Low rank of the matrix LASSO under RIP with consequences for fast large-scale algorithms

Andrew D. McRae

Institute of Mathematics, EPFL

EUROPT, June 26, 2024

Low-rank matrix recovery by LASSO

Problem: estimate low-rank matrix $M_* \in \mathbf{R}^{d_1 \times d_2}$ from

$$y = \mathcal{A}(M_*) + \xi,$$

- $\mathcal{A} \colon \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^n$ known linear measurement operator
- ξ represents noise/error

Matrix LASSO estimate (Rohde and Tsybakov, 2011; Candès and Plan, 2011; Negahban and Wainwright, 2011):

$$\widehat{M} = \operatorname*{arg\,min}_{M \in \mathbf{R}^{d_1 \times d_2}} \frac{1}{2} \|y - \mathcal{A}(M)\|^2 + \lambda \|M\|_*$$

Nuclear norm penalty promotes low solution rank

- Matches (presumed) structure of ground truth M_*

Benefits of low rank at large scale

Difficulty to estimate/use rank-r $M \in \mathbf{R}^{d_1 \times d_2}$ scales with #DOF $\approx r(d_1 + d_2)$:

- Statistics (#measurements and error)
- Algorithms (number of variables **if** we optimize directly over rank-r matrices)
- Storage/multiplication cost

Why the LASSO?

Estimate of low-rank M_*:

$$\widehat{M} = \underset{M \in \mathbf{R}^{d_1 \times d_2}}{\arg\min} \ \frac{1}{2}\|y - \mathcal{A}(M)\|^2 + \lambda \|M\|_* \text{ where } y = \mathcal{A}(M_*) + \xi$$

Convex and has (provably) great **statistical** properties, **but...**

- Optimization over **full-rank** matrices
- Direct solvers **scale poorly**
- Few theoretical guarantees that $\widehat{M}$ has low rank (potentially **costly storage**)

LASSO rank bound

$$\widehat{M} = \arg\min_{M \in \mathbf{R}^{d_1 \times d_2}} \frac{1}{2}\|y - \mathcal{A}(M)\|^2 + \lambda\|M\|_* \text{ where } y = \mathcal{A}(M_*) + \xi \qquad \text{(LASSO)}$$

Key requirement: $\mathcal{A}$ has (r, δ) restricted isometry property (RIP) if

$$(1-\delta)\|M\|_F^2 \le \|\mathcal{A}(M)\|^2 \le (1+\delta)\|M\|_F^2 \quad \text{whenever} \quad \operatorname{rank}(M) \le r \qquad \text{(RIP)}$$

Theorem

Suppose

- $\operatorname{rank}(M_*) = r_*$
- $\mathcal{A}$ *has* $(2r_*, \delta)$ *RIP for sufficiently small* $\delta > 0$
- $\|\mathcal{A}^*(\xi)\|_{\mathrm{op}} \lesssim \lambda$

Then the LASSO solution $\widehat{M}$ *is unique, and*

$$\operatorname{rank}(\widehat{M}) \le \left(1 + c\left[\delta + \frac{\|\mathcal{A}^*(\xi)\|_{\mathrm{op}}}{\lambda}\right]^2\right) r_* \le 1.1 r_*$$

Rank bound context

Theorem

Suppose

- $\operatorname{rank}(M_*) = r_*$
- $\mathcal{A}$ *has* $(2r_*, \delta)$ ***RIP*** *for sufficiently small* $\delta > 0$
- $\|\mathcal{A}^*(\xi)\|_{\mathrm{op}} \lesssim \lambda$

Then the LASSO solution $\widehat{M}$ *is unique, and*

$$\operatorname{rank}(\widehat{M}) \leq \left(1 + c\left[\delta + \frac{\|\mathcal{A}^*(\xi)\|_{\mathrm{op}}}{\lambda}\right]^2\right) r_* \leq 1.1 r_*$$

- Classical assumptions in statistical theory of low-rank recovery, e.g., Candès and Plan (2011) and Negahban and Wainwright (2011)
- **First explicit rank bound** (without exact recovery or stronger structural assumptions)

Proof idea that doesn't quite work

Original LASSO (sparse recovery):

$$\hat{x} = \arg\min_{x \in \mathbf{R}^d} \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 \text{ where } y = Ax_* + \xi$$

How to show $\hat{x}$ is sparse?

- Sufficient to show **support recovery**
- Support is discrete + ℓ_1 penalty $\implies$ robust to noise (Wainwright, 2009)

How to translate to low-rank matrix M_*?

- Support of x_* $\longrightarrow$ row/column spaces of M_*
- Continuous objects: **no longer robust** to perturbation!

Proof idea that does work

Matrix LASSO and subgradient optimality condition:

$$\widehat{M} = \arg\min_{M \in \mathbf{R}^{d_1 \times d_2}} \frac{1}{2}\|\xi + \mathcal{A}(M_* - M)\|^2 + \lambda \|M\|_*$$

$$\widehat{E} := \frac{1}{\lambda}\mathcal{A}^*\mathcal{A}(M_* - \widehat{M}) + \mathcal{A}^*(\xi) \in \partial\|\widehat{M}\|_*$$

Compare to solution/subgradient of **idealized problem**:

$$M_\lambda = \arg\min_{M \in \mathbf{R}^{d_1 \times d_2}} \frac{1}{2}\|M_* - M\|_{\mathrm{F}}^2 + \lambda \|M\|_*$$

$$E_\lambda := \frac{1}{\lambda}(M_* - M_\lambda) \in \partial\|M_\lambda\|_*$$

Proof steps:

- $\operatorname{rank}(\widehat{M}) \leq \#\{\ell : \sigma_\ell(\widehat{E}) = 1\}$ (property of subgradient)
- $\operatorname{rank}(E_\lambda) \leq \operatorname{rank}(M_*) = r_*$ (formula in terms of SVD of M_*)
- **Hard part:** show $\widehat{E} \approx E_\lambda$ (statistical analysis)

Algorithmic consequences

We showed LASSO solution $\widehat{M}$ has low rank (good for **storage/computation**)

What about an efficient **algorithm** to solve the LASSO?

- #variables can be reduced by optimizing directly over low-rank matrices:

$$\min_{\substack{M \in \mathbf{R}^{d_1 \times d_2} \\ \operatorname{rank}(M) \leq r}} \frac{1}{2}\|y - \mathcal{A}(M)\|^2 + \lambda \|M\|_*$$

- Equivalent if $\operatorname{rank}(\widehat{M}) \leq r$ (true by our rank bound)
- However, constrained problem **nonconvex**

Algorithmic result 1

Rank-constrained problem:

$$\min_{\substack{M \in \mathbf{R}^{d_1 \times d_2} \\ \operatorname{rank}(M) \le r}} \overbrace{\frac{1}{2}\|y - \mathcal{A}(M)\|^2}^{f(M)} + \lambda \|M\|_*$$

Projected proximal gradient descent (stepsize $\eta > 0$):

$$M_{t+1} = \mathop{\arg\min}_{\substack{M \in \mathbf{R}^{d_1 \times d_2} \\ \operatorname{rank}(M) \le r}} \langle M, \nabla f(M_t) \rangle + \lambda \|M\|_* + \frac{1}{2\eta}\|M - M_t\|_F^2 \qquad \text{(PPGD)}$$

Computed by truncated SVD (**fast randomized algorithms**)

Theorem (Informal)

Under the conditions of the rank bound, with appropriate η, the iterates of (PPGD) ***converge linearly*** *to the LASSO solution $\widehat{M}$ from* ***any*** *initialization.*

- Resembles previous results (usually w/o $\|M\|_*$), e.g., Zhang, Bi, and Lavaei (2021)
- Key requirements: bound on $\operatorname{rank}(\widehat{M})$ and RIP

Algorithmic result 2

Rank-constrained problem:

$$\min_{\substack{M \in \mathbf{R}^{d_1 \times d_2} \\ \operatorname{rank}(M) \le r}} \quad \frac{1}{2}\|y - \mathcal{A}(M)\|^2 + \lambda \|M\|_*$$

Equivalent Burer-Monteiro factored formulation (Srebro, Rennie, and Jaakkola, 2004):

$$\min_{\substack{U \in \mathbf{R}^{d_1 \times r} \\ V \in \mathbf{R}^{d_2 \times r}}} \quad \frac{1}{2}\|y - \mathcal{A}(UV^T)\|^2 + \lambda \frac{\|U\|_F^2 + \|V\|_F^2}{2}. \tag{BM}$$

Smooth optimization over exactly $r(d_1 + d_2)$ variables.

Theorem (Informal)

*Under the conditions of the rank bound, every **second-order critical point** (U, V) of (BM) (zero gradient and PSD Hessian) satisfies $UV^T = \widehat{M}$.*

- Many previous results assuming RIP and low-rank $\widehat{M}$
- Follows from previous result (PPGD) by an argument of Ha, Liu, and Barber (2020)

Limitations and future work

Restricted isometry property (RIP) quite **strong assumption** in matrix setting

$$(1-\delta)\|M\|_F^2 \le \|\mathcal{A}(M)\|_2^2 \le (1+\delta)\|M\|_F^2 \text{ if } \operatorname{rank}(M) \le 2r \tag{RIP}$$

Common choice: $\mathcal{A}(M)_i = \langle X_i, M\rangle$, $X_1, \ldots, X_n$ i.i.d. random matrices

- Rank-1 X_i (good for computation) don't give RIP (too heavy-tailed)
- Dense random X_i (e.g., Gaussian entries) give RIP but computationally impractical

Weaker assumptions: ℓ_2 lower isometry, ℓ_1 isometry...

- Sufficient for good statistical recovery
- **Q:** Sufficient for rank bound/landscape results?

Conclusion

LASSO algorithm for low-rank matrix recovery:

$$\widehat{M} = \arg\min_{M \in \mathbf{R}^{d_1 \times d_2}} \frac{1}{2}\|y - \mathcal{A}(M)\|^2 + \lambda \|M\|_* \text{ where } y = \mathcal{A}(M_*) + \xi$$

Contributions:

- Guarantee $\widehat{M}$ **has low rank** under classical assumptions (RIP, $\|\mathcal{A}^*(\xi)\|_{\mathrm{op}} \lesssim \lambda$)
- Consequently, **fast** low-rank algorithms (PPGD, BM) **find the solution**

Preprint: Andrew D. McRae (2024). "Low solution rank of the matrix LASSO under RIP with consequences for rank-constrained algorithms". In: arXiv: 2404.12828 [math.OC]

Swiss National Science Foundation

References I

Candès, Emmanuel J. and Y Plan (2011). "Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements". In: *IEEE Trans. Inf. Theory* 57.4, pp. 2342–2359.

Ha, Wooseok, Haoyang Liu, and Rina Foygel Barber (2020). "An Equivalence between Critical Points for Rank Constraints Versus Low-Rank Factorizations". In: *SIAM J. Optim.* 30.4, pp. 2927–2955.

McRae, Andrew D. (2024). "Low solution rank of the matrix LASSO under RIP with consequences for rank-constrained algorithms". In: arXiv: 2404.12828 [math.OC].

Negahban, Sahand and Martin J. Wainwright (2011). "Estimation of (near) low-rank matrices with noise and high-dimensional scaling". In: *Ann. Stat.* 39.2.

Rohde, Angelika and Alexandre B. Tsybakov (2011). "Estimation of high-dimensional low-rank matrices". In: *Ann. Stat.* 39.2.

Srebro, Nathan, Jason Rennie, and Tommi Jaakkola (Dec. 2004). "Maximum-Margin Matrix Factorization". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vol. 17. Vancouver, Canada.

References II

- Wainwright, Martin J. (2009). "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE Trans. Inf. Theory* 55.5, pp. 2183–2202.
- Zhang, Haixiang, Yingjie Bi, and Javad Lavaei (Dec. 2021). "General Low-rank Matrix Optimization: Geometric Analysis and Sharper Bounds". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual conference, pp. 27369–27380.