Optimization: Algorithms, Complexity & Approximations

Anastasios Kyrillidis *

*Instructor, Computer Science at Rice University

Gaurav Gupta, Yi Lin (2020 scriber)

Chapter 9

Just like the previous lecture where we talked about sparsity constraints (non-convex) we will consider another case of non-convex constraint: low-rankness of the matrix. The problem is very similar in nature, where the sparsity of the vector is analogous to the low-rankness of the matrix. We will provide motivation, background and solutions of low rank optimization problem which we will define as Matrix Sensing.

Low rankness, Matrix sensing

Motivation-Quantum state Tomography. We will define the problem with a motivation from Quantum state tomography (QST), where the goal is to test weather the algorithm in quantum computer works well. A quantum computer is a non-deterministic machine, where we don't know the final state, unless we measure it (this is where Schroedinger's cat come into the picture!). From QST we only have measurements of final state, we cannot see the intermediate state as observation can destroy change these intermediate states. But if we perform the steps "correctly", w.h.p. we measure the anticipated state we can repeat the measurement many times, we keep the data, and we try to inverse the procedure to get the value of intermediate state.

The intermediate state is represented by the density matrix $X^* \in \mathbb{R}^{2^q \times 2^q}$ of the quantum circuits, also called q-cubit state. The measurements are the expected value of q-cubit Pauli's observables $A_i \in \mathbb{C}^{2^q \times 2^q}$. The measurement vector $y \in \mathbb{R}^m$ is:

$$y = Tr(A_i.X^*) + e_i, \quad i = 1, ...m$$

for some error e_i.
Classical quantum state tomography is like solving linear equations; if we have a $O(4q)$ object to recover, we need that many measurements. Currently the best computer in the world is using $q = 53$. This makes the size of X^*, very very very large! If we don't assume anything about state X, then the number of measurements needed is so large $(2^{53} \times 2^{53})$ we can't even imagine!

But many times the state X is positive semi-definite with low rank. These states are called pure states which can be considered as a first step before going into more mixed states. Theoretically, we can assume rank-1 constructed density matrices; noise + other phenomena increases the rank in practice. We will see that for rank r, $m \times m$ matrix we need $O(mr)$ measurements.

To give an idea of measurement model:

$$A_i = \sigma_{i_1} \otimes \sigma_{i_2} \otimes \cdots \otimes \sigma_{i_q}$$

$A_i's$ are knonecker product of Pauli's operators, which looks like:

$$\sigma_I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

It is also important to note that with this formulation $A_i's$ follow Restricted Isometry Property. We will see why it is important later in the lecture. The generative model of QST is given by:

$$y = Tr(A_i.X^*) + e_i, \quad i = 1, ...m \qquad [1]$$

such that $X \succeq 0, \mathrm{Tr}(X) \leq 1, \mathrm{rank}(X) \leq r$ and $A_i \in \mathbb{R}^{p \times p}$.

Matrix sensing. Formally, the matrix sensing problem is as follows. Given a measurement mechanism $\mathcal{A}$ with m measurement matrices A_i, matrix sensing requires a solution to an optimization problem.

$$X = \arg \min_{rank(\mathbf{X}) \leq r} \frac{1}{2} \sum_{i=0}^{m-1} (y_i - \langle \mathbf{A}_i, \mathbf{X} \rangle)^2$$

Where the linear measurements are y_i, $i \in [0, m]$, which is assumed to be generated by the model $y_i = \langle A_i, X^\star \rangle$, where $X^\star \in \mathbb{R}^{p \times p}$. Without constraints or a structural assumption on $\mathbf{X}^\star$, the problem is under-determined with infinite solutions. However, given a rank r matrix and a sufficiently large number of measurements, there is a unique solution that may be found via optimization.

Restricted Isometry Property. A linear operator $A : \mathbb{R}^{p \times p} \to R^m$ satisfies the RIP on rank-r matrices, with parameter $\delta_r \in (0, 1)$, if the following holds for all rank-r $X \in \mathbb{R}^{p \times p}$:

$$(1 - \delta_r)\|X\|_F^2 \leq \|A(X)\|_2^2 \leq (1 + \delta_r)\|X\|_F^2$$

Isometry is a property which says that the all the distances are preserved after the transformation by matrix A on X. Restricted isometry is relaxed version of isometry where the distances are preserved with an error of $(1 \pm \delta_r)\|X\|_F^2$. A measurement matrix is useful for the matrix sensing problem if it is a near-isometry (satisfies RIP).
Without the assumption of A_i being RIP, finding low rank solution in matrix sensing problem is NP-hard. It's like checking all the low rank solutions in combinatorial way. But with the statistical assumption like RIP we hope to find solution in polynomial time.

To solve the matrix sensing problem (for now consider QST formulation 1 without psd and trace constraint) we have two approaches:

1) Convexification- Nuclear Norm Minimization
We have the objective:

$$\min_{X \in \mathbb{R}^{p \times p}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle A_i, X \rangle)^2$$

Given the convex constraint, $\|X\|_* \leq \lambda$.
The **nuclear norm** of a matrix is represented by the summation of singular values, $\|X\|_* = \sum_{i=0}^{r} \sigma_i$. Note that the

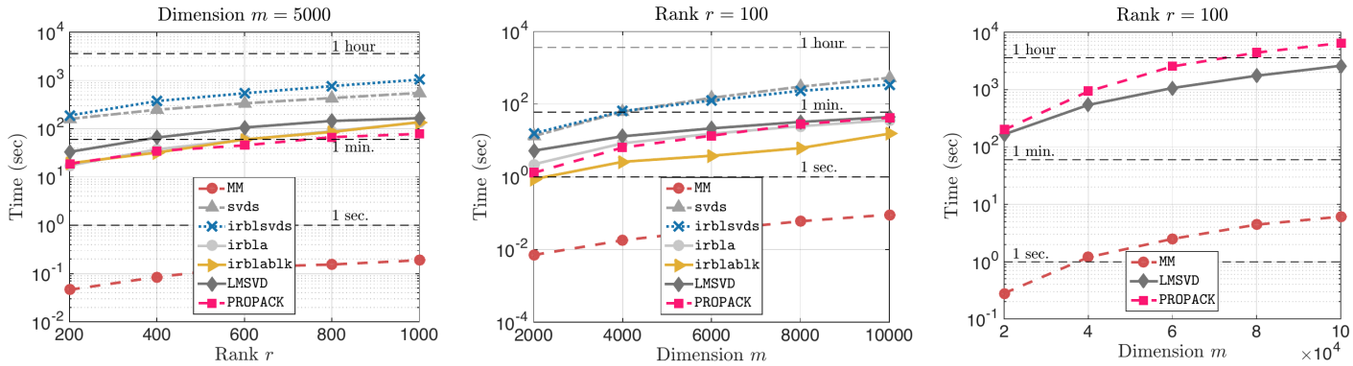


Fig. 1. Comparison between matrix multiplication (MM), i.e. $X.U$ where $X \in \mathbb{R}^{m\times m}$ and $U \in \mathbb{R}^{m\times r}$ and SVD (all methods of truncated SVD) of X. In both the scenarios (varying rank r and varying dimension m), there is a big gap in SVD and matrix multiplication calculation. Theoretically the complexity is same for both operation but there are many matrix multiplications involved in SVD.

nuclear norm is the L1 norm of the vector of singular values. This will help in giving a direct analogy to finding sparse vector solution.

In each iteration, the gradient descent step is followed by projection over the Nuclear Norm constraint set (it is basically projected gradient descent). The update looks like:

$$X_{t+1} = \Pi_{\|\cdot\|_* \le \lambda}\left(X_t - \eta\nabla f\left(X_t\right)\right)$$

The projection is taken by calculating SVD (or by power iteration) at each step.

Why this surrogate can help?

First Nuclear Norm is convex. Second, as the singular values are non-negative and follows an exponential decay distribution, by truncating sum of the singular values we are not pushing all values down but killing the small singular values. Note that in practice, we might not get the exact rank r solution from this. But the singular values profile will be similar to desired solution. Another thing to note here, we need to know λ equivalent of r.

2) Iterative Hard Thresholding

Here we keep the rank-constraint and directly project onto low rank solutions by calculating SVD (or truncated SVD) at each step.

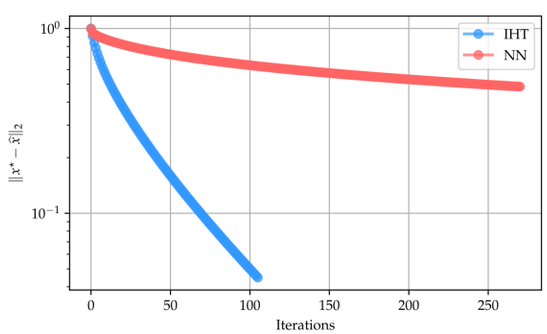


Fig. 2. Demo on Iterative Hard Thresholding and Nuclear norm minimization. $p = 128$, $r = 2$. Refer to pyhton notebook

We have the objective:

$$\min_{X\in\mathbb{R}^{p\times p}} \quad \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \langle A_i, X\rangle\right)^2$$
$$\text{s.t.} \qquad \operatorname{rank}(X) \le r$$

Given the convex constraint,

$$X_{t+1} = \Pi_{\operatorname{rank}(X)\le r}\left(X_t - \eta\nabla f\left(X_t\right)\right)$$

Projection onto low-rank matrices

$$\widehat{X} \in \min_{X} \quad \frac{1}{2}\|X - Y\|_F^2$$
$$\text{s.t.} \quad \operatorname{rank}(X) \le r$$

Here we use the Matrix IHT by changing notations from vectors to matrices, and compute the top r SVD decompositions of the given matrix:

$$X_{t+1} = H_r\left(X_t - \eta\nabla f\left(X_t\right)\right)$$
$$\text{where } H_r\left(Z\right) \in \min_{X\in\mathbb{R}^{p\times p}} \|X - Z\|_F^2$$
$$\text{s.t. } \operatorname{rank}(X) \le r$$

The theorems of IHT regarding step size, initial point and sparsity in previous lectures still hold here.

The SVD calculation is a costly process. Figure 1 is restating this fact. We can get around calculating SVD at each step by using the matrix factorization approach. To understand this we will consider a simpler rank-1 case.

Rank-1 Matrix Approximation Through Rank-1 PCA. Consider a simpler objective of matrix factorization where we are minimizing the following objective.

$$\min_{X\in\mathbb{R}^m, W\in\mathbb{R}^n} \left\|M - XW^\top\right\|_F^2, \quad M \in \mathbb{R}^{m\times n}$$

Where M is rank-1 matrix, $\|X\|_2 = \|W\|_2 = 1$. The objective can also be vectorized:

$$\left\|M - XW^\top\right\|_F^2 = \left\|\operatorname{vec}(M) - \operatorname{vec}\left(XW^\top\right)\right\|_2^2$$
$$= |Y - \mathcal{A}(X)\|_2^2$$

Where, $Y = \operatorname{vec}(M)$. Hence it is a matrix sensing formulation with rank-1 approximation. To simplify this problem

$$\left\|M - XW^\top\right\|_F^2 = \|M\|_F^2 - 2X^\top MW + \left\|XW^\top\right\|_F^2$$
$$= \|M\|_F^2 - 2X^\top MW + \|X\|_2^2 \cdot \|W\|_2^2$$

Hence our objective is:

$$f(x) = \min_{X\in\mathbb{R}^m, W\in\mathbb{R}^n} -2X^\top MW + \|X\|_2^2 \cdot \|W\|_2^2$$

Is a minimization over X and W. So we can do alternate minimization (fix one and minimize over other).
Assuming we know X optimal minimizing over W is convex.
Let

$$f(x) = \min_{W\in\mathbb{R}^n} -2X^\top MW + \|X\|_2^2 \cdot \|W\|_2^2,$$

where $-2X^\top MW$ is linear term and $\|X\|_2^2 \cdot \|W\|_2^2$ is quadratic, thus by convexity

$$\nabla_w = 0 \Rightarrow -2M^T X + 2\|X\|_2^2 \cdot W$$

$$W = \frac{M^T X}{\|X\|_2^2}$$

Substituting W in original problem.

$$f(X) = -\frac{2X^\top MM^\top X}{\|X\|_2^2} + \|X\|_2^2 \cdot \frac{X^\top MM^\top X}{\|X\|_2^L}$$

$$f(X) = -\frac{X^\top MM^\top X}{\|X\|_2^2}$$

Thus the original problem is equivalent to:

$$\min_{X\in\mathbb{R}^m} f(X) := -\frac{X^\top MM^\top X}{\|X\|_2^2}$$

Length of X does not matter, only its direction matters. To see this, define temporarily $Y = \frac{X}{\|X\|_2^2}$.
Then

$$\min_{x\in\mathbb{R}^m} f(x) \equiv \min_{y\in\mathbb{R}^m, \|y\|=1} -y^\top MM^\top y$$

This problem is non-convex now, where the objective can be perceived as finding the max eigen value of $MM^\top$.
How the objective looks like?: To maximize on a bowl (it not a perfect bowl because of different eigen-values) with a constraint of a ring on it. The solution is unique because we assume no two eigen value are same.

Via the inner product expression:

$$\langle x, u_1\rangle = \cos(\theta) \cdot \|u_1\| \cdot \|x\|_2$$

Since θ depends on x, and $\|u_1\|_2 = 1$, we have:

$$\theta(x) = \cos^{-1}\left(\frac{1}{\|x\|_2} < x, u_1 >\right)$$

To solve this optimisation problem we will do gradient descent (can't do SVD/power-iteration as we are trying to avoid it).

$$X_{t+1} = X_t - \eta \cdot \nabla f(X_t)$$

Applying quotient rule

$$\nabla f(X_t) = \frac{1}{\|X\|_2^4} \cdot \left[\nabla_X\left(-X^\top MM^\top X\right)\|X\|_2^2 + X^\top MM^\top X \cdot \nabla_X \|X\|_2^2\right]$$

$$= \frac{1}{\|X\|_2^4}\left[-2\|X\|_2^2 \cdot MM^\top X + 2\left(X^\top MM^\top X\right)\cdot X\right]$$

$$= \frac{2}{\|X\|_2^4}\left[\left(X^\top MM^\top X\right)X - \|X\|_2^2 \cdot MM^\top X\right]$$

We can write M as

$$M = \sum_{i=1}^{min\{m,n\}} \sigma_i \cdot U_i V_i^T$$

$$\sigma_1 > \sigma_2 \geqslant \ldots \geqslant 0$$

. The solution corresponds to the biggest/first singular value.

Key observation for gradient descent on PCA is that if $\langle X_t, u_1\rangle = 0$, then $\langle X_{t+1}, u_1\rangle = 0$, implies we are going orthogonal to the eigen vector u_1. To prove it:

$$\langle X_{t+1}, u_1\rangle = \langle X_t - \eta\nabla f(X_t), u_1\rangle = -\eta\langle \nabla f(X_t), u_1\rangle$$

$$\langle \nabla f(X_t), u_1\rangle = \frac{2}{\|X_t\|_2^4}\left[\left(X_t^\top MM^\top X_t\right)X_t^\top u_1 - \|X_t\|_2^2 X_t^\top MM^\top u_1\right]$$

$$= -\frac{2}{\|X_t\|_2^4}\cdot\|X_t\|_2^2 \quad X_t^\top MM^\top u_1$$

$$= -\frac{2}{\|X_t\|_2^2}\cdot X_t^\top \cdot \sum_i \sigma_i^2 u_i u_i^\top u_1$$

$$= -\frac{2}{\|X_t\|_2^2}\sigma_1^2 X_t^\top u_1 = 0$$

i.e.,

$$\langle X_{t+1}, u_1\rangle = -\frac{2}{\|X_t\|_2^2}\sigma_1^2 X_t^\top u_1 = 0$$

Remark:

i) If X_t is orthogonal to u_1 then X_{t+1} is also orthogonal to u_1. This is a no improvement state.
ii) This further means that if we start from a point such that $\langle X_0, u_1\rangle = 0$, we fail to recover u_1. We will be trapped in saddle point here.
iii) However, maybe there is a hope, if we start from any point not orthogonal to U_1. This is like selecting a point not from the vector span of $u_i, i \neq 1$. A randomly selected point $X_0 \in \mathbb{R}^m$ almost surely has non-zero component on the span of u_1.

We want to study the behavior of the potential function.
Define a potential function

$$\psi_{t+1} = 1 - \frac{\langle X_{t+1}, u_1\rangle^2}{\|X_{t+1}\|_2^2}$$

Intuition: if $\psi_{t+1} \to 0$, x_{t+1} aligns with u_1
and

$$\frac{\langle X_{t+1}, u_1\rangle^2}{\|X_{t+1}\|_2^2} \to 1,$$

which is the optimal thing to achieve for normalized vectors.

We have the following:

$$\|x_{t+1}\|_2^2 = \|x_t - y\nabla f(x_t)\|_2^2$$

$$= \|X_t\|_2^2 - 2\eta X_i^\top \nabla f(X_t) + \eta^2 \cdot \|\nabla f(X_0)\|_2^2$$

Observe that:

$$\begin{aligned} X_t^\top \nabla f(X_t) &= \frac{2}{\|X_t\|_2^4}\left(\left(X_t^\top M M^\top X_t\right)\cdot \|X_t\|_2^2\right. \\ &\quad \left. - \|X_t\|_2^2\left(X_t^\top M M^\top X_t\right)\right) \\ &= 0 \end{aligned}$$

Hence

$$\begin{aligned} \|X_{t+1}\|_2^2 &= \|X_t\|_2^2 - 2\eta X_i^\top \nabla f(X_t) + \eta^2 \cdot \|\nabla f(X_0)\|_2^2 \\ &= \|X_t\|_2^2 + \eta^2 \cdot \|\nabla f(X_t)\|_2^2 \end{aligned}$$

Then

$$\Psi_{t+1} = 1 - \frac{\langle X_{t+1}, u_1\rangle^2}{\|X_{t+1}\|_2^2}$$

$$= 1 - \frac{\langle X_t, u_1\rangle^2 - 2\eta\cdot\langle X_t, u_1\rangle\langle \nabla f(X_t), u_1\rangle + \eta^2.\langle \nabla f(X_1\rangle, u_1\rangle}{\|X_t\|_2^2 + \eta^2\cdot\|\nabla f(X_t)\|_2^2}$$

We know that

$$\frac{\langle X_t, u_1\rangle^2}{\|X_{t+1}\|_2^2} = \left\langle \frac{X_{t+1}}{\|X_{t+1}\|_2}, u_1 \right\rangle^2 = \cos^2(\theta(X_{t+1}))$$

And $1 - \cos^2(\theta(X+1)) = \sin^2(\theta(X_t+1))$.

Using these facts, it turns out that

$$\begin{aligned} \sin^2(\theta(X_{t+1})) &\le \sin^2\theta(X_t) + \frac{\eta^2}{\|X_t\|_2^2}\cdot\|\nabla f(X_t)\|_2^2 \\ &\quad + \frac{2\eta}{\|X_t\|_2^2}\cdot\langle X_t, u_1\rangle\langle \nabla f(X_t), u_1\rangle. \end{aligned}$$

1) For the inner product $\langle \nabla f(x_t), u_1\rangle$ we have:

$$\langle \nabla f(X_t), u_1\rangle \leqslant -\frac{2}{\|X_t\|_2}\left(\sigma_1^2 - \sigma_2^2\right)\cdot\sin^2\theta(X_t)\cdot\cos\theta(X_t) \le 0$$

2) For $\|\nabla f(X_t)\|_2^2$ we have:

$$\|\nabla f(X_t)\|_2^2 \le \frac{4}{\|X_t\|_2^2}\left(\sigma_1^4 + \sigma_2^4\right)\cdot\sin^2\theta(X_t)$$

Then we will get:

$$\begin{aligned} \sin^2(\theta(X_{t+1})) &\le \sin^2(\theta(X_t))\left(1 + \frac{4\eta^2}{\|X_1\|_2^4}\left(\sigma_1^4 + \sigma_2^4\right)\right. \\ &\quad \left. - \frac{4\eta}{\|X_t\|_2^2}\cdot\left(\sigma_1^2 - \sigma_2^2\right)\cdot\frac{\langle X_t, u_1\rangle^2}{\|X_t\|_2^2}\right) \end{aligned}$$

We will provide local convergence guarantees if given a proper initialization, we get convergence to global minimum.

1) If $\frac{\langle X_t, u_1\rangle^2}{\|X_t\|_2^2} \ge c$, such that $0 \le c < 1$,

we obtain:

$$\sin^2\theta(X_{t+1}) \le \sin^2\theta(X_t)\left(1 + \frac{4\eta^2}{\|X_t\|_2^4}(\sigma_1^4 + \sigma_2^4) - \frac{4\eta}{\|X_t\|_2^2}(\sigma_2^2 - \sigma_2^2)c\right)$$

2) Select $\eta = \frac{c}{2}\cdot\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^4 + \sigma_2^4}\cdot\|x_t\|_2^2$

Then

$$p = 1 + c^2\cdot\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2} - 2\cdot c^2\cdot\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2} = 1 - c^2\cdot\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2} < 1$$

Thus

$$\sin^2(\theta(X_{t+1})) \le \rho\sin^2(\theta(X_t))$$

Where $\rho < 1$. We achieve linear convergence $O(\log\frac{1}{\epsilon})$, without using SVD at any step.

Some properties of the proof:
1) Initialization does matter: e.g., for PCA there are initializations that do not lead to convergence.
2) After proper initialization, one can prove convergence to global minimum. Despite this, such convergence results are called local convergence guarantees.
3) Often the theory dictates how to set the step size, in order to obtain convergence. For some cases it is a range of values, in other cases we just rely on a specific step size.
It gives a good motivation that matrix factorization is useful.

Alternate minimization. Back to original problem, we have matrix sensing objective

$$\min_{X\in\mathbb{R}^{p\times p}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle A_i, X\rangle)^2$$

With low rank constraint $rank(X) < r$. Instead of that we can represent X as

$$X = UV^T$$

The objective now is constraint free

$$X = \arg\min_{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{p\times r}} \frac{1}{2}\sum_{i=0}^{m-1}(y_i - \langle \mathbf{A}_i, \mathbf{UV^T}\rangle)^2$$

Key differences with PCA: 1) Number of observations are less than number of parameters, 2) Mapping $\mathcal{A}$ is not identity, but satisfies a restricted isometry property.
Now if we not restrict the objective to least squares:

$$X = \arg\min_{U\in\mathbb{R}^{m\times r_r}, V\in\mathbb{R}^{n\times r}} f\left(UV^\top\right)$$

Here Restricted isometry can be substituted by Restricted Strong Convexity.

To solve this perform, we perform alternate minimization (not in true sense, as we are not using updated U_{i+1} from first step in the second step). The method is also called **Factored Gradient descent**.

$$U_{i+1} = U_i - \eta\nabla f\left(U_i V_i^\top\right)\cdot V_i$$

$$V_{i+1} = V_i - \eta\nabla f\left(\left(U_i V_i^\top\right)^\top\cdot U_i\right)$$

Although we have constraint less optimization problem now, factorization brings another problem. Objective is not convex. There are new saddle points, global and local minimas introduced.

$$X^\star = U^\star V^{\star T} = U^\star R\cdot R^T V^{\star T} = \widehat{U}^\star\widehat{V}^{\star T}$$

For all R such that $RR^T = I$
For example, if:

$$f(X) = \frac{1}{2} \cdot \|y - \text{vec}(A \cdot X)\|_2^2$$

Where

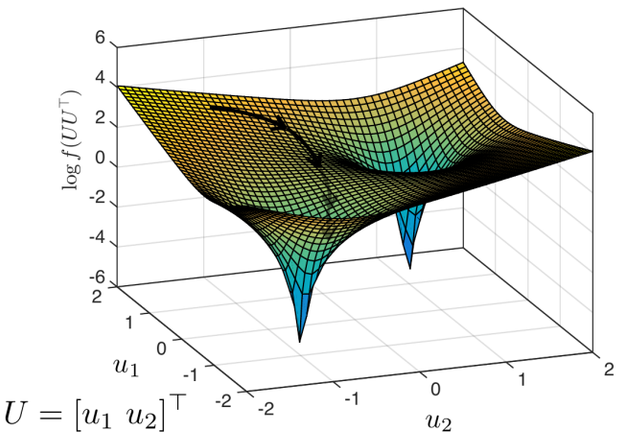
$$X^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

is a unique solution with $r = 1$

$$U^\star = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top \text{ or } [-1 - 1]^\top$$

Multiple factorizations are possible. Hence it ruins convexity.

$$f(UU^\top) = \tfrac{1}{2}\|y - \texttt{vec}(A \cdot UU^\top)\|_2^2$$

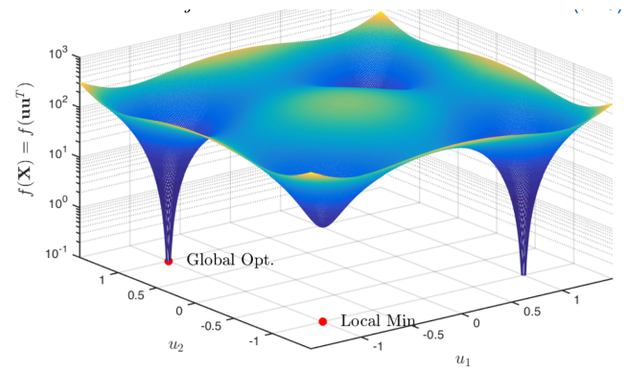


Another example:
Weighted low-rank approximation

$$f\left(uu^\top\right) = \sum_{ij} W_{ij} \cdot \left(X^\star_{ij} - u_i u_j\right)^2$$

where

$$X^\star = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} 100 & 1 \\ 1 & 100 \end{bmatrix}$$



As there is non-convexity introduced, proper initialization is the key.

To find some guarantees on convergence: We will start with general recipe of proving convergence.

$$\begin{aligned} \|x_{t+1} - x^\star\|_\#^2 &= \|x_t - \eta \nabla f(x_t) - x^\star\|_\#^2 \\ &= \|x_t - x^\star\|_\#^2 - 2\eta \langle \nabla f(x_t), x_t - x^\star \rangle + \\ &\eta^2 \|\nabla f(x_t)\|_\#^2 \end{aligned}$$

Where # is norm, indicates a general class of distance functions. The geometric intuition of $\langle \nabla f(x_t), x_t - x^\star \rangle$:

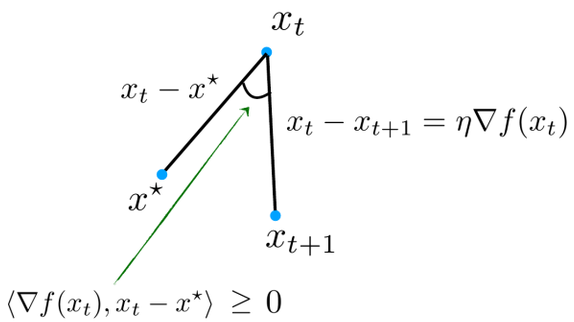


Fig. 3. Angle between the direction of gradient and correct direction should be less than $\pi/2$

We need the following to hold true to **bound** $\|x_{t+1} - x^\star\|_\#^2$

$$\langle \nabla f(x_t), x_t - x^\star \rangle \geq \alpha \|x_t - x^\star\|_\#^2 + \beta \|\nabla f(x_t)\|_\#^2$$

for sufficient $\alpha, \beta \geq 0$ such that

$$\|x_t - x^\star\|_\#^2 - 2\eta \langle \nabla f(x_t), x_t - x^\star \rangle + \eta^2 \|\nabla f(x_t)\|_\#^2$$

$$\leq \|x_t - x^\star\|_\#^2 - c\alpha\eta \|x_t - x^\star\|_\#^2 - \left(c\eta\beta - \eta^2\right) \|\nabla f(x_t)\|_\#^2$$

Now this has some connections with the convex optimization problem we have seen so far.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

If $y = x^*$ and since $\nabla f(x^*) = 0$

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{\mu L}{\mu + L} \|x - x^*\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x)\|_2^2$$

It encourages that our approach for get a bound is correct.

For simplicity now consider X to be positive semi definite. Hence $X = UU^\top$.
Define a **distance metric**, where the distance between any arbitrary matrix U and U^*, $DIST$ is defined as:

$$DIST(U, U^*) = \min_{R : R \in O_r} \|U - U^* R\|_F$$

O is the set of $r \times r$ orthonormal matrices R, such that $R^T R = I$. There can be infinite U^* but we need an optimal U where distance is one with the closest U^* upto one rotation R. The DIST will help us getting a good initial point.

Then we have

$$\begin{aligned} \text{DIST}\left(U_{t+1}, U^*\right)^2 = \min_{R \in O_r} \|U_{t+1} - U^* R\|_F^2 \leqslant \|U_{t+1} - U^* R_t\|_F^2 \\ = \|U_{t+1} - U_t + U_t - U^* R_t\|_F^2 \\ = \|U_{t+1} - U_t\|_F^2 + \|U_t - U^* R_t\|_F^2 + 2\langle U_{t+1} - U_t, U_t - U^* R_t \rangle \\ = \eta^2 \|\nabla f(U_t U_t^\tau) U_t\|_F^2 + \|U_t - U^* R_t\|_F^2 \\ -2\eta \langle \nabla f(U_t U_t^*) U_t, U_t - U^* R_t \rangle \end{aligned}$$

Key result is the fact that we can prove a regulatory condition:

$$\left\langle \nabla f\left(UU^*\right)U, U - U^k R\right\rangle \geqslant \frac{2}{3}\eta \cdot \left\|\nabla f\left(UU^\top\right)U\right\|_F^2 + \frac{3\mu}{20}\sigma_r\left(X^*\right)\cdot \mathrm{DIST}\left(U_t, U^*\right)^2$$

Using the last two equations, we have:

$$DIST\left(U_{t+1}, U^*\right)^2 \leqslant DIST\left(U_t - U^* R_t\right)^2 + \eta^2 \cdot \left\|\nabla f\left(U_t U_t^\top\right)U_t\right\|_F^2$$

$$-\frac{4}{3}\eta^2\left\|\nabla f\left(U_t U_t^\top\right)U_t\right\|_F^2 - \frac{6\mu\eta}{20}\sigma_r\left(X^*\right)\cdot DIST\left(U_t, U^*\right)^2$$

$$\leqslant \left(1 - \frac{3\mu\eta}{10}\sigma_r\left(X^*\right)\right)\cdot DIST\left(U_t, U^*\right)^2$$

This defines the step size η.

In practice, the paper "Dropping Convexity for Faster Semidefinite Optimization" has more sophisticated but more practical η.

However, in order to prove the regulatory condition,
we require

$$DIST(U_t, U^*) \leq \rho \cdot \sigma_r(X^*)^{\frac{1}{2}}$$

for all t,
which means

$$DIST(U_t, U^*) \leq \rho \cdot \sigma_r(X^*)^{\frac{1}{2}}$$

leads to good initialization.

As we have gone from convex to non-convex regime, now we have created a dependence over the singular values of X^*, which we do not know.

At the end it gives the following convergence guarantee:

THEOREM: LOCAL CONVERGENCE

Theorem 1. *If f is a "nice" function and (U_i, V_i) are sufficiently close to $U^\star, V^\star$) , then non-convex alternating gradient descent i) converges to $(U^\star, V^\star)$, and ii) achieves the same convergence guarantees with convex optimization:*

Theorem 2. *Global convergence with better initialization: If the function f is "well-conditioned" , then non-convex alternating gradient descent converges to the global optimum / optima.*

i.e in $O(\frac{1}{\epsilon})$ or in $O(\log \frac{1}{\epsilon})$ we will have

$$f\left(\widehat{U}\widehat{V}^\top\right) - f\left(U^\star V^{\star\top}\right) \leq \varepsilon$$

Goal: Initialize such that (U_0, V_0) is sufficiently close to (U^*, V^*)

Proposed initialization

1) Compute $X_0 \propto \nabla f(0_{n\times p})$
2) Perform one SVD calculation:

$$X_0 = U_0 V_0^T$$

If the function f is "well-conditioned" , then non-convex alternating gradient descent converges to the global optimum / optima.

The impact here will be: instead of SVD at each step we calculate SVD for first step. The guarantees are weak, but often it works in practice!

Appendix

1. **Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. Dropping convexity for faster semi-definite optimization. In Conference on Learning Theory, pp. 530–582, 2016a.**

2. **Sébastien Bubeck, Convex optimization: Algorithms and complexity, Foundations and Trends® in Machine Learning, 8.3-4 (2015), pp. 231-357.**