Sublinear-Time Quadratic Minimization via Spectral Decomposition of Matrices

Amit Levi[1]
University of Waterloo, Canada
amit.levi@uwaterloo.ca
https://orcid.org/0000-0002-8530-5182

Yuichi Yoshida[2]
National Institute of Informatics, Tokyo, Japan
yyoshida@nii.ac.jp
https://orcid.org/0000-0001-8919-8479

Abstract

We design a sublinear-time approximation algorithm for quadratic function minimization problems with a better error bound than the previous algorithm by Hayashi and Yoshida (NIPS'16). Our approximation algorithm can be modified to handle the case where the minimization is done over a sphere. The analysis of our algorithms is obtained by combining results from graph limit theory, along with a novel spectral decomposition of matrices. Specifically, we prove that a matrix A can be decomposed into a structured part and a pseudorandom part, where the structured part is a block matrix with a polylogarithmic number of blocks, such that in each block all the entries are the same, and the pseudorandom part has a small spectral norm, achieving better error bound than the existing decomposition theorem of Frieze and Kannan (FOCS'96). As an additional application of the decomposition theorem, we give a sublinear-time approximation algorithm for computing the top singular values of a matrix.

2012 ACM Subject Classification Theory of computation → Sketching and sampling, Theory of computation → Probabilistic computation

Keywords and phrases Qudratic function minimization, Approximation Algorithms, Matrix spectral decomposition, Graph limits

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2018.17

1 Introduction

Quadratic function minimization/maximization is a versatile tool used in machine learning, statistics, and data mining and can represent many fundamental problems such as linear regression, k-means clustering, principal component analysis (PCA), support vector machines, kernel machines and more (see [17]). In general, quadratic function minimization/maximization is NP-Hard. When the problem is convex (for minimization) or concave (for maximization), we can solve it by solving a system of linear equations, which requires $O(n^3)$ time, where n is the number of variables. There are faster approximation methods based on stochastic gradient descent [3], and the multiplicative update algorithm [5]. However, these methods still require $\Omega(n)$ time, which is prohibitive when we need to handle a huge number of variables.

[1] Research supported by NSERC Discovery grant and the David R. Cheriton Graduate Scholarship. Part of this work was done while the author was visiting NII Tokyo.
[2] Research supported by JSPS KAKENHI Grant Number JP17H04676 and JST ERATO Grant Number JPMJER1201.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018).
Editors: Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer; Article No. 17; pp. 17:1–17:19
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Quadratic function minimization over a sphere is also an important problem. This minimization problem is often called the *trust region subproblem* since it must be solved in each step of a *trust region algorithm*. Trust region algorithms are among the most important tools in solving nonlinear programming problems, as they are robust and can be applied to ill-conditioned problems. In addition, trust region subproblems are useful in many other problems such as constrained eigenvalue problems [9], least-square problems [24], combinatorial optimization problems [4] and many more. While the problem is non-convex, it has been shown that the problem exhibit strong duality properties and is known to be solved in polynomial time (see [2, 23]). In particular, it was shown to be equivalent to some semidefinite programming optimization problems that can be solved in polynomial time ([19, 1]). As in the non-constrained case, there are approximation algorithms based on gradient descent [18] and on reducing the problem to a sequence of eigenvalues computations [12]. However, as in the unconstrained case, these methods require running time which is linear in the number of the non-zero elements of the matrix (which might be linear in n).

1.1 Our Contributions

In this work, we provide sublinear-time approximation algorithms for minimizing quadratic functions, assuming random access to the entries of the input matrix and the vector.

First, we consider unconstraind minimization. Specifically, for a matrix $A \in \mathbb{R}^{n\times n}$ and vectors $\boldsymbol{d}, \boldsymbol{b} \in \mathbb{R}^n$, we consider the following quadratic function minimization problem:

$$\min_{\boldsymbol{v}\in\mathbb{R}^n} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}), \text{ where } \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) = \langle \boldsymbol{v}, A\boldsymbol{v}\rangle + n\langle \boldsymbol{v}, \operatorname{diag}(\boldsymbol{d})\boldsymbol{v}\rangle + n\langle \boldsymbol{b}, \boldsymbol{v}\rangle \ . \tag{1}$$

Here $\operatorname{diag}(\boldsymbol{d}) \in \mathbb{R}^{n\times n}$ is a matrix whose diagonal entries are specified by $\boldsymbol{d}$ and $\langle\cdot,\cdot\rangle$ denotes the standard inner product.

▶ **Theorem 1.** *Fix $\epsilon > 0$ and let $\boldsymbol{v}^*$ and z^* be an optimal solution and the optimal value, respectively, of Problem (1). Let S be a random set such that each index $i \in \{1, 2, \ldots, n\}$ is taken to S independently w.p k/n with*

$$k = \max\left\{O\left(\frac{\log^2 n}{\epsilon^2}\right), \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}\right\}.$$

Then, the following holds with probability at least $2/3$: Let $\tilde{\boldsymbol{v}}^*$ and $\tilde{z}^*$ be an optimal solution and the optimal value, respectively, of the problem

$$\min_{\boldsymbol{v}\in\mathbb{R}^{|S|}} \psi_{|S|,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v}) \ ,$$

where $\cdot|_S$ is an operator that extracts a submatrix (or subvector) specified by an index set S. Then,

$$\left|\frac{1}{|S|^2}\tilde{z}^* - \frac{1}{n^2}z^*\right| \le \epsilon L \max\left\{\frac{\|\tilde{\boldsymbol{v}}^*\|_2^2}{|S|}, \frac{\|\boldsymbol{v}^*\|_2^2}{n}\right\} \ ,$$

where $L = \max\{\max_{i,j}|A_{ij}|, \max_i |d_i|, \max_i |b_i|\}$.

Recently, Hayashi and Yoshida [10] proposed a constant-time sampling method for this problem with an additive error of $O(\epsilon L K_\infty^2 n^2)$ for $K_\infty = \max\{\|\boldsymbol{v}^*\|_\infty, \|\tilde{\boldsymbol{v}}^*\|_\infty\}$, where $\boldsymbol{v}^*$ and $\tilde{\boldsymbol{v}}^*$ are the optimal solutions to the original and sampled problems, respectively. Although their algorithm runs in constant time, the guarantee is not meaningful when $K_\infty = \omega(1)$ because the optimal value is always of order $O(Ln^2)$. Theorem 1 shows that we can improve the additive error to $O(\epsilon L K_2^2 n^2)$, where $K_2 = \max\{\|\boldsymbol{v}^*\|_2/\sqrt{n}, \|\tilde{\boldsymbol{v}}^*\|_2/\sqrt{s}\}$, as long as the number of samples s is polylogarithmic (or more). We note that we always have $K_2 \leq K_\infty$ and the difference is significant when $\boldsymbol{v}^*$ and $\tilde{\boldsymbol{v}}^*$ are sparse. For example, if $\boldsymbol{v}^*$ and $\tilde{\boldsymbol{v}}^*$ have only $O(1)$ non-zero elements, then we have $K_2 = O(K_\infty/\sqrt{s})$. Our new bound provides a trade off between the additive error and the time complexity, which was unclear from the argument by Hayashi and Yoshida [10].

Moreover, we consider minimization over a sphere. Specifically, given a matrix $A \in \mathbb{R}^{n\times n}$, vectors $\boldsymbol{b}, \boldsymbol{d} \in \mathbb{R}^n$ and $r > 0$, we consider the following quadratic function minimization problem over a sphere of radius r:

$$\min_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \ . \tag{2}$$

We give the first sublinear-time approximation algorithm for this problem.

▶ **Theorem 2.** *Let $\boldsymbol{v}^*$ and z^* be an optimal solution and optimal value, respectively, of Problem (2). Let $\epsilon > 0$ and let S be a random set such that each index $i \in \{1, 2, \ldots, n\}$ is taken to S independently w.p k/n with*

$$k = \max\left\{O\left(\frac{\log^2 n}{\epsilon^2}\right), \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}\right\}.$$

Then, the following holds with probability at least $2/3$: Let $\tilde{\boldsymbol{v}}^*$ and $\tilde{z}^*$ be an optimal solution and the optimal value, respectively, of the problem

$$\min_{\|v\|_2\leq\sqrt{\frac{|S|}{n}}r} \psi_{|S|,A|_S,d|_S,b|_S}(\boldsymbol{v}) \ .$$

Then,

$$\left|\frac{1}{|S|^2}\tilde{z}^* - \frac{1}{n^2}z^*\right| \leq \frac{\epsilon L r^2}{n} \ ,$$

where $L = \max\{\max_{i,j}|A_{ij}|, \max_i |d_i|, \max_i |b_i|\}$.

We can design a constant-time algorithm for (2) by using the result of [10], but the resulting error bound will be $O(\epsilon L r^2)$, which is n times worse than the bound in Theorem 2.

The proofs of Theorems 1 and 2 rely on a novel decomposition theorem of matrices, which will be discussed in Section 1.3. As another application of this decomposition theorem, we show that for any (small) t, we can approximate the t-th largest singular values of a matrix $A \in [-L, L]^{n\times m}$ (denoted $\sigma_t(A)$) to within an additive error of $O(L\sqrt{\epsilon t n m})$ in time

$$\max\left\{O\left(\frac{\max\{\log^2 n, \log^2 m\}}{\epsilon^2}\right), \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}\right\}.$$

Our algorithms are very simple to implement, and do not require any structure in the input matrix. However, similar results (with better running time) can be obtained by applying known sampling techniques from [8]. Formally, we prove the following.

▶ **Theorem 3.** *Given a matrix $A \in [-L, L]^{n\times m}$, $\epsilon \in (0, 1)$, let*

$$k = \max\left\{O\left(\frac{\max\{\log^2 n, \log^2 m\}}{\epsilon^2}\right), \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}\right\}.$$

Then, for every $t = O(k)$, there is an algorithm that runs in $\mathrm{poly}(k)$ time, and outputs a value z such that with probability at least $2/3$,

$$|\sigma_t(A) - z| \leq L\sqrt{\epsilon t n m}.$$

We note that since the $\sigma_t(A) \leq L\sqrt{\frac{nm}{t}}$ (see Fact 5), the *relative* error the algorithm achieves is at least $\sqrt{\epsilon} \cdot t$. Therefore, to get meaningful approximation, one must have that $\sqrt{\epsilon} \cdot t < 1$. So, if we wish to set $\epsilon = O(1)$ then we must have $t = O(1)$. We refer the reader to the full version for the singular values algorithms and their analysis.

1.2 Related work

In machine learning context, Clarkson *et al.* [5] considered several machine learning optimization problems and gave sublinear-time approximation algorithms for those problems. In particular, they considered approximate minimization of a quadratic function over the unit simplex $\Delta = \{\boldsymbol{x} \in \mathbb{R}^n \mid x_i \geq 0,\ \sum_i x_i = 1\}$. Namely, given a positive semidefinite matrix $A \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$, they showed that it is possible to obtain an approximate solution to $\min_{x \in \Delta} \boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{b}$ (up to an additive error of ϵ) in $\widetilde{O}(n/\epsilon^2)$ time, which is sublinear in the input size $\Theta(n^2)$. In contrast, our algorithms run in polylogarithmic time and are much more efficient. Hayashi and Yoshida [11] proposed a constant-time approximation algorithm for Tucker decomposition of tensors, which can be seen as minimizing low-degree polynomials.

In addition to the work of Hayashi and Yoshida [10] mentioned above, an additional line of relevant work is constant-time approximation algorithms for the max cut problem on dense graphs [6, 16]. Let $L_G \in \mathbb{R}^{n \times n}$ be the Laplacian matrix of a graph G on n vertices. Then, the max cut problem can be seen as maximizing $\langle \boldsymbol{x}, L_G \boldsymbol{x} \rangle$ subject to $x_i \in \{-1/\sqrt{n}, 1/\sqrt{n}\}$, and these methods approximate the optimal value to within $O(\epsilon n)$. Our method for approximating the largest singular values can be seen as an extension of these methods to a continuous setting.

1.3 Techniques

The main ingredient in our proof is a novel *spectral decomposition theorem* of matrices, which may be of independent interest. The theorem states that we can decompose a matrix $A \in \mathbb{R}^{n \times m}$ into a structured matrix $A^{\mathrm{str}} \in \mathbb{R}^{n \times m}$ and a pseudorandom matrix $A^{\mathrm{psd}} \in \mathbb{R}^{n \times m}$. Here, A^{str} is structured in the sense that it is a block matrix with a polylogarithmic number of blocks such that the entries in each block are equal. Also, A^{psd} is pseudorandom in the sense that it has a small spectral norm. Formally, we prove the following. For a matrix $A \in \mathbb{R}^{n \times m}$, we define its max norm as $\|A\|_{\max} = \max_{1 \leq i \leq n} \max_{1 \leq j \leq m} |A_{ij}|$.

▶ **Theorem 4.** *For any matrix $A \in [-L, L]^{n \times m}$ and $\gamma \in (0, 1)$, there exists a decomposition $A = A^{str} + A^{psd}$ with the following properties for $N = \sqrt{nm}$:*

1. *A^{str} is* structured *in the sense that it is a block matrix with $\left(\frac{1}{\gamma}\right)^{O(1/\gamma^2)}$ blocks, such that the entries in each block are equal.*
2. *$\|A^{psd}\|_2 \leq \gamma N L$.*
3. *$\|A^{str}\|_{\max} = L/\gamma^{O(1)}$.*

Our decomposition theorem is a strengthening of the matrix decomposition result of Frieze and Kannan [7, 6]. In particular, they showed that any matrix $A \in \mathbb{R}^{n \times m}$ can be decomposed to $D_1 + \cdots + D_s + W$ for $s = O(1/\gamma^2)$, where the matrices D_i are block matrices and $\|W\|_C \leq \gamma nm \|A\|_{\max}$. Here, $\|W\|_C$ is the cut norm, which is defined as

$$\max_{S \subseteq \{1,\ldots,n\}} \max_{T \subseteq \{1,\ldots,m\}} \left| \sum_{i \in S} \sum_{j \in T} W_{ij} \right| .$$

By using a result of Nikiforov [20] that $\|W\|_2 = O(\sqrt{nm \cdot \|W\|_{\max} \cdot \|W\|_C})$ and the fact that the Frieze-Kannan result implies $\|W\|_{\max} \leq \sqrt{s}\|A\|_{\max}$, we get that $\|W\|_2 = O(nm \cdot \|A\|_{\max})$, which is too loose, and thus insufficient for our applications.

Given our decomposition theorem, we can conclude the following. When approximating (1) and (2), we can disregard the pseudorandom part A^{psd}. This will not affect our approximation by much, since A^{psd} has a small spectral norm. In addition, as A^{str} consists of a polylogarithmic number of blocks, such that the entries in each block are equal, we can hit all the blocks by sampling a polylogarithmic number of indices. Hence, we can expect that $A|_S$ is a good approximation to A. To formally define the distance between A and $A|_S$ and to show it is small, we exploit graph limit theory, initiated by Lovász and Szegedy [14] (refer to [13] for a book).

2 Preliminaries

For an integer n we let $[n] \stackrel{\text{def}}{=} \{1, 2, \dots, n\}$. Given a set of indices $S = \{i_1, \dots, i_k\}$, and a vector $\boldsymbol{v} \in \mathbb{R}^n$, we let $\boldsymbol{v}|_S \in \mathbb{R}^k$ be the *restriction* of $\boldsymbol{v}$ to S; that is, $(\boldsymbol{v}|_S)_j = v_{i_j}$, for every $i \in [k]$. Similarly, for a matrix $A \in \mathbb{R}^{n \times m}$ and sets $S_R = \{i_1, \dots, i_{k_R}\} \subseteq [n]$ and $S_C = \{j_1, \dots, j_{k_C}\} \subseteq [m]$, we denote the *restriction* of A to $S_R \times S_C$ by $A|_{S_R \times S_C} \in \mathbb{R}^{k_R \times k_C}$; that is, $(A|_{S_R \times S_C})_{j\ell} = A_{i_j i_\ell}$, for every $j \in [k_R]$ and $\ell \in [k_C]$. When $S_R = S_C = S$ we often use $A|_S$ as a shorthand for $A|_{S \times S}$. We use the notation $x = y \pm z$ as a shorthand for $y - z \leq x \leq y + z$.
Given a matrix $A \in \mathbb{R}^{n \times m}$ we define the *Frobenius norm* of A as $\|A\|_F = \sqrt{\sum_{(i,j) \in [n] \times [m]} A_{ij}^2}$ and the *max norm* of A as $\|A\|_{\max} = \max_{i \in [n], j \in [m]} |A_{ij}|$. For a matrix $A \in \mathbb{R}^{n \times m}$, we let $\sigma_\ell(A)$ denote the ℓ-th largest singular value of A. It is well known that the largest singular value can be evaluated using the following.

$$\sigma_1(A) = \max_{\boldsymbol{v} \in \mathbb{R}^m \colon \|\boldsymbol{v}\|_2 \leq 1} \|A\boldsymbol{v}\|_2 \,.$$

In addition, we state the following fact regarding the singular values.

▶ **Fact 5.** *Let $A \in \mathbb{R}^{n \times m}$, and consider the singular values of A: $\sigma_1(A) \geq \ldots \geq \sigma_{\min\{n,m\}}(A)$. Then, for every $1 \leq \ell \leq \min\{n, m\}$, $\sigma_\ell(A) \leq \frac{\|A\|_F}{\sqrt{\ell}}$.*

3 Spectral Decomposition Theorem

In this section we will prove the following decomposition theorem.

▶ **Theorem 6** (Spectral decomposition, restatement of Theorem 4). *For any matrix $A \in [-L, L]^{n \times m}$ and $\gamma \in (0, 1)$, there exists a decomposition $A = A^{\text{str}} + A^{\text{psd}}$ with the following properties for $N = \sqrt{nm}$:*

1. A^{str} *is* structured *in the sense that it is a block matrix with $O\left(\left(\frac{1}{\gamma^{10}}\right)^{3/\gamma^2}\right)$ blocks, such that the entries in each block are equal.*
2. $\|A^{\text{psd}}\|_2 \leq 7\gamma N L$.
3. $\|A^{\text{str}}\|_{\max} \leq \frac{2}{\gamma^{11}} L$.

The above theorem will serve as a central tool in the analysis of our algorithms. The fact that A^{str} is a block matrix with polylogarithmic number of blocks, such that the entries in each block are equal, implies that by using polylogarithmic number of samples, we can query (with high probability) an entry from each of the blocks. In addition, the fact that A^{psd} has a small spectral norm allows us to disregard it, which only paying a small cost in the error of our approximation.

In order to prove the theorem, we introduce the following definition, two lemmas and a claim.

▶ **Definition 7.** We say that a partition $\mathcal{Q}$ is a *refinement* of a partition $\mathcal{P} = \{V_1, \dots, V_p\}$, if $\mathcal{Q}$ is obtained from $\mathcal{P}$, by splitting some sets V_i into one or more parts.

▶ **Lemma 8.** *Given a matrix $A \in [-L, L]^{n\times m}$ and $\gamma \in (0, 1)$, there exists a block matrix $A^{\mathrm{str}} \in \mathbb{R}^{n\times m}$ with $O\left(\left(\frac{1}{\gamma^{10}}\right)^{3/\gamma^2}\right)$ blocks such that the entries in each block are equal and $\|A - A^{\mathrm{str}}\|_2 \le 7\gamma NL$, where $N = \sqrt{nm}$.*

In order to prove Lemma 8, we will need to prove the following.

▶ **Lemma 9.** *Given a matrix $A \in [-L, L]^{n\times m}$ and $\gamma \in (0, 1)$, let $A' = \sum_{\ell:\sigma_\ell \ge \gamma NL} \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top$. Then, $\|A'\|_{\max} \le \frac{L}{\gamma^3}$.*

Proof. Assume that A has s singular values such that $\sigma_\ell \ge \gamma NL$. For any $\ell \in [s]$, let $M_\ell = \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top$, and let B denote $\sum_{\ell:\, \sigma_\ell < \gamma NL} \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top$. Then, we can write A as,

$$A = \underbrace{\sigma_1 \boldsymbol{u}^1 (\boldsymbol{v}^1)^\top + \cdots + \sigma_s \boldsymbol{u}^s (\boldsymbol{v}^s)^\top}_{A'} + \sum_{\ell:\, \sigma_\ell < \gamma NL} \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top = M_1 + \cdots + M_s + B \; .$$

Consider any $\ell \in [s]$, $(i, j) \in [n] \times [m]$, and let $\beta^\ell_{ij} \stackrel{\text{def}}{=} |\sigma_\ell u^\ell_i v^\ell_j|$.

Let c^A_j denote the j-th column of the matrix A. So, $c^A_j = c^{M_1}_j + \cdots + c^{M_s}_j + c^B_j$. For any $\ell \in [s]$, $c^{M_\ell}_j$ is perpendicular to c^B_j and all $\{c^{M_t}_j\}_t$ such that $t \ne \ell$, and thus,

$$\|c^A_j\|_2 \ge \|c^{M_\ell}_j\|_2 = \|\sigma_\ell \boldsymbol{u}^\ell v^\ell_j\|_2 = \sigma_\ell |v^\ell_j| \; .$$

Let r^A_i denote the i-th row of the matrix A. Then similarly, we have $\|r^A_i\|_2 \ge \sigma_\ell |u^\ell_i|$, and it follows that $\|c^A_j\|_2 \|r^A_i\|_2 \ge \sigma_\ell \beta^\ell_{ij}$.

On the other hand, since $A \in [-L, L]^{n\times m}$, we have $\|c^A_j\|_2 \le \sqrt{n}L$ and $\|r^A_i\|_2 \le \sqrt{m}L$, and therefore,

$$\beta^\ell_{ij} \le \frac{\sqrt{nm}L^2}{\sigma_\ell} \le \frac{\sqrt{nm}L^2}{\gamma NL} = \frac{L}{\gamma} \; .$$

By the fact that $s \le \frac{1}{\gamma^2}$, we get that $\|A'\|_{\max} \le \frac{L}{\gamma^3}$, which concludes the proof. ◀

With this lemma at hand, we are ready to prove Lemma 8.

Proof of Lemma 8: Recall that A can be written as,

$$A = \sum_\ell \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top \; ,$$

where $\sigma_1 \ge \cdots \ge \sigma_{\min\{n,m\}} \ge 0$ are the singular values of A and $\boldsymbol{u}^1, \dots, \boldsymbol{u}^{\min\{n,m\}} \in \mathbb{R}^n$ and $\boldsymbol{v}^1, \dots, \boldsymbol{v}^{\min\{n,m\}} \in \mathbb{R}^m$ are the corresponding left and right singular vectors. If we let A' be such that

$$A' = \sum_{\ell:\, \sigma_\ell \ge \gamma NL} \sigma_\ell \boldsymbol{u}^\ell (\boldsymbol{v}^\ell)^\top \; ,$$

then we have that $\frac{\|(A-A')\boldsymbol{x}\|_2}{\|x\|_2} \leq \gamma NL$ for any $\boldsymbol{x} \in \mathbb{R}^m$.

Next we show the existence of A^{str}, which is a block matrix (with $O\left(\left(\frac{1}{\gamma^{10}}\right)^{3/\gamma^2}\right)$ blocks), such that A^{str} has the same value on every block. We construct A^{str} as follows.

Let $\epsilon = \epsilon(\gamma)$ be determined later and let $J = \frac{1}{\gamma^2}$. Let $\delta_n \stackrel{\text{def}}{=} \frac{\epsilon}{J\sqrt{n}}$ and $T \stackrel{\text{def}}{=} (\frac{J}{\epsilon})^{3/2}$, and partition the interval $[0,1]$ into buckets $B_1^n, \dots, B_T^n, B_{\text{Large}}^n$ such that

$$B_t^n \stackrel{\text{def}}{=} [(t-1)\delta_n, t\delta_n) \;\; \forall t \in [T] \qquad \text{and} \qquad B_{\text{Large}}^n \stackrel{\text{def}}{=} [\sqrt{J/\epsilon n}, 1] \,.$$

For every $\boldsymbol{u}^\ell$ such that $\sigma_\ell \geq \gamma NL$, we define a partition $\mathcal{P}_R^\ell = \{P_{R,1}^\ell \dots, P_{R,T}^\ell, \Sigma_R^\ell\}$ of the indices in $[n]$ so that $P_{R,t}^\ell = \{i \in [n] \mid |u_i^\ell| \in B_t^n\}$ for each $t \in [T]$ and $\Sigma_R^\ell = \{i \in [n] \mid |u_i^\ell| \in B_{\text{Large}}^n\}$. We eliminate emptysets from $\mathcal{P}_R^\ell$ if exist. Note that by the definition of Σ_R^ℓ we have that $|\Sigma_R^\ell| \leq \epsilon n/J$. Next, for every ℓ such that $\sigma_\ell \geq \gamma NL$, we define $\hat{\boldsymbol{u}}^\ell$ as follows.

$$\hat{u}_i^\ell = \begin{cases} 0, & \text{if } |u_i^\ell| \in B_{\text{Large}}^n \,, \\ \delta_n(t-1), & \text{if } |u_i^\ell| \in B_t^n \text{ for some } t \in [T] \,. \end{cases} .$$

Next, let $\Sigma_R = \bigcup_{\ell:\sigma_\ell \geq \gamma NL} \Sigma_R^\ell$ and let $\mathcal{P}_R$ be a partition of $[n] \setminus \Sigma_R$ that refines all $\{P_{R,t}^\ell \mid P_{R,t}^\ell \neq \emptyset\}$ for which ℓ is such that $\sigma_\ell \geq \gamma NL$.

Similarly define $\hat{\boldsymbol{v}}^\ell$ from $\boldsymbol{v}^\ell$ by setting $\delta_m = \frac{\epsilon}{J\sqrt{m}}$, and defining $\mathcal{P}_C^\ell = \{P_{C,1}^\ell, \dots, P_{C,T}^\ell, \Sigma_C^\ell\}$ analogously for each ℓ. Let $\Sigma_C = \bigcup_{\ell:\sigma_\ell \geq \gamma NL} \Sigma_C^\ell$ and let $\mathcal{P}_C$ be a partition of $[m] \setminus \Sigma_C$ that refines all $\{P_{C,t}^\ell \mid P_{C,t}^\ell \neq \emptyset\}$ for which ℓ is such that $\sigma_\ell \geq \gamma NL$.

Define,

$$A^{\text{str}} \stackrel{\text{def}}{=} \sum_{\ell:\, \sigma_\ell \geq \gamma NL} \sigma_\ell \hat{\boldsymbol{u}}^\ell (\hat{\boldsymbol{v}}^\ell)^\top \qquad \text{and} \qquad A^{\text{psd}} \stackrel{\text{def}}{=} A - A^{\text{str}}.$$

By Fact 5 we have that $\sigma_\ell \leq \frac{NL}{\sqrt{\ell}}$ for all $\ell \geq 1$. Therefore, we can have at most $J = 1/\gamma^2$ indices such that $\sigma_\ell \geq \gamma NL$. Thus, the partition $\mathcal{P}_R$ satisfies $|\mathcal{P}_R| \leq \left((\frac{J}{\epsilon})^{\frac{3}{2\gamma^2}}\right) = \left(\frac{1}{\gamma^2\epsilon}\right)^{(3/2\gamma^2)}$. Similarly, we have $|\mathcal{P}_C| \leq \left(\frac{1}{\gamma^2\epsilon}\right)^{(3/2\gamma^2)}$. Therefore, the resulting matrix is a block matrix with $O\left(\left(\frac{1}{\gamma^2\epsilon}\right)^{\frac{3}{\gamma^2}}\right)$ many blocks, such that all the entries in each block are the same. We refer to $\mathcal{P}_R$ and $\mathcal{P}_C$ the *row partition of* A^{str} and the *column partition of* A^{str} respectively.

Next, we have that

$$\begin{aligned} \|A^{\text{psd}}\boldsymbol{x}\| = \|(A - A^{\text{str}})\boldsymbol{x}\|_2^2 &\leq \|(A - A')\boldsymbol{x}\|_2^2 + \|(A' - A^{\text{str}})\boldsymbol{x}\|_2^2 \\ &\leq \gamma^2 N^2 L^2 \|\boldsymbol{x}\|_2^2 + \|(A' - A^{\text{str}})\boldsymbol{x}\|_2^2 \,. \end{aligned}$$

Consider the “error” term $\|(A' - A^{\text{str}})\boldsymbol{x}\|_2^2$.

$$\begin{aligned} \|(A' - A^{\text{str}})\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\|_2^2 \sum_{(i,j)\in[n]\times[m]} \left(A'_{ij} - A_{ij}^{\text{str}}\right)^2 = \|\boldsymbol{x}\|_2^2 \Bigg(\sum_{(i,j)\in\bar{\Sigma}_R\times\bar{\Sigma}_C} \left(A'_{ij} - A_{ij}^{\text{str}}\right)^2 \\ + \sum_{(i,j)\in(\Sigma_R\times\bar{\Sigma}_C)\cup(\bar{\Sigma}_R\times\Sigma_C)} \left(A'_{ij} - A_{ij}^{\text{str}}\right)^2 + \sum_{(i,j)\in\Sigma_R\times\Sigma_C} \left(A'_{ij} - A_{ij}^{\text{str}}\right)^2 \Bigg), \end{aligned}$$

where $\bar{\Sigma}_R = [n] \setminus \Sigma_R$ and $\bar{\Sigma}_C = [m] \setminus \Sigma_C$.

We analyze each of these terms separately. First, note that

$$\left(A'_{ij} - A^{\text{str}}_{ij}\right)^2 = \left(\sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell (u_i^\ell v_j^\ell - \hat{u}_i^\ell \hat{v}_j^\ell)\right)^2 .$$

So,

$$\begin{aligned}\sum_{(i,j)\in\bar{\Sigma}_R\times\bar{\Sigma}_C} \left(A'_{ij} - A^{\text{str}}_{ij}\right)^2 &= \sum_{(i,j)\in\bar{\Sigma}_R\times\bar{\Sigma}_C} \left(\sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell (u_i^\ell v_j^\ell - \hat{u}_i^\ell \hat{v}_j^\ell)\right)^2 \\ &\overset{(*)}{\leq} \sum_{(i,j)\in\bar{\Sigma}_R\times\bar{\Sigma}_C} \left(\frac{2\sqrt{\epsilon}}{N} \sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell\right)^2 \\ &\overset{(**)}{\leq} \sum_{(i,j)\in\bar{\Sigma}_R\times\bar{\Sigma}_C} \left(\frac{2\sqrt{\epsilon}}{N} \cdot \frac{2NL}{\gamma}\right)^2 \leq \frac{16\epsilon N^2 L^2}{\gamma^2} .\end{aligned}$$

Here, $(*)$ follows from the fact that for two indices $i \in \bar{\Sigma}_R, j \in \bar{\Sigma}_C$ we have that $\hat{u}_i^\ell \hat{v}_j^\ell = (u_i^\ell \pm \delta_n)(v_j^\ell \pm \delta_m) \leq u_i^\ell v_j^\ell \pm \frac{2\sqrt{\epsilon}}{N}$. $(**)$ follows from the fact that there can be at most $1/\gamma^2$ indices ℓ, such that $\sigma_\ell \geq \gamma NL$, and therefore

$$\sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell \leq \sum_{\ell=1}^{1/\gamma^2} \sigma_\ell \leq \sum_{\ell=1}^{1/\gamma^2} \frac{NL}{\sqrt{\ell}} \leq \frac{2NL}{\gamma} . \tag{3}$$

Next, we have that,

$$\begin{aligned}&\sum_{(i,j)\in(\Sigma_R\times\bar{\Sigma}_C)\cup(\bar{\Sigma}_R\times\Sigma_C)} \left(A'_{ij} - A^{\text{str}}_{ij}\right)^2 \\ &= \sum_{(i,j)\in(\Sigma_R\times\bar{\Sigma}_C)\cup(\bar{\Sigma}_R\times\Sigma_C)} \left(\sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell (u_i^\ell v_j^\ell - \hat{u}_i^\ell \hat{v}_j^\ell)\right)^2 \\ &\overset{(***)}{=} \sum_{(i,j)\in(\Sigma_R\times\bar{\Sigma}_C)\cup(\bar{\Sigma}_R\times\Sigma_C)} \left(A'_{ij}\right)^2 \leq \sum_{(i,j)\in(\Sigma_R\times\bar{\Sigma}_C)\cup(\bar{\Sigma}_R\times\Sigma_C)} \left(\frac{L}{\gamma^3}\right)^2 \leq \frac{2\epsilon N^2 L^2}{\gamma^6} .\end{aligned}$$

Here, $(* * *)$ follows from the fact that when one of the indices is in Σ_R or Σ_C we set the corresponding entry in the rounded vector to 0. In addition, the last inequality follows from the fact that when we remove the lower singular part of the matrix, we can only increase the value by at most factor of $1/\gamma^3$ (see Lemma 9). Finally,

$$\begin{aligned}\sum_{(i,j)\in\Sigma_R\times\Sigma_C} \left(A'_{ij} - A^{\text{str}}_{ij}\right)^2 &= \sum_{(i,j)\in\Sigma_R\times\Sigma_C} \left(\sum_{\ell:\sigma_\ell \geq \gamma NL} \sigma_\ell (u_i^\ell v_j^\ell - \hat{u}_i^\ell \hat{v}_j^\ell)\right)^2 \\ &= \sum_{(i,j)\in\Sigma_R\times\Sigma_C} \left(A'_{ij}\right)^2 \leq \frac{\epsilon^2 N^2 L^2}{\gamma^6}\end{aligned}$$

Combining all the three terms and setting $\epsilon = \gamma^8$ gives,

$$\|(A' - A^{\text{str}})\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\|_2^2 \left(\frac{16\epsilon}{\gamma^2} + \frac{2\epsilon}{\gamma^6} + \frac{\epsilon^2}{\gamma^6}\right) N^2 L^2 \leq 19\gamma^2 \|\boldsymbol{x}\|_2^2 N^2 L^2.$$

Therefore, we get that,

$$\begin{aligned}\|(A - A^{\mathrm{str}})\boldsymbol{x}\|_2^2 &\le 2\|(A - A')\boldsymbol{x}\|_2^2 + 2\|(A' - A^{\mathrm{str}})\boldsymbol{x}\|_2^2 \\ &\le 2(\gamma^2 N^2 L^2 \|\boldsymbol{x}\|_2^2 + 19\gamma^2 \|\boldsymbol{x}\|_2^2 N^2 L^2) \le 40\gamma^2 \|\boldsymbol{x}\|_2^2 N^2 L^2 ,\end{aligned}$$

and the lemma follows. ◀

We are left with bounding the max norm of A^{str}.

▶ **Claim 10.** *Given a matrix $A \in [-L, L]^{n\times m}$ and $\gamma \in (0, 1)$, let $A^{\mathrm{str}} \in \mathbb{R}^{n\times m}$ be the block approximation matrix defined above. Then, $\|A^{\mathrm{str}}\|_{\max} \le \frac{2L}{\gamma^{11}}$*

Proof. By the definition of A^{str} we have that $|A^{\mathrm{str}}_{ij}| = |\sum_{\ell:\sigma_\ell \ge \gamma N L} \sigma_\ell \hat{u}^\ell_i \hat{v}^\ell_j|$. By the definition of the rounding process, we have that for every $i \in [n]$ we have that $|\hat{u}^\ell_i| \le \sqrt{\frac{J}{\epsilon n}} = \frac{1}{\gamma^5\sqrt{n}}$ (recall that $J = 1/\gamma^2$ and $\epsilon = \gamma^8$). Similarly, for every $j \in [m]$ we have that $|\hat{v}^\ell_j| \le \frac{1}{\gamma^5\sqrt{m}}$. Therefore,

$$|A^{\mathrm{str}}_{ij}| = \left| \sum_{\ell:\sigma_\ell \ge \gamma N L} \sigma_\ell \hat{u}^\ell_i \hat{v}^\ell_j \right| \le \frac{1}{\gamma^{10} N} \sum_{\ell:\sigma_\ell \ge \gamma N L} \sigma_\ell \le \frac{2L}{\gamma^{11}} ,$$

where the last inequality uses (3). ◀

Proof of Theorem 6: The proof follows directly from Lemma 8 and Claim 10. ◀

4 Dikernels and Sampling Lemmas

In this section we will formalize the idea that $A|_{S_R\times S_C}$ is a good approximation of A when S_R and S_C are uniformly random subsets of indices. (The proof for $A|_S$, where S is uniformly random subset of indices, is almost identical and we omit it.) We start by providing some background on dikernels and their connection to matrices and then move on to proving our sampling lemmas.

4.1 Dikernels and Matrices

We call a (measurable) function $f\colon [0, 1]^2 \to \mathbb{R}$ a *dikernel*. We can regard a dikernel as a matrix whose index is specified by a real value in $[0, 1]$. For two functions $f, g : [0, 1] \to \mathbb{R}$, we define their inner product as $\langle f, g\rangle \stackrel{\mathrm{def}}{=} \int_0^1 f(x)g(x)dx$. For a dikernel $\mathcal{A} : [0, 1]^2 \to \mathbb{R}$ and a function $f : [0, 1] \to \mathbb{R}$, we define the function $\mathcal{A}f\colon [0, 1] \to \mathbb{R}$ as $(\mathcal{A}f)(x) = \langle \mathcal{A}(x, \cdot), f\rangle$. In addition, we define the *spectral norm* of $\mathcal{A}$ as $\|\mathcal{A}\|_2 \stackrel{\mathrm{def}}{=} \sup_{f\colon [0,1]\to\mathbb{R}} \frac{\|\mathcal{A}f\|_2}{\|f\|_2}$, and the *Frobenius norm* of $\mathcal{A}$ as $\|\mathcal{A}\|_F \stackrel{\mathrm{def}}{=} \sqrt{\int_0^1\int_0^1 \mathcal{A}(x, y)^2 dxdy}$.

For an integer $n \in \mathbb{N}$, let $I^n_1 = [0, \frac{1}{n}]$, and for every $1 < k \le n$, let $I^n_k = (\frac{k-1}{n}, \frac{k}{n}]$. For $x \in [0, 1]$, we define $i^n(x)$ as the unique integer $k \in [n]$ such that $x \in I^n_k$.

▶ **Definition 11.** Given a matrix $A \in \mathbb{R}^{n\times m}$, we construct the corresponding dikernel $\mathcal{A}$ as $\mathcal{A}(x, y) = A_{i^n(x),i^m(y)}$. In addition, given two sets of indices $S_R \subseteq [n]$ and $S_C \subseteq [m]$, when we write $\mathcal{A}|_{\mathcal{S}_R\times \mathcal{S}_C}$, we first extract the matrix $A|_{S_R\times S_C}$ and then consider its corresponding dikernel.

The following lemma shows that the spectral norms of A and $\mathcal{A}$ are essentially the same up to normalization. The proof can be found in Appendix A.1.

▶ **Lemma 12.** *Let $A \in \mathbb{R}^{n\times m}$ be a matrix. Then, we have*

$$\max_{\boldsymbol{v}\in\mathbb{R}^m} \frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} = nm \cdot \sup_{f\colon [0,1]\to\mathbb{R}} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2} .$$

▶ **Corollary 13.** *Let $A \in \mathbb{R}^{n\times m}$ be a matrix. Then,*

$$\|A\|_2 = \sqrt{nm} \cdot \|\mathcal{A}\|_2 .$$

Proof. The proof is immediate by the definition of the spectral norm and Lemma 12. ◀

▶ **Definition 14.** Let μ be a Lebesgue measure. A map $\pi\colon [0,1] \to [0,1]$ is *measure preserving* if the pre-image $\pi^{-1}(X)$ is measurable for every measurable set X and $\mu(\pi^{-1}(X)) = \mu(X)$. A *measure preserving bijection* is a measure preserving map whose inverse map exists and is also measurable. For a measure preserving bijection π and a dikernel $\mathcal{A}$, we define the dikernel $\pi(\mathcal{A})$, as $\pi(\mathcal{A})(x,y) = \mathcal{A}(\pi(x),\pi(y))$.

4.2 Sampling Lemmas

In this subsection, we will prove that given matrices $A_1,\dots,A_T \in [-L,L]^{n\times m}$, we obtain a good approximation of their corresponding dikernels, by sampling a small number of elements. The next lemma states that there is a way to "align" the sampled matrices with the original matrices. We refer the reader to Appendix A.1 for the full proofs.

▶ **Lemma 15.** *Given matrices $A_1,\dots,A_T \in [-L,L]^{n\times n}$ and $\gamma \in (0,1)$, let $A_1^{str},\dots,A_T^{str}$ be the block approximation matrices as in Lemma 8. In addition, for $t\in[T]$, let $\mathcal{P}_R^{A_t} = \{V_1^{A_t},\dots,V_p^{A_t}\}$ and $\mathcal{P}_C^{A_t} = \{V_1'^{A_t},\dots,V_q'^{A_t}\}$ be the row and column partitions of A_t^{str} (from Lemma 8). Let S_R be a set of size s_R, generated by picking each element in $[n]$ independently with probability k_R/n, and let S_C be a set of size s_C, generated by picking each element in $[m]$ independently with probability k_C/m for some $k_R,k_C>0$.*

Then, there exists a measure preserving bijection $\pi\colon [0,1]\to[0,1]$ such that for every $t\in[T]$

$$\mathop{\mathbf{E}}_{S_R,S_C}\left[\|\mathcal{A}_t^{str} - \pi(\mathcal{A}_t^{str}|_{S_R\times S_C})\|_2\right] = O\left(\frac{L}{\gamma^{11}}\cdot\max\left(\sqrt{\frac{p^{T/2}}{s_R^{1/2}}},\sqrt{\frac{q^{T/2}}{s_C^{1/2}}}\right)\right) .$$

In the following lemma we prove concentration around the mean.

▶ **Lemma 16.** *Let $\gamma>0$, and $A_1,\dots,A_T \in [-L,L]^{n\times m}$. Let S_R be a set generated by picking each element in $[n]$ independently with probability k_R/n and let S_C be a set generated by picking each element in $[m]$ independently with probability k_C/m for some $k_R,k_C>0$. Then, with probability at least $89/100$ there exists a measure preserving bijection $\pi\colon[0,1]\to[0,1]$ such that for every $t\in[T]$,*

$$\|\mathcal{A}_t - \pi\left(\mathcal{A}_t|_{S_R\times S_C}\right)\|_2 \le 210\gamma LT + 10LT\left(\sqrt{\frac{8\log n}{k_C}} + \sqrt{\frac{8\log m}{k_R}} + \sqrt{\frac{4\log n\log m}{k_Rk_C}}\right. \\ \left. + O\left(\frac{LT}{\gamma^{11}}\left(\frac{1}{\gamma^{10}}\right)^{\frac{3T}{4\gamma^2}}\cdot\max\left(\frac{1}{k_R^{1/4}},\frac{1}{k_C^{1/4}}\right)\right)\right),$$

where $N=\sqrt{nm}$.

Algorithm 1 Minimization Algorithm(A, n, ϵ, k).

1: Let $S \subseteq \{1, 2, \ldots, n\}$ such that each index i is taken to S independently w.p k/n.
2: **if** $|S| > 2k$ **then**
3: **Abort**
4: **return** $\min_{\boldsymbol{v} \in \mathbb{R}^{|S|}} \psi_{|S|, A|_S, \boldsymbol{d}|_S, \boldsymbol{b}|_S}(\boldsymbol{v})$.

5 Applications

5.1 Quadratic Function Minimization

In this section, we show that we can approximately solve quadratic function minimization problems in polylogarithmic time.

Recall that we are given a matrix $A \in \mathbb{R}^{n \times n}$ and vectors $\boldsymbol{d}, \boldsymbol{b} \in \mathbb{R}^n$, and consider the following quadratic function minimization problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}), \text{ where } \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) = \langle \boldsymbol{v}, A\boldsymbol{v} \rangle + n \langle \boldsymbol{v}, \operatorname{diag}(\boldsymbol{d})\boldsymbol{v} \rangle + n \langle \boldsymbol{b}, \boldsymbol{v} \rangle \ . \tag{4}$$

Here $\operatorname{diag}(\boldsymbol{d}) \in \mathbb{R}^{n \times n}$ is a matrix whose diagonal entries are specified by $\boldsymbol{d}$.

First, we describe our algorithm for minimizing quadratic functions. We first sample a set of indices $S \subseteq \{1, 2, \ldots, n\}$ with each index included with probability k/n, where k is some constant. If $|S|$ is too large, we immediately stop the process by claiming that the algorithm has failed. Otherwise, we solve the problem on $A|_S$, $\boldsymbol{d}|_S$, $\boldsymbol{b}|_S$ and then output the optimal solution. The detail is given in Algorithm 1.

Due to our extensive use of dikernels in the analysis, we introduce a continuous version of problem (4). The real valued function $\Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}$ on function $f\colon [0,1] \to \mathbb{R}$ is defined as

$$\Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f) = \langle f, \mathcal{A}f \rangle + \left\langle f^2, \mathcal{D}\mathbf{1} \right\rangle + \langle f, \mathcal{B}\mathbf{1} \rangle \ ,$$

where $\mathcal{D}$ and $\mathcal{B}$ are the corresponding dikernels of $\boldsymbol{d} \cdot \mathbf{1}^\top$ and $\boldsymbol{b} \cdot \mathbf{1}^\top$ respectively, $f^2\colon [0,1] \to \mathbb{R}$ is a function such that $f^2(x) = f(x)^2$ for every $x \in [0,1]$ and $\mathbf{1}\colon [0,1] \to \mathbb{R}$ is the constant function that has the value 1 everywhere.

In order to prove Theorem 1, we prove that the minimizations of $\psi_{n,A,\boldsymbol{d},\boldsymbol{b}}$ and $\Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}$ are essentially equivalent. The proof of the lemma can be found in Appendix A.2.

▶ **Lemma 17.** *Let $A \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b}, \boldsymbol{d} \in \mathbb{R}^n$. Then, for any $r > 0$*

$$\min_{\boldsymbol{v}: \|\boldsymbol{v}\|_2 \le r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) = n^2 \cdot \inf_{f: \|f\|_2 \le \frac{r}{\sqrt{n}}} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f) \ .$$

With the above lemma, we are ready to prove our main result.

Proof of Theorem 1: By applying Chernoff bounds, we have that with probability at least $1 - o(1)$, the size of S is at most $2k$. As before, we apply Lemma 16 with $\gamma = O(\epsilon)$,

$$k = \max \left\{ O\left(\frac{\log^2 n}{\epsilon^2} \right), \left(\frac{1}{\epsilon} \right)^{O(1/\epsilon^2)} \right\},$$

and $S_C = S_R = S$[3]. Then, with probability at least $2/3$ there exists a measure preserving bijection $\pi\colon [0,1] \to [0,1]$ such that for any function $f\colon [0,1] \to \mathbb{R}$,

$$\max \left\{ \Big| \langle f, (\mathcal{A} - \pi(\mathcal{A}|_{\mathcal{S}}))f \rangle \Big|, \Big| \langle f, (\mathcal{D} - \pi(\mathcal{D}|_{\mathcal{S}}))f \rangle \Big|, \Big| \langle f, (\mathcal{B} - \pi(\mathcal{B}|_{\mathcal{S}}))f \rangle \Big| \right\} \le \frac{\epsilon L}{3} \|f\|_2^2 \ .$$

[3] We note that in this case, the sets S_R and S_C are dependent. However, the proof of Lemma 16 can be easily modified to the case where the matrix is in $[-L, L]^{n \times n}$, where for this case $S_R = S_C = S$.

Algorithm 2 Minimization Algorithm Over a Ball(A, n, ϵ, k, r).

```
Let S ⊆ {1, 2, ..., n} such that each index i is taken to S independently w.p k/n.
if |S| > 2k then
    Abort
return min_{‖v‖_2 ≤ sqrt(|S|/n) r} ψ_{|S|, A|_S, d|_S, b|_S}(v).
```

Let $R = \max\left\{\frac{\|\tilde{\boldsymbol{v}}^*\|_2}{\sqrt{|S|}}, \frac{\|\boldsymbol{v}^*\|_2}{\sqrt{n}}\right\}$. Then, by using Lemma 17:

$$\begin{aligned}
\tilde{z}^* &= \min_{\boldsymbol{v}\in\mathbb{R}^{|S|}} \psi_{|S|,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v}) = \min_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\le R\sqrt{|S|}} \psi_{|S|,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(\boldsymbol{v}) \\
&= |S|^2 \cdot \min_{f:\|f\|_2\le R} \Psi_{|S|,A|_S,\boldsymbol{d}|_S,\boldsymbol{b}|_S}(f) \\
&= |S|^2 \cdot \min_{f:\|f\|_2\le R} \Big\{ \langle f, (\pi(\mathcal{A}|_{\mathcal{S}}) - \mathcal{A})\, f\rangle + \langle f, \mathcal{A}f\rangle + \left\langle f^2, (\pi(\mathcal{D}|_{\mathcal{S}}) - \mathcal{D})\,\mathbf{1}\right\rangle + \left\langle f^2, \mathcal{D}\mathbf{1}\right\rangle \\
&\qquad\qquad + \langle f, (\pi(\mathcal{B}|_{\mathcal{S}}) - \mathcal{B})\,\mathbf{1}\rangle + \langle f, \mathcal{B}\mathbf{1}\rangle \Big\} \\
&\le |S|^2 \cdot \min_{f:\|f\|_2\le R} \left\{\langle f, \mathcal{A}f\rangle + \left\langle f^2, \mathcal{D}\mathbf{1}\right\rangle + \langle f, \mathcal{B}\mathbf{1}\rangle \pm \epsilon L\|f\|_2^2\right\} \\
&\le |S|^2 \cdot \min_{f:\|f\|_2\le R} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f) \pm \epsilon L|S|^2R^2 \\
&= \frac{|S|^2}{n^2} \cdot \min_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\le\sqrt{n}R} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \pm \epsilon L|S|^2R^2 \\
&= \frac{|S|^2}{n^2} \cdot \min_{\boldsymbol{v}\in\mathbb{R}^n} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \pm \epsilon L|S|^2R^2 = \frac{|S|^2}{n^2} z^* \pm \epsilon L|S|^2R^2 \ .
\end{aligned}$$

By rearranging the inequality and applying the union bound the theorem follows. ◀

As a corollary we show that we can obtain an approximation algorithm for minimizing a quadratic function over a ball of radius r with better error bounds compared than the one obtained by Hayashi and Yoshida [10]. The proof of correctness is similar to the proof of Theorem 1.

▶ **Corollary 18** (Restatement of Theorem 2). *Let $\boldsymbol{v}^*$ and z^* be an optimal solution and optimal value, respectively, of $\min_{\|\boldsymbol{v}\|_2\le r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v})$. Let $\epsilon > 0$ and let S be a random set generated as in Algorithm 2 with*

$$k = \max\left\{O\left(\frac{\log^2 n}{\epsilon^2}\right), \left(\frac{1}{\epsilon}\right)^{O(1/\epsilon^2)}\right\}.$$

Then, we have that with probability at least $2/3$, the following hold: Let $\tilde{\boldsymbol{v}}^*$ and $\tilde{z}^*$ be an optimal solution and the optimal value, respectively, of the problem

$$\min_{\|v\|_2\le\sqrt{\frac{|S|}{n}}r} \psi_{|S|,A|_S,d|_S,b|_S}(\boldsymbol{v}) \ .$$

Then,

$$\left|\frac{1}{|S|^2}\tilde{z}^* - \frac{1}{n^2}z^*\right| \le \frac{\epsilon L r^2}{n} \ ,$$

where $L = \max\{\max_{i,j}|A_{ij}|, \max_i |d_i|, \max_i |b_i|\}$.

References

1 Farid Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51, 1995.

2 Aharon Ben-Tal and Marc Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Mathematical Programming*, 72(1):51–63, 1996.

3 Léon Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning*, pages 146–168. Springer, 2004.

4 Stanislav Busygin. A new trust region technique for the maximum weight clique problem. *Discrete Applied Mathematics*, 154(15):2080–2096, 2006.

5 Kenneth L. Clarkson, Elad Hazan, and David P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM*, 59(5):23:1–23:49, 2012.

6 A Frieze and R Kannan. The regularity lemma and approximation schemes for dense problems. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 12–20, 1996.

7 Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

8 Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

9 Walter Gander, Gene H Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its applications*, 114:815–839, 1989.

10 Kohei Hayashi and Yuichi Yoshida. Minimizing quadratic functions in constant time. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2217–2225, 2016.

11 Kohei Hayashi and Yuichi Yoshida. Fitting low-rank tensors in constant time. In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2473–2481, 2017.

12 Elad Hazan and Tomer Koren. A linear-time algorithm for trust region problems. *Mathematical Programming*, 158(1-2):363–381, 2016.

13 L Lovász. *Large Networks and Graph Limits.* American Mathematical Society, 2012.

14 László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

15 David G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1997.

16 Claire Mathieu and Warren Schudy. Yet another algorithm for dense max cut: go greedy. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 176–182, 2008.

17 Kevin P Murphy. *Machine learning: a probabilistic perspective.* The MIT Press, 2012.

18 Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

19 Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming.* SIAM, 1994.

20 Vladimir Nikiforov. Cut-norms and spectra of matrices. *arXiv:0912.0336*, 2009.

21 Anthony L. Peressini, Francis E. Sullivan, and J. J. Jr. Uhl. *The Mathematics of Nonlinear Programming.* Springer, 1993.

22 Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8:1–230, 2015.

23 Yinyu Ye and Shuzhong Zhang. New results on quadratic minimization. *SIAM Journal on Optimization*, 14(1):245–267, 2003.

24 Hongchao Zhang, Andrew R Conn, and Katya Scheinberg. A derivative-free algorithm for least-squares minimization. *SIAM Journal on Optimization*, 20(6):3555–3576, 2010.

A Omitted Technical Lemmas and Proofs

A.1 Dikernels and Sampling Lemmas

Proof of Lemma 12: Before starting the proof we introduce some notations and an important observation. We note that $\sup_{f:[0,1]\to\mathbb{R}} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2}$ has a minimizer, since the objective function is weakly continuous and we may assume that $\|f\|_2 \le 1$ which is weakly compact.

For every $x \in [0,1]$ and $j \in [m]$, we let $\mathcal{A}(x, I_j^m) = A_{i^n(x)j}$. Also for every $(i,j) \in [n]\times[m]$, we let $\mathcal{A}(I_i^n, I_j^m) = A_{ij}$.

We start by showing that $\max_{\boldsymbol{v}\in\mathbb{R}^m} \frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \le nm \cdot \sup_{f:[0,1]\to\mathbb{R}} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2}$. Given a vector $\boldsymbol{v} \in \mathbb{R}^m$, we define the function $f : [0,1] \to \mathbb{R}$ as $f(x) = v_{i^m(x)}$. Then,

$$\begin{aligned}
\|\mathcal{A}f\|_2^2 &= \int_0^1 \left(\int_0^1 \mathcal{A}(x,y)f(y)dy\right)^2 dx = \int_0^1 \left(\sum_{j\in[m]}\int_{I_j^m} \mathcal{A}(x,y)f(y)dy\right)^2 dx \\
&= \int_0^1 \left(\frac{1}{m}\sum_{j\in[m]} \mathcal{A}(x,I_j^m)v_j\right)^2 dx = \frac{1}{n}\sum_{i\in[n]}\left(\frac{1}{m}\sum_{j\in[m]}\mathcal{A}(I_i^n,I_j^m)v_j\right)^2 \\
&= \frac{1}{n}\sum_{i\in[n]}\left(\frac{1}{m}\sum_{j\in[m]}A_{ij}v_j\right)^2 = \frac{1}{nm^2}\|A\boldsymbol{v}\|_2^2 .
\end{aligned}$$

In addition,

$$\|f\|_2^2 = \int_0^1 f(x)^2 dx = \sum_{j\in[m]}\int_{I_j^m} f(x)^2 dx = \frac{1}{m}\sum_{j\in[m]} v_j^2 = \frac{1}{m}\|\boldsymbol{v}\|_2^2 .$$

Next, we will prove that $\max_{\boldsymbol{v}\in\mathbb{R}^m} \frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \ge nm \cdot \sup_{f:[0,1]\to\mathbb{R}} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2}$. Let $f : [0,1] \to \mathbb{R}$ be a measurable function. Then, for any $x \in [0,1]$, consider the partial derivative,

$$\frac{\partial}{\partial f(x)}\frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2} = \frac{\|f\|_2^2 \cdot \frac{\partial}{\partial f(x)}\|\mathcal{A}f\|_2^2 - \|\mathcal{A}f\|_2^2 \cdot \frac{\partial}{\partial f(x)}\|f\|_2^2}{\|f\|_2^4} .$$

Note that,

$$\frac{\partial}{\partial f(x)}\|\mathcal{A}f\|_2^2 = \frac{\partial}{\partial f(x)}\langle \mathcal{A}f, \mathcal{A}f\rangle = \frac{\partial}{\partial f(x)}\langle f, (\mathcal{A}^*\mathcal{A})f\rangle ,$$

where $\mathcal{A}^*(x,y) = \mathcal{A}^*(y,x)$. So, for $\mathcal{M} = \mathcal{A}^*\mathcal{A}$, we have

$$\begin{aligned}
\frac{\partial}{\partial f(x)}\|\mathcal{A}f\|_2^2 &= \frac{\partial}{\partial f(x)}\langle f, \mathcal{M}f\rangle \\
&= \sum_{j\in[m]}\int_{I_j^m}\mathcal{M}(I_j^m, i^m(x))f(y)dy + \sum_{j\in[m]}\int_{I_j^m}\mathcal{M}(i^m(x), I_j^m)f(y)dy .
\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial f(x)} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2}$$
$$= \frac{\|f\|_2^2 \left(\sum_{j \in [m]} \int_{I_j^m} \mathcal{M}(I_j^m, i^m(x)) f(y) dy + \sum_{j \in [m]} \int_{I_j^m} \mathcal{M}(i^n(x), I_j^m) f(y) dy \right)}{\|f\|_2^4}$$
$$- \frac{2\|\mathcal{A}f\|_2^2 \cdot f(x)}{\|f\|_2^4} .$$

Consider the optimal solution f^*. By the form of the partial derivative, it holds that $f^*(z) = f^*(z')$ for almost all $z, z' \in [0,1]$ such that $i_n(z) = i_n(z')$. That is, f^* is almost constant on each of the intervals $I_1^m, \ldots, I_m^m$. Hence, we can define $\boldsymbol{v} \in \mathbb{R}^m$ as $v_j = f^*(x)$, where x is the dominant value in I_j^m. Then,

$$\|A\boldsymbol{v}\|_2^2 = \sum_{i \in [n]} \left(\sum_{j \in [m]} A_{ij} v_j \right)^2 = nm^2 \int_0^1 \left(\sum_{j \in [m]} \int_{I_j^m} \mathcal{A}(x, I_j^m) f^*(y) dy \right)^2 dx$$
$$= nm^2 \int_0^1 \left(\int_0^1 \mathcal{A}(x,y) f^*(y) dy \right)^2 dx = nm^2 \|\mathcal{A}f^*\|_2^2$$

Moreover,

$$\|f^*\|_2^2 = \int_0^1 f^N(x)^2 dx = \sum_{j \in [m]} \int_{I_j^m} f^*(x)^2 dx = \frac{1}{m} \sum_{j \in [m]} v_j^2$$

Therefore, we have $\max_{v \in \mathbb{R}^m} \frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \geq nm \cdot \sup_{f:[0,1] \to \mathbb{R}} \frac{\|\mathcal{A}f\|_2^2}{\|f\|_2^2}$. ◀

Proof of Lemma 15: We first note that for any $t \in [T]$, any refinement of the row or column partitions of A_t^{str} will give the same block approximation matrix A_t^{str}. Therefore, let $\mathcal{P}_R = \{V_1^R, \ldots, V_P^R\}$ be a partition which refines $\mathcal{P}_R^{A_1}, \ldots, \mathcal{P}_R^{A_T}$ and its size is $P = O(p^T)$. Similarly, let $\mathcal{P}_C = \{V_1^C, \ldots, V_Q^C\}$ be a partition which refines $\mathcal{P}_C^{A_1}, \ldots, \mathcal{P}_C^{A_T}$ and its size is $Q = O(q^T)$.

We denote by z_i^R the number of elements of S_R that falls into the set V_i^R. Then,

$$\mathop{\mathbf{E}}_{S_R}[z_i^R] = s_R \cdot \mu(V_i^R) \text{ and } \mathbf{Var}[z_i^R] = \mu(V_i^R)(1 - \mu(V_i^R)) s_R .$$

Similarly, we denote by z_j^C the number of elements of S_C that falls into the set V_j^C. Then,

$$\mathop{\mathbf{E}}_{S_C}[z_j^C] = s_C \cdot \mu(V_j^C) \text{ and } \mathbf{Var}[z_j^C] = \mu(V_j^C)(1 - \mu(V_j^C)) s_C .$$

We next construct a measure preserving bijection. We define the following two partitions of the $[0,1]$ interval.

Let $\{V_1^{R'}, \ldots, V_P^{R'}\}$ be a partition such that $\mu(V_i^{R'}) = z_i^R / s_R$ and $\mu(V_i^R \cap V_i^{R'}) = \min(\mu(V_i^R), z_i^R / s_R)$, and let $\{V_1^{C'}, \ldots, V_Q^{C'}\}$ be a partition such that $\mu(V_j^{C'}) = z_j^C / s_C$ and $\mu(V_j^C \cap V_j^{C'}) = \min(\mu(V_j^C), z_j^C / s_C)$. We construct the dikernel $\mathcal{Y} : [0,1]^2 \to \mathbb{R}$ such that the value of $\mathcal{Y}$ on $V_i^{R'} \times V_j^{C'}$ is the same as the value of $\mathcal{A}^{\mathrm{str}}$ on $V_i^R \times V_j^C$. Therefore, the dikernel $\mathcal{Y}$ agrees with $\mathcal{A}^{\mathrm{str}}$ on the set $Y = \bigcup_{(i,j) \in [P] \times [Q]} (V_i^R \cap V_i^{R'}) \times (V_j^C \cap V_j^{C'})$. Then, there exists a bijection π such that $\pi(\mathcal{A}^{\text{str}}|_{S_R\times S_C}) = \mathcal{Y}$. Then,

$$\begin{aligned}
1-\mu(Y) &\le 1-\left(\sum_{i\in[P]}\min\left(\mu(V_i^R),\frac{z_i^R}{s_R}\right)\right)\left(\sum_{j\in[Q]}\min\left(\mu(V_j^C),\frac{z_j^C}{s_C}\right)\right)\\
&= 1-\left(1-\frac{1}{2}\sum_{i\in[P]}\left|\mu(V_i^R)-\frac{z_i^R}{s_R}\right|\right)\left(1-\frac{1}{2}\sum_{j\in[Q]}\left|\mu(V_j^C)-\frac{z_j^C}{s_C}\right|\right)\\
&\le \max\left\{\left(P\sum_{i\in[P]}\left(\mu(V_i^R)-\frac{z_i^R}{s_R}\right)^2\right)^{1/2},\left(Q\sum_{j\in[Q]}\left(\mu(V_j^C)-\frac{z_j^C}{s_C}\right)^2\right)^{1/2}\right\}.
\end{aligned}$$

Therefore, we have that

$$(1-\mu(Y))^2 \le \max\left\{P\sum_{i\in[P]}\left(\mu(V_i^R)-\frac{z_i^R}{s_R}\right)^2, Q\sum_{j\in[Q]}\left(\mu(V_j^C)-\frac{z_j^C}{s_C}\right)^2\right\}.$$

Taking expectation (over the choice of S_R and S_C) yields,

$$\mathop{\mathbf{E}}_{S_R,S_C}[1-\mu(Y)] \le \max\left(\sqrt{\frac{Q}{s_C}},\sqrt{\frac{P}{s_R}}\right).$$

Let $\mathcal{U} = \mathcal{A}^{\text{str}} - \mathcal{Y}$ and consider a corresponding matrix U. U is an $N\times M$ matrix, where $N = \text{lcm}(n\cdot\mu(V_1^R\triangle V_1^{R'}),\ldots,n\cdot\mu(V_P^R\triangle V_P^{R'}))$ and $M = \text{lcm}(m\cdot\mu(V_1^C\triangle V_1^{C'}),\ldots,m\cdot\mu(V_Q^C\triangle V_Q^{C'}))$. By Claim 10, the absolute value of an entry in the matrix A^{str} is bounded by $\frac{2}{\gamma^{11}}L$, and thus the absolute value of an entry in U is bounded by $\frac{4}{\gamma^{11}}L$.

Then,

$$\begin{aligned}
\mathop{\mathbf{E}}_{S_R,S_C}[\|U\|_2^2] &\le \mathop{\mathbf{E}}_{S_R,S_C}[\|U\|_F^2] = \mathop{\mathbf{E}}_{S_R,S_C}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}U_{ij}^2\right]\\
&\le \mathop{\mathbf{E}}_{S_R,S_C}\left[NM(1-\mu(Y))\cdot\max_{(i,j)\in[N]\times[M]}(U)_{ij}^2\right]\\
&\le NM\cdot\max_{(i,j)\in[N]\times[M]}U_{ij}^2\cdot\mathop{\mathbf{E}}_{S_R,S_C}[1-\mu(Y)]\\
&\le NM\cdot\left(\frac{4}{\gamma^{11}}\right)^2L^2\cdot\max\left(\sqrt{\frac{Q}{s_C}},\sqrt{\frac{P}{s_R}}\right).
\end{aligned}$$

Using Corollary 13 we get $\mathop{\mathbf{E}}_{S_R,S_C}[\|\mathcal{U}\|_2] \le \frac{4L}{\gamma^{11}}\max\left(\sqrt{\frac{p^{T/2}}{s_R^{1/2}}},\sqrt{\frac{q^{T/2}}{s_C^{1/2}}}\right)$, and the lemma follows. ◀

In the remaining part of the subsection we will prove Lemma 16. In order to prove the lemma we introduce the following result regarding a random submatrix from [22] section 5.2.2.

▶ **Lemma 19.** *Given a matrix $A\in[-L,L]^{n\times m}$, let $P=\text{diag}(\chi_1,\ldots,\chi_n)$ be the diagonal matrix where $\{\chi_i\}$'s are Bernoulli(k_R/n) random variables for $k_R>0$. In addition, let $R=\text{diag}(\xi_1,\ldots,\xi_m)$ be the diagonal matrix where $\{\xi_j\}$'s are Bernoulli(k_C/m) random variables for $k_C>0$. Then,*

$$\mathbf{E}[\|PAR\|_2^2] \le \frac{3k_R\cdot k_C}{nm}\|A\|_2^2 + 2k_RL^2\log n + 2k_CL^2\log m + L^2\log n\log m.$$

The above lemma shows that a random submatrix of size roughly $k_R \times k_C$ gets its "fair share" of the spectral norm of A.

Proof of Lemma 16: Since S_R and S_C are set of indices generated by choosing each index to S_R (or S_C) with probability k_R/n (k_C/m), we have that with probability at least 99/100,

$$|S_R| \geq k_R/2 \text{ and } |S_C| \geq k_C/2 .$$

We henceforth condition on that.

For any measure preserving bijection $\pi : [0,1] \to [0,1]$, and $t \in [T]$ we have

$$\mathop{\mathbf{E}}_{S_R,S_C}[\|\mathcal{A}_t - \pi(\mathcal{A}_t|_{S_R \times S_C})\|_2] \leq \|\mathcal{A}_t - \mathcal{A}_t^{\text{str}}\|_2 + \mathop{\mathbf{E}}_{S_R,S_C}\left[\|\mathcal{A}_t^{\text{str}} - \pi(\mathcal{A}_t^{\text{str}}|_{S_R \times S_C})\|_2\right] + \mathop{\mathbf{E}}_{S_R,S_C}\left[\|\pi(\mathcal{A}_t^{\text{str}}|_{S_R \times S_C}) - \pi(\mathcal{A}_t|_{S_R \times S_C})\|_2\right] .$$

Where A_t^{str} is the matrix obtained by Lemma 8.

By Lemma 8 and Corollary 13, we have that for any $t \in [T]$

$$\|\mathcal{A}_t - \mathcal{A}_t^{\text{str}}\|_2 \leq 7\gamma L .$$

By the facts that $|S_R| \geq k_R/2$, $|S_C| \geq k_C/2$, $p = O\left(\left(\frac{1}{\gamma^{10}}\right)^{3/\gamma^2}\right)$ and $q = O\left(\left(\frac{1}{\gamma^{10}}\right)^{3/\gamma^2}\right)$, Lemma 15 yields that for any $t \in [T]$:

$$\mathop{\mathbf{E}}_{S_R,S_C}\left[\|\mathcal{A}_t^{\text{str}} - \pi(\mathcal{A}_t^{\text{str}}|_{S_R \times S_C})\|_2\right] = O\left(\frac{L}{\gamma^{11}}\left(\frac{1}{\gamma^{10}}\right)^{\frac{3T}{4\gamma^2}} \cdot \max\left(\frac{1}{k_R^{1/4}}, \frac{1}{k_C^{1/4}}\right)\right)$$

We are left with bounding $\mathop{\mathbf{E}}_{S_R,S_C}[\|\pi(\mathcal{A}_t^{\text{str}}|_{S_R \times S_C}) - \pi(\mathcal{A}_t|_{S_R \times S_C})\|_2]$. We apply Lemma 19 on $(A_t^{\text{str}} - A_t)|_{S_R \times S_C}$ to get

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S_R,S_C}\left[\|(A_t^{\text{str}} - A_t)|_{S_R \times S_C}\|_2^2\right] \\
&\leq \frac{3k_C \cdot k_R}{nm}\|A^{\text{str}} - A\|_2^2 + 2L^2 k_R \log n + 2L^2 k_C \log m + L^2 \log n \log m \\
&\leq \frac{3k_C \cdot k_R}{nm} \cdot 16\gamma^2 L^2 nm + 2L^2 k_R \log n + 2L^2 k_C \log m + L^2 \log n \log m \\
&\leq 48L^2\gamma^2 k_R \cdot k_C + 2L^2 k_R \log n + 2L^2 k_C \log m + L^2 \log n \log m ,
\end{aligned}$$

which implies that,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S_R,S_C}\left[\|(A_t^{\text{str}} - A_t)|_{S_R \times S_C}\|_2\right] \\
&\leq 7\gamma L\sqrt{k_C \cdot k_R} + L\sqrt{2k_R \log n} + L\sqrt{2k_C \log m} + L\sqrt{\log m \log n} .
\end{aligned}$$

By applying Corollary 13 and using the fact that the dimension of $(A^{\text{str}} - A)|_{S_R \times S_C}$ at least $k_R \cdot k_C/4$, we get

$$\mathop{\mathbf{E}}_{S_R,S_C}\left[\|\mathcal{A}_t^{\text{str}}|_{S_R \times S_C} - \mathcal{A}_t|_{S_R \times S_C}\|_2\right] \leq 14\gamma L + L\sqrt{\frac{8\log n}{k_c}} + L\sqrt{\frac{8\log m}{k_R}} + L\sqrt{\frac{4\log m \log n}{k_R \cdot k_C}} .$$

Putting everything together,

$$\mathop{\mathbf{E}}_{S_R,S_C}[\|\mathcal{A}_t - \pi(\mathcal{A}_t|_{S_R \times S_C})\|_2] \leq 21\gamma L + L\left(\sqrt{\frac{8\log n}{k_C}} + \sqrt{\frac{8\log m}{k_R}} + \sqrt{\frac{4\log n \log m}{k_R k_C}}\right) + O\left(\frac{L}{\gamma^{11}}\left(\frac{1}{\gamma^{10}}\right)^{\frac{3T}{4\gamma^2}} \cdot \max\left(\frac{1}{k_R^{1/4}}, \frac{1}{k_C^{1/4}}\right)\right) .$$

By Markov inequality, with probability at least $9/10T$,

$$\|\mathcal{A}_t - \pi(\mathcal{A}_t|_{S_R \times S_C})\|_2 \leq 210\gamma LT + 10LT\left(\sqrt{\frac{8\log n}{k_C}} + \sqrt{\frac{8\log m}{k_R}} + \sqrt{\frac{4\log n \log m}{k_R k_C}}\right) + O\left(\frac{LT}{\gamma^{11}}\left(\frac{1}{\gamma^{10}}\right)^{\frac{3T}{4\gamma^2}} \cdot \max\left(\frac{1}{k_R^{1/4}}, \frac{1}{k_C^{1/4}}\right)\right) .$$

By using a union bound the lemma follows. ◀

A.2 Quadratic Function Minimization

Proof of Lemma 17: In contrast to the proof of Lemma 12, in this case we have to deal with constrained optimization, and therefore must consider the KKT optimality conditions. We start by showing that $\min_{\boldsymbol{v}:\|\boldsymbol{v}\|_2 \leq r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \geq n^2 \cdot \inf_{f:\|f\|_2 \leq \frac{r}{\sqrt{n}}} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f)$. Given a vector $\boldsymbol{v} \in \mathbb{R}^n$ such that $\|\boldsymbol{v}\|_2 \leq r$, we define the function $f : [0,1] \to \mathbb{R}$ as $f(x) = v_{i^n(x)}$. Then,

$$\langle f, \mathcal{A}f \rangle = \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} A_{ij} f(x) f(y) dx dy = \frac{1}{n^2} \langle \boldsymbol{v}, A\boldsymbol{v} \rangle$$

$$\left\langle f^2, \mathcal{D}\mathbf{1} \right\rangle = \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} d_i f(x)^2 dx dy = \frac{1}{n} \langle \boldsymbol{v}, \mathrm{diag}(\boldsymbol{d})\boldsymbol{v} \rangle$$

$$\langle f, \mathcal{B}\mathbf{1} \rangle = \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} b_i f(x) dx dy = \frac{1}{n} \langle \boldsymbol{v}, \boldsymbol{b} \rangle$$

In addition,

$$\|f\|_2^2 = \int_0^1 f(x)^2 dx = \sum_{j \in [n]} \int_{I_j^n} f(x)^2 dx = \frac{1}{n} \sum_{j \in [n]} v_j^2 = \frac{1}{n}\|\boldsymbol{v}\|_2^2 \leq \frac{r^2}{n} .$$

Next, we show that $\min_{\boldsymbol{v}:\|\boldsymbol{v}\|_2 \leq r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \leq n^2 \cdot \inf_{f:\|f\|_2 \leq \frac{r}{\sqrt{n}}} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f)$. First, we note that the latter problem has a minimizer $f : [0,1] \to \mathbb{R}$ because it is weakly continuous and coercive (See, e.g., [21]). According to the generalized KKT conditions (see, e.g., Section 9.4 of [15]), there exists λ such that:

- (Stationarity) $\frac{\partial}{\partial f^*(x)} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f^*(x)) - \lambda \frac{\partial}{\partial f^*(x)}(\|f^*\|_2 - r/\sqrt{n}) = 0$ for almost all x.
- (Primal feasibility) $\|f^*\|_2 - r/\sqrt{n} \leq 0$
- (Complementary slackness) $\lambda \cdot (\|f^*\|_2 - r/\sqrt{n}) = 0$

The stationarity condition yields:

$$\frac{\partial}{\partial f^*(x)} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f^*(x)) - \lambda \frac{\partial}{\partial f^*(x)}(\|f^*\|_2 - r/\sqrt{n})$$
$$= \sum_{i \in n} \int_{I_i^n} A_{ii^n(x)} f^*(y) dy + \sum_{j \in n} \int_{I_j^n} A_{i^n(x)j} f^*(y) dy + 2d_{i^n(x)} f^*(x) + b_{i^n(x)} - 2\lambda f^*(x) ,$$

By the form of the partial derivatives, $f^*(z) = f^*(z')$ for almost all $z, z' \in [0,1]$ such that $i^n(z) = i^n(z')$. That is, f^* is almost constant on each of the intervals $I_1^n, \ldots, I_n^n$. Therefore, we define $\boldsymbol{v} \in \mathbb{R}^n$ as $v_j = f^*(x)$, where x is the dominant value in I_j^n. Then,

$$\langle \boldsymbol{v}, A\boldsymbol{v} \rangle = \sum_{i,j \in [n]} A_{ij} v_i v_j = n^2 \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} A_{ij} f^*(x) f^*(y) dx dy = n^2 \langle f^*, \mathcal{A} f^* \rangle \ .$$

$$\langle \boldsymbol{v}, \mathrm{diag}(\boldsymbol{d}) \boldsymbol{v} \rangle = \sum_{i \in [n]} d_i v_i^2 = n \sum_{i \in [n]} \int_{I_i^n} d_i f^*(x)^2 dx = n \left\langle (f^*)^2, \mathcal{D}\mathbf{1} \right\rangle \ .$$

$$\langle \boldsymbol{v}, \boldsymbol{b} \rangle = \sum_{i \in [n]} b_i v_i = n \sum_{i \in [n]} \int_{I_i^n} b_i f^*(x) dx = n \langle f^*, \mathcal{B}\mathbf{1} \rangle \, .$$

In addition, $\|f^*\|_2^2 = \int_0^1 f^*(x)^2 dx = \sum_{j \in [n]} \int_{I_j^n} f^*(x)^2 dx = \frac{1}{n} \|\boldsymbol{v}\|_2^2 \leq \frac{r^2}{n}$.
Hence, we get that

$$\min_{\boldsymbol{v} : \|\boldsymbol{v}\|_2 \leq r} \psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(\boldsymbol{v}) \leq n^2 \cdot \inf_{f : \|f\|_2 \leq \frac{r}{\sqrt{n}}} \Psi_{n,A,\boldsymbol{d},\boldsymbol{b}}(f) \ ,$$

and the lemma follows. ◀