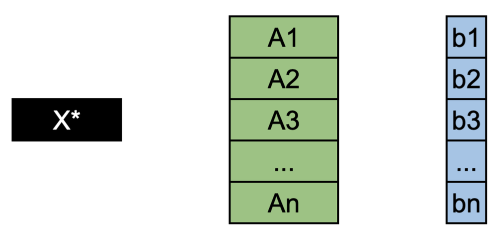
Robust Matrix Sensing in the Semi-Random Model

Xing Gao[1], Yu Cheng[2]

[1]University of Illinois at Chicago, [2]Brown University

Low-Rank Matrix Sensing

Goal: Recover a ground-truth matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ where $\text{rank}(X^*) \le r$.
Input: n sensing matrices $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$ and
linear measurements $b_i = \langle A_i, X^* \rangle$ = entrywise inner product of A_i and X^*.



We define a sensing operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ where $\mathcal{A}[X] \in \mathbb{R}^n$ and $(\mathcal{A}[X])_i = \langle A_i, X \rangle$.

Standard Assumptions and Prior Approaches

Standard assumption (matrix RIP):
The sensing matrices $(A_i)_{i=1}^n$ satisfy (r, L)-Restricted Isometry Property (**RIP**) if:

$$\frac{1}{L} \|X\|_F^2 \le \frac{1}{n} \sum_{i=1}^n \langle A_i, X \rangle^2 \le L \|X\|_F^2 \quad \text{for all } X \text{ with } \text{rank}(X) \le r.$$

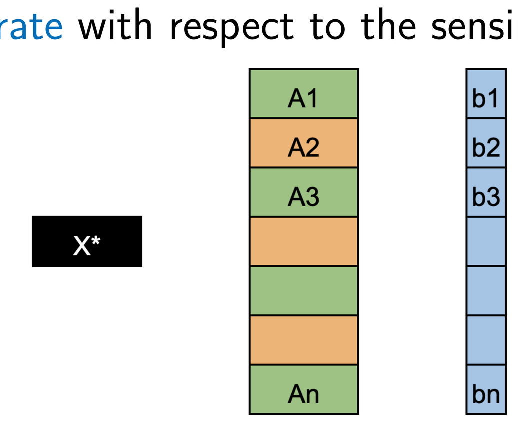
Notably, $n = \Omega(dr)$ Gaussian sensing matrices satisfy $(r, O(1))$-RIP with high probability.

Prior approaches:

- Convex approaches, e.g., nuclear norm minimization [1]: $\min_{X \in \mathbb{R}^{d_1 \times d_2}} \|X\|_* \ \ s.t. \ \ \mathcal{A}[X] = b$.
- Non-convex approaches, e.g., alternating minimization [1, 2]: $\min_{\substack{U \in \mathbb{R}^{d_1 \times r} \\ V \in \mathbb{R}^{d_2 \times r}}} \|\mathcal{A}[UV^\top] - b\|_2^2$.

Our Model: Semi-Random Matrix Sensing

An unknown subset of "good" sensing matrices satisfy RIP.
The remaining (arbitrarily many) "bad" sensing matrices are chosen adversarially.
Linear measurements are accurate with respect to the sensing matrices.



Our assumption: weighted RIP (wRIP):
There exist unknown weights $(w_i^*)_{i=1}^n$ where the weighted sensing matrices is (r, L)-RIP:

$$\frac{1}{L} \|X\|_F^2 \le \sum_{i=1}^n w_i^* \langle A_i, X \rangle^2 \le L \|X\|_F^2 \quad \text{for all } X \text{ with } \text{rank}(X) \le r.$$

Challenges:

- Finding weights to satisfy RIP seems hard, as verifying RIP is NP-hard.
- Natural non-convex objective functions can have bad local optima [2].

Our Results and Contributions

Theorem 1.1 (informal) Suppose $X^* \in \mathbb{R}^{d_1 \times d_2}$ has $\text{rank}(X^*) \le r$ and $\|X^*\|_F \le \text{poly}(d)$.
Let $A_1, \dots, A_n$ be the sensing matrices and $b_i = \langle A_i, X^* \rangle$ be the linear measurements.
Suppose $A_1, \dots, A_n$ satisfy the wRIP condition.
Our algorithm can output $X \in \mathbb{R}^{d_1 \times d_2}$ in time $\widetilde{O}(nd^{\omega+1} r \log(1/\epsilon))$ such that with high probability $\|X - X^*\|_F \le \epsilon$, where $\widetilde{O}(\cdot)$ hides logarithmic factors in its parameters, $d = \max(d_1, d_2)$, and $\omega < 2.373$ is the matrix multiplication exponent.

Our Contributions:

- We pose and study matrix sensing in the semi-random model.
- We present an efficient algorithm that provably recovers the ground-truth matrix X^*.
- We exploit the connection between sparsity in vectors and low-rankness in matrices, which may have other applications in designing robust algorithms for sparse/low-rank problems.

Overview of Our Approach

We build on the framework proposed in [3] for semi-random sparse linear regression.

In each iteration, our algorithm

- Calls a weight oracle which finds a set of weights based on the current solution.
- Takes a weighted gradient step that is guaranteed to work well locally.

Roadmap:

Main algorithm (Algorithm 1)

Start with $X = 0$ and $\|X^*\|_F = R_0$.
Run an algorithm that reduces the error by half with each call.
After $\log(R_0/\epsilon)$ calls, the error is reduced to $\|X - X^*\|_F \le \epsilon$.

Halving algorithm (Algorithm 2)

Given X_{in}, Algorithm 2 finds X_{out} via projected gradient descent such that $\|X_{\text{out}} - X^*\|_F \le \frac{1}{2}\|X_{\text{in}} - X^*\|_F$.
In each iteration, Algorithm 2:
(1) **Reweights** the objective function with weights $w_t \in \mathbb{R}^n$ provided by a weight oracle:
$f_t(X) = \frac{1}{2} \sum_{i=1}^n (w_t)_i \langle A_i, X - X^* \rangle^2$.
(2) **Take a gradient step** $-\eta G_t$ where
$G_t = \sum_{i=1}^n (w_t)_i \langle A_i, X_t - X^* \rangle A_i$.
(3) **Project** $(X_t - \eta G_t)$ onto a $\|\cdot\|_*$ ball to obtain X_{t+1}, so that good weights exist for measuring $(X_{t+1} - X^*)$.

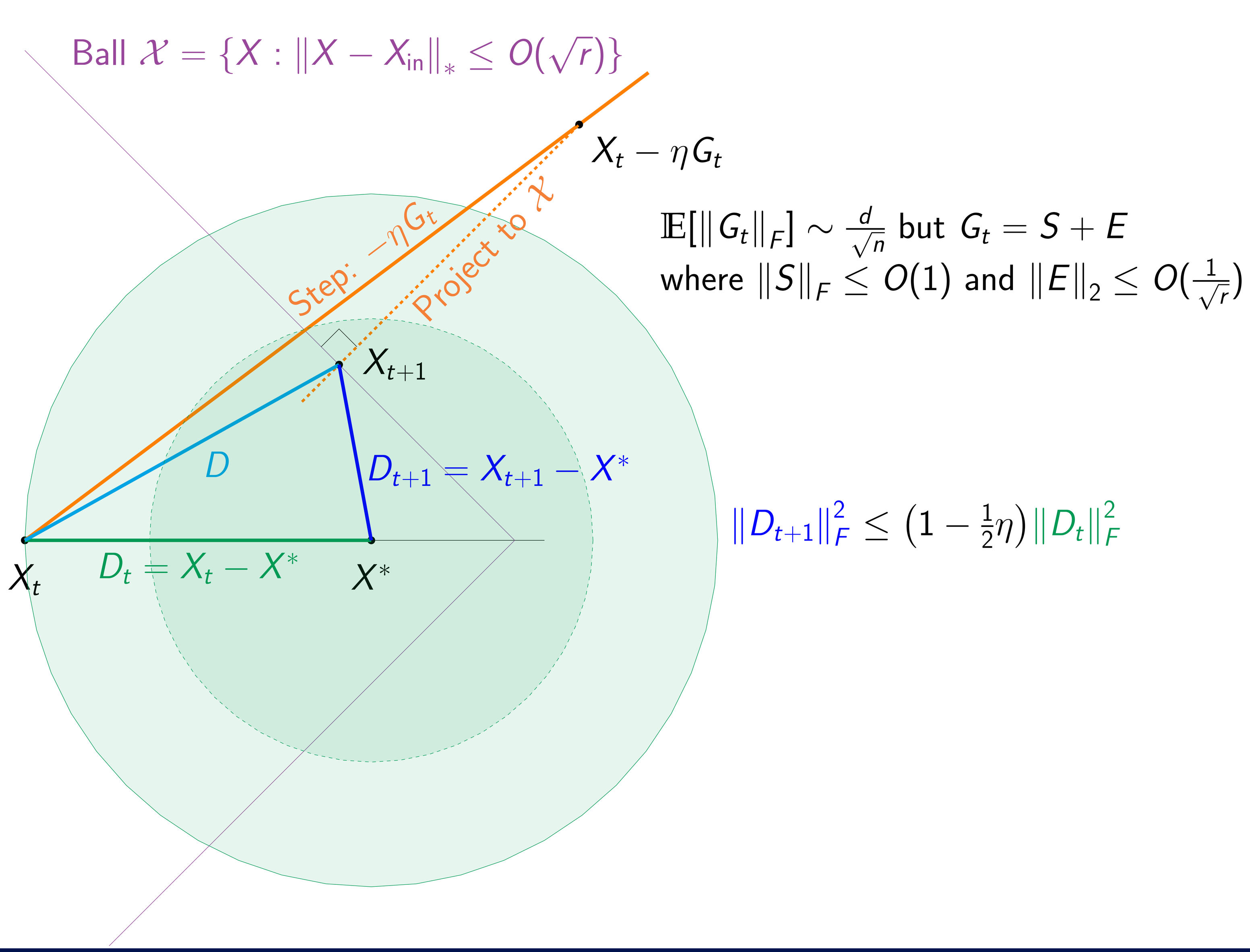
Weight oracle (Algorithm 3)

Given X_t, Algorithm 3 computes locally good weights.
The output of Algorithm 3 satisfies
(1) **Progress guarantee:**
$\langle G_t, X_t - X^* \rangle$ is large.
In particular, $\langle G_t, X_t - X^* \rangle \ge 1$.
(2) **Decomposition guarantee:**
$G_t = S + E$ where $\|S\|_F$ and $\|E\|_2$ are small.
In particular, $\|S\|_F \le 12L^2, \|E\|_2 \le \frac{1}{6\sqrt{r}}$.
Algorithm 3 uses a potential function Φ to guide the choice of w_t.
Intuitively, Φ encourages increasing the weight on A_i if $\langle A_i, X_t - X^* \rangle$ is large or if it improves the decomposability of G_t.

wRIP

The existence of good weights follows from the wRIP condition.

Intuition for the halving algorithm (Algorithm 2)



Halving Algorithm (Algorithm 2)

Algorithm 2: HalveError($X_{\text{in}}, \mathcal{A}, b, \delta$)

Input: X_{in} with $\text{rank}(X_{\text{in}}) \le r$ and $\|X_{\text{in}} - X^*\|_F \le 1$.
Output: X_{out} with $\text{rank}(X_{\text{out}}) \le r$ and $\|X_{\text{out}} - X^*\|_F \le \frac{1}{2}$ w.p. $\ge 1 - \delta$.
for $0 \le t < T$ **do**
  $u_t \leftarrow \mathcal{A}[X_t] - b$
  $w_t \leftarrow \text{WeightOracle}(\mathcal{A}, u_t)$
  if w_t satisfies the **progress** and **decomposition** guarantees on u_t **then**
    $G_t \leftarrow \sum_{i=1}^n (w_t)_i (u_t)_i A_i$
    $X_{t+1} \leftarrow \text{argmin}_{X \in \mathcal{X}} \|X - (X_t - \eta G_t)\|_F^2$
  else
    Return $X_{\text{out}} \leftarrow$ rank-r approximation of X_t
  end if
end for

Proof sketch (Lemma 4.2)

- Projection $\implies$ $\|D_t\|_F^2 - \|D_{t+1}\|_F^2 \ge 2\eta \langle G_t, D_t \rangle + 2\eta \langle G_t, D \rangle + \langle D, D \rangle$.
- Progress guarantee $\implies$ $2\eta \langle G_t, D_t \rangle = 2\eta \sum_{i=1}^n (w_t)_i (u_t)_i^2 \ge 2\eta$.
- Decomposition guarantee $\implies$ $G_t = S_t + E_t$ where $\|S_t\|_F \le 12L^2$ and $\|E_t\|_2 \le \frac{1}{6\sqrt{r}}$,

$$\begin{aligned} 2\eta \langle G_t, D \rangle + \langle D, D \rangle &= 2\eta \langle E_t, D \rangle + 2\eta \langle S_t, D \rangle + \langle D, D \rangle \\ &\ge -2\eta \|E_t\|_2 \|D\|_* - \eta^2 \langle S_t, S_t \rangle \ge -\frac{3}{2}\eta \, . \end{aligned}$$

$\implies$ Contraction in each step: $\|D_{t+1}\|_F^2 \le \left(1 - \frac{\eta}{2}\right) \cdot \|D_t\|_F^2$.

Weight Oracle (Algorithm 3)

Algorithm 3: WeightOracle($\mathcal{A}, u$), simplified

Input: $u = \mathcal{A}[V] \in \mathbb{R}^n$ for some V satisfying $\|V\|_F \in [\frac{1}{4}, 1]$ and $\|V\|_* \le 2\sqrt{2r}$.
Output: $w \in \mathbb{R}^n$ that satisfies the progress and decomposition guarantees w.p. $\ge \frac{1}{2}$.
$w_0 \leftarrow 0$
for $0 \le t \le N$ **do**
  if $\Phi_{\text{prog}}(w_t) \ge 1$ **then**
    Return $w \leftarrow w_t$
  else
    for a random coordinate i : $s_t \leftarrow \text{argmax}_{s \in [0, \eta]} \Phi(w_t + s e_i)$, $w_{t+1} \leftarrow w_t + s_t e_i$
  end if
end for

Potential functions:
Parameters: $C = 108$, $\mu = \frac{1}{\sqrt{Cr \log d}}$, $\eta = o(\frac{1}{r\rho^2 \log d})$, $N = O(\frac{Ln}{\eta})$.
Progress potential: $\Phi_{\text{prog}}(w) = \sum_{i=1}^n w_i u_i^2$.
Decomposition potential: $\Phi_{\text{dc}}(w) = \min_{\|S\|_F \le L\|w\|_1} \left(\mu^2 \log\left[F(G_w - S)\right] \right) + \frac{\|w\|_1}{4CLr}$,
where $G_w = \sum_{i=1}^n w_i u_i A_i$, $F(E) = \text{tr} \exp\left(\frac{E^\top E}{\mu^2}\right)$.
Overall potential: $\Phi(w) = \Phi_{\text{prog}}(w) - Cr\Phi_{\text{dc}}(w)$.

Proof sketch (Lemma B.1, B.2, B.3)

- Progress guarantee:
  $\Phi_{\text{prog}}(w_t) \ge 1$ is satisfied upon termination.
- Decomposition guarantee:
  by maximizing overall potential $\Phi(w)$, upon termination $\Phi_{\text{dc}}(w_t) \le \frac{3}{Cr}$,
  which implies existence of decomposition $\|S\|_F \le 12L^2$ and $\|E\|_2 \le \frac{1}{6\sqrt{r}}$.
- Success probability:
  overall potential $\Phi(w) = \Phi_{\text{prog}}(w) - Cr\Phi_{\text{dc}}(w) \ge 0 \implies \Phi_{\text{prog}}(w) \ge 1$,
  and expected increase per round in $\Phi(w)$ is lower bounded by $\frac{\eta}{2Ln}$.

References

[1] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[2] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. "Global optimality of local search for low rank matrix recovery". In: *Advances in Neural Information Processing Systems* 29 (2016).

[3] Jonathan Kelner et al. "Semi-random sparse recovery in nearly-linear time". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 2352–2398.