Lecture #24: No Spurious Local Minima

Last time: Strict saddle property $\Rightarrow$ first/second order methods find a local minimum

But why is finding a local minimum good enough?

Today: No spurious local minima in matrix sensing, following [Ge, Jin, Zheng]

Matrix Sensing: Unknown, low-rank matrix M (not necessarily incoherent)

We observe linear measurements

$$\langle A_i, M \rangle = b_i \quad \text{for } i = 1 \text{ to } m$$

Can we recover M exactly?

def: Sensing matrices $A_1 \dots A_m$ satisfy the (r, δ)-matrix RIP condition, if

$$(1-\delta)\|M\|_F^2 \leq \frac{1}{m}\sum_i \langle A_i, M \rangle^2 \leq (1+\delta)\|M\|_F^2$$

for all M with $\text{rank}(M) \leq r$

When $A_i = \text{diag}(a_i)$ then condition is equivalent to:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$ having the usual vector RIP property for all r-sparse vectors

Note: Matrix completion is the special case of matrix sensing when each A_i has one 1 and the rest are zeros

Q: Do such A_i's have the matrix RIP property?

No, can set M to have r 1s in locations where all A_i's are zero

But dense, random A_i's do have matrix RIP:

Theorem [Candes, Plan] If each A_i has i.i.d $N(0,1)$ entries after normalization whp the A_i's satisfy $(2r, \frac{1}{20})$-matrix RIP for $m \geq c(n_1+n_2)r$

(n_1, n_2: dimensions of M)

We will prove the following:

Theorem [informal] If $A_1 \ldots A_m$ have the $(2r, \frac{1}{20})$ matrix RIP then matrix sensing has no spurious local minima

Which objective function for matrix sensing are we talking about?

$$f_1(\hat{U}) = \frac{1}{2} \sum_i \langle A_i, \underbrace{\hat{U}\hat{U}^T}_{n \times r} - \underbrace{M}_{\text{let's assume } M \text{ is symmetric}} \rangle^2$$

i.e. $M = UU^T$

We will show $\hat{U}$ is close to being a local minima of $f_1 \Rightarrow \hat{U}\hat{U}^T$ is close to M

We will also work with the unconstrained objective:

$$f_2(\underbrace{\hat{M}}_{\hat{U}\hat{U}^T}) = \frac{1}{2}\sum_i \langle A_i, \hat{M} - M\rangle^2$$

where we want to minimize f_2 over rank-r matrices

Idea: For any $\hat{U}$ <u>far from</u>* U where the gradient $\nabla f_1(\hat{U})$ is small, that the Hessian

$$\nabla^2 f_1(\hat{U})$$

has a negative eigenvalue.

First, what do we mean by far? For any orthogonal matrix Z, we have

$$(UZ)(UZ)^T = UU^T = M$$

Let $R = \underset{\text{orthogonal } Z}{\operatorname{argmin}} \|\hat{U} - UZ\|_F^2$

Now how do we prove the Hessian has a negative eigenvalue?

We will find a direction of improvement, set

$$\Delta = \hat{U} - UR$$

then we will show that $\langle \Delta, \nabla^2 f_1(\hat{U})\Delta\rangle < 0$

We need the following helper lemma:

Lemma: With Δ defined as above

$$\|\Delta\Delta^T\|_F^2 \leq 2\|\hat{M} - M\|_F^2$$

and $\sigma_r \|\Delta\|_F^2 \leq 2\|\hat{M} - M\|_F^2$

σ_r: r^{th} smallest singular value of M

This is straight forward to prove. Consider the special case when

$$\hat{U}\hat{U}^T = UU^T$$

then use the Cholesky decomposition to show

$$\hat{U} = UR \quad \text{for some } R$$

and prove R is orthogonal. The lemma is a quantitative version of this type of argument

Now the main lemma:

Lemma [Main]

$$\langle \Delta, \nabla^2 f_1(\hat{U})\Delta \rangle = \tfrac{1}{2}\langle \Delta\Delta^T, \nabla^2 f_2(\hat{M})\Delta\Delta^T \rangle$$

$$- \tfrac{3}{2}\langle \hat{M} - M, \nabla^2 f_2(\hat{M})(\hat{M} - M) \rangle + 2\langle \nabla f_1(\hat{U}), \Delta \rangle$$

Proof: It's easiest to compute $f_1(\hat{U} + Z)$ for some matrix Z and pull out linear and quadratic terms in Z to get gradient/Hessian

Then $f_1(\hat{U}+Z) =$

$$\sum_i \langle A_i, \hat{U}Z^T + Z\hat{U}^T \rangle \langle A_i, \hat{M} - M \rangle \quad (1)$$

$$(2) \quad + \frac{1}{2}\sum_i \langle A_i, \hat{U}Z^T + Z\hat{U}^T \rangle^2 + \sum_i \langle A_i, ZZ^T \rangle \langle A_i, \hat{M} - M \rangle$$

then $(1) = \langle \nabla f_1(\hat{U}), Z \rangle$ and

$(2) = \langle Z, \nabla^2 f_1(\hat{U}) Z \rangle$

Now plugging in $Z = \Delta = \hat{U} - UR$ and observe

$$\hat{M} - M + \Delta\Delta^T = \hat{U}\hat{U}^T - UU^T + \hat{U}\hat{U}^T + URR^TU^T$$
$$- \hat{U}R^TU^T - URU^T$$

$$= \hat{U}\hat{U}^T - \hat{U}R^TU^T + \hat{U}\hat{U}^T - URU^T$$

$$= \hat{U}\Delta^T + \Delta\hat{U}^T$$

Thus we have: $\hat{M} - M + \Delta\Delta^T$

$$\langle \Delta, \nabla^2 f_1(\hat{U})\Delta \rangle = \frac{1}{2}\sum_i \langle A_i, \hat{U}\Delta^T + \Delta\hat{U}^T \rangle^2$$
$$+ \sum_i \langle A_i, \Delta\Delta^T \rangle \langle A_i, \hat{M} - M \rangle$$

$$= \frac{1}{2}\sum_i \langle A_i, \Delta\Delta^T\rangle^2 + 2\sum_i \langle A_i, \hat{M}-M+\Delta\Delta^T\rangle \langle A_i, \hat{M}-M\rangle$$

$$-\frac{3}{2}\sum_i \langle A_i, \hat{M}-M\rangle^2$$

$$= 4\langle \nabla f_1(\hat{U}), \Delta\rangle$$

Now what is $\nabla^2 f_2(\hat{M})$?

$$\langle Z, \nabla^2 f_2(\hat{M})Z\rangle = \frac{1}{2}\sum_i \langle A_i, Z\rangle^2$$

and in the formula we only ever apply it to rank $\leq 2r$ matrices

In the main lemma's formula if $\nabla f_1(\hat{U}) = 0$ and the A_i's are $(2r,0)$-matrix RIP we get

$$\langle \Delta, \nabla^2 f_1(\hat{U})\Delta\rangle = \frac{1}{2}\|\Delta\Delta^T\|_F^2 - \frac{3}{2}\|\hat{M}-M\|_F^2$$

using the helper lemma

$$\leq -\frac{1}{2}\|\hat{M}-M\|_F^2 < 0$$

as desired. Thus the only local minima are when

$$\hat{M} = M$$

Can carry through the $(2r, \frac{1}{20})$ term and get a quantitative bound for the strict saddle property

Notice the following theme in linear inverse problems

We always use the direction to M as part of the analysis

(1) in convex programs, to construct the dual certificate

(2) in iterative methods, to show our direction of movement has reasonable inner product with $M - \hat{M}$

(3) to rule out spurious local minima by showing if the gradient is zero and we are far from M, there is a negative eigenvalue (quadratic form of $M - \hat{M}$ on Hessian)