Experimental Design for Any p-Norm

Lap Chi Lau
David R. Cheriton School of Computer Science, University of Waterloo, Canada

Robert Wang
David R. Cheriton School of Computer Science, University of Waterloo, Canada

Hong Zhou[1]
School of Mathematics and Statistics, Fuzhou University, China

Abstract

We consider a general p-norm objective for experimental design problems that captures some well-studied objectives (D/A/E-design) as special cases. We prove that a randomized local search approach provides a unified algorithm to solve this problem for all nonnegative integer p. This provides the first approximation algorithm for the general p-norm objective, and a nice interpolation of the best known bounds of the special cases.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis; Theory of computation → Rounding techniques

Keywords and phrases Approximation Algorithm, Optimal Experimental Design, Randomized Local Search

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2023.4

Category APPROX

Related Version *Full Version*: `https://arxiv.org/abs/2305.01942`

Funding *Lap Chi Lau*: Supported in part by NSERC Discovery Grant.
Robert Wang: Supported in part by NSERC Discovery Grant and Canada Graduate Scholarship.
Hong Zhou: Supported in part by Fuzhou University startup funding (XRC-22056) and Natural Science Foundation of Fujian Province of China (2023J01414).

Acknowledgements We thank Mohit Singh for bringing the Φ_p objective function to our attention. We also thank anonymous reviewers of an earlier version of this manuscript for helpful suggestions.

1 Introduction

In experimental design problems, we are given vectors $v_1, \ldots, v_n \in \mathbb{R}^d$ and a budget $k \geq d$, and the goal is to choose a subset S of k vectors so that $\sum_{i \in S} v_i v_i^\top$ optimizes some objective function that measures the "diversity" of the input data. The most popular and well-studied objective functions are:

- D-design: Maximizing $(\det(\sum_{i \in S} v_i v_i^\top))^{\frac{1}{d}}$.
- A-design: Minimizing $\frac{1}{d}\operatorname{tr}((\sum_{i \in S} v_i v_i)^{-1})$.
- E-design: Maximizing $\lambda_{\min}(\sum_{i \in S} v_i v_i^\top)$.

Experimental design problems have a long history and wide applications, from statistics to machine learning to numerical linear algebra to graph algorithms. For more information on these applications, we refer the reader to [13, 11, 1, 7, 6] and the references therein.

Although the objectives of D/A/E-design look quite different, we observe that there is a natural generalization using eigenvalues that captures all three objectives as special cases.

[1] Corresponding author

▶ **Definition 1** (p-Norm of Inverse Eigenvalues). *Given a d-dimensional real-symmetric matrix A with eigenvalues $\lambda_1, \ldots, \lambda_d > 0$ and a natural number $0 \le p \le \infty$, we define*

$$\Phi_p(A) := \left(\frac{1}{d}\operatorname{tr}\left(A^{-p}\right)\right)^{\frac{1}{p}} = \left(\frac{1}{d}\sum_{i=1}^{d}\lambda_i^{-p}\right)^{\frac{1}{p}}, \tag{1}$$

with $\Phi_0(A) := \lim_{p\to 0^+}\Phi_p(A)$ and $\Phi_\infty(A) := \lim_{p\to+\infty}\Phi_p(A)$.

Given $p \ge 0$, we refer to the experimental design problem with the objective function Φ_p as Φ_p-design. To see that Φ_p-design is a generalization of D/A/E-design, let $A = \sum_{i\in S} v_i v_i^T$ and note that:

- For $p = \infty$, $\Phi_\infty(A) = \lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A)$, which is the inverse of the E-design objective;
- For $p = 1$, $\Phi_1(A) = \frac{1}{d}\operatorname{tr}(A^{-1})$ is exactly the A-design objective;
- For $p = 0$, $\Phi_0(A)$ is the inverse of the D-design objective, as

  $$\Phi_0(A) = \lim_{p\to 0^+}\left(\frac{1}{d}\sum_{i=1}^{d}\lambda_i^{-p}\right)^{1/p} = \left(\prod_{i=1}^{d}\lambda_i^{-1}\right)^{1/d} = \det(A)^{-1/d},$$

  where the second equality is a well-known fact (see, e.g., Exercise 28 in Chapter 5 of [16]).

It is known that $\Phi_p(A)$ is convex in A for any given $0 \le p \le \infty$, and so the following is a natural convex programming relaxation for Φ_p-design:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & \Phi_p\left(\sum_{i=1}^{n} x(i)\cdot v_i v_i^\top\right) \\ \text{subject to} \quad & \sum_{i=1}^{n} x(i) \le k, \\ & 0 \le x(i) \le 1, \qquad \text{for } 1 \le i \le n. \end{aligned} \tag{2}$$

To the best of our knowledge, there are no known approximation algorithms for the general Φ_p-design problem, other than the special cases $p = 0, 1, \infty$ which we summarize as follows (the notation $x \gtrsim y$ denotes that $x \ge cy$ for some large enough constant c):

- There is a $(1+\varepsilon)$-approximation algorithm for Φ_0-design (D-design) when $k \gtrsim d/\varepsilon$ [13, 9, 11, 6].
- There is a $(1+\varepsilon)$-approximation algorithm for Φ_1-design (A-design) when $k \gtrsim d/\varepsilon$ [9, 11, 6].
- There is a $(1+\varepsilon)$-approximation algorithm for Φ_∞-design (E-design) when $k \gtrsim d/\varepsilon^2$ [1, 7].

These results are tight in the sense that they match the known integrality gap lower bound of the convex programming relaxation (2) (see [11] for integrality gap examples).

Note that there is a d/ε vs d/ε^2 gap between the relaxations for D/A-design ($p = 0, 1$) and for E-design ($p = \infty$). The main question that we study in this paper is: How does the integrality gap of the convex programming relaxation (2) change with varying value of p? In particular, where does the transition from d/ε to d/ε^2 happen?

1.1 Main Result

Our main result is that, when $k \gtrsim \min\{dp/\varepsilon, d/\varepsilon^2\}$, there is a $(1+\varepsilon)$-approximation algorithm for Φ_p-design.

▶ **Theorem 2.** *Given an integer* $p \geq 1$*, let* $x \in [0,1]^n$ *be an optimal fractional solution to* (2)*. For any* $\varepsilon \in (0,1)$*, let* $\gamma = \max\{\varepsilon, 1/p\}$*, if* $k \gtrsim d/(\gamma\varepsilon)$*, then there is a randomized polynomial time algorithm that returns an integral solution* $Z = \sum_{i=1}^{n} z(i) \cdot v_i v_i^\top$ *with* $z(i) \in \{0,1\}$ *for* $1 \leq i \leq n$ *such that*

$$\Phi_p\left(\sum_{i=1}^{n} z(i) \cdot v_i v_i^\top\right) \leq (1+\varepsilon) \cdot \Phi_p\left(\sum_{i=1}^{n} x(i) \cdot v_i v_i^\top\right) \quad \text{and} \quad \sum_{i=1}^{n} z(i) \leq k,$$

with probability at least $1 - O\left(\frac{(d^2+d/\gamma)^2}{\varepsilon^2 p^2} \cdot e^{-\Omega(\gamma\sqrt{d})}\right) - e^{-\Omega(\varepsilon d/\gamma)}$*.*

▶ Remark. Theorem 2 can be generalized to all real $p \geq 1$, see Remark 11 for more details. The $p = 0$ case can also be covered by Theorem 2, but with a different analysis from [6].

This is the first approximation algorithm for Φ_p-design for general p. Theorem 2 shows that Φ_p-design for constant p admits as good an approximation algorithm as for D/A-design, and there is a unifying algorithm to achieve this guarantee.

Note that, when $p \to +\infty$, Φ_p becomes the E-design objective and $\gamma = \max\{\varepsilon, 1/p\} = \varepsilon$. Thus, Theorem 2 provides a nice interpolation between the d/ε bound for D/A-design and the d/ε^2 bound for E-design.

We further remark that our results can be generalized to the weighted setting to handle multiple budget/knapsack constraints as in [6], but we omit the details to keep the presentation cleaner as they are the same as in [6].

The proof of Theorem 2 is built on the randomized local search approach in [7] and [6], but several new technical ideas are needed to handle higher moments that are introduced by the higher p-norm. In Section 2, we will review the background and previous work, present the algorithm and the overall structure of the analysis, and explain the new ideas in this work. Then, in Section 3, we will present the details of the Φ_p experimental design.

1.2 Discussions and Future Directions

Our proof of Theorem 2 does not address the range $p \in [0,1)$. However, with a similar analysis as in [6], exactly the same algorithm in this paper can achieve a $(1+\varepsilon)$-approximation for $p = 0$ (i.e., D-design) when $k \gtrsim d/\varepsilon$. It would be interesting to see whether the randomized local search approach can be extended to Φ_p-design with $p \in (0,1)$. An obstacle to the current analysis is that the Lieb-Thirring inequality used in the proof of Lemma 10 goes in the wrong direction for $p \in (0,1)$.

If we plot the minimum required k for achieving the $(1+\varepsilon)$-approximation for the Φ_p-design as a function of p, Theorem 2 has ruled out a sharp transition of the curve at $p = 1$, from d/ε to d/ε^2. However, we do not know whether Theorem 2 is tight or not, in particular, in the range of $p \geq 1/\varepsilon$ (see Figure 1 for a demonstration). It would be interesting to fully characterize the whole curve.

1.3 Related Work

In this paper, we focus on generalizing the D/A/E-design with the p-norm objective in (1). However, there are other different ways to extend D/A/E-design. For example, the Bayesian framework of experimental design [2] extends the problem by adding a fixed matrix B (which encodes some prior information, e.g., a multiple of identity) to the covariance matrix before applying the objective function. When $B = 0$, we recover classical experimental design problems. Tantipongpipat [14] provided an approximation algorithm for Bayesian A-design

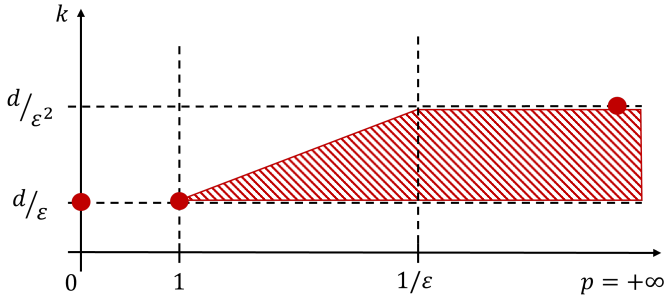

Figure 1 Known results on the minimum k required to achieve $(1+\varepsilon)$-approximation for the Φ_p-design (ignoring constant factors). The exact curve should lie within the shadowed area.

problem when B is a multiple of identity. Dereziński et. al. [3] studied approximation algorithms for Bayesian A/D-design (together with C/V-design, i.e., two other objectives) with a general PSD matrix B. Another example is using elementary symmetric polynomials to generalize A/D-design [10]. Nikolov et. al. [11] extended their approximation algorithm for A-design to tackle this family of generalized ratio objectives.

All these are interesting generalizations of D/A/E-design. It would be interesting to see whether the randomized local search approach in [7, 6] and this paper can be applied to these settings to obtain better results.

2 The Framework

In this section, we first review the randomized local search approach in Section 2.1. Then, we present the full algorithm in Section 2.2 and state the main technical theorem. Then, in Section 2.3, we provide the overall proof plan and the precise statements for analyzing the randomized exchange algorithm, and then put together the statements to prove the main technical theorem. Finally, we discuss the main difficulty and the new ideas needed to analyze the Φ_p objective in Section 3.

2.1 Randomized Local Search Approach

The proof of Theorem 2 is built on the randomized local search approach in [6], which is based on the regret minimization framework developed in [1] for experimental design problems and the randomized spectral rounding techniques in [7]. We review this approach in this subsection.

In [6], the first step is to solve (2) for D/A/E-design to obtain a solution $x \in \mathbb{R}^n$, and then to normalize the vectors v_i's so that $\sum_{i=1}^n x(i) \cdot v_i v_i^\top = I$ for using the regret minimization framework in [1]. Then, the rounding algorithm starts from a random initial solution set S_0 that is independently sampled according to x. Using the density matrix A_t maintained by the regret minimization framework at each step t, the algorithm randomly chooses a pair of vectors v_{i_t} and v_{j_t} with the following probability distributions:

$$\Pr(i_t = i) \propto \big(1 - x(i)\big) \cdot \Big(1 - \alpha \langle v_i v_i^\top, A_t^{1/2} \rangle\Big) \quad \text{and} \quad \Pr(j_t = j) \propto x(j) \cdot \Big(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle\Big), \tag{3}$$

and set $S_{t+1} \leftarrow S_t - v_{i_t} + v_{j_t}$. Using the above randomized local search strategy, it can be shown that the size of the solution is expected to stay around k, while the potential function from the regret minimization framework [1] related to the minimum eigenvalue is expected to improve as long as the minimum eigenvalue is less than $1 - \varepsilon$. Freedman's martingale inequality and a new concentration inequality for non-martingales are used in [7] to prove that all these quantities are close to their expected values with high probability. This randomized local search algorithm provides a $(1 \pm \varepsilon)$-approximate solution for D/A/E-design when $k \gtrsim d/\varepsilon^2$.

The main contribution in [6] is to prove that the randomized local search algorithm can be adapted to achieve $(1 \pm \varepsilon)$-approximation for D/A-design when $k \gtrsim d/\varepsilon$, thus providing a unifying approach to achieve the optimal integrality gap for D/A/E-design. Essentially, the algorithm is the same as the one for E-design but only require that the solution to have minimum eigenvalue $3/4$ rather than $1 - \varepsilon$. For the analysis, the randomized local search algorithm is conceptually divided into two phases. In the first phase, the algorithm will find a solution with minimum eigenvalue at least $3/4$ with high probability when $k \gtrsim d/\varepsilon$. In the second phase, the minimum eigenvalue will maintain to be at least $1/4$ with high probability, and the objective value for D/A-design will improve to $(1 \pm \varepsilon)$ times the optimal objective value in polynomial time with high probability. The analysis of the first phase follows directly from earlier work on spectral rounding in [7]. The analysis of the second phase includes two main parts: (1) to show that, in expectation, the probability distributions in (3) for E-design are also good for making progress towards D/A-design objectives; (2) to show that the progress in the objective value is concentrated around the expectation with a martingale concentration argument. The condition that the minimum eigenvalue is at least $1/4$ is very important in both parts in the second phase, and the optimality conditions for the convex program (2) is crucially used in the second part in the second phase.

In this paper, we will extend the algorithm and the analysis described above and show that the randomized local search algorithm provides a unifying approach for Φ_p-design for all p.

2.2 The Algorithm

We present the full algorithm for Φ_p-design in this subsection.

Algorithm 1 Randomized Exchange Algorithm.

Input: m vectors $u_1, ..., u_m \in \mathbb{R}^d$, a budget $k \geq d$, an accuracy parameter $\varepsilon \in (0, 1)$.

1. Solve the convex programming relaxation (2) and obtain an optimal solution $x \in [0, 1]^m$ with at most $d^2 + 1$ fractional entries, i.e. $|\{i \in [m] \mid 0 < x(i) < 1\}| \leq d^2 + 1$. Let $X = \sum_{i=1}^m x(i) \cdot u_i u_i^\top$.
2. Preprocessing: Let $v_i \leftarrow X^{-1/2} u_i$ for all $i \in [m]$, so that $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$.
3. Initialization: $t \leftarrow 1$, $S_0 \leftarrow \emptyset$, $\gamma = \max\{\frac{\varepsilon}{6}, \frac{1}{6p}\}$, $\kappa = \max\{\frac{\varepsilon}{2}, \frac{1}{2p}\}$, $M \leftarrow \frac{d}{\gamma} + d^2 + 1$, and $\alpha \leftarrow \frac{\sqrt{d}}{\gamma}$.
4. Add i into S_0 independently with probability $x(i)$ for each $i \in [m]$. Let $Y_1 \leftarrow \sum_{i \in S_0} u_i u_i^\top$ and $Z_1 \leftarrow \sum_{i \in S_0} v_i v_i^\top$.
5. While the termination condition $(\mathrm{tr}(Y_t^{-p}))^{\frac{1}{p}} \leq (1 + \varepsilon)(\mathrm{tr}(X^{-p}))^{\frac{1}{p}}$ is not satisfied and $t \leq \frac{2M}{\gamma} + \frac{2M}{\varepsilon p}$, do the following:
   a. $S_t \leftarrow \mathrm{Exchange}(S_{t-1})$.
   b. Set $Y_{t+1} \leftarrow \sum_{i \in S_t} u_i u_i^\top$, $Z_{t+1} \leftarrow \sum_{i \in S_t} v_i v_i^\top$ and $t \leftarrow t + 1$.
6. Return S_{t-1} as the solution.

The main algorithm is almost the same as the one in [6], but with two additional parameters γ, κ that will depend on the value of p. For A-design (when $p = 1$), the algorithm in [6] is just a special case of the randomized exchange algorithm with parameter $\gamma = \frac{1}{8}$ and $\kappa = 1$. In this paper, the parameter γ is used to adjust the learning rate α of the regret minimization framework and the parameter κ will be used in the exchange subroutine.

Algorithm 2 Exchange Subroutine.

Input: the current solution set S_{t-1}.

1. Compute the action matrix $A_t := (\alpha Z_t - c_t I)^{-2}$, where $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ and c_t is the unique scalar such that $A_t \succ 0$ and $\mathrm{tr}(A_t) = 1$.
2. Define $S'_t := \left\{ i \in S_{t-1} \mid \alpha \langle v_i v_i^\top, A_t^{1/2} \rangle \le \frac{1}{2} \text{ and } \langle v_i v_i^\top, Z_t^{-1} \rangle \le \kappa \right\}$.
3. Sample $j_t \in [m] \backslash S_{t-1}$ from the following probability distribution

$$\Pr(j_t = j) = \frac{x(j)}{M} \cdot \left(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle\right), \text{ for } j \in [m] \backslash S_{t-1} \text{ and}$$
$$\Pr(j_t = \emptyset) = 1 - \sum_{j \in [m] \backslash S_{t-1}} \frac{x(j)}{M} \cdot \left(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle\right).$$

4. Sample $i_t \in S'_{t-1}$ from the following probability distribution

$$\Pr(i_t = i) = \frac{1 - x(i)}{M} \cdot \left(1 - \alpha \langle v_i v_i^\top, A_t^{1/2} \rangle\right), \text{ for } i \in S'_{t-1} \text{ and}$$
$$\Pr(i_t = \emptyset) = 1 - \sum_{i \in S'_{t-1}} \frac{1 - x(i)}{M} \cdot \left(1 - \alpha \langle v_i v_i^\top, A_t^{1/2} \rangle\right).$$

5. Return $S_t := S_{t-1} \cup \{j_t\} \backslash \{i_t\}$.

The exchange subroutine is also almost the same as in [6]. The key difference is to use the new parameter κ to further restrict the set of vectors that are allowed to remove from the current solution.

▶ Remark 3. Using the new analysis in this paper, the same algorithm in [6] (without changing the parameters γ and κ) can achieve $(1+\varepsilon)$-approximation for Φ_p-design when $k \gtrsim 2^{O(p)} d/\varepsilon$, with an exponential dependence on p in the budget requirement, much worse than $k \gtrsim \min\{dp/\varepsilon, d/\varepsilon^2\}$ in Theorem 2.

Main Technical Theorem

To prove Theorem 2, we will prove that the randomized exchange algorithm is a bicriteria approximation algorithm, such that it returns a solution that is $(1+\varepsilon)$-approximate in the Φ_p objective and the size of the solution is not much larger than k.

▶ **Theorem 4.** *Given $\varepsilon \in (0, 1)$, if $k \gtrsim \frac{d}{\gamma\varepsilon}$, then the randomized exchange algorithm returns a solution set S within $\frac{2M}{\gamma} + \frac{2M}{\varepsilon p}$ iterations such that*

$$\left(\mathrm{tr} \left(\left(\sum_{i \in S} u_i u_i^\top \right)^{-p} \right) \right)^{\frac{1}{p}} \le (1+\varepsilon) \cdot \left(\mathrm{tr}\left(X^{-p} \right) \right)^{1/p}$$

with probability at least $1 - O\left(\frac{M^2}{\varepsilon^2 p^2} \cdot e^{-\Omega(\gamma\sqrt{d})}\right)$, where X is an optimal fractional solution to (2). Moreover, the solution set S satisfies $|S| \le (1+\varepsilon)k + O\left(\frac{d}{\gamma} + \frac{d}{\kappa}\right)$ with probability at least $1 - e^{-\Omega(\varepsilon d/\min\{\gamma,\kappa\})}$.

With Theorem 4, we can prove Theorem 2 by first noticing that both γ and κ are chosen in the order of $\Theta(\max\{\varepsilon, 1/p\})$ and then turning the bicriteria approximation result into a true approximation with a scaling argument as in [6]. We omit the standard proof and refer the readers to [6] for more details.

2.3 The Proof Plan

In this subsection, we provide the overall plan and the precise statements for analyzing the randomized exchange algorithm, and then put together the statements to obtain the main technical theorem.

2.3.1 Well-Defined Algorithm

First, we prove that the randomized exchange algorithm is well-defined. In particular, we need to show that a fractional optimal solution to the convex relaxation (2) with at most $O(d^2)$ fractional entries can be found in polynomial time, and also the probability distributions in the exchange subroutine are well-defined for $M = d^2 + \frac{d}{\gamma} + 1$. These can be established using the same arguments (with a different value for γ) as in Lemma 4.2 and Claim 4.4 in [6]. Note that the modified exchange subroutine does not affect these arguments.

The following simple observation (Observation 4.3 in [6]) will be useful in the analysis of the algorithm.

▶ **Observation 5.** *For any $t \geq 0$, it holds that $i \in S_t$ for all i with $\mathsf{x}(i) = 1$ and $j \in [n] \backslash S_t$ for all j with $\mathsf{x}(j) = 0$. This further implies that $\Pr(i_t = i) = 0$ for all i with $\mathsf{x}(i) \in \{0, 1\}$ and $\Pr(j_t = j) = 0$ for all j with $\mathsf{x}(j) \in \{0, 1\}$.*

2.3.2 Solution Size Bound

Then, we show that the algorithm returns a solution set S of size not much larger than k with high probability.

▶ **Theorem 6** (Variant of Theorem 3.12 of [7])**.** *Let $\alpha = \sqrt{d}/\gamma$ and κ be the parameters used in the randomized exchange algorithm. Suppose that the solution S_t of the randomized exchange algorithm satisfies $\lambda_{\min}\left(\sum_{i \in S_t} \mathsf{v}_i \mathsf{v}_i^\top\right) < 1$ for all $t \in [\tau]$. Then, for any given $\delta \in [0, 1]$,*

$$\Pr\left[|S_\tau| \leq (1+\delta) \cdot \sum_{i=1}^{n} \mathsf{x}(i) + \Big(\frac{12d}{\gamma} + \frac{2d}{\kappa}\Big)\right] \geq 1 - \exp\Big(-\Omega\Big(\frac{\delta d}{\min\{\gamma, \kappa\}}\Big)\Big).$$

When $\kappa = 1$, the above theorem follows directly from the one-sided spectral rounding result in [7]. With smaller κ, we are restricting the set of vectors that can be swapped out from the current solution. This would increase the chance of not removing a vector, and thus increasing the size of the solution. Fortunately, we can show that the increase of the solution size can be bounded by an additive d/κ term. The proof idea is similar to the one in [7], and the main difference is a modified bound on the expected change of size. Please refer to the appendices of the full arxiv version [5] for more details.

2.3.3 Approximation Guarantee

The most technical part of the proof is to establish the approximation guarantee. We follow the analysis in [6] to conceptually divide the execution of the algorithm into two phases as described in Section 2.1.

In the first phase, we show that the minimum eigenvalue will reach $1 - 2\gamma$ in $O(M/\gamma)$ iterations with high probability, which follows from the spectral rounding result in [7].

▶ **Proposition 7** (Counterpart of Proposition 4.5 in [6]). *The probability that the randomized exchange algorithm has terminated successfully within $2M/\gamma$ iterations or there exists $\tau_1 \leq 2M/\gamma$ with $\lambda_{\min}(Z_{\tau_1}) \geq 1 - 2\gamma$ is at least $1 - \exp(-\Omega(\sqrt{d}))$.*

In the second phase, the minimum eigenvalue will be at least $1 - 5\gamma$ during the next $\Theta\left(\frac{M}{\varepsilon p}\right)$ iterations with good probability.

▶ **Proposition 8** (Counterpart of Proposition 4.6 in [6]). *Suppose $\lambda_{\min}(Z_{\tau_1}) \geq 1 - 2\gamma$ for some τ_1. In the randomized exchange algorithm, the probability that $\lambda_{\min}(Z_t) \geq 1 - 5\gamma$ for all $\tau_1 \leq t \leq \tau_1 + \frac{2M}{\varepsilon p}$ is at least $1 - \frac{4M^2}{\varepsilon^2 p^2} \cdot e^{-\Omega(\gamma\sqrt{d})}$.*

Both of the proofs of Proposition 7 and Proposition 8 follow same arguments in [6] (with a new parameter γ). We remark that the modified exchange subroutine with a restricted S'_t does not affect these arguments, as removing less vectors only helps to improve the minimum eigenvalue of the solution. We omit the proofs and refer the readers to [6].

The main technical contribution in this paper is to prove that the Φ_p objective will improve to at most $(1 + \varepsilon)$ times the optimal value during the second phase when the minimum eigenvalue is at least $1 - 5\gamma$.

▶ **Theorem 9.** *Given $\varepsilon \in (0, 1)$, if $p \leq 1/\varepsilon$ and $k \gtrsim \frac{pd}{\varepsilon}$ for some $\varepsilon \in (0, 1)$, then the probability that the following three events happen simultaneously during the execution of the randomized exchange algorithm is at most $\exp(-\Omega(\gamma\sqrt{d}))$.*

1. $\lambda_{\min}(Z_{\tau_1}) \geq 1 - 2\gamma$ *for some* τ_1;
2. $\lambda_{\min}(Z_t) \geq 1 - 5\gamma$ *for all* $\tau_1 \leq t \leq \tau_1 + \frac{2M}{\varepsilon p}$;
3. *the randomized exchange algorithm has not terminated by time* $\tau_1 + \frac{2M}{\varepsilon p}$.

2.3.4 Proof of Theorem 4

We put together the statements in this subsection to obtain Theorem 4.

Proof of Theorem 4. We start with analyzing the approximation guarantee in the theorem.

Firstly, consider the easier case $p \geq 1/\varepsilon$, which implies that $\gamma = \varepsilon/6$. By Proposition 7, there exists $\tau_1 \leq 2M/\gamma$ such that $\lambda_{\min}(Z_{\tau_1}) \geq (1 - \varepsilon/3)$ with probability $1 - \exp(-\Omega(\sqrt{d}))$. We note that $\lambda_{\min}(Z_{\tau_1}) \geq (1 - \varepsilon/3)$ is equivalent to $Y_{\tau_1} \succcurlyeq (1 - \varepsilon/3)X$, which is sufficient to establish

$$(\operatorname{tr}(Y_{\tau_1}^{-p}))^{1/p} \leq (1 + \varepsilon)(\operatorname{tr}(X^{-p}))^{1/p},$$

i.e., the approximation guarantee in the theorem. We remark that we do not need the assumption $k \gtrsim d/(\gamma\varepsilon)$ in the proof of this case.

Then, we consider the case of $p \leq 1/\varepsilon$ and define the bad events for the randomized exchange algorithm:

- B_1: the algorithm has not terminated successfully within $2M/\gamma$ iterations and $\tau_1 > 2M/\gamma$ where τ_1 is the first time such that $\lambda_{\min}(Z_{\tau_1}) \geq 1 - 2\gamma$.
- B_2: there exists some $\tau_1 \leq t \leq \tau_1 + \frac{2M}{\varepsilon p}$ such that $\lambda_{\min}(Z_t) < 1 - 5\gamma$.
- B_3: the termination condition $(\operatorname{tr}(Y_t^{-p}))^{\frac{1}{p}} \leq (1 + \varepsilon)(\operatorname{tr}(X^{-p}))^{\frac{1}{p}}$ is not satisfied for all $\tau_1 \leq t \leq \tau_1 + \frac{2M}{\varepsilon p}$.

If none of the bad events happens, then either the algorithm has terminated successfully within $2M/\gamma$ iterations or the termination condition will be satisfied at some time $t \leq \tau_1 + \frac{2M}{\varepsilon p} \leq \frac{2M}{\gamma} + \frac{2M}{\varepsilon p}$. So, the probability that the randomized exchange algorithm has not satisfied the termination condition within $\frac{2M}{\gamma} + \frac{2M}{\varepsilon p}$ iterations is upper bounded by

$$\begin{aligned}\Pr[B_1 \cup B_2 \cup B_3] &= \Pr[B_1] + \Pr[B_2 \cap \neg B_1] + \Pr[B_3 \cap \neg B_2 \cap \neg B_1] \\ &\leq O\left(e^{-\Omega(\sqrt{d})}\right) + O\left(\frac{M^2}{\varepsilon^2 p^2} \cdot e^{-\Omega(\gamma\sqrt{d})}\right) + O\left(e^{-\Omega(\gamma\sqrt{d})}\right) \\ &\leq O\left(\frac{M^2}{\varepsilon^2 p^2} \cdot e^{-\Omega(\gamma\sqrt{d})}\right),\end{aligned}$$

where $\Pr[B_1]$ is bounded in Proposition 7, $\Pr[B_2 \cap \neg B_1]$ is bounded in Proposition 8, and $\Pr[B_3 \cap \neg B_2 \cap \neg B_1]$ is bounded in Theorem 9 (note that we need the assumption $p \leq 1/\varepsilon$ and $k \gtrsim pd/\varepsilon$ here). The termination condition implies the approximation guarantee directly.

Finally, we consider the size of the returned solution. Note that if $\lambda_{\min}(Z_t) \geq 1$ then $Y_t \succcurlyeq X$, which further implies that the termination condition is met at time t. Hence, we can assume $\lambda_{\min}(Z_t) < 1$ before the algorithm terminates. Therefore, we can apply Theorem 6 to conclude that the returned solution S satisfies $|S| \leq (1+\varepsilon)k + O\big(\frac{d}{\gamma} + \frac{d}{\kappa}\big)$ with probability at least $1 - \exp(-\Omega(\frac{\varepsilon d}{\min\{\gamma,\kappa\}}))$. ◀

2.4 New Ideas

The key in proving Theorem 9 is to bound the change of the objective value after an exchange. For A-design ($p = 1$), there is a simple inequality bounding the change of the objective as

$$\operatorname{tr}\left(\left(Y - vv^\top + ww^\top\right)^{-1}\right) \leq \operatorname{tr}\left(Y^{-1}\right) + \underbrace{\frac{v^\top Y^{-2} v}{1 - \langle vv^\top, Y^{-1}\rangle} - \frac{w^\top Y^{-2} w}{1 + \langle ww^\top, Y^{-1}\rangle}}_{\text{progress}}.$$

For general Φ_p-design, the change of the Φ_p function under rank-two updates is considerably more complicated. Using Sherman-Morrison formula and Lieb-Thirring inequality, we can bound the change of the Φ_p objective (in fact, the p-th power of the Φ_p objective) as follows:

$$\begin{aligned}&\operatorname{tr}\left((Y + ww^\top - vv^\top)^{-p}\right) \\ &\leq \operatorname{tr}(Y^{-p}) + \sum_{i=1}^{p} \binom{p}{i} \left((-1)^i \frac{(w^\top Y^{-1} w)^{i-1} \cdot w^\top Y^{-p-1} w}{(1 + w^\top Y^{-1} w)^i} + \frac{(v^\top Y^{-1} v)^{i-1} \cdot v^\top Y^{-p-1} v}{(1 - v^\top Y^{-1} v)^i}\right).\end{aligned}$$

There are many higher order terms introduced by the p-norm, and dealing with these is the main technical difficulty in this paper.

As discussed in Remark 3, if we use the same algorithm in [6], with some careful manipulations including applying Hölder's inequality appropriately, we can achieve $(1+\varepsilon)$-approximation but with the much worst requirement that $k \gtrsim 2^{O(p)} d/\varepsilon$. The reason is that removing some "influential" vectors (even with relatively small probability) from the current solution will blow up the expectation of the change of the objective function due to the higher order terms in the above inequality.

To overcome this issue, we introduce the parameter κ and modify the randomized exchange algorithm by restricting those vectors (in S_t') that are allowed to swap out of the current solution. This helps us to effectively bound those higher order terms in the above inequality about the change of the objective function. But, with smaller κ, we are restricting the set of vectors that can be swapped out from the current solution. This would increase the chance of not removing a vector, and thus increasing the size of the solution. Fortunately, we can show that the increase of the solution size can be bounded by an additive d/κ term as described in Theorem 6.

The analysis for A-design in [6] contains two parts: (1) bound the expected progress; (2) prove the concentration of the total progress. The condition that the minimum eigenvalue is at least 1/4 is very important in both parts, and the optimality conditions for the convex program (2) is crucially used in the concentration argument. Interestingly, the optimality condition of the convex program (2) is also crucial in bounding the expected progress for Φ_p objective with higher p. Much effort in this paper is used to get the expectations of the objective value right, while it was relatively easy for D/A-design (when $p = 0, 1$).

3 The Analysis of Φ_p Objective

The goal of this section is to prove Theorem 9. Since we are focusing on the case of $p \le 1/\varepsilon$ in Theorem 9, we can assume without loss of generality that

$$\gamma = \max\left\{\frac{\varepsilon}{6}, \frac{1}{6p}\right\} = \frac{1}{6p} \qquad \text{and} \qquad \kappa = \max\left\{\frac{\varepsilon}{2}, \frac{1}{2p}\right\} = \frac{1}{2p}$$

in the remaining of this section. For ease of notation, we also assume the start time of the second phase is $\tau_1 = 1$. To analyze progress of the algorithm in terms of the objective function, we upper bound the change of $\operatorname{tr}(Y_t^{-p})$ after a swap using the following lemma, for which we will provide a proof in Section 3.1.

▶ **Lemma 10.** *Let $Y \succ 0$ be a d-dimensional positive definite matrix and $p \ge 1$ be an integer. For any $w \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ such that $v^\top Y^{-1} v < 1$,*

$$\begin{aligned}&\operatorname{tr}\left((Y + ww^\top - vv^\top)^{-p}\right)\\ &\quad \le \operatorname{tr}(Y^{-p}) + \sum_{i=1}^{p} \binom{p}{i} \left((-1)^i \frac{(w^\top Y^{-1} w)^{i-1} \cdot w^\top Y^{-p-1} w}{(1 + w^\top Y^{-1} w)^i} + \frac{(v^\top Y^{-1} v)^{i-1} \cdot v^\top Y^{-p-1} v}{(1 - v^\top Y^{-1} v)^i} \right).\end{aligned}$$

In the randomized exchange algorithm, we swap vectors u_{i_t} and u_{j_t} in each iteration where u_{i_t} is in the current solution. Thus, $u_{i_t}^\top Y_t^{-1} u_{i_t} \le 1$ always holds, and in fact it will be clear later that $u_{i_t}^\top Y_t^{-1} u_{i_t}$ is strictly less than 1. Hence, Lemma 10 can be applied repeatedly to obtain that, for any $\tau \ge 1$,

$$\begin{aligned}\operatorname{tr}(Y_{\tau+1}^{-p}) - \operatorname{tr}(Y_1^{-p}) \le \sum_{t=1}^{\tau} \Bigg(&\underbrace{\sum_{i=1}^{p} \binom{p}{i} \frac{(u_{i_t}^\top Y_t^{-1} u_{i_t})^{i-1} \cdot u_{i_t}^\top Y_t^{-p-1} u_{i_t}}{(1 - u_{i_t}^\top Y_t^{-1} u_{i_t})^i}}_{\text{loss}} \\ &- \underbrace{\sum_{i=1}^{p} \binom{p}{i} (-1)^{i+1} \frac{(u_{j_t}^\top Y_t^{-1} u_{j_t})^{i-1} \cdot u_{j_t}^\top Y_t^{-p-1} u_{j_t}}{(1 + u_{j_t}^\top Y_t^{-1} u_{j_t})^i}}_{\text{gain}} \Bigg).\end{aligned} \tag{4}$$

We define gain g_t, loss l_t, and progress Γ_t in the t-th iteration as follows

$$\begin{aligned}g_t &:= \sum_{i=1}^{p} \binom{p}{i} (-1)^{i+1} \frac{(u_{j_t}^\top Y_t^{-1} u_{j_t})^{i-1} \cdot u_{j_t}^\top Y_t^{-p-1} u_{j_t}}{(1 + u_{j_t}^\top Y_t^{-1} u_{j_t})^i},\\ l_t &:= \sum_{i=1}^{p} \binom{p}{i} \frac{(u_{i_t}^\top Y_t^{-1} u_{i_t})^{i-1} \cdot u_{i_t}^\top Y_t^{-p-1} u_{i_t}}{(1 - u_{i_t}^\top Y_t^{-1} u_{i_t})^i},\\ \Gamma_t &:= g_t - l_t.\end{aligned}$$

In Section 3.2, we will prove that if x is a fractional optimal solution to (2) and Z_t has lower-bounded minimum eigenvalue and the objective value of the current solution is far from optimal, then the expected progress in the t-th iteration is large. Then, in Section 3.3, we will prove that the total progress is concentrated around its expectation. Finally, we complete the proof of Theorem 9 in Section 3.4.

3.1 Change of Objective Value in One Step

In [6], a rank-two update formula is used to compute the change of the objective value in one step when $p = 1$. For general Φ_p-design, the rank-two update formula becomes considerably more complicated, and instead we do the update in two smaller steps: We first use a rank-one update to add u_{j_t} to the current solution, then use another rank-one update to remove u_{i_t} from the current solution.

Proof of Lemma 10. Let $Y_1 = Y + ww^\top$. By Sherman-Morrison formula [12], it holds that

$$\operatorname{tr}(Y_1^{-p}) = \operatorname{tr}\left(\left(Y^{-1} - \frac{Y^{-1}ww^\top Y^{-1}}{1 + w^\top Y^{-1} w}\right)^p\right) = \operatorname{tr}\left(\left(Y^{-1/2}\left(I - \frac{Y^{-1/2}ww^\top Y^{-1/2}}{1 + w^\top Y^{-1} w}\right)Y^{-1/2}\right)^p\right).$$

Then, we can apply Lieb-Thirring inequality [8] to show that

$$\operatorname{tr}(Y_1^{-p}) \le \operatorname{tr}\left(Y^{-p/2}\left(I - \frac{Y^{-1/2}ww^\top Y^{-1/2}}{1 + w^\top Y^{-1} w}\right)^p Y^{-p/2}\right) = \operatorname{tr}\left(Y^{-p}\left(I - \frac{Y^{-1/2}ww^\top Y^{-1/2}}{1 + w^\top Y^{-1} w}\right)^p\right).$$

Expanding by the binomial theorem,

$$\begin{aligned} \operatorname{tr}(Y_1^{-p}) &\le \sum_{i=0}^{p} (-1)^i \binom{p}{i} \operatorname{tr}\left(Y^{-p}\left(\frac{Y^{-1/2}ww^\top Y^{-1/2}}{1 + w^\top Y^{-1} w}\right)^i\right) \\ &= \operatorname{tr}(Y^{-p}) + \sum_{i=1}^{p} (-1)^i \binom{p}{i} \frac{(w^\top Y^{-1} w)^{i-1} \cdot w^\top Y^{-p-1} w}{(1 + w^\top Y^{-1} w)^i}. \end{aligned} \tag{5}$$

For $Y_2 = Y_1 - vv^\top$, we can apply similar argument to show that

$$\operatorname{tr}(Y_2^{-p}) \le \operatorname{tr}(Y_1^{-p}) + \sum_{i=1}^{p} \binom{p}{i} \frac{(v^\top Y_1^{-1} v)^{i-1} \cdot v^\top Y_1^{-p-1} v}{(1 - v^\top Y_1^{-1} v)^i}.$$

Notice that $Y_1 = Y + ww^\top \succcurlyeq Y$ and $v^\top Y^{-1} v < 1$, thus it holds that

$$\operatorname{tr}(Y_2^{-p}) \le \operatorname{tr}(Y_1^{-p}) + \sum_{i=1}^{p} \binom{p}{i} \frac{(v^\top Y^{-1} v)^{i-1} \cdot v^\top Y^{-p-1} v}{(1 - v^\top Y^{-1} v)^i}. \tag{6}$$

The lemma follows by combining (5) and (6). ◀

▶ Remark 11. Lemma 10 can be generalized to all real $p \ge 1$ by invoking Newton's generalized binomial theorem in the proof. To guarantee the convergence of the generalized binomial theorem, we need to ensure a stronger condition $v^\top Y^{-1} v \le \frac{1}{2}$. This is not an issue for our application, as our algorithm always removes vectors from the restricted set S_t', which guarantees that $v^\top Y^{-1} v \le \frac{1}{2}$ is satisfied. Given the new version of Lemma 10 with real p, we can generalize the main result in this paper (i.e., Theorem 2) to all real $p \ge 1$ with essentially the same analysis.

3.2 Expected Progress

To analyze the expected progress, we need to use the following two lemmas. The first one is an implication of the lower-bounded minimum eigenvalue condition, which is an analog of Lemma 4.13 in [6], and we refer the readers to the appendices of the arxiv version of this paper [5] for a proof.

▶ **Lemma 12.** *For $\gamma \leq \frac{1}{6}$, if $Z_t \succcurlyeq (1-5\gamma)I$, then*

$$\langle v_i v_i^\top, Z_t^{-1}\rangle \leq \alpha \langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \leq \alpha \lambda_{\min}(Z_t)\langle v_i v_i^\top, Z_t^{-1}\rangle \qquad \forall i \in [n].$$

The other one is an implication of the optimality condition of (2). The proof of the lemma is similar to the one in [6] for A-design, see the appendices of the full arxiv version [5] for more details.

▶ **Lemma 13.** *Let $x \in [0,1]^n$ be an optimal fractional solution of the convex programming relaxation (2) for the p-norm problem. Then, for each $1 \leq i \leq n$ with $0 < x(i) < 1$,*

$$\langle X^{-p-1}, u_i u_i^\top \rangle \leq \frac{1}{k} \cdot \mathrm{tr}(X^{-p}).$$

Now, we are ready to lower bound the expected progress. We will first handle the expected loss and expected gain separately in Lemma 14 and Lemma 15. Then, combine the two parts to lower bound the expected progress in Lemma 17. For simplicity, we denote $\mathbb{E}_t[\cdot]$ as the conditional expectation given what had happened up to time t, that is, $\mathbb{E}[\cdot \mid S_{t-1}]$.

3.2.1 Expected Loss

The minimum eigenvalue lower bound (Lemma 12), the optimality condition (Lemma 13), and the introduction of the new parameter κ in the randomized exchange algorithm are all crucial in the following lemma.

▶ **Lemma 14** (Expected Loss). *Let S_{t-1} be the solution set at time t and $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ for $1 \leq t \leq \tau$. Suppose x is an optimal solution of (2), $\lambda_{\min}(Z_t) \in [1-5\gamma, 1)$, $\gamma = 1/6p$, and $\kappa = 1/2p$. Then*

$$\mathbb{E}_t[l_t] \leq \frac{p}{M}\Big(\mathrm{tr}(Y_t^{-p}) - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big) + O\Big(\frac{p^2 d}{kM}\Big) \cdot \mathrm{tr}(X^{-p}),$$

where we denote $X_S := \sum_{i \in S} x(i) u_i u_i^\top$ for any set $S \subseteq [n]$.

Proof. There are p terms in the loss term l_t. We deal with the linear term and higher order terms separately. Consider the linear term:

$$\begin{aligned}\mathbb{E}_t\left[\frac{p \cdot u_{i_t}^\top Y_t^{-p-1} u_{i_t}}{1 - u_{i_t}^\top Y_t^{-1} u_{i_t}}\right] &= \sum_{i \in S'_{t-1}} \frac{1-x(i)}{M}\Big(1 - \alpha\langle v_i v_i^\top, A_t^{1/2}\rangle\Big) \cdot \frac{p \cdot u_i^\top Y_t^{-p-1} u_i}{1 - u_i^\top Y_t^{-1} u_i} \\ &= \sum_{i \in S'_{t-1}} \frac{1-x(i)}{M}\Big(1 - \alpha\langle v_i v_i^\top, A_t^{1/2}\rangle\Big) \cdot \frac{p \cdot u_i^\top Y_t^{-p-1} u_i}{1 - \langle v_i v_i^\top, Z_t^{-1}\rangle},\end{aligned}$$

where the second line follows by the definitions of Y_t and Z_t, which implies that

$$\langle v_i v_i^\top, Z_t^{-1}\rangle = u_i^\top Y_t^{-1} u_i. \tag{7}$$

Note that $\gamma = 1/6p \le 1/6$. Thus, we can apply the first inequality in Lemma 12 and then relax S'_{t-1} to S_{t-1} to obtain that

$$\mathbb{E}_t\left[\frac{p \cdot u_{i_t}^\top Y_t^{-p-1} u_{i_t}}{1 - u_{i_t}^\top Y_t^{-1} u_{i_t}}\right] \le \frac{p}{M} \sum_{i \in S_{t-1}} (1 - x(i)) \cdot u_i^\top Y_t^{-p-1} u_i$$
$$= \frac{p}{M}\Big(\operatorname{tr}(Y_t^{-p}) - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big). \tag{8}$$

Then, we consider the remaining $p-1$ higher order loss terms.

$$\underbrace{\mathbb{E}_t\left[\sum_{l=2}^{p}\binom{p}{l}\frac{(u_{i_t}^\top Y_t^{-1} u_{i_t})^{l-1} \cdot u_{i_t}^\top Y_t^{-p-1} u_{i_t}}{(1 - u_{i_t}^\top Y_t^{-1} u_{i_t})^l}\right]}_{①}$$
$$= \sum_{i \in S'_{t-1}} \frac{1 - x(i)}{M}\Big(1 - \alpha\langle v_i v_i^\top, A_t^{1/2}\rangle\Big)\sum_{l=2}^{p}\binom{p}{l}\frac{(u_i^\top Y_t^{-1} u_i)^{l-1} \cdot u_i^\top Y_t^{-p-1} u_i}{(1 - u_i^\top Y_t^{-1} u_i)^l}.$$

Again, by applying Lemma 12, it holds that

$$① \le \sum_{i \in S'_{t-1}} \frac{1 - x(i)}{M} \cdot u_i^\top Y_t^{-p-1} u_i \cdot \sum_{l=2}^{p}\binom{p}{l}\frac{(u_i^\top Y_t^{-1} u_i)^{l-1}}{(1 - u_i^\top Y_t^{-1} u_i)^{l-1}}.$$

Notice that

$$\lambda_{\min}(Z_t) \ge 1 - 5\gamma \qquad \Longleftrightarrow \qquad Y_t \succcurlyeq (1 - 5\gamma) X. \tag{9}$$

Thus, by the assumption $\lambda_{\min}(Z_t) \ge 1 - 5\gamma$, it follows that

$$① \le \frac{(1 - 5\gamma)^{-p-1}}{M} \sum_{i \in S'_{t-1}} (1 - x(i)) \cdot u_i^\top X^{-p-1} u_i \cdot \sum_{l=2}^{p}\binom{p}{l}\frac{(u_i^\top Y_t^{-1} u_i)^{l-1}}{(1 - u_i^\top Y_t^{-1} u_i)^{l-1}}.$$

Using the fact that x is a fractional optimal solution to (2) and then applying Lemma 13, it holds that

$$① \le \frac{(1 - 5\gamma)^{-p-1}}{kM} \cdot \operatorname{tr}(X^{-p}) \sum_{i \in S'_{t-1}} (1 - x(i)) \cdot \sum_{l=2}^{p}\binom{p}{l}\frac{(u_i^\top Y_t^{-1} u_i)^{l-1}}{(1 - u_i^\top Y_t^{-1} u_i)^{l-1}}$$
$$\le \frac{(1 - 5\gamma)^{-p-1}}{kM} \cdot \operatorname{tr}(X^{-p}) \sum_{l=2}^{p}\binom{p}{l}\sum_{i \in S'_{t-1}} u_i^\top Y_t^{-1} u_i \cdot \frac{(u_i^\top Y_t^{-1} u_i)^{l-2}}{(1 - u_i^\top Y_t^{-1} u_i)^{l-1}}.$$

Due to the definition of the set S'_{t-1} and (7), it holds that $u_i^\top Y_t^{-1} u_i \le \kappa$ for all $i \in S'_{t-1}$. Thus,

$$① \le \frac{(1 - 5\gamma)^{-p-1}}{kM} \cdot \operatorname{tr}(X^{-p}) \sum_{l=2}^{p}\binom{p}{l}\sum_{i \in S'_{t-1}} u_i^\top Y_t^{-1} u_i \cdot \left(\frac{\kappa}{1-\kappa}\right)^{l-2} \cdot \frac{1}{1-\kappa}.$$

Since $\sum_{i \in S'_{t-1}} u_i^\top Y_t^{-1} u_i \le \langle Y_t, Y_t^{-1}\rangle = d$, we can further upper bound the ① term by

$$① \le \frac{(1 - 5\gamma)^{-p-1} d}{kM} \cdot \operatorname{tr}(X^{-p}) \cdot \sum_{l=2}^{p}\binom{p}{l}\left(\frac{\kappa}{1-\kappa}\right)^{l} \cdot \frac{1-\kappa}{\kappa^2}$$
$$= \frac{(1 - 5\gamma)^{-p-1} d}{kM} \cdot \operatorname{tr}(X^{-p}) \cdot \frac{1-\kappa}{\kappa^2}\left(\frac{1}{(1-\kappa)^p} - 1 - \frac{p\kappa}{1-\kappa}\right)$$
$$\le \frac{(1 - 5\gamma)^{-p-1} d}{kM} \cdot \operatorname{tr}(X^{-p}) \cdot \frac{1}{\kappa^2} \cdot \frac{1}{(1-\kappa)^p}. \tag{10}$$

Combining (8) and (10), the expected loss can be bounded by

$$\begin{aligned}\mathbb{E}_t[l_t] &\leq \frac{p}{M}\Big(\operatorname{tr}(Y_t^{-p}) - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big) + \frac{(1-5\gamma)^{-p-1}d}{\kappa^2(1-\kappa)^p kM} \cdot \operatorname{tr}(X^{-p}) \\ &\leq \frac{p}{M}\Big(\operatorname{tr}(Y_t^{-p}) - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big) + O\Big(\frac{p^2 d}{kM}\Big) \cdot \operatorname{tr}(X^{-p}),\end{aligned}$$

where the last inequality follows by the choice of γ and κ. ◀

3.2.2 Expected Gain

The analysis of the expected gain is slightly more complicated since the sampling probability does not cancel out with the denominator term in g_t as nicely as in the analysis of loss l_t, and so we will divide into two cases and use different arguments. Again, the lower bound of the minimum eigenvalue (Lemma 12) and the optimality condition (Lemma 13) are crucial in the following lemma, the proof of which is deferred to Appendix A due to the space limit.

▶ **Lemma 15** (Expected Gain). *Let S_{t-1} be the solution set at time t and $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ for $1 \leq t \leq \tau$. Suppose x is an optimal solution to (2), $\lambda_{\min}(Z_t) \in [1-5\gamma, 1)$, and $\gamma = 1/6p$. Then*

$$\mathbb{E}_t[g_t] \geq \frac{p}{M}\Big(\langle X, Y_t^{-p-1}\rangle - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big) - O\Big(\frac{p^2 d}{kM}\Big) \cdot \operatorname{tr}(X^{-p}).$$

3.2.3 Expected Progress

Finally, we apply Hölder's inequality appropriately to compare the gain term and loss term.

▶ **Lemma 16.** *Given positive definite matrices $A, B \in \mathbb{S}_{++}^d$ and an integer $p \geq 1$, it holds that*

$$\langle A, B^{-p-1}\rangle \geq \left(\frac{\operatorname{tr}(B^{-p})}{\operatorname{tr}(A^{-p})}\right)^{1/p} \cdot \operatorname{tr}(B^{-p}).$$

Proof. Let $A = \sum_{i=1}^d a_i v_i v_i^\top$ be the eigendecomposition of A, and $B = \sum_{i=1}^d b_i w_i w_i^\top$ be the eigendecomposition of B. Then,

$$\begin{aligned}\operatorname{tr}(B^{-p}) = \sum_{i=1}^d \frac{1}{b_i^p} &= \sum_{1 \leq i,j \leq d} \frac{1}{b_i^p}\langle v_i, w_j\rangle^2 \\ &= \sum_{1 \leq i,j \leq d} \frac{a_j^{p/(p+1)}}{b_i^p}\langle v_i, w_j\rangle^{2p/(p+1)} \cdot \frac{1}{a_j^{p/(p+1)}}\langle v_i, w_j\rangle^{2/(p+1)} \\ &\leq \Big(\sum_{1 \leq i,j \leq d} \frac{a_j}{b_i^{p+1}}\langle v_i, w_j\rangle^2\Big)^{p/(p+1)} \cdot \Big(\sum_{1 \leq i,j \leq d} \frac{1}{a_j^p}\langle v_i, w_j\rangle^2\Big)^{1/(p+1)} \\ &= \langle A, B^{-p-1}\rangle^{p/(p+1)} \cdot \operatorname{tr}(A^{-p})^{1/(p+1)},\end{aligned}$$

where the inequality follows by Hölder's inequality. Then, the lemma follows by taking the $(p+1)/p$'s power of both sides and rearranging the terms. ◀

Now, we lower bound the expected progress by combining Lemma 14 and Lemma 15.

▶ **Lemma 17.** *Let S_{t-1} be the solution set at time t and $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ for $1 \le t \le \tau$. Suppose x is an fractional optimal solution to (2), and $\lambda_{\min}(Z_t) \in [1-5\gamma, 1)$, $(\operatorname{tr}(Y_t^{-2}))^{1/p} \ge \lambda \cdot (\operatorname{tr}\left(X^{-2}\right))^{1/p}$ for some $\lambda \ge 1$ and for $1 \le t \le \tau$. Then*

$$\mathbb{E}_t[\Gamma_t] \ge \left(\frac{p(\lambda-1)\lambda^p}{M} - O\Big(\frac{p^2 d}{kM}\Big)\right) \cdot \operatorname{tr}(X^{-p}).$$

In particular, when $\lambda > 1+\varepsilon$ and $k \gtrsim \frac{pd}{\varepsilon}$, the expected progress is positive.

Proof. Combining the expected loss Lemma 14 and the expected gain Lemma 15, it follows that

$$\mathbb{E}_t[\Gamma_t] \ge \frac{p}{M}\Big(\langle X, Y_t^{-p-1}\rangle - \operatorname{tr}(Y_t^{-p})\Big) - O\Big(\frac{p^2 d}{kM}\Big) \cdot \operatorname{tr}(X^{-p}).$$

Applying Lemma 16, we derive that

$$\mathbb{E}_t[\Gamma_t] \ge \frac{p}{M}\left(\left(\frac{\operatorname{tr}(Y_t^{-p})}{\operatorname{tr}(X^{-p})}\right)^{1/p} \cdot \operatorname{tr}(Y_t^{-p}) - \operatorname{tr}(Y_t^{-p})\right) - O\Big(\frac{p^2 d}{kM}\Big) \cdot \operatorname{tr}(X^{-p}).$$

By the assumption that $(\operatorname{tr}(Y_t^{-p}))^{1/p} \ge \lambda(\operatorname{tr}(X^{-p}))^{1/p}$, or equivalently $\operatorname{tr}(Y_t^{-p}) \ge \lambda^p \operatorname{tr}(X^{-p})$, we arrive at the final bound that

$$\mathbb{E}_t[\Gamma_t] \ge \frac{p(\lambda-1)}{M} \cdot \operatorname{tr}(Y_t^{-p}) - O\Big(\frac{p^2 d}{kM}\Big) \cdot \operatorname{tr}(X^{-p}) \ge \left(\frac{p(\lambda-1)\lambda^p}{M} - O\Big(\frac{p^2 d}{kM}\Big)\right) \cdot \operatorname{tr}(X^{-p}). \blacktriangleleft$$

3.3 Martingale Concentration Argument

In this subsection, we prove that the total progress is concentrated around the expectation. The proof uses the minimum eigenvalue assumption and the optimality characterization in Lemma 13 to bound the variance of the random process. The proof idea is similar to the one in [6], but we need some additional efforts to take care of the higher order terms that are introduced by higher p-norm. We defer the detailed proof to Appendix A.

▶ **Lemma 18.** *For any $\eta > 0$, it holds that*

$$\Pr\left[\sum_{t=1}^{\tau} \Gamma_t \le \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] - \eta \bigcap \min_{1 \le t \le \tau} \lambda_{\min}(Z_t) \ge 1-5\gamma\right] \\ \le \exp\left(-\Omega\left(\frac{\eta^2 kM}{\tau p^3 \sqrt{d}(\operatorname{tr}(X^{-p}))^2 + \eta p M \operatorname{tr}(X^{-p})}\right)\right).$$

3.4 Proof of Theorem 9

We are ready to prove Theorem 9. Let $\tau = \frac{2M}{\varepsilon p}$. We want to upper bound the probability that the following three events happen simultaneously:

- E_1: The randomized exchange algorithm entered the second phase, i.e., $\lambda_{\min}(Z_1) \ge 1-2\gamma$ using the notation in this subsection.
- E_2: $\min_{1 \le t \le \tau} \lambda_{\min}(Z_t) \ge 1-5\gamma$.
- E_3: The second phase of the algorithm has not terminated by time τ.

Suppose the event E_3 happens. Then $\lambda = \min_{1 \le t \le \tau+1} (\operatorname{tr}(Y_t^{-p}))^{1/p} / (\operatorname{tr}(X^{-p}))^{1/p} > (1+\varepsilon)$. If the event E_2 also happens, then Lemma 17 implies that

$$\begin{aligned} \sum_{t=1}^{\tau} \mathbb{E}_t\left[\Gamma_t\right] &\ge \tau \cdot \left(\frac{p(\lambda-1)\lambda^p}{M} - O\Big(\frac{p^2 d}{kM}\Big)\right) \cdot \operatorname{tr}(X^{-p}) \\ &\ge \frac{2M}{\varepsilon p} \cdot \left(\frac{\varepsilon p}{M} - \frac{\varepsilon p}{2M}\right) \cdot \operatorname{tr}(X^{-p}) \ge \operatorname{tr}\left(X^{-p}\right), \end{aligned} \tag{11}$$

where the second inequality holds for $k \gtrsim pd/\varepsilon$ with large enough constant.

On the other hand, the initial solution of the second phase satisfies $Z_1 \succcurlyeq (1-2\gamma)I$ (follows by E_1), which implies that $Y_1 \succcurlyeq (1-2\gamma)X$. Thus, $\mathrm{tr}(Y_1^{-p}) \le (1-2\gamma)^{-p}\,\mathrm{tr}(X^{-p}) \le (1-1/3p)^{-p}\,\mathrm{tr}(X^{-p}) \le 3\,\mathrm{tr}(X^{-p})/2$, for $p \ge 1$. When the event E_2 happens, we know from Lemma 12 that $\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle \le \alpha \cdot \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}}\rangle < \frac{1}{2}$, and so we can apply (4) to deduce that

$$\mathrm{tr}(Y_{\tau+1}^{-p}) \le \mathrm{tr}(Y_1^{-p}) - \sum_{t=1}^{\tau} \Gamma_t \le \frac{3}{2} \cdot \mathrm{tr}\left(X^{-p}\right) - \sum_{t=1}^{\tau} \Gamma_t.$$

Since the algorithm has not terminated by time τ,

$$\mathrm{tr}(X^{-p}) \le \mathrm{tr}(Y_{\tau+1}^{-p}) \quad \Longrightarrow \quad \sum_{t=1}^{\tau} \Gamma_t \le \frac{1}{2} \cdot \mathrm{tr}\left(X^{-p}\right). \tag{12}$$

Combining (11) and (12), $E_1 \cap E_2 \cap E_3$ implies a large deviation of the progress from the expectation such that

$$\sum_{t=1}^{\tau} \Gamma_t - \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] < -\frac{1}{2} \cdot \mathrm{tr}(X^{-p}).$$

Thus, we can apply Lemma 18 with $\eta = \frac{1}{2} \cdot \mathrm{tr}(X^{-p})$ and $\tau = \frac{2M}{\varepsilon p}$ to conclude that

$$\begin{aligned}
\Pr\left[E_1 \cap E_2 \cap E_3\right] &\le \Pr\left[\sum_{t=1}^{\tau} \Gamma_t < \sum_{t=1}^{\tau} \mathbb{E}_t\left[\Gamma_t\right] - \frac{1}{2} \cdot \mathrm{tr}(X^{-p}) \bigcap E_2\right] \\
&\le \exp\left(-\Omega\left(\frac{(\mathrm{tr}(X^{-p}))^2 \cdot kM}{\left(\frac{2M}{\varepsilon p}\right) p^3 \sqrt{d} \cdot (\mathrm{tr}(X^{-p}))^2 + pM \cdot (\mathrm{tr}(X^{-p}))^2} \right)\right) \\
&\le \exp\left(-\Omega\left(\frac{\varepsilon k}{p^2 \sqrt{d}} \right)\right) \le \exp\left(-\Omega\left(\frac{\sqrt{d}}{p} \right)\right) = \exp\left(-\Omega\left(\gamma \sqrt{d} \right)\right),
\end{aligned}$$

where the last inequality holds by the assumption $k \gtrsim pd/\varepsilon$, and the last equality follows as $\gamma = 1/6p$.

References

1. Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: a regret minimization approach. *Math. Program.*, 186(1-2, Ser. A):439–478, 2021. `doi:10.1007/s10107-019-01464-2`.
2. Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: a review. *Statist. Sci.*, 10(3):273–304, 1995.
3. Michal Derezinski, Feynman T. Liang, and Michael W. Mahoney. Bayesian experimental design using regularized determinantal point processes. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 3197–3207. PMLR, 2020. URL: `http://proceedings.mlr.press/v108/derezinski20a.html`.
4. David A. Freedman. On tail probabilities for martingales. *Ann. Probability*, 3:100–118, 1975. `doi:10.1214/aop/1176996452`.
5. Lap Chi Lau, Robert Wang, and Hong Zhou. Experimental design for any p-norm. *CoRR*, abs/2305.01942, 2020. `arXiv:2305.01942`.
6. Lap Chi Lau and Hong Zhou. A local search framework for experimental design. *SIAM J. Comput.*, 51(4):900–951, 2022. `doi:10.1137/20M1386542`.

7 Lap Chi Lau and Hong Zhou. A spectral approach to network design. *SIAM J. Comput.*, 51(4):1018–1064, 2022. `doi:10.1137/20M1330762`.

8 Elliott H. Lieb and Walter E. Thirring. *Inequalities for the moments of the eigenvalues of the Schrödinger Hamiltonian and their relation to Sobolev inequalities*, pages 269–304. Princeton University Press, 1976.

9 Vivek Madan, Mohit Singh, Uthaipon Tantipongpipat, and Weijun Xie. Combinatorial algorithms for optimal design. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2210–2258, Phoenix, USA, 25–28 June 2019. PMLR.

10 Zelda Mariet and Suvrit Sra. Elementary symmetric polynomials for optimal experimental design. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2136–2145, Red Hook, NY, USA, 2017. Curran Associates Inc.

11 Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 1369–1386, Philadelphia, PA, USA, 2019. Society for Industrial and Applied Mathematics.

12 Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statistics*, 21:124–127, 1950. `doi:10.1214/aoms/1177729893`.

13 Mohit Singh and Weijun Xie. Approximate positive correlated distributions and approximation algorithms for D-optimal design. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2240–2255, USA, 2018. Society for Industrial and Applied Mathematics.

14 Uthaipon Tantipongpipat. λ-regularized A-optimal design and its approximation by λ-regularized proportional volume sampling. *CoRR*, abs/2006.11182, 2020. `arXiv:2006.11182`.

15 Joel A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011. `doi:10.1214/ECP.v16-1624`.

16 Richard L. Wheeden and Antoni Zygmund. *Measure and integral.* Pure and Applied Mathematics (Boca Raton). CRC Press, Boca Raton, FL, second edition, 2015. An introduction to real analysis.

A Omitted Proofs in Section 3

Proof of Lemma 15 (Expected Gain). We consider two separate cases:

$$S_1 := \{j \in [m]\backslash S_{t-1} \mid \alpha\langle v_j v_j^\top, A_t^{1/2}\rangle \geq p \cdot u_j^\top Y_t^{-1} u_j\} \quad \text{and}$$
$$S_2 := \{j \in [m]\backslash S_{t-1} \mid \alpha\langle v_j v_j^\top, A_t^{1/2}\rangle < p \cdot u_j^\top Y_t^{-1} u_j\}.$$

Then, the expected gain can be written as

$$\mathbb{E}_t[g_t] = \sum_{j \in S_1} \frac{x(j)}{M} \cdot \left(1 + \alpha\langle v_j v_j^\top, A_t^{1/2}\rangle\right) \cdot \sum_{l=1}^{p} \binom{p}{l} (-1)^{l+1} \frac{(u_j^\top Y_t^{-1} u_j)^{l-1} \cdot u_j^\top Y_t^{-p-1} u_j}{(1 + u_j^\top Y_t^{-1} u_j)^l} \tag{13}$$
$$+ \sum_{j \in S_2} \frac{x(j)}{M} \cdot \left(1 + \alpha\langle v_j v_j^\top, A_t^{1/2}\rangle\right) \cdot \sum_{l=1}^{p} \binom{p}{l} (-1)^{l+1} \frac{(u_j^\top Y_t^{-1} u_j)^{l-1} \cdot u_j^\top Y_t^{-p-1} u_j}{(1 + u_j^\top Y_t^{-1} u_j)^l}. \tag{14}$$

We analyze (13) and (14) separately. First, we consider (13) and rearrange the g_t term a bit, which will be easier for the analysis of (13).

$$
\begin{aligned}
g_t &= \frac{u_{j_t}^\top Y_t^{-p-1} u_{j_t}}{u_{j_t}^\top Y_t^{-1} u_{j_t}} \cdot \sum_{i=1}^{p} \binom{p}{i} (-1)^{i+1} \left(\frac{u_{j_t}^\top Y_t^{-1} u_{j_t}}{1 + u_{j_t}^\top Y_t^{-1} u_{j_t}} \right)^i \\
&= \frac{u_{j_t}^\top Y_t^{-p-1} u_{j_t}}{u_{j_t}^\top Y_t^{-1} u_{j_t}} \left(1 - \left(1 - \frac{u_{j_t}^\top Y_t^{-1} u_{j_t}}{1 + u_{j_t}^\top Y_t^{-1} u_{j_t}} \right)^p \right) \qquad (15) \\
&= \frac{u_{j_t}^\top Y_t^{-p-1} u_{j_t}}{u_{j_t}^\top Y_t^{-1} u_{j_t}} \left(1 - \frac{1}{(1 + u_{j_t}^\top Y_t^{-1} u_{j_t})^p} \right). \qquad (16)
\end{aligned}
$$

Thus, we can rewrite (13) as

$$
(13) = \sum_{j \in S_1} \frac{x(j)}{M} \cdot \left(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle \right) \cdot \frac{u_j^\top Y_t^{-p-1} u_j}{u_j^\top Y_t^{-1} u_j} \left(1 - \frac{1}{(1 + u_j^\top Y_t^{-1} u_j)^p} \right).
$$

By the definition of S_1, it holds that

$$
(13) \geq \sum_{j \in S_1} \frac{x(j)}{M} \cdot u_j^\top Y_t^{-p-1} u_j \cdot \frac{1 + p \cdot u_j^\top Y_t^{-1} u_j}{u_j^\top Y_t^{-1} u_j} \left(1 - \frac{1}{(1 + u_j^\top Y_t^{-1} u_j)^p} \right).
$$

Let $x = u_j^\top Y_t^{-1} u_j > 0$ (as $Y_t \succ 0$). Then, it holds that

$$
\frac{1 + px}{x} \cdot \left(1 - \frac{1}{(1+x)^p} \right) = \frac{1 + px}{x} \cdot \frac{(1+x)^p - 1}{(1+x)^p} = \frac{px(1+x)^p + (1+x)^p - 1 - px}{x(1+x)^p} \geq p.
$$

Thus,

$$
(13) \geq \frac{p}{M} \sum_{j \in S_1} x(j) \cdot u_j^\top Y_t^{-p-1} u_j. \qquad (17)
$$

Then, we consider (14). As in the proof of Lemma 14, we separate (14) into two parts, ② concerning the linear term and ③ concerning the remaining $p-1$ higher order terms.

$$
\begin{aligned}
(14) = &\underbrace{\sum_{j \in S_2} \frac{x(j)}{M} \cdot \left(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle \right) \cdot \frac{p \cdot u_j^\top Y_t^{-p-1} u_j}{1 + u_j^\top Y_t^{-1} u_j}}_{②} \\
&- \underbrace{\sum_{j \in S_2} \frac{x(j)}{M} \cdot \left(1 + \alpha \langle v_j v_j^\top, A_t^{1/2} \rangle \right) \cdot \sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1} \cdot u_j^\top Y_t^{-p-1} u_j}{(1 + u_j^\top Y_t^{-1} u_j)^l}}_{③}.
\end{aligned}
$$

The linear term ② is easy to bound, we can control it by Lemma 12 (combined with (7)) so that

$$
② \geq \frac{p}{M} \sum_{j \in S_2} x(j) \cdot u_j^\top Y_t^{-p-1} u_j. \qquad (18)
$$

Then, we upper bound the higher order terms ③ (notice that ③ does not contain the minus sign). To upper bound ③, for each $j \in S_2$, we can assume $\sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1} \cdot u_j^\top Y_t^{-p-1} u_j}{(1 + u_j^\top Y_t^{-1} u_j)^l} \geq 0$ without loss of generality, as otherwise we can simply ignore the j-th term. With this assumption and by the definition of S_2, it follows that

$$
\begin{aligned}
③ &\leq \frac{1}{M} \sum_{j \in S_2} x(j)(1 + p u_j^\top Y_t^{-1} u_j) \sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1} \cdot u_j^\top Y_t^{-p-1} u_j}{(1 + u_j^\top Y_t^{-1} u_j)^l} \\
&= \frac{1}{M} \sum_{j \in S_2} x(j) u_j^\top Y_t^{-p-1} u_j \cdot (1 + p u_j^\top Y_t^{-1} u_j) \sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1}}{(1 + u_j^\top Y_t^{-1} u_j)^l}.
\end{aligned}
$$

Using the assumption $\lambda_{\min}(Z_t) \geq 1 - 5\gamma$ and (9), it holds that

$$
③ \leq \frac{(1-5\gamma)^{-p-1}}{M} \sum_{j \in S_2} x(j) u_j^\top X^{-p-1} u_j \cdot (1 + p u_j^\top Y_t^{-1} u_j) \sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1}}{(1 + u_j^\top Y_t^{-1} . u_j)^l}.
$$

Since x is an optimal solution to (2), we can apply Lemma 13 and derive that

$$
\begin{aligned}
③ &\leq \frac{(1-5\gamma)^{-p-1}}{kM} \cdot \operatorname{tr}(X^{-p}) \sum_{j \in S_2} x(j) \cdot (1 + p u_j^\top Y_t^{-1} u_j) \sum_{l=2}^{p} \binom{p}{l} (-1)^l \frac{(u_j^\top Y_t^{-1} u_j)^{l-1}}{(1 + u_j^\top Y_t^{-1} u_j)^l} \\
&= \frac{(1-5\gamma)^{-p-1}}{kM} \cdot \operatorname{tr}(X^{-p}) \cdot \\
&\qquad \sum_{j \in S_2} x(j) \cdot \frac{1 + p u_j^\top Y_t^{-1} u_j}{u_j^\top Y_t^{-1} u_j} \left(\left(1 - \frac{u_j^\top Y_t^{-1} u_j}{1 + u_j^\top Y_t^{-1} u_j}\right)^p - 1 + \frac{p u_j^\top Y_t^{-1} u_j}{1 + u_j^\top Y_t^{-1} u_j} \right).
\end{aligned}
$$

Let $x = u_j^\top Y_t^{-1} u_j$, we want to upper bound

$$
\begin{aligned}
&\frac{1+px}{x} \left(\left(1 - \frac{x}{1+x}\right)^p - 1 + \frac{px}{1+x} \right) = \\
&\qquad \frac{1}{x} \left(\left(1 - \frac{x}{1+x}\right)^p - 1 + \frac{px}{1+x} \right) + p \left(\left(1 - \frac{x}{1+x}\right)^p - 1 + \frac{px}{1+x} \right).
\end{aligned}
$$

For any $y \in [0,1]$, it holds that $(1-y)^p \leq 1 - py + \binom{p}{2} y^2$. Thus, it follows that

$$
\begin{aligned}
&\frac{1}{x} \left(\left(1 - \frac{x}{1+x}\right)^p - 1 + \frac{px}{1+x} \right) + p \left(\left(1 - \frac{x}{1+x}\right)^p - 1 + \frac{px}{1+x} \right) \\
&\leq \frac{1}{x} \cdot \binom{p}{2} \cdot \frac{x^2}{(1+x)^2} + p \cdot \frac{px}{1+x} = \binom{p}{2} \cdot \frac{x}{1+x} + \frac{p^2 x}{1+x} \leq 2p^2 x.
\end{aligned}
$$

Therefore, the ③ term can be further bounded by

$$
\begin{aligned}
③ &\leq \frac{2(1-5\gamma)^{-p-1} \cdot p^2}{kM} \cdot \operatorname{tr}(X^{-p}) \sum_{j \in S_2} x(j) \cdot u_j^\top Y_t^{-1} u_j \\
&\leq \frac{2(1-5\gamma)^{-p-1} \cdot p^2}{kM} \cdot \operatorname{tr}(X^{-p}) \langle X, Y_t^{-1} \rangle.
\end{aligned}
$$

Using (9), it holds that

$$
③ \leq \frac{2(1-5\gamma)^{-p-2} \cdot p^2 d}{kM} \cdot \operatorname{tr}(X^{-p}) \leq O\left(\frac{p^2 d}{kM}\right) \cdot \operatorname{tr}(X^{-p}), \tag{19}
$$

where the last inequality follows by the choice of γ.

Combining (17), (18) and (19), we can lower bound the expected gain by

$$
\begin{aligned}
\mathbb{E}_t[g_t] &\geq \frac{p}{M} \left(\sum_{j \in [m] \setminus S_{t-1}} x(j) u_j^\top Y_t^{-p-1} u_j \right) - O\left(\frac{p^2 d}{kM}\right) \cdot \operatorname{tr}(X^{-p}) \\
&= \frac{p}{M} \left(\langle X, Y_t^{-p-1} \rangle - \langle X_{S_{t-1}}, Y_t^{-p-1} \rangle \right) - O\left(\frac{p^2 d}{kM}\right) \cdot \operatorname{tr}(X^{-p}).
\end{aligned}
$$

◀

Proof of Lemma 18. We define two sequences of random variables $\{X_t\}_t$ and $\{Y_t\}_t$, where $X_t := \mathbb{E}_t[\Gamma_t] - \Gamma_t$ and $Y_t := \sum_{l=1}^{t} X_l$. It is easy to check that $\{Y_t\}_t$ is a martingale with respect to $\{S_t\}_t$. We will use Freedman's inequality (see, e.g., [4, 15]) to bound the probability $\Pr[Y_\tau \geq \eta \cap \min_{1\leq t\leq \tau} \lambda_{\min}(Z_t) \geq 1-5\gamma]$.

In the following, we first show that if the event $\min_{1\leq t\leq\tau}\lambda_{\min}(Z_t) \geq 1-5\gamma$ happens, then we can upper bound X_t and $\mathbb{E}_t[X_t^2]$ so that we can apply Freedman's inequality. To upper bound X_t, we first prove an upper bound on g_t and l_t.

Note that, if the event $\lambda_{\min}(Z_t) \geq 1-5\gamma$ happens, then $Y_t \succeq (1-5\gamma)X$, which implies

$$u_{i_t}^\top Y_t^{-p-1} u_{i_t} \leq (1-5\gamma)^{-p-1} u_{i_t}^\top X^{-p-1} u_{i_t} \quad \text{and} \quad u_{j_t}^\top Y_t^{-p-1} u_{j_t} \leq (1-5\gamma)^{-p-1} u_{j_t}^\top X^{-p-1} u_{j_t}.$$

By Observation 5, for all the exchange pairs i_t, j_t, it holds that $x(i_t), x(j_t) \in (0,1)$. Thus, we can apply Lemma 13 to show that $\langle X^{-p-1}, u_{i_t}u_{i_t}^\top\rangle \leq \frac{1}{k}\cdot \operatorname{tr}(X^{-p})$ and $\langle X^{-p-1}, u_{j_t}u_{j_t}^\top\rangle \leq \frac{1}{k}\cdot\operatorname{tr}(X^{-p})$. Therefore, since $\gamma = 1/6p$,

$$u_{i_t}^\top Y_t^{-p-1} u_{i_t} \leq O\Big(\frac{1}{k}\Big)\cdot \operatorname{tr}(X^{-p}) \quad \text{and} \quad u_{j_t}^\top Y_t^{-p-1} u_{j_t} \leq O\Big(\frac{1}{k}\Big)\cdot\operatorname{tr}(X^{-p}). \tag{20}$$

We first give a deterministic bound on g_t. Let $x = u_{j_t}^\top Y_t^{-p-1} u_{j_t}$, according to (15), the gain term g_t can be written as

$$g_t = u_{j_t}^\top Y_t^{-p-1} u_{j_t} \cdot \frac{1}{x}\cdot\left(1 - \Big(1-\frac{x}{1+x}\Big)^p\right).$$

Since $(1-y)^p \geq 1-py$ for $y\in[0,1]$ and $p\geq 1$, we can bound g_t by

$$0 \leq g_t \leq u_{j_t}^\top Y_t^{-p-1} u_{j_t}\cdot\frac{1}{x}\cdot\frac{px}{1+x} \leq p\cdot u_{j_t}^\top Y_t^{-p-1}u_{j_t} \leq O\Big(\frac{p}{k}\Big)\cdot\operatorname{tr}(X^{-p}),$$

where the last inequality follows from (20).

Then, we give an deterministic bound on l_t. By the definition of S'_{t-1} and (7), it holds that $0 < u_{i_t}^\top Y_t^{-1} u_{i_t} = \langle v_{i_t}v_{i_t}^\top, Z_t^{-1}\rangle \leq \kappa$. Thus, we can bound l_t by

$$\begin{aligned} 0 \leq l_t = \sum_{i=1}^{p}\binom{p}{i}\frac{(u_{i_t}^\top Y_t^{-1}u_{i_t})^{i-1}\cdot u_{i_t}^\top Y_t^{-p-1}u_{i_t}}{(1-u_{i_t}^\top Y_t^{-1}u_{i_t})^i} &\leq u_{i_t}^\top Y_t^{-p-1}u_{i_t}\cdot\sum_{i=1}^{p}\binom{p}{i}\frac{\kappa^{i-1}}{(1-\kappa)^i} \\ &= u_{i_t}^\top Y_t^{-p-1}u_{i_t}\cdot\frac{1}{\kappa}\cdot\left(\frac{1}{(1-\kappa)^p}-1\right). \end{aligned}$$

For $\kappa = \frac{1}{2p}$, we can control l_t such that

$$l_t \leq O(p)\cdot u_{i_t}^\top Y_t^{-p-1}u_{i_t} \leq O\Big(\frac{p}{k}\Big)\cdot\operatorname{tr}(X^{-p}),$$

where the last inequality follows from (20).

With the above bounds on g_t and l_t, we can control the size of the martingale increment by

$$|X_t| = |\mathbb{E}_t[\Gamma_t]-\Gamma_t| \leq g_t + l_t \leq O\Big(\frac{p}{k}\Big)\cdot\operatorname{tr}(X^{-p}).$$

Next, we upper bound $\mathbb{E}_t[X_t^2]$ by

$$\mathbb{E}_t[X_t^2] \leq |X_t|\cdot\mathbb{E}_t[|X_t|] \leq O\Big(\frac{p}{k}\Big)\cdot\operatorname{tr}(X^{-p})\cdot\Big(\mathbb{E}_t[g_t]+\mathbb{E}_t[l_t]\Big).$$

Using Lemma 14, we bound the expected loss term by

$$\begin{aligned}\mathbb{E}_t[l_t] &\le \frac{p}{M}\Big(\operatorname{tr}(Y_t^{-p}) - \langle X_{S_{t-1}}, Y_t^{-p-1}\rangle\Big) + O\Big(\frac{p^2 d}{kM}\Big)\cdot \operatorname{tr}(X^{-p})\\ &\le O\Big(\frac{p}{M}\Big)\cdot \operatorname{tr}(X^{-p}) + O\Big(\frac{p^2 d}{kM}\Big)\cdot \operatorname{tr}(X^{-p})\\ &\le O\Big(\frac{p}{M}\Big)\cdot \operatorname{tr}(X^{-p}),\end{aligned}$$

where the second inequality follows by the assumption that $\lambda_{\min}(Z_t) \ge 1 - 5\gamma$ happens (and (9)), the last inequality follows by the assumption on k.

Then, with the expression of g_t in (15), we write the expected gain as

$$\mathbb{E}_t[g_t] = \sum_{j\in[m]\setminus S_{t-1}} \frac{x(j)}{M}\cdot\Big(1+\alpha\langle v_j v_j^\top, A_t^{\frac12}\rangle\Big)\cdot\frac{u_j^\top Y_t^{-p-1}u_j}{u_j^\top Y_t^{-1}u_j}\left(1-\left(1-\frac{u_j^\top Y_t^{-1}u_j}{1+u_j^\top Y_t^{-1}u_j}\right)^p\right).$$

Using the fact that $(1-y)^p \ge 1 - py$ for $y \in [0,1]$ and $p \ge 1$, the expected gain can be bounded by

$$\mathbb{E}_t[g_t] \le \sum_{j\in[m]\setminus S_{t-1}} \frac{p}{M}\cdot x(j)\cdot u_j^\top Y_t^{-p-1}u_j\cdot\frac{1+\alpha\langle v_j v_j^\top, A_t^{\frac12}\rangle}{1+u_j^\top Y_t^{-1}u_j}.$$

By (7), $u_j^\top Y_t^{-1}u_j = \langle v_j v_j^\top, Z^{-1}\rangle$. Then, by the second inequality in Lemma 12, it holds that

$$\mathbb{E}_t[g_t] \le \frac{p}{M}\cdot\alpha\lambda_{\min}(Z_t)\cdot\sum_{j=1}^m x(j)\cdot u_j^\top Y_t^{-p-1}u_j \le \frac{p}{M}\cdot\alpha\cdot\langle X, Y_t^{-p-1}\rangle,$$

where the last inequality holds as $\lambda_{\min}(Z_t) < 1$ before the termination of the algorithm. By the assumption that $\lambda_{\min}(Z_t) \ge 1-5\gamma$ happens (and (9)) and the choice of $\alpha = \sqrt{d}/\gamma$ and $\gamma = 1/6p$, we obtain the bound that

$$\mathbb{E}_t[g_t] \le \frac{p\sqrt{d}}{\gamma(1-5\gamma)^{p+1}M}\cdot\operatorname{tr}(X^{-p}) \le O\Big(\frac{p^2\sqrt{d}}{M}\Big)\cdot\operatorname{tr}(X^{-p}).$$

Therefore,

$$\mathbb{E}_t[X_t^2] \le O\Big(\frac{p}{k}\Big)\cdot O\Big(\frac{p^2\sqrt{d}}{M}\Big)\cdot\big(\operatorname{tr}(X^{-p})\big)^2 = O\Big(\frac{p^3\sqrt{d}}{kM}\Big)\cdot\big(\operatorname{tr}(X^{-p})\big)^2,$$

which implies

$$W_t := \sum_{l=1}^t \mathbb{E}_l[X_l^2] \le O\Big(\frac{\tau p^3\sqrt{d}}{kM}\Big)\cdot\big(\operatorname{tr}(X^{-p})\big)^2, \quad \forall t\in[\tau].$$

Finally, we can apply Freedman's martingale inequality Theorem (see, e.g., [4, 15]) with

$$R = O\Big(\frac{p}{k}\Big)\cdot\operatorname{tr}(X^{-p}) \qquad \text{and} \qquad \sigma^2 = O\Big(\frac{p^3\sqrt{d}\tau}{kM}\Big)\cdot\big(\operatorname{tr}(X^{-p})\big)^2$$

to conclude that

$$\begin{aligned}&\Pr\left[Y_\tau \ge \eta \bigcap \min_{1\le t\le\tau}\lambda_{\min}(Z_t) \ge 1-5\gamma\right]\\ \le\ &\Pr[\exists t\in[\tau] : Y_t \ge \eta \cap W_t \le \sigma^2] \le \exp\left(-\frac{\eta^2/2}{\sigma^2 + R\eta/3}\right)\\ =\ &\exp\left(-\Omega\left(\frac{\eta^2 kM}{\tau p^3\sqrt{d}(\operatorname{tr}(X^{-p}))^2 + \eta pM\operatorname{tr}(X^{-p})}\right)\right).\end{aligned}$$

The lemma follows by noting that $\sum_{t=1}^\tau \Gamma_t \le \sum_{t=1}^\tau \mathbb{E}_t[\Gamma_t] - \eta$ is equivalent to $Y_\tau \ge \eta$. ◀