AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

April 23, 2026

Outline

Low-Rank Optimization

Matrix Completion

The Frank–Wolfe Method

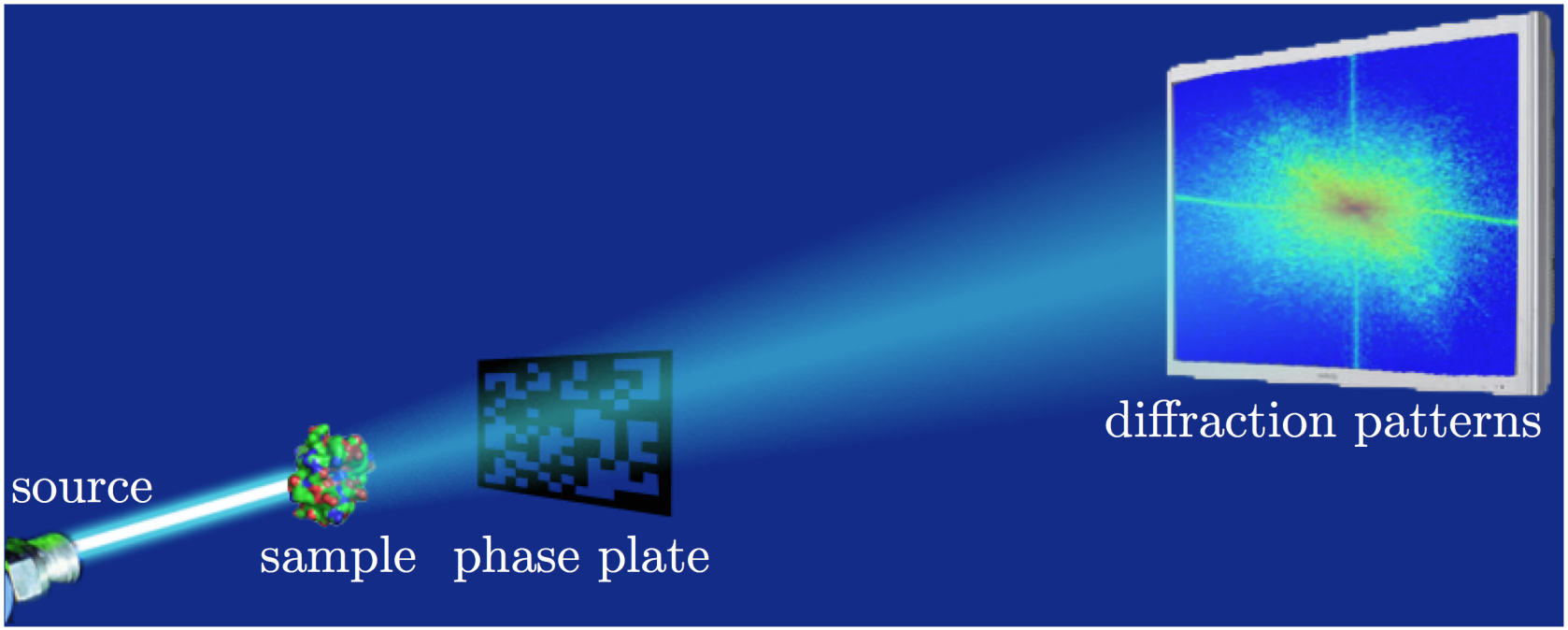
Nonconvex Factorization: Burer–Monteiro

Low Rank Structure is Ubiquitous

Recommender systems: the Netflix rating matrix is approximately low-rank because few latent factors explain user preferences.

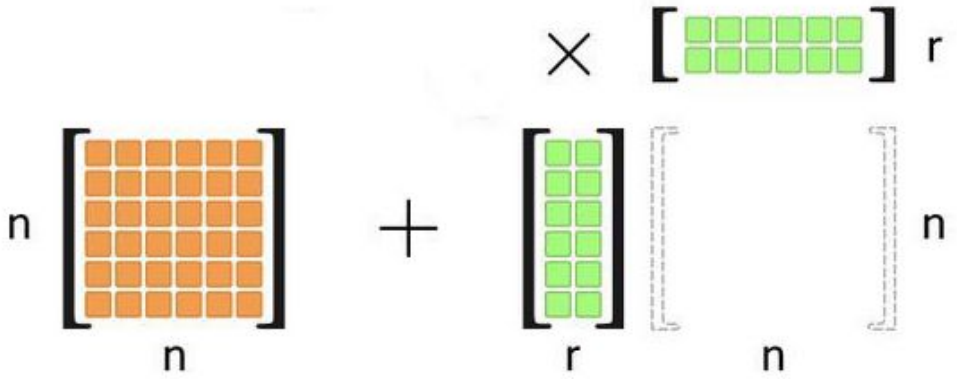
Low Rank Structure is Ubiquitous

Phase retrieval: reconstruct a rank-1 matrix $xx^\top$ from quadratic measurements.



Low Rank Structure is Ubiquitous

LoRA fine-tuning: adapt large language models by adding a low-rank update to downstream applications.



$$W = W_0 + \underline{A}\,\underline{B}$$

Problem Formulation

Generic low-rank optimization:

$$\min_{X \in \mathbb{R}^{m \times n}} \quad f(X) \quad \text{s.t.} \quad \operatorname{rank}(X) \leq r.$$

Bad news:

- this problem is NP-hard in general.
- the rank function is discontinuous and nonconvex.

Easy case: if $f(X) = \frac{1}{2}\|X - A\|_F^2$, then the optimal X is the truncated SVD of A, i.e., $X^\star = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$.

Two general routes:

- Convex relaxation: replace $\operatorname{rank}(X)$ by the nuclear norm $\|X\|_*$.
- Nonconvex factorization: parametrize $X = UV^\top$ with $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$. This is called Burer–Monteiro method.

Nuclear Norm

Definition: let $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq 0$ be singular values of $X \in \mathbb{R}^{m \times n}$, then the nuclear norm of X is

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X).$$

Intuition: $\|X\|_*$ is the ℓ_1-norm on the singular values. Just as $\|x\|_1$ promotes sparse x, $\|X\|_*$ promotes sparse spectrum, i.e., low rank.

Equivalent characterizations:

- Spectral: $\|X\|_* = \sum_i \sigma_i(X)$
- Dual: $\|X\|_* = \max_{\|Z\|_{\mathrm{op}} \leq 1} \langle Z, X \rangle$, where $\|Z\|_{\mathrm{op}} = \sigma_1(Z)$
- Factorization: $\|X\|_* = \min_{U,V} \ \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$ s.t. $X = UV^\top$

The third characterization is the bridge to the Burer–Monteiro factorization at the end of the lecture.

Low-rank approximation

Singular value decomposition: any $X \in \mathbb{R}^{m \times n}$ admits

$$X = U \Sigma V^\top = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^\top,$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ and U, V having orthonormal columns.

Rank from SVD: $\operatorname{rank}(X)$ = number of nonzero singular values.

Low-rank Approximation Theorem

The best rank-r approximation of A is the truncated SVD:

$$\min_{\operatorname{rank}(X) \leq r} \|A - X\|_F^2 = \sum_{i>r} \sigma_i(A)^2,$$

attained at $X^\star = \sum_{i=1}^{r} \sigma_i(A)\, u_i v_i^\top$.

Nuclear Norm as Convex Envelope of Rank

The rank function is nonconvex and discontinuous. We want a convex surrogate that promotes low rank. Candidate: the nuclear norm.

Theorem

On the spectral-norm unit ball $\{X : \|X\|_{\text{op}} \leq 1\}$, the nuclear norm $\|X\|_*$ is the convex envelope of $\text{rank}(X)$, i.e., the largest convex function that lower bounds rank.

Sanity check: on the unit ball, every singular value satisfies $\sigma_i \leq 1$, so

$$\|X\|_* = \sum_i \sigma_i(X) \;\leq\; \sum_{i:\sigma_i>0} 1 = \text{rank}(X).$$

The nuclear norm is a valid lower bound on rank. The theorem says it is the tightest such convex lower bound.

Consequence: replacing rank by $\|\cdot\|_*$ gives the tightest convex relaxation. This is the foundation of nuclear-norm-based methods.

Subdifferential of the Nuclear Norm

Since $\|\cdot\|_*$ is convex but not differentiable, we need the subdifferential.

Subdifferential Formula

Let $X \in \mathbb{R}^{m\times n}$ have SVD $X = U\Sigma V^\top$ with $\operatorname{rank}(X) = r$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ are the singular vectors of nonzero singular values. Then

$$\partial \|X\|_* = \{UV^\top + W \ : \ U^\top W = 0, \ \ WV = 0, \ \ \|W\|_{\text{op}} \le 1\}.$$

Vector analogy: for $\|\cdot\|_1$ on $\mathbb{R}^n$,

$$\partial \|x\|_1 = \{\operatorname{sign}(x) + w \ : \ w_i = 0 \text{ if } x_i \neq 0, \ \ \|w\|_\infty \le 1\}.$$

Outline

Low-Rank Optimization

Matrix Completion

The Frank–Wolfe Method

Nonconvex Factorization: Burer–Monteiro

The Matrix Completion Problem

Setup: we have a partially observed matrix. Let $\Omega \subset [m] \times [n]$ be the set of observed indices, and let Y_{ij} be the observed values for $(i, j) \in \Omega$.

Goal: recover the full matrix $M \in \mathbb{R}^{m \times n}$.

Netflix Problem:

- Rows = users, columns = movies, entries = ratings.
- Missing entries: a user has rated only a tiny fraction of movies.
- Predicting missing ratings = recommending movies.

Without structure, the missing entries could be anything. The key assumption is that M is low rank: a few latent factors (genre, mood, era) explain most of the variability in user preferences.

Notation: let $\mathcal{P}_\Omega : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ be the projection onto Ω,

$$[\mathcal{P}_\Omega(X)]_{ij} = \begin{cases} X_{ij} & (i, j) \in \Omega \\ 0 & \text{otherwise.} \end{cases}$$

The Optimization Problem

Convex relaxation formulation:

$$\min_{X \in \mathbb{R}^{m \times n}} \ \frac{1}{2} \|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(Y)\|_F^2 \quad \text{s.t.} \quad \|X\|_* \leq \tau.$$

Objective: measure the squared error only on observed entries.

Constraint: $\|X\|_* \leq \tau$ promotes low rank.

Gradient: let $f(X) = \frac{1}{2} \|\mathcal{P}_\Omega(X - Y)\|_F^2$. By direct calculation,

$$\nabla f(X) = \mathcal{P}_\Omega(X - Y).$$

Observation: the gradient $\nabla f(X)$ is sparse: with $|\Omega|$ nonzero entries.

In typical applications $|\Omega| \ll mn$ (e.g., Netflix: $|\Omega|/mn \approx 1\%$).

Computing the Gradient

Since $\mathcal{P}_\Omega$ keeps observed entries and zeros out the rest,

$$f(X) = \frac{1}{2}\|\mathcal{P}_\Omega(X - Y)\|_F^2 = \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij} - Y_{ij})^2.$$

Differentiate entrywise:

- If $(k,l) \in \Omega$, only the term $\frac{1}{2}(X_{kl} - Y_{kl})^2$ depends on X_{kl}, so
$$\frac{\partial f}{\partial X_{kl}} = X_{kl} - Y_{kl}.$$
- If $(k,l) \notin \Omega$, no term in the sum involves X_{kl}, so
$$\frac{\partial f}{\partial X_{kl}} = 0.$$

Matrix form:

$$[\nabla f(X)]_{kl} = \begin{cases} X_{kl} - Y_{kl} & (k,l) \in \Omega \\ 0 & (k,l) \notin \Omega \end{cases} = \mathcal{P}_\Omega(X - Y).$$

Low Rank Is Essential

Question: without the rank constraint, what happens?

Example: a 3×3 matrix with two missing entries (marked ?):

$$Y = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & ? \\ 3 & 6 & ? \end{bmatrix}.$$

The unobserved $(2,3)$ and $(3,3)$ entries can be anything: infinitely many completions match Y on the seven observed entries.

But: if we have $\mathrm{rank}(X) \leq 1$, then we have the unique completion

$$X = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \implies X^\star = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}.$$

When Does Nuclear Norm Recovery Succeed?

Setup: $M \in \mathbb{R}^{n \times n}$ has rank r, and we observe entries on Ω chosen uniformly at random with $|\Omega| = m$. Solve the convex program

$$\min_X \|X\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M).$$

Theorem

Suppose M satisfies an incoherence condition with parameter μ. If

$$m \geq C\,\mu^2\,nr\,\log^2 n$$

for an absolute constant C, then with high probability M is the unique solution to the nuclear norm program.

Lesson: we need $\widetilde{O}(nr)$ samples to recover an $n \times n$ rank-r matrix. Compare to n^2 total entries, a huge saving when $r \ll n$.

Incoherence

Let $M = U\Sigma V^\top$ where $U, V \in \mathbb{R}^{n\times r}$. We say M is μ-incoherent if

$$\max_{1\le i\le n} \|U^\top e_i\|_2^2 \le \frac{\mu r}{n}, \qquad \max_{1\le j\le n} \|V^\top e_j\|_2^2 \le \frac{\mu r}{n}.$$

Meaning: the left or right singular vectors are spread out over coordinates; no row or column carries too much information.

Good example (incoherent):

$$M = \mathbf{1}\mathbf{1}^\top$$

has singular vectors proportional to $\mathbf{1}$, so every coordinate is similar.

Bad example (coherent):

$$M = e_1 e_1^\top$$

has all its mass on one row and one column. If entry $(1,1)$ is never sampled, exact recovery is impossible.

First Attempt: Projected Gradient Descent

PGD. For the constrained problem $\min_{\|X\|_* \le \tau} f(X)$,

$$X_{k+1} = \Pi_{\|\cdot\|_* \le \tau}(X_k - \eta \nabla f(X_k)).$$

Compute projection. We need to solve

$$\Pi_{\|\cdot\|_* \le \tau}(A) = \arg \min_{\|X\|_* \le \tau} \|X - A\|_F^2 .$$

Unitary invariance: the Frobenius norm depends only on singular values, i.e., $\|A\|_F^2 = \|\Sigma\|_F^2$ if $A = U\Sigma V^\top$ is SVD.

The problem reduces to:

$$\min_{s \in \mathbb{R}^n_+} \|s - \sigma(A)\|^2 \quad \text{s.t.} \quad \sum_i s_i \le \tau.$$

Solution: $s_i^* = (\sigma_i(A) - \theta)_+$, where θ is chosen so $\sum_i (\sigma_i - \theta)_+ = \tau$.

This is the singular value thresholding (SVT) operator.

Problem of PGD

Cost of one projected gradient iteration:

- Compute $\nabla f(X_k) = \mathcal{P}_\Omega(X_k - Y)$. Cost: $O(|\Omega|)$.
- Compute $A = X_k - \eta \nabla f(X_k)$. Cost: $O(mn)$.
- Project A onto the nuclear norm ball: full SVD of A. Cost: $O(\min(m,n)^2 \max(m,n))$.

The dilemma. The gradient is cheap because it is sparse. The projection destroys the "low rank + sparse" structure of A by requiring an SVD.

Question

Can we design a first-order method that avoids the projection step entirely, while still respecting the constraint $\|X\|_* \leq \tau$?

Answer: the Frank–Wolfe (conditional gradient) method.

Outline

Low-Rank Optimization

Matrix Completion

The Frank–Wolfe Method

Nonconvex Factorization: Burer–Monteiro

The Frank–Wolfe Idea

Goal. Avoid projection onto nuclear norm ball.

Insight. Projection is expensive because it solves a quadratic subproblem over $\Omega = \{X : \|X\|_* \leq \tau\}$. What if solve a linear subproblem?

Frank–Wolfe (conditional gradient). At iterate x_t, build a linear approximation of f and minimize it over Ω:

$$\bar{x}_t = \arg\min_{x \in \Omega} \left[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle \right] = \arg\min_{x \in \Omega} \langle \nabla f(x_t), x \rangle .$$

Then move toward $\bar{x}_t$:

$$x_{t+1} = x_t + \eta_t(\bar{x}_t - x_t).$$

No projection! Since $x_t, \bar{x}_t \in \Omega$ and Ω is convex, $x_{t+1} \in \Omega$ automatically for $\eta_t \in [0, 1]$.

The Frank–Wolfe Algorithm

Frank–Wolfe (Conditional Gradient)

Start from $x_0 \in \Omega$. For $t = 0, 1, 2, \ldots$ with step size $\eta_t = \frac{2}{t+2}$,

$$\begin{cases} \bar{x}_t = \arg\min_{x \in \Omega} \ \langle \nabla f(x_t), x \rangle \\ \\ x_{t+1} = (1 - \eta_t)\, x_t + \eta_t\, \bar{x}_t \end{cases}$$

Difference:

- Projected GD needs a quadratic subproblem.
- Frank–Wolfe needs a linear subproblem.

Observation. Linear subproblems are often much cheaper than quadratic ones, especially when the extreme points of Ω have simple structure.

Intuition

Two facts about Frank–Wolfe iterates:

- The update has the form
$$x_{t+1} = (1-\gamma_t)x_t + \gamma_t \bar{x}_t, \qquad \gamma_t \in [0,1],$$
so x_{t+1} is a convex combination of x_t and $\bar{x}_t$.
- Unrolling gives
$$x_t \in \operatorname{conv}\{x_0, \bar{x}_0, \ldots, \bar{x}_{t-1}\}.$$
So the iterates inherit the structure of the extreme points of Ω.

For the nuclear norm ball, the extreme points are rank-1 matrices:
$$\Omega = \{X : \|X\|_* \le \tau\}, \qquad \text{ext}(\Omega) = \{\tau uv^\top : \|u\| = \|v\| = 1\}.$$

Therefore Frank–Wolfe iterates are built from rank-1 atoms,
$$\operatorname{rank}(X_t) \le \operatorname{rank}(X_0) + t.$$
Starting from $X_0 = 0$, we get $\operatorname{rank}(X_t) \le t$.

Frank–Wolfe on the Nuclear Norm Ball

We now specialize Frank–Wolfe to the constrained problem

$$\min_{\|X\|_* \le \tau} \; f(X), \qquad f(X) = \frac{1}{2} \|\mathcal{P}_\Omega(X - Y)\|_F^2 .$$

At iterate X_t, let

$$G_t := \nabla f(X_t) = \mathcal{P}_\Omega(X_t - Y).$$

Then the Frank–Wolfe subproblem is

$$\bar{X}_t \;=\; \arg \min_{\|X\|_* \le \tau} \langle G_t, X \rangle .$$

So everything reduces to one question:

How do we minimize a linear function over the nuclear norm ball?

The key is that the nuclear norm ball is generated by rank-1 atoms.

Rank-1 Atoms

Key Fact:

$$\mathcal{B}_\tau^* = \operatorname{conv}(\mathcal{A}_\tau), \quad \text{where} \quad \begin{aligned} \mathcal{B}_\tau^* &:= \{X \in \mathbb{R}^{m\times n} : \|X\|_* \le \tau\}, \\ \mathcal{A}_\tau &:= \{\tau uv^\top : \|u\|_2 = \|v\|_2 = 1\}. \end{aligned}$$

Therefore, for any matrix G,

$$\min_{X\in\mathcal{B}_*(\tau)} \langle G, X\rangle = \min_{X\in\mathcal{A}_\tau} \langle G, X\rangle .$$

Proof: let $X = \sum_i \sigma_i u_i v_i^\top$ be an SVD. If $\|X\|_* = \sum_i \sigma_i \le \tau$, then

$$X = \sum_i \frac{\sigma_i}{\tau}\,(\tau u_i v_i^\top) + \Big(1 - \frac{\|X\|_*}{\tau}\Big)0.$$

Since $0 \in \operatorname{conv}(\mathcal{A}_\tau)$, it follows that $X \in \operatorname{conv}(\mathcal{A}_\tau)$.

Conversely, each atom satisfies $\left\|\tau uv^\top\right\|_* = \tau$, so every convex combination of atoms belongs to $\mathcal{B}_*(\tau)$.

Linear Minimization Oracle

Let $G \in \mathbb{R}^{m \times n}$. We need to solve

$$\min_{\|X\|_* \leq \tau} \langle G, X \rangle .$$

By duality between nuclear norm and operator norm, a lower bound:

$$\langle G, X \rangle \geq - \|G\|_{\text{op}} \, \|X\|_* \geq -\tau \, \|G\|_{\text{op}} .$$

Let u_1, v_1 be top left and right singular vectors of G, so

$$Gv_1 = \sigma_1(G) u_1, \qquad u_1^\top G v_1 = \sigma_1(G) = \|G\|_{\text{op}} .$$

Choose $X^\star = -\tau u_1 v_1^\top$, then $\|X^\star\|_* = \tau$ and

$$\langle G, X^\star \rangle = -\tau \, u_1^\top G v_1 = -\tau \, \sigma_1(G) = -\tau \, \|G\|_{\text{op}} .$$

Therefore

$$\boxed{\arg \min_{\|X\|_* \leq \tau} \langle G, X \rangle = -\tau u_1 v_1^\top}$$

Frank–Wolfe for Matrix Completion

We can now write Frank–Wolfe explicitly for

$$\min_{\|X\|_* \le \tau} \ \frac{1}{2} \|\mathcal{P}_\Omega(X - Y)\|_F^2 .$$

At iterate X_t, compute the gradient

$$G_t := \nabla f(X_t) = \mathcal{P}_\Omega(X_t - Y).$$

Then compute a top singular-vector pair (u_t, v_t) of G_t, and set

$$\bar{X}_t = -\tau u_t v_t^\top .$$

The Frank–Wolfe update is

$$X_{t+1} = (1 - \eta_t) X_t + \eta_t \bar{X}_t, \qquad \eta_t = \frac{2}{t+2}.$$

Each iteration needs:

- the sparse gradient $G_t = \mathcal{P}_\Omega(X_t - Y)$,
- only a top singular-vector computation,
- no projection onto the nuclear norm ball.

Frank–Wolfe Preserves Low Rank

Recall that each $\bar{X}_t$ has rank 1. Since

$$X_{t+1} = (1 - \eta_t)X_t + \eta_t \bar{X}_t,$$

we have

$$\operatorname{rank}(X_{t+1}) \le \operatorname{rank}(X_t) + \operatorname{rank}(\bar{X}_t) \le \operatorname{rank}(X_t) + 1.$$

Therefore, if we start from $X_0 = 0$, then

$$\operatorname{rank}(X_t) \le t.$$

Interpretation: Frank–Wolfe builds the iterate by adding one rank-1 atom at a time. This is why it is so natural for low-rank optimization.

FW vs. PGD: PGD keeps feasibility by projection, but the projection requires a full SVD. FW keeps feasibility by convex combination, and the iterate remains explicitly low rank.

Convergence of Frank–Wolfe

Convergence of Frank–Wolfe

Let $\mathcal{D} \subset \mathbb{R}^{m\times n}$ be convex and compact, and assume f is convex and L-smooth on $\mathcal{D}$. Define $D := \sup_{X,Z\in\mathcal{D}} \|X - Z\|_F$. Consider

$$S_t \in \arg\min_{S\in\mathcal{D}} \langle \nabla f(X_t), S\rangle, \quad X_{t+1} = (1-\eta_t)X_t + \eta_t S_t, \quad \eta_t = \tfrac{2}{t+2}.$$

Then, for all $t \geq 1$,

$$f(X_t) - f(X^\star) \leq \frac{2LD^2}{t+2}.$$

Proof: Let $h_t := f(X_t) - f(X^\star)$. By L-smoothness,

$$f(X_{t+1}) \leq f(X_t) + \eta_t \langle \nabla f(X_t), S_t - X_t\rangle + \tfrac{L}{2}\eta_t^2 \|S_t - X_t\|_F^2.$$

Since S_t is the minimizer, $X^\star \in \mathcal{D}$, and f is convex,

$$\langle \nabla f(X_t), S_t - X_t\rangle \leq \langle \nabla f(X_t), X^\star - X_t\rangle \leq -h_t$$

Proof Continued

Also, $\|S_t - X_t\|_F \leq D$. Therefore

$$h_{t+1} \leq (1 - \eta_t)h_t + \frac{LD^2}{2}\eta_t^2.$$

With $\eta_t = \frac{2}{t+2}$,

$$h_{t+1} \leq \frac{t}{t+2}h_t + \frac{2LD^2}{(t+2)^2}. \qquad (*)$$

For $t = 0$, the smoothness bound $(*)$ gives $h_1 \leq \frac{LD^2}{2}$.

Now assume $h_t \leq \frac{2LD^2}{t+2}$, then by $(*)$,

$$h_{t+1} \leq \frac{t}{t+2} \cdot \frac{2LD^2}{t+2} + \frac{2LD^2}{(t+2)^2} = \frac{2LD^2(t+1)}{(t+2)^2} \leq \frac{2LD^2}{t+3}.$$

Hence, by induction, $h_t \leq \frac{2LD^2}{t+2}$ for all $t \geq 1$.

Outline

Low-Rank Optimization

Matrix Completion

The Frank–Wolfe Method

Nonconvex Factorization: Burer–Monteiro

Burer–Monteiro Factorization

Recall the generic low-rank problem

$$\min_{X \in \mathbb{R}^{m \times n}} f(X) \qquad \text{s.t.} \qquad \operatorname{rank}(X) \leq r.$$

A natural idea to ensure $\operatorname{rank}(UV^\top) \leq r$ is to parameterize

$$X = UV^\top, \qquad U \in \mathbb{R}^{m \times r}, \quad V \in \mathbb{R}^{n \times r}.$$

So we obtain the factorized problem

$$\min_{U \in \mathbb{R}^{m \times r},\, V \in \mathbb{R}^{n \times r}} f(UV^\top).$$

Benefits:

- We optimize over $r(m+n)$ variables instead of mn.
- The iterate is low rank by construction.
- We avoid projection onto the nuclear norm ball.

This is often called the Burer–Monteiro (BM) approach.

BM for Matrix Completion

Applying this to matrix completion leads to

$$\min_{U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}} \frac{1}{2}\left\|\mathcal{P}_\Omega(UV^\top - Y)\right\|_F^2.$$

Pros & Cons:

- the problem becomes an unconstrained and smooth;
- the number of variables drops from mn to $r(m+n)$;
- but the objective is now nonconvex in (U, V).

Let

$$F(U,V) := \frac{1}{2}\left\|R(U,V)\right\|_F^2, \qquad R(U,V) := \mathcal{P}_\Omega(UV^\top - Y).$$

Gradients:

$$\nabla_U F(U,V) = R(U,V)\,V, \qquad \nabla_V F(U,V) = R(U,V)^\top U.$$

Gradient Formula

Write the objective entrywise:

$$F(U,V) = \frac{1}{2} \sum_{(i,j)\in\Omega} (u_i^\top v_j - Y_{ij})^2,$$

where $u_i^\top$ is the i-th row of U, and $v_j^\top$ is the j-th row of V.

Now perturb U in a direction $H \in \mathbb{R}^{m\times r}$: $U \mapsto U + \varepsilon H$.

If $h_i^\top$ denotes the i-th row of H, then

$$(u_i + \varepsilon h_i)^\top v_j = u_i^\top v_j + \varepsilon h_i^\top v_j.$$

So

$$F(U+\varepsilon H, V) = \frac{1}{2} \sum_{(i,j)\in\Omega} \left(u_i^\top v_j + \varepsilon h_i^\top v_j - Y_{ij}\right)^2.$$

Differentiating with respect to ε at $\varepsilon = 0$,

$$dF(U,V)[H] = \sum_{(i,j)\in\Omega} (u_i^\top v_j - Y_{ij})\, h_i^\top v_j.$$

Gradient Formula

Define the residual matrix $R(U,V) := \mathcal{P}_\Omega(UV^\top - Y)$, where

$$R_{ij}(U,V) = \begin{cases} u_i^\top v_j - Y_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases},$$

Therefore the directional derivative becomes

$$dF(U,V)[H] = \sum_{(i,j)\in\Omega} R_{ij}(U,V)\, h_i^\top v_j.$$

Group the terms by row index i:

$$dF(U,V)[H] = \sum_{i=1}^{m} h_i^\top \left(\sum_{j:(i,j)\in\Omega} R_{ij}(U,V)\, v_j \right).$$

But the vector

$$\sum_{j:(i,j)\in\Omega} R_{ij}(U,V)\, v_j$$

is exactly the i-th row of the matrix $R(U,V)V$. Hence

$$dF(U,V)[H] = \langle R(U,V)V, H \rangle \implies \nabla_U F(U,V) = R(U,V)V.$$

Gradient Descent on the Factors

Starting from (U_0, V_0), plain gradient descent applies

$$U_{t+1} = U_t - \eta_t \nabla_U F(U_t, V_t), \qquad V_{t+1} = V_t - \eta_t \nabla_V F(U_t, V_t).$$

Using the gradient formulas, this becomes

$$R_t := \mathcal{P}_\Omega(U_t V_t^\top - Y),$$
$$U_{t+1} = U_t - \eta_t R_t V_t, \qquad V_{t+1} = V_t - \eta_t R_t^\top U_t.$$

Remarks:

- This is an unconstrained smooth optimization problem.
- Each step only uses the observed entries through R_t.
- We never compute an SVD or projection onto the nuclear norm ball.

Pros & Cons

Assume the observed index set Ω is sparse, with $|\Omega| \ll mn$. To compute

$$R_t = \mathcal{P}_\Omega(U_t V_t^\top - Y),$$

we only need the entries

$$(U_t V_t^\top)_{ij} = u_i^\top v_j \qquad \text{for } (i,j) \in \Omega.$$

Thus the dominant cost per iteration is roughly $O(|\Omega| r)$ rather than working with all mn entries of the matrix.

Pros:

- we store U_t and V_t, not a full $m \times n$ matrix;
- we use first-order updates directly on the factors;
- we avoid the expensive projection or SVD step.

Cons: the objective is nonconvex, and the factorization is not unique:

$$UV^\top = (UR)(VR^{-\top})^\top \qquad \text{for any invertible } R \in \mathbb{R}^{r \times r}.$$

Initialization

Since the factorized problem is nonconvex, initialization can matter a lot.

A common choice is spectral initialization. Let

$$p := \frac{|\Omega|}{mn}, \qquad A_0 := \frac{1}{p}\mathcal{P}_\Omega(Y).$$

Then A_0 is a crude estimate of the full matrix.

Compute its rank-r truncated SVD: $A_0 \approx \widehat{U}_r \Sigma_r \widehat{V}_r^\top$.

Initialize $U_0 := \widehat{U}_r \Sigma_r^{1/2}$ and $V_0 := \widehat{V}_r \Sigma_r^{1/2}$.

Why this is reasonable?

- If the samples are uniform, then $\mathcal{P}_\Omega(Y)$ contains an unbiased but incomplete version of the data.
- The truncated SVD extracts its leading low-rank structure.
- This gives a sensible starting point for gradient-based refinement.

Theory and Practice

The BM model is nonconvex, but often works very well.

- The convex nuclear norm model gives a global certificate, but each iteration can be expensive.
- The BM formulation is nonconvex, so global convergence is no longer automatic.
- However, under suitable assumptions (such as low rank, enough random samples, and a good initialization), gradient-based methods can provably recover the underlying matrix.
- In practice, factorized methods are often much more scalable for large matrix completion problems.

Regularizations

The basic BM model is

$$\min_{U,V} \ \frac{1}{2}\left\|\mathcal{P}_\Omega(UV^\top - Y)\right\|_F^2.$$

In practice, one may add regularization, for example

$$\frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) \quad \text{or} \quad \frac{\mu}{4}\left\|U^\top U - V^\top V\right\|_F^2.$$

Interpretation:

- $\|U\|_F^2 + \|V\|_F^2$: related to nuclear-norm regularization;
- $\left\|U^\top U - V^\top V\right\|_F^2$: encourages a balanced factorization.

These regularizers are not for low rank. Low rank already comes from the factorization $X = UV^\top$.

Wrap-up and Next Week

Week 11: Key takeaways

- Low-rank optimization framework.
- Nuclear norm: convex relaxation for low-rank optimization.
- Frank–Wolfe: projection-free, rank-1 updates, convergence.
- Burer–Monteiro: low-rank by construction, but nonconvex.

Thank you for your attendance!

Questions?

Next week (Week 12): sparse optimization