Streaming PCA Summer Findings

William Guo

July 22, 2024

Contents

1 Introduction **2**
- 1.1 Paper summary . . . 2
- 1.2 Overview of new results . . . 6

2 Tightening Bounds **7**
- 2.1 Vector growth is tight . . . 7
- 2.2 Lower bound on inductive lemma . . . 8
- 2.3 Lower bounding projection error . . . 9
- 2.4 Vector growth is tight (part 2) . . . 10

3 Improving Growth Bounds **11**
- 3.1 Synthetic approaches . . . 11
- 3.2 Bus lines puzzle . . . 12
  - 3.2.1 Puzzle statement . . . 12
  - 3.2.2 Explanation . . . 12
  - 3.2.3 Approaches to the puzzle . . . 16

4 Generalizing Oja's algorithm to k eigenvectors **21**
- 4.1 2-Oja's . . . 21

5 Gaussian skip **23**
- 5.1 Bound on the change in the orthogonal direction . . . 23
- 5.2 Growth bound . . . 26
- 5.3 Bounds on C . . . 26

Chapter 1

Introduction

1.1 Paper summary

Oja's algorithm is used to approximate the top k eigenvectors of $X^\top X$ for some data matrix $X \in \mathbb{R}^{n \times d}$ inputted as a stream of d-dimensional vectors. In the $k = 1$ instance, Oja's algorithm begins with a random vector v_0, iteratively updating with the rule $v_i = v_{i-1} + \eta x_i x_i^\top v_{i-1}$.

The key algorithm used in the paper (with minor modifications for theoretical purposes) is as follows:

Oja's Algorithm

Initialize $v_0 \in S^{d-1}$
for i from 1 to n **do**
    $v_i \leftarrow v_{i-1} + \eta x_i x_i^\top v_{i-1}$
    $\hat{v}_i \leftarrow \frac{v_i}{||v_i||}$
end for
if $||v_n|| \leq 10 \log d$ **then**
    return null
else
    return $\hat{v}_n$
end if

Oja's algorithm is typically analyzed under a stochastic environment, in which each x_i comes from a nice distribution, allowing one to predict the movement of v_i at each iteration. However, in the adversarial setting where the data points are arbitrary, an iterative approach is much less feasible.

The key results from this paper show that even in the adversarial setting, one can show that $\hat{v}_n$ does indeed approach our top eigenvector, v^*. Denoting σ_1, σ_2 as the top 2 eigenvectors of matrix $\eta X^\top X$, so long as $\sigma_1 > C \log d$ and $\sigma_2 < \frac{1}{C \log n}$ for some $C > 0$, we have that

$$||P\hat{v}_n|| \leq \sqrt{\sigma_2} + \frac{||Pv_0||}{||v_n||}$$

where $P = I - v_* v_*^\top$ is the projection matrix onto the subspace orthogonal to v_*.

The roadmap of the paper can be summarized as follows:

1. **Growth & warm start**. By expanding the iterative definition of $||v_n||$, one can show that

   $$||v_n||^2 \geq e^{\sum_{i=1}^n \langle x_i, \hat{v}_{i-1} \rangle^2}$$

   The paper then shows that growth of $||v_n||$ implies that the output $\hat{v}_n$ has low error. The underlying motivation for this claim is that if v_n has any resemblance to v_*, over the course of the algorithm, it should have grown rapidly as the $\langle x_i, \hat{v}_{i-1} \rangle$ terms should be on the order of σ_1. In particular, we seek to show

   $$||v_n||^2 \geq e^{O(\sigma_1)}$$

   As it turns out, we can assume $v_0 = v_*$ by the Gaussian lemma (lemma 3.7):

   Lemma 1 (Gaussian lemma)**.** *For Gaussian $a \sim N(0, 1)$, with probability $\geq 1 - \delta$,*

   $$||av + u|| \geq \delta \sqrt{\pi/2} ||v||$$

   which can be proven by noting that in the worst case, u is $\vec{0}$, so we want to bound $|a|||v||$. Since the absolute normal distribution has density at most $\sqrt{2/\pi}$, $|a| < k$ with probability at most $k\sqrt{2\pi}$ for all $k > 0$. Setting $k = \delta\sqrt{\pi/2}$, we get $|a| \geq \delta\sqrt{\pi/2}$ with probability $\geq 1 - \delta$, as desired.

   Now, denote $B = \prod_{i=1}^n I + \eta x_i x_i^\top$, and rewrite $v_0 = \frac{v_0'}{||v_0'||}$ for Gaussian vector $v_0' \sim N(0, I)$. We then rewrite $v_0' = av_* + u$, where $u \perp v_*$. Then, applying the Gaussian lemma to v_n, we get

   $$||v_n|| = ||Bv_0|| = \frac{||B(av_* + u)||}{||v_0'||}$$

   $$\geq \frac{1}{||v_0'||} \cdot \delta\sqrt{\pi/2} ||Bv_*||$$

with probability $1 - \delta$. As such, $||v_n||$ is lower bounded by $O(||Bv_*||)$, where Bv_* is the output of Oja's algorithm when run with $v_0 = v_*$, so it suffices to show $||v_n||$ is large with this initialization.

2. **Error bound.** The error of $\hat{v}_n$ can be expressed as $||P\hat{v}_n||$. We can expand the unnormalized error as

$$||Pv_n|| = ||Pv_n - Pv_0 + Pv_0|| \leq ||Pv_n - Pv_0|| + ||Pv_0||$$

by the triangle inequality. Then, denoting $w = \frac{Pv_n - Pv_0}{||Pv_n - Pv_0||}$,

$$||Pv_n - Pv_0||^2 \leq \langle v_n - v_0, Pv_n - Pv_0 \rangle = \langle v_n - v_0, \frac{Pv_n - Pv_0}{||Pv_n - Pv_0||} \rangle^2$$

$$= \langle v_n - v_0, w \rangle^2$$

By expansion of v_n, one finds

$$\langle v_n - v_0, w \rangle^2 = \langle v_{n-1} + \eta x_n x_n^\top v_{n-1} - v_0, w \rangle^2$$

$$= \left(\langle v_{n-1} - v_0, w \rangle + \eta \langle x_n, w \rangle \langle x_n, v_{n-1} \rangle \right)^2$$

$$= \ldots = \left(\eta \sum_{i=1}^{n} \langle x_i, w \rangle \langle x_i, v_{i-1} \rangle \right)^2$$

$$\leq \left(\eta \sum_{i=1}^{n} \langle x_i, w \rangle^2 \right) \left(\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2 \right)$$

where the final inequality arises from Cauchy-Schwarz[1]. By inductive lemma (claim 3.1), we know that

$$\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2 = \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 ||v_{i-1}||^2 \leq \eta \sum_{i=1}^{n} \log \frac{||v_i||^2}{||v_{i-1}||^2} \left(e^{\sum_{j=1}^{i} \log \frac{||v_j||^2}{||v_{j-1}||^2}} ||v_0||^2 \right)$$

$$\leq ||v_0||^2 \left(e^{\sum_{i=1}^{n} \log \frac{||v_i||^2}{||v_{i-1}||^2}} - 1 \right) = ||v_n||^2 - ||v_0||^2$$

[1] I have yet to understand what makes this inequality strong. There are many instances where I've tried to use Cauchy-Schwarz only to get a fairly loose bound; in this particular case, it ends up working quite nicely.

and since $w \perp v_*$, $\eta \sum_{i=1}^{n} \langle x_i, w \rangle^2 \leq \sigma_2$, which gives us

$$\langle v_n - v_0, w \rangle^2 \leq \sigma_2(||v_n||^2 - ||v_0||^2) \leq \sigma_2 ||v_n||^2$$
$$\therefore ||Pv_n - Pv_0|| \leq ||v_n|| \sqrt{\sigma_2}$$
$$\therefore ||P\hat{v}_n|| \leq \sqrt{\sigma_2} + \frac{||Pv_0||}{||v_n||}.$$

This confirms that growth of $||v_n||$ decreases our error. If we assume $v_0 = v_*$, this becomes

$$||P\hat{v}_n|| \leq \sqrt{\sigma_2}$$

3. **Bounding growth with matrix lemma.** Now, to show $||v_n||^2$ is large, we should lower bound $\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2$ in hopes of showing it is $\Omega(\sigma_1)$. With the idea that $\langle x_i, \hat{v}_{i-1} \rangle = \langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle + \langle x_i, P\hat{v}_{i-1} \rangle$ and repelled by the idea of having to work with $\langle x_i, v_* \rangle \langle x_i, P\hat{v}_i \rangle$, we use the inequality $(x+y)^2 \geq \frac{1}{2}x^2 - y^2$ to get

$$\langle x_i, \hat{v}_{i-1} \rangle^2 \geq \frac{1 - ||P\hat{v}_{i-1}||^2}{2} \langle x_i, v_* \rangle^2 - \langle x_i, P\hat{v}_{i-1} \rangle^2$$

Since $||P\hat{v}_{i-1}||^2 \leq \sigma_2$ ($\leq 1/2$, reasonably speaking), and $\eta \sum_{i=1}^{n} \langle x_i, v_* \rangle^2 = \sigma_1$, the first term is easy to manage. We now seek to bound the second term, which we do via the matrix lemma (lemmas 3.4/3.5):

Lemma 2 (Matrix lemma)**.** *For $v_0 = v_*$,*

$$\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \leq 4\sigma_2^2 (1 + \log_2 n)^2 \log ||v_n||$$

The technical details of this lemma are left to the bus lines puzzle section, where this problem is abstracted to a standard textbook discrete math/algorithms problem.

For $\sigma_2 \leq \frac{1}{C \log n}$, this gives us growth bound

$$\log ||v_n|| \geq \frac{1}{2} \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 \geq \frac{1}{8} \sigma_1 - 2C^2 \log ||v_n||$$
$$\therefore \log ||v_n|| \geq \frac{\sigma_1}{8(1 + 2C^2)}$$

which is $\Theta(\sigma_1)$ as desired.

4. **Bringing it all together** Theorem 1.1 of the paper states that with probability $1 - d^{-\Omega(C)}$, Oja's algorithm outputs vector v_n fulfilling

$$||Pv_n|| \leq \sqrt{\sigma_2} + d^{-10}$$

which directly follows from the fact that $||v_n|| \geq e^{O(\sigma_1)} \geq e^{10 \log d}$, which can be substituted into the error bound on $||Pv_n||$.

1.2 Overview of new results

Much of my work has tightened the existing lemmas and theorems. I have shown that the exponential growth of $\log ||v_n||$ is tight to $O(\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2)$, and have provided a new lower bound on this summation along with a lower bound on $\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2$.

I spent a copious amount of time trying to improve the matrix lemma by abstracting the lemma into a puzzle that could be shared with my friends & fellow Epic interns to take a stab at. Unfortunately, none of us have been able to improve the paper's $(\log_2(n))^2$ coefficient. Doing so would directly relieve the constraint on σ_2, but as is discussed in the bus lines puzzle section, it is probable that a different approach is needed to improve this coefficient. Nonetheless, significant effort was put into understanding this particular lemma and its applications.

Some other areas I've explored include the Gaussian skip, in which the assumption that $v_0 = v_*$ is removed to increase the lower bound on the probability of success, and k-Oja's which is a natural extension of Oja's algorithm to finding the top k eigenvectors of the covariance matrix, not just the top 1. As of writing this, I'm very optimistic about the Gaussian skip, which is complete minus a proof on $||Pv_0|| \leq (\sqrt{2} - 1)\sqrt{\sigma_2}||v_n||$. Completing this proof would essentially prove that Oja's algorithm is deterministic (works with 100% certainty) as opposed to the high probability of $1 - d^{-\Omega(C)}$ provided by the paper, which is a novel result. k-Oja's has proven to be a difficult field as existing literature on k-Oja's is exclusively in the stochastic setting, where each iteration can be analyzed, a technique not feasible in the adversarial setting.

Chapter 2

Tightening Bounds

2.1 Vector growth is tight

We can work in a very similar fashion to the existing paper to get

$$\begin{aligned}
||v_n||^2 = \langle v_n, v_n\rangle &= \langle v_{n-1} + \eta x_n x_n^\top v_{n-1}, v_{n-1} + \eta x_n x_n^\top v_{n-1}\rangle \\
&= ||v_{n-1}||^2 + 2\eta\langle x_n, v_{n-1}\rangle^2 + \eta^2||x_n||^2\langle x_n, v_{n-1}\rangle^2 \\
&= ||v_{n-1}||^2(1 + 2\eta\langle x_n, \hat{v}_{n-1}\rangle^2 + \eta^2||x_n||^2\langle x_n, \hat{v}_{n-1}\rangle^2) \\
&\leq ||v_{n-1}||^2(1 + 3\eta\langle x_n, \hat{v}_{n-1}\rangle^2) \\
&\leq ||v_{n-1}||^2 e^{3\eta\langle x_n, \hat{v}_{n-1}\rangle^2}
\end{aligned}$$

where the first inequality arises from $\eta||x_n||^2 \leq 1$. Hence, we get

$$\log \frac{||v_n||^2}{||v_{n-1}||^2} \leq 3\eta\langle x_n \hat{v}_{n-1}\rangle^2$$

By a telescoping product of $\log \frac{||v_i||^2}{||v_{i-1}||^2}$ for $1 \leq i \leq n$, we get

$$\log \frac{||v_n||^2}{||v_0||^2} = \log ||v_n||^2 \leq 3\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1}\rangle^2$$

By the paper, we know $\log ||v_n||^2 \geq \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1}\rangle^2$, implying

$$\log ||v_n||^2 = \Theta\left(\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1}\rangle^2\right)$$

2.2 Lower bound on inductive lemma

The inductive lemma is used to bound $\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2 = \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 e^{\eta \sum_{j<i} \langle x_j, \hat{v}_{j-1} \rangle^2}$. This expression is seen frequently in expansions involving v_n, and is used directly to bound the error term $||Pv_n||$.

As of now, I have only discovered a weak lower bound for the inductive lemma as follows:

Claim 1. *For* $0 \leq a_1, a_2, \cdots, a_n \leq 1$,

$$\sum_{i=1}^{n} a_i e^{\sum_{j<i} a_j} \geq e^{\frac{1}{2} \sum_{i=1}^{n} a_i} - 1$$

Proof. We begin with induction. For $n = 0$, $0 \geq 1 - 1$. Then, observe that

$$\sum_{i=1}^{n} a_i e^{\sum_{j<i} a_j} = a_n e^{\sum_{i<n} a_i} + \sum_{i=1}^{n-1} a_i e^{\sum_{j<i} a_j}$$

$$\geq a_n e^{\sum_{i<n} a_i} + e^{\frac{1}{2} \sum_{i=1}^{n-1} a_i} - 1$$

$$\geq a_n e^{\frac{1}{2} \sum_{i<n} a_i} + e^{\frac{1}{2} \sum_{i=1}^{n-1} a_i} - 1$$

$$\geq e^{\frac{1}{2} \sum_{i=1}^{n-1} a_i} (1 + a_n) - 1$$

$$\geq e^{\frac{1}{2} \sum_{i=1}^{n} a_i} - 1$$

where the last equality uses $1 + 2x \geq e^x$ for $0 \leq x \leq 1$. □

This inequality is fairly loose, as applying it to $\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2$ we get

$$\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2 = \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 ||v_{i-1}||^2$$

$$\geq \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 e^{\eta \sum_{j<i} \langle x_j, \hat{v}_{j-1} \rangle^2} \geq e^{\frac{1}{2} \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2} - 1$$

$$\geq ||v_n||^{\frac{1}{6}} - 1$$

which is off by a power of 12 compared to the current upper bound of $||v_n||^2 - ||v_0||^2$ on this same expression.

At a high level, if we find a stronger lower bound for $\eta \sum_{i=1}^{n} \langle x_i, v_{i-1} \rangle^2$, we can attempt inverting this lemma to bound $\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2$, which would directly help lower bound $||v_n||$.

2.3 Lower bounding projection error

A key quantity to bound in the paper is $\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2$, which shows up as an error term when lower bounding $\eta \sum_{i-1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2$.

We are motivated by

$$\langle x_i, \hat{v}_{i-1} \rangle = \langle x_i, P\hat{v}_{i-1} \rangle + \langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle$$

which, given $(a+b)^2 \leq 2a^2 + 2b^2$ by the AM-QM inequality, yields up upper bound

$$\begin{aligned} \langle x_i, \hat{v}_{i-1} \rangle^2 &\leq 2\langle x_i, P\hat{v}_{i-1} \rangle^2 + 2\langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle^2 \\ &\leq 2\langle x_i, P\hat{v}_{i-1} \rangle^2 + 2\langle x_i, v_* \rangle^2 \end{aligned}$$

Summing yields

$$\begin{aligned} \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 &\leq 2\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 + \langle x_i, v_* \rangle^2 \\ &= 2\sigma_1 + 2\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \end{aligned}$$

Using our upper bound on $\log ||v_n||$ from earlier and rearranging, we get

$$\begin{aligned} \frac{1}{3} \log ||v_n|| - 2\sigma_1 &\leq 2\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \\ \implies \frac{1}{6} \log ||v_n|| - \sigma_1 &\leq \eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \end{aligned}$$

This is still fairly loose, as for $\log ||v_n|| < 6\sigma_1$ the left side is negative. Nonetheless, we compare this bound with the (implied) lower bound in the paper obtained by rearranging

$$\begin{aligned} 2 \log ||v_n|| &\geq \eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 \geq \frac{1}{4}\sigma_1 - \eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \\ \implies \eta \sum_{i-1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 &\geq \frac{1}{4}\sigma_1 - 2 \log ||v_n|| \end{aligned}$$

to get that our bound is stronger when

$$\frac{1}{6}\log||v_n|| - \sigma_1 \geq \frac{1}{4}\sigma_1 - 2\log||v_n||$$

$$\implies \log||v_n|| \geq \frac{15}{26}\sigma_1$$

which hasn't been shown to be guaranteed, but is likely given $\log||v_n|| = \Omega(\sigma_1)$.

2.4 Vector growth is tight (part 2)

Conversely, we can rearrange the above inequality as

$$\log||v_n|| \leq 6\sigma_1 + 6\eta\sum_{i=1}^{n}\langle x_i, P\hat{v}_{i-1}\rangle^2$$

$$\leq 6\sigma_1 + 24\sigma_2^2(1+\log_2 n)^2\log||v_n||$$

$$\implies \log||v_n||\left(1 - 24\sigma_2^2(1+\log_2 n)^2\right) \leq 6\sigma_1$$

For $\sigma_2 \leq \frac{1}{C\log_2 n} \simeq \frac{1}{C(1+\log_2 n)}$, note that $1 - 24\sigma_2^2(1+\log_2 n)^2 \leq 1 - \frac{24}{C^2}$. For $C \leq \sqrt{24}$, this term is negative, so our inequality is useless. However, for $C > \sqrt{24}$, we have that the growth of $\log||v_n||$ is indeed tight to $\Theta(\sigma_1)$.

To given an example for how reasonable this bound is on C, consider when $n, d \approx 10^6 \to \log n, \log d > 20$; we'd need $\sigma_2 \leq \frac{1}{\sqrt{24}*20} \approx 0.01$ and $\sigma_1 \geq \sqrt{24} * 20 \approx 98$.

There is good reason to believe C is relatively large in the paper, which assumes in lemma 3.7 that

$$\frac{||v_n||}{||v_0||} \geq \frac{\delta^{3/2}e^{\frac{C\log d}{8(1+\frac{2}{C^2})}}}{\sqrt{d}} \geq d^{10}$$

w.p. 1 - 2δ. For $\delta = d^{-2/3}$,

$$d^{\frac{C}{8(1+\frac{2}{C^2})}} \geq d^{11.5}$$

$$\therefore \frac{C^3}{8(C^2+2)} \geq 11.5$$

$$\therefore C > 92$$

Chapter 3

Improving Growth Bounds

3.1 Synthetic approaches

In the paper, the known bounds on $\eta \sum_{i=1}^{n} \langle x_i, v_* \rangle^2$ and $\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2$ are used to bound $\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2$. We can attempt to bound the exact value of this summation as follows:

$$\eta \sum_{i=1}^{n} \langle x_i, \hat{v}_{i-1} \rangle^2 = \eta \sum_{i=1}^{n} \left(\langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle + \langle x_i, P\hat{v}_{i-1} \rangle \right)^2$$

$$= \eta \sum_{i=1}^{n} (1 - ||P\hat{v}_{i-1}||^2) \langle x_i, v_* \rangle^2 + \langle x_i, P\hat{v}_{i-1} \rangle^2 + 2 \langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle \langle x_i, P\hat{v}_{i-1} \rangle$$

Within this summation, first term is virtually identical to the term used in the paper, and the second term is nonnegative for all we are concerned. It remains to be shown that $2\eta \sum_{i=1}^{n} \langle x_i, \sqrt{1 - ||P\hat{v}_{i-1}||^2} v_* \rangle \langle x_i, P\hat{v}_{i-1} \rangle$ cannot be $-\Omega(\sigma_1)$.

A naive bound on this expression gives us

$$\sqrt{1 - ||P\hat{v}_{i-1}||^2} \langle x_i, v_* \rangle \langle x_i, P\hat{v}_{i-1} \rangle \geq \sqrt{1 - \sigma_2} (-||x_i||^2)$$

$$-2\eta \sqrt{1 - \sigma_2} \sum_{i=1}^{n} ||x_i||^2 = -2\sqrt{1 - \sigma_2} \sum_{i=1}^{d} \sigma_i$$

by definition of the Frobenius norm. Ultimately, this bound is $-\Omega(\sigma_1)$, which is not helpful.

Nonetheless, we are left to show either that each individual term $\langle x_i, v_* \rangle \langle x_i, P\hat{v}_{i-1} \rangle$ is sufficiently small ($\ll \frac{\sigma_1}{\eta n}$) or that in aggregate the summation is $o(\sigma_1)$. Unlike

previous summations, it is hard to lower bound this product (Cauchy-Schwarz is only helpful in upper bounding this quantity). At a high level, I believe it is unlikely for all $\langle x_i, Pv_{i-1} \rangle$ and $\langle x_i, v_* \rangle$ to be of different signs, but have yet to quantify this intuition in a succinct manner.

3.2 Bus lines puzzle

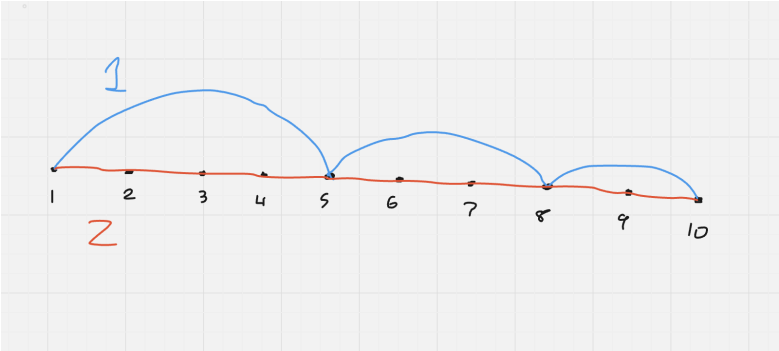
3.2.1 Puzzle statement

Consider the following puzzle:

You have n bus stops, labeled 1 through n, and k bus lines. Each bus line starts at n, ends at 1, and visits some subset of $[n]$ in decreasing order. Each stop is visited by at least 1 line.

Let the "distance" of each stop to n be defined as the minimum number of stops a traveler would have to traverse to get to 1 (including 1 but not the initial stop), potentially switching lines along the way at no cost.

What is the optimal configuration to minimize the max distance of any stop, & what is the distance in terms of n and k? Below is a diagram for $n = 10$, $k = 2$:



In the above configuration, the max distance is $d = 3$. For instance, a traveler at 7 could use line 2 to go $7 \to 6 \to 5$, then take line 1 to get to 1.

3.2.2 Explanation

I claim that this puzzle is an abstraction of the paper's approach to the matrix lemma.

Lemma 3 (Matrix lemma, simplified (3.4))**.** *Given a sequence of reals* $a_1, a_2, ..., a_n$ *with* $a_1 = 0$*, for any index* i*,*

$$a_i^2 \leq (1 + \log_2(n)) \sum_{k=0}^{\log_2 n} \sum_{j>0} (a_{1+2^k(i)} - a_{1+2^k(i-1)})^2$$

Proof. Denote $i^{(k)}$ as the index obtained when i is rounded down to the nearest integer that is $1 \bmod 2^k$; for example, $i^{(0)} = i$ and $i^{(\log_2 n)} = 1$. Then, since $a_{i^{(\log_2 n)}} = 0$,

$$a_i^2 = \left(\sum_{k=0}^{\log_2 n} a_{i^{(k)}} - a_{i^{(k+1)}} \right)^2$$

$$\leq (1 + \log_2 n) \sum_{k=0}^{\log_2 n} (a_{i^{(k)}} - a_{i^{(k+1)}})^2 \leq (1 + \log_2 n) \sum_{k=0}^{\log_2 n} \sum_{j>0} (a_{1+2^k j} - a_{1+2^k(j-1)})^2$$

where the first inequality is from Cauchy-Schwarz, and the second inequality sums over all possible differences of 2^k between values that are $1 \bmod 2^k$, making it a superset of the first inequality. □

This lemma is used to bound $\langle x_i, P\hat{v}_{i-1} \rangle^2$ as a telescoping sum beginning at $\langle x_i, P\hat{v}_n \rangle^2$ and ending at $\langle x_i, Pv_0 \rangle^2$. In particular, with the help of lemma 3.3 (see Gaussian skip for the proof) which states $||P\hat{v}_b - P\hat{v}_a||^2 \leq 4\sigma_2 \log \frac{||v_b||}{||v_a||}$, lemma 3.5 shows that

$$\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1} \rangle^2 \leq \eta(1+\log_2 n) \sum_{i=1}^{n} \sum_{k=0}^{\log_2 n} \sum_{j>0} \langle x_i, P\hat{v}_{(1+2^k j)-1} - P\hat{v}_{(1+2^k(j-1))-1} \rangle^2$$

$$= \eta(1 + \log_2 n) \sum_{k=0}^{\log_2 n} \sum_{j>0} \sum_{i=1}^{n} \langle x_i, P\hat{v}_{(1+2^k j)-1} - P\hat{v}_{(1+2^k(j-1))-1} \rangle^2$$

By flipping the order of our summations, we can now sum the expression $\langle x_i, w \rangle^2$ across i, where $w = \frac{P\hat{v}_{(1+2^k j)-1} - P\hat{v}_{(1+2^k(j-1))-1}}{||P\hat{v}_{(1+2^k j)-1} - P\hat{v}_{(1+2^k(j-1))-1}||}$, $w \perp v_*$, which gives us

$$\leq (1 + \log_2 n) \sum_{k=0}^{\log_2 n} \sum_{j>0} \sigma_2 ||P\hat{v}_{(1+2^k j)-1} - P\hat{v}_{(1+2^k(j-1))-1}||^2$$

$$\leq (1 + \log_2 n) \sum_{k=0}^{\log_2 n} \sum_{j>0} 4\sigma_2^2 \log \frac{||v_{(1+2^k j)-1}||}{||v_{(1+2^k(j-1))-1}||}$$

Assuming $\log_2 n \in \mathbb{Z}^+$, the inner summation telescopes to $\log ||v_n||$, so this becomes

$$= (1 + \log_2 n) \sum_{k=0}^{\log_2 n} 4\sigma_2^2 \log ||v_n||$$

$$= 4\sigma_2^2 (1 + \log_2 n)^2 \log ||v_n||.$$

The matrix lemmas (3.4/3.5) heavily rely on telescoping sums; I pose a couple of key insights into the length and quantity of such sums used.

Observation 1. The use of Cauchy-Schwarz in the matrix lemma adds the extra factor of $1 + \log_2 n$, which is the max possible length of a telescoping sum needed to express i in terms of $i^{(k)}$s. In general, for any telescoping sum of an element a_i using $i^{x_1}, i^{x_2}, ..., i^{x_c}$ for some indices $x_1, ..., x_c$, we add a factor of c to our inequality. If we are to improve this lemma, we might want to consider reducing the length of the max telescoping sum.

Observation 2. There are $1 + \log_2 n$ distinct possible differences between indices. In other words, $a_{i^{(k)}}$ and $a_{i^{(k+1)}}$ can differ by 1, 2, 2^2, ..., $2^{\log_2 n}$; this is represented by the outer summation $\sum_{k=0}^{\log_2 n}[\cdot]$. Within this summation, each telescoping sum has a unique difference, so we can interpret this information as "there are $1 + \log_2 n$ distinct telescoping sums".

Observation 3. The $(1+\log_2 n)^2$ term in the result of lemma 3.5 is the product of these two observations: that is, the coefficient of $\sigma_2^2 \log ||v_n||$ is the **number** of telescoping sums used times the **max length** of any telescoping sum.

This is critical: the coefficient of $\sigma_2^2 \log ||v_n||$ directly corresponds to the necessary bound on σ_2. For example, if we find a configuration with $O(1)$ distinct sums and $O(1)$ max length of the telescoping sum for any element, then $\sigma_2 \leq O(1)$ is sufficient. As such, we are heavily incentivized to improve either the number of telescoping sums used, or the max length of the sum used to express any element, as doing so directly improves our bound on σ_2.

Observation 4. Every $\langle x_i, P\hat{v}_{i-1} \rangle$ can be expressed as a telescoping sum. Furthermore, every index can be expressed as a telescoping sum that is a *subset of a telescoping sum from* $\langle x_i, P\hat{v}_n \rangle$ *to* $\langle x_i, P\hat{v}_0 \rangle$. In abstract terms, a telescoping sum

used to express a_i is included (i.e. subset-ed) in at least 1 telescoping sequence with first term a_n and last term a_0.

The need for our telescoping sums to begin at a_n and end at a_0 is by nature of lemma 3.3: if we can telescope $||P\hat{v}_b - P\hat{v}_a||^2$ from n to 0, it becomes $\log ||v_n||$. It is possible to avoid this constraint if a new lemma is uncovered that directly gives us the $\log ||v_n||$ term.

To give an example of how these observations come into play, consider the matrix lemma setup where $a_1 = 0$. It is evident that

$$a_i^2 = (a_i - a_1)^2 \leq (a_n - a_i)^2 + (a_i - a_1)^2$$

$$\leq \sum_{k=1}^{n} \left((a_n - a_i)^2 + (a_i - a_1)^2\right)$$

In other words, we can express every a_i as a "sum" of 1 term, $a_i - a_1$. And, $(a_i - a_1)^2$ is a subset of the sum of length 2 $(a_n - a_i)^2 + (a_i - a_1)^2$.

How would it look if we used this lemma in place of the matrix lemma? By the above observations, we would expect to get an upper bound of $O(\sigma_2^2 n \log ||v_n||)$ since each sum is of length 1, but we needed n of them to encompass all possible elements. Indeed,

$$\eta \sum_{i=1}^{n} \langle x_i, P\hat{v}_{i-1}\rangle^2 \leq \eta \sum_{i=1}^{n} \sum_{k=1}^{n} \langle x_i, P\hat{v}_n - P\hat{v}_{k-1}\rangle^2 + \langle x_i, P\hat{v}_{k-1} - P\hat{v}_0\rangle^2$$

$$\leq \sum_{k=1}^{n} ||P\hat{v}_n - P\hat{v}_{k-1}||^2\sigma_2 + ||P\hat{v}_{k-1} - P\hat{v}_0||^2\sigma_2 \leq \sum_{k=1}^{n} 4\sigma_2^2 \log ||v_n||$$

$$= 4\sigma_2^2 n \log ||v_n||$$

as expected.

Hence, the puzzle emerges as follows:

- The **elements** in our sequence can be represented as **bus stops**. In particular, the ith bus stop represents the term $\langle x_i, P\hat{v}_{i-1}\rangle$.
- Each **bus line** represents the elements included in a **telescoping sum** from $\langle x_i, P\hat{v}_n\rangle$ to $\langle x_i, P\hat{v}_0\rangle$. This is such that when lemma 3.3 is applied, each telescoping sum ends up with a $\log ||v_n||$ term.

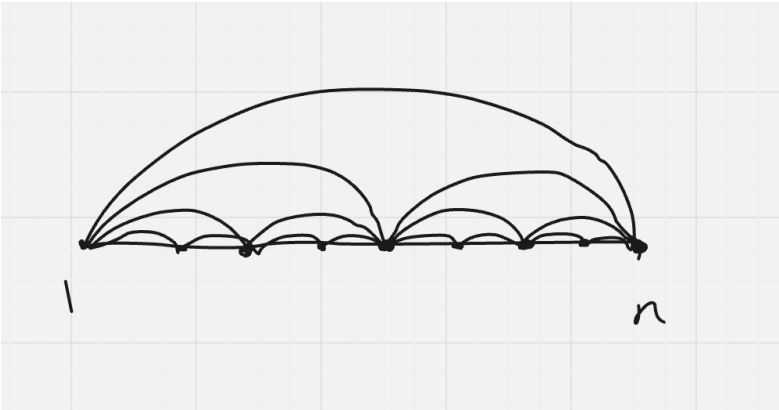
- If we consider our bus lines/elements as edges/vertices, each element must have a path to bus stop 1, aka v_0, and this path must be a subset of some path from v_n to v_0 (note this path doesn't necessarily have to be one telescoping sum - it can use a mix of edges from different bus lines).
- We seek to **minimize # of bus lines $\times$ max distance of any bus stop to** 1.

3.2.3 Approaches to the puzzle

Denote the stops a_1 through a_n. Let k be the number of bus lines and d be the max distance of any stop; we seek to minimize kd, or alternatively fix n and k and minimize d. The terms "gap", "jump", and "step size" all refer to the number of stops skipped between bus stops by a bus line (i.e. a bus line that only visits $a_{n/2}$ and a_1 leaves gaps of size $\sim n/2$ between a_1 and $a_{n/2}$, as well as between $a_{n/2}$ and a_n).

The paper's current approach to this puzzle is depicted as follows:

1. Bus line 1 traverses all n bus stops.
2. Bus line 2 goes from a_n straight to a_1.
3. Bus line 3 visits 2 stops: $a_{n/2}$ and a_1.
4. Bus line 4 visits 2^2 stops: $a_{3n/4}, a_{n/2}, a_{n/4}, a_1$
5. ...
6. Bus line k visits 2^{k-2} stops, from $a_{(2^{k-2}-1)n/2^{k-2}}$ to a_1 with step size $n/2^{k-2}$.

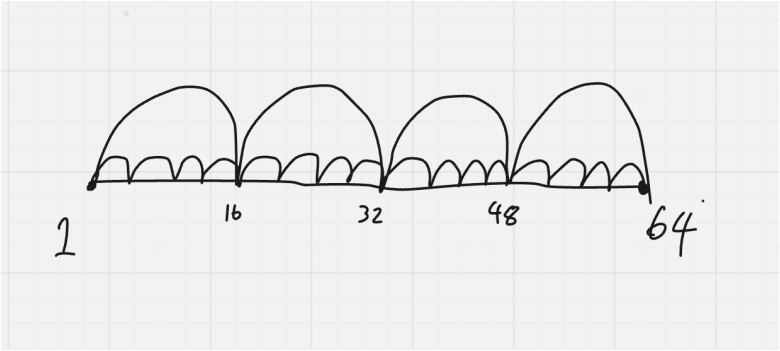


In general, the stops in the interval $[(2^{k-2}-1)n/2^{k-2}-1, n-1]$ have to make $\leq n/2^{k-2}$ jumps to get to stop $(2^{k-2}-1)n/2^{k-2}$, which then requires $k-1$ jumps to get to 1. This gives us a max distance of $n/2^{k-2}+k-1$, which is optimized when $k = O(\log n)$. Indeed, the paper essentially uses $k = \log_2 n$ to get a max distance of $n/(n/4) + \log_2 n - 1 = O(\log_2 n))$ for a product of $(\log_2 n)^2$.

This begs the question: is $kd = (\log_2 n)^2$ optimal? We consider a few configurations that seem to suggest yes.

Polynomial approach: The following approach gets $d = k\sqrt[k]{n}$.

- Bus 1 stops at $\sqrt[k]{n}$ evenly-spaced stops i.e. starting at n, it visits every $\sqrt[k]{n^{k-1}}$th stop.
- Bus 2 stops at $\sqrt[k]{n^2}$ evenly-spaced stops i.e. starting at n, it visits every $\sqrt[k]{n^{k-2}}$th stop. In other words, it splits each interval of $\sqrt[k]{n^{k-1}}$ stops skipped by bus 1 into $\sqrt[k]{n}$ more stops.
- ...
- Bus k stops at every stop.



Example for $n = 64$, $k = 3$. Note that each additional bus line splits the gaps created by previous lines into $\sqrt[k]{n} = 4$ smaller gaps.

Bus 1 only takes $n^{1/k}$ steps to get from n to 1, and in general any traveler riding bus line i will take $< n^{1/k}$ steps to get to a stop where they can transfer to bus line $i-1$, which has a much larger step size and can get to 1 faster. Hence, the max distance with this approach $O(kn^{1/k})$.

Taking the first derivative of kd with respect to k, we find that $k^2 n^{1/k}$ is optimized at $k = O(\log_2 n)$, which gives us $kd = (\log_2 n)^2 2^{(1/\log_2 n)} = (\log_2 n)^2$, just

like before.

Generalized splitting approach: we don't necessarily need each bus line split each interval into $n^{1/k}$ smaller intervals: we can instead make each bus split the previous bus's gaps into x parts until the gap size is 1.

Now, we seek to express k and d in terms of x. Since we are splitting n into x parts repeatedly, we have that $k = \log_x n$. And, a traveler on bus line i must traverse at most x stops to get to a stop where they can transfer to bus line $i-1$, for a max distance of $d = x \log_x n$.This gives us $kd = x^2 \log_x n$.

Taking the derivative of kd with respect to x gives a minimum at $x = e^2$, which would give us $kd = \frac{e^2}{4}(\log_2 n)^2$, which gives a bound $\approx 11.3\%$ better than the paper (albeit asymptotically the same).

Suboptimal approach: So far, we have found that it is optimal for the max distance to equal the number of bus lines up to a constant factor, and for both to be $O(\log n)$. Is there a better configuration using $o(\log n)$ bus lines with max distance $\Omega(\log n)$, or a configuration with $\Omega(\log n)$ bus lines with max distance $o(\log n)$?

I've played around a bit with both of these hypotheses to no avail. One promising approach was to slightly alter the gap sizes of our bus lines: instead of having $\log_2 n$ buses with gap sizes $2^0, 2^1, 2^2, ..., 2^{\log_2 n}$, we can consider buses with gap sizes $2^0, 2^1, 2^3, 2^6, \cdots, 2^{\binom{i}{2}}, \cdots, 2^{\log_2 n}$.

In this case, the number of bus lines k fulfills $\binom{k}{2} = \log_2 n \implies k \approx \sqrt{2 \log_2 n}$. Now, any traveler on bus line i with jump size $2^{\binom{j}{2}}$ must take 2^j jumps to get to bus line $i-1$ with jump size $2^{\binom{j+1}{2}}$ ($j \approx \sqrt{2 \log_2 n} - i$ is abbreviated), so the max distance would be

$$2^0 + 2^1 + 2^3 + \cdots + 2^{\sqrt{2 \log n} - 1} = O(2^{\sqrt{\log n}})$$

This gives $kd = \sqrt{2 \log n} 2^{\sqrt{\log n}}$. Unfortunately, $2^{\sqrt{\log n}} = \omega((\log n)^c)$ for all $c > 0$, so this is worse than the existing findings.

Ultimately, I haven't been able to increase the number of bus lines by only a constant factor while decreasing the max distance by asymptotically more than a constant factor, and vice versa.

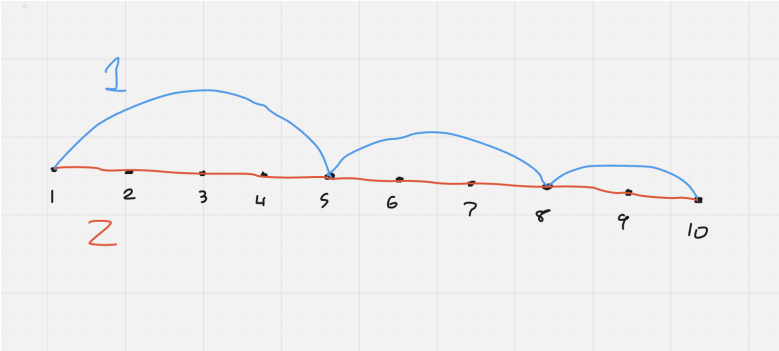
Egg drop puzzle? One lasting thought is that for each incremental bus line, it seems that we cannot asymptotically improve the max distance by anything but a constant factor; that is, it may be optimal for each new bus line to partition the gaps created by the previous bus line into a constant number of smaller gaps.

Informally, the paper approach has a general recurrence for distance as

$$d(n) = d(\frac{n}{2}) + 1$$

which evidently gives a distance logarithmic in n (assuming $k \geq \log_2 n$).

So, can we do better than halve the max distance for each incremental bus line? As it turns out, for small k, we can. As shown in the image at the start of this section, we can obtain a distance of 3 with $n = 10, k = 2$. With the recurrence above, one would expect the best distance to be 5.



The setup used in the very first diagram is resemblant of the egg-dropping puzzle[1]:

Suppose you have two eggs and you want to determine from which floors in a one hundred floor building you can drop an egg such that is doesn't break. You are to determine the minimum number of attempts you need in order to find the critical floor in the worst case while using the best strategy.

The optimal strategy uses at most $O(\sqrt{n})$ drops to find the critical floor. In particular, we drop from the $\sqrt{2n}$th floor; if the egg breaks, we check at most $\sqrt{2n}$ floors. If it doesn't break, we drop from the $\sqrt{2n} + \sqrt{2n} - 1$th floor, and continue the process.

A similar strategy is employed for the $k = 2$ bus lines: one bus line starting at n and makes its jumps iteratively bigger and bigger, while the other traverses all the bus stops which is identical to the linear runthrough once we've broken our first egg.

Now, by no means is egg dropping bijective to this puzzle; by $n = 14, k = 3$, we already have different values for d [2]. Indeed, the recursive function generated

[1] https://brilliant.org/wiki/egg-dropping/

[2] https://leetcode.com/problems/super-egg-drop/description/

by the egg dropping puzzle for n floors, k eggs is

$$D(n,k) = 1 + \min \max_i (D(i-1,k-1), D(n-i,k))$$

but I'm not too sure if a recurrence exists for this bus lines puzzle, let alone if the above recurrence holds.

As it turns out, the egg drop puzzle is binomial, meaning it is actually exponential up to a point, suggesting $d = O(\log_2 n)$ is indeed optimal. Nonetheless, if it were possible to repeatedly implement the egg drop puzzle, it may be possible to obtain a max distance of $\log_2 n$ with much fewer than $\log_2 n$ bus lines.

Chapter 4

Generalizing Oja's algorithm to k eigenvectors

The k-Oja's algorithm is significantly more complex than the 1-Oja's algorithm provided above. Below is the formal definition used by most papers[1]:

k-Oja's

initialize random orthonormal $U_0 \in \mathbb{R}^{d \times k}$
for i from 1 to n **do**
    $\tilde{U}_i \leftarrow U_{i-1} + \eta x_i x_i^\top U_{i-1}$
    $U_i \leftarrow$ orthonormalized $\tilde{U}_i$
end for
return U_n

Since we are working with arbitrary data in the adversarial setting, we can't make assumptions on x_i and hence any U_i for that matter. However, due to the orthonormalization occurring at each step, it is difficult to analyze the aggregate output U_n in the same way we can separate our analyses for v_n and $\hat{v}_n$ under 1-Oja's.

4.1 2-Oja's

The easiest way to interpret this algorithm is to analyze $k = 2$. In this case, we iterate on $U_i \in \mathbb{R}^{d \times 2}$. At each iteration, we can choose how to orthonormalize U_i, so consider a standard Gram-Schmidt process in which the first column u_1 is fixed

[1] https://arxiv.org/pdf/1607.07837, https://arxiv.org/pdf/2104.00512

while the second column u_2 is modified to fulfill $u_2 \perp u_1$.

With this approach, note that u_1 is isolated from u_2, meaning the final output $(u_1)_n$ may as well have had 1-Oja's run on it. The same cannot be said for u_2; as it turns out, we must consider both the orthogonalization and normalization of u_2 at each iteration. Consider two iterations on u_2 at any given i:

$$(u_2)_i = ((I + \eta x_i x_i^\top)(u_2)_{i-1}) - c_1(u_1)_i$$

where c_1 is such that $(u_2)_i \perp (u_1)_i$. Without normalization, we can find the next term as

$$(u_2)_{i+1} = (I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - c_2(u_1)_{i+1}$$

$$\therefore (\hat{u_2})_{i+1} = \frac{(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - c_2(u_1)_{i+1}}{||(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - c_2(u_1)_{i+1}||}$$

However, the expression obtained where $(u_2)_i$ is normalized at the first step would be

$$\hat{(u_2)}_{i+1} = \frac{(I + \eta x_{i+1} x_{i+1}^\top)(\hat{u}_2)_i - c_2(u_1)_{i+1}}{||(I + \eta x_{i+1} x_{i+1}^\top)(\hat{u}_2)_i - c_2(u_1)_{i+1}||}$$

$$= \frac{(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - ||(u_2)_i|| c_2(u_1)_{i+1}}{||(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - ||(u_2)_i|| c_2(u_1)_{i+1}||}$$

$$\neq \frac{(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - c_2(u_1)_{i+1}}{||(I + \eta x_{i+1} x_{i+1}^\top)(u_2)_i - c_2(u_1)_{i+1}||}$$

Hence, in order to analyze the movement of u_2, we must consider the orthogonalization at each step. Given this, ignoring the normalization at each iteration (as done in the proofs for 1-Oja's) is unlikely for k-Oja's, so a radically different approach from 1-Oja's will be necessary to get a good guarantee on u_2.

Chapter 5

Gaussian skip

Currently, the Gaussian lemma (lemma 3.7) is used to lower bound $||v_n|| \geq ||Bv_*||$ with high probability, giving our initial vector v_0 a "warm start"; without this lemma, we cannot assume $v_0 = v_*$.

If we can show the Gaussian lemma is unnecessary, it is possible to show that $||v_n|| \geq \Omega(\sigma_1)$ with 100% certainty. In practice, is a marginal change from probability $1-d^{-\Omega(c)}$, but does make the error-bound deterministic, which is novel even if the error bound is higher than with the warm-start assumption.

The logic of the paper is identical with the exception of the proofs that assume $v_0 = v_* \implies ||Pv_n|| \leq \sqrt{\sigma_1}$. The two proofs and their adapted versions are shown below.

5.1 Bound on the change in the orthogonal direction

Lemma 3.3 establishes the logarithmic bound on the change in the orthogonal direction of any two v_i which is critical in the matrix lemma to telescope to a $\log ||v_n||$ term.

Lemma 4. *For $v_0 = v_*$ and any indices $0 \leq a < b \leq n$,*

$$||P\hat{v}_b - P\hat{v}_a||^2 \leq 4\sigma_2 \log \frac{||v_b||}{||v_a||}$$

Proof. We are motivated by the fact that for any $||w|| = 1, w \perp v_*$, $\langle v_b - v_a, w\rangle^2 \leq \sigma_2(||v_b||^2 - ||v_a||^2)$ (which is a generalization of the proven claim that $\langle v_n -$

$v_0, w\rangle^2 \leq \sigma_2(||v_n||^2 - ||v_0||^2))$, and can try to rewrite the current expression in terms of this like so:

$$||P\hat{v}_b - P\hat{v}_a||^2 = \langle \hat{v}_b - \hat{v}_a, P\hat{v}_b - P\hat{v}_a\rangle = \langle \hat{v}_b - \hat{v}_a, \frac{P\hat{v}_b - P\hat{v}_a}{||P\hat{v}_b - P\hat{v}_a||}\rangle^2$$

$$= \langle \hat{v}_b - \hat{v}_a, w\rangle^2 = \langle \hat{v}_b - \frac{||v_a||}{||v_b||}\hat{v}_a + \frac{||v_a||}{||v_b||}v_a - \hat{v}_a, w\rangle^2$$

$$\leq 2\langle \hat{v}_b - \frac{||v_a||}{||v_b||}\hat{v}_a, w\rangle^2 + 2\langle \frac{||v_a||}{||v_b||}\hat{v}_a - \hat{v}_a, w\rangle^2$$

$$= \frac{2}{||v_b||^2}\langle v_b - v_a, w\rangle^2 + 2\left(\frac{||v_a||}{||v_b||} - 1\right)^2 \langle \hat{v}_a, w\rangle^2$$

where the inequality arises from $(x+y)^2 \leq 2x^2 + 2y^2 \; \forall x, y \in \mathbb{R}$. Hence, we are left with terms that we have already bounded (we know $\langle \hat{v}_a, w\rangle^2 \leq ||P\hat{v}_a||^2 \leq \sigma_2$), so we substitute these to get

$$||P\hat{v}_b - P\hat{v}_a||^2 \leq \frac{2}{||v_b||^2}\sigma_2(||v_b||^2 - ||v_a||^2) + 2\left(\frac{||v_a||}{||v_b||} - 1\right)^2 \sigma_2$$

$$= \sigma_2\left(4 - 4\frac{||v_a||}{||v_b||}\right)$$

$$\leq 4\sigma_2 \log \frac{||v_b||}{||v_a||}$$

where the final inequality uses $1 - 1/x \leq \log x$ for all $x > 0$. □

For our purposes, the only change to this proof for us to skip the Gaussian lemma is the assumption that $||P\hat{v}_a||^2 \leq \sigma_2$. Without this assumption, we must use $||P\hat{v}_a|| \leq \sqrt{\sigma_2} + \frac{||Pv_0||}{||v_n||}$. The analysis of the first term is the exact same, but the second term is a bit more complicated:

$$||P\hat{v}_b - P\hat{v}_a||^2 \leq \frac{2}{||v_b||^2}\sigma_2(||v_b||^2 - ||v_a||^2) + 2\left(\frac{||v_a||}{||v_b||} - 1\right)^2 \left(\sigma_2 + 2\sqrt{\sigma_2}\frac{||Pv_0||}{||v_n||} + \frac{||Pv_0||^2}{||v_n||^2}\right)$$

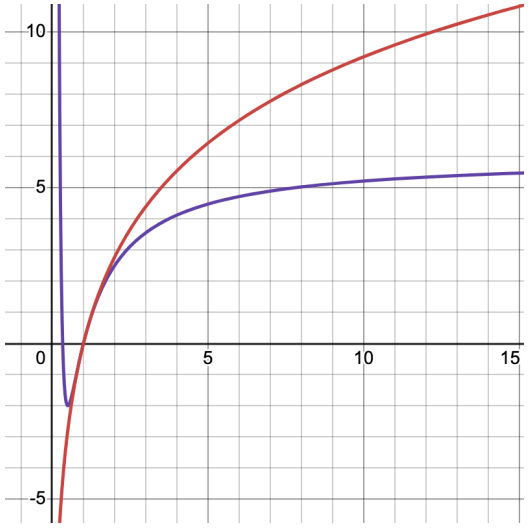
$$= 4\sigma_2\left(1 - \frac{||v_a||}{||v_b||}\right) + 2\left(\frac{||v_a||}{||v_b||} - 1\right)^2 \left(2\sqrt{\sigma_2}\frac{||Pv_0||}{||v_n||} + \frac{||Pv_0||^2}{||v_n||^2}\right)$$

In other words, we have an extra $2\sqrt{\sigma_2}\frac{||Pv_0||}{||v_n||} + \frac{||Pv_0||^2}{||v_n||^2}$ term to deal with. Assume this quantity is at most $c\sigma_2$ for some $c > 0$; we seek to find how loose c can be.

$$||P\hat{v}_b - P\hat{v}_a||^2 \leq 4\sigma_2\left(1 - \frac{||v_a||}{||v_b||}\right) + 2c\sigma_2\left(\frac{||v_a||}{||v_b||} - 1\right)^2$$
$$= \sigma_2\left((4+2c) - (4+4c)\frac{||v_a||}{||v_b||} + 2c\frac{||v_a||^2}{||v_b||^2}\right) \stackrel{?}{\leq} O\left(\sigma_2 \log\frac{||v_b||}{||v_a||}\right)$$

We'd like the $\log\frac{||v_b||}{||v_a||}$ term to remain in our end result, as it is used to telescope a summation in the matrix lemma. Hence, it remains to be shown that $(4+2c) - (4+4c)\frac{||v_a||}{||v_b||} + 2c\frac{||v_a||^2}{||v_b||^2} \leq O(\log\frac{||v_b||}{||v_a||})$.

Let $x = ||v_b||/||v_a||$; since $||v_i||$ is strictly growing, $x \geq 1$. Now, we seek to find a c such that $(4+2c) - \frac{(4+4c)}{x} + \frac{2c}{x^2} \leq c' \log x$ for some $c' > 0$ and all $x \geq 1$. After playing with this in Desmos, I noticed that for $c' = 4$, $c = 1$ is the supremum. For the time being, we'll set $c' = 4$ so as not to inflate the coefficient of our inequality, but it is still reasonable to do so.



In short, even without the assumption, we obtain the same result in the paper: $||P\hat{v}_b - P\hat{v}_a||^2 \leq 4\sigma_2 \log\frac{||v_b||}{||v_a||}$.

It now remains to be shown that $2\sqrt{\sigma_2}\frac{||Pv_0||}{||v_n||} + \frac{||Pv_0||^2}{||v_n||^2} \leq c\sigma_2 = \sigma_2$. Upon using the quadratic formula with respect to $\frac{||Pv_0||}{||v_n||}$, we get that $\frac{||Pv_0||}{||v_n||} \leq (\sqrt{2}-1)\sqrt{\sigma_2}$.

As of writing this, I have attempted to show that $||Pv_0|| \leq \frac{\sqrt{\sigma_2}}{3}||v_n||$, which would suffice. Given our bounds on $||v_n||$, I would assume this conclusion is true, and have continued to use this conjecture in the next proof.

Since our result is the same as the one in the paper, the matrix lemma ends up being the same as well, so our bound remains

$$\sum_{i=1}^{n}\langle x_i, P\hat{v}_{i-1}\rangle^2 \leq 4\sigma_2^2(1+\log_2 n)^2 \log||v_n||$$

5.2 Growth bound

When bringing our results back into the growth bound, we get

$$2\log||v_n|| \geq \eta\sum_{i=1}^{n}\langle x_i, \hat{v}_{i-1}\rangle^2 \geq \eta\sum_{i=1}^{n}\frac{1-||P\hat{v}_{i-1}||^2}{2}\langle x_i, v_*\rangle^2 - \langle x_i, P\hat{v}_{i-1}\rangle^2$$

$$\geq \eta\sum_{i=1}^{n}\frac{1-\left(\sqrt{\sigma_2}+||Pv_0||/||v_n||\right)^2}{2}\langle x_i, v_*\rangle^2 - 4\sigma_2^2(1+\log_2 n)^2\log||v_n||$$

$$\therefore \log||v_n|| \geq \frac{\left(1-(\sqrt{\sigma_2}+||Pv_0||/||v_n||)^2\right)\sigma_1}{4(1+2\sigma_2^2(1+\log_2 n)^2)}$$

$$\geq \frac{\left(1-\frac{16}{9}\sigma_2\right)\sigma_1}{4+8\sigma_2^2(1+\log_2 n)^2} \geq \frac{\sigma_1}{36+72\sigma_2^2(1+\log_2 n)^2}$$

by our assumptions that $\frac{||Pv_0||}{||v_n||} \leq \sqrt{\sigma_2}/3$ and $\sigma_2 \leq 1/2$.

For $\sigma_2 \leq \frac{1}{C\log n}$, this lower bound on $\log||v_n||$ is $\Omega(\sigma_1)$, as desired.

5.3 Bounds on C

We now consider how strict our constraints on C must be for $\sigma_1 \geq C\log d$ and $\sigma_2 \leq \frac{1}{C\log n}$. For $(1+\log_2 n)^2 \approx \log_2 n^2$, we have that

$$\log||v_n|| \geq \frac{C\log d}{36+\frac{72}{C^2}} = \frac{C^3}{36C^2+72}\log d$$

To obtain $\log||v_n|| \geq 10\log d$ which is necessary for the algorithm to germinate, we'd need $C > 360$. Even without this constraint, to guarantee $||P\hat{v}_n|| \leq \sqrt{\sigma_2} + d^{-1}$ with probability $1-d^{-1/3}$, we'd need $\log||v_n|| \geq 2\log d \implies C > 72$.

Ultimately, these bounds on C are still within reason compared to the bound of $C > 92$ discussed in the tightness of $||v_n||$, which completes the Gaussian skip.