1 Proof of main theorem

The main result of this section is

Theorem 1. *There exists an absolute constant $C > 0$ such that for every $\delta \in (0, 1)$, every integer $1 \leq m \leq n^4$ and every matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, if $B \geq \frac{1}{\delta}C(\log n)^4 \cdot d^2 \cdot m^{1/2}$, $S \in \mathbb{R}^{B\times n}$ is a random CountSketch matrix, and $G \in \mathbb{R}^{m\times B}$ and $\tilde{G} \in \mathbb{R}^{m\times n}$ are matrices of i.i.d. unit variance Gaussians, then the total variation distance between the joint distribution GSU and $\tilde{G}U$ is less than δ.*

Remark 2. Note that we restrict the range of values of m in Theorem 1 to $[1 : n^4]$. This is because if $m > n^4$, the theorem requires $B \gg \frac{1}{\delta}n^2$, at which point the CountSketch matrix S becomes an isometry of $\mathbb{R}^n$ with high probability and the theorem follows immediately. At the same time restricting m to be bounded by a small polynomial of n simplifies the proof of Theorem 1 notationally.

Recall that a CountSketch matrix $S \in \mathbb{R}^{B\times n}$ is a matrix all of whose columns have exactly one nonzero in a random location, and the value of the nonzero element is independently chosen to be -1 or $+1$. All random choices are made independently. Throughout this section we denote the number of rows in the CountSketch matrix by B. Note that the matrix S is a random variable. Let G denote an $m \times B$ matrix of independent Gaussians. For an $n \times d$ matrix U with orthonormal columns let $q : \mathbb{R}^d \to \mathbb{R}_+$ denote the p.d.f. of the random variable G_1SU, where G_1 is the first row of G (all rows have the same distribution and are independent). We note that G_1SU is a mixture of Gaussians. Indeed, for any fixed S the distribution of G_1SU is normal with covariance matrix $(G_1SU)^T(G_1SU) = U^TS^TSU$. We denote the distribution of G_1SU given S by

$$q_S(x) := \frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^TM^{-1}x}.$$

Throughout this section we use the notation $M := U^TS^TSU$. Note that since S is a random variable, M is as well. With this notation in place we have for any $x \in \mathbb{R}^d$

$$q(x) = \mathbf{E}_S\left[q_S(x)\right]. \tag{1}$$

Let $p : \mathbb{R}^d \to \mathbb{R}_+$ denote the pdf of the isotropic Gaussian distribution, i.e. for all $x \in \mathbb{R}^d$

$$p(x) = \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2}x^Tx}. \tag{2}$$

Before giving a proof of Theorem 1, which is somewhat involved, we give a simple proof of a weaker version of the theorem, where the number of buckets B of our CountSketch matrix is required to be $\approx \frac{1}{\delta}d^2m$ as opposed to $\approx \frac{1}{\delta}d^2\sqrt{m}$:

Theorem 3. *There exists an absolute constant $C > 0$ such that for every $\delta \in (0, 1)$, every integer $m \geq 1$ and every matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns if $B \geq \frac{1}{\delta^2}Cd^2 \cdot m$, $S \in \mathbb{R}^{B\times n}$ is a random CountSketch matrix, and $G \in \mathbb{R}^{m\times B}$ and $\tilde{G} \in \mathbb{R}^{m\times n}$ are matrices of i.i.d. unit variance Gaussians, then the total variation distance between the joint distribution GSU and $\tilde{G}U$ is less than δ.*

We will use the following measures of distance between two distribution in the proof of our main theorem (Theorem 1) as well as the proof of Theorem 3.

Definition 4 (Kullback-Leibler divergence)**.** The Kullback-Leibler (KL) divergence between two random variables P, Q with probability density functions $p(x), q(x) \in \mathbb{R}^d$ is given by $D_{KL}(P||Q) = \int_{\mathbb{R}^d} p(x) \ln \frac{p(x)}{q(x)} dx$

Definition 5 (Total variation distance)**.** The total variation distance between two random variables P, Q with probability density functions $p(x), q(x) \in \mathbb{R}^d$ is given by $D_{TV}(P, Q) = \frac{1}{2}\int_{\mathbb{R}^d} |p(x) - q(x)|dx$.

Theorem 6 (Pinsker's inequality)**.** *For any two random variables P, Q with probability density functions $p(x), q(x) \in \mathbb{R}^d$ one has $D_{TV}(P, Q) \leq \sqrt{\frac{1}{2}D_{KL}(P||Q)}$.*

The proof of Theorem 3 uses the following simple claim.

Claim 7 (KL divergence between multivariate Gaussians)**.** *Let $X \sim N(0, I_d)$ and $Y \sim N(0, \Sigma)$. Then $D_{KL}(X||Y) = \frac{1}{2}\text{Tr}(\Sigma^{-1} - I) + \frac{1}{2}\ln\det\Sigma$.*

Proof. One has

$$\begin{aligned}
D_{KL}(X||Y) &= \mathbf{E}_{X\sim N(0,I_d)}[-\frac{1}{2}X^TX + \frac{1}{2}X^T\Sigma^{-1}X + \frac{1}{2}\ln\det\Sigma] \\
&= \mathbf{E}_{X\sim N(0,I_d)}[\frac{1}{2}X^T(\Sigma^{-1} - I)X + \frac{1}{2}\ln\det\Sigma] \\
&= \frac{1}{2}\mathbf{E}_{X\sim N(0,I_d)}[\text{Tr}((\Sigma^{-1} - I)XX^T)] + \frac{1}{2}\ln\det\Sigma \\
&= \frac{1}{2}\text{Tr}(\Sigma^{-1} - I) + \frac{1}{2}\ln\det\Sigma,
\end{aligned}$$

where we used the fact that for a vector X of independent Gaussians of unit variance one has $\mathbf{E}_X[X^TAX] = \text{Tr}(A)$ for any symmetric A (by rotational invariance of the Gaussian distribution). □

We can now give

Proof of Theorem 3: One has by Lemma 21, **(1)** (see below; this is a standard property of the CountSketch matrix) that for any $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, and $B \geq 1$, if S is a random CountSketch matrix and $M = U^TS^TSU$, then $\mathbf{E}_S[||M - I||_F^2] = O(d^2/B)$. By Markov's inequality $\mathbf{Pr}_S[||I - M||_F > (2/\delta)\cdot O(d^2/B)] < \delta/2$. Let $\mathcal{E}$ denote the event that $||I - M||_F \leq (2/\delta)\cdot O(d^2/B)$. We condition on $\mathcal{E}$ in what follows. Since $B \geq \frac{1}{\delta^3}Cd^2m$ for a sufficiently large absolute constant $C > 1$, we have, conditioned on $\mathcal{E}$, that

$$||I - M||_F^2 \leq (2/\delta)\cdot O(d^2/B) = (2/\delta)\cdot\delta^3/(Cm) \leq 2\delta^2/(Cm). \tag{3}$$

Note that in particular we have $||I - M|| \leq ||I - M||_F < 1/2$ conditioned on $\mathcal{E}$ as long as $C > 1$ is larger than an absolute constant.

By Claim 7 we have $D_{KL}(X||Y) = \frac{1}{2}\text{Tr}(I - \Sigma^{-1}) + \frac{1}{2}\ln\det\Sigma$. We now use Taylor expansions of matrix inverse and $\log\det$ provided by Claim 9 and Claim 10 (see below) to obtain

$$\begin{aligned}
D_{KL}(X||Y) &= \frac{1}{2}\text{Tr}(M^{-1} - I) + \frac{1}{2}\ln\det M \\
&= \frac{1}{2}\text{Tr}\left(\sum_{k\geq 1}(I - M)^k\right) + \frac{1}{2}\sum_{k\geq 1}\left(-\text{Tr}((I - M)^k)/k\right) \\
&= \frac{1}{2}\text{Tr}\left(\sum_{k\geq 2}(I - M)^k\right) + \frac{1}{2}\sum_{k\geq 2}\left(-\text{Tr}((I - M)^k)/k\right) \\
&= O(\text{Tr}((I - M)^2)) \qquad \text{(since } ||I - M||_2 \leq ||I - M||_F < 1/2) \\
&= O(||I - M||_F^2) \\
&= O(2\delta^2/(Cm)) \qquad \text{(by (3))} \\
&\leq (\delta/4)^2/m
\end{aligned} \tag{4}$$

as long as $C > 1$ is larger than an absolute constant. This shows that for every $S \in \mathcal{E}$ one has $D_{KL}(p||q_S) \leq (\delta/4)^2/m$, and thus $D_{KL}(p||\tilde{q}|\mathcal{E})) \leq (\delta/4)^2/m$, where we let $\tilde{q}(x) := \mathbf{E}_S[q_S(x)|\mathcal{E}]$.

We now observe that the vectors $(G_iSU)_{i=1}^m$ and $(\tilde{G}_iU)_{i=1}^m$ are vectors of independent samples from distributions $q(x)$ and $p(x)$ respectively. We denote the corresponding product distributions by q^m and p^m. Since the good event $\mathcal{E}$ constructed above occurs with probability at least $1 - \delta/2$, it suffices to consider the distributions $\tilde{q}(x)$ and $p(x)$, as

$$D_{TV}(q^m, p^m) \leq \mathbf{Pr}[\bar{\mathcal{E}}] + D_{TV}(q^m, p^m|\mathcal{E}) = \mathbf{Pr}[\bar{\mathcal{E}}] + D_{TV}(\tilde{q}^m, p^m), \tag{5}$$

where $D_{TV}(q^m, p^m|\mathcal{E}) = D_{TV}(\tilde{q}^m, p^m)$ stands for the total variation distance between the distribution of $(\tilde{G}_i U)_{i=1}^m$ and the distribution of $(G_i SU)_{i=1}^m$ conditioned on $S \in \mathcal{E}$. We can now use the estimate from (4) to get

$$\begin{aligned} D_{TV}(\tilde{q}^m, p^m) &\leq \sqrt{\frac{1}{2} D_{KL}(p^m || \tilde{q}^m)} \quad \text{(by Pinsker's inequality)} \\ &= \sqrt{\frac{m}{2} D_{KL}(p||\tilde{q})} \quad \text{(by additivity of KL divergence over product spaces)} \\ &\leq \sqrt{\frac{m}{2} \cdot (\delta/4)^2/m} \quad \text{(by (4))} \\ &\leq \delta/4. \end{aligned} \tag{6}$$

□

The main source of hardness in proving the stronger result provided by Theorem 1 comes from the fact that unlike the setting of Theorem 3, where most elements in the mixture are close to isotropic Gaussians in KL divergence, in the setting of Theorem 1 most elements of the mixture are too far from isotropic Gaussians to establish our result directly (this can be seen by verifying that the bounds of Theorem 3 on the KL divergence of q_S to p are essentially tight). Thus, the main technical challenge in proving Theorem 1 consists of analyzing the effect of averaging over random CountSketch matrices that is involved in the definition of $q(x)$ in (1). The core technical result behind the proof of Theorem 1 is Lemma 8, stated below. Ideally, we would like a lemma that states that the ratio of the pdfs $q(x)/p(x)$ is very close to 1 for 'typical' values of x (for appropriate definition of a set of 'typical' x). Unfortunately, it is not clear how to achieve this result for the distribution $q(x)$ defined in (1). The problem is that some choices of CountSketch matrices S may lead to degenerate Gaussian distributions that are hard to analyze. For example, when S is not a subspace embedding, the matrix M may even be rank-deficient, and the inverse M^{-1} is then ill-defined. To avoid these issues, we work with an alternative definition. Specifically, instead of averaging the distributions $\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x}$ over all CountSketch matrices, we define a high probability event $\mathcal{E}$ in the space of matrices S (see Lemma 8 for the definition) and reason about the modified distribution $\tilde{q}(x)$ defined as

$$\tilde{q}(x) = \mathbf{E}_S \left[\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \middle| \mathcal{E} \right]. \tag{7}$$

For technical reasons it turns out to be useful to define yet another distribution

$$q'(x) = \mathbf{E}_S \left[\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \in \mathcal{T}(S, U)] \middle| \mathcal{E} \right] + \xi \cdot p(x), \tag{8}$$

where $\xi = \mathbf{E}_S \left[\mathbf{Pr}_{X \sim q_S}[X \notin \mathcal{T}(S, U)] | \mathcal{E} \right] \leq n^{-20}$ and for each $S \in \mathcal{E}$ and U with orthonormal columns the set $\mathcal{T}(S, U)$ (see Definition 12) is an appropriately defined set of $x \in \mathbb{R}^d$ that are 'typical' for S and U. We first note that q' is indeed the p.d.f. of a distribution. First, it is clear that $q'(x) \geq 0$ for all x. Second, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} q'(x)dx &= \int_{\mathbb{R}^d} \mathbf{E}_S\left[\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] \middle| \mathcal{E}\right] + \xi \cdot \int_{\mathbb{R}^d} p(x)dx \\
&= 1 - \int_{\mathbb{R}^d} \mathbf{E}_S\left[\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \notin \mathcal{T}(S,U)] \middle| \mathcal{E}\right] + \xi \\
&= 1 - \mathbf{E}_S\left[\int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \notin \mathcal{T}(S,U)]dx \middle| \mathcal{E}\right] + \xi \\
&= 1 - \mathbf{E}_S\left[\int_{\mathbb{R}^d \setminus \mathcal{T}(S,U)} \frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} dx \middle| \mathcal{E}\right] + \xi \\
&= 1 - \mathbf{E}_S\left[\mathbf{Pr}_{X\sim q_S}[X \notin \mathcal{T}(S,U)] \middle| \mathcal{E}\right] + \xi \\
&= 1, \qquad \text{(by definition of } \xi)
\end{aligned}$$

as required.

As we show below, the total variation distance between q' and $\tilde{q}$ is a small n^{-10}, so working with q' suffices. The main argument of our proof shows that the distribution $q'(x)$ is close to $p(x)$ for 'typical' $x \in \mathbb{R}^d$. Then since q' is close to $\tilde{q}$ and the event $\mathcal{E}$ occurs with high probability, this suffices for a proof of Theorem 1. Formally, the core technical result behind the proof of Theorem 1 is

Lemma 8. *There exists an absolute constant $C > 0$ such that for every $\delta \in (0,1)$ and every matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns if $B \geq \frac{1}{\delta}C(\log n)^4 d^2$ there exists a set $\mathcal{E}$ of CountSketch matrices and a subset $\mathcal{T}^* \subseteq \mathbb{R}^d$ that satisfies $\mathbf{Pr}_{X\sim p}[X \notin \mathcal{T}^*] \leq n^{-10}$ and $\mathbf{Pr}_{X\sim\tilde{q}}[X \notin \mathcal{T}^*] \leq n^{-10}$ such that if $S \in \mathbb{R}^{B\times n}$ is a random CountSketch matrix, then* **(1)** *$\mathbf{Pr}_S[\mathcal{E}] \geq 1 - \delta/3$, and* **(2)** *for all $x \in \mathcal{T}^*$ one has*

$$\left|\frac{q'(x)}{p(x)} - 1\right| \leq O((d^2 \log^4 n)/B) + O(n^{-10}).$$

We now prove Theorem 1 assuming Lemma 8 and Claim 15. After this, we then prove Lemma 8 and Claim 15. We now give

Proof of Theorem 1: The proof relies on the observation that the vectors $(G_i SU)_{i=1}^m$ and $(\tilde{G}_i U)_{i=1}^m$ are vectors of independent samples from distributions $q(x)$ and $p(x)$ respectively. We denote the corresponding product distributions by q^m and p^m. Since the good event $\mathcal{E}$ constructed in Lemma 8 occurs with probability at least $1 - \delta/3$, it suffices to consider the distributions $\tilde{q}(x)$ and $p(x)$, as

$$D_{TV}(q^m, p^m) \leq \mathbf{Pr}[\bar{\mathcal{E}}] + D_{TV}(\tilde{q}^m, p^m|\mathcal{E}), \tag{9}$$

where $D_{TV}(\tilde{q}^m, p^m|\mathcal{E})$ stands for the total variation distance between the distribution of $(\tilde{G}_i U)_{i=1}^m$ and the distribution of $(G_i SU)_{i=1}^m$ conditioned on $S \in \mathcal{E}$. Further, we have by the triangle inequality

$$D_{TV}(\tilde{q}^m, p^m|\mathcal{E}) \leq D_{TV}((q')^m, p^m|\mathcal{E}) + D_{TV}(\tilde{q}^m, (q')^m|\mathcal{E}) \leq D_{TV}((q')^m, p^m|\mathcal{E}) + m \cdot n^{-10}, \tag{10}$$

since $D_{TV}(\tilde{q}^m, (q')^m|\mathcal{E}) \leq m D_{TV}(\tilde{q}, q'|\mathcal{E}) \leq mn^{-10}$, where $D_{TV}(\tilde{q}, q'|\mathcal{E}) \leq n^{-10}$ by Claim 15 below.

We first prove, using Lemma 8, that the KL divergence between $p(x)$ and $q'(x)$ restricted to the set $\mathcal{T}^*$ (whose existence is guaranteed by Lemma 8) is bounded by $O(((d \log n)^2/B)^2)$. Specifically, let

$$p_*(x) := \begin{cases} p(x)/\mathbf{Pr}_{X\sim p}[\mathcal{T}^*] & \text{if } x \in \mathcal{T}^* \\ 0 & \text{o.w.} \end{cases} \tag{11}$$

and

$$q'_*(x) := \begin{cases} q'(x)/\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*] & \text{if } x \in \mathcal{T}^* \\ 0 & \text{o.w.} \end{cases} \tag{12}$$

Since $\mathcal{T}^*$ occurs with probability at least $1 - 1/n^{10}$ under both $\tilde{q}(x)$ and $p(x)$ by Lemma 19, it suffices to bound the total variation distance between the product of m independent copies of $q'_*(x)$ and m independent copies of $p_*(x)$. Specifically,

$$\begin{aligned} D_{TV}((q')^m, p^m|\mathcal{E}) &\le D_{TV}((q'_*)^m, p_*^m|(\mathcal{T}^*)^m) + m\mathbf{Pr}[q'(\mathbf{R}^d \setminus \mathcal{T}^*)] + m\mathbf{Pr}[p(\mathbb{R}^d \setminus \mathcal{T}^*)] \\ &\le D_{TV}((q'_*)^m, p_*^m) + 2mn^{-10}, \qquad \text{(by Lemma 19)} \end{aligned} \tag{13}$$

where we used the fact that q'_* and p_* are supported on $\mathcal{T}^*$. Note that both distributions are still product distributions. By Pinkser's inequality and the product structure we thus get

$$\begin{aligned} D_{TV}((q'_*)^m, p_*^m) &\le \sqrt{\frac{1}{2} D_{KL}((q'_*)^m || p_*^m)} \quad \text{(by Pinsker's inequality)} \\ &= \sqrt{\frac{m}{2} D_{KL}(q'_* || p_*)} \quad \text{(by additivity of KL divergence over product spaces)} \end{aligned} \tag{14}$$

In what follows we bound $D_{KL}(q'_*||p_*)$. By Lemma 8 we have for every $x \in \mathcal{T}^*$ that

$$|q'(x)/p(x) - 1| \le O((d^2 \log^4 n)/B) + O(n^{-10}), \tag{15}$$

so

$$\begin{aligned} |q'_*(x)/p_*(x) - 1| &= \left|(q'(x)/p(x)) \cdot \frac{\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*]}{\mathbf{Pr}_{X\sim p}[\mathcal{T}^*]} - 1\right| = \frac{\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*]}{\mathbf{Pr}_{X\sim p}[\mathcal{T}^*]} \cdot \left|(q'(x)/p(x)) - \frac{\mathbf{Pr}_{X\sim p}[\mathcal{T}^*]}{\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*]}\right| \\ &\le \frac{\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*]}{\mathbf{Pr}_{X\sim p}[\mathcal{T}^*]} \cdot \left(|q'(x)/p(x) - 1| + \left|1 - \frac{\mathbf{Pr}_{X\sim p}[\mathcal{T}^*]}{\mathbf{Pr}_{X\sim q'}[\mathcal{T}^*]}\right|\right) \\ &= (1 + O(n^{-10})) \cdot \left(|q'(x)/p(x) - 1| + O(n^{-10})\right) \\ &= O((d^2 \log^4 n)/B) + O(n^{-10}). \quad \text{(by (15))} \end{aligned}$$

Since $B \ge \frac{1}{\delta} C d^2 \log^4 n$ for a sufficiently large constant $C > 0$ by assumption of the theorem, we get that

$$O((d^2 \log^4 n)/B) + O(n^{-10}) < O(1/C) + O(n^{-10}) < 1/2.$$

We thus get, using the bound $|1/(1+x) - 1| \le 2|x|$ for $|x| \le 1/2$,

$$\begin{aligned} |p_*(x)/q'_*(x) - 1| &= \left|\frac{1}{q'_*(x)/p_*(x)} - 1\right| = \left|\frac{1}{1 + (q'_*(x)/p_*(x) - 1)} - 1\right| \\ &= O\left(|q'_*(x)/p_*(x) - 1|\right) \\ &= O((d^2 \log^4 n)/B) + O(n^{-10}) \end{aligned} \tag{16}$$

We now use the fact that $|\ln(1+x) - x| \le 2x^2$ for all $x \in (-1/10, 1/10)$ to upper bound $D_{KL}(q'_*||p_*)$. Specifically, we have

$$\begin{aligned} D_{KL}(q'_*||p_*) = \mathbf{E}_{X\sim q'_*}[\ln(q'_*(X)/p_*(X))] &\le -\mathbf{E}_{X\sim q'_*}[\ln(p_*(X)/q'_*(X))] \\ &\le -\mathbf{E}_{X\sim q'_*}[(p_*(x)/q'_*(x) - 1) - (p_*(x)/q'_*(x) - 1)^2] \\ &\le -\mathbf{E}_{X\sim q'_*}[p_*(x)/q'_*(x) - 1] + \mathbf{E}_{X\sim q'_*}[(p_*(x)/q'_*(x) - 1)^2] \\ &= -(1 - 1) + \mathbf{E}_{X\sim q'_*}[(p_*(x)/q'_*(x) - 1)^2] \\ &= \mathbf{E}_{X\sim q'_*}[(p_*(x)/q'_*(x) - 1)^2] \\ &= O(((d^2 \log^4 n)/B)^2 + n^{-10}) \qquad \text{(by (16))} \end{aligned} \tag{17}$$

Since $B \ge \frac{1}{\delta} C (\log n)^4 d^2 \cdot m^{1/2}$ for a sufficiently large constant $C > 0$ by assumption of the theorem, substituting the bound of (17) into (14), we get

$$D_{TV}((q'_*)^m, p_*^m) \le \sqrt{\frac{m}{2} D_{KL}(q'_*||p_*)} \le \sqrt{\frac{m}{2} \cdot O(((d^2 \log^4 n)/B)^2 + n^{-10})} \le \sqrt{\frac{m}{2} \cdot \delta^2/(8m)} \le \delta/2.$$

Putting this together with (13), (10) and (9) using the assumption that $m \leq n^4$ gives the result. $\square$

The rest of the section is devoted to proving Lemma 8, i.e. bounding

$$q'(x)/p(x) = \mathbf{E}_S\left[\exp\left(\frac{1}{2}x^Tx - \frac{1}{2}x^TM^{-1}x - \frac{1}{2}\log\det M\right)\cdot \mathbf{I}[x \in \mathcal{T}(S,U)]\middle|\,\mathcal{E}\right] + \xi, \tag{18}$$

where $\xi = \mathbf{E}_S\left[\mathbf{Pr}_{X\sim q_S}[X \notin \mathcal{T}(S,U)]|\,\mathcal{E}\right] \leq n^{-20}$, for 'typical' x sampled from the Gaussian distribution (i.e. $x \in \mathcal{T}^*$ – see formal definition below).

Organization. The rest of this section is organized as follows. We start by defining the set $\mathcal{E}$ of 'nice' CountSketch matrices in section 1.1, and proving that a random CountSketch matrix is likely to be 'nice'. We will in fact define a parameterized set $\mathcal{E}(\gamma)$ in terms of a parameter γ. In section 1.2 we define, for each matrix U (which can be thought of as fixed throughout our analysis) with orthonormal columns and CountSketch matrix S, a set $\mathcal{T}(S,U)$ of $x \in \mathbb{R}^d$ that are 'typical' for S and U. The ratio of pdfs in (18) can be approximated well by a Taylor expansion **for such 'typical'** $x \in \mathcal{T}(S,U)$. These Taylor expansions are developed in section 1.3 and form the basis of our proof. Unfortunately, these Taylor expansions are valid only for $x \in \mathcal{T}(S,U)$, i.e. for x that are 'typical' with respect to a given S. To complete the proof, we need to construct a universal 'typical' set $\mathcal{T}^*(U,\gamma)$ of $x \in \mathbb{R}^d$, again parameterized in terms of a parameter γ, that will allow for approximation via Taylor expansions for **all** $x \in \mathcal{T}^*(U,\gamma)$ **and** $S \in \mathcal{E}(\gamma)$. We construct such a set $\mathcal{T}^*(U,\gamma)$ in section 1.4. Finally, the proof of Lemma 8 is given in section 1.5.

1.1 Typical set $\mathcal{E}$ of CountSketch matrices and its properties

Our analysis of (18) starts by Taylor expanding M^{-1} and $\det M$ around the identity matrix. We now state the Taylor expansions, and the define a (family of) high probability events $\mathcal{E}(\gamma)$ (equivalently, sets of 'typical' CountSketch matrices) such that the Taylor expansions are valid for matrices $M \in \mathcal{E}(\gamma)$ for all sufficiently small γ.[1] The Taylor expansions that we use are given by

Claim 9. *For any matrix M with $||I-M|| < 1/2$ one has $M^{-1} = (I-(I-M))^{-1} = \sum_{k\geq 0}(I-M)^k$.*

Claim 10. *For any matrix M with $||I-M|| < 1/2$ one has $\log\det M = \log\det(I-(I-M)) = \sum_{k\geq 1} -\mathrm{Tr}((I-M)^k)/k$.*

For a parameter $\gamma \in (0,1)$ that we will later set to $1/\mathrm{poly}(\log n)$, define event $\mathcal{E}(\gamma)$ as

$$\mathcal{E}(\gamma) := \left\{||I-M||_F^2 \leq \gamma^2 \quad \text{and} \quad |\mathrm{Tr}(I-M)| \leq \gamma\right\}. \tag{19}$$

The events $\mathcal{E}(\gamma)$ occur with high probability even for fairly small γ as long as B is sufficiently large:

Claim 11. *For any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, any $B \times n$ CountSketch matrix S we have $\mathbf{Pr}[\mathcal{E}(\gamma)] \geq 1 - 3(d/\gamma)^2/B$.*

Proof. By Lemma 21 below, we have $\mathbf{E}_S[||I-M||_F^2] \leq 2d^2/B$. Applying Markov's inequality to $||I-M||_F^2$, we get

$$\mathbf{Pr}[||I-M||_F^2 \geq \gamma^2] \leq \mathbf{Pr}[||I-M||_F^2 \geq \gamma^2(B/(2d^2))\cdot\mathbf{E}[||I-M||_F^2]] \leq 2(d/\gamma)^2/B$$

as required.

We also have by Lemma 21 (fifth bound) that $\mathbf{E}_S[(\mathrm{Tr}(I-M))^2] \leq d^2/B$. Applying Markov's inequality to $(\mathrm{Tr}(I-M))^2$, we get

$$\mathbf{Pr}[|\mathrm{Tr}(I-M)| \geq \gamma] = \mathbf{Pr}[(\mathrm{Tr}(I-M))^2 \geq \gamma^2] \leq \mathbf{Pr}[(\mathrm{Tr}(I-M))^2 \geq \gamma^2(B/(d^2))\cdot\mathbf{E}[(\mathrm{Tr}(I-M))^2]] \leq (d/\gamma)^2/B.$$

A union bound over the two events gives the result. $\square$

[1] Note that we use the notation $S \in \mathcal{E}(\gamma)$ and $M \in \mathcal{E}(\gamma)$ interchangeably. This is fine since $M = U^TS^TSU$ and the matrix U is fixed.

1.2 Typical sets $\mathcal{T}(S,U)$ and their properties

In order to construct a single typical set $\mathcal{T}^*$, we will need the following simple definitions of sets $\mathcal{T}(S,U)$ of $x \in \mathbb{R}^d$ that are 'typical' for a given CountSketch matrix (as opposed to the set $\mathcal{T}^*$ whose existence is guaranteed by Lemma 8, which contains x that are 'typical' for **all matrices** $S \in \mathcal{E}$ **simultaneously**). We will use

Definition 12 (Typical x)**.** For any orthonormal matrix $U \in \mathbb{R}^{n\times d}$ and CountSketch matrix S we define

$$\mathcal{T}(S,U) := \left\{ x \in \mathbb{R}^d : |x^T(I-M)x| \le \frac{1}{100} \text{ and } |x^T(I-M)^2 x| \le \frac{1}{100} \right\},$$

The following claim will be useful in what follows. Its (simple) proof is given in the appendix:

Claim 13. *For any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns and any CountSketch matrix $S \in \mathbb{R}^{B\times n}$ one has $||I-M||_F^2 \le 4n^3$.*

The following claim is crucial to our analysis. A detailed proof is given in the appendix.

Claim 14. *For any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, every $\gamma \le 1/\log^2 n$, every CountSketch matrix $S \in \mathcal{E}(\gamma)$ one has* **(1)** $\mathbf{Pr}_{X\sim N(0,I_d)}[X \notin \mathcal{T}(S,U)] < n^{-40}$ *and* **(2)** *for any CountSketch matrix $S' \in \mathcal{E}(\gamma)$, $M' = U^T S'^T S' U$ one has $\mathbf{Pr}_{X\sim N(0,M')}[X \notin \mathcal{T}(S,U)] < n^{-40}$ for sufficiently large n.*

Using the claim above we get

Claim 15. *The total variation distance between $\tilde{q}$ (defined in* (7)*) and q' (defined in* (8)*) is at most n^{-10}. Further, $\xi \le n^{-40}$.*

Proof. We have

$$\begin{aligned}
D_{TV}(\tilde{q}, q') &\le 2\xi \le 2\int_{\mathbb{R}^d} \mathbf{E}_S\left[\frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \notin \mathcal{T}(S,U)] \middle| \mathcal{E}(\gamma) \right] dx \\
&= 2\mathbf{E}_S\left[\int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d \det M}} e^{-\frac{1}{2}x^T M^{-1} x} \cdot \mathbf{I}[x \notin \mathcal{T}(S,U)] dx \middle| \mathcal{E}(\gamma) \right] \\
&= 2\mathbf{E}_S\left[\mathbf{Pr}_{X\sim N(0,M)}[x \notin \mathcal{T}(S,U)] \middle| \mathcal{E}(\gamma) \right] \\
&\le 2n^{-40} \le n^{-10} \qquad \text{(by Claim 14)}
\end{aligned}$$

as required. ∎

1.3 Basic Taylor expansions

In this section we define the basic Taylor expansions of $\tilde{q}(x)/p(x)$ that form the foundation of our analysis. Our analysis of (18) proceeds by first Taylor expanding M^{-1} and $\det M$ around the identity matrix using Claims 9 and 10, which is valid since for any $S \in \mathcal{E}(\gamma)$ for $\gamma < 1/2$ one has $||I-M||_2 \le ||I-M||_F \le 1/2$. This gives

$$\begin{aligned}
\tilde{q}(x)/p(x) &= \mathbf{E}_S\left[\exp\left(\frac{1}{2}x^T x - \frac{1}{2}\left(\sum_{k\ge 0} x^T (I-M)^k x \right) + \frac{1}{2}\sum_{k\ge 1} \mathrm{Tr}((I-M)^k)/k \right) \middle| \mathcal{E} \right] \\
&= \mathbf{E}_S\left[\exp\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - \frac{1}{2}\sum_{k\ge 2}\left(x^T(I-M)^k x - \mathrm{Tr}((I-M)^k)/k \right) \right) \middle| \mathcal{E} \right] \\
&= \mathbf{E}_S\left[\exp\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - R(x) \right) \middle| \mathcal{E} \right],
\end{aligned} \tag{20}$$

where $R(x) := \frac{1}{2}\sum_{k\geq 2}\left(x^T(I-M)^k x - \mathrm{Tr}((I-M)^k)/k\right)$.

The rationale behind the definition of $\mathcal{E}(\gamma)$ is that for all $S \in \mathcal{E}(\gamma)$ the residual $R(x)$ above is (essentially) dominated by the quadratic terms, i.e. $||I-M||_F^2$ and $x^T(I-M)^2x$ (for 'typical' values of x – see Lemma 18 below), i.e. we can truncate the Taylor expansion to the first and second terms and control the error. This is made formal by the following three lemmas.

Lemma 16. *For every $\gamma \in (0,1)$, conditioned on $\mathcal{E}(\gamma)$ we have $\mathrm{Tr}((I-M)^k) \leq \gamma^{k-2} \cdot ||I-M||_F^2$ for all $k \geq 2$.*

Proof. $|\mathrm{Tr}((I-M)^k)| \leq ||I-M||_2^{k-2} \cdot \mathrm{Tr}((I-M)^2) \leq ||I-M||_F^{k-2} \cdot ||I-M||_F^2 \leq \gamma^k$ as required, since $||A||_2 \leq ||A||_F$ and $\mathrm{Tr}(A^TA) = ||A||_F^2$ for all $A \in \mathbb{R}^{d\times d}$. □

Lemma 17. *For any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, any $\gamma \in (0,1/2)$, for any $x \in \mathbb{R}^d$ one has, for any CountSketch matrix $S \in \mathcal{E}(\gamma)$, $x^T(I-M)^kx \leq \gamma^{k-2}x^T(I-M)^2x$ for any $k \geq 2$.*

Proof. We have, for any $x \in \mathbb{R}^d$ and any $S \in \mathcal{E}(\gamma)$ $|x^T(I-M)^kx| \leq ||I-M||_2^{k-2} \cdot x^T(I-M)^2x \leq \gamma^{k-2} \cdot x^T(I-M)^2x$, as $||I-M||_2 \leq ||I-M||_F$.

□

Lemma 18. *For any $\gamma \in (0,1/2)$, any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, any CountSketch matrix $S \in \mathcal{E}(\gamma)$ and any $x \in \mathcal{T}(S,U)$ one has*

$$|R(x)| \leq \sum_{k\geq 2}|x^T(I-M)^kx| + |\mathrm{Tr}((I-M)^k)|/k \leq C||I-M||_F^2 + Cx^T(I-M)^2x,$$

where $C > 0$ is an absolute constant.

Proof. We have by combining Lemma 16 and Lemma 17

$$\begin{aligned}\sum_{k\geq 2}|x^T(I-M)^kx| + |\mathrm{Tr}((I-M)^k)|/k &\leq \sum_{k\geq 2}[\gamma^{k-2}x^T(I-M)^2x + \gamma^{k-2}\cdot ||I-M||_F^2/k] \\ &\leq C(x^T(I-M)^2x + ||I-M||_F^2)\end{aligned}$$

for an absolute constant $C' > 0$, as $\gamma < 1/2$ by assumption of the lemma. □

1.4 Constructing the universal set $\mathcal{T}^*(U,\gamma)$ of typical x

The main result of this section is the following lemma:

Lemma 19. *For every matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, for every $\gamma \in (0, 1/\log^2 n)$ and any $\delta > 0$ if*

$$\begin{aligned}\mathcal{T}^*(U,\gamma) := \big\{x \in \mathbb{R}^d \text{ s.t. } &||x||_\infty \leq C\sqrt{\log n} \text{ and} \\ &|(Ux)_a| \leq O(\sqrt{\log n})||U_a||_2 \text{ for all } a \in [n] \text{ and} \\ &\mathbf{E}_S\left[\mathbf{I}[x \notin \mathcal{T}(S,U)] \,|\, \mathcal{E}(\gamma)\right] < 1/n^{25}.\big\},\end{aligned}$$

then **(a)** $\mathbf{Pr}_{X\sim N(0,I_d)}[X \in \mathcal{T}^*(U,\gamma)] \geq 1 - n^{-10}$ *and* **(b)** $\mathbf{Pr}_{X\sim\tilde{q}}[X \in \mathcal{T}^*(U,\gamma)] \geq 1 - n^{-10}$.

Note that the lemma guarantees the existence of a universal set $\mathcal{T}^* \subseteq \mathbb{R}^d$ that captures most of the probability mass of both the normal distribution $N(0,I_d)$ and the mixture $\tilde{q}$.

Proof of Lemma 19:

Let

$$\mathcal{T}_1^* := \{x \in \mathbb{R}^d : \mathbf{E}_S\left[\mathbf{I}[x \notin \mathcal{T}(S,U)] \,|\, \mathcal{E}(\gamma)\right] < 1/n^{25}\}.$$

$$\mathcal{T}_2^* := \{x \in \mathbb{R}^d : ||x||_\infty \leq C\sqrt{\log n}\}.$$

$$\mathcal{T}_3^* := \{x \in \mathbb{R}^d : |(Ux)_a| \leq C\sqrt{\log n}||U_a||_2 \text{ for all } a \in [n]\}.$$

We prove that $\mathcal{T}_i^*, i = 1,2,3$ occur with high probability under both distributions. As we show below, the result then follows by a union bound.

Showing that $\mathcal{T}_1^*$ occurs with high probability. We first show that $\mathcal{T}_1^*$ occurs with high probability under the isotropic Gaussian distribution $X \sim N(0, I_d)$, and then show that it also occurs with high probability under the mixture of Gaussians distribution $\tilde{q}$. In both cases the proof proceeds by applying Claim 14 followed by Markov's inequality.

Step 1: bounding $\mathbf{Pr}_{X\sim N(0,I_d)}[\mathcal{T}_1^*]$. We have by Claim 14, **(1)** that $\mathbf{Pr}_{X\sim N(0,I_d)}[\mathbf{I}[X \notin \mathcal{T}(S,U)]] < n^{-40}$, and hence

$$\mathbf{E}_S\left[\mathbf{E}_{X\sim N(0,I_d)}\left[\mathbf{I}[X \notin \mathcal{T}(S,U)]\right]\middle|\, \mathcal{E}(\gamma)\right] < 1/n^{40},$$

implying that $\mathbf{E}_{X\sim N(0,I_d)}\left[\mathbf{E}_S\left[\mathbf{I}[X \notin \mathcal{T}(S,U)]\right]\middle|\, \mathcal{E}(\gamma)\right] < 1/n^{40}$. We thus get by Markov's inequality that $\mathbf{Pr}_{X\sim N(0,I_d)}[\mathcal{T}_1^*] \geq 1 - n^{-15}$.

Step 2: bounding $\mathbf{Pr}_{X\sim\tilde{q}}[\mathcal{T}_1^*]$. We have by Claim 14, **(2)** that for any $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, any pair of matrices $S, S' \in \mathcal{E}(\gamma)$, if $M' = U^T S^T S U$, then $\mathbf{Pr}_{X\sim N(0,M')}[X \notin \mathcal{T}(S,U)] < n^{-40}$. We thus have

$$\begin{aligned}
\mathbf{E}_{X\sim\tilde{q}}[\mathbf{E}_S[\mathbf{I}[X \notin \mathcal{T}(S,U)]|\mathcal{E}(\gamma)]] &= \mathbf{E}_{S'}\left[\mathbf{E}_{X\sim q_{S'}}[\mathbf{E}_S[\mathbf{I}[X \notin \mathcal{T}(S,U)]]|\mathcal{E}(\gamma)]\middle|\, \mathcal{E}(\gamma)\right] \\
&= \mathbf{E}_S\left[\mathbf{E}_{S'}[\mathbf{E}_{X\sim q_{S'}}[\mathbf{I}[X \notin \mathcal{T}(S,U)]]|\mathcal{E}(\gamma)]\middle|\, \mathcal{E}(\gamma)\right] \\
&= \mathbf{E}_S\left[\mathbf{Pr}_{X\sim\tilde{q}}[\mathbf{I}[X \notin \mathcal{T}(S,U)]]\middle|\, \mathcal{E}(\gamma)\right] \\
&\leq n^{-40}.
\end{aligned}$$

By Markov's inequality applied to the expression in the first line we thus have

$$\mathbf{Pr}_{X\sim\tilde{q}}[\mathbf{E}_S[\mathbf{I}[X \notin \mathcal{T}(S,U)]|\mathcal{E}(\gamma)] > n^{-25}] < n^{-15}.$$

Showing that $\mathcal{T}_2^*$ occurs with high probability. The fact that

$$\mathbf{Pr}_{X\sim N(0,I_d)}\left[||X||_\infty \leq C\sqrt{\log n}\right] \geq 1 - n^{-40}$$

follows by standard properties of Gaussian random variables. Thus, it remains to show that $\mathcal{T}_2^*$ occurs with high probability under $X \sim \tilde{q}$. For any $U \in \mathbb{R}^{n\times d}$ and $S \in \mathcal{E}(\gamma)$ we now prove that for $M = U^T S^T S U$

$$\mathbf{Pr}_{X\sim N(0,M)}\left[||X||_\infty \leq C\sqrt{\log n}\right] \geq 1 - n^{-40} \tag{21}$$

It is convenient to let $X = M^{1/2}Y$, where $Y \sim N(0, I_d)$ is a vector of independent Gaussians of unit variance. Then we need to bound

$$\mathbf{Pr}_{X\sim N(0,M)}\left[||X||_\infty \geq C\sqrt{\log n}\right] = \mathbf{Pr}_{Y\sim N(0,I_d)}\left[||M^{1/2}Y||_\infty \geq C\sqrt{\log n}\right]$$

By 2-stability of the Gaussian distribution we have that for each $i = 1, \ldots, d$ the random variable $(M^{1/2}Y)_i$ is Gaussian with variance at most $||M^{1/2}||_F^2$, which we bound by

$$
\begin{aligned}
||M^{1/2}||_F = ||(I + (M-I))^{1/2}||_F &= \left|\left|\sum_{t=0}^{\infty} \binom{1/2}{t} (I-M)^t\right|\right|_F \\
&\leq \sum_{t=0}^{\infty} \left|\binom{1/2}{t}\right| \cdot \left|\left|(I-M)^t\right|\right|_F \\
&\leq \sum_{t=0}^{\infty} \left|\binom{1/2}{t}\right| \cdot ||I-M||_F^t \\
&\leq \sum_{t=0}^{\infty} ||I-M||_F^t \\
&\leq \sum_{t=0}^{\infty} (1/2)^t \\
&\leq 2
\end{aligned}
$$

Thus, for each $i \in [n]$ the random variable $(M^{1/2}Y)_i$ is Gaussian with variance at most 4, and (21) follows by standard properties of Gaussian random variables as long as $C > 0$ is a sufficiently large constant.

Showing that $\mathcal{T}_3^*$ occurs with high probability. The fact that

$$\mathbf{Pr}_{X \sim N(0, I_d)} \left[|(UX)_a| \leq C\sqrt{\log n} \cdot ||U_a||_2 \text{ for all } a \in [n] \right] \geq 1 - n^{-40}$$

follows by standard properties of Gaussian random variables and a union bound over all $a \in [n]$.

Thus, it remains to show that $\mathcal{T}_3^*$ occurs with high probability under $X \sim \tilde{q}$. For any $U \in \mathbb{R}^{n \times d}$ and $S \in \mathcal{E}(\gamma)$ we now prove that for $M = U^T S^T S U$

$$\mathbf{Pr}_{X \sim N(0, M)} \left[|(UX)_a| \leq C\sqrt{\log n} ||U_a||_2 \text{ for all } a \in [n] \right] \geq 1 - n^{-40}$$

It is convenient to let $X = M^{1/2}Y$, where $Y \sim N(0, I_d)$ is a vector of independent Gaussians of unit variance. Then we need to bound, for each $a \in [n]$

$$\mathbf{Pr}_{X \sim N(0,M)} \left[|(UX)_a| \geq C\sqrt{\log n} ||U_a||_2 \right] = \mathbf{Pr}_{Y \sim N(0, I_d)} \left[|(UM^{1/2}Y)_a| \geq C\sqrt{\log n} ||U_a||_2 \right]$$

By 2-stability of the Gaussian distribution we have that for each $a = 1, \ldots, n$ the random variable $U_a M^{1/2} Y$ is Gaussian with variance at most $||U_a M^{1/2}||_2^2 \leq 4||U_a||_F^2$ (since $\gamma < 1/\log^2 n$ by assumption of the lemma), and hence the result follows by standard properties of Gaussian random variables and a union bound.

Finally, we let $\mathcal{T}^* := \mathcal{T}_1^* \cap \mathcal{T}_2^* \cap \mathcal{T}_3^*$. By a union bound applied to the bounds above we have that $\mathcal{T}^*$ occurs with probability at least $1 - n^{-10}$ under both distributions, as required. □

1.5 Proof of Lemma 8

We first prove

Lemma 20. *There exists an absolute constant $C > 0$ such that for every $\gamma \in (0, 1/\log n)$, any matrix $U \in \mathbb{R}^{n \times d}$ with orthonormal columns and any CountSketch matrix $S \in \mathcal{E}(\gamma)$ and $x \in \mathcal{T}(S, U)$ one has, letting*

$$L(x) := -\frac{1}{2} x^T (I - M) x + \frac{1}{2} \mathrm{Tr}(I - M) - \frac{1}{8} x^T (I - M) x \cdot \mathrm{Tr}(I - M)$$

and

$$Q(x) := ((x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2x + ||I-M||_F^2),$$

that

$$\left|1 + L(x) - \exp\left(\frac{1}{2}x^Tx - \frac{1}{2}x^TM^{-1}x - \frac{1}{2}\log\det M\right)\right| \leq C \cdot Q(x).$$

The proof is given in section A.

We will need the following two lemmas, whose proofs are provided in section A.2

Lemma 21. *For any $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, and $B \geq 1$, if S is a random CountSketch matrix and $M = U^TS^TSU$, then*

(1) $\mathbf{E}_S[||M-I||_F^2] \leq 2d^2/B$

(2) *for all $x \in \mathcal{T}^*$ one has* $\mathbf{E}_S[x^T(I-M)^2x] = O(d^2(\log^2 n)/B)$

(3) *for all $x \in \mathcal{T}^*$ one has* $\mathbf{E}_S[(x^T(I-M)x)^2] = O(d^2(\log^2 n)/B)$

(4) *for all $x \in \mathcal{T}^*$ one has* $\mathbf{E}_S[(x^T(I-M)x)\cdot \mathrm{Tr}(I-M)] = O(d^2(\log n)/B)$

(5) *one has* $\mathbf{E}_S[(\mathrm{Tr}(I-M))^2] = O(d^2/B)$

and

Lemma 22 (Variance bound)**.** *For any matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns if $\gamma \in (0, 1/2)$ and $\mathcal{T}^*(U,\gamma) \subseteq \mathbb{R}^d$ is as defined in Lemma 19, then for any $x \in \mathcal{T}^*(U,\gamma)$ one has, for*

$$L(x) := -\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - \frac{1}{8}x^T(I-M)x \cdot \mathrm{Tr}(I-M)$$

and

$$Q(x) := ((x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2x + ||I-M||_F^2),$$

that for any constant C

$$\mathbf{E}_S\left[(L(x) + C\cdot Q(x))^2\right] = O(d^2(\log^2 n)/B),$$

where S is a uniformly random CountSketch matrix and $M = U^TS^TSU$.

We will use the following lemma, whose proof is given in section A:

Lemma 23. *For any random variable Z and any event $\mathcal{E}$ with $\Pr[\mathcal{E}] \geq 1/2$, if $\epsilon := \mathbf{E}[(Z-1)^2]$, then*

$$|\mathbf{E}[Z] - \mathbf{E}[Z|\mathcal{E}]| \leq 2(1+\mathbf{E}[Z])\mathbf{Pr}[\bar{\mathcal{E}}] + 2\sqrt{\epsilon\mathbf{Pr}[\bar{\mathcal{E}}]}.$$

Equipped with the bounds above, we can now prove Lemma 8:

Lemma 8 (Restated) *There exists an absolute constant $C > 0$ such that for every $\delta \in (0,1)$ and every matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns if $B \geq \frac{1}{\delta}C(\log n)^4d^2$ there exists a set $\mathcal{E}$ of CountSketch matrices and a subset $\mathcal{T}^* \subseteq \mathbb{R}^d$ that satisfies $\mathbf{Pr}_{X\sim p}[X \notin \mathcal{T}^*] \leq n^{-10}$ and $\mathbf{Pr}_{X\sim\tilde{q}}[X \notin \mathcal{T}^*] \leq n^{-10}$ such that if $S \in \mathbb{R}^{B\times n}$ is a random CountSketch matrix, then* **(1)** $\mathbf{Pr}_S[\mathcal{E}] \geq 1-\delta/3$, *and* **(2)** *for all $x \in \mathcal{T}^*$ one has*

$$\left|\frac{q'(x)}{p(x)} - 1\right| \leq O((d^2\log^4 n)/B) + O(n^{-10}).$$

Proof. Let $\mathcal{T}^*(U,\gamma) \subseteq \mathbb{R}^d$ be as defined in Lemma 19, and let $\gamma := 1/\log^2 n$. Let $\mathcal{E} := \mathcal{E}(\gamma)$, and note that $\mathbf{Pr}[\mathcal{E}] \geq 1 - \delta/3$ by Claim 11 as long as C is a large enough constant, as required.

We now bound

$$\frac{q'(x)}{p(x)} = \mathbf{E}_S\left[\frac{q_S(x)}{p(x)} \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] \middle| \mathcal{E}(\gamma)\right] + \xi,$$

for $x \in \mathcal{T}^*(U,\gamma)$, where $\xi = \mathbf{E}_S[\mathbf{Pr}_{X\sim q_S}[X \in \mathcal{T}(S,U)]] \leq n^{-40}$ by definition and Claim 15, **(2)**. For each $S \in \mathcal{E}(\gamma)$ and $x \in \mathcal{T}(S,U)$ we have by Lemma 20

$$\left|\frac{q_S(x)}{p(x)} - (1 + L(x))\right| = \left|\exp\left(\frac{1}{2}x^T x - \frac{1}{2}x^T M^{-1} x - \frac{1}{2}\log\det M\right) - (1 + L(x))\right| \leq C \cdot Q(x),$$

where

$$L(x) := -\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - \frac{1}{8}x^T(I-M)x \cdot \mathrm{Tr}(I-M)$$

denotes the 'linear' term and

$$Q(x) := (x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2 x + ||I-M||_F^2$$

denotes the 'quadratic' term.

Taking expectations, we get

$$\begin{aligned}
&\mathbf{E}_S\left[(L(x) - C\cdot Q(x)) \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right] \\
&\leq \mathbf{E}_S\left[\left(\exp\left(\frac{1}{2}x^T x - \frac{1}{2}x^T M^{-1} x - \frac{1}{2}\log\det M\right) - 1\right) \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] \middle| \mathcal{E}(\gamma)\right] \\
&\leq \mathbf{E}_S\left[(L(x) + C\cdot Q(x)) \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right].
\end{aligned}$$

Thus, it suffices to show that

$$\left|\mathbf{E}_S\left[(L(x) \pm C\cdot Q(x)) \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right]\right| = O((Cd\log n)^2/B) + O(n^{-10}),$$

which we do now. We only provide the analysis for the case when the sign in front of the constant C is a plus, as the other part is analogous.

We first show that removing the multiplier $\mathbf{I}[x \in \mathcal{T}(S,U)]$ from the equation above only changes the expectation slightly. Specifically, note that

$$\begin{aligned}
&\left|\mathbf{E}_S\left[(L(x) + C\cdot Q(x)) \cdot \mathbf{I}[x \in \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right] - \mathbf{E}_S\left[L(x) + C\cdot Q(x) | \mathcal{E}(\gamma)\right]\right| \\
&\leq \mathbf{E}_S\left[|L(x) + C\cdot Q(x)| \cdot \mathbf{I}[x \notin \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right].
\end{aligned} \tag{22}$$

By Claim 13 we have $||I-M||_F^2 \leq 4n^3$ for all S and U, so every element of the matrix $I-M$ is upper bounded by $2n^2$. Similarly, we have $||(I-M)^2||_F \leq ||I-M||_F^2$, and so every element of $(I-M)^2$ is upper bounded by $4n^3$. Thus, for any $x \in \mathcal{T}^*(U,\gamma)$ one has

$$\begin{aligned}
&|L(x) + CQ(x)| \\
&\leq (|x^T(I-M)x| + |\mathrm{Tr}(I-M)| + |x^T(I-M)x \cdot \mathrm{Tr}(I-M)| \\
&+ C((x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2 x + ||I-M||_F^2)) \\
&= O(\log n)(2n^2d^2 + d\cdot(2n^2) + (2n^2)^2 d^3 + (2n^2d^2)^2 + (d\cdot 2n^2)^2 + 4n^4d^2 + 4n^3) \leq n^{10}
\end{aligned}$$

as long as n is sufficiently large, where we used the fact that $||x||_\infty \leq O(\sqrt{\log n})$ for all $x \in \mathcal{T}^*(U,\gamma)$.

Furthermore, by Lemma 19 we have for $x \in \mathcal{T}^*(U,\gamma)$ that

$$\mathbf{E}_S\left[\mathbf{I}[x \notin \mathcal{T}(S,U)] | \mathcal{E}(\gamma)\right] < 1/n^{25}.$$

Substituting these two bounds into (22), we get

$$\mathbf{E}_S\left[|L(x)+C\cdot Q(x)|\cdot \mathbf{I}[x\notin \mathcal{T}(S,U)]|\,\mathcal{E}(\gamma)\right]\leq n^{-10} \tag{23}$$

so it remains to bound

$$\mathbf{E}_S\left[L(x)+C\cdot Q(x)|\,\mathcal{E}(\gamma)\right].$$

We bound the expectation above by relating it to the corresponding unconditional expectation. Let $Z := 1+(L(x)+C\cdot Q(x))$, and note that

$$\mathbf{E}_S[Z]=1-\mathbf{E}_S[\frac{1}{8}x^T(I-M)x\cdot \mathrm{Tr}(I-M)]+C\cdot \mathbf{E}_S[Q(x)]=1+O((C\log n)^2d^2/B) \tag{24}$$

by Lemma 21. Let $\epsilon := \mathbf{E}_S[(Z-1)^2]$. We note that by Lemma 22 that $\epsilon\leq O(d^2(\log^2 n)/B)$, and hence since $\mathcal{E}(\gamma)\geq 1/2$ by Claim 11, by Lemma 23 we have

$$|\mathbf{E}[Z]-\mathbf{E}[Z|\mathcal{E}(\gamma)]|\leq 2(1+\mathbf{E}[Z])\mathbf{Pr}[\bar{\mathcal{E}}(\gamma)]+2\sqrt{\epsilon\mathbf{Pr}[\bar{\mathcal{E}}(\gamma)]}.$$

Since $\mathbf{Pr}[\bar{\mathcal{E}}(\gamma)]\leq 3(d/\gamma)^2/B$ by Claim 11 and using the assumption that $B\geq(\log^2 n)d^2$, we get

$$|\mathbf{E}[Z]-\mathbf{E}[Z|\mathcal{E}(\gamma)]|\leq O((d/\gamma)^2/B)+2\sqrt{O(d^2\log^2 n/B)\cdot(d/\gamma)^2/B)}=O((\frac{1}{\gamma^2}+\frac{1}{\gamma}\log n)d^2/B)=O((d/\gamma)^2/B), \tag{25}$$

where we used the assumption that $\gamma\leq 1/\log^2 n$. Combining (25), (24) with (22) and (23), we get

$$\left|\frac{q'(x)}{p(x)}-1\right|=\left|\mathbf{E}_S\left[\frac{q(x)}{p(x)}\cdot \mathbf{I}[x\in\mathcal{T}(U,S)]\middle|\,\mathcal{E}(\gamma)\right]+\xi-1\right|\leq O((d^2\log^4 n)/B)+O(1/n^{10}).$$

□

A Proofs omitted from the main body

A.1 Proof of Claim 14 and Claim 13

We will use

Theorem 24 (Bernstein's inequality)**.** *Let $X_1,\dots,X_n$ be independent zero mean random variables such that $|X_i|\leq L$ for all i with probability 1, and let $X:=\sum_{i=1}^n X_i$. Then*

$$\mathbf{Pr}[X>t]<\exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^n \mathbf{E}[X_i^2]+\frac{1}{3}Lt}\right).$$

Proof of Claim 14:

Proving (1). The bound follows by standard concentration inequalities, as we now show. Since the normal distribution is rotationally invariant, we have that

$$X^T(I-M)X=\sum_{i=1}^d(\lambda_i-1)Y_i^2=\mathrm{Tr}(M-I)+\sum_{i=1}^d(\lambda_i-1)(Y_i^2-1), \tag{26}$$

where $Y\sim N(0,I_d)$ and λ_i are the eigenvalues of M. We now apply Bernstein's inequality (Theorem 24) to random variables $(\lambda_i-1)(Y_i^2-1)$ (note that they are zero mean). We also have $\mathbf{E}[(\lambda_i-1)^2(Y_i^2-1)^2]\leq$

$O((\lambda_i-1)^2)$. We later combine it with the fact that $|\mathrm{Tr}(I-M)| \le \gamma \le \frac{1}{2}\cdot\frac{1}{100}$ for all $S \in \mathcal{E}(\gamma)$ to obtain the result. We also have $|(\lambda_i-1)Y_i| \le ||I-M||_F C\sqrt{\log n} \le \gamma \cdot C\sqrt{\log n}$ for all i with probability at least $1-n^{-40}/4$ as long as $C>0$ is larger than an absolute constant. We thus have by applying Theorem 24 to random variables clipped at $\gamma C\sqrt{\log n}$ in magnitude, which we denote by event $\mathcal{F}$, to conclude for all $t \ge 0$,

$$\mathbf{Pr}[\sum_{i=1}^{d}(\lambda_i-1)(Y_i^2-1) > t \mid \mathcal{F}] < 2\exp\left(-\frac{\frac{1}{2}t^2}{O(\sum_{i=1}^{n}(\lambda_i-1)^2)+(\frac{1}{3}\gamma C\sqrt{\log n})t}\right).$$

Note the random variables are still independent and zero-mean conditioned on $\mathcal{F}$, and $\mathbf{E}[(\lambda_i-1)^2(Y_i^2-1)^2] \le O((\lambda_i-1)^2)$ continues to hold, since the clipping changes the expectation by at most a factor of $(1+O(n^{-40}))$. By a union bound we can remove the conditioning on $\mathcal{F}$,

$$\mathbf{Pr}[\sum_{i=1}^{d}(\lambda_i-1)(Y_i^2-1) > t] < 2\exp\left(-\frac{\frac{1}{2}t^2}{O(\sum_{i=1}^{n}(\lambda_i-1)^2)+(\frac{1}{3}\gamma C\sqrt{\log n})t}\right)+\frac{n^{-40}}{4}.$$

Setting $t=\frac{1}{100}$, and using the fact that $\sum_i(\lambda_i-1)^2 = ||I-M||_F^2 \le \gamma^2$, we get

$$\begin{aligned}\mathbf{Pr}[\sum_{i=1}^{d}(\lambda_i-1)(Y_i^2-1) > \frac{1}{2}\cdot\frac{1}{100}] &< 2\exp\left(-\frac{\frac{1}{2}(\frac{1}{2}\cdot\frac{1}{100})^2}{O(\gamma^2)+(\frac{1}{3}\cdot(\frac{1}{2}\cdot\frac{1}{100})\gamma C\sqrt{\log n})}\right)+\frac{n^{-40}}{4}\\ &= \exp(-\Omega(1/(\gamma\sqrt{\log n})))+\frac{n^{-40}}{4}\\ &< \frac{n^{-40}}{2},\end{aligned}$$

since $\gamma \le 1/\log^2 n$ by assumption, for a sufficiently large n. Combining this with (26), we get, using the fact that $|\mathrm{Tr}(I-M)| \le \gamma < \frac{1}{2}\cdot\frac{1}{100}$ for $S \in \mathcal{E}(\gamma)$ that

$$\mathbf{Pr}[X^T(I-M)X > \frac{1}{100}] \le \mathbf{Pr}[|\sum_{i=1}^{d}(\lambda_i-1)(Y_i^2-1)| > \frac{1}{2}\frac{1}{100}] < n^{-40}/2,$$

as required.

We also have

$$X^T(I-M)^2X = \sum_{i=1}^{d}(\lambda_i-1)^2Y_i^2 \le ||I-M||_F^2\cdot\max_{i\in[d]}|Y_i|^2 \le O(\log n)\cdot||I-M||_F^2 = O(\log n\gamma^2) \le \frac{1}{100}$$

with probability at least $1-n^{-40}/2$ by standard properties of Gaussian random variables. Putting the two estimates together and taking a union bound over the failure events now shows that $\mathbf{Pr}_{X\sim N(0,I_d)}[X \notin \mathcal{T}(S,U)] < n^{-40}$, as required.

Proving (2). Recall that $\mathcal{T}(S,U) = \left\{x \in \mathbb{R}^d : |x^T(I-M)x| \le \frac{1}{100} \text{ and } x^T(I-M)^2x \le \frac{1}{100}\right\}$. For any S' we have that $X \sim N(0,M')$, where $M' = (S'U)^TS'U$, so $X = M'^{1/2}Y$, where $Y = N(0,I_d)$. We thus have

$$X^T(I-M)X = (M'^{1/2}Y)^T(I-M)(M'^{1/2}Y) = Y^TM'^{1/2}(I-M)M'^{1/2}Y.$$

We now show that

$$\mathbf{Pr}_{Y\sim N(0,I_d)}\left[\left|Y^TM'^{1/2}(I-M)M'^{1/2}Y\right| > \frac{1}{100}\right] < 1/n^{20} \tag{27}$$

Let $Q := M'^{1/2}(I-M)M'^{1/2}$, and let $1-\tilde{\lambda}_i, i=1,\ldots,d$ denote the eigenvalues of Q. We have

$$Y^T M'^{1/2}(I-M)M'^{1/2}Y = \sum_{i=1}^{d}(1-\tilde{\lambda}_i)Z_i^2,$$

where $Z \sim N(0, I_d)$. Note that

$$\begin{aligned}
|\sum_{i=1}^{d}(1-\tilde{\lambda}_i)| &= |\mathrm{Tr}(Q)| = |\mathrm{Tr}(M'^{1/2}(I-M)M'^{1/2})| \\
&= |\mathrm{Tr}(M'(I-M))| = |\mathrm{Tr}((I-(I-M'))(I-M))| \\
&\leq |\mathrm{Tr}(I-M)| + |\mathrm{Tr}\left((I-M')\,(I-M)\right)| \\
&= \gamma + |\mathrm{Tr}\left((I-M')\,(I-M)\right)| \qquad (\text{since } |\mathrm{Tr}(I-M)| \leq \gamma \text{ for all } S \in \mathcal{E}(\gamma)) \\
&\leq \gamma + ||I-M'||_F \cdot ||M-I||_F \qquad (\text{by von Neumann and Cauchy-Schwarz inequalities}) \\
&\leq \gamma + \gamma^2
\end{aligned} \tag{28}$$

We thus have

$$\begin{aligned}
Y^T M'^{1/2}(I-M)M'^{1/2}Y &= \sum_{i=1}^{d}(1-\tilde{\lambda}_i)Z_i^2 \\
&= \sum_{i=1}^{d}(1-\tilde{\lambda}_i) + \sum_{i=1}^{d}(1-\tilde{\lambda}_i)(Z_i^2-1)
\end{aligned} \tag{29}$$

We now use a calculation analogous to the above for **(1)** to show that $|\sum_{i=1}^{d}(1-\tilde{\lambda}_i)(Z_i^2-1)| \leq \frac{1}{2}\cdot\frac{1}{100}$ with probability at least $1-n^{-40}/4$. Indeed, we first verify that the variance is bounded by

$$\begin{aligned}
O(\sum_{i=1}^{d}(1-\tilde{\lambda}_i)^2) &= O(||Q||_F^2) \\
&= O(||M'^{1/2}(I-M)M'^{1/2}||_F^2) \\
&\leq O(||M'||_2^2)||I-M||_F^2 \qquad (\text{by sub-multiplicativity}) \\
&\leq O((||I||_2 + ||M'-I||_F)^2)||I-M||_F^2 \\
&\leq O(||I-M||_F^2) \\
&= O(\gamma^2).
\end{aligned} \tag{30}$$

We also have

$$\begin{aligned}
|(1-\tilde{\lambda}_i)Y_i| &\leq ||Q||_F C\sqrt{\log n} \\
&\leq ||M'||_2||I-M||_F C\sqrt{\log n} \qquad (\text{by sub-multiplicativity}) \\
&\leq (||I||_2 + ||M'-I||_F)||I-M||_F C\sqrt{\log n} \\
&\leq 2||I-M||_F C\sqrt{\log n} \\
&\leq 2\gamma \cdot C\sqrt{\log n},
\end{aligned}$$

for all i with probability at least $1-1/n^{40}/5$ as long as $C>0$ is larger than an absolute constant. We thus have by Theorem 24 (applied to clipped variables and then unclipping by a union bound as in (**1**)) for all $t \geq 0$ that

$$\mathbf{Pr}[|Y^T M'^{1/2}(I-M)M'^{1/2}Y - \sum_{i=1}^{d}(1-\tilde{\lambda}_i)| > t] < \exp\left(-\frac{\frac{1}{2}t^2}{O(\sum_{i=1}^{n}(1-\tilde{\lambda}_i)^2) + (\frac{1}{3}2\gamma C\sqrt{\log n})t}\right) + n^{-40}/5.$$

Setting $t = \frac{1}{2}\frac{1}{100}$, and using the upper bound $O(\sum_i (1-\tilde{\lambda}_i)^2) = O(\gamma^2)$ obtained in (30), we get

$$\mathbf{Pr}[|Y^T M'^{1/2}(I-M)M'^{1/2}Y - \sum_{i=1}^d (1-\tilde{\lambda}_i)| > \frac{1}{2}\cdot\frac{1}{100}] < \exp\left(-\frac{\frac{1}{2}(\frac{1}{2}\cdot\frac{1}{100})^2}{C\gamma^2 + (\frac{1}{3}\cdot\frac{1}{2}\frac{1}{100}\gamma C\sqrt{\log n})}\right) + n^{-40}/5$$
$$= \exp(-\Omega(1/(\gamma\sqrt{\log n}))) + n^{-40}/5 < n^{-40}/4$$

since $\gamma \le 1/\log^2 n$ by assumption, for a sufficiently large n. Since $|\sum_{i=1}^d (1-\tilde{\lambda}_i)| \le \gamma + 2\gamma^2 \le \frac{1}{2}\cdot\frac{1}{100}$ by (28), we get by triangle inequality that

$$\mathbf{Pr}_{X\sim N(0,M')}[|X^T(I-M)X| > \frac{1}{100}] \le n^{-40}/4.$$

Similarly to **(1)** above, we have, when $X \sim N(0, M'), X = M'^{1/2}Y, Y \sim N(0, I_d)$,

$$\begin{aligned} X^T(I-M)^2X = Y^T M'^{1/2}(I-M)^2 M'^{1/2} Y &= \sum_{i=1}^d \tilde{\tau}_i Z_i^2 \\ &\le \mathrm{Tr}(M'^{1/2}(I-M)^2 M^{1/2}) \cdot \max_{i\in[d]} Z_i^2 \\ &\le O(\log n)\cdot \mathrm{Tr}(M'^{1/2}(I-M)^2 M'^{1/2}) \end{aligned}$$

with probability at least $1 - n^{-40}/2$ over the choice of X, as $\max_{i\in[d]} Z_i^2 \le C\log n$ with high probability if C is a sufficiently large constant by standard properties of Gaussian random variables. Since $\mathrm{Tr}(M'^{1/2}(I-M)^2 M'^{1/2}) = \mathrm{Tr}(M'(I-M)^2) \le 2||I-M||_F^2$ (as $\gamma < 1/\log^2 n < 1/3$ by assumption of the lemma), we get

$$X^T(I-M)^2X \le O(\log n)\cdot \mathrm{Tr}(M'^{1/2}(I-M)^2M'^{1/2}) \le O(\log n)\cdot\gamma^2 \le \frac{1}{100} \quad (\text{since } \gamma < 1/\log^2 n)$$

with probability at least $1 - n^{-40}/4$. A union bound over the failure events yields $\mathbf{Pr}_{X\sim N(0,M')}[X \notin \mathcal{T}(S,U)] < n^{-40}$, as required.

This completes the proof. □

Proof of Lemma 20: By assumption that $S \in \mathcal{E}(\gamma)$ we have that $||I-M||_2 \le \gamma$, so Taylor expansion is valid and gives

$$\frac{1}{2}x^Tx - \frac{1}{2}x^TM^{-1}x - \frac{1}{2}\log\det M = -\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + R(x),$$

where for all $x \in \mathcal{T}(S,U)$ one has $R(x) \le \sum_{k\ge 2} x^T(I-M)^2x + \mathrm{Tr}(I-M)^k$.

We have by Lemma 18 that $R(x) \le C(x^T(I-M)^2x + ||I-M||_F^2)$ for an absolute constant $C > 0$, for all $x \in \mathcal{T}(S,U)$ and $S \in \mathcal{E}(\gamma)$. We thus have

$$\begin{aligned} &e^{-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - C(x^T(I-M)^2x + ||I-M||_F^2)} \\ &\le e^{-\frac{1}{2}x^Tx + \frac{1}{2}\mathrm{Tr}(I-M) - \frac{1}{2}x^TM^{-1}x - \frac{1}{2}\log\det M} \\ &\le e^{-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + C(x^T(I-M)^2x + ||I-M||_F^2)} \end{aligned} \tag{31}$$

for all such M and x.

We now Taylor expand $e^{-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + A(x^T(I-M)^2x + ||I-M||_F^2)}$, where A is any constant (positive or negative), getting

$$\begin{aligned} &e^{-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + A(x^T(I-M)^2x + ||I-M||_F^2)} \\ &= \sum_{k\ge 1}\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + A(x^T(I-M)^2x + ||I-M||_F^2)\right)^k / k!. \end{aligned} \tag{32}$$

For $k = 2$ we have

$$\left|\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + x^T(I-M)^2x + ||I-M||_F^2\right)^2/2 + \frac{1}{8}x^T(I-M)x\cdot\mathrm{Tr}(I-M)\right| \\ \leq C\left((x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2x + ||I-M||_F^2\right), \tag{33}$$

where we used the fact $|x^T(I-M)x| \leq \frac{1}{100}$ for $x \in \mathcal{T}(S, U)$ and $|\mathrm{Tr}(I-M)| \leq \gamma < \frac{1}{100}$ for $S \in \mathcal{E}(\gamma)$.

For all $k \geq 3$ we use the bound

$$\begin{aligned}
&|\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + x^T(I-M)^2x + ||I-M||_F^2\right)^k|\\
&\leq \left(|x^T(I-M)x| + \frac{1}{2}|\mathrm{Tr}(I-M)| + x^T(I-M)^2x + ||I-M||_F^2\right)^k\\
&\leq \left(|x^T(I-M)x| + \frac{1}{2}|\mathrm{Tr}(I-M)| + x^T(I-M)^2x + ||I-M||_F^2\right)^3\\
&\leq C((x^T(I-M)x)^2 + (\mathrm{Tr}(I-M))^2 + x^T(I-M)^2x + ||I-M||_F^2),
\end{aligned} \tag{34}$$

where we used the bound $|x^T(I-M)x| + \frac{1}{2}|\mathrm{Tr}(I-M)| + x^T(I-M)^2x + ||I-M||_F^2 \leq 1$ to go from the second line to the third, and the last line follows from the observation that every term in the expansion of $\left(|x^T(I-M)x| + \frac{1}{2}|\mathrm{Tr}(I-M)| + x^T(I-M)^2x + ||I-M||_F^2\right)^3$ contains either at least a square of one of the first two terms or at least one of the last two.

Substituting these bounds into (32), we get

$$\begin{aligned}
&e^{-\frac{1}{2}x^T(I-M)x+\frac{1}{2}\mathrm{Tr}(I-M)+A(x^T(I-M)^2x+||I-M||_F^2)}\\
&= \sum_{k\geq 1}\left(-\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + A(x^T(I-M)^2x + ||I-M||_F^2)\right)^k/k!\\
&\leq -\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) - \frac{1}{8}x^T(I-M)x\cdot\mathrm{Tr}(I-M)\\
&+ C((x^T(I-M)x)^2 + x^T(I-M)^2x + \mathrm{Tr}(I-M)^2 + ||I-M||_F^2) \quad \text{(for a constant } C > 0 \text{ that may depend on } A)\\
&+ \sum_{k\geq 3}(A+1)^k((x^T(I-M)x)^2 + \mathrm{Tr}(I-M)^2 + x^T(I-M)^2x + ||I-M||_F^2)/k!\\
&\leq -\frac{1}{2}x^T(I-M)x + \frac{1}{2}\mathrm{Tr}(I-M) + C''(x^T(I-M)^2x + (\mathrm{Tr}(I-M))^2 + x^T(I-M)x^2 + ||I-M||_F^2)
\end{aligned}$$

for an absolute constant $C'' > 0$. The provides the upper bound in the claimed result. The lower bound is provided by a similar calculation, which we omit. □

Proof of Lemma 23: Since $\mathbf{E}[(Z-1)^2] \leq \epsilon$ by assumption of the lemma, for any event $\mathcal{E}$ one has $\mathbf{E}[(Z-1)^2 \cdot \mathbf{I}_{\bar{\mathcal{E}}}] \leq \epsilon$, where $\mathbf{I}_{\bar{\mathcal{E}}}$ is the indicator of $\bar{\mathcal{E}}$, the complement of $\mathcal{E}$. This also means that

$$\mathbf{E}[(Z-1)^2|\bar{\mathcal{E}}] \leq \epsilon/\mathbf{Pr}[\bar{\mathcal{E}}].$$

On the other hand, by Jensen's inequality

$$\mathbf{E}[|Z-1||\bar{\mathcal{E}}] \leq \left(\mathbf{E}[(Z-1)^2|\bar{\mathcal{E}}]\right)^{1/2},$$

and putting these two bounds together we get

$$\mathbf{E}[|Z-1|\cdot\mathbf{I}[\bar{\mathcal{E}}]] = \mathbf{E}[|Z-1||\bar{\mathcal{E}}]\cdot\mathbf{Pr}[\bar{\mathcal{E}}] \leq \mathbf{Pr}[\bar{\mathcal{E}}]\cdot\left(\mathbf{E}[(Z-1)^2|\bar{\mathcal{E}}]\right)^{1/2} \leq \mathbf{Pr}[\bar{\mathcal{E}}]\cdot\left(\epsilon/\mathbf{Pr}[\bar{\mathcal{E}}]\right)^{1/2} = \sqrt{\epsilon\cdot\mathbf{Pr}[\bar{\mathcal{E}}]}.$$

This means that

$$
\begin{aligned}
|\mathbf{E}[Z]-\mathbf{E}[Z|\mathcal{E}]| &\leq \left|\mathbf{E}[Z]-\frac{1}{\mathbf{Pr}[\mathcal{E}]}\mathbf{E}[Z\cdot \mathbf{I}_{\mathcal{E}}]\right| \\
&\leq \left|\mathbf{E}[Z]-\frac{1}{\mathbf{Pr}[\mathcal{E}]}\mathbf{E}[Z]+\frac{1}{\mathbf{Pr}[\mathcal{E}]}\mathbf{E}[Z\cdot \mathbf{I}_{\bar{\mathcal{E}}}]\right| \\
&\leq \mathbf{E}[Z]\left(\frac{1}{1-\mathbf{Pr}[\bar{\mathcal{E}}]}-1\right)+\left|\frac{1}{\mathbf{Pr}[\mathcal{E}]}\mathbf{E}[Z\cdot \mathbf{I}_{\bar{\mathcal{E}}}]\right| \\
&\leq \mathbf{E}[Z]\cdot 2\mathbf{Pr}[\bar{\mathcal{E}}]+2\mathbf{E}[Z\cdot \mathbf{I}_{\bar{\mathcal{E}}}] \quad \left(\text{since } \frac{1}{1-x}-1\leq 2x \text{ for } x\in(0,1/2)\right) \\
&\leq \mathbf{E}[Z]\cdot 2\mathbf{Pr}[\bar{\mathcal{E}}]+2(\mathbf{Pr}[\bar{\mathcal{E}}]+\mathbf{E}[|Z-1|\cdot \mathbf{I}_{\bar{\mathcal{E}}}]) \\
&\leq 2(1+\mathbf{E}[Z])\mathbf{Pr}[\bar{\mathcal{E}}]+2\sqrt{\epsilon \mathbf{Pr}[\bar{\mathcal{E}}]}.
\end{aligned}
$$

□

A.2 Proofs of moment bounds (Lemma 21 and Lemma 22)

Proof of Lemma 21 and Lemma 22: We start by noting that for every $i,j\in[1:d]$ the matrix $M=U^TS^TSU$ satisfies

$$
\begin{aligned}
M_{ij} &= \sum_{r=1}^{B}\sum_{a=1}^{n}\sum_{b=1}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j} \\
&= \sum_{a=1}^{n} U_{a,i}U_{a,j}\left(\sum_{r=1}^{B} S_{r,a}^2\right)+\sum_{r=1}^{B}\sum_{a=1}^{n}\sum_{b=1,b\neq a}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j} \\
&= \delta_{i,j}+\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j},
\end{aligned}
$$

where $\delta_{i,j}$ equals 1 if $i=j$ and equals 0 otherwise. We thus have, for every $i,j\in[1:d]$, that

$$
(M-I)_{ij}=\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j},
$$

which in particular means that

$$
\begin{aligned}
\mathrm{Tr}(I-M) = -\sum_i (M-I)_{ii} &= -\sum_i\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,i}, \\
&= -\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}S_{r,b}\cdot U_aU_b^T,
\end{aligned}
\tag{35}
$$

(note that it immediately follows that $\mathbf{E}_S[\mathrm{Tr}(I-M)]=0$, as $\mathbf{E}_S[S_{r,a}S_{r,b}]=0$ for $a\neq b$) and

$$
\begin{aligned}
x^T(I-M)x = -\sum_{ij}(M-I)_{ij}x_ix_j &= -\sum_{i,j}\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}x_ix_j \\
&= -\sum_{r=1}^{B}\sum_{\substack{a,b=1,\\ a\neq b}}^{n} S_{r,a}S_{r,b}(Ux)_a(Ux)_b
\end{aligned}
\tag{36}
$$

(note that it immediately follows that $\mathbf{E}_S[x^T(I-M)x] = 0$ for all x, as $\mathbf{E}_S[S_{r,a}S_{r,b}] = 0$ for $a \neq b$).

We also have

$$(M-I)^2_{ij} = \sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{r'=1}^{B} \sum_{\substack{c,d=1,\\ c\neq d}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}S_{r',c}U_{c,i}S_{r',d}U_{d,j}$$

and hence

$$\begin{aligned}
||I-M||_F^2 = \sum_{ij}(M-I)^2_{ij} &= \sum_{ij}\sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{r'=1}^{B} \sum_{\substack{c,d=1,\\ c\neq d}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}S_{r',c}U_{c,i}S_{r',d}U_{d,j} \\
&= \sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{r'=1}^{B} \sum_{\substack{c,d=1,\\ c\neq d}}^{n} S_{r,a}S_{r,b}S_{r',c}S_{r',d}(\sum_i U_{a,i}U_{c,i})(\sum_j U_{b,j}U_{d,j}) \\
&= \sum_{r=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{r'=1}^{B} \sum_{\substack{c,d=1,\\ c\neq d}}^{n} S_{r,a}S_{r,b}S_{r',c}S_{r',d}\cdot U_aU_c^T\cdot U_bU_d^T \\
&= \sum_{r_1=1}^{B} \sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n} \sum_{r_2=1}^{B} \sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n} S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}\cdot U_{a_1}U_{a_2}^T\cdot U_{b_1}U_{b_2}^T
\end{aligned} \tag{37}$$

We also need

$$\begin{aligned}
x^T(I-M)^2x = ||(I-M)x||_2^2 &= \sum_{i=1}^{d}\left(\sum_{j=1}^{d}(I-M)_{ij}x_j\right)^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{\bar{j}=1}^{d} x_jx_{\bar{j}}\cdot \sum_{r=1}^{B}\sum_{\bar{r}=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{\substack{\bar{a},\bar{b}=1,\\ \bar{a}\neq \bar{b}}}^{n} S_{r,a}U_{a,i}S_{r,b}U_{b,j}\cdot S_{\bar{r},\bar{a}}U_{\bar{a},i}S_{\bar{r},\bar{b}}U_{\bar{b},\bar{j}} \\
&= \sum_{r=1}^{B}\sum_{\bar{r}=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{\substack{\bar{a},\bar{b}=1,\\ \bar{a}\neq \bar{b}}}^{n} S_{r,a}S_{r,b}S_{\bar{r},\bar{a}}S_{\bar{r},\bar{b}}\cdot(\sum_{i=1}^{d}U_{a,i}U_{\bar{a},i})(\sum_{j=1}^{d}U_{b,j}x_j)(\sum_{\bar{j}}U_{\bar{b},\bar{j}}x_{\bar{j}}) \\
&= \sum_{r=1}^{B}\sum_{\bar{r}=1}^{B} \sum_{\substack{a,b=1,\\ a\neq b}}^{n} \sum_{\substack{\bar{a},\bar{b}=1,\\ \bar{a}\neq \bar{b}}}^{n} S_{r,a}S_{r,b}S_{\bar{r},\bar{a}}S_{\bar{r},\bar{b}}\cdot U_aU_{\bar{a}}^T\cdot(Ux)_b(Ux)_{\bar{b}} \\
&= \sum_{r_1=1}^{B}\sum_{r_2=1}^{B} \sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n} \sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n} S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}\cdot U_{a_1}U_{a_2}^T\cdot(Ux)_{b_1}(Ux)_{b_2}
\end{aligned} \tag{38}$$

Bounding $\mathbf{E}_S[||I-M||_F^2], \mathbf{E}_S[(x^T(I-M)x)^2], \mathbf{E}_S[x^T(I-M)^2x], \mathbf{E}_S[(x^T(I-M)x)\text{Tr}(I-M)], \mathbf{E}_S[\text{Tr}(I-M)^2]$ We first note that for for any r_1, r_2 and $a_1 \neq b_1, a_2 \neq b_2$ the quantity

$$\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}]$$

is only nonzero when $r_1 = r_2$ and $\{a_1, b_1, a_2, b_2\}$ contains two distinct elements, each with multiplicity 2 (let $\mathbf{I}_*(\{a_q, b_q\}_{q=1}^2)$ denote the indicator of the latter condition). In that case one has $\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}] =$

$1/B^2$. Note that the expression above appears in all of $\mathbf{E}_S[(x^T(I-M)x)^2], \mathbf{E}_S[x^T(I-M)^2x], \mathbf{E}_S[(x^T(I-M)x)\mathrm{Tr}(I-M)], \mathbf{E}_S[(\mathrm{Tr}(I-M))^2]$. Specifically, all of these expressions can be written as

$$\sum_{r_1=1}^{B}\sum_{r_2=1}^{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}]$$
$$\cdot (U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B\cdot((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D\cdot((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F\cdot((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H,$$

where $A,B,C,D,E,F,G,H\in\{0,1\}$ and $A+B+C+D+E+F+G+H=2$. We thus have

$$|\sum_{r_1=1}^{B}\sum_{r_2=1}^{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}]\cdot$$
$$\cdot (U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B\cdot((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D\cdot((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F\cdot((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H]|$$
$$\leq \frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^2)|U_{a_1}U_{a_2}^T|^A|U_{b_1}U_{b_2}^T|^B\cdot|(Ux)_{a_1}(Ux)_{a_2}|^C|(Ux)_{b_1}(Ux)_{b_2}|^D\cdot$$
$$\cdot|(Ux)_{a_1}(Ux)_{b_1}|^E|U_{a_1}U_{b_1}^T|^F\cdot|(Ux)_{a_2}(Ux)_{b_2}|^G|U_{a_2}U_{b_2}^T|^H.$$

We have $|U_aU_b^T|\leq ||U_a||_2\cdot||U_b||_2$ by Cauchy-Schwarz, and $|(Ux)_a|\leq ||U_a||_2\cdot O(\sqrt{\log n})$ since $x\in\mathcal{T}^*$ by assumption of the lemma, so

$$\frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^2)|U_{a_1}U_{a_2}^T|^A|U_{b_1}U_{b_2}^T|^B\cdot|(Ux)_{a_1}(Ux)_{a_2}|^C|(Ux)_{b_1}(Ux)_{b_2}|^D\cdot|(Ux)_{a_1}(Ux)_{b_1}|^E|U_{a_2}U_{b_2}^T|^F$$
$$\leq (O(\log n))^{C+D+E+G}\frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^2)(||U_{a_1}||_2||U_{a_2}||_2)^A\cdot(||U_{b_1}||_2||U_{b_2}||_2)^B\cdot(||U_{a_1}||_2||U_{a_2}||_2)^C$$
$$\cdot(||U_{b_1}||_2||U_{b_2}||_2)^D\cdot(||U_{a_1}||_2||U||_{b_1}||_2)^E(||U_{a_1}||_2||U_{b_1}||_2)^F\cdot(||U_{a_2}||_2||U||_{b_2}||_2)^G(||U_{a_2}||_2||U_{b_2}||_2)^H.$$

Since we are only summing over $\{a_1,a_2,b_1,b_2\}$ that contain two distinct elements, we have

$$(O(\log n))^{C+D+E+G}\frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^2)(||U_{a_1}||_2||U_{a_2}||_2)^A\cdot(||U_{b_1}||_2||U_{b_2}||_2)^B\cdot(||U_{a_1}||_2||U_{a_2}||_2)^C$$
$$\cdot(||U_{b_1}||_2||U_{b_2}||_2)^D\cdot(||U_{a_1}||_2||U||_{b_1}||_2)^E(||U_{a_1}||_2||U_{b_1}||_2)^F\cdot(||U_{a_2}||_2||U||_{b_2}||_2)^G(||U_{a_2}||_2||U_{b_2}||_2)^H$$
$$\leq (O(\log n))^{C+D+E+G}\frac{1}{B}\sum_{a_1,b_1=1}^{n}||U_{a_1}||_2^2||U_{a_2}||_2^2$$
$$\leq (O(\log n))^{C+D+E+G}\frac{1}{B}(\sum_{a_1=1}^{n}||U_{a_1}||_2^2)^2$$
$$\leq (O(\log n))^{C+D+E+G}\frac{d^2}{B},$$

where we used the fact that $\sum_a ||U_a||_2^2 = d$. Noting that $C+D+E+G=0$ for $\mathbf{E}_S[||I-M||_F^2]$ and $C+D+E+G=1$ for $\mathbf{E}_S[x^T(I-M)x\mathrm{Tr}(I-M)]$ completes the proof.

Bounding $\mathbf{E}_S[(x^T(I-M)x)^2\text{Tr}(I-M)], \mathbf{E}_S[x^T(I-M)^2x\cdot\text{Tr}(I-M)], \mathbf{E}_S[||I-M||_F^2\cdot\text{Tr}(I-M)], \mathbf{E}_S[(x^T(I-M)x)^2\cdot x^T(I-M)x], \mathbf{E}_S[x^T(I-M)^2x\cdot x^T(I-M)x], \mathbf{E}_S[||I-M||_F^2\cdot x^T(I-M)x]$ All of the above expressions can be written as

$$\sum_{r_1=1}^{B}\sum_{r_2=1}^{B}\sum_{r_3=1}^{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}]$$
$$\cdot (U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B\cdot((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D\cdot((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F\cdot((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H$$
$$\cdot((Ux)_{a_3}(Ux)_{b_3})^I(U_{a_3}U_{b_3}^T)^J$$

where $A, B, C\dots$ are in $\{0,1\}$ and $A+B+C+D+E+F+G+H+I+J=3$.

We first note that for for any r_1, r_2, r_3 and $a_1\neq b_1, a_2\neq b_2, a_3\neq b_3$ the quantity

$$\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}]$$

is only nonzero when $r_1 = r_2 = r_3$ and $\{a_1, b_1, a_2, b_2, a_3, b_3\}$ contains three distinct elements, each with multiplicity 2. Let $\mathbf{I}_*(\{a_q, b_q\}_{q=1}^3)$ denote the indicator of the latter condition. In that case one has $\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}] = 1/B^3$. Note we cannot have $a_1 = a_2 = a_3$ and $b_1 = b_2 = b_3$ since the expectation is 0 in that case.

Similarly to the above, it thus suffices to bound

$$\frac{1}{B^2}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^3)$$
$$\cdot|(U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B\cdot((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D\cdot((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F\cdot((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H$$
$$\cdot((Ux)_{a_3}(Ux)_{b_3})^I(U_{a_3}U_{b_3}^T)^J|$$
$$\leq (O(\log n))^{C+D+E+G+I}\frac{1}{B^2}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^3)\cdot$$
$$\cdot(||U_{a_1}||_2||U_{a_2}||_2)^A(||U_{b_1}||_2||U_{b_2}||_2)^B\cdot(||U_{a_1}||_2||U_{a_2}||_2)^C(||U_{b_1}||_2||U_{b_2}||_2)^D\cdot(||U_{a_1}||_2||U_{b_1}||_2)^E(||U_{a_1}||_2||U_{b_1}||_2)^F$$
$$\cdot(||U_{a_2}||_2||U_{b_2}||_2)^G(||U_{a_2}||_2||U_{b_2}||_2)^H\cdot(||U_{a_3}||_2||U_{b_3}||_2)^I(||U_{a_3}||_2||U_{b_3}||_2)^J$$

where we used Cauchy-Schwarz and the assumption that $x\in\mathcal{T}^*$ (and hence x is not correlated with any of the rows of U too much), as above.

Since we are only summing over $\{a_1, a_2, a_3, b_1, b_2, b_3\}$ that contain three distinct elements, the expression above is upper bounded by

$$\begin{aligned}&(O(\log n))^{C+D+E+G+I}\frac{1}{B^2}\sum_{a,c,b}^{n}||U_a||_2^2||U_b||_2^2||U_c||_2^2\\ &\leq (O(\log n))^{C+D+E+G+I}\frac{d^3}{B^2}\\ &\leq (O(\log n))^2\frac{d^2}{B},\end{aligned}$$

where we used the fact that $\sum_a ||U_a|_2^2 = d$ and that in all cases, $C+D+E+G+I\leq 2$.

Bounding $\mathbf{E}_S[((x^T(I-M)x)^2 + x^T(I-M)^2x + ||I-M||_F^2 + (\mathrm{Tr}(I-M))^2)^2]$ All of the pairwise products arising in the expansion of the above expressions can be written as

$$\sum_{r_1=1}^{B}\sum_{r_2=1}^{B}\sum_{r_3=1}^{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\sum_{\substack{a_4,b_4=1,\\ a_4\neq b_4}}^{n}\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}S_{r_4,a_4}S_{r_4,b_4}]$$
$$\cdot (U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B \cdot ((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D \cdot ((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F \cdot ((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H$$
$$\cdot (U_{a_3}U_{a_4}^T)^{A'}(U_{b_3}U_{b_4}^T)^{B'} \cdot ((Ux)_{a_3}(Ux)_{a_4})^{C'}((Ux)_{b_3}(Ux)_{b_4})^{D'} \cdot ((Ux)_{a_3}(Ux)_{b_3})^{E'}(U_{a_3}U_{b_3}^T)^{F'} \cdot ((Ux)_{a_4}(Ux)_{b_4})^{G'}(U_{a_4}U_{b_4}^T)^{H'},$$

where $A,B,C,D,E,F,G,H,A',B',C',D',E',F',G',H' \in \{0,1\}$ and add up to 4.

We now need to consider two cases.

Case 1: the number of distinct elements in $\{a_1,b_1,a_2,b_2,a_3,b_3,a_4,b_4\}$ is four, each occurring with multiplicity 2 (let $\mathbf{I}_*(\{a_q,b_q\}_{q=1}^4)$ denote the indicator of the latter condition) Then

$$\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}S_{r_4,a_4}S_{r_4,b_4}]$$

contributes $1/B^4$. In this case the number of distinct elements in $\{r_1,r_2,r_3,r_4\}$ cannot be larger than 2.

It thus suffices to bound

$$\frac{1}{B^2}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\sum_{\substack{a_4,b_4=1,\\ a_4\neq b_4}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^4)\cdot$$
$$\cdot |(U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B \cdot ((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D \cdot ((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F \cdot ((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H$$
$$\cdot (U_{a_3}U_{a_4}^T)^{A'}(U_{b_3}U_{b_4}^T)^{B'} \cdot ((Ux)_{a_3}(Ux)_{a_4})^{C'}((Ux)_{b_3}(Ux)_{b_4})^{D'} \cdot ((Ux)_{a_3}(Ux)_{b_3})^{E'}(U_{a_3}U_{b_3}^T)^{F'} \cdot ((Ux)_{a_4}(Ux)_{b_4})^{G'}(U_{a_4}U_{b_4}^T)^{H'}|$$
$$\leq (O(\log n))^2\frac{1}{B^2}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^4)\cdot$$
$$\cdot (||U_{a_1}||_2||U_{a_2}||_2)^A(||U_{b_1}||_2||U_{b_2}||_2)^B \cdot (||U||_{a_1}||U_{a_2}||)^C(||U_{b_1}||_2||U_{b_2}||_2)^D \cdot (||U_{a_1}||_2||U_{b_1}||_2)^E(||U_{a_1}||_2||U_{b_1}||_2)^F$$
$$\cdot (||U_{a_2}||||U_{b_2}||)^G(||U_{a_2}||_2||U_{b_2}||_2)^H$$
$$\cdot (||U_{a_3}||_2||U_{a_4}||_2)^{A'}(||U_{b_3}||_2||U_{b_4}||_2)^{B'} \cdot (||U_{a_3}||_2||U_{a_4}||_2)^{C'}(||U_{b_3}||_2||U_{b_4}||_2)^{D'} \cdot (||U_{a_3}||_2||U_{b_3}||_2)^{E'}(||U_{a_3}||_2||U_{b_3}||_2)^{F'}$$
$$\cdot (||U_{a_4}||_2||U_{b_4}||_2)^{G'}(||U_{a_4}||_2||U_{b_4}||_2)^{H'}$$

where we used Cauchy-Schwarz and the assumption that $x \in \mathcal{T}^*$ (and hence x is not correlated with any of the rows of U too much), as above.

Since we are only summing over $\{a_1,a_2,a_3,a_4,b_1,b_2,b_3,b_4\}$ that contain three distinct elements, each of multiplicity two, the expression above is upper bounded by

$$\begin{aligned}&(O(\log n))^2\frac{1}{B^2}\sum_{a,b,c,d}^{n}||U_a||_2^2||U_b||_2^2||U_c||_2^2||U_d||_2^2\\ &\leq (O(\log n))^2\frac{d^4}{B^2}\\ &\leq (O(\log n))^2\frac{d^2}{B},\end{aligned}$$

where we used the fact that $\sum_a ||U_a|_2^2 = d$.

Case 2: the number of distinct elements in $\{a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4\}$ is two, each occurring with multiplicity 4 (let $\mathbf{I}_*(\{a_q, b_q\}_{q=1}^4)$ denote the indicator of the latter condition) Then

$$\mathbf{E}_S[S_{r_1,a_1}S_{r_1,b_1}S_{r_2,a_2}S_{r_2,b_2}S_{r_3,a_3}S_{r_3,b_3}S_{r_4,a_4}S_{r_4,b_4}]$$

contributes $1/B^2$. In this case the number of distinct elements in $\{r_1, r_2, r_3, r_4\}$ has to be one, since each column of S has a single non-zero entry and necessarily $a_1 = a_2 = a_3 = a_4$ and $b_1 = b_2 = b_3 = b_4$.

It thus suffices to bound

$$\begin{aligned}
&\frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\sum_{\substack{a_4,b_4=1,\\ a_4\neq b_4}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^4)\cdot\\
&\cdot|(U_{a_1}U_{a_2}^T)^A(U_{b_1}U_{b_2}^T)^B\cdot((Ux)_{a_1}(Ux)_{a_2})^C((Ux)_{b_1}(Ux)_{b_2})^D\cdot((Ux)_{a_1}(Ux)_{b_1})^E(U_{a_1}U_{b_1}^T)^F\cdot((Ux)_{a_2}(Ux)_{b_2})^G(U_{a_2}U_{b_2}^T)^H\\
&\cdot(U_{a_3}U_{a_4}^T)^{A'}(U_{b_3}U_{b_4}^T)^{B'}\cdot((Ux)_{a_3}(Ux)_{a_4})^{C'}((Ux)_{b_3}(Ux)_{b_4})^{D'}\cdot((Ux)_{a_3}(Ux)_{b_3})^{E'}(U_{a_3}U_{b_3}^T)^{F'}\cdot((Ux)_{a_4}(Ux)_{b_4})^{G'}(U_{a_4}U_{b_4}^T)^{H'}|\\
&\leq (O(\log n))^2\frac{1}{B}\sum_{\substack{a_1,b_1=1,\\ a_1\neq b_1}}^{n}\sum_{\substack{a_2,b_2=1,\\ a_2\neq b_2}}^{n}\sum_{\substack{a_3,b_3=1,\\ a_3\neq b_3}}^{n}\sum_{\substack{a_4,b_4=1,\\ a_4\neq b_4}}^{n}\mathbf{I}_*(\{a_q,b_q\}_{q=1}^4)\cdot\\
&\cdot(||U_{a_1}||_2||U_{a_2}||_2)^A(||U_{b_1}||_2||U_{b_2}||_2)^B\cdot(||U||_{a_1}||U_{a_2}||)^C(||U_{b_1}||_2||U_{b_2}||_2)^D\cdot(||U_{a_1}||_2||U_{b_1}||_2)^E(||U_{a_1}||_2||U_{b_1}||_2)^F\\
&\cdot(||U_{a_2}||||U_{b_2}||)^G(||U_{a_2}||_2||U_{b_2}||_2)^H\\
&\cdot(||U_{a_3}||_2||U_{a_4}||_2)^{A'}(||U_{b_3}||_2||U_{b_4}||_2)^{B'}\cdot(||U_{a_3}||_2||U_{a_4}||_2)^{C'}(||U_{b_3}||_2||U_{b_4}||_2)^{D'}\cdot(||U_{a_3}||_2||U_{b_3}||_2)^{E'}(||U_{a_3}||_2||U_{b_3}||_2)^{F'}\\
&\cdot(||U_{a_4}||_2||U_{b_4}||_2)^{G'}(||U_{a_4}||_2||U_{b_4}||_2)^{H'}
\end{aligned}$$

where we used Cauchy-Schwarz and the assumption that $x \in \mathcal{T}^*$ (and hence x is not correlated with any of the rows of U too much), as above.

Since we are only summing over $\{a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4\}$ that contain two distinct elements, each of multiplicity four, the expression above is upper bounded by

$$\begin{aligned}
&(O(\log n))^2\frac{1}{B}\sum_{a,b}^{n}||U_a||_2^4||U_b||_2^4\\
&=(O(\log n))^2\frac{1}{B}\sum_{a,b}^{n}||U_a||_2^2||U_b||_2^2 \qquad \text{(since } ||U_a||_2\leq 1 \text{ for all } a)\\
&\leq (O(\log n))^2\frac{d^2}{B},
\end{aligned}$$

where we used the fact that $\sum_a ||U_a|_2^2 = d$.

□