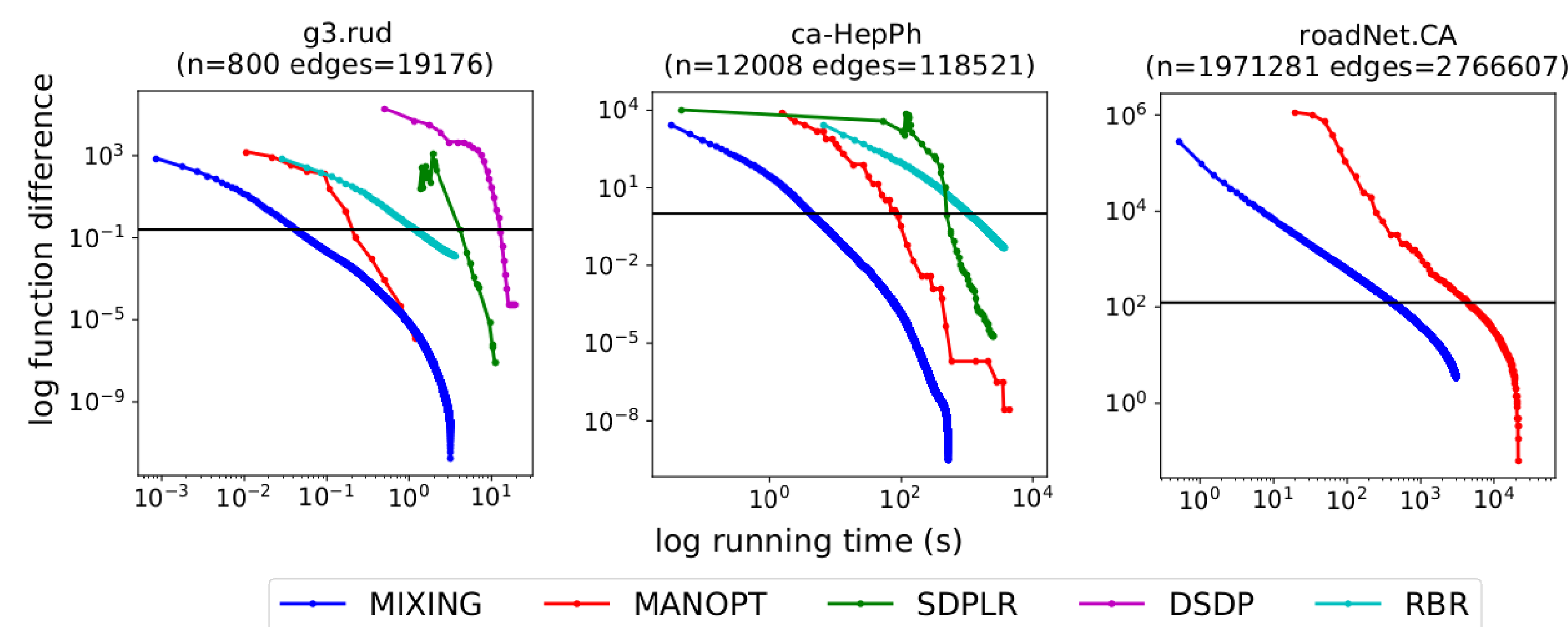
- Mixing method with step-size is a diffeomorphism (1-to-1 and $\mathcal{C}^1$).
- Apply center-stable manifold thm: basin to unstable is 0 measure.
- That is, random initialization never converges to unstable criticals.
- Converge to critical and never to unstable $\Rightarrow$ converge to global opt.

Theoretical results

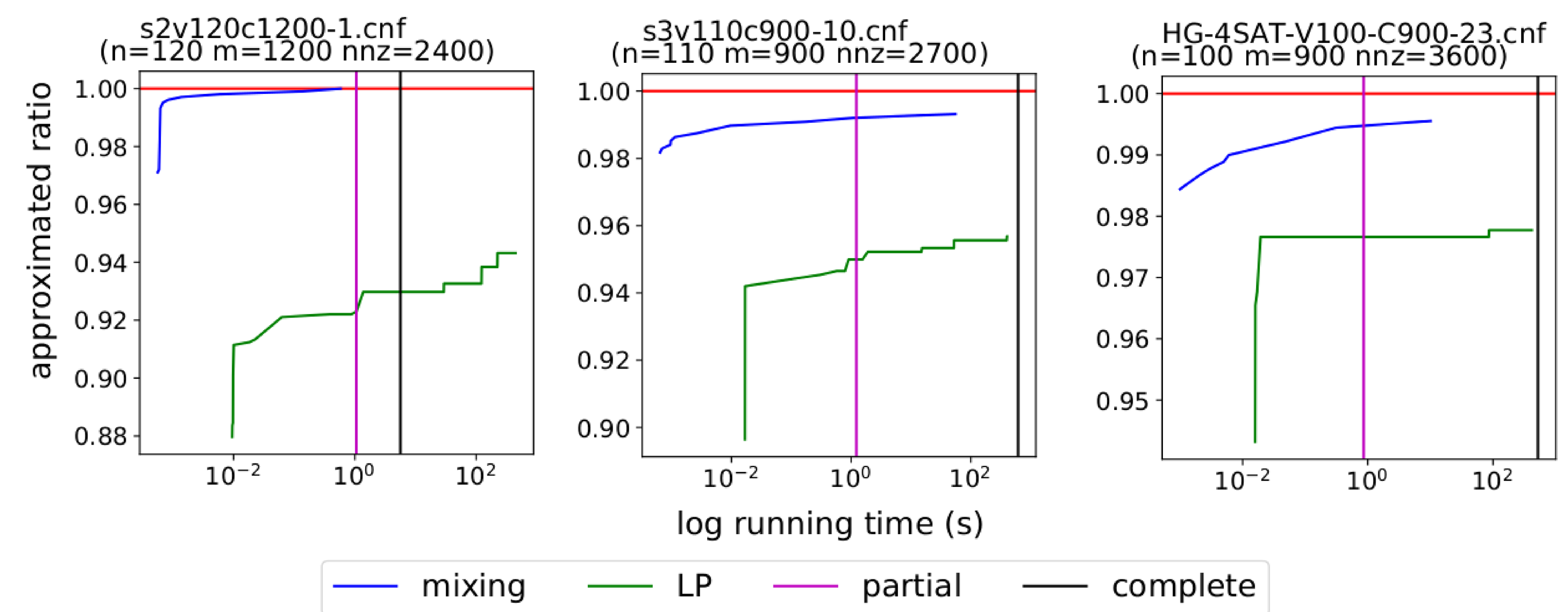
- **(All non-optimals are unstable)** Pick $k > \sqrt{2n}$. For the Mixing method with step size (no assumption) or without step size (nondegeneracy assumption) all non-optimal first-order critical points are unstable fixed points for almost all C.
- **(Convergence to global optimum)** Take $k > \sqrt{2n}$ and $\theta \in (0, \frac{1}{\max_i \|c_i\|_1})$. Then for almost every C, the Mixing method with a step size converges a.s. to a global optimum under random init.
- **(Local linear rate)** The Mixing methods converge linearly to the global optimum when close enough, with step size (no assumption) or without step size (nondegeneracy) $\Rightarrow$ Overall $O(m\sqrt{n}\log(1/\epsilon))$.

Experiments

- MAXCUT SDP (10∼100x faster than state-of-the-art, million vars)



- MAXSAT approximation ratio (avg 0.978, faster & better than LP)



[1] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[2] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[3] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[4] M. Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.