Roadmap In the appendix, we present the complete version of our proposed algorithm and main theorem. as well as rigorous proof. In Section A, we list our notations and some widely used mathematical results. In Section B we discuss coordinate-wise embedding – the sketching technique we propose in this work. We present some commonly used sketching matrix and their corresponding results. In Section C we discuss how to deal with the matrix-vector multiplication bottleneck through sketching rigorously. We also present our main Algorithm 6. We summarize our main Theorem D.1 in Section D. In Section E, we move on to discuss how to deal with the projection maintenance bottleneck through the lazy update and low-rank update. In Section F, we present the strength of our approach compared to previous state of the art results. We discuss the benefits of being feasible and oblivious of our approach. In Section H, we compare our sketching approach to the classical "sketch and solve" approach and discuss the reasons why the classical approach doesn't work in our setting.

A. Preliminaries

A.1. Notations

For notation convenience, we assume the number of variables $n \geq 10$ and there is no redundant constraints. In particular, this implies that the constraint matrix A is full rank and $n \geq d$.

For a positive integer n, let $[n]$ denote the set $\{1, 2, \cdots, n\}$.

For any function f, we define $\widetilde{O}(f)$ to be $f \cdot \log^{O(1)}(f)$. In addition to $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq Cg$ (resp. $\geq$) for some absolute constant C.

We use $\sinh x$ to denote $\frac{e^x - e^{-x}}{2}$ and $\cosh x$ to denote $\frac{e^x + e^{-x}}{2}$.

For vectors $a, b \in \mathbb{R}^n$ and accuracy parameter $\epsilon \in (0, 1)$, we use $a \approx_\epsilon b$ to denote that $(1-\epsilon)b_i \leq a_i \leq (1+\epsilon)b_i, \forall i \in [n]$. Similarly, for any scalar t, we use $a \approx_\epsilon t$ to denote that $(1-\epsilon)t \leq a_i \leq (1+\epsilon)t, \forall i \in [n]$.

For a vector $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^n$, we use xs to denote a length n vector with the i-th coordinate $(xs)_i$ is $x_i \cdot s_i$. Similarly, we extend other scalar operations to vector coordinate-wise.

Given vectors $x, s \in \mathbb{R}^n$, we use X and S to denote the diagonal matrix of those two vectors. We use $\frac{X}{S}$ to denote the diagonal matrix given $(\frac{X}{S})_{i,i} = x_i / s_i$. Similarly, we extend other scalar operations to diagonal matrix diagonal-wise. Note that matrix $\sqrt{\frac{X}{S}} A^\top (A \frac{X}{S} A^\top)^{-1} A \sqrt{\frac{X}{S}}$ is an orthogonal projection matrix.

A.2. Inequalities

Lemma A.1 ((Cohen et al., 2019b)). *Let x and y are (possibly dependent) random variables such that $|x| \leq c_x$ and $|y| \leq c_y$ almost surely. Then, we have*

$$\mathbf{Var}[xy] \leq 2c_x^2 \cdot \mathbf{Var}[y] + 2c_y^2 \cdot \mathbf{Var}[x].$$

A.3. Probability tools

Lemma A.2 (Chernoff bound (Chernoff, 1952)). *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbf{E}[X] = \sum_{i=1}^n p_i$. Then*
1. $\Pr[X \geq (1+\delta)\mu] \leq \exp(-\delta^2 \mu/3)$, $\forall \delta > 0$;
2. $\Pr[X \leq (1-\delta)\mu] \leq \exp(-\delta^2 \mu/2)$, $\forall 0 < \delta < 1$.

Lemma A.3 (Hoeffding bound (Hoeffding, 1963)). *Let $X_1, \cdots, X_n$ denote n independent bounded variables in $[a_i, b_i]$. Let $X = \sum_{i=1}^n X_i$, then we have*

$$\Pr[|X - \mathbf{E}[X]| \geq t] \leq 2 \exp\left(- \frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right).$$

Lemma A.4 (Bernstein inequality (Bernstein, 1924)). *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^n X_i > t \right] \leq \exp\left(- \frac{t^2/2}{\sum_{j=1}^n \mathbf{E}[X_j^2] + Mt/3} \right).$$

We state Khintchine's inequality

Lemma A.5 (Khintchine's inequality, (Khintchine, 1923; Haagerup, 1981)). *Let $\sigma_1, \cdots, \sigma_n$ be i.i.d. sign random variables, and let $z_1, \cdots, z_n$ be real numbers. Then there are constants $C > 0$ so that for all $t > 0$*

$$\Pr\left[\left|\sum_{i=1}^n z_i \sigma_i\right| \geq t\|z\|_2\right] \leq \exp(-Ct^2).$$

We state Hason-wright inequality here

Lemma A.6 (Hason-wright inequality (Hanson & Wright, 1971; Rudelson & Vershynin, 2013)). *Let $x \in \mathbb{R}^n$ denote a random vector with independent entries x_i with $\mathbf{E}[x_i] = 0$ and $|x_i| \leq K$. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\Pr[|x^\top A x - \mathbf{E}[x^\top A x]| > t] \leq 2 \cdot \exp(-c \min\{t^2/(K^4\|A\|_F^2), t/(K^2\|A\|)\}).$$

Lemma A.7 (Lemma 1 on page 1325 of Laurent and Massart (Laurent & Massart, 2000)). *Let $X \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero mean and σ^2 variance. Then*

$$\Pr[X - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t),$$
$$\Pr[k\sigma^2 - X \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t).$$

Lemma A.8 (Tail bound for sub-exponential distribution (Foss et al., 2011)). *We say $X \in \mathrm{SE}(\sigma^2, \alpha)$ with parameters $\sigma > 0, \alpha > 0$ if:*

$$\mathbf{E}[e^{\lambda X}] \leq \exp(\lambda^2\sigma^2/2), \quad \forall |\lambda| < 1/\alpha.$$

Let $X \in \mathrm{SE}(\sigma^2, \alpha)$ and $\mathbf{E}[X] = \mu$, then:

$$\Pr[|X - \mu| \geq t] \leq \exp(-0.5 \min\{t^2/\sigma^2, t/\alpha\}).$$

Lemma A.9 (Matrix Chernoff bound (Tropp, 2011; Lu et al., 2013)). *Let $\mathcal{X}$ be a finite set of positive-semidefinite matrices with dimension $d \times d$, and suppose that*

$$\max_{X \in \mathcal{X}} \lambda_{\max}(X) \leq B.$$

Sample $\{X_1, \cdots, X_n\}$ uniformly at random from $\mathcal{X}$ without replacement. We define $\mu_{\min}$ and $\mu_{\max}$ as follows:

$$\mu_{\min} := n \cdot \lambda_{\min}(\mathop{\mathbf{E}}_{X \sim \mathcal{X}}[X]) \text{ and } \mu_{\max} := n \cdot \lambda_{\max}(\mathop{\mathbf{E}}_{X \sim \mathcal{X}}[X]).$$

Then

$$\Pr\left[\lambda_{\min}(\sum_{i=1}^n X_i) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \exp(-\delta^2\mu_{\min}/B) \text{ for } \delta \in [0, 1),$$
$$\Pr\left[\lambda_{\max}(\sum_{i=1}^n X_i) \geq (1+\delta)\mu_{\max}\right] \leq d \cdot \exp\left(-\delta^2\mu_{\max}/(4B)\right) \text{ for } \delta \geq 0.$$

A.4. Fast matrix multiplication

In this work, we use the following fast matrix multiplication results:

- Multiplication of two matrices of size $n \times n$ requires $n^{\omega+o(1)}$ running time, where ω is the exponent of matrix multiplication. Current value of ω is roughly 2.373 (Williams, 2012; Le Gall, 2014).
- Inverse of a matrix of size $n \times n$ also requires $n^{\omega+o(1)}$ running time.
- Multiplication of two matrices of size $n \times n$ and $n \times n^a$ requires $n^{2+o(1)}$ running time if $a \in [0, \alpha)$, where α is the dual exponent of matrix multiplication. Current value of α is roughly 0.314 (Le Gall & Urrutia, 2018).

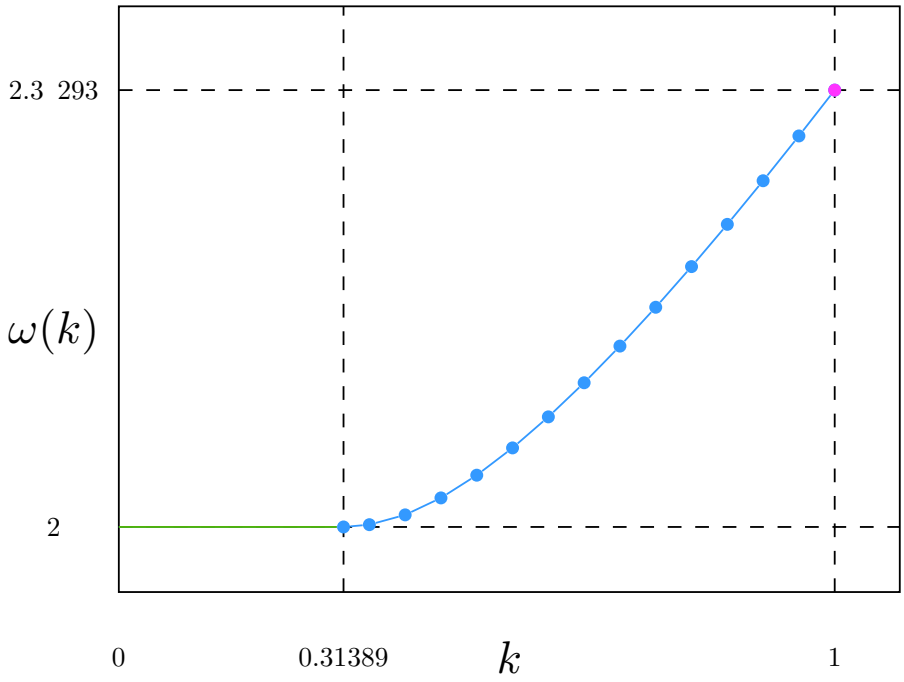


Figure 2: Current matrix multiplication time: the blue and green line represents current running time $\omega(k)$ of multiplying matrices of size $n \times n^k$ and $n^k \times n$ for $k \in [0, 1]$: when $k = 1$, multiplication of two square matrices needs roughly $n^{2.373}$ running time; when $k < 0.314$, multiplication needs $n^{2+o(1)}$ running time

B. Sketching

In this section, we discuss the (α, β, δ)-coordinate wise embedding property we proposed in this work through several commonly used sketching matrices.

We consider several standard sketching matrices:

1. Random Gaussian matrices.
2. Subsampled randomized Hadamard/Fourier transform matrices (Lu et al., 2013).
3. AMS sketch matrices (Alon et al., 1999), random $\{-1, +1\}$ per entry.
4. Count-Sketch matrices (Charikar et al., 2002), each column only has one non-zero entry, and is $-1, +1$ half probability each.
5. Sparse embedding matrices (Nelson & Nguyên, 2013), each column only has s non-zero entries, and each entry is $-\frac{1}{\sqrt{s}}, +\frac{1}{\sqrt{s}}$ half probability each.
6. Uniform sampling matrices.

We list the definitions and results of above sketching matrices for coordinate-wise embedding in Table 3.

B.1. Definition

Definition B.1 (k-wise independence)**.** *$\mathcal{H} = \{h : [m] \to [l]\}$ is a k-wise independent hash family if $\forall i_1 \neq i_2 \neq \cdots \neq i_k \in [n]$ and $\forall j_1, \cdots, j_k \in [l]$,*

$$\Pr_{h \in \mathcal{H}}[h(i_1) = j_1 \wedge \cdots \wedge h(i_k) = j_k] = \frac{1}{l^k}.$$

Sketching matrix	Definition	Expectation	Variance	Inner Product	Concentration
Random Gaussian	Definition B.2	Lemma B.11	Lemma B.13	Lemma B.18	Lemma B.24
SRHT	Definition B.3	Lemma B.11	Lemma B.12	Lemma B.19	Lemma B.23
AMS	Definition B.4	Lemma B.11	Lemma B.12	Lemma B.20	Lemma B.23
Count-sketch	Definition B.5	Lemma B.11	Lemma B.14	Lemma B.21	Lemma B.25
Sparse embedding	Definition B.6,B.7	Lemma B.11	Lemma B.15	Lemma B.22	Lemma B.28
Uniform sampling	Definition B.8	Lemma B.11	Lemma B.16		Lemma B.29

Table 3: Roadmap of the results for coordinate-wise embedding

Definition B.2 (Random Gaussian matrix)**.** *We say $R \in \mathbb{R}^{b\times n}$ is a random Gaussian matrix if all entries are sampled from $\mathcal{N}(0, 1/b)$ independently.*

Definition B.3 (Subsampled randomized Hadamard/Fourier transform matrix (Lu et al., 2013))**.** *We say $R \in \mathbb{R}^{b\times n}$ is a subsampled randomized Hadamard transform (SRHT) matrix[5] if it is of the form $R = \sqrt{n/b}SHD$, where $S \in \mathbb{R}^{b\times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, $H \in \mathbb{R}^{n\times n}$ is a normalized Walsh-Hadamard matrix, and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.*

Definition B.4 (AMS sketch matrix (Alon et al., 1999))**.** *Let $h_1, h_2, \cdots, h_b$ be b random hash functions picking from a 4-wise independent hash family $\mathcal{H} = \{h : [n] \to \{-\frac{1}{\sqrt{b}}, +\frac{1}{\sqrt{b}}\}\}$. Then $R \in \mathbb{R}^{b\times n}$ is a AMS sketch matrix if we set $R_{i,j} = h_i(j)$.*

Definition B.5 (Count-sketch matrix (Charikar et al., 2002))**.** *Let $h : [n] \to [b]$ be a random 2-wise independent hash function and $\sigma : [n] \to \{-1, +1\}$ be a random 4-wise independent hash function. Then $R \in \mathbb{R}^{b\times n}$ is a count-sketch matrix if we set $R_{h(i),i} = \sigma(i)$ for all $i \in [n]$ and other entries to zero.*

Definition B.6 (Sparse embedding matrix I (Nelson & Nguyên, 2013))**.** *We say $R \in \mathbb{R}^{b\times n}$ is a sparse embedding matrix with parameter s if each column has exactly s non-zero elements being $\pm 1/\sqrt{s}$ uniformly at random, whose locations are picked uniformly at random without replacement (and independent across columns) [6].*

Definition B.7 (Sparse embedding matrix II (Nelson & Nguyên, 2013))**.** *Let $h : [n] \times [s] \to [b/s]$ be a a ramdom 2-wise independent hash function and $\sigma : [n] \times [s] \to \{-1, 1\}$ be a 4-wise independent. Then $R \in \mathbb{R}^{b\times n}$ is a sparse embedding matrix II with parameter s if we set $R_{(j-1)b/s+h(i,j),i} = \sigma(i, j)/\sqrt{s}$ for all $(i, j) \in [n] \times [s]$ and all other entries to zero.[7]*

Definition B.8 (Uniform sampling matrix)**.** *We say $R \in \mathbb{R}^{b\times n}$ is a uniform sampling matrix if it is of the form $R = \sqrt{n/b}SD$, where $S \in \mathbb{R}^{b\times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.*

B.2. Coordinate wise embedding

We define coordinate-wise embedding as follows:

Definition B.9 ((α, β, δ)-coordinate wise embedding)**.** *We say a randomized matrix $R \in \mathbb{R}^{b\times n}$ satisfying (α, β, δ)-coordinate wise embedding if*

$$
\begin{aligned}
&1.\ \mathop{\mathbf{E}}_{R\sim\Pi}[g^\top R^\top Rh] = g^\top h, \\
&2.\ \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \le (g^\top h)^2 + \frac{\alpha}{b}\|g\|_2^2\|h\|_2^2, \\
&3.\ \Pr_{R\sim\Pi}\left[|g^\top R^\top Rh - g^\top h| \ge \frac{\beta}{\sqrt{b}}\|g\|_2\|h\|_2\right] \le \delta.
\end{aligned}
$$

Remark B.10. *Given a randomized matrix $R \in \mathbb{R}^{b\times n}$ satisfying (α, β, δ)-coordinate wise embedding and any orthogonal*

[5] In this case, we require $\log n$ to be an integer.

[6] For our purposes the signs need only be $O(\log d)$-wise independent, and each column can be specified by a $O(\log d)$-wise independent permutation, and the seeds specifying the permutations in different columns need only be $O(\log d)$-wise independent.

[7] This definition has the same behavior as sparse embedding matrix I for our purpose.

projection $P \in \mathbb{R}^{n \times n}$, above definition implies

$$
\begin{aligned}
&1.\ \mathop{\mathbf{E}}_{R\sim\Pi}[PR^\top Rh] = Ph, \\
&2.\ \mathop{\mathbf{E}}_{R\sim\Pi}[(PR^\top Rh)_i^2] \leq (Ph)_i^2 + \frac{\alpha}{b}\|h\|_2^2, \\
&3.\ \Pr_{R\sim\Pi}\left[|(PR^\top Rh)_i - (Ph)_i| \geq \frac{\beta}{\sqrt{b}}\|h\|_2\right] \leq \delta.
\end{aligned}
$$

since $\|P\|_2 \leq 1$ implies $\|P_{i,:}\|_2 \leq 1$ for all $i \in [n]$.

B.3. Expectation and variance

Lemma B.11. *Let $R \in \mathbb{R}^{b\times n}$ denote any of the random matrix in Definition B.2, B.3, B.4, B.6, B.7, B.8. Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ denote two fixed vectors. Then the following properties hold:*

$$
\mathop{\mathbf{E}}_{R\sim\Pi}[g^\top R^\top Rh] = g^\top h
$$

Proof.

$$
\mathop{\mathbf{E}}_{R\sim\Pi}[g^\top R^\top Rh] = g^\top \mathop{\mathbf{E}}_{R\sim\Pi}[R^\top R]h = g^\top Ih = g^\top h.
$$

□

Lemma B.12. *Let $R \in \mathbb{R}^{b\times n}$ be a SRHT or AMS sketch matrix as in Definition B.3, B.4. Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ denote two vectors. Then the following properties hold:*

$$
\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{2}{b}\|g\|_2^2 \cdot \|h\|_2^2.
$$

Proof. If $\mathbf{E}_a[a] = b$, it is easy to see that

$$
\mathop{\mathbf{E}}_{a}[(a-b)^2] = \mathop{\mathbf{E}}_{a}[a^2 - 2ab + b^2] = \mathop{\mathbf{E}}_{a}[a^2 - b^2]
$$

We can rewrite it as follows:

$$
\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2 - (g^\top h)^2] = \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top (R^\top R - I)h)^2],
$$

It can be bounded as follows:

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top (R^\top R - I)h)^2] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b (Rg)_k (Rh)_k - g^\top h\right)^2\right] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b \sum_{i=1}^n R_{k,i} g_i \cdot \sum_{j\in[n]\setminus\{i\}} R_{k,j} h_j\right)^2\right] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b \sum_{i=1}^n R_{k,i} g_i \cdot \sum_{j\in[n]\setminus\{i\}} R_{k,j} h_j\right) \cdot \left(\sum_{k'=1}^b \sum_{i'=1}^n R_{k',i'} g_{i'} \cdot \sum_{j'\in[n]\setminus\{i'\}} R_{k',j'} h_{j'}\right)\right] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b \sum_{i=1}^n R_{k,i}^2 g_i^2 \cdot \sum_{j\in[n]\setminus\{i\}} R_{k,j}^2 h_j^2\right) + \left(\sum_{k=1}^b \sum_{i=1}^n R_{k,i}^2 g_i h_i \cdot \sum_{j\in[n]\setminus\{i\}} R_{k,j}^2 g_j h_j\right)\right]
\end{aligned}
$$

$$= \frac{1}{b}\left(\sum_{i=1}^n g_i^2 \sum_{j\in[n]\setminus\{i\}} h_j^2\right) + \frac{1}{b}\left(\sum_{i=1}^n g_i h_i \sum_{j\in[n]\setminus\{i\}} g_j h_j\right)$$
$$\leq \frac{2}{b}\|g\|_2^2\|h\|_2^2,$$

where the second step follows from $R_{k,i}^2 = 1/b, \forall k, i \in [b]\times[n]$, the forth step follows from $\mathbf{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] \neq 0$ only if $i = i'$, $j = j'$, $k = k'$ or $i = j'$, $j = i'$, $k = k'$, the fifth step follows from $R_{k,i}$ and $R_{k,j}$ are independent if $i \neq j$ and $R_{k,i}^2 = R_{k,j}^2 = 1/b$, and the last step follows from Cauchy-Schwartz inequality.

Therefore,

$$\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2 - (g^\top h)^2] = \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top(R^\top R - I)h)^2] \leq \frac{2}{b}\|g\|_2^2\|h\|_2^2.$$

□

Lemma B.13. *Let $R \in \mathbb{R}^{b\times n}$ be a random Gaussian matrix as in Definition B.2. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{3}{b}\|g\|_2^2 \cdot \|h\|_2^2.$$

Proof. Note

$$\begin{aligned}
&\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \\
=& \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}g_i \cdot \sum_{j=1}^n R_{k,j}h_j\right)^2\right] \\
=& \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}g_i \cdot \sum_{j=1}^n R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^b\sum_{i'=1}^n R_{k',i'}g_{i'} \cdot \sum_{j'=1}^n R_{k',j'}h_{j'}\right)\right] \\
=& \mathop{\mathbf{E}}_{R\sim\Pi}\Big[\left(\sum_{k=1}^b\sum_{k'\in[b]\setminus\{k\}}\sum_{i=1}^n\sum_{i'=1}^n R_{k,i}^2R_{k',i'}^2 g_ih_ig_{i'}h_{i'}\right) + \left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}^4 g_i^2h_i^2\right) \\
&+ \left(\sum_{k=1}^b\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} R_{k,i}^2R_{k,j}^2 g_i^2h_j^2\right) + \left(\sum_{k=1}^n\sum_{i=1}^n\sum_{i'\in[n]\setminus\{i\}} R_{k,i}^2R_{k,i'}^2 g_ih_ig_{i'}h_{i'}\right) \\
&+ \left(\sum_{k=1}^b\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} R_{k,i}^2R_{k,j}^2 g_ih_jg_jh_i\right)\Big] \\
=& \frac{b-1}{b}\sum_{i=1}^n\sum_{i'=1}^n g_ih_ig_{i'}h_{i'} + \frac{3}{b}\sum_{i=1}^n g_i^2h_i^2 \\
&+ \frac{1}{b}\sum_{i=1}^n\sum_{j\in[n]\setminus[i]} g_i^2h_j^2 + \frac{1}{b}\sum_{i=1}^n\sum_{i'\in[n]\setminus[i]} g_ih_ig_{i'}h_{i'} + \frac{1}{b}\sum_{i=1}^n\sum_{j\in[n]\setminus[i]} g_ih_jg_jh_i \\
\leq& (g^\top h)^2 + \frac{3}{b}\|g\|_2^2\|h\|_2^2,
\end{aligned}$$

where the third step follows from that for independent entries of a random Gaussian matrix, $\mathbf{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] \neq 0$ only if 1. $k \neq k'$, $i = j$, $i' = j'$, or 2. $k = k'$, $i = i' = j = j'$, or 3. $k = k'$, $i = i' \neq j = j'$, or 4. $k = k'$, $i = j \neq i' = j'$, or 5. $k = k'$, $i = j' \neq i' = j$, the fourth step follows from $\mathbf{E}[R_{k,i}^2] = 1/b$ and $\mathbf{E}[R_{k,i}^4] = 3/b^2$, and the last step follows from Cauchy-Schwartz inequality. □

Lemma B.14. *Let $R \in \mathbb{R}^{b\times n}$ be a count-sketch matrix as in Definition B.5. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{3}{b}\|g\|_2^2\|h\|_2^2.$$

Proof. Note

$$\begin{aligned}
&\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}g_i\sum_{j=1}^n R_{k,j}h_j\right)^2\right] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}g_i\sum_{j=1}^n R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^b\sum_{i'=1}^n R_{k',i'}g_{i'}\sum_{j'=1}^n R_{k',j'}h_{j'}\right)\right] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\Bigg[\left(\sum_{k=1}^b\sum_{k'\in[b]\setminus\{k\}}\sum_{i=1}^n\sum_{i'\in[n]\setminus\{i\}} R_{k,i}^2R_{k',i'}^2g_ih_ig_{i'}h_{i'}\right)+\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}^4g_i^2h_i^2\right) \\
&+\left(\sum_{k=1}^b\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} R_{k,i}^2R_{k,j}^2g_i^2h_j^2\right)+\left(\sum_{k=1}^n\sum_{i=1}^n\sum_{i'\in[n]\setminus\{i\}} R_{k,i}^2R_{k,i'}^2g_ih_ig_{i'}h_{i'}\right) \\
&+\left(\sum_{k=1}^b\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} R_{k,i}^2R_{k,j}^2g_ih_jg_jh_i\right)\Bigg] \\
=\ &\frac{b-1}{b}\sum_{i=1}^n\sum_{i'\in[n]\setminus i} g_ih_ig_{i'}h_{i'}+\sum_{i=1}^n g_i^2h_i^2 \\
&+\frac{1}{b}\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} g_i^2h_j^2+\frac{1}{b}\sum_{i=1}^n\sum_{i'\in[n]\setminus\{i\}} g_ih_ig_{i'}h_{i'}+\frac{1}{b}\sum_{i=1}^n\sum_{j\in[n]\setminus\{i\}} g_ih_jg_jh_i \\
\leq\ &(g^\top h)^2+\frac{3}{b}\|g\|_2^2\|h\|_2^2,
\end{aligned}$$

where in the third step we are again considering what values of k, k', i, i', j, j' that makes $\mathbf{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] \neq 0$. Since the hash function $\sigma(\cdot)$ of the count-sketch matrix is 4-wise independent, $\forall k, k'$, when $i \neq i' \neq j \neq j'$, or $i = i' = j \neq j'$ (and the other 3 symmetric cases), we have that $\mathbf{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] = 0$. Since the count-sketch matrix has only one non-zero entry in every column, when $k \neq k'$, if $i = i'$ or $i = j'$ or $j = i'$ or $j = j'$, we also have $\mathbf{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] = 0$. Thus we only need to consider the cases: 1. $k \neq k'$, $i = j \neq i' = j'$, or 2. $k = k'$, $i = i' = j = j'$, or 3. $k = k'$, $i = i' \neq j = j'$, or 4. $k = k'$, $i = j \neq i' = j'$, or 5. $k = k'$, $i = j' \neq i' = j$. And the fourth step follows from $\mathbf{E}[R_{k,i}^2] = 1/b$ and $\mathbf{E}[R_{k,i}^4] = 1/b$, and the last step follows from Cauchy-Schwartz inequality. □

Lemma B.15. *Let $R \in \mathbb{R}^{b\times n}$ be a sparse embedding matrix as in Definition B.6, B.7. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$2.\ \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{2}{b}\|g\|_2^2\cdot\|h\|_2^2.$$

Proof. Note

$$\begin{aligned}
&\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \\
=\ &\mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^b\sum_{i=1}^n R_{k,i}g_i\sum_{j=1}^n R_{k,j}h_j\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \sum_{j=1}^{n} R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'}g_{i'} \sum_{j'=1}^{n} R_{k',j'}h_{j'}\right)\right] \\
&= \mathop{\mathbf{E}}_{R\sim\Pi}\Big[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2 g_i^2 \sum_{j\in[n]\setminus\{i\}} R_{k,j}^2 h_j^2\right) + \left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2 g_i h_i \sum_{j\in[n]\setminus\{i\}} R_{k,j}^2 g_j h_j\right) \\
&\quad + \left(\sum_{k}\sum_{i\neq i'} R_{k,i}^2 R_{k,i'}^2 g_i h_i g_{i'} h_{i'}\right) + \left(\sum_{k}\sum_{i} R_{k,i}^4 g_i^2 h_i^2\right) + \left(\sum_{k\neq k'}\sum_{i\neq i'} R_{k,i}^2 R_{k',i'}^2 g_i h_i g_{i'} h_{i'}\right) \\
&\quad + \left(\sum_{k\neq k'}\sum_{i} R_{k,i}^2 R_{k',i}^2 g_i^2 h_i^2\right)\Big] \\
&= \frac{1}{b}\sum_{i\neq j} g_i^2 h_j^2 + \frac{1}{b}\sum_{i\neq j} g_i h_i g_j h_j + \frac{1}{b}\sum_{i\neq i'} g_i h_i g_{i'} h_{i'} + \frac{1}{s}\sum_{i} g_i^2 h_i^2 + \frac{b-1}{b}\sum_{i\neq i'} g_i h_i g_{i'} h_{i'} + \frac{s-1}{s}\sum_{i} g_i^2 h_i^2 \\
&\leq (g^\top h)^2 + \frac{2}{b}\|g\|_2^2\|h\|_2^2,
\end{aligned}
$$

where the third step follows from the fact that the sparse embedding matrix has independent columns and s non-zero entry in every column, the fourth step follows from $\mathbf{E}[R_{k,i}^2] = 1/b$, $\mathbf{E}[R_{k,i}^4] = 1/(bs)$, and $\mathbf{E}[R_{k,i}^2 R_{k',i}^2] = \frac{s(s-1)}{b(b-1)}\cdot\frac{1}{s^2}, \forall k \neq k'$ and the last step follows from Cauchy-Schwartz inequality. □

Lemma B.16. *Let $R \in \mathbb{R}^{b\times n}$ be a uniform sampling matrix as in Definition B.8. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$
2.\ \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top R h)^2] \leq (g^\top h)^2 + \frac{n}{b}\|g\|_2^2\|h\|_2^2.
$$

Proof. Note

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top R h)^2] \\
&= \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \sum_{j=1}^{n} R_{k,j}h_j\right)^2\right] \\
&= \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \sum_{j=1}^{n} R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'}g_{i'} \sum_{j'=1}^{n} R_{k',j'}h_{j'}\right)\right] \\
&= \mathop{\mathbf{E}}_{R\sim\Pi}\left[\left(\sum_{k}\sum_{i} R_{k,i}^4 g_i^2 h_i^2\right) + \left(\sum_{k\neq k'}\sum_{i\neq i'} R_{k,i}^2 R_{k',i'}^2 g_i h_i g_{i'} h_{i'}\right)\right] \\
&= \frac{n}{b}\sum_{i} g_i^2 h_i^2 + \frac{(b-1)n}{(n-1)b}\sum_{i\neq i'} g_i h_i g_{i'} h_{i'} \\
&\leq (g^\top h)^2 + \frac{n}{b}\|g\|_2^2\|h\|_2^2,
\end{aligned}
$$

where the third step follows from the fact that the random sampling matrix has one non-zero entry in every row, the fourth step follows from $\mathbf{E}[R_{k,i}^2 R_{k',i'}^2] = n/((n-1)b^2)$ for $k \neq k'$, $i \neq i'$ and $\mathbf{E}[R_{k,i}^4] = n/b^2$. □

Remark B.17. *Lemma B.16 indicates that uniform sampling fails in bounding variance in some sense, since the upper bound give here involves n.*

B.4. Bounding inner product

Lemma B.18 (Gaussian)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a random Gaussian matrix (Definition B.2). Then we have:*

$$\Pr\Big[\max_{i\neq j}|\langle R_{*,i}, R_{*,j}\rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).$$

Proof. Note for $i \neq j$, $R_{*,i}, R_{*,j} \sim \mathcal{N}(0, \frac{1}{b}I_b)$ are two independent Gaussian vectors. Let $z_k = R_{k,i}R_{k,j}$ and $z = \langle R_{*,i}, R_{*,j}\rangle$. Then we have for any $|\lambda| \leq b/2$,

$$\mathbf{E}[e^{\lambda z_k}] = \frac{1}{\sqrt{1-\lambda^2/b^2}} \leq \exp(\lambda^2/b^2),$$

where the first step follows from $z_k = \frac{1}{4}(R_{k,i}+R_{k,j})^2 + \frac{1}{4}(R_{k,i}-R_{k,j})^2 = \frac{b}{2}(Q_1 - Q_2)$ where $Q_1, Q_2 \sim \chi^2_1$, and $\mathbf{E}[e^{\lambda Q}] = \frac{1}{\sqrt{1-2\lambda}}$ for any $Q \sim \chi^2_1$.

This implies $z_k \in \mathrm{SE}(2/b^2, 2/b)$ is a sub-exponential random variable. Thus, we have $z = \sum_{k=1}^b z_k \in \mathrm{SE}(2/b, 2/b)$, by sub-exponential concentration Lemma A.8 we have

$$\Pr[|z| \geq t] \leq 2\exp(-bt^2/4)$$

for $0 < t < 1$. Picking $t = \sqrt{\log(n^2/\delta)/b}$, we have

$$\Pr\Big[|\langle R_{*,i}, R_{*,j}\rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \delta/n^2.$$

Taking the union bound over all $(i,j) \in [n]\times[n]$ and $i \neq j$, we complete the proof. □

Lemma B.19 (SRHT)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a SRHT (Definition B.3). Then we have:*

$$\Pr\Big[\max_{i\neq j}|\langle R_{*,i}, R_{*,j}\rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).$$

Proof. For fixed $i \neq j$, let $X = [R_{*,i}, R_{*,j}] \in \mathbb{R}^{b\times 2}$. Then $X^\top X = \sum_{k=1}^b G_k$, where

$$G_k = [R_{k,i}, R_{k,j}]^\top [R_{k,i}, R_{k,j}] = \begin{bmatrix} \frac{1}{b} & R_{k,i}R_{k,j} \\ R_{k,i}R_{k,j} & \frac{1}{b} \end{bmatrix}.$$

Note the eigenvalues of G_k are 0 and $\frac{2}{b}$ and $\mathbf{E}[X^\top X] = b\cdot\mathbf{E}[G_k] = I_2$ for all $k \in [b]$. Thus, applying matrix Chernoff bound A.9 to $X^\top X$ we have

$$\Pr\Big[\lambda_{\max}(X^\top X) \leq 1-t\Big] \leq 2\exp\left(-t^2 b/2\right) \text{ for } t \in [0,1), \text{ and}$$
$$\Pr\Big[\lambda_{\max}(X^\top X) \geq 1+t\Big] \leq 2\exp\left(-t^2 b/8\right) \text{ for } t \geq 0.$$

which implies the eigenvalues of $X^\top X$ are between $[1-t, 1+t]$ with probability $1 - 4\exp\left(-\frac{t^2 b}{8}\right)$. So the eigenvalues of $X^\top X - I_2$ are between $[-t, t]$ with probability $1 - 4\exp\left(-\frac{t^2 b}{8}\right)$. Picking $t = \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}$, we have

$$\Pr\Big[\|X^\top X - I_2\| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \frac{\delta}{n^2}.$$

Note

$$X^\top X - I_2 = \begin{bmatrix} 0 & \langle R_{*,i}, R_{*,j}\rangle \\ \langle R_{*,i}, R_{*,j}\rangle & 0 \end{bmatrix},$$

whose spectral norm is $|\langle R_{*,i}, R_{*,j}\rangle|$. Thus, we have

$$\Pr\Big[|\langle R_{*,i}, R_{*,j}\rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \delta/n^2.$$

Taking a union bound over all pairs $(i,j) \in [n]\times[n]$ and $i \neq j$, we complete the proof. □

Lemma B.20 (AMS). *Let $R \in \mathbb{R}^{b\times n}$ be a random AMS matrix (Definition B.4). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\Pr\Big[\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).$$

Proof. Note for any fixed $i \neq j$, $\overline{R}_{*,i}$ and $\overline{R}_{*,j}$ are independent. By Hoeffding inequality (Lemma A.3), we have

$$\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq t\Big] \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^b (\frac{1}{b} - (-\frac{1}{b}))^2}\Big) \leq 2e^{-t^2 b/2}$$

Choosing $t = \sqrt{2\log(2n^2/\delta)}/\sqrt{b}$, we have

$$\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \sqrt{2\log(2n^2/\delta)}/\sqrt{b}\Big] \leq \frac{\delta}{n^2}.$$

Taking a union bound over all pairs $(i,j) \in [n]\times[n]$ and $i \neq j$, we complete the proof. □

Lemma B.21 (Count-Sketch). *Let $R \in \mathbb{R}^{b\times n}$ be a count-sketch matrix (Definition B.5). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \leq 1.$$

Proof. Directly follow the definition of count-sketch matrices. □

Lemma B.22 (Sparse embedding). *Let $R \in \mathbb{R}^{b\times n}$ be a sparse embedding matrix with parameter s (Definition B.6 and B.7). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\Pr\Big[\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{s}}\Big] \leq \Theta(\delta).$$

Proof. Note for fixed $i \neq j$, $\overline{R}_{*,i}$ and $\overline{R}_{*,j}$ are independent. Assume $R_{*,i}$ and $R_{*,j}$ has u non-zero elements at the same positions, where $0 \leq u \leq s$, then by Hoeffding inequality (Lemma A.3), we have

$$\Pr[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^u (\frac{1}{s} - (-\frac{1}{s}))^2}\right) \leq 2\exp(-t^2 s^2/(2u)) \tag{9}$$

Let $t = \sqrt{(2u/s^2)\log(2n^2/\delta)}$, we have

$$\begin{aligned}\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \sqrt{2s^{-1}\log(2n^2/\delta)}\Big] &\leq \Pr\Big[|\langle R_{*,i}, R_{*,j}\rangle| \geq \sqrt{2us^{-2}\log(2n^2/\delta)}\Big] \\ &\leq \delta/n^2 \end{aligned} \tag{10}$$

since $u \leq s$. By taking a union bound over all $(i,j) \in [n]\times[n]$ and $i \neq j$, we complete the proof. □

B.5. Infinite norm bound

Lemma B.23 (SRHT and AMS). *Let $R \in \mathbb{R}^{b\times n}$ be a SRHT (Definition B.3) or AMS sketching matrix (Definition B.4). Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ be two fixed vectors. Then, the following properties hold:*

$$\Pr_{R\sim\Pi}\Big[|(g^\top R^\top R h) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{b}}\|g\|_2\|h\|_2\Big] \leq \Theta(\delta).$$

Proof. We can rewrite $(g^\top R^\top Rh) - (g^\top h)$ as follows:,

$$\begin{aligned}(g^\top R^\top Rh) - (g^\top h) &= \sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle R_{*,i}, R_{*,j}\rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ &= \sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j}\rangle.\end{aligned}$$

where σ_i's are independent Rademacher random variables and $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$, and the second step follows from $\|R_{*,i}\|_2^2 = 1, \forall i \in [n]$.

We define matrix $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ as follows:

$$\begin{aligned}A_{i,j} &= g_i h_j \cdot \langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle, && \forall i \in [n], j \in [n] \\ B_{i,j} &= g_i h_j \cdot \max_{i'\neq j'} |\langle \overline{R}_{*,i'}, \overline{R}_{*,j'}\rangle| && \forall i \in [n], j \in [n]\end{aligned}$$

We define $A^\circ \in \mathbb{R}^{n\times n}$ to be the matrix $A \in \mathbb{R}^{n\times n}$ with removing diagonal entries, applying Hason-wright inequality (Lemma A.6), we have

$$\Pr_\sigma[|\sigma^\top A^\circ \sigma| \geq \tau] \leq 2 \cdot \exp(-c \min\{\tau^2/\|A^\circ\|_F^2, \tau/\|A^\circ\|\})$$

We can upper bound $\|A^\circ\|$ and $\|A^\circ\|_F$.

$$\begin{aligned}\|A^\circ\| &\leq \|A^\circ\|_F \\ &\leq \|A\|_F \\ &\leq \|B\|_F \\ &= \|g\|_2 \cdot \|h\|_2 \cdot \max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \\ &\leq \|g\|_2 \cdot \|h\|_2 \cdot \max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle|.\end{aligned}$$

where the forth step follows from B is rank-1.

For SRHT, note $\overline{R}$ has the same distribution as R. By Lemma B.19 (for AMS, we use Lemma B.20) with probability at least $1 - \Theta(\delta)$, we have :

$$\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}.$$

Conditioning on the above event holds.

Choosing $\tau = \|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(n/\delta)/\sqrt{b}$, we can show that

$$\Pr\left[\left|(g^\top R^\top Rh) - (g^\top h)\right| \geq \|g\|_2 \cdot \|h\|_2 \frac{\log^{1.5}(n/\delta)}{\sqrt{b}}\right] \leq \Theta(\delta).$$

Thus, we complete the proof. □

Lemma B.24 (Random Gaussian)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a random Gaussian matrix (Definition B.2). Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ denote two fixed vectors. Then, the following properties hold:*

$$\Pr_{R\sim\Pi}\left[|(g^\top R^\top Rh) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{b}} \|g\|_2 \|h\|_2\right] \leq \Theta(\delta).$$

Proof. We follow the same procedure as proving Lemma B.23.

We can rewrite $(g^\top R^\top Rh) - (g^\top h)$ as follows:,

$$\begin{aligned}(g^\top R^\top Rh) - (g^\top h) &= \sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle R_{*,i}, R_{*,j}\rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ &= \sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j}\rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1). \end{aligned} \tag{11}$$

where σ_i's are independent Rademacher random variables and $\overline{R}$ has the same distribution as R.

To bound the first term $\sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j}\rangle$, we define matrix $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ as follows:

$$\begin{aligned} A_{i,j} &= g_i h_j \cdot \langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle, && \forall i \in [n], j \in [n] \\ B_{i,j} &= g_i h_j \cdot \max_{i'\neq j'} |\langle \overline{R}_{*,i'}, \overline{R}_{*,j'}\rangle| && \forall i \in [n], j \in [n] \end{aligned}$$

We define $A^\circ \in \mathbb{R}^{n\times n}$ to be the matrix $A \in \mathbb{R}^{n\times n}$ with removing diagonal entries, applying Hason-wright inequality (Lemma A.6), we have

$$\Pr_\sigma[|\sigma^\top A^\circ \sigma| \geq \tau] \leq 2 \cdot \exp(-c \min\{\tau^2/\|A^\circ\|_F^2, \tau/\|A^\circ\|\})$$

We can upper bound $\|A^\circ\|$ and $\|A^\circ\|_F$.

$$\begin{aligned} \|A^\circ\| &\leq \|A^\circ\|_F \\ &\leq \|A\|_F \\ &\leq \|B\|_F \\ &= \|g\|_2 \cdot \|h\|_2 \cdot \max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \\ &\leq \|g\|_2 \cdot \|h\|_2 \cdot \max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle|. \end{aligned}$$

where the forth step follows from B is rank-1.

Using Lemma B.18 with probability at least $1 - \Theta(\delta)$, we have :

$$\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}.$$

Conditioning on the above event holds.

Choosing $\tau = \|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(n/\delta)/\sqrt{b}$, we can show that

$$\Pr\left[\left|\sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j}\rangle\right| \geq \|g\|_2 \cdot \|h\|_2 \frac{\log^{1.5}(n/\delta)}{\sqrt{b}}\right] \leq \Theta(\delta). \tag{12}$$

To bound the second term $\sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1)$, note that $b\|R_{*,i}\|_2^2 \sim \chi_b^2$ for every $i \in [n]$. Applying Lemma A.7, we have

$$\Pr\left[\left|\|R_{*,i}\|_2^2 - 1\right| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\right] \leq \delta/n.$$

which implies

$$\Pr\left[\sum_{i=1}^n g_i h_i \left|\|R_{*,i}\|_2^2 - 1\right| \geq \|g\|_2 \|h\|_2 \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\right] \leq \Theta(\delta). \tag{13}$$

Plugging the bounds Eq. (12) and (13) back to Eq. (11), we complete the proof. □

Lemma B.25 (Count-sketch)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a count-sketch matrix (Definition B.5). Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ be two fixed vectors. Then, the following properties hold:*

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \log(1/\delta)\|g\|_2\|h\|_2\right] \leq \Theta(\delta).$$

Proof. We follow the identical procedure as proving Lemma B.23 to apply Hason-wright inequality (Lemma A.6).

Then note Lemma B.21 shows

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \leq 1$$

Thus, choosing $\tau = c\|g\|_2 \cdot \|h\|_2 \cdot \log(1/\delta)$, we can show that

$$\Pr\left[|(g^\top R^\top R h) - (g^\top h)| \geq c\|g\|_2 \cdot \|h\|_2 \log(1/\delta)\right] \leq \delta.$$

which completes the proof. □

Lemma B.26 (Count-sketch 2)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a count-sketch matrix (Definition B.5). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \frac{1}{\sqrt{b\delta}}\|g\|_2\|h\|_2\right] \leq \Theta(\delta).$$

Proof. It is known that a count-sketch matrix with $b = \epsilon^{-2}\delta^{-1}$ rows satisfies the $(\epsilon, \delta, 2)$-JL moment property (Definition G.6) (see e.g. Theorem 14 of (Woodruff, 2014)). Using Markov's inequality, $(\epsilon, \delta, 2)$-JL moment property implies

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \epsilon\|g\|_2\|h\|_2\right] \leq \Theta(\delta),$$

where $\epsilon = \frac{1}{\sqrt{b\delta}}$. □

Remark B.27. *In LP solver, we need $\delta = 1/\operatorname{poly}(n)$, thus Lemma B.25 is stronger than Lemma B.26.*

Lemma B.28 (Sparse embedding)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a sparse-embedding matrix (Definition B.6 and B.7). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$3.\ \Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{s}}\|g\|_2\|h\|_2\right] \leq \Theta(\delta).$$

Proof. We follow the identical procedure as proving Lemma B.23 to apply Hason-wright inequality (Lemma A.6).

Then note Lemma B.22 shows with probability at least $1 - \delta$ we have

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \leq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{s}}.$$

Conditioning on the above event holds, choosing $\tau = c'\|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(1/\delta)$, we can show that

$$\Pr\left[|(g^\top R^\top R h) - (g^\top h)| \geq \frac{c' \log^{1.5}(n/\delta)}{\sqrt{s}}\|g\|_2 \cdot \|h\|_2\right] \leq \Theta(\delta).$$

Thus, we complete the proof. □

Lemma B.29 (Uniform sampling)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a uniform sampling matrix (Definition B.8). Let $h \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ denote two fixed vectors. Then, the following properties hold:*

$$3.|(g^\top R^\top R h) - (g^\top h)| \leq (1 + \frac{n}{b})\|g\|_2\|h\|_2$$

where $I \subset [n]$ be the subset of indexes chosen by the uniform sampling matrix.

Proof. We can rewrite $(g^\top R^\top Rh) - (g^\top h)$ as follows:

$$\begin{aligned}(g^\top R^\top Rh) - (g^\top h) = & \sum_{i=1}^n \sum_{j\in[n]\setminus i}^n g_i h_j \langle R_{*,i}, R_{*,j}\rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ = & \frac{n}{b}\sum_{i\in I} g_i h_i - \sum_{i=1}^n g_i h_i.\end{aligned}$$

where the second step follows from the uniform sampling matrix has only one nonzero entry in each row.

Let $I \subset [n]$ be the subset chosen by the uniform sampling matrix, then $\|R_{*,i}\|_2^2 = n/b$ for $i \in I$ and $\|R_{*,i}\|_2^2 = 0$ for $i \in [n] \setminus I$. So we have

$$\begin{aligned}|(g^\top R^\top Rh) - (g^\top h)| = & \Big|\sum_{i\in I} g_i h_i (\frac{n}{b} - 1) - \sum_{i\in[n]\setminus I} g_i h_i\Big| \\ \leq & (1 + \frac{n}{b})\|g\|_2 \|h\|_2.\end{aligned}$$

□

C. Sketching Central Path Method

We remark the proof of this section is similar to (Jiang et al., 2021). The major difference is they (Jiang et al., 2021) apply sketching matrix on the left of the projection matrix, and in this work we apply the sketching matrix on the right of the projection matrix. For the completeness, we still provide a proof.

Algorithm 6 Our main algorithm

procedure MAIN(A, b, c, δ) ▷ Theorem D.1
  $\epsilon \leftarrow \frac{1}{40000\log n}$, $\epsilon_{\mathrm{mp}} \leftarrow \frac{1}{40000}$, $b_{\mathrm{sketch}} \leftarrow \frac{1000\epsilon\sqrt{n}\log^2 n}{\epsilon_{\mathrm{mp}}}$.
  $\lambda \leftarrow 40\log n$, $\delta \leftarrow \min(\frac{\delta}{2}, \frac{1}{\lambda})$, $a \leftarrow \min(\alpha, 2/3)$.
  Modify the linear program and obtain an initial x and s according to (Ye et al., 1994).
  MAINTAINPROJECTION mp
  mp.INITIALIZE(A, $\frac{x}{s}$, $\epsilon_{\mathrm{mp}}, a, b_{\mathrm{sketch}}$) ▷ Algorithm 8
  $t \leftarrow 1$ ▷ Initialize t
  while $t > \delta^2/(32n^3)$ **do** ▷ We stopped once the precision is good
    $t^{\mathrm{new}} \leftarrow (1 - \frac{\epsilon}{3\sqrt{n}})t$
    $\mu \leftarrow xs$
    $\delta_\mu \leftarrow (\frac{t^{\mathrm{new}}}{t} - 1)xs - \frac{\epsilon}{2}\cdot t^{\mathrm{new}} \cdot \frac{\nabla\Phi_\lambda(\mu/t-1)}{\|\nabla\Phi_\lambda(\mu/t-1)\|_2}$ ▷ Φ_λ is defined in Lemma C.12
    $(x^{\mathrm{new}}, s^{\mathrm{new}}) \leftarrow$ STOCHASTICSTEP(mp, $x, s, \delta_\mu, b, \epsilon$) ▷ Algorithm 7
    if $\Phi_\lambda(\mu^{\mathrm{new}}/t^{\mathrm{new}} - 1) > n^3$ **then** ▷ When potential function is large
      $(x^{\mathrm{new}}, s^{\mathrm{new}}) \leftarrow$ CLASSICALSTEP(x, s, t^{new}) ▷ (Vaidya, 1989)
      mp.INITIALIZE(A, $\frac{x^{\mathrm{new}}}{s^{\mathrm{new}}}$, $\epsilon_{\mathrm{mp}}, a$) ▷ Restart the data structure
    end if
    $(x, s) \leftarrow (x^{\mathrm{new}}, s^{\mathrm{new}})$, $t \leftarrow t^{\mathrm{new}}$
  end while
  return an approximate solution of the original linear program according to (Ye et al., 1994) .
end procedure

To decouple the proof in both parts, we will make the following assumption in first part. It will be verified in the second part.

For simplicity, we assume the sketching matrix $R \in \mathbb{R}^{b\times n}$ is of (α, β, δ)-coordinate wise embedding with $\alpha = 1$, $\beta = \log^{1.5}(n/\delta)$, which corresponds to the case of random Gaussian, SRHT, AMS matrices. For other random matrices we discuss in the paper, the results extends directly.

Assumption C.1. *Assume the following for the input of the procedure* STOCHASTICSTEP *(see Algorithm 7):*

Algorithm 7

procedure StochasticStep($\mathrm{mp}, x, s, \delta_\mu, b, \epsilon$) ▷ Lemma C.2,C.3,C.8
  $w \leftarrow \frac{x}{s}, \widetilde{v} \leftarrow \mathrm{mp}.\textsc{Update}(w)$ ▷ Algorithm 8
  $\overline{x} \leftarrow x\sqrt{\frac{\widetilde{v}}{w}}, \overline{s} \leftarrow s\sqrt{\frac{w}{\widetilde{v}}}$ ▷ It guarantees that $\frac{\overline{x}}{\overline{s}} = \widetilde{v}$ and $\overline{xs} = xs$
  repeat
    $p_x, p_s \leftarrow \mathrm{mp}.\textsc{Query}(\frac{1}{\sqrt{\overline{XS}}}\delta_\mu)$ ▷ Algorithm 8
    $\widetilde{\delta}_s \leftarrow \frac{\overline{S}}{\sqrt{\overline{XS}}}p_s$ ▷ According to Eq. (16)
    $\widetilde{\delta}_x \leftarrow \frac{\overline{X}}{\sqrt{\overline{XS}}}p_x$ ▷ According to Eq. (15)
  until $\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty \leq \frac{1}{100\log n}$ and $\|\overline{x}^{-1}\widetilde{\delta}_x\|_\infty \leq \frac{1}{100\log n}$
  return $(x + \widetilde{\delta}_x, s + \widetilde{\delta}_s)$
end procedure

- $xs \approx_{0.1} t$ *with* $t > 0$.
- $\mathrm{mp}.\textsc{Update}(w)$ *outputs* $\widetilde{v}$ *such that* $w \approx_{\epsilon_{\mathrm{mp}}} \widetilde{v}$ *with* $\epsilon_{\mathrm{mp}} \leq 1/40000$.
- $\|\delta_\mu\|_2 \leq \epsilon t$ *with* $0 < \epsilon < 1/(40000\log n)$.
- $b \geq 1000\epsilon\sqrt{n}\log^2 n/\epsilon_{\mathrm{mp}}$.

C.1. Bounding each quantity of stochastic step

First, we give an explicit formula for our step, which will be used in all subsequent calculations.

Specifically, we show our update can be viewed as an **exact** solution of the following linear system:

$$\begin{aligned} \overline{X}\widetilde{\delta}_s + \overline{S}\widetilde{\delta}_x &= \widetilde{\delta}_\mu, \\ A\widetilde{\delta}_x &= 0, \\ A^\top\widetilde{\delta}_y + \widetilde{\delta}_s &= 0, \end{aligned} \tag{14}$$

where

$$\widetilde{\delta}_\mu = \sqrt{XS}R^\top R\frac{1}{\sqrt{XS}}\delta_\mu.$$

Lemma C.2. *The procedure* StochasticStep($\mathrm{mp}, x, s, \delta_\mu, b, \epsilon$) *(see Algorithm 7) finds a solution* $\widetilde{\delta}_x, \widetilde{\delta}_s \in \mathbb{R}^n$, $\widetilde{\delta}_y \in \mathbb{R}^d$ *to Eq.* (14) *by the formula*

$$\widetilde{\delta}_x = \frac{\overline{X}}{\sqrt{\overline{XS}}}(I - \overline{P})R^\top R\frac{1}{\sqrt{\overline{XS}}}\delta_\mu \tag{15}$$

$$\widetilde{\delta}_s = \frac{\overline{S}}{\sqrt{\overline{XS}}}\overline{P}R^\top R\frac{1}{\sqrt{\overline{XS}}}\delta_\mu \tag{16}$$

$$\widetilde{\delta}_y = -(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\sqrt{\frac{\overline{X}}{\overline{S}}}R^\top R\frac{1}{\sqrt{\overline{XS}}}\delta_\mu \tag{17}$$

and

$$\overline{P} = \sqrt{\frac{\overline{X}}{\overline{S}}}A^\top\left(A\frac{\overline{X}}{\overline{S}}A^\top\right)^{-1}A\sqrt{\frac{\overline{X}}{\overline{S}}}. \tag{18}$$

Proof. For the first equation of Eq. (14), we multiply $A\overline{S}^{-1}$ on both sides,

$$A\overline{S}^{-1}\overline{X}\widetilde{\delta}_s + A\widetilde{\delta}_x = A\overline{S}^{-1}\widetilde{\delta}_\mu.$$

Since the second equation gives $A\widetilde{\delta}_x = 0$, then we know that $A\overline{S}^{-1}\overline{X}\widetilde{\delta}_s = A\overline{S}^{-1}\widetilde{\delta}_\mu$.

Multiplying $A\overline{S}^{-1}\overline{X}$ on both sides of the third equation of Eq. (14), we have

$$-A\overline{S}^{-1}\overline{X}A^\top\widetilde{\delta}_y = A\overline{S}^{-1}\overline{X}\widetilde{\delta}_s = A\overline{S}^{-1}\widetilde{\delta}_\mu.$$

Thus,

$$\begin{aligned}
\widetilde{\delta}_y =& -(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\overline{S}^{-1}\widetilde{\delta}_\mu, \\
\widetilde{\delta}_s =& A^\top(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\overline{S}^{-1}\widetilde{\delta}_\mu, \\
\widetilde{\delta}_x =& \overline{S}^{-1}\widetilde{\delta}_\mu - \overline{S}^{-1}\overline{X}A^\top(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\overline{S}^{-1}\widetilde{\delta}_\mu.
\end{aligned}$$

Recall we define $\overline{P}$ as Eq. (18) and $\widetilde{\delta}_\mu$ as Eq. (6), then we have

$$\widetilde{\delta}_s = \frac{\overline{S}}{\sqrt{\overline{X}\overline{S}}}\cdot\sqrt{\frac{\overline{X}}{\overline{S}}}A^\top(A\frac{\overline{X}}{\overline{S}}A^\top)^{-1}A\sqrt{\frac{\overline{X}}{\overline{S}}}\cdot\frac{1}{\sqrt{\overline{X}\overline{S}}}\widetilde{\delta}_\mu = \frac{\overline{S}}{\sqrt{\overline{X}\overline{S}}}\overline{P}\frac{1}{\sqrt{\overline{X}\overline{S}}}\widetilde{\delta}_\mu = \frac{\overline{S}}{\sqrt{\overline{X}\overline{S}}}\overline{P}R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu,$$

and

$$\begin{aligned}
\widetilde{\delta}_x =& \overline{S}^{-1}\widetilde{\delta}_\mu - \frac{\overline{X}}{\sqrt{\overline{X}\overline{S}}}\cdot\sqrt{\frac{\overline{X}}{\overline{S}}}A^\top(A\frac{\overline{X}}{\overline{S}}A^\top)^{-1}\sqrt{\frac{\overline{X}}{\overline{S}}}\cdot\frac{1}{\sqrt{\overline{X}\overline{S}}}\widetilde{\delta}_\mu \\
=& \frac{\overline{X}}{\sqrt{\overline{X}\overline{S}}}(I-\overline{P})\frac{1}{\sqrt{\overline{X}\overline{S}}}\widetilde{\delta}_\mu = \frac{\overline{X}}{\sqrt{\overline{X}\overline{S}}}(I-\overline{P})R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu,
\end{aligned}$$

and

$$\widetilde{\delta}_y = -(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\sqrt{\frac{\overline{X}}{\overline{S}}}\frac{1}{\sqrt{\overline{X}\overline{S}}}\widetilde{\delta}_\mu = -(A\overline{S}^{-1}\overline{X}A^\top)^{-1}A\sqrt{\frac{\overline{X}}{\overline{S}}}R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu,$$

which match Eq. (16), Eq. (15) and Eq. (17).

To see why the StochasticStep outputs $\widetilde{\delta}_x$, $\widetilde{\delta}_s$ satisfying Eq. (16) and Eq. (15), we note that

$$\begin{aligned}
p_x =& (I - \sqrt{\widetilde{V}}A^\top\left(A\frac{\overline{X}}{\overline{S}}A^\top\right)^{-1}A\sqrt{\widetilde{V}})R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu = (I-\overline{P})R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu \\
p_s =& \sqrt{\widetilde{V}}A^\top\left(A\frac{\overline{X}}{\overline{S}}A^\top\right)^{-1}A\sqrt{\widetilde{V}}R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu = \overline{P}R^\top R\frac{1}{\sqrt{\overline{X}\overline{S}}}\delta_\mu
\end{aligned}$$

because of Theorem E.1. □

Using the explicitly formula, we are ready to bound all quantities we needed in the following two subsubsections.

C.2. Bounding $\widetilde{\delta}_s$, $\widetilde{\delta}_x$ and $\widetilde{\delta}_\mu$

Lemma C.3. *Under the Assumption C.1, the two vectors $\widetilde{\delta}_x$ and $\widetilde{\delta}_s$ found by* StochasticStep *satisfy :*
1. $\|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\|_2 \leq 2\epsilon, \|\mathbf{E}[\overline{x}^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon, \|\mathbf{E}[s^{-1}\widetilde{\delta}_s]\|_2 \leq 2\epsilon, \|\mathbf{E}[x^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon, \|\mathbf{E}[\mu^{-1}\widetilde{\delta}_\mu]\|_2 \leq 4\epsilon.$
2. $\mathbf{Var}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b, \mathbf{Var}[\overline{x}_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b, \mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b, \mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b, \mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{\mu,i}] \leq 8\epsilon^2/b.$
3. $\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty \leq \epsilon, \|s^{-1}\widetilde{\delta}_s\|_\infty \leq 2\epsilon, \|\overline{x}^{-1}\widetilde{\delta}_x\|_\infty \leq \epsilon, \|x^{-1}\widetilde{\delta}_x\|_\infty \leq 2\epsilon, \|\mu^{-1}\widetilde{\delta}_\mu\|_\infty \leq 2\epsilon.$

Remark C.4. *For notational simplicity, the* $\mathbf{E}$ *and* $\mathbf{Var}$ *in the proof are for the case without resketching (Line 8). Since the all the additional terms due to resketching are polynomially bounded and since we can set failure probability to an arbitrarily small inverse polynomial (see Claim C.7), the proof does not change and the result remains the same.*

Claim C.5 (Part 1, bounding the ℓ_2 norm of expectation)**.**

$$\|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\|_2 \leq 2\epsilon, \|\mathbf{E}[\overline{x}^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon, \|\mathbf{E}[s^{-1}\widetilde{\delta}_s]\|_2 \leq 2\epsilon, \|\mathbf{E}[x^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon, \|\mathbf{E}[\mu^{-1}\widetilde{\delta}_\mu]\|_2 \leq 4\epsilon.$$

Proof. For $\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty$, we consider the i-th coordinate of the vector

$$\overline{s}_i^{-1}\widetilde{\delta}_{s,i} = \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\sum_{j=1}^n (\overline{P}R^\top R)_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}}.$$

Then, we have

$$\mathbf{E}\left[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}\right] = \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\sum_{j=1}^n (\mathbf{E}[\overline{P}R^\top R])_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}} = \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\sum_{j=1}^n \overline{P}_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}}, \tag{19}$$

where the second step follows by $\mathbf{E}[R^\top R] = I$. Since $xs \approx_{0.1} t$ and $\|\delta_\mu\|_2 \leq \epsilon t$, we have $\|\frac{\delta_\mu}{\sqrt{xs}}\|_2 \leq \frac{1.1\epsilon t}{\sqrt{t}}$. Since $\overline{P}$ is an orthogonal projection matrix, we have $\|\overline{P}\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2 \leq \|\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2$. Putting all above facts together, we can show

$$\begin{aligned}
\left\|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\right\|_2^2 &= \sum_{i=1}^n \left(\frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\sum_{j=1}^n \overline{P}_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}}\right)^2 \\
&= \sum_{i=1}^n \frac{1}{\overline{x}_i\overline{s}_i}\left(\sum_{j=1}^n \overline{P}_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}}\right)^2 \\
&\leq \frac{1}{0.9t}\sum_{i=1}^n \left(\sum_{j=1}^n \overline{P}_{i,j}\frac{\delta_{\mu,j}}{\sqrt{\overline{x}_j\overline{s}_j}}\right)^2 \\
&= \frac{1}{0.9t}\|\overline{P}\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2^2 \\
&\leq \frac{1}{0.9t}\|\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2^2 \\
&\leq \frac{(1.1)^2}{0.9t}\cdot\frac{(\epsilon t)^2}{t} \\
&\leq 1.4\epsilon^2,
\end{aligned}$$

where the third step follows by $\overline{xs} = xs \approx_{0.1} t$. It implies that

$$\left\|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\right\|_2 \leq 1.2\epsilon. \tag{20}$$

Notice that the proof for x is identical to the proof for s because $(I - \overline{P})$ is also a projection matrix. Further, since $\overline{s} \approx_{0.1} s$ and $\overline{x} \approx_{0.1} x$ by Assumption C.1, the next two inequalities in Claim C.5 can be easily shown.

Now, we are ready to bound $\|\mathbf{E}[\mu^{-1}\widetilde{\delta}_\mu]\|_2$ by

$$\|\mathbf{E}[\mu^{-1}\widetilde{\delta}_\mu]\|_2 = \|\mathbf{E}[\overline{s}^{-1}\overline{x}^{-1}(\overline{x}\widetilde{\delta}_s + \overline{s}\widetilde{\delta}_x)]\|_2 \leq \|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\|_2 + \|\mathbf{E}[\overline{x}^{-1}\widetilde{\delta}_x]\|_2 \leq 4\epsilon.$$

where the first step follows by $\mu = xs = \overline{xs}$ and $\overline{x}\widetilde{\delta}_s + \overline{s}\widetilde{\delta}_x = \widetilde{\delta}_\mu$ defined in Eq. (14), the second step follows by triangle inequality, and last step follows by $\|\mathbf{E}[\overline{s}^{-1}\widetilde{\delta}_s]\|_2, \|\mathbf{E}[\overline{x}^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon$. □

Claim C.6 (Part 2, bounding the variance per coordinate)**.**

$$\mathbf{Var}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b, \mathbf{Var}[\overline{x}_i^{-1}\widetilde{\delta}_{x,i}] \leq \epsilon^2/b, \mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b, \mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b, \mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{\mu,i}] \leq 8\epsilon^2/b.$$

Proof. Consider the i-th coordinate of the vector

$$\overline{s}_i^{-1}\widetilde{\delta}_{s,i} = \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\Big(\overline{P}R^\top R\frac{\delta_\mu}{\sqrt{\overline{xs}}}\Big)_i.$$

For variance of $\overline{s}_i^{-1}\widetilde{\delta}_{s,i}$, we have

$$\begin{aligned}
\mathbf{Var}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] \leq\ & \mathbf{E}[(\overline{s}_i^{-1}\widetilde{\delta}_{s,i})^2] - \mathbf{E}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}]^2 \\
\leq\ & \frac{1}{\overline{x}_i\overline{s}_i}\Big(\overline{P}\frac{\delta_\mu}{\sqrt{\overline{xs}}}\Big)_i^2 + \frac{1}{\overline{x}_i\overline{s}_i}\frac{\|\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2^2}{b} - \frac{1}{\overline{x}_i\overline{s}_i}\Big(\overline{P}\frac{\delta_\mu}{\sqrt{\overline{xs}}}\Big)_i^2 \\
\leq\ & \frac{1}{\overline{x}_i\overline{s}_i}\frac{\|\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2^2}{b} \\
\leq\ & \frac{1.3}{t^2}\frac{\|\delta_\mu\|_2^2}{b} \\
\leq\ & \frac{1.3\epsilon^2}{b}, && \text{by } \overline{x}_i\overline{s}_i = x_is_i \approx_{1/10} t
\end{aligned}$$

where the second step follows by $(1, \log^{1.5}(n/\delta), \delta)$-coordinate wise embedding and Eq. (19), the fourth step follows by $\overline{x}_i\overline{s}_i = x_is_i \approx_{0.1} t$, and the last step follows by $\|\delta_\mu\|_2 \leq \epsilon t$ in Assumption C.1.

The proof for the next three inequalities in Claim C.6 are identical, which are omit here.

For the variance of $\mu_i^{-1}\widetilde{\delta}_{\mu,i}$,

$$\begin{aligned}
\mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{\mu,i}] =\ & \mathbf{Var}[\overline{x}_i^{-1}\overline{s}_i^{-1}(\overline{x}_i\widetilde{\delta}_{s,i} + \overline{s}_i\widetilde{\delta}_{x,i})] \\
\leq\ & 2\,\mathbf{Var}[\overline{x}_i^{-1}\overline{x}_i\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] + 2\,\mathbf{Var}[\overline{s}_i^{-1}\overline{s}_i\overline{x}_i^{-1}\widetilde{\delta}_{x,i}] \\
=\ & 2\,\mathbf{Var}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] + 2\,\mathbf{Var}[\overline{x}_i^{-1}\widetilde{\delta}_{x,i}] \\
\leq\ & 8\epsilon^2/b.
\end{aligned}$$

where the first step follows by $\mu = xs = \overline{xs}$ and $\overline{x}\widetilde{\delta}_s + \overline{s}\widetilde{\delta}_x = \widetilde{\delta}_\mu$ defined in Eq. (14), the second step follows by triangle inequality, and the last step follows by $\mathbf{Var}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}], \mathbf{Var}[\overline{x}_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b$. □

Claim C.7 (Part 3, bounding the probability of success)**.** *Let $b \geq 10\log^3(2n^2/\delta)$. Without resketching, the following holds with probability $1-\delta$.*

$$\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty \leq \epsilon, \|s^{-1}\widetilde{\delta}_s\|_\infty \leq 2\epsilon, \|\overline{x}^{-1}\widetilde{\delta}_x\|_\infty \leq \epsilon, \|x^{-1}\widetilde{\delta}_x\|_\infty \leq 2\epsilon, \|\mu^{-1}\widetilde{\delta}_\mu\|_\infty \leq 2\epsilon.$$

With resketching, it always holds.

Proof. Note by Eq. (19), we have

$$\overline{s}_i^{-1}\widetilde{\delta}_{s,i} - \mathbf{E}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}] = \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}((\overline{P}R^\top R - \overline{P})\frac{\delta_\mu}{\sqrt{\overline{xs}}})_i$$

Using ℓ_∞ bound in $(1, \log^{1.5}(n/\delta), \delta)$-coordinate wise embedding property, we have

$$\Pr\left[|\overline{s}_i^{-1}\widetilde{\delta}_{s,i} - \mathbf{E}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}]| \geq \frac{1}{\sqrt{\overline{x}_i\overline{s}_i}}\frac{\log^{1.5}(n/\delta)}{\sqrt{b}} \cdot \|\frac{\delta_\mu}{\sqrt{\overline{xs}}}\|_2\right] \leq \delta$$

As long as

$$b \geq 10 \log^3(2n^2/\delta)$$

we have with probability $1 - \delta/2n$,

$$|\overline{s}_i^{-1}\widetilde{\delta}_{s,i} - \mathbf{E}[\overline{s}_i^{-1}\widetilde{\delta}_{s,i}]| \leq \frac{1}{0.9\sqrt{t}} \cdot 0.5 \cdot \frac{\epsilon t}{0.9\sqrt{t}} \leq \epsilon.$$

Taking a union bound, we can show

$$\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty \leq \epsilon.$$

The next three inequalities can be shown in identical way. To show the last inequality, we have

$$|\mu_i^{-1}\widetilde{\delta}_{\mu,i}| = |\overline{x}_i^{-1}\overline{s}_i^{-1}(\overline{x}_i\widetilde{\delta}_{s,i} + \overline{s}_i\widetilde{\delta}_{x,i})| = |\overline{s}_i^{-1}\widetilde{\delta}_{s,i}| + |\overline{x}_i^{-1}\widetilde{\delta}_{x,i}| \leq 2\epsilon,$$

where the first step follows by $\mu = xs = \overline{x}\overline{s}$ and $\overline{x}\widetilde{\delta}_s + \overline{s}\widetilde{\delta}_x = \widetilde{\delta}_\mu$ defined in Eq. (14), the second step follows by triangle inequality, and the last step follows by $\|\overline{s}^{-1}\widetilde{\delta}_s\|_\infty,\ \|\overline{x}^{-1}\widetilde{\delta}_x\|_\infty \leq \epsilon$.

□

C.3. Bounding $\mu^{\mathrm{new}} - \mu$

Lemma C.8. *Let $\epsilon \leq \epsilon_{\mathrm{mp}}$ and $b \geq \sqrt{n}$. Under the Assumption C.1, the vector $\mu_i^{\mathrm{new}} \stackrel{\mathrm{def}}{=} (x_i + \widetilde{\delta}_{x,i})(s_i + \widetilde{\delta}_{s,i})$ satisfies*
1. $\|\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu)]\|_2 \leq 10\epsilon_{\mathrm{mp}} \cdot \epsilon$ and $\|\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}} - \mu)]\|_2 \leq 5\epsilon$.
2. $\mathbf{Var}[\mu_i^{-1}\mu_i^{\mathrm{new}}] \leq 50\epsilon^2/b$ for all i.
3. $\|\mu^{-1}(\mu^{\mathrm{new}} - \mu)\|_\infty \leq 3\epsilon$.

Claim C.9 (Part 1 of Lemma C.8)**.**

$$\|\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu)]\|_2 \leq 10\epsilon_{\mathrm{mp}} \cdot \epsilon,\ \textit{and}\ \|\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}} - \mu)]\|_2 \leq 5\epsilon.$$

Proof. We write

$$\mu^{\mathrm{new}} = (x + \widetilde{\delta}_x)(s + \widetilde{\delta}_s) = \mu + x\widetilde{\delta}_s + s\widetilde{\delta}_x + \widetilde{\delta}_x\widetilde{\delta}_s = \mu + \underbrace{\overline{x}\widetilde{\delta}_s + \overline{s}\widetilde{\delta}_x}_{\widetilde{\delta}_\mu} + \underbrace{(x - \overline{x})\widetilde{\delta}_s + (s - \overline{s})\widetilde{\delta}_x + \widetilde{\delta}_x\widetilde{\delta}_s}_{\epsilon_\mu}.$$

Taking the expectation on both sides, we have

$$\mathbf{E}[\mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu] = (x - \overline{x})\,\mathbf{E}[\widetilde{\delta}_s] + (s - \overline{s})\,\mathbf{E}[\widetilde{\delta}_x] + \mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s].$$

Hence, we have

$$\begin{aligned}
&\|\mu^{-1}\,\mathbf{E}[\mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu]\|_2 \\
\leq\ &\|\mu^{-1}(x - \overline{x})s \cdot s^{-1}\,\mathbf{E}[\widetilde{\delta}_s]\|_2 + \|\mu^{-1}(s - \overline{s})x \cdot x^{-1}\,\mathbf{E}[\widetilde{\delta}_x]\|_2 + \|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2 \\
\leq\ &\|\mu^{-1}(x - \overline{x})s\|_\infty \cdot \|s^{-1}\,\mathbf{E}[\widetilde{\delta}_s]\|_2 + \|\mu^{-1}(s - \overline{s})x\|_\infty \cdot \|x^{-1}\,\mathbf{E}[\widetilde{\delta}_x]\|_2 + \|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2 \\
\leq\ &\epsilon_{\mathrm{mp}} \cdot \|s^{-1}\,\mathbf{E}[\widetilde{\delta}_s]\|_2 + \epsilon_{\mathrm{mp}} \cdot \|x^{-1}\,\mathbf{E}[\widetilde{\delta}_x]\|_2 + \|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2 \\
\leq\ &4\epsilon_{\mathrm{mp}} \cdot \epsilon + \|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2, \qquad (21)
\end{aligned}$$

where the first step follows by triangle inequality, the second step follows by $\|ab\|_2 \leq \|a\|_\infty \cdot \|b\|_2$, the third step follows by $\|\mu^{-1}(x - \overline{x})s\|_\infty \leq \epsilon_{\mathrm{mp}}$ and $\|\mu^{-1}(s - \overline{s})x\|_\infty \leq \epsilon_{\mathrm{mp}}$ (since $\overline{x} \approx_{\epsilon_{\mathrm{mp}}} x$, $\overline{s} \approx_{\epsilon_{\mathrm{mp}}} s$), the last step follows by $\|\mathbf{E}[s^{-1}\widetilde{\delta}_s]\|_2 \leq 2\epsilon$ and $\|\mathbf{E}[x^{-1}\widetilde{\delta}_x]\|_2 \leq 2\epsilon$ (Part 1 of Lemma C.3).

To bound the last term, note $\mathbf{E}[\widetilde{\delta}_s] = \delta_s$ and $\mathbf{E}[\widetilde{\delta}_x] = \delta_x$, so we have

$$\mathbf{E}[\widetilde{\delta}_{x,i}\widetilde{\delta}_{s,i}] = \delta_{x,i}\delta_{s,i} + \mathbf{E}[(\widetilde{\delta}_{x,i} - \delta_{x,i})(\widetilde{\delta}_{s,i} - \delta_{s,i})].$$

Hence,

$$\begin{aligned}
\|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2 &\leq \|\mu^{-1}\delta_x\delta_s\|_2 + \left(\sum_{i=1}^n \left(\mathbf{E}\left[x_i^{-1}(\widetilde{\delta}_{x,i}-\delta_{x,i})\cdot s_i^{-1}(\widetilde{\delta}_{s,i}-\delta_{s,i})\right]\right)^2\right)^{1/2} \\
&\leq 4\epsilon^2 + \frac{1}{2}\left(\sum_{i=1}^n \left(\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] + \mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}]\right)^2\right)^{1/2} \\
&\leq 4\epsilon^2 + \frac{1}{2}\left(\sum_{i=1}^n 2(\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}])^2 + 2(\mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}])^2\right)^{1/2} \\
&\leq 4\epsilon^2 + 2\sqrt{n\cdot\epsilon^4/b^2} \\
&\leq 4\epsilon^2 + 2\epsilon^2 \\
&\leq 6\epsilon^2, \qquad (22)
\end{aligned}$$

where the first step follows by triangle inequality, the second step follows by $\|\mu^{-1}\delta_x\delta_s\|_2 \leq \|x^{-1}\delta_x\|_2 \cdot \|s^{-1}\delta_s\|_2 \leq 4\epsilon^2$ (Part 1 of Lemma C.3) and $2ab \leq a^2+b^2$, the third step follows by $(a+b)^2 \leq 2a^2+2b^2$, the fourth step follows by $\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b$ and $\mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b$ (Part 2 of Lemma C.3), the last step follows by $b \geq \sqrt{n}$.

Combining Eq. (21) and Eq. (22), we have that

$$\begin{aligned}
\|\mu^{-1}(\mathbf{E}[\mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu])\|_2 &\leq 4\epsilon_{\mathrm{mp}}\cdot\epsilon + \|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_x\widetilde{\delta}_s]\|_2 \\
&\leq 4\epsilon_{\mathrm{mp}}\cdot\epsilon + 6\epsilon^2 \\
&\leq 10\epsilon_{\mathrm{mp}}\cdot\epsilon.
\end{aligned}$$

where we used $\epsilon \leq \epsilon_{\mathrm{mp}}$ in Assumption C.1.

From Part 1 of Lemma C.3, we know that $\|\mu^{-1}\,\mathbf{E}[\widetilde{\delta}_\mu]\|_2 \leq 4\epsilon$. Thus using triangle inequality, we know

$$\|\mu^{-1}(\mathbf{E}[\mu^{\mathrm{new}} - \mu])\|_2 \leq 10\epsilon_{\mathrm{mp}}\cdot\epsilon + 4\epsilon \leq 5\epsilon.$$

□

Claim C.10 (Part 2 of Lemma C.8)**.** $\mathbf{Var}[\mu_i^{-1}\mu_i^{\mathrm{new}}] \leq 50\epsilon^2/b$ *for all* i.

Proof. Recall that

$$\mu^{\mathrm{new}} = \mu + \widetilde{\delta}_\mu + (x-\overline{x})\widetilde{\delta}_s + (s-\overline{s})\widetilde{\delta}_x + \widetilde{\delta}_x\widetilde{\delta}_s.$$

We can upper bound the variance of $\mu_i^{-1}\mu_i^{\mathrm{new}}$ by,

$$\begin{aligned}
\mathbf{Var}[\mu_i^{-1}\mu_i^{\mathrm{new}}] &\leq 4\,\mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{\mu,i}] + 4\,\mathbf{Var}[\mu_i^{-1}(x_i-\overline{x}_i)\widetilde{\delta}_{s,i}] + 4\,\mathbf{Var}[\mu_i^{-1}(s_i-\overline{s}_i)\widetilde{\delta}_{x,i}] + 4\,\mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{x,i}\widetilde{\delta}_{s,i}] \\
&\leq 32\frac{\epsilon^2}{b} + 4\frac{\epsilon^2}{b} + 4\frac{\epsilon^2}{b} + \mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{x,i}\widetilde{\delta}_{s,i}] \\
&= 40\frac{\epsilon^2}{b} + \mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}\cdot s_i^{-1}\widetilde{\delta}_{s,i}] \\
&\leq 40\frac{\epsilon^2}{b} + 2\,\mathbf{Sup}[(x_i^{-1}\widetilde{\delta}_{x,i})^2]\cdot\mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] + 2\,\mathbf{Sup}[(y_i^{-1}\widetilde{\delta}_{y,i})^2]\cdot\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] \\
&\leq 40\frac{\epsilon^2}{b} + 2\cdot(2\epsilon)^2\cdot\frac{\epsilon^2}{b} + 2\cdot(2\epsilon)^2\cdot\frac{\epsilon^2}{b} \\
&\leq 50\frac{\epsilon^2}{b}.
\end{aligned}$$

where the first step follows by Cauchy-Schwartz inequality, the second step follows by $\mathbf{Var}[\mu_i^{-1}\widetilde{\delta}_{\mu,i}] \leq 8\epsilon^2/b$ (Part 2 of Lemma C.3) and

$$\mathbf{Var}[\mu_i^{-1}(x_i - \overline{x}_i)\widetilde{\delta}_{s,i}] = \mathbf{Var}[x_i^{-1}(x_i - \overline{x}_i)s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon_{\mathrm{mp}}^2 \mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq \epsilon^2/b.$$

and a similar inequality for $\mathbf{Var}[\mu_i^{-1}(s_i - \overline{s}_i)\widetilde{\delta}_{x,i}] \leq \epsilon^2/b$, the third step follows by $\mu = xs$, the fourth step follows by $\mathbf{Var}[xy] \leq 2\,\mathbf{Sup}[x^2]\,\mathbf{Var}[y] + 2\,\mathbf{Sup}[y^2]\,\mathbf{Var}[x]$ (Lemma A.1) with $\mathbf{Sup}$ denoting the deterministic maximum of the random variable, the fifth step follows by $\mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b$ and $\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] \leq 2\epsilon^2/b$ (Part 2 of Lemma C.3). □

Claim C.11 (Part 3 of Lemma C.8)**.** $\|\mu^{-1}(\mu^{\mathrm{new}} - \mu)\|_\infty \leq 3\epsilon$.

Proof. We again note that

$$\mu^{\mathrm{new}} = \mu + \widetilde{\delta}_\mu + (x - \overline{x})\widetilde{\delta}_s + (s - \overline{s})\widetilde{\delta}_x + \widetilde{\delta}_x\widetilde{\delta}_s.$$

Hence, we have

$$\begin{aligned}
&|\mu_i^{-1}(\mu_i^{\mathrm{new}} - \mu_i - \widetilde{\delta}_{\mu,i})| \\
\leq\ &|(x - \overline{x})_i\mu_i^{-1}\widetilde{\delta}_{s,i}| + |(s - \overline{s})_i\mu_i^{-1}\widetilde{\delta}_{x,i}| + |\mu_i^{-1}\widetilde{\delta}_{x,i}\widetilde{\delta}_{s,i}| \\
=\ &|(x - \overline{x})_i x_i^{-1}| \cdot |s_i^{-1}\widetilde{\delta}_{s,i}| + |(s - \overline{s})_i s_i^{-1}| \cdot |x_i^{-1}\widetilde{\delta}_{x,i}| + |x_i^{-1}\widetilde{\delta}_{x,i}| \cdot |s_i^{-1}\widetilde{\delta}_{s,i}| \\
\leq\ &\epsilon_{\mathrm{mp}}|s_i^{-1}\widetilde{\delta}_{s,i}| + \epsilon_{\mathrm{mp}}|x_i^{-1}\widetilde{\delta}_{x,i}| + |s_i^{-1}\widetilde{\delta}_{s,i}||x_i^{-1}\widetilde{\delta}_{x,i}| \\
\leq\ &\epsilon_{\mathrm{mp}} \cdot (2\epsilon) + \epsilon_{\mathrm{mp}} \cdot (2\epsilon) + (2\epsilon)^2 \\
\leq\ &\epsilon,
\end{aligned}$$

where the first step follows by triangle inequality, the second step follows by $\mu_i = x_i s_i$, the third step follows by $x \approx_{\epsilon_{\mathrm{mp}}} \overline{x}$ and $s \approx_{\epsilon_{\mathrm{mp}}} \overline{s}$, the fifth step follows by $|s_i^{-1}\widetilde{\delta}_{s,i}| \leq 2\epsilon$ and $|x_i^{-1}\widetilde{\delta}_{x,i}| \leq 2\epsilon$ (Part 3 of Lemma C.3).

Since we know that $|\mu_i^{-1}\widetilde{\delta}_{\mu,i}| \leq 2\epsilon$ (Part 3 of Lemma C.3), we have

$$|\mu_i^{-1}(\mu_i^{\mathrm{new}} - \mu_i)| \leq \epsilon + 2\epsilon \leq 3\epsilon.$$

□

C.4. Stochastic central path

We state a tool from previous work ((Cohen et al., 2019b)). It gives us several basic properties of potential function Φ_λ.

Lemma C.12 (Basic properties of potential function, (Cohen et al., 2019b))**.** *Let $\Phi_\lambda(r) = \sum_{i=1}^n \cosh(\lambda r_i)$ for some $\lambda > 0$. For any vector $r \in \mathbb{R}^n$,*
1. For any vector $\|v\|_\infty \leq 1/\lambda$, we have that

$$\Phi_\lambda(r + v) \leq \Phi_\lambda(r) + \langle \nabla\Phi_\lambda(r), v\rangle + 2\|v\|^2_{\nabla^2\Phi_\lambda(r)}.$$

2. $\|\nabla\Phi_\lambda(r)\|_2 \geq \frac{\lambda}{\sqrt{n}}(\Phi_\lambda(r) - n)$.
3. $\left(\sum_{i=1}^n \lambda^2 \cosh^2(\lambda r_i)\right)^{1/2} \leq \lambda\sqrt{n} + \|\nabla\Phi_\lambda(r)\|_2$.

The following lemma shows that the potential Φ is decreasing in expectation when Φ is large.

Lemma C.13. *Let $n \geq b \geq \sqrt{n}$ and $\lambda\epsilon \leq 1/1000$. Under the Assumption C.1, we have*

$$\mathbf{E}\left[\Phi_\lambda\left(\frac{\mu^{\mathrm{new}}}{t^{\mathrm{new}}} - 1\right)\right] \leq \Phi_\lambda\left(\frac{\mu}{t} - 1\right) - \frac{\lambda\epsilon}{15\sqrt{n}}\left(\Phi_\lambda\left(\frac{\mu}{t} - 1\right) - 10n\right).$$

Proof. Let $\epsilon_\mu = \mu^{\mathrm{new}} - \mu - \widetilde{\delta}_\mu$. From the definition, we have

$$\mu^{\mathrm{new}} - t^{\mathrm{new}} = \mu + \widetilde{\delta}_\mu + \epsilon_\mu - t^{\mathrm{new}},$$

which implies

$$\begin{aligned}\frac{\mu^{\mathrm{new}}}{t^{\mathrm{new}}} - 1 &= \frac{\mu}{t^{\mathrm{new}}} + \frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu) - 1 \\ &= \frac{\mu}{t}\frac{t}{t^{\mathrm{new}}} + \frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu) - 1 \\ &= \frac{\mu}{t} + \frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu) - 1 \\ &= \frac{\mu}{t} - 1 + \underbrace{\frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu)}_{v}. \end{aligned} \tag{23}$$

Let $v = \frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu)$, we have

$$\begin{aligned}\mathbf{E}[v] &= \frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}(\mathbf{E}[\widetilde{\delta}_\mu] + \mathbf{E}[\epsilon_\mu]) \\ &= \frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}(\delta_\mu + \mathbf{E}[\epsilon_\mu]) \\ &= \frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1) + \frac{1}{t^{\mathrm{new}}}\left(\left((\frac{t^{\mathrm{new}}}{t} - 1)\mu - \frac{\epsilon}{2}t^{\mathrm{new}}\frac{\nabla\Phi_\lambda(\mu/t - 1)}{\|\nabla\Phi_\lambda(\mu/t-1)\|_2}\right) + \mathbf{E}[\epsilon_\mu]\right) \\ &= -\frac{\epsilon}{2}\frac{\nabla\Phi_\lambda(\mu/t-1)}{\|\nabla\Phi_\lambda(\mu/t-1)\|_2} + \frac{1}{t^{\mathrm{new}}}\mathbf{E}[\epsilon_\mu], \end{aligned} \tag{24}$$

where the third step follows by definition of δ_μ defined in Algorithm 6.

Next, we bound the $\|v\|_\infty$ as follows:

$$\begin{aligned}\|v\|_\infty &\leq \left\|\frac{\mu}{t}(\frac{t}{t^{\mathrm{new}}} - 1)\right\|_\infty + \left\|\frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_\mu + \epsilon_\mu)\right\|_\infty \\ &\leq \frac{\epsilon}{\sqrt{n}} + \frac{\|\mu^{-1}(\mu^{\mathrm{new}} - \mu)\|_\infty}{0.9} \\ &\leq \frac{\epsilon}{\sqrt{n}} + \frac{3\epsilon}{0.9} \\ &\leq 4\epsilon \\ &\leq \frac{1}{\lambda}. \end{aligned}$$

where the second step follows by definition of t^{new} defined in Algorithm 6 and Part 1 of Assumption C.1, the third step follows by Part 3 of Lemma C.8, and the last step follows by $\epsilon \leq \frac{1}{4\lambda}$.

Since $\|v\|_\infty \leq \frac{1}{\lambda}$, we can apply Part 1 of Lemma C.12 and get

$$\begin{aligned}&\mathbf{E}\left[\Phi_\lambda\left(\frac{\mu^{\mathrm{new}}}{t^{\mathrm{new}}} - 1\right)\right] = \mathbf{E}[\Phi_\lambda(\mu/t + v - 1)] \\ &\leq \Phi_\lambda(\mu/t - 1) + \langle\nabla\Phi_\lambda(\mu/t-1), \mathbf{E}[v]\rangle + 2\,\mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}] \\ &= \Phi_\lambda(\mu/t - 1) - \frac{\epsilon}{2}\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + \frac{t}{t^{\mathrm{new}}}\langle\nabla\Phi_\lambda(\mu/t-1), \mathbf{E}[t^{-1}\epsilon_\mu]\rangle + 2\,\mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}] \\ &\leq \Phi_\lambda(\mu/t - 1) - \frac{\epsilon}{2}\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + \frac{t}{t^{\mathrm{new}}}\|\nabla\Phi_\lambda(\mu/t-1)\|_2 \cdot \|\,\mathbf{E}[t^{-1}\epsilon_\mu]\|_2 + 2\,\mathbf{E}[\|[v]\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}] \\ &\leq \Phi_\lambda(\mu/t - 1) - \frac{\epsilon}{2}\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + 10\epsilon_{\mathrm{mp}}\cdot\epsilon\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + 2\,\mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}], \end{aligned}$$

where the third step follows by substituting $\mathbf{E}[v]$ by Eq. (24), the fourth step follows by $\langle a, b\rangle \leq \|a\|_2 \cdot \|b\|_2$, the fifth step follows by $\|\,\mathbf{E}[t^{-1}\epsilon_\mu]\|_2 \leq 10\epsilon_{\mathrm{mp}}\cdot\epsilon$ (from Part 1 of Lemma C.8 and $\mu \approx_{0.1} t$).

To bound the last term $\mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}]$, we first bound $\mathbf{E}[v_i^2]$,

$$\begin{aligned}
\mathbf{E}[v_i^2] \leq & \, 2\,\mathbf{E}\left[\left(\frac{\mu_i}{t}(\frac{t}{t^{\mathrm{new}}}-1)\right)^2\right] + 2\,\mathbf{E}\left[\left(\frac{1}{t^{\mathrm{new}}}(\widetilde{\delta}_{\mu,i}+\widehat{\delta}_{\mu,i})\right)^2\right] \\
\leq & \, \epsilon^2/n + 2.5\,\mathbf{E}\left[((\mu_i^{\mathrm{new}}-\mu_i)/\mu_i)^2\right] \\
= & \, \epsilon^2/n + 2.5\,\mathbf{Var}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i] + 2.5(\mathbf{E}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i])^2 \\
\leq & \, \epsilon^2/n + 125\epsilon^2/b + 2.5(\mathbf{E}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i])^2 \\
\leq & \, 126\epsilon^2/b + 3(\mathbf{E}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i])^2, \qquad (25)
\end{aligned}$$

where the first step follows by definition of v (see Eq. (23)), the second step follows by $\mu \approx_{0.1} t$ and $(t/t^{\mathrm{new}}-1)^2 \leq \epsilon^2/(4n)$, the third step follows by $\mathbf{E}[x^2] = \mathbf{Var}[x] + (\mathbf{E}[x])^2$, the fourth step follows by Part 2 of Lemma C.8, and the last step follows by $n \geq b$.

Now, we are ready to bound $\mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}]$

$$\begin{aligned}
& \mathbf{E}[\|v\|^2_{\nabla^2\Phi_\lambda(\mu/t-1)}] \\
= & \, \lambda^2 \sum_{i=1}^n \mathbf{E}[\Phi_\lambda(\mu/t-1)_i v_i^2] \\
\leq & \, \lambda^2 \sum_{i=1}^n \Phi_\lambda(\mu/t-1)_i \cdot (126\epsilon^2/b + 3(\mathbf{E}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i])^2) \\
= & \, 126\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) + 3\lambda^2 \sum_{i=1}^n \Phi_\lambda(\mu/t-1)_i \cdot (\mathbf{E}[(\mu_i^{\mathrm{new}}-\mu_i)/\mu_i])^2 \\
\leq & \, 126\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) + 3\lambda \left(\sum_{i=1}^n \lambda^2 \Phi_\lambda(\mu/t-1)_i^2\right)^{1/2} \cdot \|\,\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}}-\mu)]\|_4^2 \\
\leq & \, 126\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) + 3\lambda\left(\lambda\sqrt{n} + \|\nabla\Phi_\lambda(\mu/t-1)\|_2\right)\cdot(5\epsilon)^2,
\end{aligned}$$

where the first step follows by defining $\Phi_\lambda(x)_i = \cosh(\lambda x_i)$, the second step follows from Eq. (25), the fourth step follows from Cauchy-Schwarz inequality, the fifth step follows from Part 3 of Lemma C.12 and the fact that $\|\,\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}}-\mu)]\|_4^2 \leq \|\,\mathbf{E}[\mu^{-1}(\mu^{\mathrm{new}}-\mu)]\|_2^2 \leq (5\epsilon)^2$ (Lemma C.8).

Plugging back, we have

$$\begin{aligned}
& \mathbf{E}\left[\Phi_\lambda\left(\frac{\mu^{\mathrm{new}}}{t^{\mathrm{new}}}-1\right)\right] = \mathbf{E}[\Phi_\lambda(\mu/t+v-1)] \\
\leq & \, \Phi_\lambda(\mu/t-1) - (\frac{\epsilon}{2} - 10\epsilon_{\mathrm{mp}}\cdot\epsilon)\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + 252\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) \\
& + 150\lambda^2\epsilon^2\sqrt{n} + 150\lambda\epsilon^2\|\Phi_\lambda(\mu/t-1)\|_2 \\
\leq & \, \Phi_\lambda(\mu/t-1) - \frac{\epsilon}{3}\|\nabla\Phi_\lambda(\mu/t-1)\|_2 + 252\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) + 150\lambda^2\epsilon^2\sqrt{n} \\
\leq & \, \Phi_\lambda(\mu/t-1) - \frac{\lambda\epsilon}{3\sqrt{n}}(\Phi_\lambda(\mu/t-1)-n) + 252\frac{\lambda^2\epsilon^2}{b}\Phi_\lambda(\mu/t-1) + 150\lambda^2\epsilon^2\sqrt{n} \\
\leq & \, \Phi_\lambda(\mu/t-1) - \frac{\lambda\epsilon}{3\sqrt{n}}(\Phi_\lambda(\mu/t-1)/5 - 2n),
\end{aligned}$$

where the third step follows from $1000\lambda\epsilon \leq 1$ and $1000\epsilon_{\mathrm{mp}} \leq 1$, the fourth step follows from Part 2 of Lemma C.12, and the last step follows from $b \geq 1000\sqrt{n}\lambda\epsilon$. □

As a corollary, we have the following:

Lemma C.14. *During the* MAIN *algorithm, Assumption C.1 is always satisfied. Furthermore, the* CLASSICALSTEP *happens with probability $O(\frac{1}{n^2})$ each step.*

Proof. The second and the fourth assumptions simply follow from the choice of ϵ_{mp} and b.

Let $\Phi^{(k)}$ be the potential at the k-th iteration of the MAIN. The CLASSICALSTEP ensures that $\Phi^{(k)} \leq n^3$ at the end of each iteration. By the definition of Φ and the choice of λ in MAIN, we have that

$$\left\| \frac{xs}{t} - 1 \right\|_\infty \leq \frac{\ln(2n^3)}{\lambda} \leq 0.1.$$

This proves the first assumption $xs \approx_{0.1} t$ with $t > 0$.

For the third assumption, we note that

$$\begin{aligned}
\|\delta_\mu\|_2 &= \left\| \left(\frac{t^{\mathrm{new}}}{t} - 1 \right) xs - \frac{\epsilon}{2} \cdot t^{\mathrm{new}} \cdot \frac{\nabla \Phi_\lambda(\mu/t - 1)}{\|\nabla \Phi_\lambda(\mu/t-1)\|_2} \right\|_2 \\
&\leq \left| \frac{t^{\mathrm{new}}}{t} - 1 \right| \|xs\|_2 + \frac{\epsilon}{2} t^{\mathrm{new}} \\
&\leq \frac{\epsilon}{3\sqrt{n}} \cdot 1.1\sqrt{n}t + 1.01 \cdot \frac{\epsilon}{2} t \leq \epsilon t,
\end{aligned}$$

where we used $xs \approx_{0.1} t$ and the formula of t^{new}. Hence, we proved all assumptions in Assumption C.1.

Now, we bound the probability that CLASSICALSTEP happens. In the beginning of the MAIN, (Ye et al., 1994) is used to modify the linear program with parameter $\min(\frac{\delta}{2}, \frac{1}{\lambda})$. Hence, the initial point x and s satisfies $xs \approx_{1/\lambda} 1$. Therefore, we have $\Phi^{(0)} \leq 10n$. Lemma C.13 shows $\mathbf{E}[\Phi^{(k+1)}] \leq (1 - \frac{\lambda\epsilon}{15\sqrt{n}}) \mathbf{E}[\Phi^{(k)}] + \frac{\lambda\epsilon}{15\sqrt{n}} 10n$. By induction, we have that $\mathbf{E}[\Phi^{(k)}] \leq 10n$ for all k. Since the potential is positive, Markov inequality shows that for any k, $\Phi^{(k)} \geq n^3$ with probability at most $O(\frac{1}{n^2})$. □

C.5. Analysis of cost per iteration

To apply the data structure for projection maintenance (Theorem E.1), we need to first prove the input vector w does not change too much for each step.

Lemma C.15. *Let $x^{\mathrm{new}} = x + \widetilde{\delta}_x$ and $s^{\mathrm{new}} = s + \widetilde{\delta}_s$. Let $w = \frac{x}{s}$ and $w^{\mathrm{new}} = \frac{x^{\mathrm{new}}}{s^{\mathrm{new}}}$. Then we have*

$$\sum_{i=1}^n \left(\mathbf{E}[\ln w_i^{\mathrm{new}}] - \ln w_i \right)^2 \leq 64\epsilon^2, \quad \sum_{i=1}^n \left(\mathbf{Var}[\ln w_i^{\mathrm{new}}] \right)^2 \leq 1000\epsilon^2.$$

Proof. From the definition, we know that

$$\frac{w_i^{\mathrm{new}}}{w_i} = \frac{1}{s_i^{-1} x_i} \frac{x_i + \widetilde{\delta}_{x,i}}{s_i + \widetilde{\delta}_{s,i}} = \frac{1 + x_i^{-1}\widetilde{\delta}_{x,i}}{1 + s_i^{-1}\widetilde{\delta}_{s,i}}.$$

Part 1. For each $i \in [n]$, we have

$$\begin{aligned}
\mathbf{E}[\ln w_i^{\mathrm{new}}] - \ln w_i &= \mathbf{E}\left[\ln(1 + x_i^{-1}\widetilde{\delta}_{x,i}) - \ln(1 + s_i^{-1}\widetilde{\delta}_{s,i}) \right] \\
&\leq 2 |\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i} - s_i^{-1}\widetilde{\delta}_{s,i}]| && \text{by } |s_i^{-1}\widetilde{\delta}_{s,i}|, |x_i^{-1}\widetilde{\delta}_{x,i}| \leq 0.2, \text{Lemma C.3} \\
&\leq 2 |\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i}]| + 2 |\mathbf{E}[s_i^{-1}\widetilde{\delta}_{s,i}]|. && \text{by triangle inequality}
\end{aligned}$$

Thus, summing over all the coordinates gives

$$\sum_{i=1}^n \left(\mathbf{E}[\ln w_i^{\mathrm{new}}] - \ln w_i \right)^2 \leq \sum_{i=1}^n 8(\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i}])^2 + 8(\mathbf{E}[s_i^{-1}\widetilde{\delta}_{s,i}])^2 \leq 64\epsilon^2.$$

where the first step follows by Cauchy-Schwartz inequality, the last step follows by $\| \mathbf{E}[s^{-1}\widetilde{\delta}_s]\|_2^2, \| \mathbf{E}[x^{-1}\widetilde{\delta}_x]\|_2^2 \leq 4\epsilon^2$ (Part 1 of Lemma C.3).

Part 2. For each $i \in [n]$, we have

$$\begin{aligned}
\mathbf{Var}[w_i^{\mathrm{new}}] \leq & \mathbf{E}\left[(\ln w_i^{\mathrm{new}} - \ln w_i)^2\right] \\
= & \mathbf{E}\left[\left(\ln \frac{1 + x_i^{-1}\widetilde{\delta}_{x,i}}{1 + s_i^{-1}\widetilde{\delta}_{s,i}}\right)^2\right] \\
\leq & 2\,\mathbf{E}[(x_i^{-1}\widetilde{\delta}_{x,i} - s_i^{-1}\widetilde{\delta}_{s,i})^2] \\
\leq & 2\,\mathbf{E}[2(x_i^{-1}\widetilde{\delta}_{x,i})^2 + 2(s_i^{-1}\widetilde{\delta}_{s,i})^2] \\
= & 4\,\mathbf{E}[(x_i^{-1}\widetilde{\delta}_{x,i})^2] + 4\,\mathbf{E}[(s_i^{-1}\widetilde{\delta}_{s,i})^2] \\
= & 4\,\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}] + 4(\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i}])^2 + 4\,\mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] + 4(\mathbf{E}[s_i^{-1}\widetilde{\delta}_{s,i}])^2 \\
\leq & 16\epsilon^2/b + 4(\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i}])^2 + 4(\mathbf{E}[s_i^{-1}\widetilde{\delta}_{s,i}])^2,
\end{aligned}$$

where the last step follows by $\mathbf{Var}[x_i^{-1}\widetilde{\delta}_{x,i}], \mathbf{Var}[s_i^{-1}\widetilde{\delta}_{s,i}] \leq 2\epsilon^2/b$ (Part 2 of Lemma C.3).

Thus summing over all the coordinates

$$\begin{aligned}
\sum_{i=1}^n (\mathbf{Var}[w_i^{\mathrm{new}}])^2 \leq & \frac{512n\epsilon^4}{b^2} + 64\sum_{i=1}^n \left((\mathbf{E}[x_i^{-1}\widetilde{\delta}_{x,i}])^4 + (\mathbf{E}[s_i^{-1}\widetilde{\delta}_{s,i}])^4\right) \\
\leq & \frac{512n\epsilon^4}{b^2} + 2048\epsilon^4 \leq 1000\epsilon^2,
\end{aligned}$$

where the last step follows by $\| \mathbf{E}[s^{-1}\widetilde{\delta}_s]\|_2^2, \| \mathbf{E}[x^{-1}\widetilde{\delta}_x]\|_2^2 \leq 4\epsilon^2$ and $b \geq \sqrt{n}\epsilon$.

□

Now, we analyze the cost per iteration in procedure MAIN. This is a direct application of our projection maintenance result.

Lemma C.16. *For $\epsilon \geq \frac{1}{\sqrt{n}}$, each iteration of* MAIN *(Algorithm 6) takes*

$$n^{1+a+o(1)} + \epsilon \cdot (n^{\omega-1/2+o(1)} + n^{2-a/2+o(1)})$$

expected time per iteration in amortized where $0 \leq a \leq \alpha$ controls the batch size in the data structure and $\alpha \in [0,1]$ is the dual exponent of matrix multiplication.

Proof. Lemma C.14 shows that CLASSICALSTEP happens with only $O(1/n^2)$ probability each step. Since the cost of each step only takes $\widetilde{O}(n^{2.5})$, the expected cost is only $\widetilde{O}(n^{0.5})$.

Lemma C.15 shows that the conditions in Theorem E.1 holds with the parameter $C_1 = O(\epsilon), C_2 = O(\epsilon), \epsilon_{\mathrm{mp}} = \Theta(1)$.

In the procedure STOCHASTICSTEP, Theorem E.1 shows that the amortized time per iteration is mainly dominated by two steps:

1. mp.UPDATE(w): $O(\epsilon \cdot (n^{\omega-1/2+o(1)} + n^{2-a/2+o(1)}))$.

2. mp.QUERY($\frac{1}{\sqrt{XS}}\widetilde{\delta}_\mu$): $O(n^{1+b+o(1)} + n^{1+a+o(1)})$.

□

D. Main result

The goal of this section is to putting everything together and prove the following main theorem:

Theorem D.1 (Main result)**.** *Given a linear program* $\min_{Ax=b,x\geq 0} c^\top x$ *with no redundant constraints. Assume that the polytope has diameter* R *in* ℓ_1 *norm, namely, for any* $x \geq 0$ *with* $Ax = b$*, we have* $\|x\|_1 \leq R$*.*

Then, for any $0 < \delta \leq 1$, $\textsc{Main}(A, b, c, \delta)$ *outputs* $x \geq 0$ *such that*

$$c^\top x \leq \min_{Ax=b,x\geq 0} c^\top x + \delta \cdot \|c\|_\infty R \quad \textit{and} \quad \|Ax - b\|_1 \leq \delta \cdot (R\|A\|_1 + \|b\|_1)$$

in expected time

$$\left(n^{\omega+o(1)} + n^{2.5-\alpha/2+o(1)} + n^{2+1/6+o(1)}\right) \cdot \log(n/\delta)$$

where ω *is the exponent of matrix multiplication,* α *is the dual exponent of matrix multiplication.*

For the current value of $\omega \sim 2.38$ *and* $\alpha \sim 0.31$*, the expected time is simply* $n^{\omega+o(1)} \log(\frac{n}{\delta})$*.*

Proof. In the beginning of the $\textsc{Main}$ algorithm, (Ye et al., 1994) is called to modify the linear program. Then, we run the stochastic central path method on this modified linear program.

When the algorithm stops, we obtain a vector x and s such that $xs \approx_{0.1} t$ with $t \leq \frac{\delta^2}{32n^3}$. Hence, the duality gap is bounded by $\sum_i x_i s_i \leq (\delta/4n)^2$. (Ye et al., 1994) shows how to obtain an approximate solution of the original linear program with the guarantee needed using the x and s we just found.

Since t is decreased by $1 - \frac{\epsilon}{3\sqrt{n}}$ factor each iteration, it takes $O(\frac{\sqrt{n}}{\epsilon} \cdot \log(\frac{n}{\delta}))$ iterations in total. In Lemma C.16, we proved that each iteration takes

$$n^{1+a+o(1)} + \epsilon \cdot (n^{\omega-1/2+o(1)} + n^{2-a/2+o(1)}).$$

and hence the total runtime is

$$O(n^{2.5-a/2+o(1)} + n^{\omega+o(1)} + \epsilon^{-1} n^{1.5+a+o(1)}) \cdot \log(n/\delta).$$

Since $\epsilon = \Theta(\frac{1}{\log n})$, the total runtime is

$$O(n^{2.5-a/2+o(1)} + n^{\omega+o(1)} + n^{1.5+a+o(1)}) \cdot \log(n/\delta).$$

Finally, we note that the optimal choice of a is $\min(\frac{2}{3}, \alpha)$, which gives the promised runtime. □

E. Projection Maintenance

In this section, we present how to resolve the second bottleneck. The main idea is similar to (Cohen et al., 2019b). We need to maintain the query structure Ph, where P is the projection matrix as shown in Figure 3. We use the idea of lazy update and low-rank update as discussed in the main body. Here, we supplement the explanation of constructing a copy of W in the main body by using a 2-person chasing game, as shown in Figure 4.

E.1. Main result

The goal of this section is to prove the following theorem:

Theorem E.1 (Projection maintenance)**.** *Given a full rank matrix* $A \in \mathbb{R}^{d\times n}$ *with* $n \geq d$ *and a tolerance parameter* $0 < \epsilon_{\mathrm{mp}} < 1/4$*. Given any positive number* a *such that* $a \leq \alpha$ *where* α *is the dual exponent of matrix multiplication. Let* $R_{1,*}, \cdots, R_{L,*} \in \mathbb{R}^{n^b \times n}$ *denote a list of sketching matrices, where* $b \in [0, 1]$*. There is a deterministic data structure (Algorithm 8) that approximately maintains the projection matrices*

$$\sqrt{W}A^\top (AWA^\top)^{-1} A\sqrt{W}$$

for positive diagonal matrices W *through the following two operations:*

Algorithm 8 Projection Maintenance Data Structure

datastructure MAINTAINPROJECTION ▷ Theorem E.1

members
- $w \in \mathbb{R}^n$ ▷ Target vector
- $v, \widetilde{v} \in \mathbb{R}^n$ ▷ Approximate vector
- $A \in \mathbb{R}^{d \times n}$
- $M \in \mathbb{R}^{n \times n}$ ▷ Approximate projection matrix
- $Q \in \mathbb{R}^{n \times n^b L}$ ▷ Sketched version approximate projection matrix
- $R_{1,*}, R_{2,*}, \cdots R_{L,*} \in \mathbb{R}^{n^b \times n}$ ▷ Sketching matrices
- $l \in \mathbb{N}_+, L \in \mathbb{N}_+$
- $\epsilon_{\mathrm{mp}} \in (0, 1/4)$ ▷ Tolerance
- $a \in (0, \alpha]$ ▷ Batch Size for Update (n^a)

end members

procedure INITIALIZE($A, w, \epsilon_{\mathrm{mp}}, a$) ▷ Lemma E.4
- $w \leftarrow w$, $v \leftarrow w$, $\epsilon_{\mathrm{mp}} \leftarrow \epsilon_{\mathrm{mp}}$, $A \leftarrow A$, $a \leftarrow a$
- $M \leftarrow A^\top (AVA^\top)^{-1} A$
- Choosing $R_{1,*}, R_{2,*} \cdots, R_{L,*} \in \mathbb{R}^{n^b \times n}$ to be sketching matrices
- $R \leftarrow [R_{*,1}, R_{*,2} \cdots, R_{*,L}]$
- $Q \leftarrow M\sqrt{V}R^\top$
- $l \leftarrow 1$

end procedure

end datastructure

1. UPDATE(w): *Output a vector $\widetilde{v}$ such that for all i,*

$$(1 - \epsilon_{\mathrm{mp}})\widetilde{v_i} \leq w_i \leq (1 + \epsilon_{\mathrm{mp}})\widetilde{v_i}.$$

2. QUERY(h): *Output $\sqrt{\widetilde{V}}A^\top (A\widetilde{V}A^\top)^{-1} A\sqrt{\widetilde{V}}(R^\top)_{*,l} R_{l,*} h$ for the $\widetilde{v}$ outputted by the last call to* UPDATE.

The data structure takes $n^2 d^{\omega-2}$ time to initialize and each call of QUERY(h) *takes time*

$$n^{1+b+o(1)} + n^{1+a+o(1)}.$$

Furthermore, if the initial vector $w^{(0)}$ and the (random) update sequence $w^{(1)}, \cdots, w^{(T)} \in \mathbb{R}^n$ satisfies

$$\sum_{i=1}^n \left(\mathbf{E}[\ln w_i^{(k+1)}] - \ln w_i^{(k)} \right)^2 \leq C_1^2 \qquad \text{and} \qquad \sum_{i=1}^n (\mathbf{Var}[\ln w_i^{(k+1)}])^2 \leq C_2^2$$

with the expectation and variance is conditional on $w_i^{(k)}$ for all $k = 0, 1, \cdots, T-1$. Then, the amortized expected time[8] *per call of* UPDATE(w) *is*

$$(C_1/\epsilon_{\mathrm{mp}} + C_2/\epsilon_{\mathrm{mp}}^2) \cdot (n^{\omega - 1/2 + o(1)} + n^{2 - a/2 + o(1)}).$$

Remark E.2. *For our linear program algorithm, we have $C_1 = O(1/\log n)$, $C_2 = O(1/\log n)$ and $\epsilon_{\mathrm{mp}} = \Theta(1)$. See Lemma C.15.*

To verify the correctness of our updates, we have the following lemma:

[8] If the input is deterministic, so is the output and the runtime.

Algorithm 9 UPDATE

datastructure MAINTAINPROJECTION

procedure UPDATE(w) ▷ Lemma E.5
  $y_i \leftarrow \ln w_i - \ln v_i, \forall i \in [n]$
  $r \leftarrow$ the number of indices i such that $|y_i| \geq \epsilon_{\mathrm{mp}}/2$.
  if $r < n^a$ **then**
    $v^{\mathrm{new}} \leftarrow v$
    $M^{\mathrm{new}} \leftarrow M$
    $l \leftarrow l + 1$
  else
    Let $\pi : [n] \to [n]$ be a sorting permutation such that $|y_{\pi(i)}| \geq |y_{\pi(i+1)}|$
    while $1.5 \cdot r < n$ and $|y_{\pi(\lceil 1.5 \cdot r \rceil)}| \geq (1 - 1/\log n)|y_{\pi(r)}|$ **do**
      $r \leftarrow \min(\lceil 1.5 \cdot r \rceil, n)$
    end while
    $v^{\mathrm{new}}_{\pi(i)} \leftarrow \begin{cases} w_{\pi(i)} & i \in \{1, 2, \cdots, r\} \\ v_{\pi(i)} & i \in \{r+1, \cdots, n\} \end{cases}$

    ▷ Compute $M^{\mathrm{new}} = A^\top (A V^{\mathrm{new}} A^\top)^{-1} A$ via Matrix Woodbury
    $\Delta \leftarrow \mathrm{diag}(v^{\mathrm{new}} - v)$ ▷ $\Delta \in \mathbb{R}^{n \times n}$ and $\|\Delta\|_0 = r$
    $\Gamma \leftarrow \mathrm{diag}(\sqrt{v^{\mathrm{new}}} - \sqrt{v})$
    Let $S \leftarrow \pi([r])$ be the first r indices in the permutation.
    Let $M_S \in \mathbb{R}^{n \times r}$ be the r columns from S of M.
    Let $M_{S,S}, \Delta_{S,S} \in \mathbb{R}^{r \times r}$ be the r rows and columns from S of M and Δ.
    $M^{\mathrm{new}} \leftarrow M - M_{*,S} \cdot (\Delta^{-1}_{S,S} + M_{S,S})^{-1} \cdot (M_{*,S})^\top$
    Re-generate R
    $Q^{\mathrm{new}} \leftarrow Q + (M^{\mathrm{new}} \cdot \Gamma) \cdot R^\top + (M^{\mathrm{new}} - M) \cdot \sqrt{V} \cdot R^\top$
    $l \leftarrow 1$
  end if
  $v \leftarrow v^{\mathrm{new}}$
  $M \leftarrow M^{\mathrm{new}}$
  $Q \leftarrow Q^{\mathrm{new}}$
  $\widetilde{v}_i \leftarrow \begin{cases} v_i & \text{if } |\ln w_i - \ln v_i| < \epsilon_{\mathrm{mp}}/2 \\ w_i & \text{otherwise} \end{cases}$
  return $\widetilde{v}$
end procedure

end datastructure

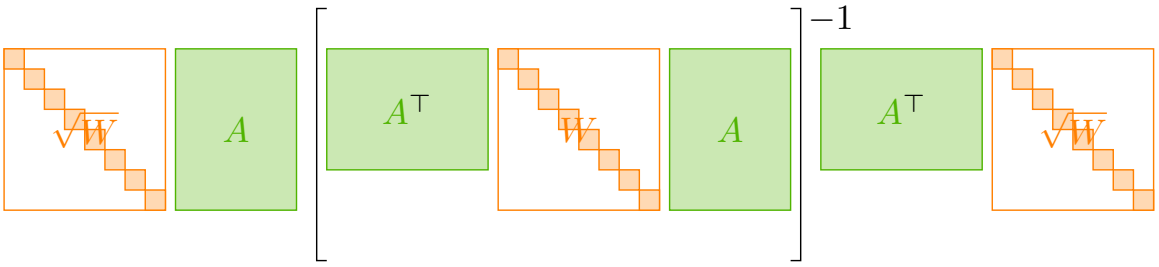


Figure 3: A visualization of projection matrix.

Lemma E.3 (Correctness of the algorithm)**.** *The output of* UPDATE(w) *in Algorithm 9 satisfies*

$$M = A^\top (AVA^\top)^{-1} A \text{ and}$$
$$Q = M\sqrt{V}R^\top$$

The output of QUERY(h) *in Algorithm 10 satisfies*

$$p_s = \widetilde{P}(R^\top)_{*,l} R_{l,*} h$$
$$p_x = (I - \widetilde{P})(R^\top)_{*,l} R_{l,*} h$$

where $\widetilde{P} = \sqrt{\widetilde{V}} A^\top (A\widetilde{V}A^\top)^{-1} A\sqrt{\widetilde{V}}$*, and* $\widetilde{V}$ *is outputted by* UPDATE(w).

Proof. For UPDATE(w) procedure, note v^{new} only differs from w in entries correspond to the set S. Thus, by Matrix Woodbury Identity and definition of M^{new}, we have

$$\begin{aligned}
A^\top (AV^{\mathrm{new}}A^\top)^{-1} A &= A^\top (A(V+\Delta)A^\top)^{-1} A \\
&= A^\top \Big((AVA^\top)^{-1} - (AVA^\top)^{-1} A_{*,S} \cdot \Big(\Delta_{S,S}^{-1} + (A^\top)_{S,*} (AVA^\top)^{-1} A_{*,S} \Big)^{-1} \\
&\quad \cdot (A^\top)_{S,*} (AVA^\top)^{-1} \Big) A \\
&= A^\top (AVA^\top)^{-1} A - A^\top (AVA^\top)^{-1} A_{*,S} \cdot \Big(\Delta_{S,S}^{-1} + (A^\top)_{S,*} (AVA^\top)^{-1} A_{*,S} \Big)^{-1} \\
&\quad \cdot (A^\top)_{S,*} (AVA^\top)^{-1} A \\
&= M - M_{*,S} \Big(\Delta_{S,S}^{-1} + M_{S,S} \Big)^{-1} M_{S,*} \\
&= M^{\mathrm{new}}.
\end{aligned}$$

Note the output $M = M^{\mathrm{new}}$ and $V = V^{\mathrm{new}}$, so we have the output satisfying $M = A^\top (AVA^\top)^{-1} A$.

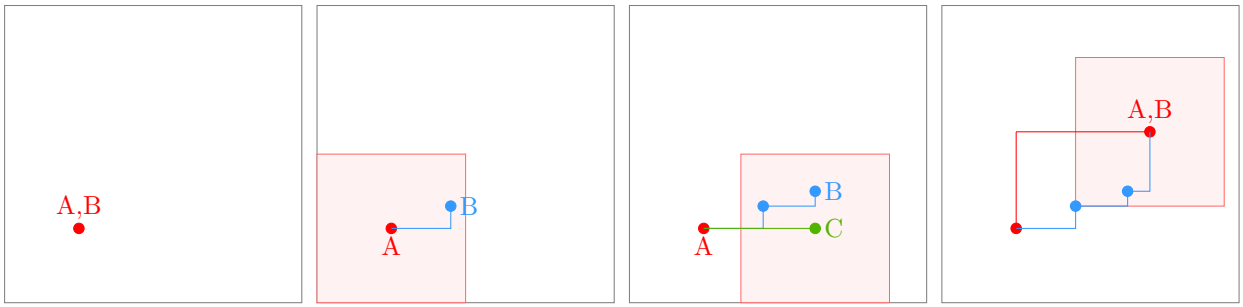
As for Q, notice by definition

$$\begin{aligned}
Q^{\mathrm{new}} &= Q + (M^{\mathrm{new}} \cdot \Gamma) \cdot R^\top + (M^{\mathrm{new}} - M) \cdot \sqrt{V} \cdot R^\top \\
&= M\sqrt{V}R^\top + M^{\mathrm{new}} (\sqrt{V^{\mathrm{new}}} - \sqrt{V}) R^\top + (M^{\mathrm{new}} - M)\sqrt{V}R^\top \\
&= M^{\mathrm{new}} \sqrt{V^{\mathrm{new}}} R^\top
\end{aligned}$$

Again, since the output $Q = Q^{\mathrm{new}}$, $M = M^{\mathrm{new}}$ and $V = V^{\mathrm{new}}$, we have the output satisfying $Q = M\sqrt{V}R^\top$.

For QUERY(h) procedure, by definition we have

$$p_m = \sqrt{\widetilde{V}} \cdot (M_{*,\widetilde{S}}) \cdot (\widetilde{\Delta}_{\widetilde{S},\widetilde{S}}^{-1} + M_{\widetilde{S},\widetilde{S}})^{-1} \cdot (Q_{\widetilde{S},l} + M_{\widetilde{S},*} \cdot \widetilde{\Gamma} \cdot (R^\top)_{*,l}) \cdot R_{l,*} \cdot h$$



(a) round 0 (b) round 1 (c) round 2 (d) round 3

Figure 4: Visualization of projection maintenance: we model the task as a game of person A chasing person B, where person A needs to report the approximate location of person B in each round while moving as little as possible. Person A brings a drone C which can only fly in one direction. The location of Person B represents the exact projection matrix, the location of Person A represents the projection matrix we store in the datastructure, the location reported in each round represents the output of our algorithm. In the beginning, they start off at the same location. At round 1, person B moves but is still close to person A. In this case, person A stays idol and reports its location. This case corresponds to the situation that the projection changes little in all coordinates, so we use the idea of lazy updates. At round 2, person B moves far away from A only in one direction. In this case, person A keeps its location and releases drone C to chase person B in the direction where person B moves a lot. And we report the location of the drone C.This case corresponds to the situation that the projection only changes a lot in few coordinates, so we use the idea of low-rank updates while keeping lazy on updating the stored projection matrix. In round 3, person B moves far away from A in all directions. In this case, person A moves to the location of person B and reports its location. This case corresponds to the situation that the projection changes a lot in many coordinates, so we update the stored projection matrix.

$$= \sqrt{\widetilde{V}} \cdot (M_{*,\widetilde{S}}) \cdot (\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + M_{\widetilde{S},\widetilde{S}})^{-1} \cdot M_{\widetilde{S},*} \cdot \widetilde{V} \cdot (R^\top)_{*,l} \cdot R_{l,*} \cdot h.$$

Thus,

$$\begin{aligned} p_s &= \sqrt{\widetilde{V}} \cdot (Q_{*,l} + M \cdot \widetilde{\Gamma} \cdot (R^\top)_{*,l}) \cdot R_{l,*} \cdot h - p_m \\ &= \sqrt{\widetilde{V}} \cdot M \cdot \widetilde{V} \cdot (R^\top)_{*,l} \cdot R_{l,*} \cdot h - p_m \\ &= \sqrt{\widetilde{V}} (M - M_{*,\widetilde{S}} (\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + M_{\widetilde{S},\widetilde{S}})^{-1} M_{\widetilde{S},*}) \widetilde{V} (R^\top)_{*,l} R_{l,*} h. \end{aligned}$$

Note $\widetilde{V}$ only differs from V in entries correspond to the set $\widetilde{S}$, again by Matrix Woodbury Identity and definition of M, we have

$$\begin{aligned} A^\top (A\widetilde{V}A^\top)^{-1} A &= A^\top (A(V + \widetilde{\Delta})A^\top)^{-1} A \\ &= A^\top \Big((AVA^\top)^{-1} - (AVA^\top)^{-1} A_{*,\widetilde{S}} \cdot \Big(\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + (A^\top)_{\widetilde{S},*} (AVA^\top)^{-1} A_{*,\widetilde{S}} \Big)^{-1} \\ &\quad \cdot (A^\top)_{\widetilde{S},*} (AVA^\top)^{-1} \Big) A \\ &= A^\top (AVA^\top)^{-1} A - A^\top (AVA^\top)^{-1} A_{*,\widetilde{S}} \cdot \Big(\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + (A^\top)_{\widetilde{S},*} (AVA^\top)^{-1} A_{*,\widetilde{S}} \Big)^{-1} \\ &\quad \cdot (A^\top)_{\widetilde{S},*} (AVA^\top)^{-1} A \\ &= M - M_{*,\widetilde{S}} \Big(\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + M_{\widetilde{S},\widetilde{S}} \Big)^{-1} M_{\widetilde{S},*}, \end{aligned}$$

which implies

$$p_s = \sqrt{\widetilde{V}} A^\top (A\widetilde{V}A^\top)^{-1} A\widetilde{V} (R^\top)_{*,l} R_{l,*} h = \widetilde{P} (R^\top)_{*,l} R_{l,*} h.$$

Further,

$$p_x = (R^\top)_{*,l} R_{l,*} h - p_s = (I - \widetilde{P})(R^\top)_{*,l} R_{l,*} h,$$

which completes the proof. □

Above lemma verifies our algorithm. Now we consider the running time of the projection maintenance, which consists of Initialization time, update time and query time, as discussed below.

E.2. Initialization time, update time

To formalize the amortized runtime proof, we first analyze the initialization time (Lemma E.4), update time (Lemma E.5), and query time (Lemma E.6) of our projection maintenance data-structure.

Lemma E.4 (Initialization time)**.** *The initialization time of data-structure* MAINTAINPROJECTION *(Algorithm 8) is* $O(n^2 d^{\omega-2})$.

Proof. Given matrix $A \in \mathbb{R}^{d \times n}$ and diagonal matrix $V \in \mathbb{R}^{n \times n}$, computing $A^\top (AVA^\top)^{-1} A$ takes $O(n^2 d^{\omega-2})$. □

Lemma E.5 (Update time)**.** *The update time of data-structure* MAINTAINPROJECTION *(Algorithm 9) is* $O(r g_r n^{2+o(1)})$ *where* r *is the number of indices we updated in* v.

Proof. The proof is identical to (Cohen et al., 2019b; Lee et al., 2019). We omit the details here.

□

E.3. Query time

Algorithm 10 QUERY

datastructure MAINTAINPROJECTION

procedure QUERY(h) ▷ Lemma E.6

Let $\widetilde{S}$ be the indices i such that $|\ln w_i - \ln v_i| \geq \epsilon_{\mathrm{mp}}/2$.

$\widetilde{\Delta} \leftarrow \widetilde{V} - V$

$\widetilde{\Gamma} \leftarrow \sqrt{\widetilde{V}} - \sqrt{V}$

$p_m \leftarrow \sqrt{\widetilde{V}} \cdot (M_{*,\widetilde{S}}) \cdot (\widetilde{\Delta}^{-1}_{\widetilde{S},\widetilde{S}} + M_{\widetilde{S},\widetilde{S}})^{-1} \cdot (Q_{\widetilde{S},l} + M_{\widetilde{S},*} \cdot \widetilde{\Gamma} \cdot (R^\top)_{*,l}) \cdot R_{l,*} \cdot h$

$p_s \leftarrow \sqrt{\widetilde{V}} \cdot (Q_{*,l} + M \cdot \widetilde{\Gamma} \cdot (R^\top)_{*,l}) \cdot R_{l,*} \cdot h - p_m$

$p_x \leftarrow (R^\top)_{*,l} \cdot R_{l,*} \cdot h - p_s$

return (p_x, p_s)

end procedure

end datastructure

Lemma E.6 (Query time)**.** *The query time of data-structure* MAINTAINPROJECTION *(Algorithm 8) is* $O(n^{1+b+o(1)} + n^{1+a+o(1)})$.

Proof. Notice by the algorithm we have $|\widetilde{S}| \leq n^a$. Thus, $\widetilde{\Gamma}$ is a sparse diagonal matrix with at most n^a non-zero elements. The running time mainly comes from three parts.

Part 1. Computing p_m:

- Compute $R_{l,*} \cdot h$: matrix-vector multiplication between matrix of size $n^b \times n$ and vector of size $n \times 1$, this takes n^{1+b} time.
- Compute $(R^\top)_{*,l} \cdot (R_{l,*} h)$: matrix-vector multiplication between matrix of size $n \times n^b$ and vector of size $n^b \times 1$, this takes n^{1+b} time.

- Compute $\widetilde{\Gamma} \cdot (R_{l,*}^\top R_{l,*} h)$: matrix-vector multiplication between sparse diagonal matrix with at most n^a non-zero elements and vector of size $n \times 1$, this takes n^a time.
- Compute $M_{\widetilde{S},*} \cdot (\widetilde{\Gamma} R_{l,*}^\top R_{l,*} h)$: matrix-vector multiplication between matrix of size at most $n^a \times n$ and sparse vector with at most n^a non-zero elements, this takes n^{2a} time.
- Compute $Q_{\widetilde{S},l} \cdot (R_{l,*} h)$: matrix-vector multiplication between matrix of size at most $n^a \times n^b$ and vector of size $n^b \times 1$, this takes n^{a+b} time.
- Compute $(\widetilde{\Delta}_{\widetilde{S},\widetilde{S}}^{-1} + M_{\widetilde{S},\widetilde{S}})^{-1}$: inverse of matrix of size at most $n^a \times n^a$, this takes $n^{a\omega}$ time.
- Compute $(\widetilde{\Delta}_{\widetilde{S},\widetilde{S}}^{-1} + M_{\widetilde{S},\widetilde{S}})^{-1} \cdot [(Q_{\widetilde{S},l} + M_{\widetilde{S},*} \widetilde{\Gamma} (R^\top)_{*,l}) R_{l,*} h]$: matrix-vector multiplication between matrix of size at most $n^a \times n^a$ and vector of size at most $n^a \times 1$, this takes n^{2a} time.
- Compute $\sqrt{\widetilde{V}} \cdot (M_{*,\widetilde{S}})$: matrix-matrix multiplication between diagonal matrix of size $n \times n$ and matrix of size at most $n \times n^a$, this takes n^{1+a} time.
- Compute $[\sqrt{\widetilde{V}} M_{*,\widetilde{S}}] \cdot [(\widetilde{\Delta}_{\widetilde{S},\widetilde{S}}^{-1} + M_{\widetilde{S},\widetilde{S}})^{-1} (Q_{\widetilde{S},l} + M_{\widetilde{S},*} \widetilde{\Gamma} (R^\top)_{*,l}) R_{l,*} h]$: matrix-vector multiplication between matrix of size at most $n \times n^a$ and vector of size at most $n^a \times 1$, this takes n^{1+a} time.

To conclude, we can compute p_m in $O(n^{1+b} + n^{a\omega} + n^{1+a})$ time.

Part 2. Computing p_s:

- Compute $R_{l,*} h$ and $\widetilde{\Gamma} R_{l,*}^\top R_{l,*} h$ in same way as in calculating p_m: take n^{1+b} and $O(n^{1+b} + n^a)$ time respectively.
- Compute $\sqrt{\widetilde{V}} \cdot Q_{*,l}$: matrix-matrix multiplication between diagonal matrix of size $n \times n$ and matrix of size $n \times n^b$, takes n^{1+b} time.
- Compute $[\sqrt{\widetilde{V}} Q_{*,l}] \cdot [R_{l,*} h]$: matrix-vector multiplication between matrix of size $n \times n^b$ and vector of size $n^b \times 1$, takes n^{1+b} time.
- Compute $M \cdot [\widetilde{\Gamma} R_{l,*}^\top R_{l,*} h]$: matrix-vector multiplication between matrix of size $n \times n$ and sparse vector with at most n^a non-zero elements, takes $O(n^{1+a})$ time.
- Compute $\sqrt{\widetilde{V}} \cdot [M \widetilde{\Gamma} R_{l,*}^\top R_{l,*} h]$: matrix-vector multiplication between diagonal matrix of size $n \times n$ and vector of size $n \times 1$, takes n time.

To conclude, we can compute p_s in $O(n^{1+b} + n^{1+a})$ time.

Part 3. Computing p_x:

- Compute $R_{l,*}^\top R_{l,*} h$ in same way as in calculating p_m: take $O(n^{1+b})$ time.

Thus, overall the running time is

$$O(n^{1+a} + n^{1+b} + n^{a\omega}).$$

Finally, we note that $\omega \leq 3 - \alpha \leq 3 - a$ (see (Cohen et al., 2019b)) and hence $a \cdot \omega \leq a(3-a) \leq 1 + a$. Therefore, the final running time it takes is $O(b^{1+b+o(1)} + n^{1+a+o(1)})$. □

References	Sampling/Sketching	How?	Feasible?	Oblivious?
(Cohen et al., 2019b)	Sampling on the right	$\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}Dh$	Yes	No
(Lee et al., 2019)	Sketching on the left	$R^\top R\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}h$	No	Yes
This work	Sketching on the right	$\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}R^\top Rh$	Yes	Yes

Table 4: Summary of different approaches to reduce dimensionality. We remark that (Brand, 2020) propose a deterministic technique called vector maintenance without using any randomized technique and achieve the same running time as this work. In (Jiang et al., 2021), they propose an algorithm that improves the term $n^{2+1/6}$ to $n^{2+1/18}$.

F. Comparison to state of the art results

In this section, we will explain the solutions of one-step central path equation (a linear system) are different in two previous works and this paper. Let us just believe adding a sparse diagonal D and using a sketching matrix R is able to reduce the computational cost. We won't explain the reason why it can reduce the computational cost.

We compare our approach to the state of the art results (Cohen et al., 2019b; Lee et al., 2019). Though all methods shares the same running time up to subpolynomial factors, they use different randomization techniques.

Note that the major question in fast central path method is how to speed up the following calculation

$$\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}\cdot h. \tag{26}$$

F.1. Feasible vs Infeasible

Lee, Song, Zhang'19. The approach of (Lee et al., 2019) can be interpreted as sketching on the left. Let $R\in\mathbb{R}^{\sqrt{n}\times n}$ be a sketching matrix. Let T be the number of iterations of the iterative algorithm. We pick T independent sketching matrices at the beginning of the algorithm. In each iteration $t\in[T]$, we are computing

$$R^\top\cdot R\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}\cdot h \tag{27}$$

which can be viewed as an **approximated** solution to the linear system in the classical central path method:

$$\begin{aligned} X\delta_s + S\delta_x &= \widetilde{\delta}_\mu, \\ A\delta_x &= 0, \\ A^\top\delta_y + \delta_s &= 0, \end{aligned} \tag{28}$$

where

$$\widetilde{\delta}_\mu = \delta_\mu$$

They choose δ_x, δ_s and δ_y as follows

$$\begin{aligned} \delta_x &= \frac{X}{\sqrt{XS}}(I - R^\top RP)\frac{1}{\sqrt{XS}}\delta_\mu \\ \delta_s &= \frac{S}{\sqrt{XS}}R^\top RP\frac{1}{\sqrt{XS}}\delta_\mu \\ \delta_y &= -(A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}\frac{1}{\sqrt{XS}}\delta_\mu \end{aligned} \tag{29}$$

Note that plugging the above solution back to Eq. (28), we can see line 2 and 3 (primal and dual feasibility conditions) of the linear system does not hold exactly, which results in an infeasible issue in each iteration. Specifically,

For the first line of Eq. (28), we have the LHS becomes

$$\text{LHS} = X\delta_s + S\delta_x.$$

The RHS becomes

$$\mathrm{RHS} = \widetilde{\delta}_\mu = X\delta_s + S\delta_x.$$

Thus, LHS $=$ RHS.

For the second line of Eq. (28), i.e., the primal feasible condition, we have the left hand side becomes

$$\begin{aligned}\mathrm{LHS} = & A\delta_x \\ = & A\frac{X}{\sqrt{XS}}(I - R^\top RP)\frac{1}{\sqrt{XS}}\delta_\mu \\ = & \frac{A}{S}\delta_\mu - A\sqrt{\frac{X}{S}}R^\top R\sqrt{\frac{X}{S}}A^\top(A\frac{X}{S}A^\top)^{-1}A\frac{1}{S}\delta_\mu\end{aligned}$$

while the right hand side is always 0 and does not match the left hand side. Therefore, Eq. (29) does not satisfy the primal feasible condition in each iteration.

For the third line of Eq. (28), i.e., the dual feasible condition, we have the left hand side becomes

$$\begin{aligned}\mathrm{LHS} = & A^\top\delta_y + \delta_s \\ = & -A^\top(A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}\frac{1}{\sqrt{XS}}\delta_\mu + \frac{S}{\sqrt{XS}}R^\top RP\frac{1}{\sqrt{XS}}\delta_\mu \\ = & A^\top(A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}\frac{1}{\sqrt{XS}}\delta_\mu + \sqrt{\frac{S}{X}}R^\top R\sqrt{\frac{X}{S}}A^\top(A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}\frac{1}{\sqrt{XS}}\delta_\mu\end{aligned}$$

while the right hand side is always 0 and does not match the left hand side. Therefore, Eq. (29) does not satisfy the dual feasible condition in each iteration.

Cohen, Lee, Song'19. The approach of (Cohen et al., 2019b) can be interpreted as sampling on the complementarity gap h. Let D denote a random diagonal sampling matrix, (Cohen et al., 2019b) approximates Eq. (26) by

$$\underbrace{\sqrt{W}A^\top(AWA^\top)^{-1}A\sqrt{W}}_{P}\cdot D\cdot h,$$

where D roughly only has $\sqrt{n}$ non-zero entries on the diagonal. The approach can also be viewed as explicit solving the linear system (Eq. (30)) in each iteration:

$$\begin{aligned}X\delta_s + S\delta_x &= \widetilde{\delta}_\mu, \\ A\delta_x &= 0, \\ A^\top\delta_y + \delta_s &= 0,\end{aligned} \tag{30}$$

where

$$\widetilde{\delta}_\mu = D\delta_\mu.$$

The above system (consists of three equations) can be solved **exactly** by:

$$\begin{aligned}\delta_x &= \frac{X}{\sqrt{XS}}(I - P)\frac{1}{\sqrt{XS}}D\delta_\mu, \\ \delta_s &= \frac{S}{\sqrt{XS}}P\frac{1}{\sqrt{XS}}D\delta_\mu, \\ \delta_y &= (A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}\frac{1}{\sqrt{XS}}D\delta_\mu.\end{aligned} \tag{31}$$

which means (Cohen et al., 2019b) is a feasible method.

This paper. Our methods sketching on the right as follows:

$$\sqrt{W}A^\top (AWA^\top)^{-1} A\sqrt{W}R^\top \cdot R \cdot h.$$

Our method can be interpreted as an **exact** solution to the new linear system we construct:

$$\begin{aligned} X\delta_s + S\delta_x &= \widetilde{\delta}_\mu, \\ A\delta_x &= 0, \\ A^\top \delta_y + \delta_s &= 0, \end{aligned} \tag{32}$$

where

$$\widetilde{\delta}_\mu = \sqrt{XS}R^\top R \frac{1}{\sqrt{XS}}\delta_\mu.$$

The above linear system (Eq. (32)) can be solved **exactly** by

$$\begin{aligned} \delta_x &= \frac{X}{\sqrt{XS}}(I-P)R^\top R\frac{1}{\sqrt{XS}}\delta_\mu, \\ \delta_s &= \frac{S}{\sqrt{XS}}PR^\top R\frac{1}{\sqrt{XS}}\delta_\mu, \\ \delta_y &= -(A\frac{X}{S}A^\top)^{-1}A\sqrt{\frac{X}{S}}R^\top R\frac{1}{\sqrt{XS}}\delta_\mu, \end{aligned} \tag{33}$$

which means our method is feasible.

F.2. Oblivious vs non-oblivious

Cohen, Lee, Song'19. The explicit construction for the sampling matrix D in Eq. (31) is given by (Cohen et al., 2019b):

$$\widetilde{\delta}_{\mu,i} = \begin{cases} \delta_{\mu,i}/p_i, & \text{with probability } p_i = k \cdot (\frac{\delta_{\mu,i}^2}{\sum_l \delta_{\mu,l}^2} + \frac{1}{n}) \\ 0, & \text{else.} \end{cases}$$

Therefore, (Cohen et al., 2019b) is a non-oblivious approach since the sampling matrix D depends on the value of δ_μ.

Lee, Song, Zhang'19. The sketching matrix R in Eq. (29) does not depend on the value of δ_μ, meaning it is an oblivious method.

This paper The sketching matrix R in our approach (see Eq. (33)) does not depend on the value of δ_μ as shown in Algorithm 6, which makes ours an oblivious method.

To conclude, we summarize above discussion in Table 4. Compare to previous results, our method is both feasible and oblivious. These advantages help to implement expensive calculations in the pre-processing stage and have a much simpler analysis, which gives the potential to generalize to other optimization problems.

G. Comparison to JL, SE and AMP

In this section we compare the guarantees of coordinate-wise embedding (CE, Definition 2.1) with three different guarantees: Johnson-Lindenstrauss embedding (JL, (Johnson & Lindenstrauss, 1984)), ℓ_2-subspace embedding (SE, (Sarlós, 2006; Woodruff, 2014)), and approximate matrix product (AMP, (Sarlós, 2006)). We also consider the JL moment property (JLM, (Kane & Nelson, 2012)) which is closely related to AMP. We first state the definitions of these embeddings and properties here.

Definition G.1 (Coordinate-wise embedding (CE), restatement of Definition 2.1)**.** *Given parameters $\alpha, \beta \in \mathbb{R}$ and $\delta \in (0,1)$, we say a randomized matrix $R \in \mathbb{R}^{b\times n}$ from a distribution Π satisfies (α, β, δ)-coordinate-wise embedding property if for any fixed vector $g, h \in \mathbb{R}^n$, we have*

1. $\mathbf{E}_{R\sim\Pi}[g^\top R^\top Rh] = g^\top h,$

$$2.\ \mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{\alpha}{b}\|g\|_2^2\|h\|_2^2,$$
$$3.\ \Pr_{R\sim\Pi}\left[|g^\top R^\top Rh - g^\top h| \geq \frac{\beta}{\sqrt{b}}\|g\|_2\|h\|_2\right] \leq \delta.$$

From now on we will refer to the three properties of CE as CE1, CE2, and CE3.

Definition G.2 (Johnson-Lindenstrauss embedding (JL) (Johnson & Lindenstrauss, 1984), restatement of Definition 3.1)**.** *Given $\epsilon, \delta \in (0,1)$, a finite point set $S \subset \mathbb{R}^n$ with $|S| = m$, we say a randomized matrix $R \in \mathbb{R}^{b\times n}$ from a distribution Π satisfies (ϵ, δ, m)-Johnson-Lindenstrauss property if*

$$\Pr_{R\sim\Pi}\left[(1-\epsilon)\|g\|_2^2 \leq \|Rg\|_2^2 \leq (1+\epsilon)\|g\|_2^2,\ \forall g \in S\right] \geq 1-\delta.$$

Definition G.3 (Subspace embedding (SE) (Sarlós, 2006), restatement of Definition 3.2)**.** *Given $\epsilon \in (0,1)$, any matrix $A \in \mathbb{R}^{n\times d}$, we say a randomized matrix $R \in \mathbb{R}^{b\times n}$ from a distribution Π satisfies (ϵ, δ, d)-subspace embedding for the column space of A if*

$$\Pr_{R\sim\Pi}\left[(1-\epsilon)\|Ax\|_2^2 \leq \|RAx\|_2^2 \leq (1+\epsilon)\|Ax\|_2^2,\ \forall x \in \mathbb{R}^d\right] \geq 1-\delta.$$

Definition G.4 (Approximate matrix product (AMP) (Sarlós, 2006))**.** *Given $\epsilon, \delta \in (0,1)$, any two matrices A, B each with n rows, we say a randomized matrix $R \in \mathbb{R}^{b\times n}$ from a distribution Π satisfies (ϵ, δ)-approximate matrix product for A and B if*

$$\Pr_{R\sim\Pi}\left[\|A^\top R^\top RB - A^\top B\|_F > \epsilon\|A\|_F\|B\|_F\right] \leq \delta.$$

Remark G.5. *More generally, we can also define a matrix C that satisfies $\|C - A^\top B\|_F \leq \epsilon$ with high probability.*

Definition G.6 (JL moment property (JLM) (Kane & Nelson, 2012))**.** *Given $\epsilon, \delta \in (0,1)$, we say a randomized matrix $R \in \mathbb{R}^{b\times n}$ from a distribution Π satisfies (ϵ, δ)-JL moment property if*

$$\mathop{\mathbf{E}}_{R\sim\Pi}[(\|Rg\|_2^2 - 1)^2] \leq \epsilon^2 \cdot \delta,\ \forall g \in \mathbb{R}^n \text{ such that } \|g\|_2 = 1.$$

Remark G.7. *More generally, we can also define (ϵ, δ, l)-JL moment property for $l \geq 2$ as*

$$\mathop{\mathbf{E}}_{R\sim\Pi}[|\|Rg\|_2^2 - 1|^l] \leq \epsilon^l \cdot \delta,\ \forall g \in \mathbb{R}^n.$$

We compare the coordinate-wise embedding defined in this paper with the other four guarantees in Remark G.8 and G.9. Then we summarize the relations of all the five guarantees in Remark G.10.

The first remark shows that CE3, JL, and SE can be viewed as the same ℓ_2-norm guarantee, but over different number of vectors.

Remark G.8 (ℓ_2-norm guarantee over different number of vectors)**.** *Consider the following ℓ_2-norm guarantee: For any fixed set $S \subseteq \mathbb{R}^n$ of certain type. Let b be a number that depends on $n, \epsilon, \delta, |S|$. There is a distribution Π over $\mathbb{R}^{b\times n}$ such that*

$$\Pr_{R\sim\Pi}\left[(1-\epsilon)\|g\|_2^2 \leq \|Rg\|_2^2 \leq (1+\epsilon)\|g\|_2^2,\ \forall g \in S\right] \geq 1-\delta.$$

CE3 *When S only contains one vector, i.e., $S = \{g\}$ for some $g \in \mathbb{R}^n$, this ℓ_2-norm guarantee is the same as the third property of coordinate-wise embedding. Note that it is equivalent to let S contain $O(1)$ vectors by losing a constant factor in δ.*

JL *When S is a finite set of vectors, i.e., $S = \{g_1, g_2, \cdots, g_m\}$ for $g_1, g_2, \cdots, g_m \in \mathbb{R}^n$, this ℓ_2-norm guarantee is the same as JL guarantee.*

SE *When S is a subspace of $\mathbb{R}^n$ and contains infinite number of vectors, i.e., $S = \{g = Ax | x \in \mathbb{R}^d\}$ for some matrix $A \in \mathbb{R}^{n\times d}$, this ℓ_2-norm guarantee is the same as the subspace embedding guarantee.*

JL *and* **SE** *parts are straightforward. The equivalence of* **CE3** *is as follows. On the one hand, if we know that for any vectors $h, g \in \mathbb{R}^n$, $|\langle Rg, Rh\rangle - \langle g, h\rangle| \leq \epsilon \|g\|_2 \|h\|_2$ is satisfied with probability at least $1-\delta$, then by setting $g = h$, we have $\|Rh\|_2^2 = (1 \pm \epsilon)\|h\|_2^2$. On the other hand, if for any $h, g \in \mathbb{R}^n$, $\|Rv\|_2 = (1 \pm \epsilon)\|v\|_2$ is satisfied for $v = h, g, (h-g)$ with probability at least $1-\delta$, without loss of generality we assume $\|h\|_2 = \|g\|_2 = 1$, then*

$$\begin{aligned}
\langle Rg, Rh\rangle &= \frac{1}{2}(\|Rg\|_2^2 + \|Rh\|_2^2 - \|Rg - Rh\|_2^2) \\
&= \frac{1}{2}\left((1\pm\epsilon)\|g\|_2^2 + (1\pm\epsilon)\|h\|_2^2 - (1\pm\epsilon)\|g-h\|_2^2\right) \\
&= \langle g, h\rangle \pm \frac{\epsilon}{2}(\|g\|_2^2 + \|h\|_2^2 + \|g-h\|_2^2) \\
&= \langle g, h\rangle \pm \epsilon,
\end{aligned}$$

which gives the other guarantee.

The second remark shows that CE2 is equivalent to JLM under the assumption of CE1.

Remark G.9 (CE1 and CE2 implies JLM)**.** *We show that together the first two properties of coordinate-wise embedding with parameter $\alpha = b\epsilon^2\delta$ implies JL moment property with parameters ϵ and δ.*

Let $h = g$ where g is any vector in $\mathbb{R}^n$ that satisfies $\|g\|_2 = 1$, CE2 with parameter $\alpha = b\epsilon^2\delta$ implies JL moment property with parameters ϵ and δ as follows:

$$\mathop{\mathbf{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{\alpha}{b}\|g\|_2^2\|h\|_2^2 \implies \mathop{\mathbf{E}}_{R\sim\Pi}[\|Rg\|_2^4] \leq 1 + \epsilon^2\delta \iff \mathop{\mathbf{E}}_{R\sim\Pi}[(\|Rg\|_2^2 - 1)^2] \leq \epsilon^2\delta,$$

where the first step follows from $h = g$ and $\|g\|_2^2 = 1$, the second step follows from CE1 that $\mathbf{E}_{R\sim\Pi}[R^\top R] = I$, and hence $\mathbf{E}_{R\sim\Pi}[\|Rg\|_2^2] = 1$.

In the third remark we summarize the relations between all five guarantees.

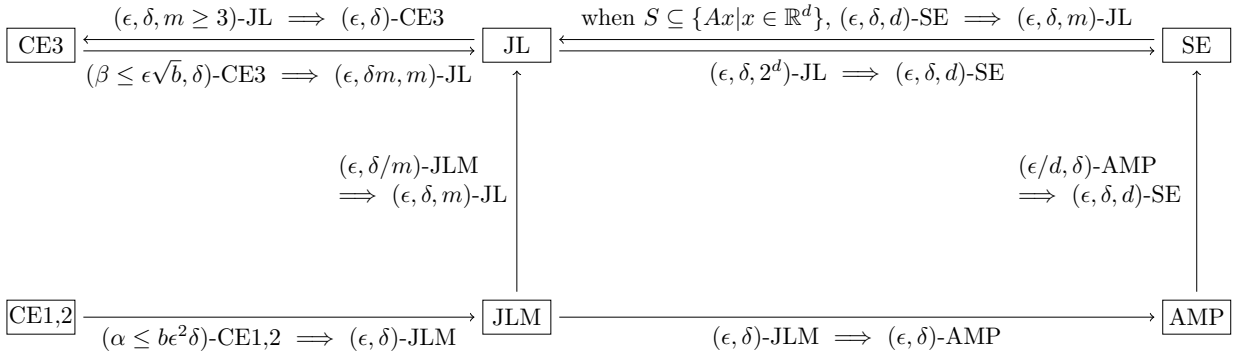


Figure 5: Summary of the relations between five guarantees. See Remark G.10.

Remark G.10 (Summary of the relations between five guarantees)**.** *We summarize the the relations between five guarantees in Figure 5.*

- **JL $\implies$ CE3.** *JL gives bound over a set which implies the CE3 bound over one vector. See Remark G.8.*
- **CE3 $\implies$ JL.** *Directly follows from Union bound over the m vectors of JL. See Remark G.8.*
- **JL $\implies$ SE.** *Suppose a matrix $R \in \mathbb{R}^{b\times n}$ satisfies the JL guarantee with size $b = b(\epsilon, \delta, m)$, where $m \in \mathbb{N}_+$ is the number of vectors, $\epsilon \in (0,1)$ is the error parameter, and $\delta \in (0,1)$ is the probability parameter. Then using the same construction of R but with size $b = b(\epsilon, \delta, 2^d)$, R satisfies the subspace embedding guarantee.*

  A proof sketch is as follows: w.l.o.g. we consider the subspace with unit vectors $S = \{y \in \mathbb{R}^n | y = Ax, x \in \mathbb{R}^d, \|y\|_2 = 1\}$. We choose a ϵ-net of S, which is a set $N \subseteq S$ that satisfies $\forall y \in S$, $\exists w \in N$ such that $\|y - w\|_2 \leq 1/2$. We build

a matrix $R \in \mathbb{R}^{b\times n}$ that satisfies the JL guarantee for the set N. Note that the size $b = b(\epsilon, \delta, 2^d)$ since the ϵ-net N has size $|N| = O(2^d)$. The ϵ-net ensures that any vector $y \in \mathbb{R}^n$ can be decomposed as $y = \sum_{i=0}^{\infty} \frac{1}{2^i} y^{(i)}$ where all $y^{(i)} \in N$. Thus the guarantee $\|Ry\|_2 = (1 \pm \epsilon)\|y\|_2$ is satisfied for all vectors $y \in \mathbb{R}^n$.

For more details see page 12-13 of (Woodruff, 2014).

- **SE $\implies$ JL.** *When the set S of JL is chosen as a subset of subspace $\{Ax | x \in \mathbb{R}^d\}$, SE trivially implies JL. See Remark G.8.*
- **CE1,2 $\implies$ JLM.** *See Remark G.9.*
- **JLM $\implies$ JL.** *Directly follows from Markov's inequality and Union bound over the m vectors of JL.*
- **JLM $\implies$ AMP.** *See Theorem 13 of (Woodruff, 2014). The AMP guarantee of different sketching matrices are usually proved from JLM, e.g., count-sketch matrix (Theorem 14 of (Woodruff, 2014)).*
- **AMP $\implies$ SE.** *See proof of Theorem 9 of (Woodruff, 2014) (page 25).*

Next we summarize the required size of the different sketching matrices to achieve CE, JL, SE, and AMP guarantees in Table 5, 6, 7, 8. We restate the definitions of the different types of sketching matrices $R \in \mathbb{R}^{b\times n}$.

Random Gaussian matrix All entries of R are sampled from $\mathcal{N}(0, 1/b)$ independently.

SRHT matrix (Lu et al., 2013) Let $R = \sqrt{n/b}SHD$, where $S \in \mathbb{R}^{b\times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, $H \in \mathbb{R}^{n\times n}$ is a normalized Walsh-Hadamard matrix, and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.

AMS sketch matrix (Alon et al., 1999) Let $R_{i,j} = h_i(j)$, where $h_1, h_2, \cdots, h_b$ are b random hash functions picking from a random hash family $\mathcal{H} = \{h : [n] \to \{-\frac{1}{\sqrt{b}}, +\frac{1}{\sqrt{b}}\}\}$.

Count-sketch matrix (Charikar et al., 2002) Let $R_{h(i),i} = \sigma(i)$ for all $i \in [n]$ and other entries to zero, where $h : [n] \to [b]$ and $\sigma : [n] \to \{-1, +1\}$ are random hash functions.

Sparse embedding matrix (Nelson & Nguyên, 2013) Let $R_{(j-1)b/s+h(i,j),i} = \sigma(i,j)/\sqrt{s}$ for all $(i,j) \in [n] \times [s]$ and all other entries to zero, where $h : [n] \times [s] \to [b/s]$ and $\sigma : [n] \times [s] \to \{-1, 1\}$ are random hash functions.

Sketching matrix for CE	α	β	Lemma
Random Gaussian	$O(1)$	$O(\log^{1.5}(n/\delta))$	B.13, B.24
SRHT	$O(1)$	$O(\log^{1.5}(n/\delta))$	B.12, B.23
AMS	$O(1)$	$O(\log^{1.5}(n/\delta))$	B.12, B.23
Count-sketch	$O(1)$	$O(\sqrt{b}\log(1/\delta)$ or $O(\frac{1}{\sqrt{\delta}})$	B.14, B.25, B.26
Sparse embedding*	$O(1)$	$O(\sqrt{b/s}\log^{1.5}(n/\delta)$	B.15, B.28
Uniform sampling	$O(n)$	$O(n/\sqrt{b})$	B.16, B.29

Table 5: Summary for different sketching matrices for coordinate embedding. Restatement of the first three columns of Table 1. The sketching matrix R has size $b \times n$. $\alpha, \beta \in \mathbb{R}$ are the two error parameters, $\delta \in (0, 1)$ is the probability parameter, and $s \in \mathbb{N}_+$ is the number of non-zero entries in each column of the sparse embedding matrices.

H. Comparison to classical "sketch and solve"

In this section, we compare our sketching approach to the classical "sketch and solve" approach.

Sketching mat. for JL	b	**Time for** $R \cdot x$	**Reference**
Random Gaussian	$\epsilon^{-2}\log(m/\delta)$	bn	Theorem 4 of (Woodruff, 2014)
SRHT	$\epsilon^{-2}\log(m/\delta)$	$n\log n + n\epsilon^{-2}\log(m/\delta)$	(Ailon & Chazelle, 2006), Page 15 of (Woodruff, 2014)
AMS	$\epsilon^{-2}\log(m/\delta)$	bn	(Achlioptas, 2003), Page 14 of (Woodruff, 2014)
Count-sketch†	$\epsilon^{-2}\delta^{-1}m$	$b+n$	Theorem 14 of (Woodruff, 2014)
Sparse embedding*	$\epsilon^{-2}\log(m/\delta)$	sn	(Kane & Nelson, 2012), Page 14 of (Woodruff, 2014)

Table 6: Summary for different sketching matrices for JL lemma. The sketching matrix R has size $b \times n$. $m \in \mathbb{N}_+$ is the number of vectors, $\epsilon \in (0,1)$ is the error parameter, and $\delta \in (0,1)$ is the probability parameter. * In sparse embedding matrices, each column has $s = \Omega(\epsilon^{-1}\log(m/\delta))$ non-zero entries. † Count-sketch matrices satisfy the (ϵ,δ)-JL moment property when $b = \Omega(\epsilon^{-2}\delta^{-1})$. Then using Markov inequality and union bound over all m vectors, we have $b = \Omega(\epsilon^{-2}\delta^{-1}m)$ suffices for JL guarantee with m vectors.

Sketching mat. for SE	b	**Time for** $R \cdot A$	**Reference**
Random Gaussian	$\epsilon^{-2}(d+\log(1/\delta))$	$\mathcal{T}_{\mathrm{mat}}(b,n,d)$	Thm. 6 of (Woodruff, 2014)
SRHT	$\epsilon^{-2}(\sqrt{d}+\sqrt{\log n})^2\log(d/\delta)$	$nd\log(\epsilon^{-1}d(\log n))$	Thm. 7 of (Woodruff, 2014)
AMS	$\epsilon^{-2}(d+\log(1/\delta))$	$\mathcal{T}_{\mathrm{mat}}(b,n,d)$	Follow from JL guarantee
Count-sketch‡	$\epsilon^{-2}\delta^{-1}d^2$	$\mathrm{nnz}(A)$	Thm. 9 of (Woodruff, 2014)
Sparse embedding*	$\epsilon^{-2}d\cdot\mathrm{poly}\log(d/(\epsilon\delta))$	$\epsilon^{-1}\mathrm{nnz}(A)\,\mathrm{poly}\log(d/(\epsilon\delta))$	Thm. 10 (2) of (Woodruff, 2014)
Sparse embedding†	$\epsilon^{-2}d^{1+\gamma}$	$\epsilon^{-1}\mathrm{nnz}(A)\,\mathrm{poly}(1/\gamma)$	Thm. 10 (1) of (Woodruff, 2014)

Table 7: Summary for different sketching matrices for subspace embedding. The sketching matrix R has size $b \times n$. The vectors are from the column subspace of matrix A with size $n \times d$. $\epsilon \in (0,1)$ is the error parameter, and $\delta \in (0,1)$ is the probability parameter. $\mathcal{T}_{\mathrm{mat}}(a,b,c)$ denotes the running time of fast matrix multiplication of two matrices with size $a \times b$ and $b \times c$. * In the first sparse embedding matrix, each column has $s \geq \epsilon^{-1}\,\mathrm{poly}\log(d/(\epsilon\delta))$ non-zero entries; † In the second sparse embedding matrix, each column has $s \geq \epsilon^{-1}\,\mathrm{poly}(1/\gamma)$ non-zero entries, $\gamma > 0$ is a tunable parameter that gives different trade-offs, and δ can be as small as $1/\,\mathrm{poly}(d)$. ‡ For count-sketch matrices, the subspace embedding guarantee is proved from JL moment property, instead of directly from JL guarantee.

"Sketch and solve" algorithm. First we explain the procedure of the "sketch and solve" approach. Consider the least squares problem as an example. Given $A \in \mathbb{R}^{n\times d}$, and $b \in \mathbb{R}^n$, we try to solve

$$\min_{x\in\mathbb{R}^d} \|Ax - b\|_2,$$

whose solution is $x^* = A^\dagger b = (A^\top A)^{-1}A^\top b$ and it takes $O(nd^{\omega-1} + d^\omega)$ running time to compute.

To speed it up, in a over-constrained case where n is much larger than d, the "sketch and solve" approach chooses a $b \times n$ random matrix R from a certain distribution Π on matrices, where $b \ll n$. Consider the following algorithm for least squares regression:

1. Sample a random matrix $R \sim \Pi$.
2. (Sketch) Compute $R \cdot A$ and $R \cdot b$.
3. (Solve) Output the exact solution x' to the regression problem $\min_{x\in\mathbb{R}^d} \|(RA)x - (Rb)\|_2$.

Analysis. To ensure the accuracy of above approach, they require the sketching matrix S to satisfy the subspace embedding guarantee (Definition G.3).

Theorem H.1 (SE gives approximate regression, Theorem 21 of (Woodruff, 2014))**.** *When $R \in \mathbb{R}^{b\times n}$ used in the "sketch and solve" algorithm satisfies the subspace embedding guarantee with parameters $\epsilon/2$ and δ, then with probability $1-\delta$, the output x' satisfies*

$$\|Ax' - b\|_2 \leq (1+\epsilon)\min_{x\in\mathbb{R}^d} \|Ax - b\|_2.$$

Sketching mat. for AMP	b	Time for $A^\top R^\top RB$	Reference
Random Gaussian	$\epsilon^{-2}\delta^{-1}$	$\mathcal{T}_{\mathrm{mat}}(d_A, b, d_B) + \mathcal{T}_{\mathrm{mat}}(d_A, n, b)$	Lem. 10 of (Boutsidis et al., 2016)(v1)
SRHT	$\epsilon^{-2}\delta^{-1}$	$n \cdot d_A \cdot \log(\epsilon^{-1} d_A \log n) + \mathcal{T}_{\mathrm{mat}}(d_A, b, d_B)$	Lem. 32 of (Clarkson & Woodruff, 2013)
AMS	$\epsilon^{-2}\delta^{-1}$	$\mathcal{T}_{\mathrm{mat}}(d_A, b, d_B) + \mathcal{T}_{\mathrm{mat}}(d_A, n, b)$	Lem. 32 of (Clarkson & Woodruff, 2013)
Count-sketch	$\epsilon^{-2}\delta^{-1}$	$\mathrm{nnz}(A) + \mathrm{nnz}(B) + \mathcal{T}_{\mathrm{mat}}(d_A, b, d_B)$	Thm. 14 of (Woodruff, 2014)
Sparse embedding	$\epsilon^{-2}\delta^{-1}$	$s \cdot \mathrm{nnz}(A) + s \cdot \mathrm{nnz}(B) + \mathcal{T}_{\mathrm{mat}}(d_A, b, d_B)$	Lem. 32 of (Clarkson & Woodruff, 2013)

Table 8: Summary for different sketching matrices for approximate matrix product. The sketching matrix R has size $b \times n$. The matrices A has size $n \times d_A$ and B has size $n \times d_B$, and w.l.o.g. assume $d_A \geq d_B$. $\epsilon \in (0,1)$ is the error parameter, and $\delta \in (0,1)$ is the probability parameter. $\mathcal{T}_{\mathrm{mat}}(a,b,c)$ denotes the running time of fast matrix multiplication of two matrices with size $a \times b$ and $b \times c$. For sparse embedding matrix, s is the number of non-zero entries in its columns. Note that these sketching matrices all have the same size, and this can be easily seen from the fact that they all have the same parameter α for CE2 (Table 1) and CE1,2 $\implies$ JLM $\implies$ AMP.

Proof. Let $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2$. We have

$$\|Ax' - b\|_2 \leq (1+\epsilon/2)\|RAx' - Rb\|_2 \leq (1+\epsilon/2)\|RAx^* - Rb\|_2 \leq (1+\epsilon)\|Ax^* - b\|_2.$$

where the first and the third steps follow from the subspace embedding guarantee, and the second step follows from $x' = \arg\min_{x \in \mathbb{R}^d} \|RAx - Rb\|_2$. □

Remark H.2 (Better regression time using AMP)**.** *We remark that by using approximate matrix product (AMP) guarantee, sometimes the size of the sketching matrix can be further reduced for the "sketch and solve" algorithm.*

Let $R \in \mathbb{R}^{b \times n}$ be a sketching matrix sampled from distribution Π. We use $b_{\mathrm{SE}}(\epsilon, \delta, d)$ to denote the minimum size of R to achieve (ϵ, δ, d)-subspace embedding, and we use $b_{\mathrm{AMP}}(\epsilon, \delta)$ to denote the minimum size of R to achieve (ϵ, δ)-approximate matrix product.

The previous theorem showed that we can solve $(1+\epsilon)$-approximate linear regression with probability $1-\delta$ using sketching matrices with size

$$b \geq b_{\mathrm{SE}}(\epsilon, \delta, d).$$

In fact, it suffices with size

$$b \geq b_{\mathrm{SE}}(1/2, \delta, d) + b_{\mathrm{AMP}}(\sqrt{\epsilon/d}, \delta).$$

For example, for count-sketch matrices, $b_{\mathrm{SE}}(\epsilon, \delta, d) = \epsilon^{-2}\delta^{-1}d^2$ (Table 7) and $b_{\mathrm{AMP}}(\epsilon, \delta) = \epsilon^{-2}\delta^{-1}$ (Table 8). Let $\delta = 0.01$. Only using SE guarantee, we need sketch size $b \geq \epsilon^{-2}d^2$. But using SE guarantee together with AMP guarantee, we can reduce the sketch size to $b \geq \epsilon^{-1}d^2$.

For more details see Theorem 23 of (Woodruff, 2014).

Comparison with our algorithm. Therefore, instead of sketching before solving the problem, we open up the iterations of the classical central path method and apply sketching inside each iteration. Our "iterate and sketch" approach in this work differs from the classical "sketch and solve" approach.