Note I: Efficient Convex Optimization Requires Superlinear Memory

[MSSV22]

Question: Is there a convex optimization task for which limited memory causes a need for more queries?

The answer is YES!

To show: an information-theoretic source of hardness ① and reduction from any first-order optimization problem in a given class ③ & Exhibiting a concrete function in that class.

The key takeaway: learning a global feature of some structure in high-dimension is difficult with little memory

For us, we focus on the null-space of a random matrix to start with ①, following "Orthogonal Vector Game"

Game 1 (OVG) Input: $d, k, B_d \subset \mathbb{R}^d, M$

Phase 1: Store Message

Oracle: For $n \leftarrow d/2$, sample $A \sim \mathrm{Unif}(B_d^n)$

Oracle: sample random bit string of length $3 \cdot 2^d$ & present to Player

Player: observing A, R, store Message $\in \{0,1\}^M$

Phase 2: Adaptive Queries

for $i \in [m]$:

Player: based on prev. queries & responses, submits $x_i \in S_d$

Oracle: responds with predetermined $g_A(x_i) =: g_i$

Phase 3: Output Player produces $(y_1, \dots, y_k)$ sequence that is approx. orthogonal to A & robustly linearly independent.

<u>Question</u>: What makes the game hard?

Roughly speaking, we attempt to "find" k-different directions in $\text{nul}(A)$: this should be incompressible in the following sense:

<u>Definition</u> (Memory-sensitive base) A sequence of sets $\{B_d\}_{d>0}$, $B_d \subset \mathbb{R}^d$ is (k, c_B)-memory-sensitive if, $\forall Z = [z_1, \dots z_k]$ projection matrices

$$\Pr\left[\|Z^T h\|_\infty \leq \tfrac{1}{2}\right] = \Pr\left[\max_{i=1}^{k} |\langle z_i, h\rangle| \leq \tfrac{1}{2}\right] \leq 2^{-c_B k}$$

additionally, $\Pr[\|h\|_2 > d] \leq 2^{-d}$.

So, locating k-vectors nearly orthonormal to A <u>substantially</u> narrows the possibilities for rows $a_1^T, \dots a_n^T$ to at most $2^{-c_B k}|B_d|$ options: encodable in $\log|B_d| - c_B k$ bits

Indeed, memory-sensitive bases exist, such as $\mathcal{H}_d = \{\pm 1\}^d$.

Now, we begin efforts to showing a lower bound to OVG.

... but first, what are the trivial strategies?

① No Queries: we must solve the problem in Message immediately

we can store the subspace of $\text{nul}(A)$ in $\tilde{O}(dk)$ bits

Further, Shannon source coding tells us we can't hope for much better

② No Memory: Simply query $m = n$ rows of A and solve directly.

We refine our lower bound to ask the following:

Is it possible to use even slightly less than dk bits of memory & make a few adaptive queries to solve OVG?

Alas, the answer is NO!

Theorem $\exists c > 0$ s.t. $\forall k \geq \Omega(\log d)$, if $M \leq cdk$, if the player wins the OVG with probability at least $1/3$, then $m \geq d/6$

proof The fact that Y is successful insinuates that it greatly reduces the entropy of A at the end of Phase 3.
[We will assume that Y is orthonormal, but show it is 'WLOG' later]
We consider "reconstructing" A using Y & oracle responses:

$$I(A; Y | G, R) = I(A; f(\text{Message}, G, R) | G, R)$$
$$\leq I(A; \text{Message}, G, R | G, R) \leq H(\text{Message}) = M.$$

by the bottleneck caused by passing Message after Phase 1.

However,
$$I(A; Y | G, R) = H(A | G, R) - H(A | G, R, Y)$$
$$\geq (n-m) \log |B_d| - (n-m)(\log |B_d| - c_B k)$$
$$= c_B (n-m) k$$
using the compression & assuming that all m oracles are distinct.

Thus, $M \geq c_B (n-m) k \Rightarrow m \geq n - \frac{M}{c_B k}$.

So, if $n = d/2$, $M = \frac{c_B}{3} \cdot dk \Rightarrow m \geq d/6$, as desired. $\square$

Note that we have oversimplified a lot of the reconstruction, but the proof is "essentially" correct.

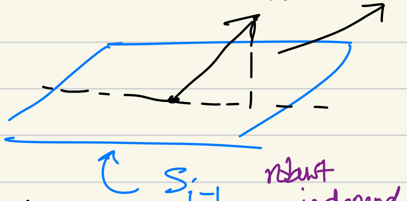
To tie some loose ends, we formalize approximate orthogonality (a.o.) & robust linear independence (r.l.i) and show that they do indeed imply an orthonormal basis of O(k).

<u>Definition</u> With respect to $A \in \mathbb{R}^{d \times n}$, a sequence $(y_1, \dots y_k)$ is

(a.o.) $\frac{\|Ay_i\|_\infty}{\|y_i\|_2} \le d^{-4} \; \forall i,$

(r.l.i) Let $S_i = \text{span}(y_1, y_2, \dots y_i)$, $S_0 = \phi$,

$\|\text{Proj}_{S_{i-1}}(y_i)\| / \|y_i\|_2 \le 1 - \frac{1}{d^2}$



<u>Lemma</u>: fix $\lambda \in (0,1]$, $n_0 \le d$. Let a sequence $(y_1, y_2, \dots y_{n_0})$ of unit vectors satisfy λ-RLI, so

$\|\text{Proj}_{S_{i-1}}(y_i)\| \le 1 - \lambda.$

Then, $\exists$ $(m_1, m_2, \dots m_{n_1})$ orthonormal s.t. $\forall a \in \mathbb{R}^d$, $\|M^T a\|_\infty \le \frac{d}{\lambda} \|Y^T a\|_\infty$

Where $M = [m_1 \; m_2 \dots m_{n_1}]$, $Y = [y_1, y_2, \dots y_{n_0}]$ and $n_1 = \lfloor n_0/2 \rfloor$

In particular, if the sequence is additionally d^{-4}-a.o. as well as d^{-2}-RLI, we have that $\|M^T a\| \le 1/d \le 1/2$, $\forall a \in \text{row}(A)$.

This completes our sketch for OVG hardness.

Next, we design an optimization task which inherently involves finding directions in the incompressible null-space of a random matrix.

We propose the following, two-part construction:

$$F(x) = \frac{1}{d^6} \max\Big\{ \underbrace{d^5 \|Ax\|_\infty - 1}_{f_A(x)},\ \underbrace{\max_{i=1}^{N} (v_i^T x - i\gamma)}_{f_V(x)} \Big\}$$

v_i — Nemirovski vector.

for $A \in \{\pm 1\}^{d/2 \times d}$, $v_i \sim \frac{1}{\sqrt{d}} \{\pm 1\}^d$ drawn uniformly at random.

We define the subgradient function as

$$g_F(x) = \begin{cases} \min\{ i : \frac{1}{d} |a_i^T x| - \frac{1}{d^6} = F(x) \} & \text{if } F(x) = d^{-6} f_A(x) \\ \min\{ i : v_i^T x - i\gamma = d^6 F(x) \} & \text{else} \end{cases}$$

We also define an informative subgradient as a query which returns a previously unseen Nemirovski vector v_i.

Later, we will show $\mathcal{F}$ (the distribution over (F, g_F)) has the structure of the following class.

<u>Definition</u> An (L, N, k, ϵ^*)-memory sensitive function class $\mathcal{F}$ can be sampled with at most 2^d bits, and, on any potentially randomized algorithm A_{rand} of Nd, the following holds w.p $\geq 2/3$.

1) The sampled function is L-Lipschitz
2) The informative queries are d^{-2}-r.l.i up to history of length k

2) (Cont) more formally, if $\{x_{t_j}\}$ are the set of informative queries with $\{t_j\} \subset [Nd]$ and $S_j = \mathrm{span}(\{x_{t_i} : \max(1, j-k) \le i \le j\})$

$$\Rightarrow \forall j \in [N], \quad \|\mathrm{Proj}_{S_{j-1}}(x_{t_j})\| / \|x_{t_j}\| \le 1 - 1/d^{2}$$

3) Receiving Nemirovski vectors implies approx. orth. to A:
Any query x with $F(x) \ne \eta \|Ax\|_\infty - \rho$ sat. either

$$g_F(x) = v_1 \quad \text{or} \quad \|Ax\|_\infty / \|x\|_2 \le d^{-4}.$$

4) If $r < N \Rightarrow \forall i \le t_r, \ F(x_i) - F(x^*) \ge \epsilon^*$,
informative queries are necessary for minimization!

Immediately, let's show why this definition is useful by the reduction

Lemma Let $A \sim \mathrm{Unif}(B_d^n)$, $(f, g_f) \sim \mathcal{F}$ for (L, N, k, ϵ^*)-sensitive class,
If $\exists$ an M-bit algorithm optimizing $F_{A,f}$ with at most $M \lfloor N/(k+1) \rfloor$ queries wp $\ge 2/3$ $\Rightarrow$ winning strategy for OVG wp $\ge 1/3$.

proof By Pigeonhole, there exists a period of M queries in which $k+1$ informative queries are observed.

We run the algorithm in Phase 1 by drawing $(f, g_f) \sim \mathcal{F}$, Constructing $F_{f,A}$, and saving in Message the memory state at the start of this period

In phase 2, we continue by resampling (f, g_f) using the same randomness, and query to the oracle. We can simulate our first-order oracle again even without access to A (how?).

Finally, we search over all k-size subsets of our queries and output a robustly independent sequence.

Correctness? We can condition on properties in the memory-sensitive class with extra failure probability of $1/3$.

The existence of a solution, the number of informative queries is direct and completes the proof ∎

Let's finish by sketching the argument for showing F is memory sensitive

The key is to show that a "resisting oracle" which draws v_j only when needed, is consistent with itself.

$\Rightarrow$ we can show that, prior to v_j being drawn, all prior queries are independent: thus the following event

$$E = \begin{cases} \forall x \text{ submitted up to phase } j, \ |\langle x, v_j \rangle| \le \sqrt{\frac{10 \log d}{d}} \\ \| \mathrm{Proj}_{S_j} v_j \| \le \sqrt{\frac{30 k \log d}{d}} \end{cases}$$

has probability $\Pr[E] \ge 1 - 1/d$, by concentration

Using this, we show that the $-i\gamma$ drop-off between terms is sufficient to discovering v_j's in order, as long as

$$\gamma \ge \sqrt{\frac{ck \log d}{d}}.$$

Other properties readily follow from algebra and conditioning on E.