Iterative Double Sketching for Faster Least-Squares Optimization

Rui Wang [1] **Yanyan Ouyang** [1] **Wangli Xu** [1]

Abstract

This work is concerned with the overdetermined linear least-squares problem for large scale data. We generalize the iterative Hessian sketching (IHS) algorithm and propose a new sketching framework named iterative double sketching (IDS) which uses approximations for both the gradient and the Hessian in each iteration. To understand the behavior of the IDS algorithm and choose the optimal hyperparameters, we derive the exact limit of the conditional prediction error of the IDS algorithm in the setting of Gaussian sketching. Guided by this theoretical result, we propose an efficient IDS algorithm via a new class of sequentially related sketching matrices. We give a non-asymptotic analysis of this efficient IDS algorithm which shows that the proposed algorithm achieves the state-of-the-art trade-off between accuracy and efficiency.

1. Introduction

We consider the overdetermined least-squares problem

$$\mathbf{x}^* := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}; \mathbf{A}, \mathbf{y}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 \right\}, \tag{1}$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_N)^\top \in \mathbb{R}^{N\times d}$ is a given data matrix and $\mathbf{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ is a vector of observations. Throughout the paper, it is assumed that $\mathbf{A}$ has full column rank, that is, $\mathrm{Rank}(\mathbf{A}) = d$.

The solution to the problem (1) has the explicit expression

$$\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{y}. \tag{2}$$

The direct computation of $\mathbf{x}^*$ using this formula costs $O(Nd^2)$ time. Note that for any initial point $\mathbf{x}_0 \in \mathbb{R}^d$,

$$\mathbf{x}^* = \mathbf{x}_0 - (\mathbf{A}^\top\mathbf{A})^{-1}\nabla f(\mathbf{x}_0; \mathbf{A}, \mathbf{y}),$$

where $\nabla f(\mathbf{x}; \mathbf{A}, \mathbf{y}) := \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})$ is the gradient of f at $\mathbf{x}$, and $\mathbf{A}^\top\mathbf{A}$ is the Hessian of f. Thus, from any initial point $\mathbf{x}_0$, one can obtain the exact solution to the problem (1) in just one Newton iteration. And the formula (2) is the output of one Newton iteration with $\mathbf{x}_0 = \mathbf{0}_d$.

When N and d are large, the direct computation via the formula (2) may be time-consuming. In this case, sketching methods are often used to obtain approximate solutions to the problem (1). For the classical sketching methods, the data $(\mathbf{A}, \mathbf{y})$ is projected to $(\mathbf{SA}, \mathbf{Sy})$ via certain random sketching matrix $\mathbf{S} \in \mathbb{R}^{r\times N}$ with $d \le r \ll N$, and the sketched least-squares problem

$$\mathbf{x}_{\mathrm{CS}} := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}; \mathbf{SA}, \mathbf{Sy}) = \frac{1}{2}\|\mathbf{SA}\mathbf{x} - \mathbf{Sy}\|^2 \right\}$$

is considered as a surrogate of the original problem; see Mahoney (2011); Woodruff (2014); Drineas & Mahoney (2016) for reviews of classical sketching methods. If the matrix $\mathbf{A}^\top\mathbf{S}^\top\mathbf{SA}$ is invertible, then we have

$$\mathbf{x}_{\mathrm{CS}} = \mathbf{x}_0 - (\mathbf{A}^\top\mathbf{S}^\top\mathbf{SA})^{-1}\nabla f(\mathbf{x}_0; \mathbf{SA}, \mathbf{Sy}). \tag{3}$$

Once the sketched data $(\mathbf{SA}, \mathbf{Sy})$ is obtained, the direct computation of $\mathbf{x}_{\mathrm{CS}}$ via the formula (3) only costs $O(rd^2)$ time which can be much faster than the direct computation of $\mathbf{x}^*$. With a small r, however, the classical sketching method can only produce a low-precision approximation of $\mathbf{x}^*$. Recently, Pilanci & Wainwright (2016) introduced the iterative Hessian sketch (IHS) algorithm which uses sketching methods in conjunction with an iteration method to achieve high-precision approximations of $\mathbf{x}^*$. They considered the following iteration formula:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{A})^{-1}\nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y}), \tag{4}$$

where $\mathbf{S}_0, \mathbf{S}_1, \ldots$ are independent and identically distributed (i.i.d.) $r\times N$ random sketching matrices. Compared with the classical sketching methods, the formula (4) of the IHS algorithm has two new features. First, the IHS algorithm only sketches the Hessian and does not sketch the gradient. Second, the IHS algorithm applies multiple Newton iterations to refine the solution. The theoretical results of Pilanci & Wainwright (2016) guarantee that with high probability, the IHS algorithm can produce a high-precision approximation of $\mathbf{x}^*$.

Since the publication of Pilanci & Wainwright (2016), the IHS algorithm has drawn much attention and several improvements are proposed. In the original iteration formula

[1]Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China. Correspondence to: Wangli Xu <wlxu@ruc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

(4) of the IHS algorithm, a refreshed Hessian sketching $\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{A}$ is used for each iteration. The computation of refreshed Hessian sketching is time-consuming. Some recent work reveals that the IHS algorithm can also work well with fixed Hessian sketching across iterations, that is, $\mathbf{S}_0 = \mathbf{S}_1 = \cdots$; see, e.g., Wang & Xu (2018); Özaslan et al. (2019); Lacotte & Pilanci (2021). With a fixed Hessian sketching, one can first compute and cache the matrix $(\mathbf{A}^\top \mathbf{S}_0^\top \mathbf{S}_0 \mathbf{A})^{-1}$. Then in each iteration, one only need to compute the gradient $\nabla f(\mathbf{x}_t, ; \mathbf{A}, \mathbf{y})$, and update $\mathbf{x}_t$ via the formula $\mathbf{x}_{t+1} = \mathbf{x}_t - (\mathbf{A}^\top \mathbf{S}_0^\top \mathbf{S}_0 \mathbf{A})^{-1} \nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y})$. In another direction of research, more general update formulas are considered to improve the IHS algorithm. For example, Özaslan et al. (2019) considered the update formula

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha (\mathbf{A}^\top \mathbf{S}_0^\top \mathbf{S}_0 \mathbf{A})^{-1} \nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y}) + \beta (\mathbf{x}_t - \mathbf{x}_{t-1}),$$

and investigated fine-grained choices of α and β. See Lacotte & Pilanci (2020; 2021) for some recent work in this direction.

The IHS algorithm and its variants have achieved great success for the problem (1). Nevertheless, there is still room for improvement. To appreciate this point, we note that the typical convergence rate of the IHS algorithms has the form

$$\|\mathbf{A}(\mathbf{x}_t - \mathbf{x}^*)\| \leq \rho^t \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|, \tag{5}$$

where $\rho \in (0, 1)$ is certain constant; see, e.g., Özaslan et al. (2019); Lacotte & Pilanci (2020; 2021). Such a convergence rate is often tight for the IHS algorithms. Hence roughly speaking, in each iteration of the IHS algorithm, the error is reduced by a constant ratio. On the other hand, for each iteration of the IHS algorithm, one need to compute the gradient $\nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y}) = \mathbf{A}^\top (\mathbf{A}\mathbf{x}_t - \mathbf{y})$ which costs $O(Nd)$ time. That is, in each iteration of the IHS algorithm, it costs at least $O(Nd)$ time to reduce the error by a constant ratio. For the first few iterations, this may not achieve a good trade-off between the accuracy and efficiency.

The computation of the gradient is a bottleneck of the IHS algorithm. To improve the IHS algorithm, we consider a general sketching framework, named *iterative double sketching* (IDS), which uses not only a sketched Hessian but also a sketched gradient in each iteration. While using a sketched gradient may negatively affect the convergence property of the algorithm, it can reduce the computing time significantly, and hence allows for more iterations within a given computing time. Hence it can be expected that there exists an IDS algorithm which can achieve much better performance than the IHS algorithm. However, there are two main challenges in the design of such an IDS algorithm.

The first challenge is how to choose the sketch sizes of gradient sketching in each iteration. We note that the classical sketching and IHS are two extreme cases in terms of the sketch sizes of gradient sketching. In fact, for the classical sketching method, the sketch size of gradient is r which is the same as the sketch size of Hessian. In contrast, for the IHS algorithm, the gradient is computed using the full data, or in other words, the sketch size for gradient sketching is N for all iterations. Unfortunately, neither of these two choices is optimal. To design a concrete IDS algorithm, we would like to choose the optimal sketch sizes of gradient sketching in each iteration.

The second challenge is how to efficiently compute the sketched gradient in each iteration. We note that the Hessian sketching can be efficiently computed using existing sketching matrices such as the Subsampled Randomized Hadamard Transform (SRHT) (Sarlós, 2006; Ailon & Chazelle, 2009) or sparse Johnson-Lindenstrauss transforms (Kane & Nelson, 2014). However, it is not efficient to use these sketching matrices to compute the sketched gradient in each iteration. In fact, for dense data matrix, the application of these sketching matrices requires accesssing each element of data matrix at least once, which costs at least $O(Nd)$ time. The computing time $O(Nd)$ is undesirable. In fact, we can even compute the exact gradient within $O(Nd)$ time. Thus, a new sketching method is required to efficiently compute the sketched gradient in each iteration.

The goal of the present work is to investigate the above two challenges and propose an efficient IDS algorithm which has guaranteed good performance.

1.1. Our Contributions

We propose the general IDS framework. To theoretically understand the behavior of the IDS algorithm, we give an asymptotic analysis of the IDS algorithm in the setting of Gaussian sketching. In this setting, we derive the exact limit of the conditional prediction error of the IDS algorithm. Based on this result, we obtain the optimal sketch sizes of gradient sketching such that the limiting conditional prediction error is minimized under the constraint of a given computational cost. While these results are interesting in theory, the Gaussian sketching is not efficient to apply. Nevertheless, these results provide a general guidance on how to choose the sketch sizes of gradient sketching.

We propose a new class of sequentially related sketching matrices named *iteration efficient sketching*. The iteration efficient sketching matrices can be efficiently applied to obtain the sketched gradient in each iteration of the IDS algorithm. We establish the embedding properties of the iteration efficient sketching matrix.

We design an efficient IDS algorithm. In the proposed algorithm, the choice of the sketch size of gradient sketching is guided by our theoretical results in the Gaussian sketching. We use the proposed iteration efficient sketching matrices to efficiently compute the sketched gradient in each iteration. We give a non-asymptotic analysis of the proposed IDS algorithm. As shown in Table 1, for a wide range of the

Table 1. The computing time of various algorithms to achieve ϵ relative error, i.e., $\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\| \leq \epsilon \|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|$. Here IHS is the algorithm in Lacotte & Pilanci (2020), PCG is the algorithm in Lacotte & Pilanci (2021) and IDS is Algorithm 3 of the present paper. We assume that $N = \Omega(d^2)$, and $\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|$ has the same order of magnitude as $\sqrt{\frac{d}{r}}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|$.

METHODS	COMPUTING TIME
IHS IN LACOTTE & PILANCI (2020)	$O\left((\log(d) + \log(\frac{1}{\epsilon}))Nd + d^3\right)$
PCG IN LACOTTE & PILANCI (2021)	$O\left(\left(\log(d) + \max\left(\sqrt{\log(\frac{1}{\epsilon})}, \frac{\log(\frac{1}{\epsilon})}{\log(\frac{N}{d^2})}\right)\right)Nd\right)$
IDS (ALGORITHM 3)	$O\left(\max\left(1, \log_2(\frac{1}{\epsilon}) - \frac{1}{2}\log_2(\frac{N}{d(\log(d))^3})\right)Nd + d^3\log(d)\right)$

error parameter ϵ, the proposed IDS algorithm improves the state-of-the-art computing time for high-precision least-squares problem. We conduct experiments to verify the good performance of the IDS algorithm in practice.

1.2. Related Work

Classical sketching methods for the problem (1) were extensively researched in the field of theoretical computer science and applied mathematics; see, e.g., Mahoney (2011); Woodruff (2014); Drineas & Mahoney (2016) and the references therein. For classical sketching methods for the problem (1), the algorithm precision is directly connected to the *subspace embedding* property of the sketching matrix $\mathbf{S}$, which refers to the norm preserving property of $\mathbf{S}$; see, e.g., Woodruff (2014). Perhaps the most classical sketching matrix is the Gaussian sketching matrix whose elements are independent normal random variables. The celebrated Johnson-Lindenstrauss lemma (Johnson & Lindenstrauss, 1984) implies that the Gaussian sketching matrix has good subspace embedding property. However, the Gaussian sketching matrix is not efficient to apply. In fact, for an $m \times N$ Gaussian sketching matrix $\mathbf{A}$, the direct computation of $\mathbf{SA}$ costs $O(mNd)$ time.

In recent years, several alternative sketching matrices have been proposed which are more efficient to apply. A popular fast sketching matrix is the SRHT; see Sarlós (2006); Ailon & Chazelle (2009). For an $m \times N$ SRHT sketching matrix $\mathbf{S}$, the computation of $\mathbf{SA}$ can be completed within $O(Nd\log(m))$ time; see Theorem 2.1 of Ailon & Liberty (2009). Also, it is known that the SRHT matrix has good subspace embedding property; see, e.g., Tropp (2011); Boutsidis & Gittens (2013); Cohen et al. (2016). Another widely used sketching matrix is the CountSketch matrix which is even faster to apply. The CountSketch matrix stems from the data stream literature (Charikar et al., 2004; Thorup & Zhang, 2004), and is later used to construct sparse subspace embedding (Dasgupta et al., 2010; Clarkson & Woodruff, 2013; 2017). The CountSketch matrix $\mathbf{S}$ has a single non-zero element per column. As a result, the sketching method based on CountSketch can achieve input-sparsity time. While the CountSketch matrix is fast to apply, it may require a large m to achieve good subspace embedding property; see, e.g., Woodruff (2014). See Meng & Mahoney (2013); Nelson & Nguyen (2013); Kane & Nelson (2014); Allen-Zhu et al. (2014); Bourgain et al. (2015); Cohen et al. (2018); Jagadeesan (2019) for further analyses on general sparse subspace embeddings.

The sketching methods can be used in conjunction with iterative algorithms to achieve to achieve high-precision approximation in the least-squares problem (1). To the best of our knowledge, the first work in this direction is made by Rokhlin & Tygert (2008) who proposed a preconditioned conjugate gradient (PCG) algorithm based on sketching. A similar idea was used in Avron et al. (2010). Lacotte & Pilanci (2021) proposed a PCG algorithm which achieves the current state-of-the-art computing time high-precision least-squares problem in the regime $N > d^2$. In another line of research, Pilanci & Wainwright (2016) proposed the IHS algorithm which has become a popular method since then. We have mentioned some recent achievements on the theoretical understanding and methodological improvement of the IHS algorithm for the least-squares problem (1). To the best of our knowledge, the fastest variant of the IHS algorithm in the literature is the algorithm in Lacotte & Pilanci (2020). In addition to the sketching algorithms, there are some general purpose optimization algorithms that can be applied to solve the least-squares problem; see, e.g., (Lan et al., 2019) and the references therein. However, to the best of our knowledge, general purpose algorithms can not yield state-of-the-art computing time for the least-squares problem. Table 1 lists the computing time of the proposed IDS algorithm and the current state-of-the-art algorithms.

2. Iterative Double Sketching Framework

In this section, we introduce the proposed IDS Framework. The IDS algorithm is an iterative sketching method which uses both gradient sketching and Hessian sketching. The Hessian sketching is fixed across all iterations. Let $\tilde{\mathbf{S}} \in \mathbb{R}^{r \times N}$ be the sketching matrix for Hessian approximation. The sketched Hessian is defined as $\tilde{\mathbf{H}} := \mathbf{A}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \mathbf{A}$.

Algorithm 1 Generic iterative double sketching

Input: $\mu, T, \tilde{\mathbf{S}}\mathbf{A}, (\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y}), t = 0, \dots, T-1$
$\tilde{\mathbf{H}}^{-1} \leftarrow (\mathbf{A}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}\mathbf{A})^{-1}$
$\mathbf{x}_0 \leftarrow \tilde{\mathbf{H}}^{-1}\mathbf{A}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}\mathbf{y}$
for $t \leftarrow 0$ **to** $T-1$ **do**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \mu \tilde{\mathbf{H}}^{-1} \nabla f(\mathbf{x}_t; \mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$
end for
Return $\mathbf{x}_T$

We use the classical sketching method based on $(\tilde{\mathbf{S}}\mathbf{A}, \tilde{\mathbf{S}}\mathbf{y})$ to obtain the initial point $\mathbf{x}_0 := \tilde{\mathbf{H}}^{-1}\mathbf{A}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}\mathbf{y}$. Let $\mathbf{S}_t \in \mathbb{R}^{m_t \times N}$ be the sketching matrix for gradient approximation when computing $\mathbf{x}_{t+1}$, $t = 0, \dots, T-1$. Here $m_0, \dots, m_{T-1}$ are the sketch sizes for gradient sketching in each step. We consider the update formula

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mu \tilde{\mathbf{H}}^{-1} \nabla f(\mathbf{x}_t; \mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y}), \tag{6}$$

where $\mu > 0$ is the step size parameter. Note that the original IHS uses the step size $\mu = 1$. Here we introduce μ to allow for fine-grained choices of the step size. We summarize the generic IDS algorithm in Algorithm 1.

For small t, $\mathbf{x}_t$ is relatively far from $\mathbf{x}^*$, and an approximate gradient may be sufficient to ensure that $\mathbf{x}_t$ moves toward $\mathbf{x}^*$. As t increases, $\mathbf{x}_t$ gets closer to $\mathbf{x}^*$, and a gradient with higher precision may be necessary to ensure that $\mathbf{x}_t$ moves toward $\mathbf{x}^*$. In this view, m_t should be an increasing function of t. Formally, we require that the sketch sizes $r, m_0, \dots, m_{T-1}$ satisfy $d \le r \le m_0 \le \cdots \le m_{T-1} \le N$. If m_t reaches N for some t, then there is no need to sketch the gradient, and we use the exact gradient instead. Formally, we require that if $m_t = N$, then $\mathbf{S}_t = \mathbf{I}_N$. Define

$$T^\dagger := \min\left(\{t : 0 \le t < T \text{ and } m_t = N\} \cup \{T\}\right).$$

Hence for $t = 0, \dots, T^\dagger - 1$, sketched data is used to approximate the gradient; for $t = T^\dagger, \dots, T-1$, the full data is used to compute the exact gradient. We call the iterations for $t = 0, \dots, T^\dagger - 1$ the IDS iteration, and call the iterations for $t = T^\dagger, \dots, T-1$ the IHS iteration. If $T^\dagger = 0$, then all iterations are IHS iterations, and Algorithm 1 reduces to the IHS algorithm.

Algorithm 1 relies on the given sketched data $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$, $t = 0, \dots, T-1$. However, to obtain these sketched data via the vanilla method, one needs to access the full data for T times, which is highly inefficient. To reduce the computational cost of the IDS algorithm, we assume that for $t = 0, \dots, T^\dagger - 2$, $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$ is a function of $(\mathbf{S}_{t+1}\mathbf{A}, \mathbf{S}_{t+1}\mathbf{y})$. In this case, we can compute the sketched data $(\mathbf{S}_{T^\dagger-1}\mathbf{A}, \mathbf{S}_{T^\dagger-1}\mathbf{y}), \dots, (\mathbf{S}_0\mathbf{A}, \mathbf{S}_0\mathbf{y})$ sequentially. We shall propose new sketching matrices such that the sketched data can be efficiently computed in this sequential manner.

Now we consider the cost of floating point operations (FLOPs) in the iteration steps in Algorithm 1. The FLOPs of basic matrix computations are counted as in Section 1.1.15 of Golub & Van Loan (2013). Given $\mathbf{x}_t$, $\tilde{\mathbf{H}}^{-1}$ and $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$, the computation of $\mathbf{x}_{t+1}$ via the update formula (6) costs $\{(4d+1)m_t + 2d^2 + 2d\}$ FLOPs, $t = 0, \dots, T-1$. Thus, given $\mathbf{x}_0$, $\tilde{\mathbf{H}}^{-1}$ and $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$, $t = 0, \dots, T-1$, the total T iterations of the IDS algorithm cost $\{(4d+1)\sum_{t=0}^{T-1} m_t + 2d^2T + 2dT\}$ FLOPs. For the IHS algorithm, where $m_t = N$, $t = 0, \dots, T-1$, the total T iterations cost $\{(4d+1)NT + 2d^2T + 2dT\}$ FLOPs. In this expression, the leading term is $(4d+1)NT$. With the same number T of iterations, the IDS algorithm reduces this term to $(4d+1)\sum_{t=0}^{T-1} m_t$. However, to achieve a given level of precision, the IDS algorithm may need more iterations than the IHS algorithm.

Compared with the IHS algorithm, the IDS framework involves additional hyperparameters $m_0, \dots, m_{T-1}$. The choice of these hyperparameters largely affects the performance of the IDS algorithm. How to choose the optimal $m_0, \dots, m_{T-1}$ is a challenge in the IDS framework. Another challenge in the IDS framework is how to design the sketching matrices $\mathbf{S}_0, \dots, \mathbf{S}_{t-1}$ such that $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$ can be efficiently computed from $(\mathbf{S}_{t+1}\mathbf{A}, \mathbf{S}_{t+1}\mathbf{y})$ and $\mathbf{S}_t$ has good embedding property. We shall deal with these two challenges in the following sections.

3. IDS with Gaussian Sketching

In this section, we investigate the asymptotic properties of the IDS algorithm with Gaussian sketching matrices. A Gaussian sketching matrix $\mathbf{S} \in \mathbb{R}^{m \times N}$ is a random matrix whose elements are independent with distribution $\mathcal{N}(0, \frac{1}{m})$. Recall that we require that the sketched data can be obtained sequentially. To meet this requirement, we construct the sketching matrices as follows. Let $\mathbf{C}$ be an $N \times N$ random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Let $\mathbf{C}_k$ denote the $k \times N$ matrix consisting of the first k rows of $\mathbf{C}$, $k = 1, \dots, N$. The sketching matrices are defined as

$$\tilde{\mathbf{S}} = \frac{1}{\sqrt{r}}\mathbf{C}_r, \quad \mathbf{S}_t = \frac{1}{\sqrt{m_t}}\mathbf{C}_{m_t}, \quad t = 0, \dots, T-1. \tag{7}$$

With the above construction, $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$ is the first m_t rows of $\frac{\sqrt{m_{t+1}}}{\sqrt{m_t}}(\mathbf{S}_{t+1}\mathbf{A}, \mathbf{S}_{t+1}\mathbf{y})$. Hence once $(\mathbf{S}_{T^\dagger-1}\mathbf{A}, \mathbf{S}_{T^\dagger-1}\mathbf{y})$ is obtained, the sketched data can be obtained sequentially.

With the above construction of sketching matrices, the sketched data are equivalent to subsamples of the preconditioned data $(\tilde{\mathbf{A}}, \tilde{\mathbf{y}}) := (\mathbf{C}\mathbf{A}, \mathbf{C}\mathbf{y})$. To see this, denote $\tilde{\mathbf{A}} = (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_N)^\top$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_N)^\top$. Let $\tilde{\mathbf{A}}_k := (\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_k)^\top$ denote the matrix of the first k rows of $\tilde{\mathbf{A}}$, and $\tilde{\mathbf{y}}_k := (\tilde{y}_1, \dots, \tilde{y}_k)^\top$ denote the vector of the first k elements of $\tilde{\mathbf{y}}$, $k = 1, \dots, N$. Then we have $(\tilde{\mathbf{S}}\mathbf{A}, \tilde{\mathbf{S}}\mathbf{y}) = \frac{1}{\sqrt{r}}(\tilde{\mathbf{A}}_r, \tilde{\mathbf{y}}_r)$ and $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y}) = \frac{1}{\sqrt{m_t}}(\tilde{\mathbf{A}}_{m_t}, \tilde{\mathbf{y}}_{m_t})$, $t = 0, \dots, T-1$. With above notations,

the update formula (6) becomes

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mu \frac{r}{m_t} (\tilde{\mathbf{A}}_r^\top \tilde{\mathbf{A}}_r)^{-1} \tilde{\mathbf{A}}_{m_t}^\top (\tilde{\mathbf{A}}_{m_t} \mathbf{x}_t - \tilde{\mathbf{y}}_{m_t}). \quad (8)$$

While the original data $(\mathbf{A}, \mathbf{y})$ is non-random, the preconditioned data $(\tilde{\mathbf{A}}, \tilde{\mathbf{y}})$ has good statistical properties. In fact, the rows of $(\tilde{\mathbf{A}}, \tilde{\mathbf{y}})$ are independent and

$$\begin{pmatrix} \tilde{y}_k \\ \tilde{\mathbf{a}}_k \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{d+1}, \begin{pmatrix} \mathbf{y}^\top \mathbf{y} & \mathbf{y}^\top \mathbf{A} \\ \mathbf{A}^\top \mathbf{y} & \mathbf{A}^\top \mathbf{A} \end{pmatrix} \right), \quad k = 1, \dots, N.$$

Equivalently,

$$\tilde{\mathbf{a}}_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{A}^\top \mathbf{A}), \quad \tilde{y}_k \mid \tilde{\mathbf{a}}_k \sim \mathcal{N}(\tilde{\mathbf{a}}_k^\top \mathbf{x}^*, \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2).$$

That is, $\tilde{\mathbf{a}}_k$ and $\tilde{y}_k$ satisfy the Gaussian linear model

$$\tilde{y}_k = \tilde{\mathbf{a}}_k^\top \mathbf{x}^* + \varepsilon_k, \quad (9)$$

where $\varepsilon_k \sim \mathcal{N}(0, \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2)$ and ε_k is independent of $\tilde{\mathbf{a}}_k$. Thus, with Gaussian preconditioning, the problem (1) is statistically equivalent to the estimation problem of Gaussian linear model (9). Hence as a by-product, we obtain a fast algorithm with update formula (8) for the estimation problem of Gaussian linear model.

Now we derive the asymptotic behavior of the IDS algorithm with Gaussian sketching matrices. We consider the asymptotic scenario where T is fixed, $N \to \infty$ and the quantities μ, d, $\mathbf{A}$, $\mathbf{y}$ and m_t, $t = 0, \dots, T-1$ are functions of N. For nonnegative integer n, let $C(n) := \frac{(2n)!}{(n+1)!n!}$ denote the nth Catalan number. By definition, $C(0) := 1$. We have the following theorem.

Theorem 3.1. *Suppose $\mathbf{x}_T$ is the output of Algorithm 1 where the sketching matrices are defined in (7). Suppose as $N \to \infty$, the iteration number T is fixed, $m_{T-1} < N$ and*

$$d \to \infty, \ \frac{d}{r} \to 0, \ \frac{r}{m_0} \to 0, \ \frac{m_t}{m_{t+1}} \to 0, \ t = 0, \dots, T-2.$$

Suppose the step size μ satisfies $|\mu - 1| = O(\frac{d}{r})$. Then as $N \to \infty$,

$$\begin{aligned} &\mathrm{E}\left\{ \left\| \mathbf{A}(\mathbf{x}_T - \mathbf{x}^*) \right\|^2 \mid \tilde{\mathbf{A}} \right\} \\ =&(1 + o_P(1)) \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2 \left\{ \left(\frac{d}{r} \right)^{T+1} C(T) + \sum_{t=0}^{T-1} \frac{g(t,T)}{m_t} \right\}, \end{aligned}$$

where

$$g(t,T) := C(T-t-1) \frac{d^{T-t}}{r^{T-t-1}}, \quad t = 0, \dots, T-1.$$

In Theorem 3.1, the assumption $m_{T-1} < N$ ensures that $\mathbf{x}_T$ is the output of an IDS iteration. Theorem 3.1 explicitly characterizes how the conditional prediction error of $\mathbf{x}_T$ depends on the sketch sizes $m_0, \dots, m_{T-1}$. Now we use this result to determine the optimal choice of $m_0, \dots, m_{T-1}$. The idea is to minimize the limiting conditional prediction error under the constraint of a given FLOPs. Recall that the T iterations of Algorithm 1 cost $\{(4d+1)\sum_{t=0}^{T-1} m_t + 2d^2T + 2dT\}$ FLOPs, which relies on $m_0, \dots, m_{T-1}$ through their sum $\sum_{t=0}^{T-1} m_t$. Thus, we consider the following constrained optimization problem

$$\begin{aligned} &\min_{m_0, \dots, m_{T-1} \in \mathbb{R}} \sum_{t=0}^{T-1} \frac{g(t,T)}{m_t} \quad (10) \\ \text{s. t.} \quad &\sum_{t=0}^{T-1} m_t \le M \quad \text{and} \quad 0 < m_t \le N, \, t = 0, \dots, T-1, \end{aligned}$$

where $M > 0$ is a prespecified upper bound of $\sum_{t=0}^{T-1} m_t$. In practice, $m_0, \dots, m_{T-1}$ should be positive integers. In (10), however, we relax this restriction and allow $m_0, \dots, m_{T-1}$ to be positive real numbers, which makes the problem easier to solve. Of course, one can round the solution for practical use. We propose an algorithm to solve the problem (10), which is summarized in Algorithm 2.

Algorithm 2 Optimal sketch sizes

Input: $g(0,T), \dots, g(T-1,T)$, M, N
$t \leftarrow T-1$
for $i \leftarrow T-1$ **to** 0 **do**
  $m_i \leftarrow \frac{M\sqrt{g(i,T)}}{\sum_{j=0}^{t} \sqrt{g(j,T)}}$
  if $m_i \ge N$ **then**
    $t \leftarrow i-1$
    $m_i \leftarrow N$
    $M \leftarrow M - N$
  end if
end for
Return $m_0, \dots, m_{T-1}$

While Theorem 3.1 assumes $m_{T-1} < N$, Algorithm 2 can also determine the optimal sketch sizes when $m_{T-1} = N$. The following proposition verifies the correctness of Algorithm 2.

Proposition 3.2. *Assume that $r \ge 4d$. Then Algorithm 2 returns exactly the solution to the problem (10).*

The output of Algorithm 2 may not have a closed-form solution. Nevertheless, Algorithm 2 implies that the optimal sketch sizes take the form

$$m_t = \min(c\sqrt{g(t,T)}, N), \quad t = 0, \dots, T-1,$$

where $c > 0$ is a constant. We have

$$\frac{m_{t+1}}{m_t} = \sqrt{\frac{r}{4d}} \sqrt{\frac{T-t}{T-t-\frac{3}{2}}}, \quad t = 0, \dots, T^\dagger - 2.$$

It can be seen that the ratio $\frac{m_{t+1}}{m_t}$ has small variation for $t = 0, \dots, T^\dagger - 2$. In fact, we have

$$\sqrt{\frac{r}{4d}} \le \frac{m_{t+1}}{m_t} \le 2\sqrt{\frac{r}{4d}}, \quad t = 0, \dots, T^\dagger - 2.$$

While the optimal sketch sizes are obtained only for Gaussian sketching matrices, our result gives a useful insight for general sketching matrices, that is, we can choose the sketch sizes such that $\frac{m_1}{m_0} \approx \frac{m_2}{m_1} \approx \cdots \approx \frac{m_{T^\dagger-1}}{m_{T^\dagger-2}}$.

Now we consider the performance of the IDS algorithm with the optimal sketch sizes. Note that there remain two constants, that is c and T, to be determine. For the convenience of analysis, we would like to choose a c such that $\frac{m_0}{r} \approx \frac{m_1}{m_0}$. Hence we take $m_0 = \frac{1}{2} r^{\frac{3}{2}} d^{-\frac{1}{2}}$. With this choice of m_0, the optimal sketch sizes are

$$m_t = \min\left(\sqrt{\frac{r^{t+3} C(T-t-1)}{4 d^{t+1} C(T-1)}}, N \right). \tag{11}$$

To be simple, we would like to choose a T such that $T = T^\dagger + 1$. That is, in addition to the IDS iterations, we would like to perform one additional IHS iteration. In this way, it is guaranteed that each observation is accessed for at least once. To achieve this goal, we define

$$T = 1 + \max\left\{ \tilde{T} : \sqrt{\frac{r^{\tilde{T}+2}}{4 d^{\tilde{T}} C(\tilde{T}-1)}} < N \right\}. \tag{12}$$

If $r > 4d$, then $\frac{r^{\tilde{T}+2}}{4d^{\tilde{T}} C(\tilde{T}-1)}$ is an increasing function of $\tilde{T}$ and hence T is well defined. We have the following theorem.

Theorem 3.3. *Suppose $\mathbf{x}_T$ is obtained from Algorithm 1 where T is defined by (12), the sketching matrices are defined in (7) and the sketch sizes for IDS iterations are defined in (11). Suppose $\frac{\log(N/r)}{\log(r/(4d))}$ is bounded. Suppose as N tends to infinity, $d \to \infty$, $\frac{d}{r} \to 0$ and $|\mu - 1| = O(\frac{d}{r})$. Then as $N \to \infty$, $\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\| = o_P\left(\sqrt{\frac{d}{N}} \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\| \right)$.*

Now we consider the computing time of Algorithm 1 under the setting of Theorem 3.3. The computation of $\tilde{\mathbf{H}}^{-1}$ and $\mathbf{x}_0$ is the same as that in the IHS algorithm, and costs $O(rd^2)$ time. Note that $\sum_{t=0}^{T^\dagger-1} m_t \le (1+o(1))N$. Then by our definition of m_t, $\sum_{t=0}^{T-1} m_t \le (1+o(1))2N$. Thus, the total T iterations of the IDS algorithm cost at most $\{(1+o(1))8Nd\}$ FLOPs. In comparison, within $\{(1+o(1))8Nd\}$ FLOPs, one can only perform two IHS iterations.

To appreciate the error rate in Theorem 3.3, we consider the case that the data $(\mathbf{A}, \mathbf{y})$ is generated from a simple Gaussian linear model. Precisely, suppose that the elements of $\mathbf{A}$ are independent standard normal random variables, and that $\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\xi}$ where $\boldsymbol{\beta}$ is an unknown d-dimensional parameter and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{I}_N)$. In this case, $\mathbf{x}^*$ is the least square estimator of the parameter $\boldsymbol{\beta}$, and the estimation error is $\|\mathbf{A}(\mathbf{x}^* - \boldsymbol{\beta})\|^2 \sim \chi^2(d) = O_P(d)$. On the other hand, $\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2 \sim \chi^2(N-d) = O_P(N)$. Hence Theorem 3.3 implies that $\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\| = o_P(\sqrt{d})$. In comparison, from the theory of IHS algorithm, the output of 2 IHS iterations has the error rate $\|\mathbf{A}(\mathbf{x}_2 - \mathbf{x}^*)\| = O_P(N^{\frac{1}{2}}(\frac{d}{r})^{\frac{3}{2}})$. Thus, if $\frac{r^3}{d^2} = o(N)$, then the IDS algorithm can achieve a smaller order of error than the IHS algorithm with comparable computing time.

In the above analysis, we only consider the computing time of the iteration procedure when the sketched data $(\tilde{\mathbf{S}}\mathbf{A}, \tilde{\mathbf{S}}\mathbf{y})$ and $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$, $t = 0, \ldots, T-1$ are given. Although these sketched data can be obtained sequentially, their computation is still very intensive. Nevertheless, the theoretical analysis of the IDS algorithm with Gaussian sketching matrices can give insights on the general behavior of the IDS algorithm, and will guide us to propose an efficient IDS algorithm.

4. IDS with Iteration Efficient Sketching

We have investigated the properties of the IDS algorithm and derived the optimal choice of sketch sizes with Gaussian sketching matrices. These results are interesting in theory. In practice, however, it is not efficient to compute the gradient sketching and Hessian sketching with Gaussian sketching matrices. In this section, we propose a new class of sketching matrices, named *iteration efficient sketching*, which are efficient to apply in the IDS framework. Based on the proposed sketching matrices, we design an efficient IDS algorithm. For general sketchng matrices, it is hard to derive the exact error of the IDS algorithm as we did in the Gaussian sketching setting. Nevertheless, our theoretical results for Gaussian sketching imply that it may be a good choice to set sketch sizes such that $\frac{m_1}{m_0} = \cdots = \frac{m_{T^\dagger-1}}{m_{T^\dagger-2}}$. For the sake of simplicity, throughout this section, we assume N is a power of 2 and $m_{t+1} = 2m_t$ for $t = 0, \ldots, T^\dagger - 1$. Then we have $T^\dagger = \log_2(\frac{N}{m_0})$. Here the constant 2 is not essential, and can be replaced by other positive integers.

We will sequentially compute the sketched data in reverse order $(\mathbf{S}_{T^\dagger-1}\mathbf{A}, \mathbf{S}_{T^\dagger-1}\mathbf{y}), \ldots, (\mathbf{S}_0\mathbf{A}, \mathbf{S}_0\mathbf{y})$. We divide the $T^\dagger$ iterations into two stages. The first stage consists of the iterations for $t = T^\dagger - 1, \ldots, T^\diamond$, and the second stage consists of the iterations for $t = T^\diamond - 1, \ldots, 0$, where $T^\diamond \in \{1, \ldots, T^\dagger\}$ will be specified later. The computation in the two stages are exactly the same. However, we perform a preconditioning operation between the two stages. This preconditioning operation allows for a theoretical guarantee on the performance of the proposed algorithm.

In the first stage, the proposed sketching matrix is motivated by CountSketch, a popular sketching matrix that is fast to apply. The CountSketch matrix $\mathbf{S} \in \mathbb{R}^{m \times N}$ is defined as follows. The columns of $\mathbf{S}$ are independent. Each column of $\mathbf{S}$ contains exactly one non-zero element for which the position is uniformly distributed on $\{1, \ldots, m\}$. The non-zero elements of $\mathbf{S}$ are independent Rademacher random variables. For CountSketch $\mathbf{S}$, the computation of $\mathbf{S}\mathbf{A}$ can be completed within $O(Nd)$ time. Unfortunately, for CountSketch matrices, it may not be easy to compute the sketched data $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$ sequentially in the IDS framework. It can

be seen that the $m \times N$ CountSketch matrix has the same distribution as $\mathbf{G}_{m,N}\mathbf{D}_N\mathbf{P}_N$, where the matrices $\mathbf{G}_{m,N}$, $\mathbf{D}_N$ and $\mathbf{P}_N$ are independent, $\mathbf{G}_{m,N}$ is defined as

$$\mathbf{G}_{m,N} := \begin{pmatrix} \mathbf{1}_{k_1}^\top & \mathbf{0}_{k_2}^\top & \cdots & \mathbf{0}_{k_m}^\top \\ \mathbf{0}_{k_1}^\top & \mathbf{1}_{k_2}^\top & \cdots & \mathbf{0}_{k_m}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{k_1}^\top & \mathbf{0}_{k_2}^\top & \cdots & \mathbf{1}_{k_m}^\top \end{pmatrix},$$

the vector $(k_1, \ldots, k_m)$ has multinomial distribution $\mathrm{Mult}(N; \frac{1}{m}, \ldots, \frac{1}{m})$, $\mathbf{D}_N \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables and $\mathbf{P}_N \in \mathbb{R}^{N\times N}$ is a uniformly distributed permutation matrix. The matrix $\mathbf{G}_{m,N}$ relies on $k_1, \ldots, k_m$ which are random. We propose to replace $(k_1, \ldots, k_m)^\top$ by its expectation $(\frac{N}{m}, \ldots, \frac{N}{m})^\top$. To be precise, let $\mathbf{G}^*_{m,N} := \mathbf{I}_m \otimes \mathbf{1}^\top_{\frac{N}{m}}$, where $\otimes$ denotes the Kronecker product of matrices. We propose the following sketching matrix:

$$\hat{\mathbf{S}}_{m,N} := \mathbf{G}^*_{m,N}\mathbf{D}_N\mathbf{P}_N. \tag{13}$$

The following theorem implies that the sketching matrix $\hat{\mathbf{S}}_{m,N}$ has a similar embedding property as CountSketch.

Theorem 4.1. *Suppose $\mathbf{U} \in \mathbb{R}^{N\times d}$ is a non-random matrix such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$, and $\epsilon, \delta \in (0,1)$. If $m \geq \frac{d(d+1)}{\delta\epsilon^2}$, then*

$$\Pr\left\{\|\mathbf{U}^\top\hat{\mathbf{S}}^\top_{m,N}\hat{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\} \leq \delta.$$

Here we remark that $\hat{\mathbf{S}}_{m,N}$ has an good property that is not shared by CountSketch. That is, if $m = N$, then $\hat{\mathbf{S}}_{N,N}$ is an orthogonal matrix. Hence the application $\hat{\mathbf{S}}_{N,N}$ does not loss any information. In comparison, the $N \times N$ CountSketch matrix may not be invertible. This phenomenon implies that $\hat{\mathbf{S}}_{m,N}$ may be more favorable for large m.

For m_1, m_2 such that $\frac{m_2}{m_1}$ is an integer, we have $\mathbf{G}^*_{m_1,N} = \big(\mathbf{I}_{m_1} \otimes \mathbf{1}^\top_{\frac{m_2}{m_1}}\big)\mathbf{G}^*_{m_2,N}$. With this property, we can compute the sketched data sequentially in the first stage. We define

$$\mathbf{S}_t := \mathbf{G}^*_{m_t,N}\mathbf{D}_N\mathbf{P}_N, \quad t = T^\diamond, \ldots, T^\dagger - 1,$$

where the random matrices $\mathbf{D}_N$ and $\mathbf{P}_N$ are defined as above. By construction, all sketching matrices $\mathbf{S}_t$ share the same random matrices $\mathbf{D}_N$ and $\mathbf{P}_N$. Note that $\mathbf{S}_t\mathbf{A} = \big(\mathbf{I}_{m_t} \otimes \mathbf{1}^\top_{\frac{m_{t+1}}{m_t}}\big)\mathbf{S}_{t+1}\mathbf{A}$. This formula allows us to efficiently compute $(\mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$ based on $(\mathbf{S}_{t+1}\mathbf{A}, \mathbf{S}_{t+1}\mathbf{y})$.

The sketching matrix $\hat{\mathbf{S}}_{m,N}$ is fast to apply. However, Theorem 4.1 implies that one must take $m = \Omega(d^2)$ to guarantee a valid subspace embedding. This phenomenon is also shared by CountSketch. In fact, for CountSketch, it is known that the $\Omega(d^2)$ sketch size is necessary for subspace embedding; see, e.g., Nelson & Nguyên (2014). Hence $\hat{\mathbf{S}}_{m,N}$ can not be used to reduce the sample size smaller than the order d^2. Nevertheless, the theoretical results of Bourgain et al. (2015) imply that if the data is preconditioned by the randomized Hadamard transform, then we can apply CountSketch with a much smaller m. Motivated by this result, we propose the following sketching matrix:

$$\check{\mathbf{S}}_{m,N} := \mathbf{G}^*_{m,N}\mathbf{D}_N\mathbf{P}_N\mathbf{W}_N\tilde{\mathbf{D}}_N,$$

where $\mathbf{D}_N$ and $\mathbf{P}_N$ are defined as in (13), $\tilde{\mathbf{D}}_N$ is an independent copy of $\mathbf{D}_N$, and $\mathbf{W}_N \in \mathbb{R}^{N\times N}$ is the Walsh-Hadamard transform defined recursively as

$$\mathbf{W}_1 := 1, \quad \mathbf{W}_N := \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \mathbf{W}_{\frac{N}{2}}.$$

The computation of $\check{\mathbf{S}}_{m,N}\mathbf{A}$ costs $O(Nd\log(N))$ time where the bottleneck is the application of the Walsh-Hadamard transform. Nevertheless, $\check{\mathbf{S}}_{m,N}\mathbf{A}$ has a better embedding property than $\hat{\mathbf{S}}_{m,N}\mathbf{A}$, as indicated by the following theorem.

Theorem 4.2. *Suppose $\mathbf{U} \in \mathbb{R}^{N\times d}$ is a non-random matrix such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$, and $\epsilon, \delta \in (0,1)$. Then for any $m \geq \gamma\epsilon^{-2}d\{\log(\frac{e^2d}{\epsilon\delta})\}^3$, where $\gamma > 0$ is an absolute constant, we have*

$$\Pr\left\{\|\mathbf{U}^\top\check{\mathbf{S}}^\top_{m,N}\check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\} \leq \delta.$$

Theorem 4.2 implies that $\check{\mathbf{S}}_{m,N}$ is a valid subspace embedding for $m = \Omega(d(\log(d))^3)$. Since $\check{\mathbf{S}}_{m,N}$ is slower to apply than $\hat{\mathbf{S}}_{m,N}$, we do not directly use $\check{\mathbf{S}}_{m,N}$. Instead, we apply it after the sample size is already reduced to $m_{T^\diamond}$ in the first stage. In the first stage, we apply $\hat{\mathbf{S}}_{T^\dagger-1}, \ldots, T^\diamond$ sequentially and obtain the reduced data $\mathbf{S}_{T^\diamond}\mathbf{A}$. Then the rows of $\mathbf{S}_{T^\diamond}\mathbf{A}$ are random vectors whose distributions are invariant under sign flipping. Consequently, in the second stage, we do not need to apply the matrix $\tilde{\mathbf{D}}_{m_{T^\diamond}}$. We define

$$\mathbf{S}_t := \mathbf{G}^*_{m_t,m_{T^\diamond}}\mathbf{D}_{m_{T^\diamond}}\mathbf{P}_{m_{T^\diamond}}\mathbf{W}_{m_{T^\diamond}}\mathbf{S}_{T^\diamond}, \ t = 0, \ldots, T^\diamond - 1,$$

where the random matrices $\mathbf{D}_{m_{T^\diamond}}$ and $\mathbf{P}_{m_{T^\diamond}}$ are independent of $\mathbf{S}_{T^\diamond}$. Then the distribution of $\mathbf{S}_t$ is the same as that of $\check{\mathbf{S}}_{m_t,m_{T^\diamond}}\mathbf{S}_{T^\diamond}$ where $\check{\mathbf{S}}_{m_t,m_{T^\diamond}}$ and $\mathbf{S}_{T^\diamond}$ are independent. In practice, the proposed two stage sketching is very easy to implement. In fact, after we obtain $\mathbf{S}_{T^\diamond}\mathbf{A}$ in the first stage, we apply $\mathbf{D}_{m_{T^\diamond}}\mathbf{P}_{m_{T^\diamond}}\mathbf{W}_{m_{T^\diamond}}$ to this matrix and then we can compute $\mathbf{S}_t$, $t = T^\diamond - 1, \ldots, 0$, exactly as in the first stage.

From Theorem 4.2, after two stages of sketching, we can eventually reach $m_0 = O(d(\log(d))^3)$. It is known that SRHT can reduce the sample size to $O(d\log(d))$; see Cohen (2016). Hence we use SRHT to further reduce the sample size to $O(d(\log(d)))$ to obtain the Hessian sketching. Specifically, we define $\tilde{\mathbf{S}} := \mathbf{S}^\dagger\mathbf{S}_0$, where $\mathbf{S}^\dagger \in \mathbb{R}^{r\times m_0}$ is an SRHT matrix which is independent of $\mathbf{S}_0$.

We summarize the proposed IDS algorithm with iteration efficient sketching matrices in Algorithm 3. The following

Algorithm 3 IDS algorithm with iteration efficient sketching matrices

Input: $\mathbf{A} \in \mathbb{R}^{N\times d}$, $\mathbf{y} \in \mathbb{R}^N$, r, m_0, $T^\diamond$, T, μ
$T^\dagger \leftarrow \log_2(\frac{N}{m_0})$
$\mathbf{A} \leftarrow \mathbf{D}_N\mathbf{P}_N\mathbf{A}$; $\mathbf{y} \leftarrow \mathbf{D}_N\mathbf{P}_N\mathbf{y}$; $\mathbf{S}_{T^\dagger} \leftarrow \mathbf{I}_N$;
for $t \leftarrow T^\dagger - 1$ **to** 0 **do**
  $\mathbf{S}_t\mathbf{A} \leftarrow (\mathbf{I}_{2^t m_0} \otimes \mathbf{1}_2^\top)\mathbf{S}_{t+1}\mathbf{A}$; $\mathbf{S}_t\mathbf{y} \leftarrow (\mathbf{I}_{2^t m_0} \otimes \mathbf{1}_2^\top)\mathbf{S}_{t+1}\mathbf{y}$
  if $t = T^\diamond$ **then**
    $\mathbf{S}_{T^\diamond}\mathbf{A} \leftarrow \mathbf{D}_{m_{T^\diamond}}\mathbf{P}_{m_{T^\diamond}}\mathbf{W}_{m_{T^\diamond}}\mathbf{S}_{T^\diamond}\mathbf{A}$
    $\mathbf{S}_{T^\diamond}\mathbf{y} \leftarrow \mathbf{D}_{m_{T^\diamond}}\mathbf{P}_{m_{T^\diamond}}\mathbf{W}_{m_{T^\diamond}}\mathbf{S}_{T^\diamond}\mathbf{y}$
  end if
end for
$\tilde{\mathbf{H}}^{-1} \leftarrow (\mathbf{A}^\top\mathbf{S}_0^\top\mathbf{S}^{\dagger\top}\mathbf{S}^\dagger\mathbf{S}_0\mathbf{A})^{-1}$
$\mathbf{x}_0 \leftarrow \tilde{\mathbf{H}}^{-1}\mathbf{A}^\top\mathbf{S}_0^\top\mathbf{S}^{\dagger\top}\mathbf{S}^\dagger\mathbf{S}_0\mathbf{y}$
for $t \leftarrow 0$ **to** $T^\dagger - 1$ **do**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \mu\tilde{\mathbf{H}}^{-1}\nabla f(\mathbf{x}_t; \mathbf{S}_t\mathbf{A}, \mathbf{S}_t\mathbf{y})$
end for
for $t \leftarrow T^\dagger$ **to** $T - 1$ **do**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \mu\tilde{\mathbf{H}}^{-1}\nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y})$
end for
Return $\mathbf{x}_T$

theorem gives a non-asymptotic bound on the convergence rate of Algorithm 3.

Theorem 4.3. *In Algorithm 3, suppose N, r, m_0 are powers of 2, $T^\diamond < T$, $|\mu - 1| \le \frac{1}{4}$, $\delta \in (0,1)$, and $\epsilon \in (0, \frac{1}{10})$. Suppose*

$$m_{T^\diamond} > \tilde{\gamma}\frac{d^2}{\delta\epsilon^2}, \quad m_0 > \tilde{\gamma}\epsilon^{-2}d\left\{\log\left(\frac{e^2 d}{\epsilon\delta}\right)\right\}^3,$$
$$r > \tilde{\gamma}\epsilon^{-2}\left(d + \log\left(\frac{m_0}{\delta}\right)\right)\log\left(\frac{ed}{\delta}\right),$$

where $\tilde{\gamma} > 0$ is an absolute constant. Let $\mathbf{x}_T$ be the output of Algorithm 3. Then with probability at least $1 - 3\delta$, for any $T \ge T^\dagger = \log_2(\frac{N}{m_0})$,

$$\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\| \le \frac{1}{2^T}\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\| + \frac{4\sqrt{5}(\sqrt{2}+1)}{2^{T-T^\dagger}\delta}\sqrt{\frac{d}{N}}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|. \quad (14)$$

The first term in the error bound (14) takes the same form as the error bound of the IHS algorithm (5). The additional term in the bound (14) is the price of using an approximation of the gradient. According to Theorem 4.3, we take $m_{T^\diamond} = O(d^2)$, $m_0 = O(d(\log(d))^3)$, $r = O(d\log(d))$.

Now we consider the computing time of Algorithm 3. Note that $\sum_{t=0}^{T^\dagger-1} m_t = O(N)$. Hence with iteration efficient sketching matrices, the computation of the sketched data $(\mathbf{S}_0\mathbf{A}, \mathbf{S}_0\mathbf{y}), \ldots, (\mathbf{S}_{T^\dagger-1}\mathbf{A}, \mathbf{S}_{T^\dagger-1}\mathbf{y})$ and $(\tilde{\mathbf{S}}\mathbf{A}, \tilde{\mathbf{S}}\mathbf{y})$ can be completed within $O((\sum_{t=0}^{T^\dagger-1} m_t + N)d) = O(Nd)$ time. Since we take $r = O(d\log(d))$, the computation of $\tilde{\mathbf{H}}^{-1}$ and $\mathbf{x}_0$ can be completed within $O(Nd + d^3\log(d))$ time. Finally, the $T^\dagger$ steps of IDS iterations and $T - T^\dagger$ steps of IHS iterations cost $O((\sum_{i=0}^{T^\dagger-1} m_t)d + (T - T^\dagger)Nd) = O((T + 1 - T^\dagger)Nd)$ time. In summary, Algorithm 3 costs $O((T + 1 - T^\dagger)Nd + d^3\log(d))$ time in total.

Theorem 4.3 allows us to analyze the trade-off between the computing time and the precision of Algorithm 3. It can be shown that $\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\| = O_P(\sqrt{\frac{d}{r}}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|)$. We make a further assumption that $\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|$ and $\sqrt{\frac{d}{r}}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|$ have the same order of magnitude. This assumption is valid under general conditions. In this case, the error bound (14) is bounded by $O(\frac{1}{2^{T-T^\dagger/2}}\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|)$. Suppose we would like to achieve an ϵ relative error, i.e., $\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\| \le \epsilon\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\|$, where $\epsilon \in (0,1)$. Then the IDS algorithm needs to take $T = \max(T^\dagger, \frac{T^\dagger}{2} + \log_2(\frac{1}{\epsilon}) + O(1))$ iterations. But $T^\dagger = \log_2(\frac{N}{m_0}) = \log_2(\frac{N}{d(\log(d))^3}) + O(1)$. Hence the IDS algorithm costs $O\left(\max\left(1, \log_2(\frac{1}{\epsilon}) - \frac{1}{2}\log_2(\frac{N}{d(\log(d))^3})\right)Nd + d^3\log(d)\right)$ time. This computing time is significantly faster than the IHS algorithm. In fact, if $\log_2(\frac{1}{\epsilon}) - \frac{1}{2}\log_2(\frac{N}{d(\log(d))^3})$ is bounded, or equivalently, $\epsilon = \Omega(\sqrt{\frac{d(\log(d))^3}{N}})$, and $d^2\log(d) = O(N)$, then the IDS algorithm can reach the ϵ relative error within $O(Nd)$ time. In this regime, the proposed IDS algorithm improves the current state-of-the-art performance for least-squares problem.

5. Numerical Experiments

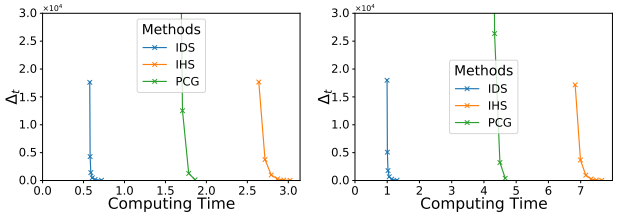
We carry out simulations of the IDS algorithm and compare it with the IHS algorithm and the PCG algorithm in (Lacotte & Pilanci, 2021). All algorithms are implemented by `C++`. No external library is used except for `C++` Standard Template Library. The matrix inverse is implemented by Gaussian elimination. The program is compiled using `gcc` version 7.5.0 with `-O2` optimization, and runs on a CPU with 3.30 GHz.

We use $\Delta_t := \|\mathbf{A}(\mathbf{x}_t - \mathbf{x}^*)\|^2$ to measure the precision of $\mathbf{x}_t$, $t = 1, \ldots, T$. In our experiments, $T = 6$ iterations are performed for all algorithms. The IDS algorithm is implemented according to Algorithm 3 with $T^\dagger = 5$, $T^\diamond = 1$, $m_0 = N/2^5$, $r = 8d$. Following the result of Özaslan et al. (2019), we adopt the step size $\mu = \frac{(1-d/r)^2}{1+d/r}$. For the IHS algorithm, we use the update formula $\mathbf{x}_{t+1} = \mathbf{x}_t - \mu(\mathbf{A}^\top\mathbf{S}_0^\top\mathbf{S}_0\mathbf{A})^{-1}\nabla f(\mathbf{x}_t; \mathbf{A}, \mathbf{y})$ where $\mathbf{S}_0$ is an $(8d)\times N$ SRHT sketching matrix and $\mu = \frac{(1-d/r)^2}{1+d/r}$. The computing time is measured in seconds. For the PCG algorithm, SRHT sketching matrix is used. The reported results are the averages of 10 independent replications.

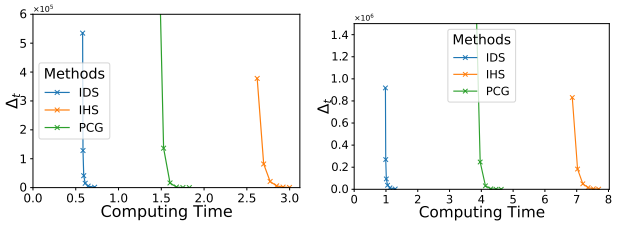
The data generation mechanism is as follows. For Model I, the elements of $\mathbf{A}$ are i.i.d. generated from the standard normal distribution, and $\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\xi}$, where $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\xi} \in \mathbb{R}^N$ are random vectors whose elements are i.i.d.

standard normal random variables. For Model II, we first generate data from Model I. Then each element of $\mathbf{A}$ and $\mathbf{y}$ is replaced by zero with probability 0.5. The ground truth $\mathbf{x}^*$ is computed via the formula $\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y}$.

Figure 1 illustrates the relationship between Δ_t and the computing time (measured by seconds) in different settings. The numerical results show that for a given computing time, the IDS algorithm can achieve a much smaller error than the IHS algorithm, which verifies our theoretical results.



(a) Model I. $N = 2^{20}$, $d = 2^6$. (b) Model I. $N = 2^{20}$, $d = 2^7$.



(c) Model II. $N = 2^{20}$, $d = 2^6$. (d) Model II. $N = 2^{20}$, $d = 2^7$.

Figure 1. Δ_t versus the computing time for IDS, IHS and PCG.

6. Discussion

In this work, we proposed the IDS algorithm for the least-squares problem which uses approximations for both the gradient and the Hessian. We investigated the theoretical properties of the IDS algorithm. The proposed IDS algorithm improves the state-of-the-art computing time in a wide range. Nevertheless, there are several problems worth further research.

In this work, we did not consider the fine-grained choice of the step size in Algorithm 3. Also, it is unknown if a momentum term is useful for the IDS algorithm. To solve these problems, one may need to analyze the fine-grained behavior of the IDS algorithm.

The present work focuses on unconstrained least-squares problem. It is interesting to apply the idea of IDS to constrained or regularized least-squares problem.

In this work, we proposed the iteration efficient sketching and investigated its subspace embedding property. This sketching method may be useful in other problems. It is also interesting to investigate the fine-grained subspace embedding properties of the iteration efficient sketching.

Acknowledgements

The authors thank anonymous reviewers for their valuable comments and suggestions. This work was supported by National Natural Science Foundation of China (No 11971478) and Beijing Natural Science Foundation (No Z200001).

References

Ailon, N. and Chazelle, B. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

Ailon, N. and Liberty, E. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

Allen-Zhu, Z., Gelashvili, R., Micali, S., and Shavit, N. Sparse sign-consistent johnson–lindenstrauss matrices: Compression with neuroscience-based constraints. *Proceedings of the National Academy of Sciences*, 111(47): 16872–16876, 2014.

Avron, H., Maymounkov, P., and Toledo, S. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*. Springer, 2010.

Bai, Z. D. and Yin, Y. Q. Convergence to the semicircle law. *The Annals of Probability*, 16(2):863–875, 1988.

Bourgain, J., Dirksen, S., and Nelson, J. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

Boutsidis, C. and Gittens, A. Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34 (3):1301–1340, 2013.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *STOC*, pp. 81–90, 2013.

Clarkson, K. L. and Woodruff, D. P. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63(6):1–45, feb 2017.

Cohen, M. B. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 278–287, 2016.

Cohen, M. B., Nelson, J., and Woodruff, D. P. Optimal Approximate Matrix Product in Terms of Stable Rank. In *ICALP*, pp. 11:1–11:14, 2016.

Cohen, M. B., Jayram, T., and Nelson, J. Simple Analyses of the Sparse Johnson-Lindenstrauss Transform. In *SOSA*, pp. 15:1–15:9, 2018.

Dasgupta, A., Kumar, R., and Sarlós, T. A sparse Johnson-Lindenstrauss transform. In *STOC*, pp. 341–350, 2010.

Drineas, P. and Mahoney, M. W. RandNLA. *Communications of the ACM*, 59(6):80–90, may 2016.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. The Johns Hopkins University Press, fourth edition, 2013.

Jagadeesan, M. Simple Analysis of Sparse, Sign-Consistent JL. In *APPROX/RANDOM*, pp. 61:1–61:20, 2019.

Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. volume 26 of *Contemporary Mathematics*, pp. 189–206. 1984.

Kane, D. M. and Nelson, J. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4:1–4:23, 2014.

Lacotte, J. and Pilanci, M. Optimal randomized first-order methods for least-squares problems. In *ICML*, pp. 5587–5597, 2020.

Lacotte, J. and Pilanci, M. Faster least squares optimization, 2021. arXiv:1911.02675.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. In *NeurIPS*, pp. 10462–10472, 2019.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014.

Mahoney, M. W. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3 (2):123–224, 2011.

Meng, X. and Mahoney, M. W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *STOC*, pp. 91–100, 2013.

Nelson, J. and Nguyen, H. L. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *FOCS*, pp. 117–126, 2013.

Nelson, J. and Nguyên, H. L. Lower bounds for oblivious subspace embeddings. In *ICALP*, volume 8572, pp. 883–894, 2014.

Özaslan, I. K., Pilanci, M., and Arikan, O. Iterative hessian sketch with momentum. In *ICASSP*, pp. 7470–7474, 2019.

Pilanci, M. and Wainwright, M. J. Iterative Hessian sketch: fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17: Paper No. 53, 38, 2016.

Rokhlin, V. and Tygert, M. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13212–13217, 2008.

Sarlós, T. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pp. 143–152, 2006.

Thorup, M. and Zhang, Y. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, pp. 615–624, 2004.

Tropp, J. A. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis. Theory and Applications*, 3(1-2):115–126, 2011.

Wainwright, M. J. *High-Dimensional Statistics*. Cambridge University Press, feb 2019.

Wang, D. and Xu, J. Large scale constrained linear regression revisited: Faster algorithms via preconditioning. In *AAAI*, pp. 1439–1446, 2018.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1-2):iv+157, 2014.

A. Proofs of Theoretical Results

A.1. Notations

We introduce some notations that will be used throughout our theoretical proofs. For a matrix $\mathbf{B}$, let $\|\mathbf{B}\|$ denote its operator norm, and $\|\mathbf{B}\|_F$ denote its Frobeneous norm. Let $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top$ denote the compact singular value decomposition of $\mathbf{A}$, where $\mathbf{U}_\mathbf{A}$ is an $N \times d$ column orthogonal matrix, $\mathbf{V}_\mathbf{A}$ is a $d \times d$ orthogonal matrix and $\mathbf{D}_\mathbf{A}$ is a $d \times d$ diagonal matrix. In the proofs, we denote $\kappa := \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2$.

A.2. Proof of Theorem 3.1

In this section, we prove Theorem 3.1. We begin with a useful lemma.

Lemma A.1. *Suppose $\mathbf{W} \in \mathbb{R}^{N\times d}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. If $N \geq d$, then for any $x > 0$,*

$$\Pr\left(\left\|\frac{1}{N}\mathbf{W}^\top\mathbf{W} - \mathbf{I}_d\right\| > 2(1+x)\sqrt{\frac{d}{N}} + (1+x)^2\frac{d}{N}\right) \leq 2\exp(-dx^2/2).$$

See, e.g., Wainwright (2019), Example 6.2 for a proof of this result.

Now we prove Theorem 3.1. To make the proof clear, we defer some technical details of the proof to Lemmas A.2, A.3 and A.4.

Proof of Theorem 3.1. From the update formula (8), we have

$$\mathbf{x}_{t+1} - \mathbf{x}^* = \left(\mathbf{I}_d - \mu\frac{r}{m_t}(\tilde{\mathbf{A}}_r^\top\tilde{\mathbf{A}}_r)^{-1}\tilde{\mathbf{A}}_{m_t}^\top\tilde{\mathbf{A}}_{m_t}\right)(\mathbf{x}_t - \mathbf{x}^*) + \mu\frac{r}{m_t}(\tilde{\mathbf{A}}_r^\top\tilde{\mathbf{A}}_r)^{-1}\tilde{\mathbf{A}}_{m_t}^\top(\tilde{\mathbf{y}}_{m_t} - \tilde{\mathbf{A}}_{m_t}\mathbf{x}^*). \tag{15}$$

Note that $\|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^*)\|^2 = \|\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_T - \mathbf{x}^*)\|^2$. Hence we only need to deal with $\|\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_T - \mathbf{x}^*)\|^2$. Define

$$\breve{\mathbf{A}}_k = (\breve{\mathbf{a}}_1, \ldots, \breve{\mathbf{a}}_k)^\top := \tilde{\mathbf{A}}_k\mathbf{V}_\mathbf{A}\mathbf{D}_\mathbf{A}^{-1}, \quad k = 1, \ldots, N.$$

Then we have $\breve{\mathbf{a}}_k \sim \mathcal{N}(0, \mathbf{I}_d)$. That is, $\breve{\mathbf{A}}_k$ is a $k \times d$ matrix with independent $\mathcal{N}(0,1)$ entries. Denote

$$\begin{aligned}\breve{\varepsilon}_i :=& (\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\breve{\mathbf{a}}_i\varepsilon_i, \quad i = 1, \ldots, N,\\ \mathbf{L}_t :=& \mathbf{I}_d - \mu\frac{r}{m_t}(\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\breve{\mathbf{A}}_{m_t}^\top\breve{\mathbf{A}}_{m_t}, \quad t = 0, \ldots, T-1.\end{aligned}$$

Then from (15), we have

$$\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_{t+1} - \mathbf{x}^*) = \mathbf{L}_t\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t - \mathbf{x}^*) + \mu\frac{r}{m_t}\sum_{i=1}^{m_t}\breve{\varepsilon}_i.$$

Note that $\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t - \mathbf{x}^*) = \sum_{k=1}^r \breve{\varepsilon}_k$. Then by induction, we obtain the following formula:

$$\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_T - \mathbf{x}^*) = (\prod_{j=1}^T \mathbf{L}_{T-j})\sum_{k=1}^r\breve{\varepsilon}_k + \sum_{i=0}^{T-1}\mu\frac{r}{m_i}(\prod_{j=1}^{T-i-1}\mathbf{L}_{T-j})\sum_{k=1}^{m_i}\breve{\varepsilon}_k, \tag{16}$$

where by convention, $\prod_{j=1}^0(\mathbf{I}_d - \mathbf{L}_{T-j}) = \mathbf{I}_d$.

Denote $\mathbf{K} := \frac{1}{r}\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r - \mathbf{I}_d$. To deal with the expression (16), our strategy is to approximate the term $\mathbf{L}_i$ by $\mathbf{K}$. Note that

$$\begin{aligned}&\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_T - \mathbf{x}^*) - \mathbf{K}^T\sum_{k=1}^r\breve{\varepsilon}_k - \sum_{i=0}^{T-1}\mu\frac{r}{m_i}\mathbf{K}^{T-i-1}\sum_{k=1}^{m_i}\breve{\varepsilon}_k\\ =&(\prod_{j=1}^T\mathbf{L}_{T-j} - \mathbf{K}^T)\sum_{k=1}^r\breve{\varepsilon}_k + \sum_{i=0}^{T-1}\mu\frac{r}{m_i}(\prod_{j=1}^{T-i-1}\mathbf{L}_{T-j} - \mathbf{K}^{T-i-1})\sum_{k=1}^{m_i}\breve{\varepsilon}_k.\end{aligned}$$

Then from Minkowski inequality,

$$\begin{aligned}&\left[\mathrm{E}\left\{\left\|\mathbf{D}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^{\top}(\mathbf{x}_T-\mathbf{x}^*)-\mathbf{K}^T\sum_{k=1}^{r}\breve{\varepsilon}_k-\sum_{i=0}^{T-1}\mu\frac{r}{m_i}\mathbf{K}^{T-i-1}\sum_{k=1}^{m_i}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}\right]^{\frac{1}{2}}\\ \leq&\left\|\prod_{j=1}^{T}\mathbf{L}_{T-j}-\mathbf{K}^T\right\|\left[\mathrm{E}\left\{\left\|\sum_{k=1}^{r}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}\right]^{\frac{1}{2}}+\sum_{i=0}^{T-1}\mu\frac{r}{m_i}\left\|\prod_{j=1}^{T-i-1}\mathbf{L}_{T-j}-\mathbf{K}^{T-i-1}\right\|\left[\mathrm{E}\left\{\left\|\sum_{k=1}^{m_i}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}\right]^{\frac{1}{2}}.\end{aligned}\tag{17}$$

From Lemma A.3, for $i=0,\ldots,T-1$,

$$\mathrm{E}\left\{\left\|\sum_{k=1}^{m_i}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}=O_P\left(\kappa\frac{dm_i}{r^2}\right).\tag{18}$$

And

$$\mathrm{E}\left\{\left\|\sum_{k=1}^{r}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}=O_P\left(\kappa\frac{d}{r}\right).\tag{19}$$

It follows from (17), (18), (19) and Lemma A.2 that

$$\begin{aligned}&\left[\mathrm{E}\left\{\left\|\mathbf{D}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^{\top}(\mathbf{x}_T-\mathbf{x}^*)-\mathbf{K}^T\sum_{k=1}^{r}\breve{\varepsilon}_k-\sum_{i=0}^{T-1}\mu\frac{r}{m_i}\mathbf{K}^{T-i-1}\sum_{k=1}^{m_i}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}\right]^{\frac{1}{2}}\\ =&o_P\left(\left(\kappa\frac{d}{r}\right)^{\frac{1}{2}}\left(\frac{d}{r}\right)^{\frac{T}{2}}+\sum_{i=0}^{T-1}\frac{r}{m_i}\left(\kappa\frac{dm_i}{r^2}\right)^{\frac{1}{2}}\left(\frac{d}{r}\right)^{\frac{T-i-1}{2}}\right)\\ =&o_P\left(\kappa^{\frac{1}{2}}\left(\frac{d}{r}\right)^{\frac{T+1}{2}}+\kappa^{\frac{1}{2}}\sum_{i=0}^{T-1}\left(\frac{d}{m_i}\right)^{\frac{1}{2}}\left(\frac{d}{r}\right)^{\frac{T-i-1}{2}}\right)\\ =&o_P\left[\kappa^{\frac{1}{2}}\left\{\left(\frac{d}{r}\right)^{T+1}C\left(T\right)+\sum_{\ell=0}^{T-1}\frac{g(i,T)}{m_i}\right\}^{\frac{1}{2}}\right].\end{aligned}\tag{20}$$

Having obtained the above approximation bound, now we deal with $\mathbf{K}^T\sum_{k=1}^{r}\breve{\varepsilon}_k+\sum_{i=0}^{T-1}\mu\frac{r}{m_i}\mathbf{K}^{T-i-1}\sum_{k=1}^{m_i}\breve{\varepsilon}_k$. From Lemma A.4, we have

$$\mathrm{E}\left\{\left\|\mathbf{K}^T\sum_{k=1}^{r}\breve{\varepsilon}_k+\sum_{i=0}^{T-1}\mu\frac{r}{m_i}\mathbf{K}^{T-i-1}\sum_{k=1}^{m_i}\breve{\varepsilon}_k\right\|^2\mid\tilde{\mathbf{A}}\right\}=(1+o_P(1))\kappa\left\{\frac{1}{r}\operatorname{tr}(\mathbf{K}^{2T})+\sum_{i=0}^{T-1}\frac{1}{m_i}\operatorname{tr}(\mathbf{K}^{2(T-i-1)})\right\}.\tag{21}$$

Since $\frac{r}{d}\to\infty$, the empirical spectral distribution of the matrix $\sqrt{\frac{r}{d}}\mathbf{K}$ converges almost surely to the semicircle law with density function $(2\pi)^{-1}\sqrt{4-x^2}\mathbf{1}_{[-2,2]}(x)$; see Bai & Yin (1988). Lemma A.1 implies that the extreme eigenvalues of $\sqrt{\frac{r}{d}}\mathbf{K}$ are bounded in probability. Hence for $i=0,\ldots,T$,

$$\operatorname{tr}\left(\mathbf{K}^{2i}\right)=d\left(\frac{d}{r}\right)^{i}\frac{1}{d}\operatorname{tr}\left\{\left(\sqrt{\frac{r}{d}}\mathbf{K}\right)^{2i}\right\}=d\left(\frac{d}{r}\right)^{i}\left(\int_{-2}^{2}x^{2i}\frac{1}{2\pi}\sqrt{4-x^2}\,\mathrm{d}x+o_P(1)\right).$$

From Lemma 2.1 of Bai & Silverstein (2010), we have

$$\int_{-2}^{2}x^{2i}\frac{1}{2\pi}\sqrt{4-x^2}\,\mathrm{d}x=C(i),\quad i=0,\ldots,T.$$

Thus,

$$\frac{1}{r}\operatorname{tr}(\mathbf{K}^{2T})+\sum_{i=0}^{T-1}\mu^2\frac{1}{m_i}\operatorname{tr}(\mathbf{K}^{2(T-i-1)})=(1+o_P(1))\left\{\left(\frac{d}{r}\right)^{T+1}C\left(T\right)+\sum_{\ell=0}^{T-1}\frac{d}{m_i}\left(\frac{d}{r}\right)^{T-i-1}C\Big(T-i-1\Big)\right\}.\tag{22}$$

The conclusion follows from (20), (21) and (22).

□

Lemma A.2. *Suppose the conditions of Theorem 3.1 hold. Then for $i = 0, \ldots, T-1$,*

$$\Big\| \prod_{j=1}^{T-i-1} \mathbf{L}_{T-j} - \mathbf{K}^{T-i-1} \Big\| = o_P\left(\left(\frac{d}{r}\right)^{\frac{T-i-1}{2}} \right).$$

and

$$\Big\| \prod_{j=1}^{T} \mathbf{L}_{T-j} - \mathbf{K}^{T} \Big\| = o_P\left(\left(\frac{d}{r}\right)^{\frac{T}{2}} \right).$$

Proof. It is straightforward to see that for $t = 0, \ldots, T-1$,

$$\mathbf{L}_t = \mathbf{K} - r(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}\mathbf{K}^2 + r(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}(\mathbf{I}_d - \frac{\mu}{m_t}\breve{\mathbf{A}}_{m_t}^\top \breve{\mathbf{A}}_{m_t}).$$

We have assumed that $d/r \to \infty$ and T is fixed. Hence from Lemma A.1,

$$\|\mathbf{K}\| = O_P\left(\sqrt{\frac{d}{r}}\right), \quad \left\| \frac{1}{m_t}\breve{\mathbf{A}}_{m_t}^\top \breve{\mathbf{A}}_{m_t} - \mathbf{I}_d \right\| = O_P\left(\sqrt{\frac{d}{m_t}}\right), \quad t = 0, \ldots, T-1.$$

As a consequence, $\|r(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}\| = 1 + o_P(1)$. Also, combine the above bound and the assumptions $|\mu - 1| = O(\frac{d}{r})$, $r/m_0 \to 0$, we have, for $t = 0, \ldots, T-1$, that

$$\left\| \frac{\mu}{m_t}\breve{\mathbf{A}}_{m_t}^\top \breve{\mathbf{A}}_{m_t} - \mathbf{I}_d \right\| \le \mu \left\| \frac{1}{m_t}\breve{\mathbf{A}}_{m_t}^\top \breve{\mathbf{A}}_{m_t} - \mathbf{I}_d \right\| + O\left(\frac{d}{r}\right) = O_P\left(\sqrt{\frac{d}{m_t}} + \frac{d}{r}\right) = o_P\left(\sqrt{\frac{d}{r}}\right).$$

Thus, for $t = 0, \ldots, T-1$,

$$\|\mathbf{L}_t - \mathbf{K}\| \le \|r(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}\| \left(\|\mathbf{K}\|^2 + \|\frac{\mu}{m_t}\breve{\mathbf{A}}_{m_t}^\top \breve{\mathbf{A}}_{m_t} - \mathbf{I}_d\| \right) = o_P\left(\sqrt{\frac{d}{r}}\right).$$

It follows that that for $i = 0, \ldots, T-1$,

$$\begin{aligned}
\Big\| \prod_{j=1}^{T-i-1} \mathbf{L}_{T-j} - \mathbf{K}^{T-i-1} \Big\| &= \Big\| \prod_{j=1}^{T-i-1} \{\mathbf{K} + (\mathbf{L}_{T-j} - \mathbf{K})\} - \mathbf{K}^{T-i-1} \Big\| \\
&\le (2^{T-i-1} - 1) \max_{\substack{k \in \{1,\ldots,T-i-1\} \\ t \in \{0,\ldots,T-1\}}} \left(\|\mathbf{L}_t - \mathbf{K}\|^k \, \|\mathbf{K}\|^{T-i-1-k} \right) \\
&= o_P\left(\left(\frac{d}{r}\right)^{\frac{T-i-1}{2}} \right).
\end{aligned}$$

Similarly, we have

$$\Big\| \prod_{j=1}^{T} \mathbf{L}_{T-j} - \mathbf{K}^{T} \Big\| = o_P\left(\left(\frac{d}{r}\right)^{\frac{T}{2}} \right).$$

This completes the proof. □

Lemma A.3. *Suppose the conditions of Theorem 3.1 hold. Then for $i = 0, \ldots, T-1$,*

$$\left\| \mathrm{E}\Big\{ \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k \Big) \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k \Big)^\top \mid \tilde{\mathbf{A}} \Big\} - \kappa \frac{m_i}{r^2} \mathbf{I}_d \right\| = o_P\left(\kappa \frac{m_i}{r^2} \right).$$

Also,

$$\left\| \mathrm{E}\Big\{ \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k \Big) \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k \Big)^\top \mid \tilde{\mathbf{A}} \Big\} - \kappa \frac{1}{r} \mathbf{I}_d \right\| = o_P\left(\kappa \frac{1}{r} \right).$$

Proof. For $i = 0, \dots, T-1$, we have

$$\mathrm{E}\left\{ \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mid \tilde{\mathbf{A}} \right\} = \kappa(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}(\breve{\mathbf{A}}_{m_i}^\top \breve{\mathbf{A}}_{m_i})(\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r)^{-1}.$$

Note that $\breve{\mathbf{A}}_{m_i}$ is a random matrix with independent $\mathcal{N}(0,1)$ entries, $i = 0, \dots, T-1$. Then from Lemma A.1, $\|\frac{1}{m_i}\breve{\mathbf{A}}_{m_i}^\top \breve{\mathbf{A}}_{m_i} - \mathbf{I}_d\| = o_P(1)$, $i = 0, \dots, T-1$. Similarly, $\|\frac{1}{r}\breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r - \mathbf{I}_d\| = o_P(1)$. For any fixed $\delta \in (0,1)$, with probability tending to 1, we have

$$(1-\delta)m_i\mathbf{I}_d \le \breve{\mathbf{A}}_{m_i}^\top \breve{\mathbf{A}}_{m_i} \le (1+\delta)m_i\mathbf{I}_d, \quad i = 0, \dots, T-1, \quad (1-\delta)r\mathbf{I}_d \le \breve{\mathbf{A}}_r^\top \breve{\mathbf{A}}_r \le (1+\delta)r\mathbf{I}_d.$$

If the above inequalities hold, then for $i = 0, \dots, T-1$,

$$\frac{(1-\delta)}{(1+\delta)^2}\kappa\frac{m_i}{r^2}\mathbf{I}_d \le \mathrm{E}\left\{ \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mid \tilde{\mathbf{A}} \right\} \le \frac{(1+\delta)}{(1-\delta)^2}\kappa\frac{m_i}{r^2}\mathbf{I}_d.$$

It follows that

$$\left\| \mathrm{E}\left\{ \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mid \tilde{\mathbf{A}} \right\} - \kappa\frac{m_i}{r^2}\mathbf{I}_d \right\| \le \max\left(\frac{3\delta-\delta^2}{(1-\delta)^2}, \frac{3\delta+\delta^2}{(1+\delta)^2} \right) \kappa\frac{m_i}{r^2}\mathbf{I}_d.$$

The first conclusion follows. The second conclusion follows from a similar argument.

□

Lemma A.4. *Suppose the conditions of Theorem 3.1 hold. Let $\tilde{\mathbf{Q}}, \mathbf{Q}_0, \dots, \mathbf{Q}_{N-1}$ be $d \times d$ symmetric matrices which are functions of $\tilde{\mathbf{A}}$. Then*

$$\mathrm{E}\left\{ \left\| \tilde{\mathbf{Q}} \sum_{k=1}^{r} \breve{\varepsilon}_k + \sum_{i=0}^{T-1} \mathbf{Q}_i \sum_{k=1}^{m_i} \breve{\varepsilon}_k \right\|^2 \mid \tilde{\mathbf{A}} \right\} = (1+o_P(1))\kappa \left\{ \frac{1}{r}\operatorname{tr}(\tilde{\mathbf{Q}}^2) + \sum_{i=0}^{T-1} \frac{m_i}{r^2}\operatorname{tr}(\mathbf{Q}_i^2) \right\}.$$

Proof. We have

$$\begin{aligned}
\mathrm{E}\left\{ \left\| \tilde{\mathbf{Q}} \sum_{k=1}^{r} \breve{\varepsilon}_k + \sum_{i=0}^{T-1} \mathbf{Q}_i \sum_{k=1}^{m_i} \breve{\varepsilon}_k \right\|^2 \mid \tilde{\mathbf{A}} \right\} &= \operatorname{tr}\Big\{ \tilde{\mathbf{Q}} \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)^\top \tilde{\mathbf{Q}} \Big\} + \sum_{i=0}^{T-1} \operatorname{tr}\Big\{ \mathbf{Q}_i \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_i \Big\} \\
&\quad + 2\sum_{i=0}^{T-1} \operatorname{tr}\Big\{ \tilde{\mathbf{Q}} \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_i \Big\} + 2\sum_{i=0}^{T-1}\sum_{j=i+1}^{T-1} \operatorname{tr}\Big\{ \mathbf{Q}_i \Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_j} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_j \Big\} \\
&=: u_1 + u_2 + u_3 + u_4.
\end{aligned}$$

Lemma A.3 implies that

$$\mathrm{E}(u_1 + u_2 \mid \tilde{\mathbf{A}}) = (1+o_P(1))\kappa\Big\{ \frac{1}{r}\operatorname{tr}(\tilde{\mathbf{Q}}^2) + \sum_{i=0}^{T-1} \frac{m_i}{r^2}\operatorname{tr}(\mathbf{Q}_i^2) \Big\}.$$

Hence we only need to show that $\mathrm{E}(u_3 \mid \tilde{\mathbf{A}})$ and $\mathrm{E}(u_4 \mid \tilde{\mathbf{A}})$ are negligible compared with $\mathrm{E}(u_1 + u_2 \mid \tilde{\mathbf{A}})$. From Cauchy-Schwarz inequality and Lemma A.3, for $i = 0, \dots, T-1$,

$$\begin{aligned}
\mathrm{E}\left[\operatorname{tr}\Big\{ \tilde{\mathbf{Q}} \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{m_i} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_i \Big\} \mid \tilde{\mathbf{A}} \right] &= \mathrm{E}\left[\operatorname{tr}\Big\{ \tilde{\mathbf{Q}} \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_i \Big\} \mid \tilde{\mathbf{A}} \right] \\
&\le \left[\mathrm{E}\left[\operatorname{tr}\Big\{ \tilde{\mathbf{Q}} \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)^\top \tilde{\mathbf{Q}} \Big\} \mid \tilde{\mathbf{A}} \right] \right]^{\frac{1}{2}} \left[\mathrm{E}\left[\operatorname{tr}\Big\{ \mathbf{Q}_i \Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)\Big(\sum_{k=1}^{r} \breve{\varepsilon}_k\Big)^\top \mathbf{Q}_i \Big\} \mid \tilde{\mathbf{A}} \right] \right]^{\frac{1}{2}} \\
&\le \Big\{ \mathrm{E}(u_1 \mid \tilde{\mathbf{A}})(1+o_P(1))\kappa\frac{1}{r}\operatorname{tr}(\mathbf{Q}_i^2) \Big\}^{\frac{1}{2}} \\
&= o_P(1)\Big\{ \mathrm{E}(u_1 \mid \tilde{\mathbf{A}})\,\mathrm{E}(u_2 \mid \tilde{\mathbf{A}}) \Big\}^{\frac{1}{2}}.
\end{aligned}$$

Thus, $\mathrm{E}(u_3 \mid \tilde{\mathbf{A}}) = o_P\{\mathrm{E}(u_1 \mid \tilde{\mathbf{A}}) + \mathrm{E}(u_2 \mid \tilde{\mathbf{A}})\}$. Similarly, for $0 \le i < j \le T-1$,

$$\begin{aligned}
\mathrm{E}\left[\mathrm{tr}\left\{\mathbf{Q}_i(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)(\sum_{k=1}^{m_j}\breve{\varepsilon}_k)^\top\mathbf{Q}_j\right\} \mid \tilde{\mathbf{A}}\right] &= \mathrm{E}\left[\mathrm{tr}\left\{\mathbf{Q}_i(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)^\top\mathbf{Q}_j\right\} \mid \tilde{\mathbf{A}}\right] \\
&\le \left[\mathrm{E}\left[\mathrm{tr}\left\{\mathbf{Q}_i(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)^\top\mathbf{Q}_i\right\} \mid \tilde{\mathbf{A}}\right]\right]^{\frac{1}{2}} \left[\mathrm{E}\left[\mathrm{tr}\left\{\mathbf{Q}_j(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)(\sum_{k=1}^{m_i}\breve{\varepsilon}_k)^\top\mathbf{Q}_j\right\} \mid \tilde{\mathbf{A}}\right]\right]^{\frac{1}{2}} \\
&\le \left\{\mathrm{E}(u_2 \mid \tilde{\mathbf{A}})(1+o_P(1))\kappa\frac{m_i}{r^2}\,\mathrm{tr}(\mathbf{Q}_j^2)\right\}^{\frac{1}{2}} \\
&\le \left(\frac{m_i}{m_j}\right)^{\frac{1}{2}} \mathrm{E}(u_2 \mid \tilde{\mathbf{A}}) \\
&= o_P\left\{\mathrm{E}(u_2 \mid \tilde{\mathbf{A}})\right\}.
\end{aligned}$$

This completes the proof. □

A.3. Proof of Proposition 3.2

In this section, we present the proof of Proposition 3.2. A technical result in the proof is deferred to Lemma A.5.

Proof of Proposition 3.2. Note that for any positive integer n,

$$\frac{C(n)}{C(n-1)} = \frac{4n-2}{n+1} < 4.$$

Hence the assumption $r \ge 4d$ implies that for $t = 1, \ldots, T-1$,

$$\frac{g(t,T)}{g(t-1,T)} = \frac{r}{d}\frac{C(T-t-1)}{C(T-t)} > 1.$$

That is, $g(t,T)$ is increasing in t. For the optimization problem (10), the objective function is strictly convex and the feasible region is convex and connected. Hence there is a unique solution, denoted by $(m_0^*, \ldots, m_{T-1}^*)$, to the problem (10).

If the restriction $m_t \le N$ is dropped, then from Cauchy-Schwarz inequality,

$$\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t} \ge \frac{1}{M}\left(\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t}\right)\left(\sum_{t=0}^{T-1} m_t\right) \ge \frac{1}{M}\left(\sum_{t=0}^{T-1}\sqrt{g(t,T)}\right)^2,$$

where the equalities hold if and only if

$$m_t = \frac{M\sqrt{g(t,T)}}{\sum_{j=0}^{T-1}\sqrt{g(j,T)}}, \quad t = 1, \ldots, T-1. \tag{23}$$

If $\frac{M\sqrt{g(T-1,T)}}{\sum_{j=0}^{T-1}\sqrt{g(j,T)}} \le N$, then (23) is exactly the solution to the problem (10).

Now we consider the case that $\frac{M\sqrt{g(T-1,T)}}{\sum_{j=0}^{T-1}\sqrt{g(j,T)}} > N$. We claim that in this case, the solution to the problem (10) satisfies $m_{T-1}^* = N$. The proof of this claim is deferred to Lemma A.5. In this case, $m_1^*, \ldots, m_{T-2}^*$ are the solution to the optimization problem

$$\min_{m_0,\ldots,m_{T-2}\in\mathbb{R}} \sum_{t=0}^{T-2}\frac{g(t,T)}{m_t} \quad \text{s. t.} \quad \sum_{t=0}^{T-2} m_t \le M - N, \quad 0 < m_t \le N,\ t = 0, \ldots, T-2.$$

This problem has the same structure as the original problem. We can apply the above arguments recursively until the algorithm is finished. □

Lemma A.5. *Suppose the conditions of Proposition 3.2 hold. If* $\frac{M\sqrt{g(T-1,T)}}{\sum_{j=0}^{T-1}\sqrt{g(j,T)}} > N$, *then the solution to the problem* (10) *satisfies* $m_{T-1}^* = N$.

Proof. For $t = 0, \dots, T-1$, define functions

$$m_t(h) := (1-h)m_t^* + h\frac{M\sqrt{g(t,T)}}{\sum_{j=0}^{T-1}\sqrt{g(j,T)}}, \quad h \in [0,1].$$

Since $(m_0(1), \dots, m_{T-1}(1))$ is the solution to the relaxed problem, we have

$$\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(1)} < \sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(0)}.$$

Then for any $h \in (0,1]$,

$$\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(h)} \le h\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(1)} + (1-h)\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(0)} < \sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(0)},$$

where the first inequality follows from Jensen's inequality. Since $(m_0(0), \dots, m_{T-1}(0))$ is the solution to the problem (10), for any $h \in (0,1]$, $(m_1(h), \dots, m_{T-1}(h))$ must violate the constraint $m_t(h) \le N$, $t = 0, \dots, T-1$. As a consequence, there exists an $t^* \in \{0, \dots, N-1\}$ such that $m_{t^*}(0) = N$. Suppose our claim does not hold, that is, $m_{T-1}(0) < N$. Then we have $t^* \neq T-1$. In this case, we define $m_t^\dagger = m_t(0)$ for $i \notin \{t^*, T-1\}$, $m_{t^*}^\dagger = m_{T-1}(0)$ and $m_{T-1}^\dagger = N$. Then

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{g(t,T)}{m_t(0)} - \sum_{t=0}^{T-1}\frac{g(t,T)}{m_t^\dagger} =& \frac{g(t^*,T)}{N} + \frac{g(T-1,T)}{m_{T-1}(0)} - \left(\frac{g(t^*,T)}{m_{T-1}(0)} + \frac{g(T-1,T)}{N}\right)\\ =& (g(t^*,T) - g(T-1,T))\left(\frac{1}{N} - \frac{1}{m_{T-1}(0)}\right) > 0,\end{aligned}$$

which contradicts the definition of $(m_0(0), \dots, m_{T-1}(0))$. This completes the proof. □

A.4. Proof of Theorem 3.3

In this section, we provide the proof of Theorem 3.3.

Proof of Theorem 3.3. By the definition (12) of T, we have

$$\sqrt{\frac{r^{T+1}}{4d^{T-1}C(T-2)}} < N, \quad \sqrt{\frac{r^{T+2}}{4d^{T}C(T-1)}} \ge N. \tag{24}$$

As a consequence of (24), we have

$$m_{T-1} = \min\left(\sqrt{\frac{r^{T+2}}{4d^{T}C(T-1)}}, N\right) = N, \quad m_{T-2} = \min\left(\sqrt{\frac{r^{T+1}}{4d^{T-1}C(T-1)}}, N\right) < N.$$

It follows that $T = T^\dagger + 1$.

From the first inequality in (24), we have

$$N > \sqrt{\frac{r^{T+1}}{4d^{T-1}C(T-2)}} \ge \sqrt{\frac{r^{T+1}}{(4d)^{T-1}}}.$$

Thus, $T < \frac{2\log(\frac{N}{r})}{\log(\frac{r}{4d})} + 1$. Hence by assumption, T is bounded. By a standard subsequence argument, we can without loss of generality and assume that T is fixed. From the definition of m_t, we have $\frac{r}{m_0} \to 0$ and $\frac{m_t}{m_{t+1}} \to 0$, $t = 0, \dots, T^\dagger - 2$.

Hence if only $T^\dagger$ iterations are performed, then the conditions of Theorem 3.1 hold. It follows that

$$\begin{aligned}
\mathrm{E}\left\{\left\|\mathbf{A}(\mathbf{x}_{T^\dagger}-\mathbf{x}^*)\right\|^2 \mid \tilde{\mathbf{A}}\right\} =&(1+o_P(1))\kappa\left\{\left(\frac{d}{r}\right)^{T^\dagger+1} C(T^\dagger)+\sum_{t=0}^{T^\dagger-1}\frac{g(t,T^\dagger)}{m_t}\right\}\\
=&(1+o_P(1))\kappa\left\{\left(\frac{d}{r}\right)^{T^\dagger+1} C(T^\dagger)+\sum_{t=0}^{T^\dagger-1}\sqrt{C(T^\dagger-1)C(T^\dagger-t-1)\frac{4d^{2T^\dagger-t+1}}{r^{2T^\dagger-t+1}}}\right\}\\
\le&(1+o_P(1))\frac{\kappa}{4}\left\{\left(\frac{4d}{r}\right)^{T^\dagger+1}+\sum_{t=0}^{T^\dagger-1}\left(\frac{4d}{r}\right)^{T^\dagger-\frac{t}{2}+\frac{1}{2}}\right\}\\
=&(1+o_P(1))\frac{\kappa}{4}\left(\frac{4d}{r}\right)^{\frac{T^\dagger}{2}+1}.
\end{aligned}$$

It follows from the above inequality and Markov's inequality that

$$\left\|\mathbf{A}(\mathbf{x}_{T^\dagger}-\mathbf{x}^*)\right\|^2 = O_P\Big(\kappa\Big(\frac{d}{r}\Big)^{\frac{T^\dagger}{2}+1}\Big). \tag{25}$$

By definition, the $(T^\dagger+1)$th iteration is an IHS iteration. Hence

$$\mathbf{x}_T = \mathbf{x}_{T^\dagger} - \mu r(\tilde{\mathbf{A}}_r^\top\tilde{\mathbf{A}}_r)^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{T^\dagger}-\mathbf{y}).$$

It follows that

$$\mathbf{x}_T-\mathbf{x}^* = \left(\mathbf{I}_d-\mu r(\tilde{\mathbf{A}}_r^\top\tilde{\mathbf{A}}_r)^{-1}\mathbf{A}^\top\mathbf{A}\right)(\mathbf{x}_{T^\dagger}-\mathbf{x}^*).$$

With the notation $\breve{\mathbf{A}}_k$ in the proof of Theorem 3.1, we have

$$\mathbf{D}_\mathbf{A}\mathbf{V}^\top(\mathbf{x}_T-\mathbf{x}^*) = \left\{\mathbf{I}_d-\mu r(\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\right\}\mathbf{D}_\mathbf{A}\mathbf{V}^\top(\mathbf{x}_{T^\dagger}-\mathbf{x}^*).$$

Thus,

$$\|\mathbf{A}(\mathbf{x}_T-\mathbf{x}^*)\| \le \left\|\mathbf{I}_d-\mu r(\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\right\|\|\mathbf{A}^\top(\mathbf{x}_{T^\dagger}-\mathbf{x}^*)\|. \tag{26}$$

From Lemma A.1 and the condition $|\mu-1|=O(d/r)$, we have

$$\left\|\mathbf{I}_d-\mu r(\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\right\| \le \mu\|r(\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r)^{-1}\|\left\|\frac{1}{r}\breve{\mathbf{A}}_r^\top\breve{\mathbf{A}}_r-\mathbf{I}_d\right\|+|\mu-1| = o_P\left(\left(\frac{d}{r}\right)^{\frac{1}{2}}\right). \tag{27}$$

Combining (25), (26) and (27) yields

$$\|\mathbf{A}(\mathbf{x}_T-\mathbf{x}^*)\|^2 = O_P\left(\kappa\left(\frac{d}{r}\right)^{\frac{T^\dagger}{2}+2}\right).$$

Then from (24),

$$\sqrt{\frac{r^{T^\dagger+3}}{4d^{T^\dagger+1}}} \ge N.$$

The above inequality implies that

$$\left(\frac{d}{r}\right)^{\frac{T^\dagger}{2}+2} \le \frac{r}{2N}\left(\frac{d}{r}\right)^{\frac{3}{2}} = \frac{d}{2N}\left(\frac{d}{r}\right)^{\frac{1}{2}} = o\left(\frac{d}{N}\right).$$

Then the conclusion follows.

□

A.5. Proof of Theorem 4.1

Our proof of Theorem 4.1 relies on the following lemma.

Lemma A.6. *Suppose $\frac{N}{m}$ is an integer. Then for any non-random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, we have*

$$\mathrm{E}\{(\mathbf{x}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{y} - \mathbf{x}^\top \mathbf{y})^2\} \leq \frac{1}{m}\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + (\mathbf{x}^\top \mathbf{y})^2\right).$$

Proof. By the definition of $\mathbf{P}_N$, there is a uniformly random permutation of $\{1, \ldots, N\}$, denote as π, such that $(\mathbf{P})_{i,\pi(i)} = 1$ and $(\mathbf{P})_{i,j} = 0$ for $j \neq \pi(i)$. By the definition of $\mathbf{D}_N$, we can write $\mathbf{D}_N = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$ where $\sigma_1, \ldots, \sigma_N$ are i.i.d. Rademacher random variables. With these notations, we have

$$\hat{\mathbf{S}}_{m,N}\mathbf{x} = \begin{pmatrix} \sum_{i=1}^{\frac{N}{m}} \sigma_i x_{\pi(i)} \\ \vdots \\ \sum_{i=\frac{m-1}{m}N+1}^{N} \sigma_i x_{\pi(i)} \end{pmatrix}.$$

Hence

$$\mathbf{x}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{y} - \mathbf{x}^\top \mathbf{y} = \sum_{k=1}^{m} \sum_{\frac{k-1}{m}N < i < j \leq \frac{k}{m}N} \sigma_i \sigma_j (x_{\pi(i)} y_{\pi(j)} + x_{\pi(j)} y_{\pi(i)}).$$

Note that for $i < j$ and $i' < j'$ such that $\{i, j\} \neq \{i', j'\}$, we have $\mathrm{E}(\sigma_i \sigma_j \sigma_{i'} \sigma_{j'}) = 0$. Thus,

$$\begin{aligned}
\mathrm{E}\{(\mathbf{x}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{y} - \mathbf{x}^\top \mathbf{y})^2\} =& \sum_{k=1}^{m} \sum_{\frac{k-1}{m}N < i < j \leq \frac{k}{m}N} \mathrm{E}\left\{(x_{\pi(i)} y_{\pi(j)} + x_{\pi(j)} y_{\pi(i)})^2\right\} \\
=& N\left(\frac{N}{m} - 1\right) \mathrm{E}(x_{\pi(1)}^2 y_{\pi(2)}^2 + x_{\pi(1)} y_{\pi(1)} x_{\pi(2)} y_{\pi(2)}) \\
=& \frac{N(\frac{N}{m} - 1)}{N(N-1)} \sum_{1 \leq i \neq j \leq N} (x_i^2 y_j^2 + x_i y_i x_j y_j) \\
\leq& \frac{1}{m} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i^2 y_j^2 + x_i y_i x_j y_j) \\
=& \frac{1}{m}\left(\|\mathbf{x}\|^2\|\mathbf{y}\|^2 + (\mathbf{x}^\top \mathbf{y})^2\right).
\end{aligned}$$

This completes the proof. □

Proof of Theorem 4.1. Let $\mathbf{u}_j \in \mathbb{R}^N$ denote the jth column of $\mathbf{U}$. From Lemma A.6,

$$\mathrm{E}\{\|\mathbf{U}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\|_F^2\} = \sum_{i=1}^{d} \mathrm{E}\{(\mathbf{u}_i^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{u}_i - 1)^2\} + \sum_{1 \leq i \neq j \leq d} \mathrm{E}\{(\mathbf{u}_i^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{u}_j)^2\} \leq \frac{d(d+1)}{m}.$$

Then Markov's inequality implies that

$$\Pr\left\{\|\mathbf{U}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\} \leq \Pr\left\{\|\mathbf{U}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\|_F > \epsilon\right\} \leq \frac{\mathrm{E}\{\|\mathbf{U}^\top \hat{\mathbf{S}}_{m,N}^\top \hat{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\|_F^2\}}{\epsilon^2} \leq \frac{d(d+1)}{m\epsilon^2}.$$

This completes the proof. □

A.6. Proof of Theorem 4.2

Mackey et al. (2014) gave a matrix Khintchine inequality for Hermitian matrices (Mackey et al. (2014), Corollary 7.3). This result can be generalized to general matrices via *Hermitian dilation*. This generalized version is a key tool in our proof of Theorem 4.2. For any matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ and $p \geq 1$, let $\|\mathbf{A}\|_{S_p} = \left[\mathrm{tr}\left\{(\mathbf{A}^\top \mathbf{A})^{\frac{p}{2}}\right\}\right]^{\frac{1}{p}}$ denote the Schatten p-norm of $\mathbf{A}$. We have the following lemma.

Lemma A.7 (Matrix Khintchine inequality)**.** *Suppose that $p = 1$ or $p \geq 1.5$. Let $\{\mathbf{A}_k\}_{k=1}^N$ be a sequence of non-random symmetric matrices. Let $\{\epsilon_i\}_{i=1}^N$ be a sequence of independent Rademacher random variables. Then*

$$\mathrm{E}\left(\left\|\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right\|_{S_{2p}}^{2p}\right) \leq \frac{1}{2}(2p)^p \left\|\sum_{i=1}^N \mathbf{A}_i \mathbf{A}_i^\top\right\|_{S_p}^p + \frac{1}{2}(2p)^p \left\|\sum_{i=1}^N \mathbf{A}_i^\top \mathbf{A}_i\right\|_{S_p}^p.$$

Proof. Let $\mathbf{B}_i = \begin{pmatrix} \mathbf{O}_{n_1,n_1} & \mathbf{A}_i \\ \mathbf{A}_i^\top & \mathbf{O}_{n_2,n_2} \end{pmatrix}$ be the Hermitian dilation of $\mathbf{A}_i$, $i = 1, \ldots, N$. By construction, $\mathbf{B}_i$ is symmetric. From Mackey et al. (2014), Corollary 7.3, we have

$$\mathrm{E}\left(\left\|\sum_{i=1}^N \epsilon_i \mathbf{B}_i\right\|_{S_{2p}}^{2p}\right) \leq (2p)^p \,\mathrm{E}\left(\left\|\sum_{i=1}^N \mathbf{B}_i^2\right\|_{S_p}^p\right).$$

Note that

$$\begin{aligned}
\left\|\sum_{i=1}^N \epsilon_i \mathbf{B}_i\right\|_{S_{2p}}^{2p} &= \mathrm{tr}\left\{\begin{pmatrix} \mathbf{O}_{n_1,n_1} & \sum_{i=1}^N \epsilon_i \mathbf{A}_i \\ \sum_{i=1}^N \epsilon_i \mathbf{A}_i^\top & \mathbf{O}_{n_2,n_2} \end{pmatrix}^{2p}\right\} \\
&= \mathrm{tr}\left\{\begin{pmatrix} \left\{\left(\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right)\left(\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right)^\top\right\}^p & \mathbf{O}_{n_1,n_2} \\ \mathbf{O}_{n_2,n_1} & \left\{\left(\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right)^\top\left(\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right)\right\}^p \end{pmatrix}\right\} \\
&= 2\left\|\sum_{i=1}^N \epsilon_i \mathbf{A}_i\right\|_{S_{2p}}^{2p}.
\end{aligned}$$

On the other hand,

$$\left\|\sum_{i=1}^N \mathbf{B}_i^2\right\|_{S_p}^p = \mathrm{tr}\left\{\begin{pmatrix} \sum_{i=1}^N \mathbf{A}_i \mathbf{A}_i^\top & \mathbf{O}_{n_1,n_2} \\ \mathbf{O}_{n_2,n_1} & \sum_{i=1}^N \mathbf{A}_i^\top \mathbf{A}_i \end{pmatrix}^p\right\} = \left\|\sum_{i=1}^N \mathbf{A}_i \mathbf{A}_i^\top\right\|_{S_p}^p + \left\|\sum_{i=1}^N \mathbf{A}_i^\top \mathbf{A}_i\right\|_{S_p}^p.$$

Hence the conclusion holds. □

The following lemma provides a standard decoupling technique.

Lemma A.8. *Suppose $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a non-random matrix whose diagonal elements equal 0. Let $\mathbf{D} \in \mathbb{R}^{N\times N}$ be a diagonal matrix whose diagonal elements are independent and have mean 0. Let $\mathbf{D}'$ be an independent copy of $\mathbf{D}$. Then for any convex function $F : \mathbb{R}^{N\times N} \to \mathbb{R}$, we have*

$$\mathrm{E}\, F(\mathbf{DAD}) \leq \mathrm{E}\, F(4\mathbf{DAD}').$$

Proof. Let $\mathbf{W} \in \mathbb{R}^{N\times N}$ be a random diagonal matrix which is independent of $\mathbf{D}$ and $\mathbf{D}'$. The diagonal elements of $\mathbf{W}$ are i.i.d. random variables taking on values 0 or 1 with equal probability. Since the diagonal elements of $\mathbf{A}$ equal 0, we have $\mathbf{A} = 4\,\mathrm{E}\{\mathbf{WA}(\mathbf{I}_N - \mathbf{W})\}$. Hence $\mathbf{DAD} = 4\,\mathrm{E}\{\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D} \mid \mathbf{D}\}$. From Jensen's inequality,

$$\mathrm{E}\, F(\mathbf{DAD}) = \mathrm{E}\, F(4\,\mathrm{E}\{\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D} \mid \mathbf{D}\}) \leq \mathrm{E}\, F(4\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}).$$

Conditioning on $\mathbf{W}$, the random matrices $\mathbf{DW}$ and $(\mathbf{I}_N - \mathbf{W})\mathbf{D}$ are independent. Thus, the distribution of $\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}$ is the same as that of $\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}'$. Hence

$$\mathrm{E}\, F(\mathbf{DAD}) \leq \mathrm{E}\, F(4\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}').$$

We have

$$\mathbf{DAD}' = \mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}' + \mathbf{DWAWD}' + \mathbf{D}(\mathbf{I}_N - \mathbf{W})\mathbf{AWD}' + \mathbf{D}(\mathbf{I}_N - \mathbf{W})\mathbf{A}(\mathbf{I}_N - \mathbf{W})\mathbf{D}'.$$

Note that conditioning on $\mathbf{W}$, the random matrices $\mathbf{DW}$, $\mathbf{D}(\mathbf{I}_N - \mathbf{W})$, $\mathbf{WD}'$ and $(\mathbf{I}_N - \mathbf{W})\mathbf{D}'$ are independent. Hence we have $\mathrm{E}\{\mathbf{DWAWD}' + \mathbf{D}(\mathbf{I}_N - \mathbf{W})\mathbf{AWD}' + \mathbf{D}(\mathbf{I}_N - \mathbf{W})\mathbf{A}(\mathbf{I}_N - \mathbf{W})\mathbf{D}' \mid \mathbf{W}, \mathbf{DW}, (\mathbf{I}_N - \mathbf{W})\mathbf{D}'\} = \mathbf{O}_{N\times N}$. Thus,

$$\mathrm{E}\{\mathbf{DAD}' \mid \mathbf{W}, \mathbf{DW}, (\mathbf{I}_N - \mathbf{W})\mathbf{D}'\} = \mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}'.$$

Then it follows from Jensen's inequality that

$$\mathrm{E}\, F(4\mathbf{DWA}(\mathbf{I}_N - \mathbf{W})\mathbf{D}') \le \mathrm{E}\, F(4\mathbf{DAD}').$$

Hence the conclusion holds. □

Proof of Theorem 4.2. Let $\ell \ge 1.5$ be a parameter to be specified. From Markov's inequality,

$$\Pr(\left\|\mathbf{U}^\top \check{\mathbf{S}}_{m,N}^\top \check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\right\| > \epsilon) \le \epsilon^{-2\ell}\, \mathrm{E}\left(\left\|\mathbf{U}^\top \check{\mathbf{S}}_{m,N}^\top \check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right).$$

We only need to bound the expectation of the right hand side.

Let $\tilde{\mathbf{U}} := \mathbf{P}_N\mathbf{W}_N\tilde{\mathbf{D}}_N\mathbf{U}$. Then $\tilde{\mathbf{U}}$ is a column orthogonal matrix. It can be seen that $\mathbf{U}^\top \check{\mathbf{S}}_{m,N}^\top \check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d = \tilde{\mathbf{U}}^\top\mathbf{D}_N\left(\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^* - \mathbf{I}_N\right)\mathbf{D}_N\tilde{\mathbf{U}}$. Note that the matrix $\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^* - \mathbf{I}_N$ has zero diagonal elements. For any given $\tilde{\mathbf{U}}$, the function $\mathbf{B} \mapsto \|\tilde{\mathbf{U}}^\top\mathbf{B}\tilde{\mathbf{U}}\|_{S_{2\ell}}^{2\ell}$ is convex. Hence we can applying Lemma A.8 conditional on $\tilde{\mathbf{U}}$ and obtain the bound

$$\begin{aligned}\mathrm{E}\left(\left\|\mathbf{U}^\top \check{\mathbf{S}}_{m,N}^\top \check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \le& 2^{4\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\left(\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^* - \mathbf{I}_N\right)\mathbf{D}_N'\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right)\\ \le& \frac{1}{2}2^{6\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^*\mathbf{D}_N'\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right) + \frac{1}{2}2^{6\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{D}_N'\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right), \end{aligned} \tag{28}$$

where $\mathbf{D}_N'$ is an independent copy of $\mathbf{D}_N$ and the last inequality follows from the triangle inequality of the Schatten norm. Let $\mathbf{D}_{N,k} \in \mathbb{R}^{\frac{N}{m}\times\frac{N}{m}}$ denote the $(k-1)\frac{N}{m}+1$ to $k\frac{N}{m}$ rows and $(k-1)\frac{N}{m}+1$ to $k\frac{N}{m}$ columns of $\mathbf{D}_N$, $k = 1,\dots,m$. We define $\mathbf{D}_{N,k}'$ similarly. Let $\tilde{\mathbf{U}}_k \in \mathbb{R}^{\frac{N}{m}\times d}$ denote the $(k-1)\frac{N}{m}+1$ to $k\frac{N}{m}$ rows of $\tilde{\mathbf{U}}$, $k = 1,\dots,m$. Then we have $\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^*\mathbf{D}_N'\tilde{\mathbf{U}} = \sum_{k=1}^m \tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}'\tilde{\mathbf{U}}_k$. Let $\sigma_1,\dots,\sigma_m$ be i.i.d. Rademacher random variables which are independent of $\mathbf{D}_N$, $\mathbf{D}_N'$ and $\tilde{\mathbf{U}}$. Note that $\mathbf{D}_{N,k}$ has the same distribution as $\sigma_k\mathbf{D}_{N,k}$, $k = 1,\dots,m$. Hence $\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^*\mathbf{D}_N'\tilde{\mathbf{U}}$ has the same distribution as $\sum_{k=1}^m \sigma_k\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}'\tilde{\mathbf{U}}_k$. Then from Lemma A.7, we have

$$\begin{aligned}&\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^*\mathbf{D}_N'\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right)\\ \le&(2\ell)^\ell\,\mathrm{E}\left(\left\|\sum_{k=1}^m\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}'\tilde{\mathbf{U}}_k\right\|^2\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}\tilde{\mathbf{U}}_k\right\|_{S_\ell}^\ell\right)\\ \le&(2\ell)^\ell\,\mathrm{E}\left\{\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}'\tilde{\mathbf{U}}_k\right\|^{2\ell}\right)\left\|\sum_{k=1}^m\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}\tilde{\mathbf{U}}_k\right\|_{S_\ell}^\ell\right\}\\ \le&(2\ell)^\ell\sqrt{\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}'\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)\mathrm{E}\left(\left\|\sum_{k=1}^m\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}\tilde{\mathbf{U}}_k\right\|_{S_\ell}^{2\ell}\right)}, \end{aligned} \tag{29}$$

where the last inequality follows from Cauchy-Schwarz inequality. Furthermore, we have

$$\begin{aligned}\mathrm{E}\left\{\left\|\sum_{k=1}^m\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}\tilde{\mathbf{U}}_k\right\|_{S_\ell}^{2\ell}\right\} \le& d\,\mathrm{E}\left\{\left\|\sum_{k=1}^m\tilde{\mathbf{U}}_k^\top\mathbf{D}_{N,k}\mathbf{1}_{\frac{N}{m}}\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}_{N,k}\tilde{\mathbf{U}}_k\right\|_{S_{2\ell}}^{2\ell}\right\}\\ =& d\,\mathrm{E}\left\{\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{G}_{m,N}^{*\top}\mathbf{G}_{m,N}^*\mathbf{D}_N\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right\}\\ =& d\,\mathrm{E}\left\{\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U}\right\|_{S_{2\ell}}^{2\ell}\right\}\\ \le&\frac{1}{2}d2^{2\ell}\left[\mathrm{E}\left\{\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U} - \mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right\} + d\right], \end{aligned} \tag{30}$$

where the first inequality follows from Cauchy-Schwarz inequality and the last inequality follows from the triangle inequality of the Schatten norm. It follows from (28), (29), and (30) that

$$\begin{aligned}
\mathrm{E}\left(\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U}-\mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \le& \frac{1}{2}2^{8\ell}\ell^\ell\sqrt{d\,\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)\left[\mathrm{E}\left\{\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U}-\mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right\}+d\right]} \\
&+\frac{1}{2}2^{6\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{D}'_N\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right) \\
\le& \frac{1}{2}2^{8\ell}\ell^\ell\sqrt{d\,\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)\mathrm{E}\left\{\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U}-\mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right\}} \\
&+\frac{1}{2}2^{8\ell}\ell^\ell d\sqrt{\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)}+\frac{1}{2}2^{6\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{D}'_N\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right).
\end{aligned}$$

Note that if $x^2 \le ax+b$ for some $a,b\ge 0$, then $x^2\le a^2+2b$. Hence the above inequality implies that

$$\begin{aligned}
\mathrm{E}\left(\left\|\mathbf{U}^\top\check{\mathbf{S}}_{m,N}^\top\check{\mathbf{S}}_{m,N}\mathbf{U}-\mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \le& \frac{1}{4}2^{16\ell}\ell^{2\ell}d\,\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right) \\
&+2^{8\ell}\ell^\ell d\sqrt{\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)}+2^{6\ell}\,\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top\mathbf{D}_N\mathbf{D}'_N\tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right). \qquad (31)
\end{aligned}$$

Now we deal with the first two terms in (31). We have

$$\mathrm{E}\left(\max_{k\in\{1,\dots,m\}}\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)\le\mathrm{E}\left(\sum_{k=1}^m\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,k}\tilde{\mathbf{U}}_k\right\|^{4\ell}\right)=m\,\mathrm{E}\left(\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,1}\tilde{\mathbf{U}}_1\right\|^{4\ell}\right),$$

where the last equality holds since $\mathbf{D}'_{N,k}$ has the same distribution as $\mathbf{D}'_{N,1}$ and $\tilde{\mathbf{U}}_k$ has the same distribution as $\tilde{\mathbf{U}}_1$. Denote $\breve{\mathbf{U}}:=\mathbf{W}_N\tilde{\mathbf{D}}_N\mathbf{U}$. Then $\tilde{\mathbf{U}}_1$ consists of $\frac{N}{m}$ uniformly sampled (without replacement) rows from $\breve{\mathbf{U}}$. Let $\delta_1,\dots,\delta_N$ be random variables taking on values 0 and 1 which indicates $\frac{N}{m}$ uniformly sampled (without replacement) elements from $\{1,\dots,N\}$. Hence the set $\{i:\delta_i=1\}$ has exactly $\frac{N}{m}$ elements. Let $\sigma_1,\dots,\sigma_N$ be i.i.d. Rademacher random variables which are independent of $\delta_1,\dots,\delta_N$ and $\breve{\mathbf{U}}$. Then $\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,1}\tilde{\mathbf{U}}_1$ has the same distribution as $\sum_{i=1}^N\sigma_i\delta_i\breve{\mathbf{u}}_i^\top$ where $\breve{\mathbf{u}}_i\in\mathbb{R}^d$ is the ith row of $\breve{\mathbf{U}}$. From Lemma (A.7),

$$\begin{aligned}
\mathrm{E}\left(\left\|\mathbf{1}_{\frac{N}{m}}^\top\mathbf{D}'_{N,1}\tilde{\mathbf{U}}_1\right\|^{4\ell}\right)=&\,\mathrm{E}\,\mathrm{E}\left(\left\|\sum_{i=1}^N\sigma_i\delta_i\breve{\mathbf{u}}_i\right\|^{4\ell}\mid\delta_1,\dots,\delta_N,\breve{\mathbf{U}}\right) \\
\le&\frac{1}{2}(4\ell)^{2\ell}\,\mathrm{E}\left\{\left(\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\right)^{2\ell}\right\}+\frac{1}{2}(4\ell)^{2\ell}\,\mathrm{E}\left(\left\|\sum_{i=1}^N\delta_i\breve{\mathbf{u}}_i\breve{\mathbf{u}}_i^\top\right\|_{S_{2\ell}}^{2\ell}\right) \\
\le&\frac{1}{2}(4\ell)^{2\ell}\,\mathrm{E}\left\{\left(\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\right)^{2\ell}\right\}+\frac{1}{2}(4\ell)^{2\ell}\,\mathrm{E}\left(\left\|\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \\
=&\frac{1}{2}(1+d)(4\ell)^{2\ell}\,\mathrm{E}\left\{\left(\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\right)^{2\ell}\right\} \\
\le&d(4\ell)^{2\ell}\,\mathrm{E}\left\{\left(\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\right)^{2\ell}\right\}.
\end{aligned}$$

Note that there are only $\frac{N}{m}$ nonzero elements among $\delta_1\|\breve{\mathbf{u}}_1\|^2,\dots,\delta_N\|\breve{\mathbf{u}}_N\|^2$. Hence from Jensen's inequality, we have

$$\mathrm{E}\left\{\left(\sum_{i=1}^N\delta_i\|\breve{\mathbf{u}}_i\|^2\right)^{2\ell}\right\}\le\left(\frac{N}{m}\right)^{2\ell-1}\sum_{i=1}^N\mathrm{E}\left(\delta_i\|\breve{\mathbf{u}}_i\|^{4\ell}\right)=\left(\frac{N}{m}\right)^{2\ell-1}\frac{1}{m}\sum_{i=1}^N\mathrm{E}\left(\|\breve{\mathbf{u}}_i\|^{4\ell}\right)=\left(\frac{N}{m}\right)^{2\ell}\mathrm{E}\left(\|\breve{\mathbf{u}}_1\|^{4\ell}\right),$$

where the second last equality holds since $\mathrm{E}(\delta_i) = \frac{1}{m}$ and the last equality holds since $\breve{\mathbf{u}}_1, \ldots, \breve{\mathbf{u}}_N$ are identically distributed. We note that $\breve{\mathbf{u}}_1$ has the same distribution as $\frac{1}{\sqrt{N}} \sum_{i=1}^N \sigma_i \mathbf{u}_i$, where $\mathbf{u}_i \in \mathbb{R}^d$ is the ith row of $\mathbf{U}$. Hence from Lemma A.7,

$$\mathrm{E}\left(\|\breve{\mathbf{u}}_1\|^{4\ell}\right) = \frac{1}{N^{2\ell}} \mathrm{E}\left(\left\|\sum_{i=1}^N \sigma_i \mathbf{u}_i\right\|^{4\ell}\right) \le \frac{(4\ell)^{2\ell}}{2N^{2\ell}} \left[\left\|\sum_{i=1}^N \mathbf{u}_i \mathbf{u}_i^\top\right\|_{S_{2\ell}}^{2\ell} + \left(\sum_{i=1}^N \|\mathbf{u}_i\|^2\right)^{2\ell}\right] = \frac{(4\ell)^{2\ell}}{2N^{2\ell}}(d + d^{2\ell}) \le \left(\frac{4\ell d}{N}\right)^{2\ell}. \quad (32)$$

Combining the above bounds yields,

$$\mathrm{E}\left(\max_{k\in\{1,\ldots,m\}} \left\|\mathbf{1}_{\frac{N}{m}}^\top \mathbf{D}'_{N,k} \tilde{\mathbf{U}}_k\right\|^{4\ell}\right) \le md(4\ell)^{2\ell} \left(\frac{N}{m}\right)^{2\ell} \left(\frac{4\ell d}{N}\right)^{2\ell} = md\left(\frac{16\ell^2 d}{m}\right)^{2\ell}. \quad (33)$$

Now we deal with the last term in (31). Let $\tilde{\mathbf{u}}_i \in \mathbb{R}^d$ denote the vector of the ith row of $\tilde{\mathbf{U}}$. Let $\sigma_1, \ldots, \sigma_N$ be i.i.d. Rademacher random variables which are independent of $\tilde{\mathbf{U}}$. Then the random matrix $\tilde{\mathbf{U}}^\top \mathbf{D}_N \mathbf{D}'_N \tilde{\mathbf{U}}$ has the same distribution as $\sum_{i=1}^N \sigma_i \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^\top$. Hence from Lemma A.7,

$$\begin{aligned}
\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top \mathbf{D}_N \mathbf{D}'_N \tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right) =& \mathrm{E}\left(\left\|\sum_{i=1}^N \sigma_i \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^\top\right\|_{S_{2\ell}}^{2\ell}\right) \\
\le& (2\ell)^\ell \mathrm{E}\left(\left\|\sum_{i=1}^N \|\tilde{\mathbf{u}}_i\|^2 \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^\top\right\|_{S_\ell}^{\ell}\right) \\
\le& (2\ell)^\ell \mathrm{E}\left(\left\|\left(\max_{i\in\{1,\ldots,N\}} \|\tilde{\mathbf{u}}_i\|^2\right)\left(\sum_{i=1}^N \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^\top\right)\right\|_{S_\ell}^{\ell}\right) \\
=& (2\ell)^\ell d \,\mathrm{E}\left(\max_{i\in\{1,\ldots,N\}} \|\tilde{\mathbf{u}}_i\|^{2\ell}\right),
\end{aligned}$$

where the last equality holds since $\sum_{i=1}^N \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^\top = \mathbf{I}_d$ and $\|\mathbf{I}_d\|_{S_\ell}^\ell = d$. We have

$$\mathrm{E}\left(\max_{i\in\{1,\ldots,N\}} \|\tilde{\mathbf{u}}_i\|^{2\ell}\right) = \mathrm{E}\left(\max_{i\in\{1,\ldots,N\}} \|\breve{\mathbf{u}}_i\|^{2\ell}\right) \le N \mathrm{E}\left(\|\breve{\mathbf{u}}_1\|^{2\ell}\right) \le N\left\{\mathrm{E}\left(\|\breve{\mathbf{u}}_1\|^{4\ell}\right)\right\}^{\frac{1}{2}} \le N\left(\frac{4\ell d}{N}\right)^\ell,$$

where the last inequality follows from (32). Thus,

$$\mathrm{E}\left(\left\|\tilde{\mathbf{U}}^\top \mathbf{D}_N \mathbf{D}'_N \tilde{\mathbf{U}}\right\|_{S_{2\ell}}^{2\ell}\right) \le Nd\left(\frac{8\ell^2 d}{N}\right)^\ell \le md\left(\frac{8\ell^2 d}{m}\right)^\ell. \quad (34)$$

It follows from (31), (33) and (34) that

$$\begin{aligned}
\mathrm{E}\left(\left\|\mathbf{U}^\top \breve{\mathbf{S}}_{m,N}^\top \breve{\mathbf{S}}_{m,N} \mathbf{U} - \mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \le& \frac{1}{4} 2^{16\ell} \ell^{2\ell} dmd \left(\frac{16\ell^2 d}{m}\right)^{2\ell} + 2^{8\ell} \ell^\ell d \sqrt{md\left(\frac{16\ell^2 d}{m}\right)^{2\ell}} + 2^{6\ell} md\left(\frac{8\ell^2 d}{m}\right)^\ell \\
=& \frac{1}{4} 2^{24\ell} md^2 \left(\frac{\ell^3 d}{m}\right)^{2\ell} + 2^{12\ell} d \sqrt{md\left(\frac{\ell^3 d}{m}\right)^{2\ell}} + 2^{9\ell} md\left(\frac{\ell^2 d}{m}\right)^\ell \\
\le& 2^{24\ell} md^2 \left\{\left(\frac{\ell^3 d}{m}\right)^{2\ell} + \left(\frac{\ell^3 d}{m}\right)^\ell\right\}.
\end{aligned}$$

Thus, for $\epsilon \in (0, 1)$,

$$\Pr\left(\left\|\mathbf{U}^\top \breve{\mathbf{S}}_{m,N}^\top \breve{\mathbf{S}}_{m,N} \mathbf{U} - \mathbf{I}_d\right\| > \epsilon\right) \le \epsilon^{-2\ell} \mathrm{E}\left(\left\|\mathbf{U}^\top \breve{\mathbf{S}}_{m,N}^\top \breve{\mathbf{S}}_{m,N} \mathbf{U} - \mathbf{I}_d\right\|_{S_{2\ell}}^{2\ell}\right) \le 2^{24\ell} md^2 \left\{\left(\frac{\ell^3 d}{m\epsilon^2}\right)^{2\ell} + \left(\frac{\ell^3 d}{m\epsilon^2}\right)^\ell\right\}. \quad (35)$$

We take $\ell = \log(\frac{e^2 d}{\epsilon\delta}) > 1.5$. Define $\check{m} := \gamma \epsilon^{-2} d \left\{\log(\frac{e^2 d}{\epsilon\delta})\right\}^3$ where $\gamma > 1$ is an absolute constant to be specified. Then for $m \ge \check{m}$, we have $\frac{\ell^3 d}{m\epsilon^2} < 1$. Hence for $m \ge \check{m}$,

$$2^{24\ell} md^2 \left\{\left(\frac{\ell^3 d}{m\epsilon^2}\right)^{2\ell} + \left(\frac{\ell^3 d}{m\epsilon^2}\right)^\ell\right\} \le 2 \cdot 2^{24\ell} \check{m} d^2 \left(\frac{\ell^3 d}{\check{m}\epsilon^2}\right)^\ell = \left[\exp\left\{\frac{\log(2) + 2\log(d) + \log(\check{m})}{\ell}\right\} \frac{2^{24}}{\gamma}\right]^\ell. \quad (36)$$

Note that

$$\frac{\log(2)+2\log(d)+\log(\check{m})}{\ell} \leq \frac{\log(2)+3\ell+\log(\gamma)+3\log(\ell)}{\ell} \leq \frac{\log(2)}{1.5}+6+\frac{\log(\gamma)}{1.5}. \tag{37}$$

Hence we take a constant γ such that $\gamma > (\frac{\log(3)}{1.5}+6+\frac{\log(\gamma)}{1.5})2^{24}e$ (obviously, such a γ exists). Then from (35), (36), (37) and the choice of γ, for $m \geq \check{m}$, we have

$$\Pr(\left\|\mathbf{U}^\top \check{\mathbf{S}}_{m,N}^\top \check{\mathbf{S}}_{m,N}\mathbf{U}-\mathbf{I}_d\right\| > \epsilon) \leq e^{-\ell} \leq e^{-\log(\frac{1}{\delta})} = \delta.$$

This completes the proof.

□

A.7. Proof of Theorem 4.3

In this section, we prove Theorem 4.3. First we introduce some notations. For $\epsilon \in (0,1)$, define the event

$$\mathcal{E}_\epsilon := \bigcap_{t=0}^{T^\dagger-1} \left\{\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}_\mathbf{A}-\mathbf{I}_d\| \leq \epsilon\right\}.$$

For $\epsilon \in (0,1)$, let $\tilde{\mathcal{E}}_\epsilon$ denote the event

$$\tilde{\mathcal{E}}_\epsilon := \left\{\|\mathbf{U}_\mathbf{A}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}\mathbf{U}_\mathbf{A}-\mathbf{I}_d\| \leq \epsilon\right\}.$$

For $t=0,\dots,T^\dagger-1$, let

$$\mathbf{x}_t^* := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{S}_t(\mathbf{A}\mathbf{x}-\mathbf{y})\|^2 = (\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{y}.$$

Lemma A.9. *Let $\epsilon,\delta \in (0,1)$. Suppose $\mathbf{S} \in \mathbb{R}^{n\times N}$ and $\mathbf{S}' \in \mathbb{R}^{m\times n}$ are sketching matrices. Suppose $\mathbf{S}$ and $\mathbf{S}'$ are independent. Then for any column orthogonal matrix $\mathbf{U} \in \mathbb{R}^{N\times d}$, on the event $\left\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\| < \frac{1}{3}\epsilon\right\}$, we have*

$$\Pr\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{S}\mathbf{U}-\mathbf{I}_d\| > \epsilon \mid \mathbf{S}\} < \sup_{\substack{\mathbf{U}'\in\mathbb{R}^{n\times d},\\ \mathbf{U}'^\top\mathbf{U}'=\mathbf{I}_d}} \Pr\left\{\|\mathbf{U}'^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}'-\mathbf{I}_d\| > \frac{1}{3}\epsilon\right\}.$$

Proof. We have

$$\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{S}\mathbf{U}-\mathbf{I}_d\| \leq \|\mathbf{U}^\top\mathbf{S}^\top(\mathbf{S}'^\top\mathbf{S}'-\mathbf{I}_d)\mathbf{S}\mathbf{U}\| + \|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\|.$$

Denote by $\mathbf{S}\mathbf{U} = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^\top$ the compact singular value decomposition of $\mathbf{S}\mathbf{U}$, where $\mathbf{U}_0 \in \mathbb{R}^{n\times d}$ and $\mathbf{V}_0 \in \mathbb{R}^{d\times d}$ are column orthogonal matrices and $\mathbf{D}_0 \in \mathbb{R}^{d\times d}$ is a diagonal matrix. Then we have

$$\|\mathbf{U}^\top\mathbf{S}^\top(\mathbf{S}'^\top\mathbf{S}'-\mathbf{I}_d)\mathbf{S}\mathbf{U}\| \leq \|\mathbf{U}_0^\top(\mathbf{S}'^\top\mathbf{S}'-\mathbf{I}_d)\mathbf{U}_0\|\|\mathbf{D}_0\mathbf{V}_0^\top\|^2 = \|\mathbf{U}_0^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}_0-\mathbf{I}_d\|\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}\|.$$

It follows that

$$\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{S}\mathbf{U}-\mathbf{I}_d\| \leq \|\mathbf{U}_0^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}_0-\mathbf{I}_d\|\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\| + \|\mathbf{U}_0^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}_0-\mathbf{I}_d\| + \|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\|.$$

Hence on the event $\left\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\| < \frac{1}{3}\epsilon\right\}$, we have

$$\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{S}\mathbf{U}-\mathbf{I}_d\| \leq 2\|\mathbf{U}_0^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}_0-\mathbf{I}_d\| + \frac{1}{3}\epsilon.$$

Thus, on the event $\left\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}-\mathbf{I}_d\| < \frac{1}{3}\epsilon\right\}$,

$$\begin{aligned}\Pr\left\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{S}\mathbf{U}-\mathbf{I}_d\| > \epsilon \mid \mathbf{S}\right\} &\leq \Pr\left\{\|\mathbf{U}_0^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}_0-\mathbf{I}_d\| > \frac{1}{3}\epsilon \mid \mathbf{S}\right\}\\ &\leq \sup_{\substack{\mathbf{U}'\in\mathbb{R}^{n\times d},\\ \mathbf{U}'^\top\mathbf{U}'=\mathbf{I}_d}} \Pr\left\{\|\mathbf{U}'^\top\mathbf{S}'^\top\mathbf{S}'\mathbf{U}'-\mathbf{I}_d\| > \frac{1}{3}\epsilon\right\},\end{aligned}$$

where the last inequality holds since $\mathbf{S}'$ and $\mathbf{U}_0$ are independent. This completes the proof. □

The following Lemma is a restatement of Lemma 4.1 of Boutsidis & Gittens (2013). It gives the embedding property of the SRHT sketching matrix.

Lemma A.10. *Let* $\mathbf{S}$ *be an* $m \times N$ *SRHT sketching matrix. Suppose* N *is a power of* 2, $\mathbf{U} \in \mathbb{R}^{N\times d}$ *is a column orthogonal matrix and* $\epsilon, \delta \in (0,1)$. *Suppose*

$$m \geq c\epsilon^{-2}\left(d + \log\left(\frac{N}{\delta}\right)\right)\log\left(\frac{ed}{\delta}\right),$$

where $c > 0$ *is an absolute constant. Then*

$$\Pr\left\{\|\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\} \leq \delta.$$

Lemma A.11. *Suppose the conditions of Theorem 4.3 holds. Then for any column orthogonal matrix* $\mathbf{U} \in \mathbb{R}^{N\times d}$, *we have*

$$\Pr\left\{\bigcup_{t=0}^{T^\dagger-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\}\right\} \leq \delta,$$

and

$$\Pr\left\{\|\mathbf{U}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U} - \mathbf{I}_d\| > \epsilon\right\} \leq \delta.$$

Proof. For sufficiently large absolute constant $\tilde{\gamma}$, we have $m_{T^\diamond} \geq 324\frac{d(d+1)}{\delta\epsilon^2}$. Then from Theorem 4.1, we have

$$\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U} - \mathbf{I}_d\| > \frac{1}{9}\epsilon\right\} \leq \frac{81d(d+1)}{m_t\epsilon^2}, \quad t = T^\diamond, \ldots, T^\dagger - 1.$$

In particular, we have

$$\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\mathbf{S}_{T^\diamond}\mathbf{U} - \mathbf{I}_d\| > \frac{1}{9}\epsilon\right\} \leq \frac{\delta}{4}. \tag{38}$$

From the union bound, we have

$$\begin{aligned}
\Pr\left\{\bigcup_{t=T^\diamond}^{T^\dagger-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U} - \mathbf{I}_d\| > \frac{1}{9}\epsilon\right\}\right\} &= \sum_{t=T^\diamond}^{T^\dagger-1}\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U} - \mathbf{I}_d\| > \frac{1}{9}\epsilon\right\} \\
&\leq \sum_{t=T^\diamond}^{T^\dagger-1}\frac{81d(d+1)}{m_t\epsilon^2} \\
&= \sum_{t=T^\diamond}^{T^\dagger-1}\frac{81d(d+1)}{2^{t-T^\diamond}m_{T^\diamond}\epsilon^2} \\
&\leq \frac{162d(d+1)}{m_{T^\diamond}\epsilon^2} \\
&\leq \frac{\delta}{2},
\end{aligned} \tag{39}$$

where the last inequality holds since $m_{T^\diamond} \geq 324\frac{d(d+1)}{\delta\epsilon^2}$.

For $t = 0, \ldots, T^\diamond - 1$, the distribution of $\mathbf{S}_t$ is the same as that of $\check{\mathbf{S}}_{m_t, m_{T^\diamond}}\mathbf{S}_{T^\diamond}$ where $\check{\mathbf{S}}_{m_t,m_{T^\diamond}}$ and $\mathbf{S}_{T^\diamond}$ are independent. For sufficiently large absolute constant $\tilde{\gamma}$, we have $m_0 > 81\gamma\epsilon^{-2}d\{\log(\frac{72e^2d}{\epsilon\delta})\}^3$. Then from Theorem 4.2, we have

$$\sup_{\substack{\mathbf{U}'\in\mathbb{R}^{m_{T^\diamond}\times d},\\ \mathbf{U}'^\top\mathbf{U}'=\mathbf{I}_d}}\Pr\left\{\|\mathbf{U}'^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}\mathbf{U}' - \mathbf{I}_d\| > \frac{1}{9}\epsilon\right\} \leq \frac{9e^2d}{\epsilon}\exp\left\{-\left(\frac{m_t\epsilon^2}{81\gamma d}\right)^{\frac{1}{3}}\right\}, \quad t = 0, \ldots, T^\diamond - 1. \tag{40}$$

Therefore,

$$\begin{aligned}
\Pr\left\{\bigcup_{t=0}^{T^\diamond-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\right\} &= \mathrm{E}\left[\Pr\left\{\bigcup_{t=0}^{T^\diamond-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\mid\mathbf{S}_{T^\diamond}\right\}\right] \\
&\le \mathrm{E}\left[\Pr\left\{\bigcup_{t=0}^{T^\diamond-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\mid\mathbf{S}_{T^\diamond}\right\}\mathbf{1}_{\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\mathbf{S}_{T^\diamond}\mathbf{U}-\mathbf{I}_d\|\le\frac{1}{9}\epsilon\right\}}\right] \\
&\quad+\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\mathbf{S}_{T^\diamond}\mathbf{U}-\mathbf{I}_d\|>\frac{1}{9}\epsilon\right\} \\
&\le \mathrm{E}\left[\sum_{t=0}^{T^\diamond-1}\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\mid\mathbf{S}_{T^\diamond}\right\}\mathbf{1}_{\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\mathbf{S}_{T^\diamond}\mathbf{U}-\mathbf{I}_d\|\le\frac{1}{9}\epsilon\right\}}\right]+\frac{\delta}{4},
\end{aligned}$$

where the last inequality follows from the union bound and (38). From Lemma A.9 and (40), on the event $\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\mathbf{S}_{T^\diamond}\mathbf{U}-\mathbf{I}_d\|\le\frac{1}{9}\epsilon\right\}$, we have

$$\begin{aligned}
\sum_{t=0}^{T^\diamond-1}\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\mid\mathbf{S}_{T^\diamond}\right\} &= \sum_{t=0}^{T^\diamond-1}\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_{T^\diamond}^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}\mathbf{S}_{T^\diamond}\mathbf{U}\|>\frac{1}{3}\epsilon\mid\mathbf{S}_{T^\diamond}\right\} \\
&\le \sum_{t=0}^{T^\diamond-1}\sup_{\substack{\mathbf{U}'\in\mathbb{R}^{m_{T^\diamond}\times d},\\ \mathbf{U}'^\top\mathbf{U}'=\mathbf{I}_d}}\Pr\left\{\|\mathbf{U}'^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}^\top\check{\mathbf{S}}_{m_t,m_{T^\diamond}}\mathbf{U}'-\mathbf{I}_d\|>\frac{1}{9}\epsilon\right\} \\
&\le \sum_{t=0}^{T^\diamond-1}\frac{9e^2d}{\epsilon}\exp\left\{-\left(\frac{m_t\epsilon^2}{81\gamma d}\right)^{\frac{1}{3}}\right\} \\
&\le \sum_{t=0}^{+\infty}\frac{9e^2d}{\epsilon}\exp\left\{-2^{\frac{t}{3}}\left(\frac{m_0\epsilon^2}{81\gamma d}\right)^{\frac{1}{3}}\right\} \\
&\le \sum_{t=0}^{+\infty}\frac{9e^2d}{\epsilon}\left(\frac{\epsilon\delta}{72e^2d}\right)^{2^{\frac{t}{3}}} \\
&\le \frac{\delta}{8}\sum_{t=0}^{+\infty}\left(\frac{\epsilon}{72e^2d}\right)^{2^{\frac{t}{3}}-1} \\
&\le \frac{\delta}{4},
\end{aligned}$$

where the fourth inequality holds since $m_0>81\gamma\epsilon^{-2}d\{\log(\frac{72e^2d}{\epsilon\delta})\}^3$. It follows that

$$\Pr\left\{\bigcup_{t=0}^{T^\diamond-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\right\}\le\frac{\delta}{2}. \tag{41}$$

From (39) and (41) and the union bound,

$$\Pr\left\{\bigcup_{t=0}^{T^\dagger-1}\left\{\|\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}-\mathbf{I}_d\|>\epsilon\right\}\right\}\le\delta.$$

Hence the first conclusion holds.

By construction, $\tilde{\mathbf{S}}=\mathbf{S}^\dagger\mathbf{S}_0$ where $\mathbf{S}^\dagger\in\mathbb{R}^{r\times m_0}$ is an SRHT matrix which is independent of $\mathbf{S}_0$. From Lemma A.10, for sufficiently large absolute constant $\tilde{\gamma}$,

$$\sup_{\substack{\mathbf{U}'\in\mathbb{R}^{m_0\times d},\\ \mathbf{U}'^\top\mathbf{U}'=\mathbf{I}_d}}\Pr\left\{\|\mathbf{U}'^\top\mathbf{S}^{\dagger\top}\mathbf{S}^\dagger\mathbf{U}'-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\le\frac{\delta}{2}.$$

From (41), we have

$$\Pr\left\{\|\mathbf{U}^\top\mathbf{S}_0^\top\mathbf{S}_0\mathbf{U}-\mathbf{I}_d\|>\frac{1}{3}\epsilon\right\}\leq\frac{\delta}{2}.$$

It follows from the above two inequalities and Lemma A.9 that

$$\Pr\left\{\|\mathbf{U}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}-\mathbf{I}_d\|>\epsilon\right\}\leq\delta.$$

Hence the second conclusion holds. □

Lemma A.12. *Suppose the conditions of Theorem 4.3 hold. Then on $\mathcal{E}_\epsilon$, for $t=0,\ldots,T^\dagger-1$,*

$$\|\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)\|\leq\frac{1}{1-\epsilon}\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|.$$

And on $\tilde{\mathcal{E}}_\epsilon$

$$\|\mathbf{A}(\mathbf{x}_0-\mathbf{x}^*)\|\leq\frac{1}{1-\epsilon}\|\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|.$$

Proof. We have

$$\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)=-\mathbf{A}(\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})=-\mathbf{U}_\mathbf{A}(\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y}).$$

It follows that

$$\|\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)\|\leq\|(\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A})^{-1}\|\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|.$$

On $\mathcal{E}_\epsilon$, we have

$$\|(\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A})^{-1}\|=\|(\mathbf{I}_d+\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}-\mathbf{I}_d)^{-1}\|\leq\frac{1}{1-\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}-\mathbf{I}_d\|}\leq\frac{1}{1-\epsilon}.$$

Hence the first conclusion holds. The proof of the second conclusion is similar. □

Lemma A.13. *Suppose the conditions of Theorem 4.3 hold. Then on $\tilde{\mathcal{E}}_\epsilon$, we have,*

$$\|\mathbf{I}_d-\mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|\leq\frac{\epsilon+|\mu-1|}{1-\epsilon}.$$

Proof. We have

$$\begin{aligned}\|\mathbf{I}_d-\mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|=&\|(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A}-\mu\mathbf{I}_d)\|\\ \leq&\|(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|\|\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A}-\mu\mathbf{I}_d\|\\ \leq&\|(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|\left(\|\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A}-\mathbf{I}_d\|+|\mu-1|\right).\end{aligned}$$

On $\tilde{\mathcal{E}}_\epsilon$, we have $\|(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|\leq\frac{1}{1-\epsilon}$. Hence the conclusion follows. □

Lemma A.14. *Suppose the conditions of Theorem 4.3 hold. Then on $\mathcal{E}_\epsilon\cap\tilde{\mathcal{E}}_\epsilon$, we have, for $t=0,\ldots,T^\dagger-1$,*

$$\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}^*)\|\leq\frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon}\|\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)\|+\frac{1+\mu\epsilon+|\mu-1|}{(1-\epsilon)^2}\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|.$$

Proof. By the triangle inequality,

$$\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}^*)\|\leq\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}_t^*)\|+\|\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)\|. \tag{42}$$

Now we deal with the term $\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}_t^*)\|$ on the right hand side of (42). From the definition of $\mathbf{x}_t^*$, we have the normal equation $\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{A}\mathbf{x}_t^* = \mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{y}$. Then from the update formula (6) we have

$$\begin{aligned}\mathbf{x}_{t+1} =&\mathbf{x}_t - \mu\tilde{\mathbf{H}}^{-1}\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}_t-\mathbf{y})\\ =&\mathbf{x}_t - \mu\tilde{\mathbf{H}}^{-1}\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{A}(\mathbf{x}_t-\mathbf{x}_t^*)\\ =&\mathbf{x}_t - \mu\mathbf{V}_\mathbf{A}\mathbf{D}_\mathbf{A}^{-1}(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t-\mathbf{x}_t^*).\end{aligned}$$

It follows that

$$\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_{t+1}-\mathbf{x}_t^*) = \left\{\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\right\}\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t-\mathbf{x}_t^*).$$

Thus,

$$\|\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_{t+1}-\mathbf{x}_t^*)\| \le \|\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\|\|\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t-\mathbf{x}_t^*)\|. \tag{43}$$

Note that

$$\begin{aligned}\|\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\| =&\|\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1} + \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}(\mathbf{I}_d - \mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A})\|\\ \le&\|\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\| + \mu\|(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\|\|\mathbf{I}_d - \mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\|.\end{aligned}$$

Thus, from Lemma A.13, on $\mathcal{E}\cap\tilde{\mathcal{E}}$, we have

$$\|\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}_\mathbf{A}\| \le \frac{\epsilon+|\mu-1|}{1-\epsilon} + \mu\frac{\epsilon}{1-\epsilon} = \frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon}.$$

From the above inequality and (43), we have, on $\mathcal{E}\cap\tilde{\mathcal{E}}$, that

$$\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}_t^*)\| \le \frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon}\left(\|\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)\| + \|\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)\|\right).$$

Combining the above inequality and (42) leads to that on $\mathcal{E}\cap\tilde{\mathcal{E}}$,

$$\begin{aligned}\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}^*)\| \le& \frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon}\|\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)\| + \frac{1+\mu\epsilon+|\mu-1|}{1-\epsilon}\|\mathbf{A}(\mathbf{x}_t^*-\mathbf{x}^*)\|\\ \le& \frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon}\|\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)\| + \frac{1+\mu\epsilon+|\mu-1|}{(1-\epsilon)^2}\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|,\end{aligned}$$

where the last inequality follows from Lemma A.12. This completes the proof. □

Lemma A.15. *Suppose the conditions of Theorem 4.3 hold. Then on $\tilde{\mathcal{E}}_\epsilon$, for $t = T^\dagger,\dots,T-1$, we have*

$$\|\mathbf{A}(\mathbf{x}_{t+1}-\mathbf{x}^*)\| \le \frac{\epsilon+|\mu-1|}{1-\epsilon}\|\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)\|.$$

Proof. For $t = T^\dagger,\dots,T-1$, we have

$$\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_{t+1}-\mathbf{x}^*) = \left\{\mathbf{I}_d - \mu(\mathbf{U}_\mathbf{A}^\top\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}}\mathbf{U}_\mathbf{A})^{-1}\right\}\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top(\mathbf{x}_t-\mathbf{x}^*).$$

Then the conclusion follows from Lemma A.13. □

Proof of Theorem 4.3. Since $0<\epsilon\le\frac{1}{10}$ and $|\mu-1|\le\frac{1}{4}$, we have

$$\frac{(\mu+1)\epsilon+|\mu-1|}{1-\epsilon} \le \frac{1}{2}, \quad \frac{1+\mu\epsilon+|\mu-1|}{(1-\epsilon)^2} \le 2.$$

From Lemma A.15, on $\tilde{\mathcal{E}}_\epsilon$, we have

$$\|\mathbf{A}(\mathbf{x}_T-\mathbf{x}^*)\| \le \frac{1}{2^{T-T^\dagger}}\|\mathbf{A}(\mathbf{x}_{T^\dagger}-\mathbf{x}^*)\|. \tag{44}$$

Lemma A.14 implies that on $\mathcal{E}_\epsilon \cap \tilde{\mathcal{E}}_\epsilon$, we have, for $t = 0, \dots, T^\dagger - 1$, that

$$\|\mathbf{A}(\mathbf{x}_{t+1} - \mathbf{x}^*)\| \leq \frac{1}{2}\|\mathbf{A}(\mathbf{x}_t - \mathbf{x}^*)\| + 2\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|.$$

Then by induction, on $\mathcal{E}_\epsilon \cap \tilde{\mathcal{E}}_\epsilon$, we have

$$\|\mathbf{A}(\mathbf{x}_{T^\dagger} - \mathbf{x}^*)\| \leq \frac{1}{2^{T^\dagger}}\|\mathbf{A}(\mathbf{x}_0 - \mathbf{x}^*)\| + \sum_{t=0}^{T^\dagger - 1} \frac{1}{2^{T^\dagger - t - 2}}\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|. \tag{45}$$

From Lemma A.11, we have $\Pr(\mathcal{E}_\epsilon^\complement) \leq \delta$ and $\Pr(\tilde{\mathcal{E}}_\epsilon^\complement) \leq \delta$. Thus, the inequalities (44) and (45) holds with probablity at least $1 - 2\delta$.

Now we deal with the second term of (45). Let $\mathbf{z}_1, \dots, \mathbf{z}_d \in \mathbb{R}^N$ denote the columns of $\mathbf{U}_\mathbf{A}$. Then for $t = T^\diamond, \dots, T^\dagger - 1$, we have

$$\mathrm{E}\left\{\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2\right\} = \sum_{j=1}^d \mathrm{E}\{(\mathbf{z}_j^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\} \leq \frac{d}{m_t}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2,$$

where the last inequality follows from Lemma A.6.

We can write $\mathbf{S}_t = \hat{\mathbf{S}}_{m_t, m_{T^\diamond}} \mathbf{W}_{m_{T^\diamond}} \mathbf{S}_{T^\diamond}$, $t = 0, \dots, T^\diamond - 1$. Hence for $t = 0, \dots, T^\diamond - 1$,

$$\begin{aligned}
\mathrm{E}\left\{\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2\right\} &= \sum_{j=1}^d \mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{W}_{m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}} \mathbf{W}_{m_{T^\diamond}} \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\right\} \\
&= \sum_{j=1}^d \mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{W}_{m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}} \mathbf{W}_{m_{T^\diamond}} \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}) - \mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\right\} \\
&\quad + \sum_{j=1}^d \mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\right\},
\end{aligned}$$

where the last equality holds since the cross term has zero mean. From Lemma A.6, we have

$$\sum_{j=1}^d \mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\right\} \leq \frac{d}{m_{T^\diamond}}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2.$$

On the other hand, Lemma A.6 implies that

$$\begin{aligned}
&\mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{W}_{m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}} \mathbf{W}_{m_{T^\diamond}} \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}) - \mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2\right\} \\
=& \mathrm{E}\,\mathrm{E}\left\{(\mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{W}_{m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}}^\top \hat{\mathbf{S}}_{m_t, m_{T^\diamond}} \mathbf{W}_{m_{T^\diamond}} \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}) - \mathbf{z}_j^\top \mathbf{S}_{T^\diamond}^\top \mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y}))^2 \mid \mathbf{S}_{T^\diamond}\right\} \\
\leq& \frac{2}{m_t}\,\mathrm{E}\left(\|\mathbf{S}_{T^\diamond}\mathbf{z}_j\|^2 \|\mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2\right) \\
=& \frac{2}{m_t}\,\mathrm{E}\left\{(\|\mathbf{S}_{T^\diamond}\mathbf{z}_j\|^2 - \|\mathbf{z}_j\|^2)(\|\mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2)\right\} + \frac{2}{m_t}\|\mathbf{z}_j\|^2\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2 \\
\leq& \frac{2}{m_t}\sqrt{\mathrm{E}\left\{(\|\mathbf{S}_{T^\diamond}\mathbf{z}_j\|^2 - \|\mathbf{z}_j\|^2)^2\right\}\mathrm{E}\left\{(\|\mathbf{S}_{T^\diamond}(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2 - \|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2)^2\right\}} + \frac{2}{m_t}\|\mathbf{z}_j\|^2\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2 \\
\leq& \left(\frac{4}{m_t m_{T^\diamond}} + \frac{2}{m_t}\right)\|\mathbf{z}_j\|^2\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2 \\
\leq& \frac{4}{m_t}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2.
\end{aligned}$$

Combining the above bounds yields

$$\mathrm{E}\left\{\|\mathbf{U}_\mathbf{A}^\top \mathbf{S}_t^\top \mathbf{S}_t(\mathbf{A}\mathbf{x}^* - \mathbf{y})\|^2\right\} \leq \frac{5d}{m_t}\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|^2, \quad t = 0, \dots, T^\diamond - 1.$$

It follows that

$$\begin{aligned}
\mathrm{E}\left\{\sum_{t=0}^{T^\dagger-1}\frac{1}{2^{T^\dagger-t-2}}\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|\right\} &\le \sum_{t=0}^{T^\dagger-1}\frac{1}{2^{T^\dagger-t-2}}\left[\mathrm{E}\left\{\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\|^2\right\}\right]^{\frac{1}{2}} \\
&\le \sum_{t=0}^{T^\dagger-1}\frac{\sqrt{5}}{2^{T^\dagger-t-2}}\sqrt{\frac{d}{m_t}}\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\| \\
&= \sum_{t=0}^{T^\dagger-1}\frac{\sqrt{5}}{2^{T^\dagger-\frac{t}{2}-2}}\sqrt{\frac{d}{m_0}}\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\| \\
&= \frac{\sqrt{5}}{2^{T^\dagger-2}}\frac{2^{\frac{T^\dagger}{2}}-1}{\sqrt{2}-1}\sqrt{\frac{d}{m_0}}\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\| \\
&\le 4\sqrt{5}(\sqrt{2}+1)\sqrt{\frac{d}{N}}\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|.
\end{aligned}$$

Then from Markov's inequality, with probablity at least $1-\delta$,

$$\sum_{t=0}^{T^\dagger-1}\frac{1}{2^{T^\dagger-t-2}}\|\mathbf{U}_\mathbf{A}^\top\mathbf{S}_t^\top\mathbf{S}_t(\mathbf{A}\mathbf{x}^*-\mathbf{y})\| \le \frac{4\sqrt{5}(\sqrt{2}+1)}{\delta}\sqrt{\frac{d}{N}}\|\mathbf{A}\mathbf{x}^*-\mathbf{y}\|.$$

Then the conclusion follows from (44), (45) and the above inequality.

□