A. Omitted Proofs of Useful Inequalities

A.1. Proof of Proposition 2

Proof. Let $h(x) = (1+2x)\ln(1+x) - x$. Since $h(0) = 0$, it suffices to show that $h'(x) > 0$. We calculate that

$$h'(x) = \frac{x}{1+x} + 2\ln(1+x).$$

Since $h'(0) = 0$, it suffices to show that $h''(x) > 0$. This can be readily verified by calculating that

$$h''(x) = \frac{3+2x}{(1+x)^2} > 0. \qquad \square$$

A.2. Proof of Proposition 4

Proof. Let $f(x,y) = \ln^2(1+x) + \ln^2(1+y) - \ln^2(1+\sqrt{x^2+y^2})$. It suffices to show that $f(x,y) \geq 0$. The inequality is clearly true when $x = 0$ or $y = 0$. Note that

$$\frac{\partial f}{\partial x} = 2\left(\frac{\log(1+x)}{1+x} - \frac{x\ln(1+\sqrt{x^2+y^2})}{x^2+y^2+\sqrt{x^2+y^2}}\right)$$
$$\frac{\partial f}{\partial y} = 2\left(\frac{\log(1+y)}{1+y} - \frac{y\ln(1+\sqrt{x^2+y^2})}{x^2+y^2+\sqrt{x^2+y^2}}\right)$$

Assuming $x, y > 0$, $\partial f/\partial x = \partial f/\partial y = 0$ implies that

$$\frac{\log(1+x)}{x(1+x)} = \frac{\log(1+x)}{y(1+y)}.$$

It is easy to verify that $\log(1+x))/(x(1+x))$ is decreasing w.r.t. x (checking the derivative and using Proposition 4), so we must have $x = y$. Now, let

$$h(x) = \frac{\partial f}{\partial x}(x,x) = \frac{2\ln(1+x)}{1+x} - \frac{\sqrt{2}\ln(1+\sqrt{2}x)}{1+\sqrt{2}x}.$$

We shall show that $h(x) > 0$ for all $x > 0$. This will imply that $f(x,y)$ has no local minimum or maximum when $x, y > 0$ and so it is easy to see that $f(x,y)$ attains the minimum at its boundary $x = 0$ or $y = 0$, yielding that $f(x,y) \geq 0$ for all $x, y \geq 0$.

To see that $h(x) > 0$, let

$$g(a) = \frac{\ln(1+ax)}{a(1+ax)}.$$

We calculate

$$g'(a) = \frac{ax - (1+2ax)\ln(1+ax)}{a^2(1+ax)^2}.$$

It follows from Proposition 2 that $g'(a) < 0$. Hence $g(a)$ is decreasing w.r.t. a and $g(\sqrt{2}) < g(1)$, which is exactly $\frac{1}{\sqrt{2}}h(x) > 0$. $\square$

A.3. Proof of Lemma 5

Proof. It is clear that the base of the logarithm does not matter and we assume that the base is e. Let $Z = \sum_i \epsilon_i a_i$ and $\sigma^2 = \sum a_i^2$. Then $\mathbb{E}\, Z^2 = \sigma^2$ and $\mathbb{E}\,|Z| \leq (\mathbb{E}\,|Z|^2)^{1/2} = \sigma$. Let $g(x) = \ln(1+x)$ and

$$Z_1 = \begin{cases} |Z|, & |Z| \geq e-1; \\ 0, & \text{otherwise,} \end{cases} \quad Z_2 = \begin{cases} 0, & |Z| \geq e-1; \\ |Z|, & \text{otherwise.} \end{cases}$$

Then $|Z| = Z_1 + Z_2$ and

$$\mathbb{E}\, g(|Z|)^2 = \mathbb{E}(g(Z_1+Z_2))^2 \le \mathbb{E}(g(Z_1)+g(Z_2))^2 \le \mathbb{E}\, 2(g(Z_1)^2 + g(Z_2)^2),$$

where the first inequality follows from Proposition 3. For the first term, we define $h(x) = g(x) \cdot \mathbf{1}_{\{x \ge e-1\}}$. Then $h(x)^2$ is concave on $[0,\infty)$. Hence

$$\mathbb{E}\, g(Z_1)^2 = \mathbb{E}\, h(Z_1)^2 = \mathbb{E}\, h(|Z|)^2 \le h(\mathbb{E}\,|Z|)^2 \le h(\sigma)^2 \le g(\sigma)^2.$$

Next we upper bound the second term. The first case is $\sigma \le e-1$. Since $\mathbb{E}\, Z^4 \le 3\sigma^4$, it holds that $\Pr\{Z_2 \ge t\sigma\} \le \Pr\{|Z| \ge t\sigma\} \le 3/t^4$. Then

$$\begin{aligned}
\mathbb{E}\, g(Z_2)^2 &\le \mathbb{E}\, g(e-1) g(Z_2) \\
&= \mathbb{E}\, g(Z_2) \\
&= \int_0^{e-1} g(x) \Pr\{Z_2 \ge x\} dx \\
&= \sigma \int_0^{(e-1)/\sigma} g(t\sigma) \Pr\{Z_2 \ge t\sigma\} dt \\
&= \sigma^2 \int_0^{(e-1)/\sigma} g(t) \Pr\{Z_2 \ge t\sigma\} dt \quad \text{(by Proposition 3)} \\
&\le \sigma^2 \left(\int_0^1 g(t) dt + 3 \int_1^{(e-1)/\sigma} \frac{g(t)}{t^4} dt \right) \\
&\le C_1 \sigma^2 \\
&\le C_1 (e-1)^2 g(\sigma)^2,
\end{aligned}$$

where $C_1 > 0$ is an absolute constant and the last inequality follows from the fact that $g(x) \ge x/(e-1)$ on $[0, e-1]$. The second case is $\sigma > e-1$. In this case,

$$\mathbb{E}\, g(Z_2)^2 \le 1 \le g(\sigma)^2.$$

Therefore, we conclude that

$$\mathbb{E}\, g(|Z|)^2 \le 2(1 + C_1(e-1)^2) g(\sigma)^2 = C_2 g\left(\sqrt{\sum_i a_i^2} \right)^2 \le C_2 \sum_i g(|a_i|)^2,$$

where the last inequality follows from Proposition 4. □

A.4. Proof of Lemma 6

Proof. We first prove the upper bound.

$$\begin{aligned}
\|f(y+z)\|_2^2 &= \sum_i f(y_i+z_i)^2 \\
&\le \sum_i [f(y_i) + f(z_i)]^2 \quad \text{(Proposition 3)} \\
&= \sum_i f(y_i)^2 + \sum_i f(z_i)^2 + \sum_i 2 f(y_i) f(z_i) \\
&\le \sum_i f(y_i)^2 + \xi^2 \sum_i f(y_i)^2 + 2 \sqrt{\sum_i f(y_i)^2} \sqrt{\sum_i f(z_i)^2} \quad \text{(Cauchy-Schwarz)} \\
&\le \left(\xi^2 + 2\xi + 1\right) \|f(y)\|_2^2 \\
&\le (1+3\xi) \|f(y)\|_2^2 . \quad (\text{since } \xi < 1)
\end{aligned}$$

Next we prove the lower bound. Let $I = \{i : y_i z_i \le 0\}$, $J_1 = \{i \in I : |y_i| \le |z_i|\}$ and $J_2 = \{i \in I : |z_i| < |y_i| \le \zeta^{-1}|z_i|\}$ for some $\zeta < 1$ to be determined. Then

$$\begin{aligned}\|f(y+z)\|_2^2 &= \sum_{i\in J_1} f(y_i+z_i)^2 + \sum_{i\in J_2} f(y_i+z_i)^2 + \sum_{i\in I\setminus(J_1\cup J_2)} f(y_i+z_i)^2 + \sum_{i\notin I} f(y_i+z_i)^2 \\ &\ge \sum_{i\in I\setminus(J_1\cup J_2)} f(y_i+z_i)^2 + \sum_{i\notin I} f(y_i)^2.\end{aligned}$$

When $i \in I \setminus (J_1 \cup J_2)$, we have $|z_i| \le \zeta|y_i|$. It then follows that

$$\log(|y_i+z_i|+1) \ge \log((1-\zeta)|y_i|+1) \ge (1-\zeta)\log(|y_i|+1),$$

where, for the last inequality, one can easily verify that $h_\epsilon(x) = \frac{\log(1+(1-\epsilon)x)}{\log(1+x)}$ is increasing on $[0,\infty)$ and $\lim_{x\to 0^+} h_\epsilon(x) = 1-\epsilon$. Hence

$$\sum_i f(y_i+z_i)^2 \ge (1-\zeta)^2 \sum_{i\in I\setminus(J_1\cup J_2)} f(y_i)^2 + \sum_{i\notin I} f(y_i)^2 \ge (1-\zeta)^2 \sum_{i\notin J_1\cup J_2} f(y_i)^2.$$

Now, note that

$$\sum_{i\in J_1} f(y_i)^2 \le \sum_{i\in J_1} f(z_i)^2 \le \|f(z)\|_2^2 \le \xi^2 \|f(y)\|_2^2$$

and (using Proposition 3)

$$\sum_{i\in J_2} f(y_i)^2 \le \zeta^{-2} \sum_{i\in J_1} f(z_i)^2 \le \zeta^{-2}\|f(z)\|_2^2 \le (\zeta^{-1}\xi)^2 \|f(y)\|_2^2.$$

It follows that

$$\begin{aligned}\sum_i f(y_i+z_i)^2 &\ge (1-\zeta)^2 \left(\|f(y)\|_2^2 - \xi^2\|f(y)\|_2^2 - (\zeta^{-1}\xi)^2\|f(y)\|_2^2\right) \\ &= (1-\zeta)^2(1-\xi^2-(\zeta^{-1}\xi)^2)\|f(y)\|_2^2.\end{aligned}$$

Choosing $\zeta = (\xi^2/(1-\xi^2))^{1/3}$ maximizes the right-hand side, yielding

$$\|f(y+z)\|_2^2 \ge (1-3\xi^{2/3})\|f(y)\|_2^2.$$ □

B. Omitted Proofs from Section 3.1

B.1. Proof of Lemma 7

Proof. Note that $|I_{\alpha\phi}| \le 1/(\alpha\phi)$. Thus, there exists a collision with probability at most

$$\frac{1}{w}\binom{1/(\alpha\phi)}{2} \le \frac{1}{2w\alpha^2\phi^2} \le 0.1,$$

provided that $w \ge 1/(0.2\cdot\alpha^2\phi^2) = 5/(\alpha^2\phi^2)$. □

B.2. Proof of Lemma 8

Proof. Let $v = h(u)$. Since h is pairwise independent, $\Pr\{h(i) = v\} = 1/w$ for all $i \ne w$. Let

$$Z_v = \sum_{i\notin(I_{\alpha\phi}\cup\{u\})} \mathbf{1}_{\{h(i)=v\}}\|f(A_i)\|_2^2.$$

then

$$\mathbb{E}\, Z_v \le \sum_{i\notin I_{\alpha\phi}} \mathbb{E}\,\mathbf{1}_{\{h(i)=v\}}\|f(A_i)\|_2^2 \le \frac{M}{w}.$$

It follows from Lemma 5 that

$$\begin{aligned}\mathop{\mathbb{E}}_{\{\epsilon_i\},h}\left\|f\left(\sum_{i\notin I_{\alpha\phi}}\mathbf{1}_{\{h(i)=v\}}\epsilon_i A_i\right)\right\|_2^2 &\leq \mathop{\mathbb{E}}_h C\sum_{i\notin I_{\alpha\phi}}\left\|f\left(\mathbf{1}_{\{h(i)=v\}}A_i\right)\right\|_2^2\\ &= C\mathop{\mathbb{E}}_h Z_v\\ &\leq C\frac{M}{w},\end{aligned}$$

where we used the fact that $f(0)=0$ and $\mathbf{1}_{\{h(i)=v\}}\in\{0,1\}$ in the second step (the equality). The result follows from Markov's inequality. □

B.3. Obtaining an Overestimate $\widehat{M}$

In this subsection we verify that $g(x)=\ln^2(1+\eta x)$ is slow-jumping, slow-dropping, and predictable, where the three properties are defined in (Braverman et al., 2016).

To show that g is slow-jumping, we shall verify that for any $\alpha>0$, $g(y)\leq\lfloor\frac{y}{x}\rfloor^{2+\alpha}x^\alpha g(x)$ for all $x<y$, whenever y is sufficiently large. (i) When $x\geq y/2$, it suffices to show that $g(y)\leq x^\alpha g(x)$. Since $g(x)$ is increasing, it reduces to showing $g(y)\leq(y/2)^\alpha g(y/2)$. This clearly holds for all large y because one can easily check that $\ln(1+y)\leq 2\ln(1+\frac{y}{2})$ when $y>0$. (ii) When $x<y/2$, we shall show that $g(y)\leq(\frac{y}{x}-1)^{2+\alpha}x^\alpha g(x)$, i.e., $g(y)\leq(\frac{y-x}{x})^2(y-x)^\alpha g(x)$. Since $x<y/2$, we have $y-x\geq y/2$ and thus it suffices to show that $g(y)\leq\frac{1}{4}(\frac{y}{x})^2(\frac{y}{2})^\alpha g(x)$, and for large y that $\frac{g(y)}{y^2}\leq\frac{g(x)}{x^2}$, which can be easily verified. This concludes the proof that g is slow-jumping.

To show that g is slow-dropping, we shall verify that for any $\alpha>0$ it holds that $g(y)\geq g(x)/x^\alpha$ for all $x<y$ whenever y is sufficiently large. This holds obviously because $g(x)$ is increasing.

To show that g is predictable, we shall verify that for any $\gamma\in(0,1)$ and subpolynomial $\epsilon(x)$, it holds that $g(y)\geq x^{-\gamma}g(x)$ for all sufficiently large x and all $y\in[1,x^{1-\gamma}]$ such that $g(x+y)>(1+\epsilon(x))g(x)$. This holds automatically because $g(2x)/g(x)\to 1$ as $x\to\infty$ and thus for any given $\epsilon(x)$,when x is sufficiently large, it would not hold that $g(x+y)>(1+\epsilon(x))g(x)$ for $y\in[1,x]$.

C. Omitted Proofs from Section 3.2

C.1. Proof of Theorem 12

Proof. For notational convenience, let $G=f(A)$. Let S be a random sample of s rows chosen from a distribution that satisfies (1). We can write the i-th sample as G_i+E_i for some error vector E_i. Consider the singular value decomposition of $G=\sum_t\sigma_t u_t v_t^\top$.

For each t, we define a random vector

$$w_t=\frac{1}{s}\sum_{i\in S}\frac{(u_t)_i}{p_i}(G_i+E_i).$$

Note that S in general consists of sampled columns of $f(A)$ with noise. The vectors w_t are clearly in the subspace generated by S. We first compute $\mathbb{E}\,w_t$. We can view w_t as the average of s i.i.d. random variables $X_1,\dots,X_s$, where each X_j has the following distribution:

$$X_j=\frac{(u_t)_i}{p_i}(G_i+E_i)\text{ with probability }p_i,\quad i=1,2,\dots n.$$

Taking expectations,

$$\mathbb{E}\,X_j=\sum_{i=1}^n\frac{(u_t)_i}{p_i}(G_i+E_i)p_i=u_t^\top(G+E)=\sigma_t v_t^\top+u_t^\top E$$

Hence

$$\mathbb{E}\,w_t=\mathbb{E}\,X_j=\sigma_t v_t^\top+u_t^\top E$$

and

$$\|\mathbb{E}\, X_j\|_2^2 = \sigma_t^2 + 2\langle \sigma_t v_t^\top, u_t^\top E\rangle + \left\|u_t^\top E\right\|_2^2 \le \sigma_t^2 + 2\langle \sigma_t v_t^\top, u_t^\top E\rangle + \|E\|_2^2 .$$

We also calculate that

$$\begin{aligned}
\mathbb{E}\,\|X_j\|_2^2 &= \sum_i \frac{(u_t)_i^2}{p_i^2}\,\|G_i + E_i\|_2^2 \cdot p_i \\
&\le \sum_i \frac{(u_t)_i^2}{p_i}(\|G_i\|_2 + \|E_i\|_2)^2 \\
&\le \sum_i (u_t)_i^2 \frac{\|G\|_F^2}{c\,\|G_i\|_2^2}(1+\gamma)^2\,\|G_i\|_2^2 \\
&= \frac{(1+\gamma)^2}{c}\,\|G\|_F^2 ,
\end{aligned}$$

where we used the assumption (1) in the third line and the fact that $\|u_t\|_2 = 1$ in the last line. It follows that

$$\begin{aligned}
\mathbb{E}\,\|w_t\|_2^2 = \mathbb{E}\left\|\frac{1}{s}\sum_j X_j\right\|_2^2 &= \frac{1}{s}\sum_j \mathbb{E}\,\|X_j\|_2^2 + \frac{1}{s^2}\sum_{j\ne\ell}\langle \mathbb{E}\,X_j, \mathbb{E}\,X_\ell\rangle \\
&\le \frac{(1+\gamma)^2}{sc}\,\|G\|_F^2 + \frac{s(s-1)}{s^2}\left(\sigma_t^2 + 2\langle \sigma_t v_t^\top, u_t^\top E\rangle + \|E\|_2^2\right),
\end{aligned}$$

and thus

$$\begin{aligned}
\mathbb{E}\left\|w_t - \sigma_t v_t^\top\right\|_2^2 &= \mathbb{E}\,\|w_t\|_2^2 - 2\langle \mathbb{E}\,w_t, \sigma_t v_t^\top\rangle + \sigma_t^2 \\
&\le \frac{(1+\gamma)^2}{sc}\,\|G\|_F^2 + \sigma_t^2 + 2\langle \sigma_t v_t^\top, u_t^\top E\rangle + \|E\|_2^2 - 2\sigma_t^2 - 2\langle u_t^T E, \sigma_t v_t^\top\rangle + \sigma_t^2 \qquad (2)\\
&= \frac{(1+\gamma)^2}{sc}\,\|G\|_F^2 .
\end{aligned}$$

If w_t were exactly equal to $\sigma_t v_t^\top$ (instead of just in expectation), we would have

$$G\sum_{t=1}^{k} v_t v_t^\top = G\sum_{t=1}^{k} w_t^\top w_t,$$

which would be sufficient to prove the theorem. We wish to carry this out approximately. To this end, define $\hat{y}_t = \frac{1}{\sigma_t} w_t^\top$ for $t = 1, 2, \ldots, s$ and let $V_1 = \operatorname{span}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_s) \subseteq V$. Let $y_1, y_2, \ldots, y_n$ be an orthonormal basis of $\mathbb{R}^n$ with $V_1 = \operatorname{span}(y_1, y_2, \ldots, y_l)$, where $l = \dim(V_1)$. Let

$$B = \sum_{t=1}^{l} G y_t y_t^\top \quad \text{and} \quad \hat{B} = \sum_{t=1}^{k} G v_t \hat{y}_t^\top .$$

The matrix B will be our candidate approximation to G in the span of S. We shall bound its error using $\hat{B}$. Note that for any $i \le k$ and $j > l$, we have $(\hat{y}_i)^\top y_j = 0$. Thus,

$$\|G - B\|_F^2 = \sum_{i=1}^{n}\left\|(G-B)y^{(i)}\right\|_2^2 = \sum_{i=l+1}^{n}\left\|Gy^{(i)}\right\|_2^2 = \sum_{i=l+1}^{n}\left\|(G-\hat{B})y^{(i)}\right\|_2^2 \le \left\|G - \hat{B}\right\|_F^2 . \qquad (3)$$

Also,

$$\|G - \hat{B}\|_F^2 = \sum_{i=1}^{n}\left\|u_i^\top (G - \hat{B})\right\|_2^2 = \sum_{i=1}^{k}\left\|\sigma_i v_i^\top - w_i\right\|_2^2 + \sum_{i=k+1}^{n}\sigma_i^2$$

Taking expectations and using (2), we obtain that

$$\mathbb{E}\left\|G-\hat{B}\right\|_F^2 \le \sum_{i=k+1}^{n} \sigma_i^2 + \frac{k(1+\gamma)^2}{sc}\left\|G\right\|_F^2. \tag{4}$$

Note that $\hat{B}$ is of rank at most k and D_k is the best rank-k approximation to G. We have

$$\left\|G-\hat{B}\right\|_F^2 \ge \left\|G-D_k\right\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

Thus $\|G-\hat{B}\|_F^2 - \|G-D_k\|_F^2$ is a non-negative random variable. It follows from (4) that

$$\Pr\left\{\left\|G-\hat{B}\right\|_F^2 - \left\|G-D_k\right\|_F^2 \ge \frac{10k(1+\gamma)^2}{sc}\left\|G\right\|_F^2\right\} \le \frac{1}{10}.$$

The result follows from (3) and the fact that $\left\|E\right\|_F^2 \le \gamma\left\|G\right\|_F^2$. □

C.2. Proof of Corollary 13

Proof. First, it follows from a Chernoff bound and a union bound that we can guarantee with probability at least 0.9 that all samples have the form $f(A_i)+E_i$ with small $\|E_i\|_2$. Then, it follows from another Chernoff bound that with probability at least 0.9, it holds that there are $s/2$ samples from A'. We apply Theorem 12 to A' and $s/2$ and obtain that

$$\left\|f(A') - f(A')\sum_j y_j y_j^\top\right\|_F^2 \le \min_{D:\mathrm{rank}(D)\le k}\left\|f(A')-D\right\|_F^2 + \frac{30k}{sc}\left\|f(A')\right\|_F^2.$$

Suppose that A'' is the submatrix of A which consists of the rows of A that are not in A'. Then $f(A)$ is the (interlacing) concatenation of $f(A')$ and $f(A'')$. Since $\|f(A'')\|_F^2 \le \epsilon\|f(A)\|_F^2$ and $y_1,\dots,y_k$ remains valid if we add more samples,

$$\begin{aligned}
&\left\|f(A) - f(A)\sum_j y_j y_j^\top\right\|_F^2 \\
&= \left\|f(A') - f(A')\sum_j y_j y_j^\top\right\|_F^2 + \left\|f(A'') - f(A'')\sum_j y_j y_j^\top\right\|_F^2 \\
&\le \min_{D:\mathrm{rank}(D)\le k}\left\|f(A')-D\right\|_F^2 + \frac{30k}{sc}\left\|f(A)\right\|_F^2 + \left\|f(A'')\right\|_F^2 \\
&\le \min_{D:\mathrm{rank}(D)\le k}\left\|f(A)-D\right\|_F^2 + \left(\frac{30k}{sc}+\epsilon\right)\left\|f(A)\right\|_F^2.
\end{aligned}$$

The overall failure probability combines that of Theorem 10, Theorem 12 and the events at the beginning of this proof.

For the second result, take $s = O(k/\epsilon)$ and rescale ϵ. □

D. Proof of Theorem 16

By Theorem 10, for every $i \in [s]$, there exists $j(i)$ such that $h_i = (f(A)_{j(i)}, b_{j(i)}) + F_{j(i)}$, where $F_i = \frac{E_i}{\sqrt{sp_i}}$. We define a new matrix S such that in the i-th row of S, $S_{i,j(i)} = \frac{1}{\sqrt{sp_{j(i)}}}$ and the other entries are zero. By Theorem 15, we have that the row-sampling probability we use is a $(1\pm O(\epsilon))$ approximation to the true sampling probability. Therefore, we define matrix $\hat{S}$ such that in the i-th row of $\hat{S}$, $\hat{S}_{i,j(i)} = \frac{1}{\sqrt{s\hat{p}_{j(i)}}}$ and the other entries are zero, and matrix $\hat{F}$ is such that $\hat{F}_i = \frac{E_i}{\sqrt{s\hat{p}_i}}$. Then, we find that $\hat{S}\begin{pmatrix} f(A) & b\end{pmatrix} + \hat{F} = T$.

Proof. For notational convenience, we let $G = f(A)$ with singular value decomposition $G = U\Sigma V^\top$. We shall show that $\left\|I_d - (\hat{S}U)^\top(\hat{S}U)\right\|_2$ is small, for which we first show $\left\|I_d - (SU)^\top(SU)\right\|_2$ is small.

Let $X_i = I_d - Y_i^T Y_i$ and $Y_i = \frac{U_{j(i)}}{\sqrt{p_{j(i)}}}$, where U_t is the t-th row of U, which means that the $j(i)$-th row of M is chosen in the i-th trial. Since

$$\mathbb{E}(X_i) = I_d - \mathbb{E}(Y_i^T Y_i) = I_d - \sum_{t=1}^{n} p_t \frac{U_t^T}{\sqrt{p_t}} \frac{U_t}{\sqrt{p_t}} = I_d - \sum_{t=1}^{n} U_t^T U_t = 0,$$

we can apply Lemma 1 to $X_1, \ldots, X_s$, for which we need to upper bound $\|X_i\|_2$ and $\|\mathbb{E}(X_i^2)\|_2$.

We first bound $\|X_i\|_2$.

$$\|X_i\|_2 = \left\|I_d - Y_i^\top Y_i\right\|_2 \le 1 + \frac{\left\|U_i^\top U_i\right\|_2}{p_i} \le 1 + \frac{\|U_i\|_2^2}{c\,\|G_i\|_2^2}\,\|G\|_F^2 \le 1 + \frac{\sigma_1^2 + \cdots + \sigma_d^2}{c\sigma_d^2} \le 1 + \frac{d\kappa^2}{c},$$

where $\sigma_1 \ge \cdots \ge \sigma_d$ are the singular values of G, and in the penultimate inequality we use the fact that $\|G_i\|_2 = \left\|U_i \Sigma V^T\right\|_2 = \|U_i\Sigma\|_2 \ge \sigma_d \|U_i\|_2$.

Next, we bound $\left\|\mathbb{E}(X_i^2)\right\|_2$. Observe that

$$\begin{aligned}
\mathbb{E}(X_i^2 + I_d) &= I_d + \mathbb{E}(I_d - Y_i^\top Y_i)(I_d - Y_i^\top Y_i) = I_d + \mathbb{E}(I_d - 2Y_i^\top Y_i + Y_i^\top Y_i Y_i^\top Y_i) \\
&= 2I_d - \mathbb{E}(Y_i^\top Y_i) + \mathbb{E}(Y_i^\top Y_i \|Y_i\|_2^2) = \mathbb{E}\left(\frac{\left\|U_{j(i)}\right\|_2^2}{p_{j(i)}} Y_i^\top Y_i\right),
\end{aligned}$$

and thus

$$\left\|\mathbb{E}(X_i^2 + I_d)\right\|_2 = \left\|\mathbb{E}\left(\frac{\left\|U_{j(i)}\right\|_2^2}{p_{j(i)}} Y_i^\top Y_i\right)\right\|_2 \le \left\|\mathbb{E}\left(\frac{\|U_i\|_2^2}{c\,\|G_i\|_2^2}\,\|G\|_F^2\, Y_i^\top Y_i\right)\right\|_2 \le \left\|\mathbb{E}\left(\frac{d\kappa^2}{c} Y_i^\top Y_i\right)\right\|_2 = \frac{d\kappa^2}{c}.$$

It follows immediately from the triangle inequality that

$$\left\|\mathbb{E}\, X_i^2\right\|_2 \le \left\|\mathbb{E}(X_i^2 + I_d)\right\|_2 + \|I_d\|_2 \le \frac{d\kappa^2}{c} + 1.$$

Invoking Lemma 1, for

$$W = \frac{1}{s}\sum_{i=1}^{s} X_i = I_d - \frac{1}{s}\sum_{i=1}^{s} Y_i^\top Y_i = I_d - (SU)^\top(SU),$$

and $\rho = \sigma^2 = 1 + d\kappa^2/c$, we have that

$$\Pr\left\{\left\|I_d - (SU)^\top(SU)\right\|_2 > \epsilon\right\} \le 2d\exp\left(-\frac{\epsilon^2 s}{\sigma^2 + \rho\epsilon/3}\right) \le 2d\exp\left(-\frac{\epsilon^2 s}{2d\kappa^2/c}\right) \le \delta$$

by our choice of s. Equivalently, with probability at least $1-\delta$, it holds that $\left\|I_d - (SU)^\top(SU)\right\|_2 \le \epsilon$, which implies that $\|SGx\|_2 = (1\pm\epsilon)\|Gx\|_2$ for all $x \in \mathbb{R}^d$. We condition on this event in the rest of the proof.

Second, we show that the error between $\left\|I_d - (SU)^\top(SU)\right\|_2$ and $\left\|I_d - (\hat{S}U)^\top(\hat{S}U)\right\|_2$ is small.

$$\begin{aligned}
\left\|I_d - (\hat{S}U)^\top(\hat{S}U)\right\|_2 &\le \left\|I_d - (SU)^\top(SU)\right\|_2 + \left\|(\hat{S}U)^\top(\hat{S}U) - (SU)^\top(SU)\right\|_2 \\
&\le \epsilon + \left\|(\hat{S}U)^\top(\hat{S}U) - (SU)^\top(SU)\right\|_2.
\end{aligned}$$

Observe that $(\hat{S}U)^\top(\hat{S}U) = \sum_{i=1}^{s} \frac{U_{j(i)}^\top U_{j(i)}}{s\hat{p}_{j(i)}} = \sum_{i=1}^{s} \frac{U_{j(i)}^\top U_{j(i)}}{(1\pm O(\epsilon))sp_{j(i)}} = \frac{(SU)^\top(SU)}{1\pm O(\epsilon)}$ and thus

$$\left\|(\hat{S}U)^\top(\hat{S}U) - (SU)^\top(SU)\right\|_2 = O(\epsilon)\left\|(SU)^\top(SU)\right\|_2.$$

We have proved that $\left\|I_d - (SU)^\top (SU)\right\|_2 \le \epsilon$, so we have $\left\|I_d - (\hat{S}U)^\top (\hat{S}U)\right\|_2 \le \epsilon + O(\epsilon)(1+\epsilon) = O(\epsilon)$. By rescaling ϵ', we can assume that $\left\|I_d - (\hat{S}U)^\top (\hat{S}U)\right\|_2 \le \epsilon$.

Now consider the subspace spanned by the columns of M together with b. For any vector $y = Gx - b$, $\left\|\hat{S}y\right\|_2 = (1 \pm \epsilon) \left\|y\right\|_2$. Recall that we have defined $\hat{F}_i = \frac{E_i}{\sqrt{s\hat{p}_i}}$, where $\hat{F}_i$ and E_i are the corresponding i-th row of F and E. Let $\hat{F}^{(1)}$ be the first d columns of $\hat{F}$ and $\hat{F}^{(2)}$ be the last column of $\hat{F}$. Hence, the original linear regression problem can be written as $\min \left\|(\hat{S}G + \hat{F}^{(1)})x - (\hat{S}b + \hat{F}^{(2)})\right\|_2$.

Note that $\tilde{x} = \arg\min_x \left\|(\hat{S}G + \hat{F}^{(1)})x - (\hat{S}b + \hat{F}^{(2)})\right\|_2$ satisfies

$$\begin{aligned}
\min_{\tilde{x}} \left\|(\hat{S}G + \hat{F}^{(1)})\tilde{x} - (\hat{S}b + \hat{F}^{(2)})\right\|_2 &\le \left\|(\hat{S}G + \hat{F}^{(1)})x^* - (\hat{S}b + \hat{F}^{(2)})\right\|_2 \\
&\le \left\|\hat{S}(Gx^* - b)\right\|_2 + \left\|\hat{F}^{(1)}x^* - \hat{F}^{(2)}\right\|_2 \\
&\le (1+\epsilon) \left\|Gx^* - b\right\|_2 + \left\|\hat{F}\right\|_2 \sqrt{\left\|x^*\right\|_2^2 + 1},
\end{aligned}$$

where the third inequality holds because $\hat{S}$ is a subspace embedding for the column space of G together with b and $x^* = \arg\min_{x \in \mathbb{R}^d} \left\|Gx - b\right\|_2$.

Now, consider the upper bound on $\left\|\hat{F}\right\|_2$. Since

$$\left\|\hat{F}_i\right\|_2^2 = \frac{\left\|E_i\right\|_2^2}{s\hat{p}_i} \le \gamma^2 \frac{\left\|G_i\right\|_2^2 + |b_i|^2}{sc(\left\|G_i\right\|_2^2 + |b_i|^2)} (\left\|G\right\|_F^2 + \left\|b\right\|_2^2) \le \frac{\gamma^2}{sc} (\left\|G\right\|_F^2 + \left\|b\right\|_2^2)$$

and

$$\left\|x^*\right\|_2 = \left\|G^\dagger b\right\|_2 \le \frac{\left\|b\right\|_2}{\sigma_{\min}(G)},$$

we have that

$$\begin{aligned}
\min_{\tilde{x}} \left\|(\hat{S}G + \hat{F}^{(1)})\tilde{x} - (\hat{S}b + \hat{F}^{(2)})\right\|_2 &\le (1+\epsilon) \left\|Gx^* - b\right\|_2 + \left\|\hat{F}\right\|_2 \sqrt{\left\|x^*\right\|_2^2 + 1} \\
&\le (1+\epsilon) \left\|Gx^* - b\right\|_2 + \frac{\gamma}{\sqrt{c}} \sqrt{\left\|G\right\|_F^2 + \left\|b\right\|_2^2} \cdot \sqrt{\frac{\left\|b\right\|_2^2}{\sigma_{\min}^2(G)} + 1} \\
&\le (1+\epsilon) \left\|Gx^* - b\right\|_2 + \frac{\gamma}{\sqrt{c}} \left(\sqrt{\left\|G\right\|_F^2 + \left\|b\right\|_2^2} + \sqrt{d + \frac{\left\|b\right\|_2^2}{\left\|G\right\|_2^2}} \kappa \left\|b\right\|_2 \right).
\end{aligned}$$

By our assumption, $c = 1 - O(\epsilon)$ and $\gamma = O(\epsilon)$. Rescaling ϵ gives the claimed bound, completing the proof of Theorem 16. □