Convex Optimization 10-725, Lecture 19: Introduction to non-convex optimization: Over-parameterization

Yuanzhi Li

Assistant Professor, Carnegie Mellon University
Visiting Researcher, Microsoft

Today

Last lecture

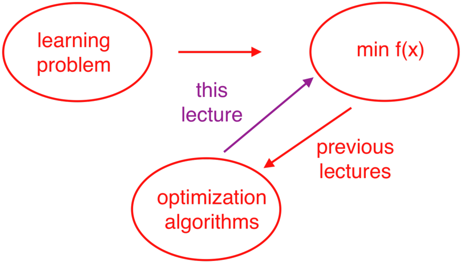
- We learnt Bayesian Optimization

This lecture

- From this lecture, we will focus on special topic series in non-convex optimization.
- Today we are going to study the first topic: over-parameterization.

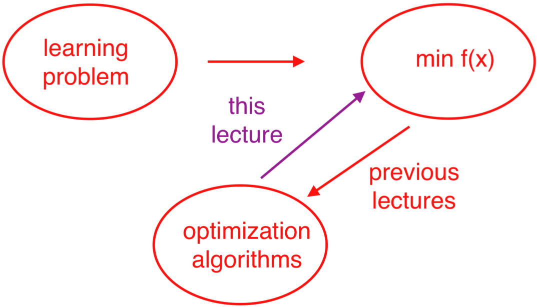
Over-parameterization: Motivation

- Through out the class, we have been focusing on how to minimize/maximize a given function f.
- However, in machine learning, our goal is to solve a learning problem, optimization is an intermediate step.
- Starting from this lecture, we are going to answer the following extremely important, and fundamental question:
- To solve solve the same learning problem, can we design a function f where minimizing f is easier?



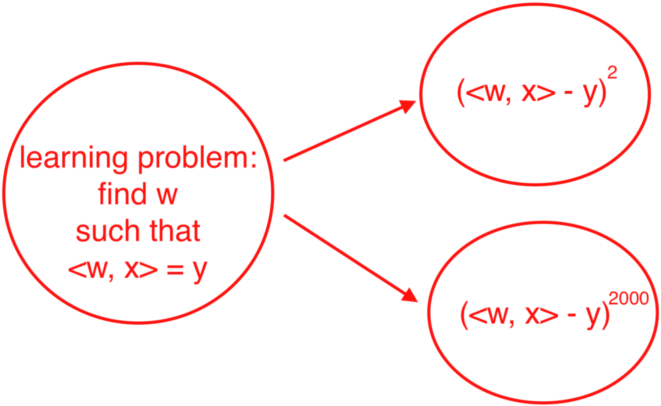
-

Over-parameterization: Motivation



-
- This is the art side of optimization.
- Today we are going to learn the first art: over-parameterization.

Over-parameterization: Motivation



-
- Different parameterizations of the learning problem can make the underlying optimization tasks easier/harder.

Parameterization: Example

- This is a practical example:
- The problem of matrix sensing (matrix completion): Given a set of sensing matrix $A_1, A_2, \cdots, A_N \in \mathbb{R}^{d\times d}$, and observations $y_1, y_2, \cdots, y_N$:
- The matrix sensing asks to find the matrix U, V with the smallest Frobenius norm such that for all $i \in [N]$:

$$\langle A_i, UV^\top \rangle \approx y_i$$

- In other words, the optimization problem associated with matrix sensing is

$$\min_{U,V} \frac{1}{N} \sum_{i\in[N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \| U \|_F^2 + \lambda \| V \|_F^2$$

Parameterization: Easy or hard?

- The optimization problem associated with matrix sensing is

$$\min_{U,V} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

- This is a non-convex optimization problem, and in general, there are (a lot of) bad local minima with large objective value.
- There is a re-parameterization with is equivalent to the original problem, but it is convex:

$$\min_{M} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, M \rangle - y_i \right)^2 + \lambda \|M\|_*$$

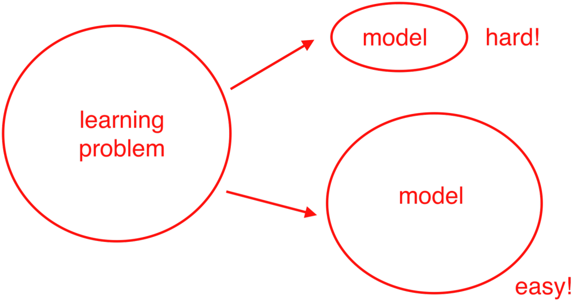
Where $\|M\|_*$ is the trace norm/nuclear norm of M (convex).

Parameterization: Easy or hard?

- For many machine learning problems, there are multiple optimization tasks that can solve the same problem.
- Some of the optimization tasks are easier, some of them are harder.
- Main question: Is there a generic routine to make the optimization task easier?
- Generic: We don't need to have special knowledges about the problem (for example, in matrix sensing, what if we do not know that the non-convex matrix product is equivalent to a nuclear norm).
- Today we are going to see the most generic routine: over-parameterization.

Over-parameterization

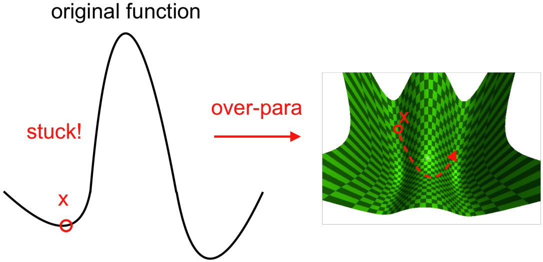
- Principle of over-parameterization: by making the number of parameters in the optimization problem much larger than necessary, it makes the underlying optimization problem easier.
- Intuitively, we change the optimization problem from $\min_{w \in \mathbb{R}^d} f(w)$ to $\min_{w \in \mathbb{R}^D} f(w)$, where $D \gg d$.
- Intuitively, we will have much more directions to search through when optimizing f, and less likely to stuck at a local minima.



-

Over-parameterization

- Principle of over-parameterization: by making the number of parameters in the optimization problem much larger than necessary, it makes the underlying optimization problem easier.
- Here, easier typically means that the new optimization has less or no bad local minima.
- Indeed, for the larger model, the per-iteration optimization cost is higher, but the quality of the solution can also be much higher.
- In most of the non-convex optimization applications, we prefer good



Over-parameterization

- Main question: How do we change the optimization problem from $\min_{w \in \mathbb{R}^d} f(w)$ to $\min_{w \in \mathbb{R}^D} f(w)$, where $D \gg d$?
- For the empirical risk minimization (ERM) problem in machine learning, recall we are given data $\{x_i, y_i\}_{i \in [N]}$, and ask to minimize

$$f(W) = \frac{1}{N} \sum_{i \in [N]} \ell(h(W, x_i), y_i) + \lambda R(W)$$

- Here, ℓ is the loss, $h(W, x)$ is the model with parameters W, R is the regularizer.
- We can simply use a larger model $h(W, x)$.
- For example, in deep learning, we can use a larger neural network.

Over-parameterization: Specific examples

- Intuitively, when we have a larger model, we have more degrees of freedom when searching through the parameter space, and less likely to stuck at a bad local minima.
- In this lecture, we are going to see a proof how over-parameterization can make the optimization problem associated with matrix sensing easy, in the next lecture, we will see a proof how over-parameterization works in deep learning beyond the neural tangent kernel approach.

Over-parameterization: Specific examples

- The optimization problem associated with matrix sensing is

$$\min_{U,V} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

- For $\lambda, \epsilon > 0$, assuming there are (unknown) matrices $U^*, V^* \in \mathbb{R}^{d \times r}$ such that

$$\frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, U^*(V^*)^\top \rangle - y_i \right)^2 + \lambda \|U^*\|_F^2 + \lambda \|V^*\|_F^2 \leq \epsilon$$

- Question: Can we find U, V efficiently such that:

$$\frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2 \leq 1.1\epsilon$$

Over-parameterization in matrix sensing

- The optimization problem associated with matrix sensing is

$$\min_{U,V} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \| U \|_F^2 + \lambda \| V \|_F^2$$

- For $\lambda, \epsilon > 0$, assuming there are (unknown) matrices $U^*, V^* \in \mathbb{R}^{d \times r}$ such that

$$\frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, U^* (V^*)^\top \rangle - y_i \right)^2 + \lambda \| U^* \|_F^2 + \lambda \| V^* \|_F^2 \le \epsilon$$

- Main theorem, in proper-parameterization case: the following objective can have many bad (second order) local minima (with objective values $\gg 1.1\epsilon$):

$$\min_{U,V \in \mathbb{R}^{d \times r}} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \| U \|_F^2 + \lambda \| V \|_F^2$$

- In other words, if you do it properly, you are not doing it properly.

Over-parameterization in matrix sensing

- The optimization problem associated with matrix sensing is

$$\min_{U,V} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

- For $\lambda, \epsilon > 0$, assuming there are (unknown) matrices $U^*, V^* \in \mathbb{R}^{d \times r}$ such that

$$\frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, U^* (V^*)^\top \rangle - y_i \right)^2 + \lambda \|U^*\|_F^2 + \lambda \|V^*\|_F^2 \le \epsilon$$

- Main theorem, in over-parameterization case: all the (second order) local minima U, V's have objective value $\le 1.1\epsilon$ when $R \gg r$, for the following optimization problem:

$$\min_{U,V \in \mathbb{R}^{d \times R}} \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2$$

Over-parameterization in matrix sensing

- The optimization problem

$$\min_{U,V\in\mathbb{R}^{d\times R}} \frac{1}{N}\sum_{i\in[N]}\left(\langle A_i, UV^\top\rangle - y_i\right)^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$$

- For $\lambda, \epsilon > 0$, assuming there are (unknown) matrices $U^*, V^* \in \mathbb{R}^{d\times r}$ such that

$$\frac{1}{N}\sum_{i\in[N]}\left(\langle A_i, U^*(V^*)^\top\rangle - y_i\right)^2 + \lambda\|U^*\|_F^2 + \lambda\|V^*\|_F^2 \le \epsilon$$

- However, when $R = r$: the objective function can still have bad (second order) local minima with objective value $\gg 1.1\epsilon$.
- When $R \gg r$, all (second order) local minima are good (with objective value $\le 1.1\epsilon$).
- $R \gg r$: over-parameterization: Much more parameters in the model than necessary.

Over-parameterization in matrix sensing

- Proof: To show that there are no (second order) local minima when the objective is larger than 1.1ϵ, we just need to show that for every U, V such that

$$f(U, V) = \frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, UV^\top \rangle - y_i \right)^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2 \geq 1.1\epsilon$$

- There exists matrices U', V' and a value $\delta > 0$ such that for all sufficiently small $\eta > 0$,

$$f(U + \eta U', V + \eta V') \leq f(U, V) - \delta \eta^2$$

Over-parameterization in matrix sensing

- Proof of the above statement: Suppose (U, V) is a (second order) local minima, then by definition: $\nabla f(U, V) = 0$ and $\nabla^2 f(U, V) \geq 0$. This implies

$$f(U + \eta U', V + \eta V')$$

$$\geq f(U, V) + \eta \langle \nabla f(U, V), (U', V') \rangle$$

$$+\frac{1}{2}\eta^2 (U', V')^\top \nabla^2 f(U, V)(U', V') - O(\eta^3)$$

$$\geq f(U, V) - O(\eta^3)$$

- Which means that there should not be U', V' and $\delta > 0$ such that for all sufficiently small $\eta > 0$:

$$f(U + \eta U', V + \eta V') \leq f(U, V) - \delta \eta^2$$

Over-parameterization in matrix sensing

- Now we just need to find these matrices U', V'. Intuitively, $U', V' \in \mathbb{R}^{d \times R}$. When $R \gg r$, we have a much larger search space and finding such matrices should be easier.
- Technically, we are actually going to construct distributions P_U, P_V over U', V' such that for all sufficiently small $\eta > 0$:

$$\mathbb{E}_{U' \sim P_U, V' \sim P_V} f\left((1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}\,U', (1-\eta^2)V + \sqrt{2\eta^2 - \eta^4}\,V'\right)$$

$$\leq f(U, V) - \delta\eta^2$$

Over-parameterization in matrix sensing

- Recall: we assume there are (unknown) matrices $U^*, V^* \in \mathbb{R}^{d \times r}$ such that

$$\frac{1}{N} \sum_{i \in [N]} \left(\langle A_i, U^* (V^*)^\top \rangle - y_i \right)^2 + \lambda \| U^* \|_F^2 + \lambda \| V^* \|_F^2 \le \epsilon$$

- We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}} (\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$

$$P_V \sim \frac{1}{\sqrt{k}} (\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.

Over-parameterization in matrix sensing

- We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}}(\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$

$$P_V \sim \frac{1}{\sqrt{k}}(\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.
- In this way, we have that for every $U', V' \sim P_U, P_V$, $\|U'\|_F = \|U^*\|_F, \|V'\|_F = \|V^*\|_F$, and

$$(U')(V')^\top = U^*(V^*)^\top$$

- We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}}(\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$

$$P_V \sim \frac{1}{\sqrt{k}}(\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.
- We first check the changes in the regularizers:
- Key observation:

$$\mathbb{E}[\|(1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}U'\|_F^2] = \|(1-\eta^2)U\|_F^2$$

$$+\mathbb{E}[2(1-\eta^2)\sqrt{2\eta^2 - \eta^4}\langle U', U\rangle] + \mathbb{E}[\|\sqrt{2\eta^2 - \eta^4}U'\|_F^2]$$

Over-parameterization in matrix sensing

- Key observation:

$$\mathbb{E}[\|(1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}U'\|_F^2] = \|(1-\eta^2)U\|_F^2$$

$$+\mathbb{E}[2(1-\eta^2)\sqrt{2\eta^2-\eta^4}\langle U', U\rangle] + \mathbb{E}[\|\sqrt{2\eta^2-\eta^4}U'\|_F^2]$$

- Key observation: $\mathbb{E}[2(1-\eta^2)\sqrt{2\eta^2-\eta^4}\langle U', U\rangle] = 0$,
  $\mathbb{E}[\|\sqrt{2\eta^2-\eta^4}U'\|_F^2] = (2\eta^2-\eta^4)\|U'\|_F^2 = (2\eta^2-\eta^4)\|U^*\|_F^2$
- Therefore, for $\eta' = 2\eta^2 - \eta^4$:

$$\mathbb{E}[\|(1-\eta^2)U + \sqrt{2\eta^2-\eta^4}U'\|_F^2] = (1-\eta')\|U\|_F^2 + \eta'\|U^*\|_F^2$$

Over-parameterization in matrix sensing

- Now we look at the change in the loss term:

$$\mathbb{E}\left[\left(\left\langle A_i, \left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top\right\rangle - y_i\right)^2\right]$$

- Key observation: for $\eta' = 2\eta^2 - \eta^4$:

$$\left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top$$

$$= (1-\eta')UV^\top + \sqrt{\eta'}(1-\eta^2)U'V^\top + \sqrt{\eta'}(1-\eta^2)U(V')^\top + \eta' U'(V')^\top$$

Over-parameterization in matrix sensing

- Recall: We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}}(\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$

$$P_V \sim \frac{1}{\sqrt{k}}(\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.
- Thus, we have:

$$U'(V')^\top = U^*(V^*)^\top$$

Over-parameterization in matrix sensing

- Recall: We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}}(\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$
$$P_V \sim \frac{1}{\sqrt{k}}(\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.
- Thus, for the *cross term* $U' V^\top$, we have: for $V = (V_1, V_2, \cdots, V_k)$:

$$U' V^\top = \frac{1}{\sqrt{k}} \sum_{i \in [k]} \tau_i U^* V_i^\top$$

- This implies that $\mathbb{E}[U' V^\top] = 0$, moreover,

$$\mathbb{E}[\|U' V^\top\|_F^2] = \frac{1}{k} \sum_{i \in [k]} \|U^*\|_F^2 \|V_i\|_F^2 = \frac{1}{k} \|U^*\|_F^2 \|V\|_F^2$$

Over-parameterization in matrix sensing

- Recall: We construct the distribution P_U, P_V as ($R = kr$ for a large integer $k > 0$):

$$P_U \sim \frac{1}{\sqrt{k}}(\tau_1 U^*, \tau_2 U^*, \cdots, \tau_k U^*)$$

$$P_V \sim \frac{1}{\sqrt{k}}(\tau_1 V^*, \tau_2 V^*, \cdots, \tau_k V^*)$$

- Where $\tau_i \sim \text{Uniform}(\{-1, 1\})$ are i.i.d. random variables.
- Now we have for the *cross term*: $\mathbb{E}[U' V^\top] = 0$ and

$$\mathbb{E}[\|U' V^\top\|_F^2] = \frac{1}{k} \sum_{i \in [k]} \|U^*\|_F^2 \|V_i\|_F^2 = \frac{1}{k} \|U^*\|_F^2 \|V\|_F^2$$

- This implies that when $k \to \infty$, $U' V^\top \to 0$. Mathematically speaking, this is the role of over-parameterization $(R = kr)$, for large k.

Over-parameterization in matrix sensing

- Back to the change in the

$$\mathbb{E}\left[\left(\left\langle A_i, \left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top\right\rangle - y_i\right)^2\right]$$

- Key observation: for $\eta' = 2\eta^2 - \eta^4$:

$$\left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top$$
$$= (1-\eta')UV^\top + \sqrt{\eta'}(1-\eta^2)U'V^\top + \sqrt{\eta'}(1-\eta^2)U(V')^\top$$
$$+\eta' U^*(V^*)^\top$$

- Recall we just showed: When $k \to \infty$, $U(V')^\top, U'V^\top \to 0$. This implies that

$$\left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top$$
$$= (1-\eta')UV^\top + \eta' U^*(V^*)^\top + \sqrt{\eta'}\xi$$

- Where $\mathbb{E}[\xi] = 0$, and $\xi \to 0$ as $k \to \infty$.

Over-parameterization in matrix sensing

- Now we have:

$$\left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top$$
$$= (1-\eta')UV^\top + \eta' U^*(V^*)^\top + \sqrt{\eta'}\xi$$

- Together, the change in the loss:

$$\mathbb{E}\left[\left(\left\langle A_i, \left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top\right\rangle - y_i\right)^2\right]$$

- Is given as:

$$\mathbb{E}\left[\left(\left\langle A_i, \left((1-\eta^2)U + \sqrt{\eta'}U'\right)\left((1-\eta^2)V + \sqrt{\eta'}V'\right)^\top\right\rangle - y_i\right)^2\right]$$
$$= \left(\langle A_i, (1-\eta')UV^\top + \eta' U^*(V^*)^\top\rangle - y_i\right)^2 + \eta' O(\mathbb{E}[\|\xi\|_F^2])$$
$$\leq (1-\eta')\left(\langle A_i, UV^\top\rangle - y_i\right)^2 + \eta'\left(\langle A_i, U^*(V^*)^\top\rangle - y_i\right)^2 + \eta' O(\mathbb{E}[\|\xi\|_F^2])$$

- The last inequality is by convexity of the function x^2.

- Together with the change in the regularizer, we have: for $\eta' = 2\eta^2 - \eta^4$:

$$\mathbb{E}_{U' \sim P_U, V' \sim P_V} f\left((1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}U', (1-\eta^2)V + \sqrt{2\eta^2 - \eta^4}V'\right)$$

$$\leq (1-\eta')\frac{1}{N}\sum_{i \in [N]}\left(\langle A_i, UV^\top\rangle - y_i\right)^2 + \eta'\frac{1}{N}\sum_{i \in [N]}\left(\langle A_i, U^*(V^*)^\top\rangle - y_i\right)^2$$

$$+\eta' O(\mathbb{E}[\|\xi\|_F^2]) + (1-\eta')\lambda\|U\|_F^2 + \eta'\lambda\|U^*\|_F^2 + (1-\eta')\lambda\|V\|_F^2 + \eta'\lambda\|V^*\|_F^2$$

$$= (1-\eta')f(U,V) + \eta' f(U^*, V^*) + \eta' O(\mathbb{E}[\|\xi\|_F^2])$$

- Where $O(\mathbb{E}[\|\xi\|_F^2]) = O\left(\frac{1}{k}\right)$ for $R = kr$.

Over-parameterization in matrix sensing

- Together we know ($\eta' = 2\eta^2 - \eta^4$):

$$\mathbb{E}_{U' \sim P_U, V' \sim P_V} f\left((1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}U', (1-\eta^2)V + \sqrt{2\eta^2 - \eta^4}V'\right)$$

$$\leq (1-\eta')f(U,V) + \eta' f(U^*, V^*) + \eta' O(\mathbb{E}[\|\xi\|_F^2])$$

- Where $O(\mathbb{E}[\|\xi\|_F^2]) = O\left(\frac{1}{k}\right)$ for $R = kr$.
- Now by assumption, $f(U,V) \geq 1.1\epsilon$, $f(U^*, V^*) \leq \epsilon$ we show that as the over-parameterization $R \geq \text{poly}(1/\epsilon)r$ so $O(\mathbb{E}[\|\xi\|_F^2]) \leq 0.01\epsilon$, we have:

$$\mathbb{E}_{U' \sim P_U, V' \sim P_V} f\left((1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}U', (1-\eta^2)V + \sqrt{2\eta^2 - \eta^4}V'\right)$$

$$\leq f(U,V) - 0.05\eta^2\epsilon$$