An $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$-Cost Algorithm for Semidefinite Programs with Diagonal Constraints

Yin Tat Lee YINTAT@UW.EDU
University of Washington, Seattle, USA

Swati Padmanabhan PSWATI@UW.EDU
University of Washington, Seattle, USA

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We provide a first-order algorithm for semidefinite programs (SDPs) with diagonal constraints on the matrix variable. Our algorithm outputs an ε-optimal solution with a run time of $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$, where m is the number of non-zero entries in the cost matrix. This improves upon the previous best run time of $\widetilde{\mathcal{O}}(m/\varepsilon^{4.5})$ by Arora and Kale (2007). As a corollary of our result, given an instance of the Max-Cut problem with n vertices and $m \gg n$ edges, our algorithm returns a $(1-\varepsilon)\alpha_{GW}$ cut in the faster time of $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$, where $\alpha_{GW} \approx 0.878567$ is the approximation ratio by Goemans and Williamson (1995). Our key technical contribution is to combine an approximate variant of the Arora-Kale framework of mirror descent for SDPs with the idea of trading off exact computations in every iteration for variance-reduced estimations in most iterations, only periodically resetting the accumulated error with exact computations. This idea, along with the constructed estimator, are of possible independent interest for other problems that use the mirror descent framework.

1. Introduction

Consider the SDP maximizing $C \bullet X \stackrel{\text{def}}{=} \operatorname{Tr}(CX)$ over the set of $n \times n$ positive semidefinite matrices with every diagonal entry bounded by a constant:

$$\max C \bullet X \text{ subject to } X \succeq 0, X_{ii} \leq 1 \text{ for all } i \in [n]. \tag{1.1}$$

We seek a matrix $\widetilde{X}^* \succeq 0$ with $\widetilde{X}^*_{ii} \leq 1$ for all $i \in [n]$ satisfying $C \bullet \widetilde{X}^* \geq C \bullet X^* - \varepsilon \sum_{i,j} |C_{ij}|$, where X^* is an optimal solution of (1.1). This is not an ε-multiplicative guarantee ($C \bullet \widetilde{X}^* \geq C \bullet X^*(1-\varepsilon)$), but a slightly weaker one, since $\sum_{i,j} |C_{ij}| \geq C \bullet X^*$. A multiplicative guarantee is not always easy to provide; indeed, many classical optimization algorithms also provide a guarantee only additive in some quantity that bounds from above the difference of the function values between the initial and optimal points. For example, gradient descent on an L-smooth convex function f over a set with diameter D returns, after k iterations, a point x_k such that $f(x_k) - f(x^*) \leq O(LD^2 k^{-1})$, where $f(x_0) - f(x^*) \leq O(LD^2)$.

To solve (1.1) as per the above accuracy criterion, it suffices to solve (1.2):

$$\min f(X) \stackrel{\text{def}}{=} -\widehat{C} \bullet X + \sum_{i=1}^{n} (X_{ii} - \rho_i)^+, \text{ subject to } X \succeq 0. \tag{1.2}$$

This problem is derived from (1.1) by promoting the diagonal constraints to the objective and appropriately scaling C to $\widehat{C} \overset{\text{def}}{=} \mathbf{diag}(1/\sqrt{\rho})C\,\mathbf{diag}(1/\sqrt{\rho})$, where $\rho \in \mathbb{R}^n$ such that $\rho_i = \sum_{j\in[n]} |C_{ij}|$. By rescaling $C_{ij} = nC_{ij}/\sum_{i,j}|C_{ij}|$, we assume $\sum_{i\in[n]} \rho_i = n$. Lemma 2 gives a solution of (1.1) from a solution of (1.2).

For (1.1), Arora and Kale (2007) have the previous best run time linear in $m \overset{\text{def}}{=} \mathbf{nnz}(C)$, the size of the input. Though there exist algorithms with better dependence on ε, their dependence on n is superlinear, as we describe in Section 1.1. In this paper, we operate in the regime of moderate ε and large n, focusing on first-order methods.

Arora and Kale (2007) use the matrix multiplicative weights (MMW) update, which can be interpreted as mirror descent in the nuclear norm[1], using the negative entropy function, $\Phi(X) = X \bullet \log X$, over the scaled simplex, $\mathcal{D} = \{X : X \succeq 0, \operatorname{Tr} X = n\}$, as the mirror map. Their iterates at iteration t are given by

$$X^{(t)} = n\frac{\exp\left(Y^{(t)}\right)}{\operatorname{Tr}\exp\left(Y^{(t)}\right)}, \quad \text{where } Y^{(t)} = \sum_{s=1}^{t-1} -\eta\nabla f(X^{(s)}), \tag{1.3}$$

with step size $\eta = \mathcal{O}(\varepsilon)$ and gradient $\nabla f(M) = \mathbf{diag}(\mathbf{1}_{M\geq\rho}) - \widehat{C}$. Computing this gradient entails only comparing the diagonal entries of the current iterate with a fixed vector. Therefore, the naïve computational cost of this method is dominated by $\Omega(n^\omega)$ for the matrix exponentiation (Pan and Chen, 1999), prohibitively expensive for a large problem dimension. Arora and Kale (2007) circumvent this by *approximating* the diagonal entries of the matrix exponential. Therefore, their overall cost is composed of the following three parts: (1) mirror descent requiring $\mathcal{O}(1/\varepsilon^2)$ iterations to converge, (2) degree $\mathcal{O}(1/\varepsilon)$ Taylor approximation of the matrix exponential, each matrix-vector product costing $\mathcal{O}(m)$, and (3) $\mathcal{O}(1/\varepsilon^2)$ random projections (Johnson and Lindenstrauss, 1984) to estimate the diagonal entries of the matrix exponential; combined, these give a run time of $\widetilde{\mathcal{O}}(m/\varepsilon^5)$, which, Allen-Zhu and Li (2017) observe, can be sped up to $\mathcal{O}(m/\varepsilon^{4.5})$ by using Chebyshev (instead of Taylor) approximation of matrix exponentials (see (Sachdeva et al., 2014)).

Our contribution. In this work, we solve (1.1) with a run time of $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$, thus speeding up the current best run time for this problem. Our result (formally stated in Theorem 3) is effected by careful technical work that incorporates variance-reduced estimators and fast products of matrix exponentials with vectors into the Arora-Kale framework of mirror descent for SDPs. We use the generalized negative entropy, $\Phi(X) = X \bullet \log(X) - \operatorname{Tr} X$, as our mirror map, and our primary high-level idea is the following: *instead of exactly computing the primal iterate in each iteration, we frequently approximate it at a low accuracy (to reduce the cost) and infrequently at a high accuracy (to "reset" the error resulting from approximation).* This idea is inspired by recent variance-reduction methods (Shalev-Shwartz and Zhang, 2013; Johnson and Zhang, 2013; Defazio et al., 2014; Hazan and Luo, 2016; Schmidt et al., 2017). The periodic high-accuracy computations and small bias and variance of estimators in the low-accuracy computations ensure sufficient closeness, in the appropriate norm, of the estimated iterates to the true ones, which, by the convergence guarantee of approximate mirror descent, leads to an ε-optimal solution. Making this variance-reduction work in the MMW setting requires several technical ideas, as follows.

[1] The nuclear norm of a matrix $X \in \mathbb{R}^{m\times n}$ is the sum of its singular values: $\|X\|_{\text{nuc}} \overset{\text{def}}{=} \sum_{i=1}^{\min(m,n)} \sigma_i(X)$.

We introduce the technical idea of expanding the domain of our mirror map by a polylogarithmic factor. Due to the expanded domain and our choice of the mirror map, the gradient step of mirror descent falls in the *interior* of this domain. The upshot of this is that the primal iterate is related to the dual via simply a matrix exponential, with no trace normalization as in Equation 1.3. Thus, the quantity for which we require an estimator is greatly simplified. Drawing on the observation of Arora and Kale (2007) that the gradient uses only the diagonal entries of the primal iterate, we build an estimator, with a small bias and variance, for the change in diagonal entries of the (dual) matrix exponential. We also prove the strong convexity parameter of our mirror map on the expanded domain by confecting classical results from convex analysis in a novel way. Due to the ubiquity of the MMW framework in optimization, efficient algorithms for SDPs, balanced separators, Ramanujan sparsifiers, packing/covering, and machine learning, we anticipate that our technical contributions will be useful for problems that hinge on the MMW foundation.

Applications. When C is a graph Laplacian, (1.1) is the SDP relaxation of the Max-Cut problem (Goemans and Williamson, 1995). An NP-complete problem (Karp, 1972), Max-Cut has seen widespread utility in circuit design (Chen et al., 1983), statistical physics (Barahona et al., 1988), semi-supervised learning (Wang et al., 2013), and phase recovery (Waldspurger et al., 2015). Another instance of (1.1) is max-norm regularization (Jaggi, 2011), a convex surrogate for rank minimization (Srebro and Shraibman, 2005) enforcing simplicity in modeling observations (Fazel et al., 2004). SDPs of the form of (1.1) have also found applications in community detection (Abbe et al., 2015; Guédon and Vershynin, 2016; Montanari and Sen, 2016b) and as relaxations to the maximum-likelihood estimator in the group synchronization problem (Singer and Shkolnisky, 2011; Bandeira et al., 2014).

1.1. Related work

We describe in this section previous work on (1.1) using first-order methods, other than that of Arora and Kale (2007). Of note is that most papers below solve problems more general than (1.1), and the run times we mention occur when specialized to (1.1).

Saddle-point formulation. Since any SDP can be instantiated as an online convex optimization problem, we apply to our setting some notable results from this area. To do so, we first reduce (1.1) to a feasibility problem following the approach of Arora et al. (2005). Recall our assumption that $\sum_{i,j}|C_{ij}| = n$. The facts $X^* \succeq 0$ and $X^*_{ii} \leq 1$ for $i \in [n]$ imply $|X^*_{ij}|^2 \leq X^*_{ii}X^*_{jj} \leq 1$, which in turn bounds the optimum from above as $OPT = \sum_{i,j} C_{ij}X^*_{ij} \leq \sum_{i,j}|C_{ij}||X^*_{ij}| \leq n$. We can also bound the optimum from below by choosing X to be the zero matrix, thus bounding OPT with $\lambda \in [0, n]$. Let $A_0 = \frac{1}{\lambda}C$, $b_0 = 1$, $A_i = -e_ie_i^\top$, and $b_i = -1$ for $i \in [n]$. Therefore, solving (1.1) requires, for each guess of λ (obtained via a binary search over its range), solving the feasibility problem:

$$\text{Find } Z \in \mathbb{S}^n_{\geq 0} \text{ subject to } A_i \bullet Z - b_i \geq 0, \text{ for all } i \in \{0, n\}, \operatorname{Tr} Z \leq n. \tag{1.4}$$

To do so, we leverage the saddle-point problem studied by (Garber and Hazan, 2016),

$$\max_{X \in \mathbb{S}^n_{\geq 0}, \operatorname{Tr} X = 1} \ \min_{p \in \mathbb{R}^m_{\geq 0}, \|p\|_1 = 1} \sum_{i=1}^m p_i(A_i \bullet X - b_i). \tag{1.5}$$

If the optimum of (1.5) is non-negative, solving it up to an additive accuracy of ε is equivalent to finding a solution in the spectrahedron that satisfies all $A_i \bullet X - b_i \geq 0$ upto an additive error of ε. For (1.4), this means the solution for (1.5) satisfies $X_{ii} \approx 1/n \pm \varepsilon$. However, due to the requirement of $X_{ii} \approx 1 \pm \varepsilon$ in (1.1), the accuracy parameter of (1.5) must be ε/n. This causes the run time of Garber and Hazan (2016) for (1.1) to be $\widetilde{\mathcal{O}}(m(n/\varepsilon)^{2.5})$. By the same reasoning, when solving (1.1) to ε multiplicative accuracy, the work of Baes et al. (2013), which uses a randomized Mirror-Prox algorithm, incurs a cost of $\widetilde{\mathcal{O}}(n^5/\varepsilon^3)$, and the recent algorithms of Follow the Compressed Leader by Allen-Zhu and Li (2017) and rank-1 sketch by Carmon et al. (2019) incur a cost of $\widetilde{\mathcal{O}}(m(n/\varepsilon)^{2.5})$. It must be noted that Garber and Hazan (2016), Allen-Zhu and Li (2017), and Carmon et al. (2019) provide algorithms satisfying ε-additive accuracy. When we translate our accuracy results to their language, the costs are not quite comparable. For instance, Carmon et al. (2019), for ε-additive accuracy for (1.1), incurs a cost of $m(n\|C\|_\infty/\varepsilon)^{2.5}$. Our algorithm, using this accuracy criterion, incurs a cost of $m(\sum_{i,j}|C_{ij}|/\varepsilon)^{3.5}$. Unless we assume additional structure on the matrix C, the comparison between these two costs is inconclusive.

Covering SDP. When $C \succeq 0$, (1.1) is a *covering* SDP:

$$\begin{aligned} \max \langle C, X\rangle \ \text{ subject to } X \in \mathbb{S}^n_{\geq 0}, \langle A_i, X\rangle \leq b_i \text{ for all } i \in [m], \\ \text{for } \{A_i\}_{i\in[d]}, C \succeq 0, b \geq 0. \end{aligned} \tag{1.6}$$

Covering SDPs constitute a class of positive SDPs that, until recently, no positive SDP solver (Peng and Tangwongsan, 2012; Jain and Yao, 2011; Allen Zhu et al., 2016) could provide efficient, width-independent algorithms for, due to the non-commutativity of matrices in general and non-monotonicity of the matrix exponential. A recent result (Jambulapati et al., 2020) breaks this barrier; for (1.1), their cost is $\widetilde{\mathcal{O}}(m/\varepsilon^{-5})$, thus still lower than ours.

Low-rank updates. When C is the graph Laplacian in (1.1), there exists an ε-accurate solution of rank $\mathcal{O}(1/\varepsilon)$ (Raghavendra and Steurer, 2009; Montanari and Sen, 2016a; Mei et al., 2017). Many papers capitalize on this fact and perform low-rank updates, which reduces cost per iteration. For example, Klein and Lu (1996) base their algorithm on the framework of Plotkin et al. (1991) in conjunction with the power method to achieve a run time of $\widetilde{\mathcal{O}}(mn/\varepsilon^3)$. As another example, Hazan (2008) incorporates into the Frank-Wolfe algorithm (Frank and Wolfe, 1956) fast computation of an approximate minimum eigenvector and provides an $\widetilde{\mathcal{O}}(mn^3/\varepsilon^3)$-algorithm. A recent noteworthy result (Yurtsever et al., 2019) returns a rank-R approximation to an ε-optimal solution at a cost $\widetilde{\mathcal{O}}(R/\varepsilon^2 + n/\varepsilon^3)$. Even though, as alluded to earlier, there exists a rank-$\mathcal{O}(1/\varepsilon)$ solution to the MaxCut SDP, perturbing such a solution by an appropriately small amount gives an ε-optimal solution that is in fact full rank. Indeed, per Theorem 6.2 of Yurtsever et al. (2019), for any $r < R$, the iterate $\widehat{X}_t$ returned by their algorithm in iteration t satisfies $\limsup_{t\to\infty} \mathbf{E}_\Omega \operatorname{dist}_*(\widehat{X}_t, \Psi_*) \leq (1 + r/(R-r-1)) \cdot \max_{X\in\Psi_*} \|X - [X]_r\|_*$, where Ω is the randomness in their algorithm, Ψ_* is the solution set, R is the rank of the iterate returned, and $[X]_r$ is an r-truncated singular value decomposition of matrix X. The existence of full-rank matrices in the solution set Ψ^* implies a possibly large bound above, so one cannot conclude that Yurtsever et al. (2019) improves upon our run time.

Polynomial mirror map. One of the contributions of Allen-Zhu and Li (2017) is a "polynomial-style" mirror map such as $\Phi(X) = \frac{1}{1+1/2p} \operatorname{Tr} X^{1+1/2p}$. The projection step with this map is $X = (Y^+)^{2p}$, where Y^+ is the matrix obtained by zeroing out the negative eigenvalues of Y, which is as expensive as matrix exponentiation.

Variance-reduction methods. Standard variance reduction algorithms such as SVRG (Johnson and Zhang, 2013) minimize an objective that is a sum of functions, employing an unbiased estimator of the gradient. Unfortunately, neither is (1.2) a sum of functions, nor is its gradient ($\mathbf{diag}(\mathbf{1}_{X>=\rho})$) cheap to estimate.

1.2. Preliminaries

Notation. We use $\mathbb{R}^n$ to denote the subspace of n-dimensional real vectors, $\mathbf{1}$ for the vector of all ones, and $\mathbf{1}_{\{\mathcal{E}\}}$ for the all-zero vector with one at coordinates where $\mathcal{E}$ is true. We use x^+ to denote the non-smooth function x when $x \geq 0$ and zero otherwise. Denote by $\mathbb{S}^n$ the subspace of $n \times n$ symmetric matrices and by I_n the $n \times n$ identity matrix. For $u \in \mathbb{R}^n$, $\mathbf{diag}(u)$ is the $n \times n$ diagonal matrix with $\mathbf{diag}(u)_{ii} = u_i$. For $A, B \in \mathbb{S}^n$, the trace inner product is $A \bullet B \stackrel{\text{def}}{=} \operatorname{Tr}(AB) = \sum_{i,j} A_{ij}B_{ij}$. We define $|\!|\!|A|\!|\!| = \sum_i |A_{ii}|$. Given a scalar function f and a vector u, we use $f(u)$ to mean that entrywise, and similarly, for a symmetric matrix $A = U\Lambda U^\top$, $f(A) = Uf(\Lambda)U^\top$. Given $A \in \mathbb{R}^{n\times n}$ and $p \in \mathbb{R}^n$, $A \geq p$ means $A_{ii} \geq p_i$ for all $i \in [n]$. For $u \in \mathbb{R}^n$, $N \in \mathbb{N}$, and vectors $\zeta_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_n)$ for $k \in [N]$, the scalar $v = \mathbf{RandProj}(u, N) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{k=1}^N (u^T\zeta_k)^2$. This implies $\mathbf{E}\, v = \|u\|_2^2$. We extend the definition to $A \in \mathbb{S}^n$ with each row of A as the vector u. Then the diagonal matrix $B = \mathbf{RandProj}(A, N)$ satisfies $\mathbf{E}\, B = \mathbf{diag}\, A^2$. We use $\widetilde{\mathcal{O}}$ to denote polylogarithmic factors. The superscript * denotes optimality for variables and Fenchel conjugate for functions.

Fact 0.1 (Allen Zhu et al. (2016)) *Given $A \succeq 0$, $B \in \mathbb{S}^n$, and $\alpha \in [0,1]$, the inequality* $\operatorname{Tr}\big(BA^\alpha BA^{1-\alpha}\big) \leq A \bullet B^2$ *holds.*

Fact 0.2 (Wilcox (1967)) *For a symmetric matrix-valued function $X(t)$ with argument scalar t, we have $\frac{d}{dt}\exp(X(t)) = \int_{\alpha=0}^1 \exp(\alpha X(t))\frac{d}{dt}X(t)\exp((1-\alpha)X(t))d\alpha$.*

Setup. Our underlying algorithm to solve (1.2) is a slight variant of lazy mirror descent (also called Nesterov's Dual Averaging Nesterov (2009)), which we term *approximate lazy mirror descent.* To solve $\min_{x\in\mathcal{X}} f(x)$ using this algorithm, select a mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ and a norm; the associated Bregman Divergence is $\mathcal{D}_\Phi(x, y) \stackrel{\text{def}}{=} \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle$; set $x^{(1)} \in \operatorname{argmin}_{\mathcal{X}\cap\mathcal{D}} \Phi(x)$ and $z^{(1)} \in \nabla^{-1}\Phi(0)$. We repeat, in succession, the gradient update, $\nabla\Phi(z^{(t+1)}) = \nabla\Phi(z^{(t)}) - \eta\nabla f(x^{(t)})$, and the *approximate* projection, finding $\widetilde{x}^{(t+1)}$ satisfying $\mathbf{E}\,\|\widetilde{x}^{(t+1)} - x^{(t+1)}\| \leq \delta$, where $x^{(t+1)} \in \operatorname{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} \mathcal{D}_\Phi(x, z^{(t+1)})$.

Theorem 1 (Convergence of Lazy Mirror Descent) *Fix a norm $\|\cdot\|$. Given an α-strongly convex mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ and a convex, G-Lipschitz objective $f : \mathcal{X} \to \mathbb{R}$, run Algorithm 3 with step size η and $\mathbf{E}\,\|x^{(t)} - \widetilde{x}^{(t)}\| \leq \delta$. Let $D \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x) - \inf_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x)$. Then, Algorithm 3 after T iterations returns $\widetilde{x}^{t^*}$, satisfying*

$$\mathbf{E}\, f(\widetilde{x}^{(t^*)}) - f(x^*) \leq \frac{D}{T\eta} + \frac{2\eta G^2}{\alpha} + \delta G. \tag{1.7}$$

Lemma 2 *Given $C \in \mathbb{R}^{n\times n}$ and $0 \preceq X$, let $\rho \in \mathbb{R}^n$ with $\rho_i = \sum_{j=1}^n |C_{ij}|$; diagonal matrix S with $S_{ii} = \min(1/\sqrt{\rho_i}, 1/\sqrt{X_{ii}})$ for $i \in [n]$; $\widehat{X} = SXS$; $\widehat{C} = \mathbf{diag}\,(1/\sqrt{\rho})\, C\, \mathbf{diag}(1/\sqrt{\rho})$. Then, $\widehat{X} \succeq 0$, $\widehat{X}_{ii} \le 1$ for all $i \in [n]$, and $\widehat{C} \bullet X - \sum_{i=1}^n (X_{ii} - \rho_i)^+ \le C \bullet \widehat{X}$.*

2. Our approach

We present our algorithm below, parameters in Table 1, and main result in Theorem 3.

Algorithm 1 Our Algorithm

```
Input: Cost matrix $C \in \mathbb{R}^{n\times n}$, accuracy $\varepsilon$
Parameters: Displayed in Table 1
Initialize $t \leftarrow 0$, $Y^{(1)} \leftarrow \mathbf{0}$. Set $\widehat{C}$ and $\rho$ from Lemma 2 and $\nabla f(X) = \mathbf{diag}(\mathbf{1}_{X \ge \rho}) - \widehat{C}$
for $T_{outer}$ iterations do
    $t \leftarrow t + 1$
    $\widetilde{\exp}(\frac{1}{2}Y^{(t)}) \leftarrow$ ChebyExp$(\frac{1}{2}Y^{(t)}, \mathrm{T}_{\mathrm{Cheby}}, \delta_{\mathrm{Cheby}})$          ▷ Defined in Corollary 34
    $\widetilde{X}^{(t)} \leftarrow$ RandProj$(\widetilde{\exp}(\frac{1}{2}Y^{(t)}), \mathrm{T}_{\mathrm{jl}})$          ▷ High-accuracy projection
    $Y^{(t+1)} \leftarrow Y^{(t)} - \eta \nabla f(\widetilde{X}^{(t)})$          ▷ Gradient update
    for $t_i = 1 \to T_{inner}$ do
        $t \leftarrow t + 1$
        $\widehat{\theta}^{(t_i)} \leftarrow$ UpdateEstimator$(\widetilde{X}^{(t-1)}, Y^{(t-1)}, \varepsilon, \eta)$          ▷ See Algorithm 2
        $\widetilde{X}^{(t)}_{jj} \leftarrow (\sqrt{\widetilde{X}^{(t-1)}_{jj} + 1} + \widehat{\theta}^{(t_i)}_j)^2 - 1$ for $j \in [n]$          ▷ Constant-accuracy projection
        $Y^{(t+1)} \leftarrow Y^{(t)} - \eta \nabla f(\widetilde{X}^{(t)})$          ▷ Gradient update
    end
end
For $t^* \overset{\text{unif.}}{\sim} \{1, 2, \ldots, t\}$, return $Y^{(t^*)}$ and $S$, where $S$ is from Lemma 2.
```

Parameter	Value	Proof
Diameter D	$K \log K$	Lemma 25
Strong convexity α	$1/(4K)$	Lemma 11
Step size η	$\frac{1}{8\times 10^4 (\log(n/\varepsilon))^{11}} \varepsilon^2$	Lemma 41
Inner iteration count T_{inner}	ε^{-2}	Section 3.4
Outer iteration count T_{outer}	$\frac{1}{\varepsilon} \cdot 24 \times 10^5 (\log(n/\varepsilon))^{11} \log n$	Lemma 10
JL projection count T_{jl}	$(2 \times 10^5) \cdot (\log n)^{21} \cdot \varepsilon^{-2}$	Lemma 41
Chebyshev approximation degree $\mathrm{T}_{\mathrm{Cheby}}$	$150 \log(n/\varepsilon) \cdot \varepsilon^{-1/2}$	Lemma 36
Chebyshev approximation accuracy δ_{Cheby}	$(\varepsilon/n)^{401}$	Lemma 36

Table 1: All Algorithm 1 parameters and where their values are set. $K = 40n(\log n)^{10}$.

Theorem 3 (Main Result) *Given $C \in \mathbb{R}^{n\times n}$ with $m \ge n$ non-zero entries and $0 < \varepsilon \le \frac{1}{2}$, we can find, in time $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$ and with high probability, a matrix $Y \in \mathbb{S}^n$ with $\mathcal{O}(m)$ non-zero entries and a diagonal matrix $S \in \mathbb{R}^{n\times n}$ so that[2] $\widetilde{X}^* \overset{\text{def}}{=} S \cdot \exp Y \cdot S$ satisfies $\widetilde{X}^* \succeq 0$, $\widetilde{X}^*_{ii} \le 1$ for $i \in [n]$, and $C \bullet \widetilde{X}^* \ge C \bullet X^* - \varepsilon \sum_{i,j} |C_{ij}|$.*

[2] Since $\widetilde{X}^*$ can be dense, we represent it implicitly by only returning the matrices Y and S.

As a corollary, for the Max-Cut problem on a graph with n nodes and m edges, our algorithm gives a cut that is $(1-\varepsilon)\alpha_{GW}$ optimal[3], in time $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$, where $\alpha_{GW} \approx 0.878567$. Before proceeding to the proof sketch of Theorem 3, we call attention to a technical concept crucial to our analysis: *we add to (1.2) the constraint* $\operatorname{Tr} X \leq K$, *where* $K = 40n(\log n)^{10}$. The optimal X^* remains valid under this constraint because $\operatorname{Tr} X^* = n$. Throughout our algorithm, this inequality remains inactive (Lemma 10). Coupled with the Legendre dual of our mirror map $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$, this results in the primal and the dual being related by $X = \exp(Y)$ (Lemma 42). Since the gradient requires only diagonal entries of the primal iterate, we need estimators only for the diagonal entries of $\exp(Y)$.

Proof [Proof Sketch of Theorem 3] In this proof sketch, we compute the run time of Algorithm 1, proving the claims in Theorem 3. In doing so, we provide intuition for the choice of parameters in Table 1. This sketch assumes that we are in iteration t and drops all superscripts, and aside from that, follows the notation of Algorithm 1.

1. To compute $\exp(Y)_{ii}$, we first approximate $\widetilde{\exp}(Y/2)$ to ε-accuracy using Chebyshev polynomials. We show in Lemma 35 that the spectrum of Y lies in the range $[-\mathcal{O}(1/\varepsilon), \widetilde{\mathcal{O}}(1)]$, which allows for Chebyshev approximation with $\widetilde{\mathcal{O}}(1/\sqrt{\varepsilon})$ terms, thus giving the cost of each projection to be $\widetilde{\mathcal{O}}(m/\sqrt{\varepsilon})$. The upper bound of $\widetilde{\mathcal{O}}(1)$ on the spectrum is critical to getting this cost, for in case of a symmetric range of $[-\mathcal{O}(1/\varepsilon), \mathcal{O}(1/\varepsilon)]$, the cost would be $\widetilde{\mathcal{O}}(1/\varepsilon)$. The $\widetilde{\mathcal{O}}(1/\sqrt{\varepsilon})$ terms is in contrast with the $\mathcal{O}(1/\varepsilon)$ required for Taylor approximation. We then estimate each $\exp(Y)_{ii}$ with $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ projections via the JL sketch in the high-accuracy steps, and $\widetilde{\mathcal{O}}(1)$ randomized projections in the T_{inner} low-accuracy steps. Therefore the total cost of the algorithm over T_{outer} iterations is roughly $T_{\text{outer}} \cdot (m/\sqrt{\varepsilon}) \cdot (1/\varepsilon^2 + T_{\text{inner}})$. From this expression, the optimal choice of T_{inner} (up to polylogarithmic factors) is $T_{\text{inner}} = 1/\varepsilon^2$.

2. Due to the small bias and variance of our estimator, after T_{inner} inner iterations, the estimated iterate is roughly within εK distance of the true iterate. Thus, the condition in Theorem 1 is satisfied, and its the error bound applies at the end of our algorithm: $\mathbf{E}\, f(\widetilde{X}^*) - f(X^*) \leq D/(T\eta) + 2\eta G^2/\alpha + \delta G$. Using D, G, and α from Table 1 and T_{inner} from Step 1 and bounding by εK, this inequality simplifies to $\varepsilon^2/(\eta T_{\text{outer}}) + \eta \leq \varepsilon$.

3. The step size η is chosen by studying the error generated in each estimation step versus the error our framework can tolerate. Estimating $(\exp(Y+\Delta))_{ii}$ from $(\exp Y)_{ii}$ via a first-order approximation accrues an error of $\operatorname{Tr}(\Delta \exp Y)$. Applying Hölder's inequality, the value of G, and the trace bound enforced by Lemma 10 yields $\operatorname{Tr}(\Delta \exp Y) \leq \eta K$. Therefore, after T_{inner} iterations, the variance of the error is $T_{\text{inner}} \eta^2 K^2$. Equivalently, the overall error after T_{inner} iterations is $\sqrt{T_{\text{inner}}}\eta K$. For this to be bounded by εK, we must have $\eta \leq \varepsilon/\sqrt{T_{\text{inner}}}$. Plugging in T_{inner} from Step 1 gives the step size: $\eta \approx \varepsilon^2$.

4. The value of η from Step 3 and the inequality from Step 2 give $T_{\text{outer}} \approx 1/\varepsilon$. Plugging this value of T_{outer} above gives the overall algorithm cost $\widetilde{\mathcal{O}}(m/\varepsilon^{3.5})$.

We boost our result to the high probability statement of Theorem 3 over multiple runs of the algorithm. We sidestep the issue of storage cost of $\widetilde{X}^*$ and cost of matrix-matrix

[3] Assuming the Unique Games Conjecture, this is the best we can hope for Max-Cut (Khot et al., 2007).

products by dimension reduction techniques. This finishes the proof of our error guarantee. Lemma 2 implies that $\widetilde{X}^* \succeq 0$ and satisfies the diagonal constraints. ■

	Arora and Kale (2007)	**Algorithm 1 (This Paper)**		
	(Previous Best)	Low accuracy steps	+	High accuracy steps
Number of iterations	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	$\widetilde{\mathcal{O}}(\varepsilon^{-3})$	+	$\widetilde{\mathcal{O}}(\varepsilon^{-1})$
Number of projections per iteration	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	$\widetilde{\mathcal{O}}(1)$	+	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$
Cost per projection	$\mathcal{O}(m\varepsilon^{-1})$	$\widetilde{\mathcal{O}}(m\varepsilon^{-1/2})$	+	$\widetilde{\mathcal{O}}(m\varepsilon^{-1/2})$
Total Cost	$\widetilde{\mathcal{O}}(m\varepsilon^{-5})$	$\widetilde{\mathcal{O}}(m\varepsilon^{-3.5})$	+	$\widetilde{\mathcal{O}}(m\varepsilon^{-3.5})$

Table 2: Comparing Arora and Kale (2007) to our algorithm.

2.1. The estimator

In this section, we consider the t_i'th iteration in the inner loop of Algorithm 1; suppose this is the t'th overall iteration. For now, we drop all superscripts and fix the notation below.

Definition 4 *Let* $\Delta = -\eta\nabla f(X)$, $Y_s = Y + s\Delta$ *for* $s \in [0,1]$, $\bar{\tau} = 1-\tau$, $\delta_{\exp} = \frac{4800\varepsilon^{401}}{n^{390}}$, $\theta_{1_i} = (\exp(Y_s)_{ii} + 1)^{-1/2}$, $\theta_{2_i} = \frac{1}{2}(\exp(\bar{\tau}Y_s)\Delta\exp((\tau-1/2)Y_s)\exp((1/2)Y_s))_{ii}$, $b_{1_i} = \theta_{1_i}(2\delta_{\exp} + \sqrt{2}(1+2\delta_{\exp})(\varepsilon/n)^{400})$, *and* $b_{2_i} = 15\delta_{\exp}\eta K$.

To construct an estimator for the update from $\exp(Y)$ to $\exp(Y+\Delta)$, we estimate the update in $\sqrt{(\exp Y)_{ii}+1}$. The motivation for this choice of function is two-fold: (1) because of the square root, the variance is controlled by the trace of the matrix exponential, bounded by Lemma 10; (2) since the derivative of square root is the inverse square root, we need $\sqrt{\exp(Y)_{ii}+1}$ instead of $\sqrt{\exp(Y)_{ii}}$ to prevent the update term from becoming unbounded. By chain rule, Fact 0.2, and the fundamental theorem of Calculus,

$$\begin{aligned}\sqrt{(\exp(Y+\Delta))_{jj}+1} &= \sqrt{(\exp(Y))_{jj}+1}\\ &+ \underbrace{\int_{s=0}^{1} \underbrace{((\exp Y_s)_{jj}+1)^{-1/2}}_{\overset{\text{def}}{=}\,\theta_{1_j};\text{ estimated using }\widehat{\theta}_{1_j}} \underbrace{\tfrac{1}{2}\Big(\int_{\tau=0}^{1}\exp(\tau Y_s)\Delta\exp(\bar{\tau}Y_s)d\tau\Big)_{jj}}_{\overset{\text{def}}{=}\,\theta_{2_j};\text{ estimated using }\widehat{\theta}_{2_j}}\,ds}_{\overset{\text{def}}{=}\,\theta_j;\text{ estimated using }\widehat{\theta}_j}.\end{aligned} \tag{2.1}$$

As indicated in Equation 2.1, we split the quantity to be estimated into two parts, separately estimating each. Estimating the first part, $\widehat{\theta}_{1_j}$, requires first estimating $\exp(Y_s)_{jj}+1$ using a JL sketch and then passing through the following Taylor approximation for the function $g(u) = u^{-1/2}$, where $g^{(k)}(x)$ is the k'th derivative of g at x,

$$\mathbf{InvSqrt}(\widetilde{X}, N) \overset{\text{def}}{=} \sum_{k=0}^{N-1} \frac{1}{k!} g^{(k)}(x_0) \prod_{j=1}^{k} (x_{k,j} - x_0), \text{ where } x_0, x_{k,j} \overset{\text{i.i.d.}}{\sim} \widetilde{X}. \tag{2.2}$$

Since $\widehat{\theta}_{1_j}$ must be *unbiased*, it is essential to do the Taylor approximation instead of simply evaluating $g(u) = u^{-1/2}$ at the estimator of $\exp(Y_s)_{jj} + 1$. Indeed, for a general f and a random variable $\widetilde{x}$ that is an unbiased estimator of x, $\mathbf{E}\, f(\widetilde{x}) = f(\mathbf{E}\,\widetilde{x})$ does not hold, as evidenced by Jensen's inequality; on the other hand, the intuition for the quantity from Equation 2.2 to be unbiased is that each term in the sum is a product of independent, unbiased random variables. Estimating θ_{2_j} is done by splitting it into carefully chosen parts and applying the JL sketch. Algorithm 2 is the complete subroutine for the estimator.

Algorithm 2 UpdateEstimator(Primal X, dual Y, accuracy ε, step size η)

```
Parameters $\mathrm{T}_{\mathrm{est_{jl}}} = 2^{22}10^4(\log(n/\varepsilon))^2$ and $\mathrm{T}_{\mathrm{est_{isq}}} = 1600\log(n/\varepsilon)$ (set in Lemma 6)
Sample $s$ and $\tau$ uniformly from $[0,1]$. Compute $\Delta$ and $Y_s$ as per Definition 4. Let $\widetilde{X}_s = \mathbf{RandProj}(\widetilde{\exp}(Y_s/2), \mathrm{T}_{\mathrm{est_{jl}}})$. Sample $\zeta \sim \mathcal{N}(0, I_n)$.
Compute $\widehat{\theta}_{1_j} = \mathbf{InvSqrt}(\widetilde{X}_{s_{jj}} + 1, \mathrm{T}_{\mathrm{est_{isq}}})$ for $j \in [n]$.
Compute $\widehat{\theta}_{2_j} = \frac{1}{2}(\widetilde{\exp}((\tau - \frac{1}{2})Y_s)\Delta\widetilde{\exp}(\bar{\tau}Y_s)\zeta)_j\,(\widetilde{\exp}(Y_s/2)\zeta)_j$ for $j \in [n]$.
Return the overall estimator, $\widehat{\theta}_j = \widehat{\theta}_{1_j}\widehat{\theta}_{2_j}$, for $j \in [n]$.    ▷ Coordinate-wise product
```

Properties of the estimator. The bounds on bias and variance of the estimator, as required by Theorem 3, are stated in Lemma 5. Since $\widehat{\theta}$ is constructed from $\widehat{\theta}_1$ and $\widehat{\theta}_2$, we first state their properties and use them to sketch a proof of Lemma 5.

Lemma 5 *The estimator $\widehat{\theta}^{(t)}$ has the following bounds on its first and second moments.*

(1) $|\,\mathbf{E}\,\widehat{\theta}_i - \int_{s=0}^{1}\int_{\tau=0}^{1}\theta_{1_i}\theta_{2_i}dsd\tau| \leq b_{1_i}\theta_{2_i} + b_{2_i}\theta_{1_i} + b_{1_i}b_{2_i}$ *for* $i \in [n]$.

(2) $\mathbf{E}\,\|\widehat{\theta}\|_2^2 \leq 19600\log(n/\varepsilon)K\eta^2 + 147000K^2\eta^2\delta_{\exp}$.

Lemma 6 *Given* $\mathrm{T}_{\mathrm{est_{isq}}} = 1600\log(n/\varepsilon)$, $\mathrm{T}_{\mathrm{est_{jl}}} = 2^{14}\mathrm{T}^2_{\mathrm{est_{isq}}}$, $Z \in \mathbb{S}^n$, *and* $\varepsilon \in (0, 1/2)$, *let* $\widetilde{Z^2} = \textbf{\textit{RandProj}}(Z, \mathrm{T}_{\mathrm{est_{jl}}})$ *and* $\widehat{\theta}_{1_i} \sim \textbf{\textit{InvSqrt}}((\widetilde{Z^2})_{ii} + 1, \mathrm{T}_{\mathrm{est_{isq}}})$ *for* $i \in [n]$. *Then,*

(1) *The first moment satisfies* $\left|\mathbf{E}\,\widehat{\theta}_{1_i} - \frac{1}{\sqrt{(Z^2)_{ii}+1}}\right| \leq \frac{\sqrt{2}(\varepsilon/n)^{400}}{\sqrt{(Z^2)_{ii}+1}}$.

(2) *The second moment satisfies* $\mathbf{E}\,|\widehat{\theta}_{1_i}|^2 \leq \frac{1}{(Z^2)_{ii}}1630\log(n/\varepsilon)$.

Lemma 7 *Consider Z_1, Z_2, Z, and Δ all in $\mathbb{S}^n$. Sample $\zeta \sim \mathcal{N}(\mathbf{0}, I_n)$, and define $\widehat{\theta}_2 \in \mathbb{R}^n$ as $\widehat{\theta}_{2_i} = (Z_1\Delta Z_2\zeta)_i\,(Z\zeta)_i$. Define $\theta_{2_i} \overset{\text{def}}{=} (Z_1\Delta Z_2 Z)_{ii}$. Then for $i \in [n]$:*

(1) *The first moment satisfies* $\mathbf{E}\,\widehat{\theta}_{2_i} = \theta_{2_i}$

(2) *The second moment satisfies* $\mathbf{E}\,|\widehat{\theta}_{2_i}|^2 \leq 3\left(Z_1\Delta Z_2^2\Delta Z_1\right)_{ii}\left(Z^2\right)_{ii}$.

Proof [Proof sketch for Lemma 5] By construction,

$$\mathbf{E}_{s,\tau,\zeta_1,\zeta_2}\,\|\widehat{\theta}\|_2^2 = \int_{s=0}^{1}\int_{\tau=0}^{1}\sum_{i=1}^{n}\mathbf{E}_{\zeta_1}\,|\widehat{\theta}_{1_i}|^2\,\mathbf{E}_{\zeta_2}\,|\widehat{\theta}_{2_i}|^2 dsd\tau.$$

Plugging in the second moment bounds from Lemma 6 and Lemma 7 gives

$$\mathbf{E}_{s,\tau,\zeta_1,\zeta_2} \|\widehat{\theta}\|_2^2 = 4890 \log(n/\varepsilon) \int_{s=0}^{1} \int_{\tau=0}^{1} \mathrm{Tr}(\widetilde{\exp}(2\bar{\tau} Y_s)\Delta \widetilde{\exp}((2\tau - 1)Y_s)\Delta) ds d\tau.$$

This step is made possible by the careful choice of split in $\widehat{\theta}_2$ that enable cancellations of $\frac{1}{(\widetilde{\exp} Y_s)_{ii}}$ and $(\widetilde{\exp} Y_s)_{ii}$. Applying Fact 0.1 and the fact that $\widetilde{\exp} Y_s$ is close to the true $\exp Y_s$, the above trace term is bounded by $\mathrm{Tr}\big(\exp(Y + s\Delta)\Delta^2\big)$ (plus a small error term). Applying Hölder's Inequality, Lemma 10, and values of η and G completes the proof. ■

To provide proof sketches of Lemma 6 and Lemma 7, we need two technical lemmas about **RandProj** and **InvSqrt**, the main workhorses for our estimators. These lemmas follow from properties of Gaussian and the scaled chi-squared distribution.

Lemma 8 *Consider a positive random variable x sampled from a distribution X with mean μ and variance σ^2. For some integer $k > 0$, construct the distribution $\mathcal{G}(X) =$ **InvSqrt**(X, k) defined in Equation 2.2. Then the random variable $g \sim \mathcal{G}(X)$ satisfies*

(1) $|\mathbf{E}\, g - \mu^{-1/2}| \le \mathbf{E} \left(\frac{|x-\mu|^k}{\min(\mu,x)^{k+1/2}} \right)$

(2) $\mathbf{E}\, |g|^2 \le k \sum_{j=0}^{k-1} \mathbf{E} \left(\frac{\left(\sigma^2 + (\mu - x)^2\right)^j}{x^{2j+1}} \right)$.

Lemma 9 *Given $u \in \mathbb{R}^n$ such that $\mu \overset{\text{def}}{=} \|u\|_2^2 \neq 0$, and positive integers $k > 1$ and $N \ge 4k + 6$, the following are true for x sampled from $X =$ **RandProj**(u, N).*

(1) $\mathbf{E}\, x = \mu$

(2) $\sigma^2 \overset{\text{def}}{=} \mathbf{E}\, (x - \mu)^2 = \frac{2\mu^2}{N}$

(3) $\mathbf{E} \left(\frac{\left(\sigma^2 + (x-\mu)^2\right)^k}{\min(x,\mu)^{2k+1}} \right) \le \frac{1}{\mu} \left(\frac{e^{N/2}}{2^{N-17k}} + \frac{2^{13k} k^{2k}}{N^k} \right)$

Proof [Proof sketches of Lemmas 6 and 7] Consider $x \sim \widetilde{Z^2}_{ii}$. By Lemma 9, $\mathbf{E}\, x = Z_{ii}^2$. This satisfies the bias requirement of Lemma 8, and therefore

$$\begin{aligned}
\left| \mathbf{E}\, \widehat{\theta}_{1_i} - \frac{1}{\sqrt{1 + (Z^2)_{ii}}} \right| &\le \mathbf{E} \left(\frac{\left| x - (Z^2)_{ii} \right|^{\mathrm{T_{est_{isq}}}}}{\min(x+1, (Z^2)_{ii} + 1)^{\mathrm{T_{est_{isq}}} + \frac{1}{2}}} \right) \\
&\le \sqrt{\mathbf{E} \frac{\left(x - (Z^2)_{ii}\right)^{2\mathrm{T_{est_{isq}}}}}{\min\left(x+1, (Z^2)_{ii} + 1\right)^{2\mathrm{T_{est_{isq}}} + 1}}} \\
&\le \sqrt{\frac{1}{(Z^2)_{ii} + 1} \left(\frac{e^{\mathrm{T_{est_{isq}}}/2}}{2^{\mathrm{T_{est_{jl}}} - 17\mathrm{T_{est_{isq}}}}} + \frac{2^{13\mathrm{T_{est_{isq}}}} {\mathrm{T_{est_{isq}}}}^{2\mathrm{T_{est_{isq}}}}}{{\mathrm{T_{est_{jl}}}}^{\mathrm{T_{est_{isq}}}}} \right)}.
\end{aligned}$$

where the first step is by Lemma 8, the second is by Jensen's inequality, and the third step is by a slight modification of (3) in Lemma 9. The values of $\mathrm{T_{est_{isq}}}$ and $\mathrm{T_{est_{jl}}}$ from Algorithm 2 give the final bias bound. The second moment bound follows similarly, and the properties of $\widehat{\theta}_2$ follow from simple properties of the Gaussian distribution. ■

2.2. Technical Concepts: Domain Expansion and Strong Convexity

In this section we state and sketch the proofs of two key technical concepts: (1) the addition of the trace constraint as described before the proof of Theorem 3, and (2) the value of the strong convexity parameter of our mirror map over this new domain.

Lemma 10 *With the choice of parameters in Algorithm 1, the iterate $\widetilde{X}^{(t)}$ at any iteration t satisfies $\operatorname{Tr}\widetilde{X}^{(t)} < K$ for $K = 40n(\log n)^{10}$.*

Proof [Proof sketch] We assume that for any iteration t, the primal iterate is close to the optimal point and satisfies $|||\widetilde{X}^{(t)} - X^*||| \leq 38n\,(\log n)^{10}$. In Algorithm 1, $Y^{(1)} = 0$ implies $\widetilde{X}^{(1)} = I$. We also know that the optimal point satisfies $\operatorname{Tr} X^* = n$. Therefore, in the base case, $|||\widetilde{X}^{(1)} - X^*||| \leq 2n \leq 38n\,(\log n)^{10}$. Suppose that the hypothesis is true for some $t = t'$. We complete the proof by first proving a weak bound for $|||\widetilde{X}^{(t)} - X^*|||$ using the triangle inequality of norms and then boosting our bound (thereby obtaining the stronger guarantee of the induction hypothesis) by invoking the strong convexity of the Bregman divergence. The full proof is presented in Section 3.6. ■

We now sketch the proof of the strong convexity parameter of our mirror map, the *generalized* negative entropy function. This mirror map is different from the negative entropy function and has recently appeared in (Allen-Zhu and Orecchia, 2015).

Lemma 11 *The function $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$ is $\frac{1}{4K}$-strongly convex with respect to the nuclear norm over the domain $\mathcal{D} = \{X : X \succeq 0, \operatorname{Tr} X \leq K\}$.*

Proof [Proof sketch] We invoke the duality between strong convexity and smoothness by Kakade et al. (2009), the characterization of matrix smooth functions by Juditsky and Nemirovski (2008), and the generalization of convexity of a permutation-invariant function on vectors to a spectral function on matrices by Lewis (1995). Our proof requires the following definition.

Definition 12 *Define the vector functions $\psi_1(y) = \sum_{i=1}^n \exp y_i$, $\psi_2(y) = 2K \log \psi_1(y) - 2K\log(2K) + 2K$, $\psi(y) = \psi_1(y)$ if $\psi_1(y) \leq 2K$ and $\psi_2(y)$ otherwise; $\Psi(Y) = \Psi_1(Y)$ if $\Psi_1(Y) \leq 2K$ and $\Psi_2(Y)$ otherwise; and $\phi(x) = \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i$. Define the corresponding matrix functions $\Psi_1(Y) = \operatorname{Tr}\exp Y$, $\Psi_2(Y) = 2K\log\Psi_1(Y) - 2K\log(2K) + 2K$, and $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$.*

Our first step is to show that Ψ, the matrix version of ψ, satisfies the property $\Psi^*(Y) = \Phi(Y)$ over $\{Y : Y \succeq 0, \operatorname{Tr} Y \leq K\}$. To prove this, we first prove that ψ and its matrix version, Ψ, are both continuously differentiable at the boundary of definition of their respective two parts. We then show that ψ_1 and ψ_2 are convex; combining this with the claim about continuous differentiability implies convexity of ψ, which immediately extends to Ψ by a result of Lewis (1995). We then show that ψ and ϕ satisfy $\psi_1^*(x) = \phi(x)$ for $x \in \mathbb{R}^n_+$, and given an input $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$, the point y attaining the optimum in computing $\psi_1^*(x)$ lies in the *interior* of the set $\{y : \psi_1(y) \leq 2K\}$. Therefore, given an input $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$, we invoke the preceeding facts to conclude that the point at which the value of $\psi^*(x)$ is attained must be the same as that for $\psi_1^*(x)$. This implies $\psi^*(x) = \psi_1^*(x)$ for $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$. By a result of Lewis (1995), this extends to $\Psi^* = \Phi$ on $\{X : X \succeq 0, \operatorname{Tr} X \leq K\}$.

We then use (Juditsky and Nemirovski, 2008) and continuous differentiability at the boundary to show that Ψ is $4K$-smooth in the operator norm which in turn implies, by (Kakade et al., 2009), that Ψ^* is $1/(4K)$-strongly convex in the nuclear norm, finishing the proof. Our full proof is in Section 3.1. ■

Acknowledgment

We would like to express our heartfelt gratitude to the following people: various anonymous reviewers of SODA 2019 for their careful reading, tough questions, and constructive suggestions on improving our presentation; Kevin Tian (Stanford) for insightful questions that strengthened our exposition and for helpful explanations of his paper (Carmon et al., 2019); Sidhanth Mohanty (UC Berkeley) for a clear explanation of bounds on the rank of the MaxCut SDP from (Raghavendra and Steurer, 2009) and (Montanari and Sen, 2016a); Harish Ramadas (UW), Kuikui Liu (UW), Shashwat Garg (TU Eindhoven), and Siva Ramamoorthy (UW) for helpful feedback on an early draft of the paper.

References

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: faster online learning of eigenvectors and faster mmwu. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 116–125, 2017.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, 2015.

Zeyuan Allen Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1824–1831, 2016.

Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 227–236, 2007.

Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 339–348. IEEE, 2005.

Michel Baes, Michael Bürgisser, and Arkadi Nemirovski. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM Journal on Optimization*, 23(2):934–962, 2013.

Afonso S Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 459–470, 2014.

Francisco Barahona, Martin Grötschel, Michael Jünger, and Gerhard Reinelt. An application of combinatorial optimization to statistical physics and circuit layout design. *Operations Research*, 36(3):493–513, 1988.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yair Carmon, John C. Duchi, Aaron Sidford, and Kevin Tian. A rank-1 sketch for matrix multiplicative weights. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 589–623, 2019.

Ruen-Wu Chen, Yoji Kajitani, and Shu-Park Chan. A graph-theoretic via minimization algorithm for two-layer printed circuit boards. *IEEE Transactions on Circuits and Systems*, 30(5):284–299, 1983.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. 2014.

M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proceedings of the 2004 American Control Conference*, 2004.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Dan Garber and Elad Hazan. Sublinear time algorithms for approximate semidefinite programming. *Mathematical Programming*, 158(1-2):329–361, 2016.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American symposium on theoretical informatics*, pages 306–316, 2008.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 2016.

Martin Jaggi. *Sparse Convex Optimization Methods for Machine Learning.* PhD thesis, ETH Zurich, 2011.

Rahul Jain and Penghui Yao. A parallel approximation algorithm for positive semidefinite programming. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 463–471, 2011.

Arun Jambulapati, Yin Tat Lee, Jerry Li, Swati Padmanabhan, and Kevin Tian. Positive semidefinite programming: mixed, parallel, and width-independent. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Applications of strong convexity–strong smoothness duality to learning with matrices. *CoRR*, abs/0910.0610, 2009.

R. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? 2007.

Philip Klein and Hsueh-I Lu. Efficient approximation algorithms for semidefinite programs arising from max cut and coloring. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, 1996.

Adrian S Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. In *Conference on Learning Theory, COLT 2017*, pages 1476–1515, 2017.

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, 2016a.

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827, 2016b.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Victor Y. Pan and Zhao Q. Chen. The complexity of the matrix eigenproblem. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA*, pages 507–516, 1999.

Richard Peng and Kanat Tangwongsan. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. In *Proceedings of the twenty-fourth annual ACM symposium on Parallelism in algorithms and architectures*, pages 101–108, 2012.

S. A. Plotkin, D. B. Shmoys, and E. Tardos. Fast approximation algorithms for fractional packing and covering problems. In *[1991] Proceedings 32nd Annual Symposium of Foundations of Computer Science*, 1991.

Prasad Raghavendra and David Steurer. How to round any csp. In *Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '09, 2009.

Sushant Sachdeva, Nisheeth K Vishnoi, et al. Faster algorithms via approximation theory. *Foundations and Trends® in Theoretical Computer Science*, 9(2):125–210, 2014.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Amit Singer and Yoel Shkolnisky. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming. *SIAM journal on imaging sciences*, 4(2):543–572, 2011.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560, 2005.

Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 2015.

Jun Wang, Tony Jebara, and Shih-Fu Chang. Semi-supervised learning using greedy maxcut. *Journal of Machine Learning Research*, 14(Mar):771–800, 2013.

R. M. Wilcox. Exponential Operators and Parameter Differentiation in Quantum Physics. *Journal of Mathematical Physics*, 1967.

Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming, 2019.

Appendices

We organize the appendix into four parts: Section 1, analysis common to Arora and Kale (2007) and us; Section 2 and Section 3, analysis of Arora and Kale (2007) and our algorithm, respectively; Section 4, general technical results.

1. Analysis Common to Both Algorithms

In this section we provide proofs for two results: the first is that a solution to the reformulated problem (1.2) is indeed ε close to that of the original; the second is the convergence guarantee of approximate lazy mirror descent, the framework for both the Arora-Kale algorithm as well as ours.

Algorithm 3 Approximate lazy mirror descent

```
Input: Objective function $f : \mathcal{X} \to \mathbb{R}$, accuracy parameter $\varepsilon$.
Parameters: Mirror map $\Phi : \mathcal{D} \to \mathbb{R}$, norm $\|\cdot\|$, step size $\eta$, iteration $T$, error bound $\delta$.
Initialize: $x^{(1)} \in \mathrm{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} \Phi(x)$, $\widetilde{x}^{(1)} = x^{(1)}$, and $z^{(1)}$ satisfying $\nabla\Phi(z^{(1)}) = 0$.
for $t = 1 \to T$ do
    $\nabla\Phi(z^{(t+1)}) \leftarrow \nabla\Phi(z^{(t)}) - \eta\nabla f(\widetilde{x}^{(t)})$    ▷ Lazy gradient update
    Find $\widetilde{x}^{(t+1)}$ such that $\mathbf{E}\,\|\widetilde{x}^{(t+1)} - x^{(t+1)}\| \leq \delta$, where $x^{(t+1)} \in \mathrm{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} \mathcal{D}_\Phi(x, z^{(t+1)})$
     ▷ Approximate projection
end
For $t^* \overset{\text{unif.}}{\sim} \{1, 2, \ldots, T\}$, return $\widetilde{x}^{(t^*)}$.
```

1.1. From the Reformulated to the Original SDP

Our claim of reformulating (1.1) as (1.2) works because once we have a solution X for the latter, we can apply the following result to obtain a matrix $\widehat{X}$ which satisfies all the required constraints of (1.1), and at which the objective value in (1.1) is better than that at X in (1.2).

Lemma 2 *Given $C \in \mathbb{R}^{n\times n}$ and $0 \preceq X$, let $\rho \in \mathbb{R}^n$ with $\rho_i = \sum_{j=1}^n |C_{ij}|$; diagonal matrix S with $S_{ii} = \min(1/\sqrt{\rho_i}, 1/\sqrt{X_{ii}})$ for $i \in [n]$; $\widehat{X} = SXS$; $\widehat{C} = \mathbf{diag}\,(1/\sqrt{\rho})\, C\, \mathbf{diag}(1/\sqrt{\rho})$. Then, $\widehat{X} \succeq 0$, $\widehat{X}_{ii} \leq 1$ for all $i \in [n]$, and $\widehat{C} \bullet X - \sum_{i=1}^n (X_{ii} - \rho_i)^+ \leq C \bullet \widehat{X}$.*

Proof We first prove the positive semidefiniteness. Observe that since $\widehat{X}$ and X are similar matrices, $X \succeq 0$ implies $\widehat{X} \succeq 0$ as well. Next, we define a matrix Y as $Y_{ij} = \frac{X_{ij}}{\sqrt{\rho_i}\sqrt{\rho_j}}$. Without loss of generality, assume $Y_{11} \geq Y_{22} \geq \ldots \geq Y_{nn}$. We also define a diagonal matrix, $\widehat{D}$ as $\widehat{D}_{ii} = \min(1, 1/\sqrt{Y_{ii}})$. If $Y_{ii} \geq 1$, then $\widehat{X}_{ii} = \frac{\rho_i Y_{ii}}{\sqrt{\rho_i Y_{ii}}\sqrt{\rho_i Y_{ii}}} = 1$; otherwise, $\widehat{X}_{ii} = Y_{ii}$. This proves that $\widehat{X}_{ii} \leq 1$ for all $1 \leq i \leq n$, which is precisely the claim bounding every diagonal entry. We now prove the claim about the objective value. By definition of

$\widehat{D}$, $\widehat{X}$ and Y, we have $\widehat{X} = \widehat{D} \cdot Y \cdot \widehat{D}$. Therefore we get

$$\begin{aligned} C \bullet (\widehat{X} - Y) - \sum_{i=1}^{n} C_{ii} Y_{ii} (\widehat{D}_{ii}^2 - 1) &= \sum_{i=1}^{n} \sum_{j \neq i} C_{ij} Y_{ij} (\widehat{D}_{ii} \widehat{D}_{jj} - 1) \\ &= 2 \sum_{i=1}^{n} \sum_{i<j} C_{ij} Y_{ij} (\widehat{D}_{ii} \widehat{D}_{jj} - 1). \end{aligned}$$

The definition of $\widehat{D}$ and the ordering assumption on $\{Y_{ii}\}$ imply $0 < \widehat{D}_{11} \leq \widehat{D}_{22} \leq \ldots \leq \widehat{D}_{nn} \leq 1$, which in turn means $\widehat{D}_{ii} \widehat{D}_{jj} \geq \widehat{D}_{ii}^2$. Further, since $X \succeq 0$ and $Y = \mathbf{diag}(1/\sqrt{\rho}) \cdot X \cdot \mathbf{diag}(1/\sqrt{\rho})$, we have $Y \succeq 0$. Therefore $Y_{ii} Y_{jj} \geq Y_{ij} Y_{ji}$. By symmetry of Y and the assumed ordering of $\{Y_{ii}\}_1^n$, this can be simplified to $Y_{ii} \geq |Y_{ij}|$ for $i < j$. These two facts simplify the above to

$$\begin{aligned} C \bullet (\widehat{X} - Y) - \sum_{i=1}^{n} C_{ii} Y_{ii} (\widehat{D}_{ii}^2 - 1) &\geq 2 \sum_{i=1}^{n} \sum_{i<j} |C_{ij}| |Y_{ij}| (\widehat{D}_{ii}^2 - 1) \\ &\geq 2 \sum_{i=1}^{n} \sum_{i<j} |C_{ij}| Y_{ii} (\widehat{D}_{ii}^2 - 1) \end{aligned}$$

Finally, since $\widehat{D}_{ii} \leq 1$, we have $\widehat{D}_{ii}^2 - 1 \leq 0$. Rearranging the terms in the last inequality, we get

$$\begin{aligned} C \bullet (\widehat{X} - Y) &\geq \sum_{i=1}^{n} C_{ii} Y_{ii} (\widehat{D}_{ii}^2 - 1) + \sum_{i=1}^{n} Y_{ii} (\widehat{D}_{ii}^2 - 1) (\sum_{j>i} |C_{ij}| + \sum_{j<i} |C_{ij}|) \\ &= \sum_{i=1}^{n} Y_{ii} (\widehat{D}_{ii}^2 - 1) \underbrace{\left(C_{ii} + \sum_{i>j} |C_{ij}| + \sum_{i<j} |C_{ij}| \right)}_{\leq \rho_i} \\ &\geq \sum_{i=1}^{n} Y_{ii} \rho_i (\widehat{D}_{ii}^2 - 1) \\ &= -\sum_{i=1}^{n} \rho_i \left(Y_{ii} - 1 \right)^+ \end{aligned}$$

where we used $\widehat{D}_{ii} = \min(1, 1/\sqrt{Y_{ii}})$ in the last step. Rearranging the terms in the last inequality gives

$$C \bullet \widehat{X} \geq C \bullet Y - \sum_{i=1}^{n} \rho_i \left(Y_{ii} - 1 \right)^+ = \widehat{C} \bullet X - \sum_{i=1}^{n} (X_{ii} - \rho_i)^+,$$

where the last step is by definition of matrix Y. ■

1.2. Analysis of Approximate Lazy Mirror Descent

We now derive the convergence bound of approximate lazy mirror descent. The proof closely follows that of Theorem 4.3 in Bubeck's monograph (Bubeck et al., 2015).

Theorem 1 (Convergence of Lazy Mirror Descent) *Fix a norm $\|\cdot\|$. Given an α-strongly convex mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ and a convex, G-Lipschitz objective $f : \mathcal{X} \to \mathbb{R}$, run Algorithm 3 with step size η and $\mathbf{E}\,\|x^{(t)} - \widetilde{x}^{(t)}\| \leq \delta$. Let $D \overset{\text{def}}{=} \sup_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x) - \inf_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x)$. Then, Algorithm 3 after T iterations returns $\widetilde{x}^{t^*}$, satisfying*

$$\mathbf{E}\, f(\widetilde{x}^{(t^*)}) - f(x^*) \leq \frac{D}{T\eta} + \frac{2\eta G^2}{\alpha} + \delta G. \tag{1.7}$$

Proof By convexity of f,

$$\sum_{t=1}^{T}(f(\widetilde{x}^{(t)}) - f(x)) \leq \sum_{t=1}^{T}\left\langle \nabla f(\widetilde{x}^{(t)}), \widetilde{x}^{(t)} - x\right\rangle = \underbrace{\sum_{t=1}^{T}\left\langle \nabla f(\widetilde{x}^{(t)}), \widetilde{x}^{(t)} - x^{(t)}\right\rangle}_{\text{Ⓐ}} + \underbrace{\sum_{t=1}^{T}\left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t)} - x\right\rangle}_{\text{Ⓑ}}. \tag{1.1}$$

The term Ⓐ can be bounded by Cauchy-Schwarz inequality and the invariant $\mathbf{E}\,\left\|x^{(t)} - \widetilde{x}^{(t)}\right\| \leq \delta$:

$$\text{Ⓐ} \leq \sum_{t=1}^{T}\left\|\Delta^{(t)}\right\|\left\|\nabla f\left(\widetilde{x}^{(t)}\right)\right\|_{*} \leq \delta G T. \tag{1.2}$$

Next, recall that Algorithm 3 initializes $x^{(1)} \in \operatorname{argmin}_{\mathcal{X}\cap\mathcal{D}} \Phi(x)$ and $z^{(1)}$ satisfying $\nabla\Phi(z^{(1)}) = 0$, and repeats the following two steps:

$$\begin{aligned}\nabla\Phi(z^{(t)}) &= \nabla\Phi(z^{(t-1)}) - \eta\nabla f(x^{(t)})\\ x^{(t)} &= \operatorname*{argmin}_{\mathcal{X}\cap\mathcal{D}} D_{\Phi}(x, z^{(t)}).\end{aligned}$$

Now consider the potential function $\widetilde{\Psi}_t(x) \overset{\text{def}}{=} \Phi(x) + \eta\left\langle x, \sum_{s=1}^{t}\nabla f(\widetilde{x}^{(s)})\right\rangle$. Applying the recursive definition of the gradient step, we can express $x^{(t+1)} = \operatorname*{argmin}_{x\in\mathcal{X}\cap\mathcal{D}} \widetilde{\Psi}_t(x)$. Since Φ is α-strongly convex, so is the potential function Ψ_t. We can express these two statements as follows:

$$\begin{aligned}\widetilde{\Psi}_t(x^{(t+1)}) - \widetilde{\Psi}_t(x^{(t)}) &\leq \underbrace{\left\langle \nabla\widetilde{\Psi}_t(x^{(t+1)}), x^{(t+1)} - x^{(t)}\right\rangle}_{\leq 0,\text{ by optimality of } x^{(t+1)}} - \tfrac{\alpha}{2}\left\|x^{(t+1)} - x^{(t)}\right\|^2\\ &\leq -\tfrac{\alpha}{2}\left\|x^{(t+1)} - x^{(t)}\right\|^2.\end{aligned} \tag{1.3}$$

We can also write a lower bound for the left hand side of Inequality 1.3 by evaluating the potential function $\widetilde{\Psi}_t$ at points $x^{(t+1)}$ and $x^{(t)}$:

$$\begin{aligned}\widetilde{\Psi}_t(x^{(t+1)}) - \widetilde{\Psi}_t(x^{(t)}) &= \Phi\left(x^{(t+1)}\right) + \eta \sum_{s=1}^{t} \left\langle \nabla f(\widetilde{x}^{(s)}), x^{(t+1)} \right\rangle - \Phi(x^{(t)}) - \eta \sum_{s=1}^{t} \left\langle \nabla f(\widetilde{x}^{(s)}), x^{(t)} \right\rangle \\ &= \underbrace{\widetilde{\Psi}_{t-1}(x^{(t+1)}) - \widetilde{\Psi}_{t-1}(x^{(t)})}_{\geq 0, \text{ since } x^{(t)} \text{ minimizes } \widetilde{\Psi}_{t-1}(x)} + \eta \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t+1)} - x^{(t)} \right\rangle \\ &\geq \eta \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t+1)} - x^{(t)} \right\rangle. \end{aligned} \tag{1.4}$$

Reverse and chain Inequalities 1.3 and 1.4, and apply Cauchy-Schwarz inequality to get

$$\frac{\alpha}{2}\left\|x^{(t+1)} - x^{(t)}\right\|^2 \leq \eta \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t)} - x^{(t+1)} \right\rangle \leq \eta G \left\|x^{(t)} - x^{(t+1)}\right\|. \tag{1.5}$$

This shows that

$$\left\|x^{(t)} - x^{(t+1)}\right\| \leq \frac{2\eta G}{\alpha}, \tag{1.6}$$

and applying this to the second part of Inequality 1.5 gives

$$\left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t)} - x^{(t+1)} \right\rangle \leq \frac{2\eta G^2}{\alpha}. \tag{1.7}$$

We now claim

$$\sum_{t=1}^{T} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t)} - x \right\rangle \leq \sum_{t=1}^{T} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t)} - x^{(t+1)} \right\rangle + \tfrac{1}{\eta}(\Phi(x) - \Phi(x^{(1)})). \tag{1.8}$$

Note that this claim immediately gives the desired error bound; this can be seen as follows: the left-hand side is exactly the term ② in Inequality 1.1; the first term of the right-hand side is bounded in Inequality 1.7, and the second one is bounded by the definition of set size D. Therefore Inequality 1.8 simplifies to

$$\text{Ⓑ} \leq \frac{2\eta G^2 T}{\alpha} + \frac{D}{\eta}. \tag{1.9}$$

Combine Inequalities 1.9 and 1.2 with 1.1, apply Jensen's inequality, and the fact that t^* is picked uniformly at random from $\{1, 2, \ldots, T\}$, to get the desired error bound. We now prove Inequality 1.8. First, we rewrite it as

$$\sum_{t=1}^{T} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t+1)} \right\rangle + \frac{\Phi(x^{(1)})}{\eta} \leq \sum_{t=1}^{T} \left\langle \nabla f(\widetilde{x}^{(t)}), x \right\rangle + \frac{\Phi(x)}{\eta}.$$

The claim is true for $T = 0$ for all $x \in \mathcal{X}$, by the choice of $x^{(1)}$. Assume it holds for all $x \in \mathcal{X}$ at time $T = t' - 1$. Therefore in particular, it holds at the point $x = x^{(t'+1)}$. This implies

$$\begin{aligned}
\sum_{t=1}^{t'} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t+1)} \right\rangle + \frac{\Phi(x^{(1)})}{\eta} &= \left\langle \nabla f(\widetilde{x}^{(t')}), x^{(t'+1)} \right\rangle + \underbrace{\sum_{t=1}^{t'-1} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t+1)} \right\rangle + \frac{\Phi(x^{(1)})}{\eta}}_{\text{Apply induction hypothesis at } x^{(t'+1)}} \\
&\leq \left\langle \nabla f(\widetilde{x}^{(t')}), x^{(t'+1)} \right\rangle + \sum_{t=1}^{t'-1} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t'+1)} \right\rangle + \frac{\Phi(x^{(t'+1)})}{\eta} \\
&= \sum_{t=1}^{t'} \left\langle \nabla f(\widetilde{x}^{(t)}), x^{(t'+1)} \right\rangle + \frac{\Phi\left(x^{(t'+1)}\right)}{\eta} \\
&= \frac{1}{\eta} \widetilde{\Psi}_{t'} \left(x^{(t'+1)}\right) \\
&\leq \frac{1}{\eta} \widetilde{\Psi}_{t'}(x) \\
&= \sum_{t=1}^{t'} \left\langle \nabla f\left(\widetilde{x}^{(t)}\right), x \right\rangle + \frac{\Phi(x)}{\eta},
\end{aligned}$$

where the last inequality is by optimality of $x^{(t'+1)}$ in minimizing $\widetilde{\Psi}_{t'}$. This completes the induction, and therefore proves Inequality 1.8, thus completing the proof of the error bound. ■

2. Analysis of the Arora-Kale Algorithm

In this section, we display Algorithm 4 in the approximate mirror descent framework and provide its analysis. In Section 2.1, we derive the values of all parameters; in Section 2.2, we derive the computational costs of the main steps. We then conclude with the correctness and cost of their algorithm. The main export of this section is the following theorem.

Theorem 13 (Run Time (Arora and Kale, 2007)) *Given $C \in \mathbb{R}^{n\times n}$ with $m \geq n$ non-zero entries and $0 < \varepsilon \leq \frac{1}{2}$, we can find, in time $\widetilde{\mathcal{O}}\left(m/\varepsilon^5\right)$, a matrix $Y \in \mathbb{S}^n$ with $\mathcal{O}(m)$ non-zero entries and a diagonal matrix $S \in \mathbb{R}^{n\times n}$ such that $\widetilde{X}^* = S \cdot \frac{K\exp(Y)}{\operatorname{Tr}\exp(Y)} \cdot S$ satisfies $\widetilde{X}^* \succeq 0$, $\widetilde{X}^*_{ii} \leq 1$ for all $i \in [n]$, and $\mathbf{E}(C \bullet \widetilde{X}^*) \geq C \bullet X^* - \varepsilon \cdot \sum_{i,j} |C|_{ij}$.*

2.1. Parameters

As can be seen in Algorithm 3, approximate lazy mirror descent requires five parameters: the set diameter, Lipschitz constant of the objective, strong convexity of the mirror map, step size, and number of iterations. The first three depend on our choice of mirror map Φ and objective f. The last two can be chosen based on these parameters and Inequality 1.7.

Lemma 14 (Set Diameter) *Given $\Phi(X) = X \bullet \log X$ and the domain $\{X : X \succeq 0, \operatorname{Tr} X = n\}$, the set diameter measured by Φ is given by $D = n \log n$.*

Algorithm 4 Reinterpreting Arora and Kale (2007)

```
Input: Cost matrix $C \in \mathbb{R}^{n\times n}$, accuracy parameter $\varepsilon$.
Parameters: $T = 256\log n/\varepsilon^2$, $T' = 10240\log n/\varepsilon^2$, $T'' = (1/\varepsilon)\cdot 64\log n$, $\eta = \varepsilon/64$. Set $\widehat{C}$
and $\rho$ as defined in Lemma 2.
Initialize $Y^{(1)} \leftarrow \mathbf{0}$.
Define $\nabla f(M) \overset{\text{def}}{=} \mathbf{diag}\,\mathbf{1}_{M\geq\rho} - \widehat{C}$.
for $t = 1$to$T$ do
    $\widetilde{\exp}\left(\frac{1}{2}Y^{(t)}\right) \leftarrow$ TaylorExp $\left(\frac{1}{2}Y^{(t)}, T''\right)$.        ▷ Approximate matrix exponential
    $\widehat{\exp}Y^{(t)} \leftarrow$ RandProj $\left(\widetilde{\exp}\left(\frac{1}{2}Y^{(t)}\right), T'\right)$.        ▷ Approximate projection
    $\widetilde{X}^{(t)} \leftarrow n\frac{\widehat{\exp}(Y^{(t)})}{\mathrm{Tr}\,\widehat{\exp}Y^{(t)}}$        ▷ Scaling due to the trace constraint
    $Y^{(t+1)} \leftarrow Y^{(t)} - \eta\nabla f(\widetilde{X}^{(t)})$.        ▷ Gradient update.
end
For $t^* \overset{\text{unif.}}{\sim} \{1, 2, \ldots, T\}$, return $Y^{(t^*)}$ and $S$, where $S$ is from Lemma 2.
```

Lemma 15 (Lipschitz constant) *The problem objective $\widehat{f}(X) = -\widehat{C} \bullet X + \sum_{i=1}^{n}(X_{ii} - \rho_i)^+$ (recall that $\rho_i = \sum_{j=1}^{n}|C_{ij}|$) is 2-Lipschitz in the nuclear norm. Recall that nuclear norm of a matrix is the sum of its singular values.*

Proof The gradient of the objective at point X is $\nabla\widehat{f}(X) = \mathbf{diag}(\mathbf{1}_{\{X\geq\rho\}}) - \widehat{C}$. By the Gershgorin Disk Theorem, we have

$$\left\|\mathbf{diag}\left(\tfrac{1}{\rho}\right)C\right\|_{\mathrm{op}} \leq \max_{i\in[n]}\left(\frac{1}{\rho_i}\cdot|C_{ii}| + \frac{1}{\rho_i}\cdot\sum_{j\neq i}|C_{ij}|\right) = \max_{i\in[n]}\left(\frac{1}{\rho_i}\cdot\sum_{j=1}^{n}|C_{ij}|\right) = 1, \tag{2.1}$$

where in the last equality we use the choice of $\rho_i = \sum_{j=1}^{n}|C_{ij}|$. Since the matrices $\mathbf{diag}(^1/_\rho)\cdot C$ and $\widehat{C} = \mathbf{diag}(^1/_{\sqrt{\rho}})\cdot C\cdot\mathbf{diag}(^1/_{\sqrt{\rho}})$ are similar, they have the same set of eigenvalues (and therefore, the same operator norm). Therefore

$$\left\|\mathbf{diag}(\mathbf{1}_{\{X\geq\rho\}}) - \widehat{C}\right\|_{\mathrm{op}} \leq \left\|\mathbf{diag}(\mathbf{1}_{\{X\geq\rho\}})\right\|_{\mathrm{op}} + \left\|\widehat{C}\right\|_{\mathrm{op}} = 1 + 1 = 2.$$

When we have $\left\|\nabla\widehat{f}\right\| \leq G$ for some G, it implies f is G-Lipschitz in $\|\cdot\|_*$ (the dual norm). Therefore, in our case, we have that $\widehat{f}$ is 2-Lipschitz in the nuclear norm (dual of the operator norm). ■

Lemma 16 (Strong Convexity) *(Kakade et al. (2009)) The mirror map $\Phi(X) = X \bullet \log X$ is $1/(2n)$-strongly convex with respect to the nuclear norm on the domain $\{X \in \mathbb{S}^n : X \succeq 0, \mathrm{Tr}(X) = n\}$.*

Lemma 17 *Choosing $\eta = \varepsilon/64$ and $T = 256\log n/\varepsilon^2$ in Algorithm 4 gives an accuracy of εn.*

Proof We show in Lemma 23 that Algorithm 4 maintains the invariant $\mathbf{E}\,\|X^{(t)} - \widetilde{X}^{(t)}\| \leq \delta = \varepsilon n/4$. Therefore we are in the framework of approximate lazy mirror descent and can use its error bound from Inequality 1.7 and bound it by εK. We plug in the parameters from Lemmas 14, 15, and 16 in the bound and get

$$\mathbf{E}\, f(\widetilde{x}^{(t^*)}) - f(x^*) \leq \frac{n \log n}{T\eta} + \frac{2\eta \cdot 2^2}{1/2n} + \left(\frac{\varepsilon n}{4}\right) \cdot 2.$$

We optimize for η by setting the first two terms equal, and get

$$\eta = \tfrac{1}{4}\sqrt{\frac{\log n}{T}}. \tag{2.2}$$

With this expression for η, setting the bound for the right-hand side above to be εn gives $T \geq 256 \log n/\varepsilon^2$; plug this back in Equation 2.2 to get $\eta = \varepsilon/64$. ■

2.2. Computational Cost

From Algorithm 4, we see that there are three main parts to be computed to get the overall cost of the Arora-Kale algorithm: the number of iterations, the number of JL projections per iteration, and the cost of approximating a matrix exponential and multiplying it with a vector. We derive these values in this section.

2.2.1. Taylor Approximation for Matrix Exponential

In Algorithm 4, before we do the randomized projection to get the diagonal entries, we approximate the matrix exponential $\widetilde{\exp}\left(Y^{(t)}/2\right) = \textbf{TaylorExp}\left(Y^{(t)}/2, T''\right)$. Here we show a bound on $\left|\frac{\exp\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\exp\left(Y^{(t)}\right)} - \frac{\widetilde{\exp}\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\widetilde{\exp}\left(Y^{(t)}\right)}\right|$ for any $1 \leq i \leq n$. We do so by first proving a bound on $\left|\frac{A_{ii}}{\operatorname{Tr} A} - \frac{B_{ii}}{\operatorname{Tr} B}\right|$ for a matrix B approximating the general matrix A; then we prove a general result on the number of terms required to approximate a matrix exponential using Taylor series; finally, we combine these results to get an appropriate choice of T_{poly} for approximating $\exp\left(Y^{(t)}/2\right)$.

Lemma 18 *Given positive definite matrices A and B such that $\|A - B\|_{op} \leq \varepsilon$, where $\varepsilon \leq \frac{1}{2n}\operatorname{Tr} A$, we have $\left|\frac{A_{ii}}{\operatorname{Tr} A} - \frac{B_{ii}}{\operatorname{Tr} B}\right| \leq 2\frac{\varepsilon(\operatorname{Tr} A + nA_{ii})}{(\operatorname{Tr} A)^2}$.*

Proof We have the following chain of inequalities.

$$\left|\frac{B_{ii}}{\operatorname{Tr} B} - \frac{A_{ii}}{\operatorname{Tr} A}\right| \overset{①}{\leq} \left|\frac{A_{ii} + \varepsilon}{\operatorname{Tr} A - n\varepsilon} - \frac{A_{ii}}{\operatorname{Tr} A}\right| = \frac{\varepsilon\left(\operatorname{Tr} A + nA_{ii}\right)}{\left(\operatorname{Tr} A\right)\left(\operatorname{Tr} A - \varepsilon n\right)} \overset{②}{\leq} 2\frac{\varepsilon\left(\operatorname{Tr} A + nA_{ii}\right)}{\left(\operatorname{Tr} A\right)^2},$$

where ① is by the worst case values for B_{ii} from the operator norm bound, and ② is by the bound on ε. ■

Lemma 19 *For $T \geq e^2\|Y\|_{op}$, we have* $\left\|\exp(Y) - \sum_{j=0}^{T} \frac{Y^j}{j!}\right\|_{op} \leq \exp(-T)$.

Proof We have the following chain:

$$\left\|\exp(Y) - \sum_{j=0}^{T}\tfrac{1}{j!}Y^j\right\|_{\mathrm{op}} \overset{\textcircled{1}}{=} \left\|\sum_{j=T+1}^{\infty}\tfrac{1}{j!}Y^j\right\|_{\mathrm{op}} \overset{\textcircled{2}}{\leq} \sum_{j=T+1}^{\infty}\left\|\tfrac{1}{j!}Y^j\right\|_{\mathrm{op}} = \sum_{j=T+1}^{\infty}\tfrac{1}{j!}\|Y\|_{\mathrm{op}}^j \overset{\textcircled{3}}{\leq} \sum_{j=T+1}^{\infty}\tfrac{e^j}{j^j}\|Y\|_{\mathrm{op}}^j, \tag{2.3}$$

where $\textcircled{1}$ is by the Taylor series expansion of the matrix exponential, $\textcircled{2}$ is by triangle inequality of norms, and $\textcircled{3}$ is by Stirling's approximation, $j! \geq (j/e)^j$. Since the right hand side of the above inequality is indexed over $j \geq T \geq e^2\|Y\|_{\mathrm{op}}$, we can bound it further to get

$$\left\|\exp Y - \sum_{j=0}^{T}\tfrac{1}{j!}Y^j\right\|_{\mathrm{op}} \leq \sum_{j=T+1}^{\infty} e^{-j} = \frac{(e^{-1})^{T+1}}{1-e^{-1}} \leq e^{-T}.$$

■

Lemma 20 *In Algorithm 4, for $n \geq 2$ and $\varepsilon \leq \frac{1}{2}$, set $T_{poly} = \frac{64\log n}{\varepsilon}$, and let $\widetilde{\exp}\left(Y^{(t)}/2\right) :=$ **TaylorExp** $\left(Y^{(t)}/2, T_{poly}\right)$. Then for each coordinate i, we have $\left|\frac{\exp\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\exp Y^{(t)}} - \frac{\left(\widetilde{\exp}Y^{(t)}\right)_{ii}}{\operatorname{Tr}\widetilde{\exp}Y^{(t)}}\right| \leq \frac{\varepsilon}{8n}$.*

Proof Let $\widetilde{\exp}\left(Y^{(t)}/2\right) = \exp\left(Y^{(t)}/2\right) + \Delta$, and $\|\Delta\|_{\mathrm{op}} = \varepsilon_1$. Then

$$\begin{aligned}\left\|\exp Y^{(t)} - \widetilde{\exp}Y^{(t)}\right\|_{\mathrm{op}} &= \left\|(\exp\Big(Y^{(t)}/2\Big))^2 - (\widetilde{\exp}(Y^{(t)}/2))^2\right\|_{\mathrm{op}} \\ &= \left\|\Delta^2 + \Delta\exp\Big(Y^{(t)}/2\Big) + \exp\Big(Y^{(t)}/2\Big)\Delta\right\|_{\mathrm{op}} \\ &\leq \varepsilon_1^2 + 2\varepsilon_1\left\|\exp\Big(Y^{(t)}/2\Big)\right\|_{\mathrm{op}}. \end{aligned} \tag{2.4}$$

Observe that in each iteration of Algorithm 4, we add $-\eta\nabla f(\widetilde{X}^{(t)})$ to the current $Y^{(t)}$ in the gradient step; therefore at the end of all the T iterations, $\left\|Y^{(t)}\right\|_{\mathrm{op}} \leq |\eta T|\|\nabla f(\widetilde{X}^{(t)})\|_{\mathrm{op}}$. From the values of η and T as set in Algorithm 4 (and explained in Section 2), the worst-case value is

$$\left\|Y^{(t)}/2\right\|_{\mathrm{op}} \leq \frac{1}{2}\cdot\frac{\varepsilon}{64}\cdot\frac{256\log n}{\varepsilon^2}\cdot 2 = \frac{4\log n}{\varepsilon}. \tag{2.5}$$

Next, from Lemma 19, we require the first $\max\left\{e^2\left\|Y^{(t)}/2\right\|_{\mathrm{op}}, \log(1/\varepsilon_1)\right\}$ terms of the Taylor series of $\exp\left(Y^{(t)}/2\right)$ to get an ε_1 accuracy in approximation. Since $T_{\mathrm{poly}} = 64\log n/\varepsilon \geq e^2\left\|Y^{(t)}/2\right\|_{\mathrm{op}}$ (from Inequality 2.5), this choice of number of Taylor series terms corresponds to an accuracy of $\varepsilon_1 = n^{-64/\varepsilon}$. From Inequality 2.5, we get that

$$\|\exp\Big(Y^{(t)}/2\Big)\|_{\mathrm{op}} \leq e^{4\log n/\varepsilon} = n^{4/\varepsilon}. \tag{2.6}$$

Then we have

$$\begin{aligned}\varepsilon_1^2 + 2\varepsilon_1 \| \exp\Big(Y^{(t)}/2\Big) \|_{\mathrm{op}} &\leq n^{-128/\varepsilon} + 2n^{-64/\varepsilon} n^{4/\varepsilon} \\ &\leq 4n^{-60/\varepsilon} \\ &\leq \frac{n^{-4/\varepsilon}}{2} \leq \frac{1}{2n} \operatorname{Tr} \exp\Big(Y^{(t)}/2\Big), \end{aligned} \tag{2.7}$$

where the last inequality is by Inequality 2.6. Chaining Inequalities 2.4 and 2.7, the condition in Lemma 18 is satisfied. Applying the result of Lemma 18,

$$\begin{aligned}\left| \frac{\left(\exp\left(Y^{(t)}\right)\right)_{ii}}{\operatorname{Tr}\exp\left(Y^{(t)}\right)} - \frac{\left(\widetilde{\exp}\left(Y^{(t)}\right)\right)_{ii}}{\operatorname{Tr}\widetilde{\exp}\left(Y^{(t)}\right)} \right| &\leq 2\left(\varepsilon_1^2 + 2\varepsilon_1 \|\exp\left(Y\right)\|_{\mathrm{op}}\right) \frac{\operatorname{Tr}\exp\left(Y^{(t)}\right) + n\exp\left(Y^{(t)}\right)_{ii}}{\left(\operatorname{Tr}\exp\left(Y^{(t)}\right)\right)^2}. \\ &\leq 2\frac{\left(\varepsilon_1^2 + 2\varepsilon_1 n^{4/\varepsilon}\right)\left(2n^{1+8/\varepsilon}\right)}{\left(n^{-8/\varepsilon}\right)^2} \\ &\leq 4\left(\frac{\varepsilon^2}{10000n^{41/\varepsilon}} + \frac{\varepsilon}{50n^{4/\varepsilon}}\right) \\ &\leq \frac{\varepsilon}{8n}\end{aligned}$$

■

2.2.2. Randomized Projections

Suppose we approximate each entry of a vector using randomized projections. Then we can state the following result about the accuracy of the function $g(x) = x_i/\|x\|_1$.

Lemma 21 *For* $0 \neq X \in \mathbb{S}^n$, *let* $\widetilde{X} = \textbf{\textit{RandProj}}(X, 10240\log n/\varepsilon^2)$. *Then* $\left|\frac{\widetilde{X}_{ii}}{\operatorname{Tr}\widetilde{X}} - \frac{X_{ii}^2}{\operatorname{Tr}X^2}\right| \leq \frac{\varepsilon}{8}$.

To prove this result, we need the Johnson-Lindenstrauss lemma.

Lemma 22 (Johnson and Lindenstrauss (1984)) *For any* $0 < \varepsilon < 1$, *and any integer* n, *let* k *be a positive integer such that* $k \geq 20\log n/\varepsilon^2$. *Then for any set* V *of* n *points in* $\mathbb{R}^d$ *and random matrix* $A \in \mathbb{R}^{k\times d}$, *with high probability, for all* $x \in V$,

$$(1-\varepsilon)\|x\|_2^2 \leq \left\|(1/\sqrt{k})Ax\right\|_2^2 \leq (1+\varepsilon)\|x\|_2^2.$$

Proof [Proof of Lemma 21] Applying Lemma 22 to $\widetilde{X} = \textbf{RandProj}\left(X, \frac{10240\log n}{\varepsilon^2}\right)$, we have that with high probability, $\left|X_{ii}^2 - \widetilde{X}_{ii}\right| \leq \frac{\varepsilon}{32}\left|X_{ii}^2\right|$. Therefore, $\operatorname{Tr}X^2\left(1-\frac{\varepsilon}{32}\right) \leq \operatorname{Tr}\widetilde{X}^2 \leq \operatorname{Tr}X^2\left(1+\frac{\varepsilon}{32}\right)$. Therefore $\frac{X_{ii}^2(1-\varepsilon/32)}{\operatorname{Tr}X^2(1+\varepsilon/32)} \leq \frac{\widetilde{X}_{ii}}{\operatorname{Tr}\widetilde{X}} \leq \frac{X_{ii}^2(1+\varepsilon/32)}{\operatorname{Tr}X^2(1-\varepsilon/32)}$ which can be simplified to $\frac{X_{ii}^2}{\operatorname{Tr}X^2}(1-\varepsilon/16) \leq \frac{\widetilde{X}_{ii}}{\operatorname{Tr}\widetilde{X}} \leq \frac{X_{ii}^2}{\operatorname{Tr}X^2}(1+\varepsilon/8)$, where the last simplification is by the inequalities $\frac{1-x}{1+x} \geq 1-2x$ and $\frac{1+x}{1-x} \leq 1+4x$ for $x \in \left(0,\frac{1}{2}\right)$. Therefore we have that $\left|\frac{\widetilde{X}_{ii}}{\operatorname{Tr}\widetilde{X}} - \frac{X_{ii}^2}{\operatorname{Tr}X^2}\right| \leq (\varepsilon/8)\frac{X_{ii}^2}{\operatorname{Tr}X^2} \leq \varepsilon/8$. ■

2.2.3. Number of Iterations

From Lemmas 20 and 21 proved above, we can infer that the choice of T'' and T' in Algorithm 4 gives us the following overall error in approximating the true primal iterate.

Lemma 23 *In Algorithm 4, we have that* $\|\!|\widetilde{X}^{(t)} - X^{(t)}\|\!| \leq \frac{\varepsilon n}{4}$.

Proof The quantity we want to bound is $\|\!|\frac{n\exp\left(Y^{(t)}\right)}{\operatorname{Tr}\exp\left(Y^{(t)}\right)} - \frac{\widetilde{X}^{(t)}}{\operatorname{Tr}\widetilde{X}^{(t)}}\|\!|$. Each term is bounded as:

$$\left|\frac{n\exp\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\exp\left(Y^{(t)}\right)} - \frac{\widetilde{X}^{(t)}_{ii}}{\operatorname{Tr}\widetilde{X}^{(t)}}\right| \leq n\underbrace{\left|\frac{\exp\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\exp\left(Y^{(t)}\right)} - \frac{\widetilde{\exp}\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\widetilde{\exp}\left(Y^{t}\right)}\right|}_{\textbf{TaylorExp}\text{ error}} + \underbrace{\left|\frac{n\widetilde{\exp}\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\widetilde{\exp}\left(Y^{(t)}\right)} - \frac{n\widehat{\exp}\left(Y^{(t)}\right)_{ii}}{\operatorname{Tr}\widehat{\exp}\left(Y^{(t)}\right)}\right|}_{\textbf{RandProj}\text{ error}}.$$

Apply the results of Lemmas 20 and 21 to the right hand side terms. ■

Corollary 24 *The number of iterations for convergence of the Arora-Kale algorithm is* $\mathcal{O}(1/\varepsilon^2)$.

Proof Since the Arora-Kale algorithm only depends on the diagonal entries of X, we can assume that $\widetilde{X}$ and X match on the off-diagonal entries. Then, $\|\!|\widetilde{X}^{(t)} - X^{(t)}\|\!| \leq \frac{\varepsilon n}{4}$ is exactly equivalent to $\|\widetilde{X}^{(t)} - X^{(t)}\|_{\text{nuc}} \leq \frac{\varepsilon n}{4}$. Therefore the algorithm meets the conditions of Algorithm 3 with $\delta = \frac{\varepsilon n}{4}$. Therefore by Theorem 1, the number of outer iterations required for convergence is $\mathcal{O}(1/\varepsilon^2)$. ■

2.2.4. Combining All the Costs

Recall from Algorithm 4 that $T' = \mathcal{O}(1/\varepsilon^2)$, $T'' = \mathcal{O}(1/\varepsilon)$, and the number of iterations is $\mathcal{O}(1/\varepsilon^2)$. The cost of a matrix-vector product is $\mathcal{O}(m)$. Therefore, multiplying these costs gives $\mathcal{O}(m/\varepsilon^5)$, the claimed cost of Arora-Kale algorithm. This completes the analysis.

3. Analysis of our Proposed Algorithm

We now analyze Algorithm 1, organizing this section as follows. In Section 3.1 we derive the values of parameters that appear in the error bounds. Next, in Section 3.2, we show how we construct a polynomial to approximate the matrix exponential. In Section 3.3, we prove properties of the constructed estimators. We derive the number of inner iterations we have in Section 3.4. In Section 3.5, we establish the crucial distance invariance between true and estimated iterates, which ensures that our error is always under control. We next show in Section 3.6 why we do not need to normalize our projection step, which enables us to have a simple projection. Finally, we derive the error bound in Section 3.7.

3.1. Parameters of Mirror Map

As before, there are two parameters of the mirror map that we need to use in Theorem 1: the diameter of the constraint set as measured by it, and its strong convexity parameter.

Lemma 25 (Set Diameter) *Given $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$ and the domain $\mathcal{D} = \{X : X \succeq 0, \operatorname{Tr} X \leq K\}$, where $K \geq n$, the set diameter measured by Φ is given by $D = K \log K$.*

Lemma 11 *The function $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$ is $\frac{1}{4K}$-strongly convex with respect to the nuclear norm over the domain $\mathcal{D} = \{X : X \succeq 0, \operatorname{Tr} X \leq K\}$.*

To prove the claimed strong convexity, we need the following tools.

Definition 26 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth in norm $\|\cdot\|$ if it is continuously differentiable and satisfies $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$ for all x and y in $\mathbf{dom}\, f$.*

For functions on symmetric matrices, we use the following equivalent definition of smoothness.

Definition 27 *A function $f : \mathbb{S}^n \to \mathbb{R}$ is L-smooth in $\|\cdot\|$ if and only if for $h : \mathbb{R} \to \mathbb{R}$ defined as $h(t) = f(X + tH)$ for $H \in \mathbb{S}^n$ such that $X + tH \in \mathbf{dom}(f)$, we have $h''(0) \leq L\|H\|^2$.*

Theorem 28 (Kakade et al. (2009)) *Assume that f is a closed and convex function. Then f is β-strongly convex with respect to a norm $\|\cdot\|$ if and only if its Fenchel dual, f^*, is $\frac{1}{\beta}$-smooth with respect to the dual norm $\|\cdot\|_*$.*

Theorem 29 (Juditsky and Nemirovski (2008)) *Let Δ be an open interval on the real axis, and f be a twice differentiable function on Δ satisfying, for a certain $\theta \in \mathbb{R}$, for all $a < b$, where $a, b \in \Delta$, $\frac{f'(b)-f'(a)}{b-a} \leq \theta \frac{f''(a)+f''(b)}{2}$. Let $\mathcal{X}_n(\Delta)$ be the set of all $n \times n$ symmetric matrices with eigenvalues belonging to Δ. Then for $X \in \mathcal{X}_n(\Delta)$, the function $F(X) = \operatorname{Tr} f(X)$ is twice differentiable, and for every $H \in \mathbb{S}^n$, we have $D^2F(X)[H, H] \leq \theta \operatorname{Tr}(H f''(X) H)$.*

Theorem 30 (Lewis (1995)) *Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric (that is, $f(\sigma x) = f(x)$ for all $x \in \mathbf{dom}\, f$ and all permutations σ). Then if f is convex and lower semicontinuous, the corresponding unitarily invariant function $f \circ \lambda$ is convex and lower semicontinuous on $\mathbb{R}^{n \times n}$*

For our proof, we use definitions from Definition 12 in the following way. We first show that Ψ satisfies

$$\Psi^*(Y) = \Phi(Y), \text{ on } \{Y : Y \succeq 0, \operatorname{Tr} Y \leq K\}, \tag{3.1}$$

where $\Phi(Y) = Y \bullet \log Y - \operatorname{Tr} Y$ is the mirror map, as defined in the statement of the lemma. We then prove that Ψ is β-smooth with respect to the operator norm for a certain $\beta > 0$. Theorem 28 then immediately implies $1/\beta$-strong convexity of Ψ^* with respect to the nuclear norm. Then Equation 3.1 implies that Φ is $1/\beta$-strongly convex with respect to the nuclear norm on the domain $\{Y : Y \succeq 0, \operatorname{Tr} Y \leq K\}$, which is to be proved. We accomplish our first goal (Equation 3.1) in the following sequence of steps.

Claim 1 proves that the function ψ and its matrix version, Ψ, are both continuously differentiable at the boundary of definition of the two pieces. Claim 2 then proves that ψ_1 and ψ_2 are convex; in conjuncation with Claim 1, this implies ψ is convex. Applying Theorem 30 extends the property of convexity to Ψ. Claim 3 proves that the vector functions ψ and ϕ satisfy $\psi_1^*(x) = \phi(x)$ for $x \in \mathbb{R}_+^n$. Claim 4 proves that given an input point $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$, the point y which attains the optimum in computing $\psi_1^*(x)$ lies in the *interior* of the set $\{y : \psi_1(y) \leq 2K\}$. Claim 5 shows that $\psi^*(x) = \psi_1^*(x)$ for $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$. This is obtained by combining the results of Claim 2 and 4.

We then use these results as follows: since on the set $\{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$, we have $\psi^* = \phi$, this implies $\Psi^* = \Phi$ on the corresponding set, $\{X : X \succeq 0, \operatorname{Tr} X \leq K\}$. Next, to show smoothness of Ψ, we use Theorem 29 to compute the smoothness constants of each part of Ψ (in Claims 6 and 7), and then combine with continuous differentiability at the boundary (from Claim 1) to get the overall smoothness constant of Ψ. By the argument at the start of this proof, this immediately proves the desired strong convexity parameter. We now proceed to prove all the claims aluded to above.

Claim 1 *The functions Ψ and ψ are both continuously differentiable at the boundary.*

Proof [Proof of Claim] One can check that $\psi_1(y) = \psi_2(y)$ at the boundary. This implies continuity of the function ψ. The derivatives of the two functions are also the same at the boundary. The i-th component of the gradient is given by $\nabla_i \psi_2(y) = \frac{2K\nabla_i\psi_1(y)}{\psi_1(y)}$. At the boundary of the two parts of the function, we have $\psi_1(y) = 2K$. Substituting this into the above gradient gives $\nabla\psi_2(y) = \nabla\psi_1(y)$. This shows that ψ is continuously differentiable at the boundary. We only used chain rule of derivatives here, which applies to matrices as well, so the exactly same reasoning also gives that Ψ is continuously differentiable at the boundary. ■

Claim 2 *The functions ψ and Ψ are convex on their domains.*

Proof The function ψ is a piecewise function, each piece composed of a standard convex function (see Boyd and Vandenberghe (2004)). Combine with continuous differentiability from Claim 1 gives convexity of ψ. Applying Theorem 30 implies convexity of Ψ. ■

Claim 3 *For any input $x \in \mathbb{R}_+^n$, we have $\psi_1^*(x) = \phi(x)$.*

Proof [Proof of Claim] By definition, we have $\psi_1^*(x) = \sup_y(x^\top y - \sum_{i=1}^n \exp(y_i))$. Observe that the domain of ψ_1^* is $\mathbb{R}_+^n$ (because if there exists an input with a negative coordinate, then the corresponding coordinate of the maximizer y^* can be made to go to $-\infty$). Therefore, given an input $x \in \mathbb{R}_+^n$, the supremum is attained at y^* satisfying $x_i = \exp(y_i^*)$. This means the maximizer is $y_i^* = \log x_i$. Therefore the conjugate is $\psi_1^*(x) = \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i = \phi(x)$. ■

Claim 4 *For any x in the set $\{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$, the point $y^* = \operatorname{argmax}\left(x^T y - \psi_1(y)\right)$ lies in* **int** *$\{y : \psi_1(y) \leq 2K\}$, where* **int** *denotes the interior of the set.*

Proof [Proof of Claim] From the proof of Claim 3, for any $x \in \mathbb{R}^n_+$, we have that $y^* = \operatorname{argmax}\left(x^T y - \psi_1(y)\right)$ satisfies $y_i^* = \log x_i$ for $1 \le i \le n$. In addition to this, the statement of the lemma also requires the input x to satisfy $x_i \ge 0, \sum_{i=1}^n x_i \le K$. Plug in the values of x in terms of y in the above inequality to get $\sum_{i=1}^n \exp y_i^* \le K$, which is the same as saying $\psi_1(y^*) \le K < 2K$. This shows that the optimum, y^*, lies in **int** $\{y : \psi_1(y) \le 2K\}$. ■

Claim 5 *We have $\psi^*(x) = \psi_1^*(x)$ on all $x \in \{x : x_i \ge 0, \sum_{i=1}^n x_i \le K\}$.*

Proof [Proof of Claim] By definition of conjugate and ψ,

$$\begin{aligned} \psi^*(x) &= \sup_y x^T y - \psi(y) \qquad (3.2)\\ &= \sup_y x^T y - \begin{cases} \psi_1(y) & \text{if } \psi_1(y) \le 2K \\ \psi_2(y) & \textit{otherwise} \end{cases} \end{aligned}$$

From Claim 2, ψ is convex. Therefore the function to be maximized in Equation 3.2 is concave. From Claim 4, for input x in the set $\{x : x_i \ge 0, \sum_{i=1}^n x_i \le K\}$, we have that the maximizer $\operatorname{argmax}_y \left(x^T y - \psi_1(y)\right)$ lies in the interior of $\{y : \psi_1(y) \le 2K\}$. Therefore for input $x \in \{x : x_i \ge 0, \sum_{i=1}^n x_i \le K\}$, the maximizer of Equation 3.2 is also the same as that of $\psi_1^*(x)$. This gives $\psi^*(x) = \psi_1^*(x)$. ■

Claim 6 *The function $\Psi_1(Y)$ defined over $\{Y : \operatorname{Tr}\exp Y \le 2K\}$ is $4K$-smooth.*

Proof [Proof of Claim] Let $g\left(u\right) \stackrel{\text{def}}{=} \exp(u)$. The function g is convex and differentiable (any number of times). In particular, g'' is convex. For any a, b, applying the Mean Value theorem to some point $\zeta \in (a, b)$, convexity of g'', and $g'' \ge 0$ (due to convexity of g) gives

$$\frac{g'\left(b\right) - g'\left(a\right)}{b - a} = g''\left(\zeta\right) \le \max\left(g''\left(a\right), g''\left(b\right)\right) \le 2\frac{g''\left(a\right) + g''\left(b\right)}{2}.$$

This satisfies the right-hand side condition for Theorem 29 with $\theta = 2$; so Theorem 29 implies that on the domain $\{Y : \operatorname{Tr}\exp Y \le K\}$, for $h\left(t\right) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g\left(\lambda_i\left(Y + tH\right)\right) = \operatorname{Tr}\exp(Y + tH)$, we have,

$$\begin{aligned} h''\left(0\right) = D^2\Psi_1(Y)[H, H] &\le 2\operatorname{Tr}\left(Hg''(Y)H\right) \\ &= 2\operatorname{Tr}\left(\exp(Y)H^2\right) \\ &\le 2\operatorname{Tr}\exp(Y) \cdot \|H\|_{\text{op}}^2 \\ &\le 2 \cdot 2K \cdot \|H\|_{\text{op}}^2 \\ &= 4K\|H\|_{\text{op}}^2, \qquad (3.3) \end{aligned}$$

where we used the domain constraint for Ψ_1 in the last inequality, and the fact that matrix exponential is positive semidefinite in the first (Hölder's) inequality. By Definition 27 then, we have the lemma. ■

Claim 7 *The smoothness constant of $\Psi_2(Y)$ over the set $\{Y : \operatorname{Tr}\exp Y \geq 2K\}$ is $4K$.*

Proof For ease of exposition, let $a \overset{\text{def}}{=} 2K$. Consider the same scalar function from Claim 6, $h(t) = \operatorname{Tr}\exp(Y + tH)$ and $\ell(t) \overset{\text{def}}{=} a\log(h(t)) + 2K - 2K\log(2K)$. Then $\ell'(t) = a\frac{h'(t)}{h(t)}$ and $\ell''(t) = a\left(\frac{h''(t)}{h(t)} - \left(\frac{h'(t)}{h(t)}\right)^2\right) \leq a\frac{h''(t)}{h(t)}$. In particular,

$$\ell''(0) \leq a\frac{h''(0)}{h(0)}. \tag{3.4}$$

We already showed in Inequality 3.3 that $h''(0) = D^2\Psi_1(Y)[H,H] \leq 4K\|H\|_{\text{op}}^2$. We also have that $h(0) = \operatorname{Tr}\exp(Y) \geq 2K$ (by assumption of the lemma). Plugging these along with the value of a into Inequality 3.4 gives us $\ell''(0) \leq 2K\frac{4K}{2K} \cdot \|H\|_{\text{op}}^2 = 4K\|H\|_{\text{op}}^2$. This implies the claimed smoothness constant. ■

Proof [Proof of Lemma 11] For the functions defined in Definition 12, we can combine Claims 3 and 5 to get that $\psi^*(x) = \phi(x)$ for $x \in \{x : x_i \geq 0, \sum_{i=1}^n x_i \leq K\}$. This implies the matrix version of this statement, $\Psi^*(X) = \Phi(X)$ for $X \in \{X : X \succeq 0, \operatorname{Tr} X \leq K\}$. Next, applying Claims 1, 6, and 7, we get that the function Ψ is continuously differentiable with smoothness constant $4K$. Invoking Theorem 28, we immediately obtain that Ψ^* is strongly convex with parameter $\frac{1}{4K}$. This implies that Φ is strongly convex with the same parameter over the set $\{X : X \succeq 0, \operatorname{Tr} X \leq K\}$. ■

3.2. Chebyshev Approximation of the Matrix Exponential

In this section, we show how to construct a polynomial approximation of our matrix exponential. The standard technique to do so involves truncating the Taylor series of the matrix exponential; however, a quadratically improved bound on the number of terms required for the computation is provided by Sachdeva and Vishnoi (Sachdeva et al., 2014) using Chebyshev polynomials. We follow their notation and summarize their main results below for the sake of completeness.

3.2.1. A Brief Summary of Chebyshev Approximation

For a non-negative integer d, we denote by $T_d(x)$ the Chebyshev polynomials of degree d, defined recursively as follows:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_d(x) &= 2xT_{d-1}(x) - T_{d-2}(x). \end{aligned}$$

Let Y_i be i.i.d. variables taking values 1 and -1 each with probability 1/2. Let $D_s = \sum_{i=1}^s Y_i$, $D_0 \overset{\text{def}}{=} 0$, and

$$p_{s,d}(x) \overset{\text{def}}{=} \mathbf{E}_{Y_1,Y_2,\ldots,Y_s}\left(T_{D_s}(x)\mathbf{1}_{|D_s|\leq d}\right). \tag{3.5}$$

We can use these to construct a polynomial with degree roughly $\sqrt{s}$ that can well approximate x^s. The formal statement follows.

Theorem 31 (Theorem 3.3 in Sachdeva et al. (2014)) *For any positive integers s and d, the degree d polynomial $p_{s,d}$ defined by Equation 3.5 satisfies*

$$\sup_{x\in[-1,1]} |p_{s,d}(x) - x^s| \le 2\exp\big(-d^2/(2s)\big).$$

Using this result, define the polynomial:

$$q_{\lambda,t,d}(x) \overset{\text{def}}{=} \exp(-\lambda)\sum_{i=0}^{t} \frac{(-\lambda)^i}{i!} p_{i,d}(x). \tag{3.6}$$

Then we can use q to approximate an exponential with the following error guarantee.

Lemma 32 (Lemma 4.2 of Sachdeva et al. (2014)) *For every $\lambda > 0$ and $\delta \in (0, 1/2]$, we can choose $t = \max(\lambda, \log(1/\delta))$ and $d = \sqrt{t\log(1/\delta)}$ such that*

$$\sup_{x\in[-1,1]} |\exp(-\lambda - \lambda x) - q_{\lambda,t,d}(x)| \le \delta.$$

This is a quadratic improvement over the standard cost (degree) of approximating an exponential using truncated Taylor series. Finally, this lemma can be used to generalize the approximation from the $[-1, 1]$ interval to the interval $[0, b]$, as stated below.

Theorem 33 (Theorem 4.1 of Sachdeva et al. (2014)) *For every $b > 0$, and $0 < \delta \le 1$, there exists a polynomial $r_{b,\delta}$ that satisfies*

$$\sup_{x\in[0,b]} |\exp(-x) - r_{b,\delta}(x)| \le \delta$$

and has degree $\mathcal{O}(\sqrt{\max(b, \log(1/\delta)) \cdot \log(1/\delta)})$.

The proof of this theorem uses $\lambda \overset{\text{def}}{=} b/2$, and t and d from Lemma 32 and the polynomial

$$r_{b,\delta}(x) \overset{\text{def}}{=} q_{\lambda,t,d}\left(\frac{1}{\lambda}(x-\lambda)\right). \tag{3.7}$$

Corollary 34 (Our Chebyshev Approximation) *For every $b > 0$, $a < b$, $0 < \delta \le 1$, and $d = \sqrt{\max\left(\frac{1}{2}(b-a), \log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{1}{\delta}\right)}$, there exists a degree-$d$ polynomial $\textbf{\textit{ChebyExp}}(u, d, \delta)$ such that*

$$\sup_{u\in[a,b]} |\exp(u) - \textbf{\textit{ChebyExp}}(u, d, \delta)| \le \delta\exp(b). \tag{3.8}$$

Proof Using a simple linear transformation, Theorem 33 generalizes to:

$$\sup_{z\in[a,b]} \left|\exp\left(-\frac{1}{2}(b-a)\right)\sum_{i=0}^{t}\frac{(-\frac{1}{2}(b-a))^i}{i!} p_{i,d}(\frac{z-(b+a)/2}{(b-a)/2}) - \exp(-(z-a))\right| \le \delta.$$

By choosing $\lambda = \frac{1}{2}(b-a)$, and transforming $-z+a = u-b$, we get

$$\sup_{u\in[a,b]} \left| q_{\frac{1}{2}(b-a),t,d}\left(\frac{-u+(b+a)/2}{(b-a)/2}\right) - \exp(u-b)\right| \leq \delta.$$

Using Equation 3.7 above gives

$$\sup_{u\in[a,b]} |\exp(b) r_{b-a,\delta}(b-u) - \exp(u)| \leq \delta \exp(b).$$

Therefore, let $d = \sqrt{\max\left(\frac{1}{2}(b-a), \log\left(\frac{1}{\delta}\right)\right)\log\left(\frac{1}{\delta}\right)}$ and $\textbf{ChebyExp}(u,d,\delta) = \exp(b) r_{b-a,\delta}(b-u)$. Substitute these into the last inequality to get the statement of the lemma. ■

3.2.2. Chebyshev Approximation in Our Algorithm

We can use the above results to approximate a matrix exponential as follows. Observe that

$$\|\exp(Y) - \textbf{ChebyExp}(Y,d,\delta)\|_{\text{op}} = \max_{i\in[n]} |\exp(\lambda_i) - \textbf{ChebyExp}(\lambda_i,d,\delta)|,$$

where λ_i are the eigenvalues of Y and **ChebyExp** is the subroutine described in Corollary 34. We only need the spectrum of Y in order to complete the approximation, and that is what we proceed to derive below. Once we have the spectrum, we simply combine it with the above results to get Lemma 36.

Lemma 35 *The spectrum of Y lies in the range $\left[-\frac{1}{\varepsilon}60\log n, \log K\right]$.*

Proof Recall that $Y = -\eta\nabla f(X)$. Since we start Algorithm 1 with $Y^{(1)} = 0$, at the t-th iteration, we have $Y^{(t)} = -\sum_{i=1}^{t}\eta\nabla f\left(X^{(t)}\right)$. Plugging in the parameters displayed in Table 1, we get that the total number of iterations of the algorithm is $T_{\text{inner}} \times T_{\text{outer}} = \frac{1}{\varepsilon^3}24\times 10^5\left(\log(n/\varepsilon)\right)^{11}\log n$, the Lipschitz constant of the objective function is $\|\nabla f\|_{\text{op}} \leq 2$, and the step size is $\eta = \frac{\varepsilon^2}{8\times 10^4(\log(n/\varepsilon))^{11}}$. Multiplying these gives

$$\left\|Y^{(t)}\right\|_{\text{op}} \leq 2\cdot\frac{\varepsilon^2}{8\times 10^4 \times (\log(n/\varepsilon))^{11}}\cdot\frac{24\times 10^5\times(\log(n/\varepsilon))^{11}\log n}{\varepsilon^3} = \frac{1}{\varepsilon}60\log n.$$

Therefore, the spectrum of $Y^{(t)}$ lies in

$$\lambda(Y^{(t)}) \in \left[-\frac{1}{\varepsilon}60\log n, \frac{1}{\varepsilon}60\log n\right]. \tag{3.9}$$

We now show a better upper bound on the spectrum. Since our algorithm maintains $\operatorname{Tr} X^{(t)} \leq K$ (see Lemma 10), and $X^{(t)} = \exp\left(Y^{(t)}\right)$, it implies $\operatorname{Tr}\exp\left(Y^{(t)}\right) \leq K$. Since the matrix exponential is positive definite, this implies $\left\|\exp\left(Y^{(t)}\right)\right\|_{\text{op}} \leq K$, and therefore,

$$\lambda_{max}(Y^{(t)}) \leq \log K. \tag{3.10}$$

Combining the inclusion 3.9 and Inequality 3.10 gives the claimed bound on the spectrum. ■

Lemma 36 *In Algorithm 4, for $n \geq 2$ and $\varepsilon \leq \frac{1}{2}$, set $\mathrm{T}_{\text{Cheby}} = \frac{150}{\sqrt{\varepsilon}} \log(n/\varepsilon)$, $\delta_{\text{Cheby}} = (\varepsilon/n)^{401}$, and let $\widetilde{\exp}\left(Y^{(t)}/2\right) :=$ **ChebyExp** $\left(Y^{(t)}/2, \mathrm{T}_{\text{Cheby}}, \delta_{\text{Cheby}}\right)$. Then for all $1 \leq i \leq n$,*

$$\left|\exp\left(Y^{(t)}\right)_{ii} - \left(\widetilde{\exp} Y^{(t)}\right)_{ii}\right| \leq \delta_{\exp} \overset{\text{def}}{=} \frac{4800\varepsilon^{401}}{n^{390}}.$$

Proof We plug into Inequality 3.8 the following bounds obtained from Lemma 35:

$$a = -\frac{60 \log n}{\varepsilon}, b = \log K$$
$$u = \lambda = \tfrac{1}{2}(b-a) = \frac{\log K}{2} + \frac{30 \log n}{\varepsilon}$$

Applying Inequality 3.8, we then get

$$\sup_{\lambda \in \left[-\frac{30 \log n}{\varepsilon}, \frac{1}{2}\log K\right]} \left| K r_{\frac{1}{2}\log K + \frac{30 \log n}{\varepsilon}, \delta}\left(\tfrac{1}{2}\log K - \tfrac{1}{2}\lambda\right) - \exp\left(\tfrac{1}{2}\lambda\right)\right| \leq \delta K$$

We have $K = 40n \left(\log n\right)^{10}$; therefore, if we want the error bound to be roughly $\frac{\varepsilon}{n}$, then we need to pick $\delta = \text{polylog}(\varepsilon, n)$. Because of technical details in Lemma 41, we choose

$$\delta_{\text{Cheby}} = \left(\frac{\varepsilon}{n}\right)^{401}. \tag{3.11}$$

This gives us the following result.

$$\left\|\exp\left(Y^{(t)}/2\right) - \textbf{ChebyExp}(Y^{(t)}/2, \mathrm{T}_{\text{Cheby}}, \delta_{\text{Cheby}})\right\|_{\text{op}} \leq 40\frac{\varepsilon^{401}}{n^{396}}.$$

From Lemma 32, we get that the degree of polynomial required to achieve this guarantee is

$$\text{Required Degree } = \sqrt{\frac{2 \times 10^4}{\varepsilon} \log n \log(n/\varepsilon)} \leq \frac{150}{\sqrt{\varepsilon}} \log(n/\varepsilon).$$

This is the value of $\mathrm{T}_{\text{Cheby}}$ that we choose. We now bound the quantity we actually care about. We can write $\widetilde{\exp}\left(\frac{1}{2}Y^{(t)}\right) = \exp\left(\frac{1}{2}Y^{(t)}\right) + \Delta$, where $\|\Delta\|_{\text{op}} = 40\frac{\varepsilon^{401}}{n^{396}}$, the error guarantee obtained above. Simplifying with the application of $\left\|\exp(Y^{(t)})\right\|_{\text{op}} \leq K$ obtained from Lemma 10 gives

$$\begin{aligned}
\left\|\exp\left(Y^{(t)}\right) - \widetilde{\exp}\left(Y^{(t)}\right)\right\|_{\text{op}} &= \left\|\left(\exp\left(\tfrac{1}{2}Y^{(t)}\right)\right)^2 - \left(\widetilde{\exp}\left(\tfrac{1}{2}Y^{(t)}\right)\right)^2\right\|_{\text{op}} \\
&= \left\|\Delta^2 + \Delta \exp\left(\tfrac{1}{2}Y^{(t)}\right) + \exp\left(\tfrac{1}{2}Y^{(t)}\right)\Delta\right\|_{\text{op}} \\
&\leq (40\frac{\varepsilon^{401}}{n^{396}})^2 + 2(40\frac{\varepsilon^{401}}{n^{396}})\left\|\exp\left(\tfrac{1}{2}Y^{(t)}\right)\right\|_{\text{op}} \\
&\leq (40\frac{\varepsilon^{401}}{n^{396}})^2 + 2(40\frac{\varepsilon^{401}}{n^{396}})K \\
&\leq 3(40\frac{\varepsilon^{401}}{n^{396}})K \\
&\leq 3 \cdot \frac{40\varepsilon^{401}}{n^{396}} \cdot 40n\left(\log n\right)^{10} \\
&\leq \frac{4800\varepsilon^{401}}{n^{390}}.
\end{aligned}$$

Substituting our assumption $n \geq 4$ above gives the claimed bound. ■

In conclusion, we showed that we can approximate our matrix exponential to ε-accuracy using $\mathcal{O}(1/\sqrt{\varepsilon})$ terms in the polynomial approximation.

3.3. Properties of Estimators

Since we have an inner loop in Algorithm 1 with estimated quantities, it is crucial for the convergence that these estimators have a small bias and variance. In this section we show that this is indeed the case. We first prove two technical results about the functions **InvSqrt** and **RandProj** which are "building blocks" of our estimators. We then apply these results in proving properties of $\widehat{\theta}_1$ and $\widehat{\theta}_2$, and subsequently those of the overall estimator $\widehat{\theta}$.

3.3.1. Two Technical Results about Estimators

Lemma 8 *Consider a positive random variable x sampled from a distribution X with mean μ and variance σ^2. For some integer $k > 0$, construct the distribution $\mathcal{G}(X) =$ **InvSqrt**(X, k) defined in Equation 2.2. Then the random variable $g \sim \mathcal{G}(X)$ satisfies*

(1) $|\mathbf{E}\, g - \mu^{-1/2}| \leq \mathbf{E}\left(\frac{|x-\mu|^k}{\min(\mu,x)^{k+1/2}}\right)$

(2) $\mathbf{E}\,|g|^2 \leq k\sum_{j=0}^{k-1}\mathbf{E}\left(\frac{\left(\sigma^2+(\mu-x)^2\right)^j}{x^{2j+1}}\right)$.

Proof Recall that given a distribution $\widetilde{X}$ with a positive support, and integer $N > 0$, we define **InvSqrt** as the approximation for $g\,(u) = u^{-1/2}$ at x_0 sampled from $\widetilde{X}$:

$$\mathbf{InvSqrt}(\widetilde{X}, N) = \sum_{k=0}^{N-1} \frac{1}{k!} g^{(k)}(x_0) \prod_{j=1}^{k} (x_{k,j} - x_0), \text{ where } x_0, x_{k,j} \overset{\text{i.i.d.}}{\sim} \widetilde{X},$$

where $g^{(k)}\,(u) = \frac{(-1)^k}{2^k} u^{-j-1/2} \prod\limits_{\ell=1}^{j} (2\ell - 1)$ denotes the k-th derivative of g evaluated at u. Then the expected value of g with respect to the distribution $\mathcal{G}(X)$ is

$$\begin{aligned}
\mathbf{E}\, g &= \mathbf{E} \sum_{j=0}^{k-1} \frac{1}{j!} g^{(j)}(x) \prod_{\ell=1}^{j} (x_{j,\ell} - x) \\
&= \mathbf{E} \sum_{j=0}^{k-1} \frac{1}{j!} g^{(j)}(x) \prod_{\ell=1}^{j} (\mathbf{E}\, x_{j,\ell} - x) \\
&= \mathbf{E} \sum_{j=0}^{k-1} \frac{1}{j!} g^{(j)}(x) \,(\mu - x)^j \,. \qquad (3.12)
\end{aligned}$$

To see how the term on the right hand side of Equation 3.12 differs from the true quantity to be estimated, we apply Taylor's remainder theorem: for some point ζ lying between μ

and x, we have

$$\left|\sum_{j=0}^{k-1} \frac{1}{j!} g^{(j)}(x)\,(\mu - x)^j - \mu^{-1/2}\right| \leq \frac{g^{(k)}\left(\zeta\right)}{k!} \left|x-\mu\right|^k$$
$$\leq \frac{\left|x-\mu\right|^k}{\min\left(x,\mu\right)^{k+1/2}},$$

where the second inequality follows from

$$\left|\frac{g^{(k)}\left(u\right)}{k!}\right| \leq u^{-k-\frac{1}{2}}, \tag{3.13}$$

and the fact that ζ lies between x and μ. Combining this with Jensen's inequality gives us the final bound on the first moment,

$$\left|\mathbf{E}\, g - \mu^{-1/2}\right| \leq \mathbf{E}\left|g - \mu^{-1/2}\right| \leq \mathbf{E}\,\frac{\left|x-\mu\right|^k}{\min\left(x,\mu\right)^{k+1/2}}. \tag{3.14}$$

To prove the bound on the second moment, we again start with the definition of **InvSqrt**,

$$\begin{aligned}\mathbf{E}\left|g\right|^2 &= \mathbf{E}\left(\sum_{j=0}^{k-1}\frac{1}{j!}g^{(j)}(x)\prod_{\ell=1}^{j}\left(x_{j,\ell}-x\right)\right)^2\\ &\overset{①}{\leq} k\,\mathbf{E}\sum_{j=0}^{k-1}\left(\frac{g^{(j)}(x)}{j!}\prod_{\ell=1}^{j}\left(x_{j,\ell}-x\right)\right)^2\\ &\overset{②}{=} k\sum_{j=0}^{k-1}\mathbf{E}\left(\left(\frac{g^{(j)}(x)}{j!}\right)^2\left(\sigma^2+\left(x-\mu\right)^2\right)^j\right)\\ &\overset{③}{\leq} k\sum_{j=0}^{k-1}\mathbf{E}\left(\frac{\left(\sigma^2+\left(x-\mu\right)^2\right)^j}{x^{2j+1}}\right).\end{aligned} \tag{3.15}$$

Here ① is by Cauchy-Schwarz inequality; ② is by using the fact that each $x_{j,\ell}$ is sampled independently and adding and subtracting μ from the term inside the square and using the definition of σ^2; ③ uses Inequality 3.13. ■

Lemma 9 *Given $u \in \mathbb{R}^n$ such that $\mu \overset{\text{def}}{=} \|u\|_2^2 \neq 0$, and positive integers $k > 1$ and $N \geq 4k+6$, the following are true for x sampled from $X = \boldsymbol{RandProj}(u, N)$.*

(1) $\mathbf{E}\, x = \mu$

(2) $\sigma^2 \overset{\text{def}}{=} \mathbf{E}\left(x-\mu\right)^2 = \frac{2\mu^2}{N}$

(3) $\mathbf{E}\left(\frac{\left(\sigma^2+(x-\mu)^2\right)^k}{\min(x,\mu)^{2k+1}}\right) \leq \frac{1}{\mu}\left(\frac{e^{N/2}}{2^{N-17k}} + \frac{2^{13k}k^{2k}}{N^k}\right)$

Before diving into this proof, we state below a tool we need about logconcave distributions.

Theorem 37 (Theorem 5.22 in Lovász and Vempala (2007)) *If $X \in \mathbb{R}^n$ is a random point sampled from a logconcave distribution, then* $(\mathbf{E}\,|X|^k)^{1/k} \leq 2k\,\mathbf{E}\,|X|$.

Proof [Proof of Lemma 9]By linearity of the Gaussian distribution, given a $\zeta \sim \mathcal{N}(0, I_n)$ and for some $u \in \mathbb{R}^n$, we have $\zeta^T u \sim \mathcal{N}(0, \|u\|_2^2)$. Therefore **RandProj**$(u, N)$ gives us a scaled chi-squared distribution, $X = \frac{\mu}{N}\chi_N^2$. For a point $x \sim X$, using the parameters of a standard chi-squared distribution gives us the following properties.

$$\mathbf{E}\, x = \frac{\mu}{N} \cdot N = \mu, \text{ and } \mathbf{Var}\, x = \left(\frac{\mu}{N}\right)^2 N\left(N+2\right) - \mu^2 = 2\frac{\mu^2}{N}, \tag{3.16}$$

which proves **(1)** and **(2)**. To prove **(3)**, we first scale the random variable x by N/μ to make it of a standard chi-squared distribution; this makes our computations easier, since we later need to use the closed-form expression of the probability density function of x. After the scaling, we have

$$\mathbf{E}_{x\sim\chi_N^2}\, x = N \qquad \mathbf{Var}_{x\sim\chi_N^2} = 2N. \tag{3.17}$$

Therefore,

$$\begin{aligned}\mathbf{E}_{x\sim X}\left(\frac{\left(\sigma^2 + (\mu - x)^2\right)^k}{\min\left(x, \mu\right)^{2k+1}}\right) &\overset{\textcircled{1}}{\leq} 2^k\, \mathbf{E}_{x\sim X}\left(\frac{\sigma^{2k} + (\mu - x)^{2k}}{\min\left(x, \mu\right)^{2k+1}}\right) \\ &\overset{\textcircled{2}}{=} 2^k \frac{N}{\mu} \underbrace{\mathbf{E}_{x\sim\chi_N^2}\left(\frac{(2N)^k + (N - x)^{2k}}{\min\left(x, N\right)^{2k+1}}\right)}_{\textcircled{A}}. \end{aligned} \tag{3.18}$$

Here $\textcircled{1}$ follows from Jensen's inequality applied to the function $g(x) = x^k$ for $k > 1$ and $x > 0$; the equation $\textcircled{2}$ follows from Equation 3.16. We now bound $\textcircled{A}$ by considering the random variable in two disjoint intervals as follows.

$$\begin{aligned}\textcircled{A} &= \mathbf{E}_{x\sim\chi_N^2}\left(\frac{(2N)^k + (N - x)^{2k}}{\min\left(x, N\right)^{2k+1}}\mathbf{1}_{\left\{x < \frac{N}{4}\right\}}\right) + \mathbf{E}_{x\sim\chi_N^2}\left(\frac{(2N)^k + (N - x)^{2k}}{\min\left(x, N\right)^{2k+1}}\mathbf{1}_{\left\{x \geq \frac{N}{4}\right\}}\right). \\ &\leq \underbrace{\mathbf{E}_{x\sim\chi_N^2}\left(\frac{(2N)^k + (N - x)^{2k}}{x^{2k+1}}\mathbf{1}_{\left\{x < \frac{N}{4}\right\}}\right)}_{\textcircled{B}} + \frac{1}{(N/4)^{2k+1}}\underbrace{\mathbf{E}_{x\sim\chi_N^2}\left((2N)^k + (N - x)^{2k}\right)}_{\textcircled{C}}. \end{aligned} \tag{3.19}$$

To bound $\textcircled{\scriptsize B}$, we divide the region $\{x < N/4\}$ into intervals of geometrically-varying lengths as follows.

$$\begin{aligned} \textcircled{\scriptsize B} &= \sum_{j=2}^{\infty} \mathbf{E}_{x\sim\chi_N^2}\left(\frac{(2N)^k + (N-x)^{2k}}{x^{2k+1}}\mathbf{1}_{\left\{\frac{N}{2^{j+1}}\le x<\frac{N}{2^j}\right\}}\right) \\ &\le \sum_{j=2}^{\infty} \frac{N^{2k}5^k}{(N/2^{j+1})^{2k+1}} \underbrace{\mathbf{Prob}\left(x < N/2^j\right)}_{\textcircled{\scriptsize D}}, \end{aligned} \tag{3.20}$$

where the inequality follows from the worst case upper bounds for the numerator and $1 + 2^k \le 5^k$ for $k \ge 1$ and the worst case lower bounds for the denominator over each interval $\{N/2^{j+1} \le x < N/2^j\}$. For $a > 0$ and a random variable $x \sim \chi_N^2$, we have the following cumulative distribution function:

$$\begin{aligned} \mathbf{Prob}\left(x \le a\right) &= \int_0^a \frac{e^{-x/2}x^{N/2-1}}{2^{(N/2)}\Gamma\left(N/2\right)}dx \\ &\le \int_0^a \frac{e^{-x/2}x^{N/2-1}}{2^{N/2}(N/2e)^{(N-1)/2}}dx \\ &\le \frac{2a^{N/2-1}e^{N/2}}{N^{(N-1)/2}}, \end{aligned}$$

where we used the Sterling approximation of Gamma function in the second inequality. Substituting $a = 2^{-j}N$ above and simplifying gives the following bound on the quantity from Inequality 3.20.

$$\textcircled{\scriptsize D} \le \frac{2^{j+1}}{\sqrt{N}}\left(\frac{e}{2^j}\right)^{\frac{N}{2}}. \tag{3.21}$$

Substitute into Inequality 3.20 to get

$$\begin{aligned} \textcircled{\scriptsize B} &\le \sum_{j=2}^{\infty} N^{2k}5^k\left(\frac{2^{j+1}}{N}\right)^{2k+1}\frac{2^{j+1}}{\sqrt{N}}\left(\frac{e}{2^j}\right)^{N/2} \\ &= \frac{5^k 2^{2k+2}e^{N/2}}{N^{3/2}}\sum_{j=2}^{\infty}\frac{1}{2^{j(N/2-2k-2)}} \\ &\le \frac{2^{5k+2}e^{N/2}}{N^{3/2}}\frac{2}{2^{N-4k-4}} \\ &\le \frac{e^{N/2}}{N^{3/2}2^{N-9k-7}}, \end{aligned} \tag{3.22}$$

where we used the condition that $N \geq 4k+6$ in the first two inequalities. Next, we bound $\textcircled{C}$.

$$\begin{aligned}
\textcircled{C} &= (2N)^k \left(\mathbf{E} \left| \frac{x-N}{\sqrt{2N}} \right|^{2k} + 1 \right) \\
&\overset{\textcircled{1}}{\leq} (2N)^k \left(2^{2k} \left(2k\right)^{2k} \left(\mathbf{E} \frac{|x-N|}{\sqrt{2N}} \right)^{2k} + 1 \right) \\
&\overset{\textcircled{2}}{\leq} (2N)^k \left(2^{2k} \left(2k\right)^{2k} \left(\frac{\sqrt{\mathbf{E}\,|x-N|^2}}{\sqrt{2N}} \right)^{2k} + 1 \right) \\
&= (2N)^k \left(2^{2k} \left(2k\right)^{2k} + 1 \right) \\
&\leq (2N)^k \left(32k^2 \right)^k, \qquad (3.23)
\end{aligned}$$

where $\textcircled{1}$ is by invoking Theorem 37, which is valid by logconcavity of chi-squared distribution, and $\textcircled{2}$ is by Jensen's inequality. Plugging Inequality 3.22 and Inequality 3.23 into Equation 3.18 gives:

$$\begin{aligned}
\mathbf{E}_{x\sim X} \left(\frac{\left(\sigma^2 + (x-\mu)^2\right)^k}{\min\left(x,\mu\right)^{2k+1}} \right) &\leq 2^k \frac{N}{\mu} \left(\frac{e^{N/2}}{N^{3/2} 2^{N-9k-7}} + \frac{4^{2k+1}}{N^{2k+1}} \left(2N\right)^k \left(32k^2\right)^k \right) \\
&\leq \frac{1}{\mu} \left(\frac{e^{N/2}}{2^{N-17k}} + \frac{2^{13k} k^{2k}}{N^k} \right),
\end{aligned}$$

which is what is to be proved. ■

3.3.2. Properties of $\widehat{\theta}_1$

We prove the bounds on first and second moments of $\widehat{\theta}_1$. Note that this is where we make our choice of $\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}$ and $\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}$ for the modules **InvSqrt** and **RandProj** used in estimating θ_1 in the subroutine **Estimator1**.

Lemma 6 *Given* $\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}} = 1600\log(n/\varepsilon)$, $\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}} = 2^{14}\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}$, $Z \in \mathbb{S}^n$, *and* $\varepsilon \in (0, 1/2)$, *let* $\widetilde{Z^2} = \textbf{\textit{RandProj}}(Z, \mathrm{T}_{\mathrm{est}_{\mathrm{jl}}})$ *and* $\widehat{\theta}_{1_i} \sim \textbf{\textit{InvSqrt}}((\widetilde{Z^2})_{ii} + 1, \mathrm{T}_{\mathrm{est}_{\mathrm{isq}}})$ *for* $i \in [n]$. *Then,*

(1) *The first moment satisfies* $\left| \mathbf{E}\,\widehat{\theta}_{1_i} - \frac{1}{\sqrt{(Z^2)_{ii}+1}} \right| \leq \frac{\sqrt{2}(\varepsilon/n)^{400}}{\sqrt{(Z^2)_{ii}+1}}$.

(2) *The second moment satisfies* $\mathbf{E}\,|\widehat{\theta}_{1_i}|^2 \leq \frac{1}{(Z^2)_{ii}} 1630\log(n/\varepsilon)$.

Proof Consider a random variable x sampled from the distribution $(\widetilde{Z^2})_{ii}$. Because of Lemma 9, we have $\mathbf{E}\,x = (Z^2)_{ii}$. Then $x+1$ satisfies the required bias condition of Lemma 8 for constructing a polynomial approximation for $1/\sqrt{1+(Z^2)_{ii}}$. Then $\widehat{\theta}_{1_i}$ satisfies

$$\begin{aligned}\left|\mathbf{E}\,\widehat{\theta}_{1_i} - \frac{1}{\sqrt{1+(Z^2)_{ii}}}\right| &\overset{\textcircled{1}}{\leq} \mathbf{E}\left(\frac{\left|x-(Z^2)_{ii}\right|^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}}{\min(x+1,(Z^2)_{ii}+1)^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}+1/2}}\right)\\ &\overset{\textcircled{2}}{\leq} \sqrt{\mathbf{E}\,\frac{\left(x-(Z^2)_{ii}\right)^{2\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}}{\min\left(x+1,(Z^2)_{ii}+1\right)^{2\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}+1}}}\\ &\overset{\textcircled{3}}{\leq} \sqrt{\frac{1}{(Z^2)_{ii}+1}\left(\frac{e^{\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}/2}}{2^{\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}-17\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}} + \frac{2^{13\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}^{2\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}}{{\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}}\right)}.\end{aligned}$$

where $\textcircled{1}$ is by Lemma 8, $\textcircled{2}$ is by Jensen's inequality, and $\textcircled{3}$ is by a slight modification of the proof of (3) in Lemma 9 (instead of scaling by N/μ, we scale by $N\mu/(\mu+1)$ in the proof). Finally, set $\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}} = 1600\log\left(\frac{n}{\varepsilon}\right)$ and $\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}} = 2^{14}\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}$ to get the claimed bias. Next, we can bound the variance as follows.

$$\begin{aligned}\mathbf{E}\,|\widehat{\theta}_{1_i}|^2 &\overset{\textcircled{1}}{\leq} \mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}\mathbf{E}\left(\frac{\left(\sigma^2+\left(x-(Z^2)_{ii}\right)^2\right)^k}{(x+1)^{2k+1}}\right)\\ &\leq \mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}\mathbf{E}\left(\frac{\left(\sigma^2+\left(x-(Z^2)_{ii}\right)^2\right)^k}{\min\left(x+1,(Z^2)_{ii}+1\right)^{2k+1}}\right)\\ &\overset{\textcircled{2}}{\leq} \frac{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}{(Z^2)_{ii}}\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}\left(\frac{e^{\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}/2}}{2^{\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}-17k}} + \frac{2^{13k}k^{2k}}{\mathrm{T}^k_{\mathrm{est}_{\mathrm{jl}}}}\right)\\ &\overset{\textcircled{3}}{=} \frac{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}{(Z^2)_{ii}}\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}\left(2^{17k}\left(\frac{\sqrt{e}}{2}\right)^{2^{14}\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}} + \frac{k^{2k}}{2^k\mathrm{T}^{2k}_{\mathrm{est}_{\mathrm{isq}}}}\right)\end{aligned}$$

where $\textcircled{1}$ is by (2) in Lemma 8, $\textcircled{2}$ is by (3) in Lemma 9, and $\textcircled{3}$ is by writing $\mathrm{T}_{\mathrm{est}_{\mathrm{jl}}}$ in terms of $\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}$. We have the simplications, $\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}2^{17k}\left(\frac{\sqrt{e}}{2}\right)^{2^{14}\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}} \leq \frac{2^{17\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}}{1.2^{2^{14}\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}}2^{16}}$ and

$$\sum_{k=0}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}-1}\left(\frac{k^2}{2\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}}\right)^k \leq 1+\frac{1}{2\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}}+\frac{4}{\mathrm{T}^4_{\mathrm{est}_{\mathrm{isq}}}}+\sum_{k=3}^{\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}/2}\left(\frac{k^2}{2\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}}\right)^k+\sum_{k>\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}/2}\left(\frac{k^2}{2\mathrm{T}^2_{\mathrm{est}_{\mathrm{isq}}}}\right)^k.$$

Finally, plug in the values of $\mathrm{T}_{\mathrm{est}_{\mathrm{isq}}}$ to get the desired bound. ■

In Algorithm 1, we construct the matrix Z as an approximation to $\exp\left(\frac{1}{2}\left(Y^{(t)}+s\Delta\right)\right)$ by the subroutine **ChebyExp** $\left(\frac{1}{2}\left(Y^{(t)}+s\Delta\right), T_{\text{Cheby}}, \delta_{\text{Cheby}}\right)$, with details as provided in Lemma 36. With this value of Z and the same rest of the notation as in the above lemma, we therefore wish to compare $\mathbf{E}\,\widehat{\theta}_{1_i}$ with $\frac{1}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}}$. Note that the above lemma only tells us that we are close to $\frac{1}{\sqrt{(Z^2)_{ii}+1}}$, but Z, as defined above in Lemma 36, is only an approximation to $\exp\left(\frac{1}{2}\left(Y^{(t-1)}+s\Delta\right)\right)$. We therefore obtain the following corollary which gives us a precise bound on the bias we care about.

Corollary 38 (Bias of $\widehat{\theta}_{1_i}$) *The estimator $\widehat{\theta}_{1_i}$ described in Algorithm 2 satisfies*

$$\left|\mathbf{E}\,\widehat{\theta}_{1_i}-\frac{1}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}}\right|\leq b_{1_i}\overset{\text{def}}{=}\frac{(1+2\delta_{\text{exp}})\sqrt{2}(\frac{\varepsilon}{n})^{400}+2\delta_{\text{exp}}}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}},$$

where $\delta_{\text{exp}}=4800\frac{\varepsilon^{401}}{n^{390}}$.

Proof From Lemma 36, we know that $Z=\mathbf{ChebyExp}\left(\frac{1}{2}\left(Y^{(t-1)}+s\Delta\right),\mathrm{T}_{\text{Cheby}},\delta_{\text{Cheby}}\right)$ satisfies

$$\left|\left(\exp\left(Y^{(t-1)+s\Delta}\right)-Z^2\right)_{ii}\right|\leq\frac{4800\varepsilon^{401}}{n^{390}}.$$

For ease of notation, let $\delta_{\text{exp}}\overset{\text{def}}{=}\frac{4800\varepsilon^{401}}{n^{390}}$. Given $a-\delta\leq b\leq a+\delta$, we use the Taylor series approximation to compute the error $\frac{1}{\sqrt{a}}-\frac{1}{\sqrt{b}}$. We have:

$$\begin{aligned}\left|\frac{1}{\sqrt{a}}-\frac{1}{\sqrt{b}}\right|&\leq\left|\frac{1}{\sqrt{a}}-\frac{1}{\sqrt{-\delta+a}}\right|\\&=\frac{1}{\sqrt{a}}\left|1-\frac{1}{\sqrt{1-\delta/a}}\right|\\&\leq\frac{1}{\sqrt{a}}\frac{2\delta}{a}=\frac{2\delta}{a^{3/2}},\end{aligned}$$

where we used the Taylor approximation of $\frac{1}{\sqrt{1-x}}$ for small x. Thus, we have, from the above and Lemma 6,

$$\begin{aligned}\left|\mathbf{E}\,\widehat{\theta}_{1_i}-\frac{1}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}}\right|&\leq\frac{\sqrt{2}(\varepsilon/n)^{400}}{\sqrt{Z_{ii}^2+1}}+\frac{2\delta}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}}\\&\leq\frac{(1+2\delta)\sqrt{2}(\varepsilon/n)^{400}+2\delta}{\sqrt{\exp\left(Y^{(t-1)}+s\Delta\right)_{ii}+1}},\end{aligned}$$

which proves the claim. ■

3.3.3. Properties of $\widehat{\theta}_2$

Lemma 7 *Consider Z_1,Z_2,Z, and Δ all in $\mathbb{S}^n$. Sample $\zeta\sim\mathcal{N}(\mathbf{0},I_n)$, and define $\widehat{\theta}_2\in\mathbb{R}^n$ as $\widehat{\theta}_{2_i}=(Z_1\Delta Z_2\zeta)_i\,(Z\zeta)_i$. Define $\theta_{2_i}\overset{\text{def}}{=}(Z_1\Delta Z_2 Z)_{ii}$. Then for $i\in[n]$:*

(1) *The first moment satisfies $\mathbf{E}\,\widehat{\theta}_{2_i}=\theta_{2_i}$*

(2) *The second moment satisfies* $\mathbf{E}\,|\widehat{\theta}_{2_i}|^2 \leq 3\left(Z_1\Delta Z_2^2\Delta Z_1\right)_{ii}\left(Z^2\right)_{ii}$.

Proof The bias is defined as

$$\begin{aligned}\mathbf{E}\,\widehat{\theta}_{2_i} &= \mathbf{1}_i^T Z_1\Delta Z_2\left(\mathbf{E}\,\zeta\zeta^T\right)Z\mathbf{1}_i\\ &= \left(Z_1\Delta Z_2 Z\right)_{ii} = \theta_{2_i},\end{aligned}$$

where the second step is from the fact that $\zeta \sim \mathcal{N}(0, I_n)$ and linearity of expectation, and the last is by definition of θ_2. Next, from Lemma 45, given $a, b \in \mathbb{R}^n$ and $\zeta \sim \mathcal{N}(0, I_n)$, we conclude that $\mathbf{E}((\zeta^T a)^2(\zeta^T b)^2) \leq 3\|a\|_2^2\|b\|_2^2$. Therefore,

$$\begin{aligned}\mathbf{E}\left|\widehat{\theta}_{2_i}\right|^2 &= \mathbf{E}(\mathbf{1}_i^T Z_1\Delta Z_2\zeta)^2)(\zeta^T Z\mathbf{1}_i)^2\\ &\leq 3\|Z_2\Delta Z_1\mathbf{1}_i\|^2\|Z\mathbf{1}_i\|^2\\ &= 3(Z_1\Delta Z_2^2\Delta Z_1)_{ii}(Z^2)_{ii}.\end{aligned}$$

This proves the bound on the second moment. ■

As before, we can obtain, as a corollary of this result, a comparison of the mean of our estimator with the quantity we actually are trying to compute.

Corollary 39 (Bias of $\widehat{\theta}_{2_i}$) *The estimator $\widehat{\theta}_{2_i}$ described in Algorithm 2 satisfies*

$$\left|\mathbf{E}\,\widehat{\theta}_{2_i} - \left(\exp\left(\bar{\tau}(Y^{(t-1)} + s\Delta)\right)\Delta\exp\left((\tau - \tfrac{1}{2})(Y^{(t-1)} + s\Delta)\right)\exp\left(\tfrac{1}{2}(Y^{(t-1)} + s\Delta)\right)\right)_{ii}\right| \leq 15\delta_{\exp}\eta K$$

where $\delta_{\exp} = \frac{4800\varepsilon^{401}}{n^{390}}$.

Proof This proof simply involves writing out some matrix products and bounds on the diagonal entries of the products (using the operator norm of the individual matrices). We show this below. Let $Z_1 = \exp\left(\bar{\tau}\left(Y^{(t-1)} + s\Delta\right)\right) + U_1$, $Z_2 = \exp\left((\tau - 1/2)\left(Y^{(t-1)} + s\Delta\right)\right) + U_2$, and $Z = \exp\left(\frac{1}{2}\left(Y^{(t-1)} + s\Delta\right)\right) + U$. From Lemma 7, we have that $\mathbf{E}\,\widehat{\theta}_{2_i} = \theta_{2_i}$. We now express θ_{2_i} in terms of the matrix exponentials we care about. For ease of notation, we use $Y_s = Y^{(t-1)} + s\Delta$.

$$\begin{aligned}\mathbf{E}\,\widehat{\theta}_{2_i} - \left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau - 1/2)Y_s\right)\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii} &= \left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau - 1/2)Y_s\right)U\right)_{ii}\\ &+ \left(\exp\left(\bar{\tau}Y_s\right)\Delta U_2\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii} + \left(\exp\left(\bar{\tau}Y_s\right)\Delta U_2 U\right)_{ii}\\ &+ \left(U_1\Delta\exp\left((\tau - 1/2)Y_s\right)\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii}\\ &+ \left(U_1\Delta\exp\left((\tau - 1/2)Y_s\right)U\right)_{ii}\\ &+ \left(U_1\Delta U_2\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii} + \left(U_1\Delta U_2 U\right)_{ii}.\end{aligned}$$

We can bound this by bounding the operator norm of each of the terms. Matrix norm is sub-multiplicative, so this in turn is bounded by the operator norm of the individual terms in the matrices. From Inequality 3.10, we know that $\|\exp\left(\alpha Y_s\right)\|_{\mathrm{op}} \leq K^\alpha$, $\|\Delta\|_{\mathrm{op}} \leq \eta G$, $\|U_1\|_{\mathrm{op}} \leq \delta_{\exp}$, $\|U_2\|_{\mathrm{op}} \leq \delta_{\exp}$, and $\|U\|_{\mathrm{op}} \leq \delta_{\exp}$, where $\delta_{\exp} = \frac{4800\varepsilon^{401}}{n^{390}}$. Substituting these values here and bounding each term by the largest of all terms gives us the bound to be proved. ■

3.3.4. Properties of the Overall Estimator, $\widehat{\theta}$

Lemma 5 *The estimator $\widehat{\theta}^{(t)}$ has the following bounds on its first and second moments.*

(1) $|\mathbf{E}\,\widehat{\theta}_i - \int_{s=0}^1\int_{\tau=0}^1 \theta_{1_i}\theta_{2_i} ds d\tau| \leq b_{1_i}\theta_{2_i} + b_{2_i}\theta_{1_i} + b_{1_i}b_{2_i}$ *for* $i \in [n]$.
(2) $\mathbf{E}\,\|\widehat{\theta}\|_2^2 \leq 19600\log(n/\varepsilon)K\eta^2 + 147000K^2\eta^2\delta_{\text{exp}}$.

Proof We can get the bound on the bias by applying the results of Corollaries 38 and 39 in $\mathbf{E}\,\widehat{\theta}_i = \mathbf{E}\,\widehat{\theta}_{1_i}\,\mathbf{E}\,\widehat{\theta}_{2_i}$. We need the following definition to concisely write out expressions in this proof.

Definition 40 *Let* $\theta_{1_i} = \frac{1}{\sqrt{\exp(Y_s)_{ii}+1}}$, $\theta_{2_i} = \frac{1}{2}\left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau-\frac{1}{2})Y_s\right)\exp\left(\frac{1}{2}Y_s\right)\right)_{ii}$, $b_{1_i} = \theta_{1_i}(2\delta_{\text{exp}} + (1+2\delta_{\text{exp}})\sqrt{2}(\varepsilon/n)^{400})$, *and* $b_{2_i} = 15\delta_{\text{exp}}\eta K$ *for* $Y_s = Y^{(t-1)} + s\Delta$.

We have the following error bound.

$$\begin{aligned}\left|\mathbf{E}\,\widehat{\theta}_i - \int_{s=0}^1 \theta_{1_i}\int_{\tau=0}^1 \theta_{2_i} d\tau ds\right| &= \left|\int_{s=0}^1 \mathbf{E}\,\widehat{\theta}_{1_i}\int_{\tau=0}^1 \mathbf{E}\,\widehat{\theta}_{2_i} d\tau ds - \int_{s=0}^1 \theta_{1_i}\int_{\tau=0}^1 \theta_{2_i} d\tau ds\right| \\ &\leq \int_{s=0}^1\int_{\tau=0}^1 \left|\mathbf{E}\,\widehat{\theta}_{1_i}\,\mathbf{E}\,\widehat{\theta}_{2_i} - \theta_{1_i}\theta_{2_i}\right| d\tau ds \\ &\leq \left|\mathbf{E}\,\widehat{\theta}_{1_i}\,\mathbf{E}\,\widehat{\theta}_{2_i} - \theta_{1_i}\theta_{2_i}\right|.\end{aligned}$$

From Corollary 38, we have $\mathbf{E}\,\widehat{\theta}_{1_i} \in [\theta_{1_i} \pm b_{1_i}]$. From Corollary 39, we have $\mathbf{E}\,\widehat{\theta}_{2_i} \in [\theta_{2_i} \pm b_{2_i}]$. Therefore, the right hand side above is bounded by:

$$\left|\mathbf{E}\,\widehat{\theta}_i - \int_{s=0}^1 \theta_{1_i}\int_{\tau=0}^1 \theta_{2_i} ds d\tau\right| \leq b_{1_i}\theta_{2_i} + b_{2_i}\theta_{1_i} + b_{1_i}b_{2_i}.$$

We now compute a quantity which will be useful later:

$$\sum_{i=1}^n\left(\mathbf{E}\,\widehat{\theta}_i - \int_{s=0}^1 \theta_{1_i}\int_{\tau=0}^1 \theta_{2_i} ds d\tau\right)^2 \leq b_{1_i}^2\sum_{i=1}^n \theta_{2_i}^2 + (2b_{1_i}b_{2_i})(1+b_{1_i})\sum_{i=1}^n \theta_{2_i} + nb_{2_i}^2(1+b_{1_i})^2. \tag{3.24}$$

Here we used the fact that $\theta_{1_i} = \frac{1}{\sqrt{\exp(Y_s)_{ii}+1}} \leq 1$. We compute each of these terms separately.

$$\begin{aligned}\sum_{i=1}^n \theta_{2_i}^2 &= \sum_{i=1}^n \left(\left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau-1/2)Y_s\right)\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii}\right)^2 \\ &\overset{\text{Ⓐ}}{\leq} \sum_{i=1}^n \left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau-1/2)Y_s\right)\exp\left(\tfrac{1}{2}Y_s\right)\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left((\tau-1/2)Y_s\right)\exp\left(\tfrac{1}{2}Y_s\right)\right)_{ii} \\ &= \operatorname{Tr}\left(\exp\left(\bar{\tau}Y_s\right)\Delta\exp\left(Y_s\right)\Delta\exp\left(\tau Y_s\right)\right) \\ &= \operatorname{Tr}\left(\exp\left(Y_s\right)\Delta\exp\left(Y_s\right)\Delta\right) \\ &\leq K^2\eta^2G^2. \end{aligned} \tag{3.25}$$

Here, $\textcircled{A}$ was because $\sum_{i=1}^n (A_{ii})^2 \leq \sum_{i=1}^n (A^2)_{ii}$, which can be checked by a simple computation. Similarly, the sum in the cross-term can be computed as follows.

$$\begin{aligned}
\sum_{i=1}^n \theta_{2_i} &= \sum_{i=1}^n \left(\exp\left(\bar{\tau} Y_s\right) \Delta \exp\left((\tau - 1/2) Y_s\right) \exp\left(\tfrac{1}{2} Y_s\right)\right)_{ii} \\
&= \mathrm{Tr}\left(\exp\left(\bar{\tau} Y_s\right) \Delta \exp\left((\tau - 1/2) Y_s\right) \exp\left(\tfrac{1}{2} Y_s\right)\right) \\
&= \mathrm{Tr}\left(\exp\left(\bar{\tau} Y_s\right) \Delta \exp\left(\tau Y_s\right)\right) \\
&= \mathrm{Tr}\left(\exp\left(Y_s\right) \Delta\right) \\
&\leq K \eta G. \qquad (3.26)
\end{aligned}$$

Substituting Inequalities 3.25 and 3.26 into Inequality 3.24, and using $\frac{1}{\sqrt{\exp(Y_s)_{ii}+1}} \leq 1$ gives us:

$$\begin{aligned}
\sum_{i=1}^n \left(\mathbf{E}\, \widehat{\theta}_i - \int_{s=0}^1 a_1 \int_{\tau=0}^1 a_2 ds d\tau\right)^2 &\leq (2\delta_{\exp} + (1 + 2\delta_{\exp})\sqrt{2}(\varepsilon/n)^{400})^2 K^2 \eta^2 G^2 \\
&\quad + 900 n \delta^2 \eta^2 K^2 \\
&\quad + 60 \eta \delta K (2\delta_{\exp} + (1 + 2\delta_{\exp})\sqrt{2}(\varepsilon/n)^{400}) K \eta G \\
&\leq 6K^2 \eta^2 (\sqrt{2}(\varepsilon/n)^{400} + 2\delta_{\exp}) \\
&\leq 400 n K^2 \eta^2 (\sqrt{2}(\varepsilon/n)^{400} + 2\delta_{\exp}). \qquad (3.27)
\end{aligned}$$

We now prove the final variance bound.

$$\begin{aligned}
\mathbf{E}_{s,\tau,\zeta_1,\zeta_2} \|\widehat{\theta}\|_2^2 &= \mathbf{E}_{s,\tau,\zeta_1,\zeta_2} \sum_{i=1}^n |\widehat{\theta}_i|^2 \\
&= \int_{s=0}^1 \int_{\tau=0}^1 \sum_{i=1}^n \mathbf{E}_{\zeta_1} |\widehat{\theta}_{1_i}|^2 \, \mathbf{E}_{\zeta_2} |\widehat{\theta}_{2_i}|^2 ds d\tau.
\end{aligned}$$

Combining Lemmas 6 and 7, we get:

$$\begin{aligned}
\mathbf{E}_{s,\tau,\zeta_1,\zeta_2} \|\widehat{\theta}\|_2^2 &= \int_{s=0}^1 \int_{\tau=0}^1 \sum_{i=1}^n \underbrace{\frac{1630 \log(n/\varepsilon)}{(Z^2)_{ii}}}_{\textcircled{1}} \cdot 3 \underbrace{\left(Z_2 \Delta Z_1^2 \Delta Z_2\right)_{ii} \left(Z^2\right)_{ii}}_{\textcircled{2}} ds d\tau, \\
&\overset{\textcircled{A}}{=} \sum_{i=1}^n \int_{s=0}^1 \int_{\tau=0}^1 4890 \log(n/\varepsilon) (Z_2 \Delta Z_1^2 \Delta Z_2)_{ii} ds d\tau \\
&= 4890 \log(n/\varepsilon) \int_{s=0}^1 \int_{\tau=0}^1 \mathrm{Tr}\left(Z_2^2 \Delta Z_1^2 \Delta\right) ds d\tau, \qquad (3.28)
\end{aligned}$$

where $Z_1 = \exp\left((\tau - 1/2)\left(Y^{(t-1)} + s\Delta\right)\right) + U_1$ and $Z_2 = \exp\left(\bar{\tau}\left(Y^{(t-1)} + s\Delta\right)\right) + U_2$ as defined in Corollary 39. The term $\textcircled{A}$ shows the significance of carefully choosing the split in the estimator $\widehat{\theta}_2$, which enabled the cancellation of $\frac{1}{(Z^2)_{ii}}$ and $(Z^2)_{ii}$. We now bound $\mathrm{Tr}\left(Z_2^2 \Delta Z_1^2 \Delta\right)$. In Lemma 36 we showed how to construct Z_1 and Z_2 as $\delta_{\exp} = 4800\varepsilon^{401}/n^{390}$

approximations to the respective matrix exponentials. Thus, writing $\|U_1\|_{\mathrm{op}} = \|U_2\|_{\mathrm{op}} = \delta_{\mathrm{exp}}$ and expanding out the product $Z_2^2 \Delta Z_1^2 \Delta$ in terms of the true matrix exponentials and the error matrices, we get the following:

$$\mathrm{Tr}\Big(Z_2^2 \Delta Z_1^2 \Delta\Big) \le \mathrm{Tr}\Big(\exp\Big(2\bar{\tau}(Y^{(t-1)} + s\Delta)\Big)\Delta \exp\Big((2\tau - 1)(Y^{(t-1)} + s\Delta)\Big)\Delta\Big) + 30\eta^2 \delta_{\mathrm{exp}} K^2.$$

Choosing $A = \exp\big(Y^{(t-1)} + s\Delta\big)$ and $B = \Delta$ and combining with the fact that matrix exponential is positive semidefinite, and Δ is a symmetric matrix since the gradient of the objective is symmetric, invoking Fact 0.1 gives:

$$\mathrm{Tr}\left(Z_2^2 \Delta Z_1^2 \Delta\right) \le \mathrm{Tr}\Big(\exp\Big(Y^{(t-1)} + s\Delta\Big)\Delta^2\Big) + 30\eta^2 \delta_{\mathrm{exp}} K^2 \le 4K\eta^2 + 30\eta^2 \delta_{\mathrm{exp}} K^2,$$

where the last inequality follows from applying Holder's inequality with the nuclear norm and operator norm. Plugging this back into Equation 3.28 and completing the integration gives

$$\mathbf{E}_{s,\tau,\zeta_1,\zeta_2} \|\widehat{\theta}\|_2^2 \le 4890 \log(n/\varepsilon) \left(4K\eta^2 + 30K^2\eta^2\delta_{\mathrm{exp}}\right) \le 19600 \log(n/\varepsilon) K\eta^2 + 147000 K^2 \eta^2 \delta_{\mathrm{exp}}.$$

■

3.4. Number of Inner Iterations

We can use the general expression for overall running time to choose a value for number of 'low-accuracy' iterations. The total computational cost of the algorithm is

$$T_{\mathrm{outer}} \times \frac{10^5 \left(\log n\right)^{21}}{\varepsilon^2} T_{\mathrm{exp}} + T_{\mathrm{outer}} \times T_{\mathrm{inner}} \times 2^{30} \left(\log\left(\frac{1}{\varepsilon}\right)\right)^4 T_{\mathrm{exp}}, \tag{3.29}$$

where the first term is the total cost of exact computations, and the second term is the total cost of approximate computations (done inside the inner loop); T_{exp} is the cost of approximating the products of matrix exponentials with a vector. This is optimal (ignoring polylogarithmic terms) when setting $T_{\mathrm{inner}} = \mathcal{O}(1/\varepsilon^2)$. We set $T_{\mathrm{inner}} = 1/\varepsilon^2$ due to technical reasons arising in Lemma 41.

3.5. Distance Bound Between Estimated and True Iterates

Since the estimators in the inner loop iterations are constructed to have a low variance, the estimated and true iterates aren't far apart, as we show now. This is also where we choose the step size η.

Lemma 41 *In Algorithm 1, after $t \le T_{inner}$ iterations, we have $\mathbf{E}\, \|X^{(t)} - \widetilde{X}^{(t)}\|_{\mathrm{nuc}} \le 1.132 n\varepsilon$. Recall, $\widetilde{X}^{(t)}$ is the approximate primal iterate, while $X^{(t)}$ is the exact iterate.*

Proof By the definition of $|||\cdot|||$ and some algebra, we have

$$\begin{aligned}
\mathbf{E}\,|||X^{(t)}-\widetilde{X}^{(t)}||| &= \mathbf{E}\sum_{i=1}^{n}\left|X_{ii}^{(t)}-\widetilde{X}_{ii}^{(t)}\right|\\
&= \mathbf{E}\sum_{i=1}^{n}\left|\left(\sqrt{X_{ii}^{(t)}+1}\right)^2-\left(\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right)^2\right|\\
&= \mathbf{E}\sum_{i=1}^{n}2\sqrt{X_{ii}^{(t)}+1}\left|\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right|+\mathbf{E}\sum_{i=1}^{n}\left|\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right|^2.
\end{aligned}$$

Next, apply Cauchy-Schwarz inequality and Lemma 10 to get

$$\begin{aligned}
\mathbf{E}\,|||X^{(t)}-\widetilde{X}^{(t)}||| &\le 2\,\mathbf{E}\,\sqrt{\operatorname{Tr}X^{(t)}+n}\sqrt{\sum_{i=1}^{n}\left(\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right)^2}+\mathbf{E}\sum_{i=1}^{n}\left(\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right)^2\\
&\le 2\sqrt{K+n}\,\underbrace{\mathbf{E}\sqrt{\sum_{i=1}^{n}\left(\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right)^2}}_{\textcircled{A}}+\underbrace{\mathbf{E}\sum_{i=1}^{n}\left(\sqrt{X_{ii}^{(t)}+1}-\sqrt{\widetilde{X}_{ii}^{(t)}+1}\right)^2}_{\textcircled{B}}.
\end{aligned}\tag{3.30}$$

We first bound $\textcircled{B}$. We can write a recursive formulation for as follows.

$$\sqrt{\widetilde{X}_{ii}^{(t)}+1}-\sqrt{X_{ii}^{(t)}+1}=\underbrace{\left(\sqrt{\widetilde{X}_{ii}^{(0)}+1}-\sqrt{X_{ii}^{(0)}+1}\right)}_{\textcircled{C}}+\underbrace{\sum_{s=1}^{t}\left(\widehat{\theta}_i^{(s)}-\sqrt{X_{ii}^{(s)}+1}+\sqrt{X_{ii}^{(s-1)}+1}\right)}_{\textcircled{D}}.$$

We invoke Johnson-Lindenstrauss lemma (restated in Lemma 22 for completeness) and choose the accuracy parameter for it to be such that $\left|X_{ii}^{(0)}-\widetilde{X}_{ii}^{(0)}\right|\le\widetilde{\varepsilon}X_{ii}^{(0)}=\frac{\varepsilon}{100(\log n)^{10}}X_{ii}^{(0)}$. Therefore, $\textcircled{C}\le\frac{\widetilde{\varepsilon}}{2}\sqrt{X_{ii}^{(0)}+1}=\frac{\varepsilon}{200(\log n)^{10}}\sqrt{X_{ii}^{(0)}+1}$. Summing over all indices and taking expectations gives

$$\begin{aligned}
\textcircled{B} &\le \mathbf{E}\sum_{i=1}^{n}\left(\frac{\varepsilon}{200\,(\log n)^{10}}\sqrt{X_{ii}^{(0)}+1}+\sum_{s=1}^{t}\left(\widehat{\theta}_i^{(s)}-\sqrt{X_{ii}^{(s)}+1}+\sqrt{X_{ii}^{(s-1)}+1}\right)\right)^2\\
&\overset{\textcircled{1}}{\le} 2\frac{\varepsilon^2}{40000\,(\log n)^{20}}(\operatorname{Tr}X^{(0)}+n)+2\,\mathbf{E}\left\|\sum_{s=1}^{t}\left(\widehat{\theta}^{(s)}-\sqrt{\mathbf{diag}\left(X^{(s)}\right)+\mathbf{1}}+\sqrt{\mathbf{diag}\left(X^{(s-1)}\right)+\mathbf{1}}\right)\right\|_2^2\\
&\overset{\textcircled{2}}{\le} \frac{K\varepsilon^2}{10000\,(\log n)^{20}}+\underbrace{2\,\mathbf{E}\left\|\sum_{s=1}^{t}\left(\widehat{\theta}^{(s)}-\sqrt{\mathbf{diag}\left(X^{(s)}\right)+\mathbf{1}}+\sqrt{\mathbf{diag}\left(X^{(s-1)}\right)+\mathbf{1}}\right)\right\|_2^2}_{\textcircled{E}},
\end{aligned}$$

where $\textcircled{1}$ is by Cauchy-Schwarz inequality, and $\textcircled{2}$ by Lemma 10. A subtle point here is that even though the very first iterate in the algorithm satisfies a stronger inequality, namely, $\operatorname{Tr} X^{(0)} \leq n$, we *cannot* use this stronger bound because we care about *all* iterations, and this stronger bound doesn't hold later on. We now bound $\textcircled{E}$ below. Note that since the random variable $\widehat{\theta}^{(s)}$ is not entirely unbiased, the term $\textcircled{E}$ is not the variance. Let $\theta^{(s)} \stackrel{\text{def}}{=} \mathbf{E}\,\widehat{\theta}^{(s)}$ and $d^{(s)} = \sqrt{\mathbf{diag}\left(X^{(s)}\right)+\mathbf{1}} - \sqrt{\mathbf{diag}\left(X^{(s-1)}\right)+\mathbf{1}}$. Then,

$$
\begin{aligned}
\textcircled{E} &= \mathbf{E}\left\|\sum_{s=1}^{t}\left(\widehat{\theta}^{(s)} - \left(\sqrt{\mathbf{diag}\left(X^{(s)}\right)+\mathbf{1}} - \sqrt{\mathbf{diag}\left(X^{(s-1)}\right)+\mathbf{1}}\right)\right)\right\|_2^2 \\
&= \mathbf{E}\left\|\sum_{s=1}^{t}\left(\widehat{\theta}^{(s)} - \theta^{(s)} + \theta^{(s)} - d^{(s)}\right)\right\|_2^2 \\
&= \mathbf{E}\sum_{i=1}^{n}\left(\sum_{s=1}^{t}\left(\widehat{\theta}_i^{(s)} - \theta_i^{(s)}\right)^2 + \sum_{s=1}^{t}\left(\theta_i^{(s)} - d_i^{(s)}\right)^2 + 2\sum_{s\neq\ell}\left(\widehat{\theta}_i^{(s)} - \theta_i^{(s)}\right)\left(\theta_i^{(\ell)} - d_i^{(\ell)}\right)\right) \\
&= \sum_{s=1}^{t}\mathbf{E}\left\|\widehat{\theta}^{(s)} - \theta^{(s)}\right\|_2^2 + \sum_{s=1}^{t}\underbrace{\sum_{i=1}^{n}\left(\theta_i^{(s)} - d_i^{(s)}\right)^2}_{\textcircled{F}} + 0 \\
&\leq \sum_{s=1}^{t}\left(\mathbf{E}\left\|\widehat{\theta}^{(s)}\right\|^2 + \textcircled{F}\right),
\end{aligned}
$$

where the last step is by the bound on variance by its second moment. Recall that we already have from Inequality 3.27, $\textcircled{F} \leq 400nK^2\eta^2(\sqrt{2}(\varepsilon/n)^{400} + 2\delta_{\text{exp}})$. Substitute this into the bound for $\textcircled{E}$ and $\textcircled{B}$, and apply the result of Lemma 5 to bound $\mathbf{E}\,\|\widehat{\theta}^{(s)}\|_2^2$; we choose $t = T_{\text{inner}} = \frac{1}{\varepsilon^2}$ and get

$$
\textcircled{B} \leq \underbrace{\frac{K\varepsilon^2}{10000\,(\log n)^{20}} + \frac{1}{\varepsilon^2}\left(\underbrace{19600\log(n/\varepsilon)K\eta^2 + 147000K^2\eta^2\delta_{\text{exp}}}_{\text{second-moment bound from Lemma 5}} + \underbrace{400nK^2\eta^2\left(\sqrt{2}(\varepsilon/n)^{400} + 2\delta\right)}_{\text{squared error in bias}}\right)}_{\textcircled{G}}. \tag{3.31}
$$

Next, we bound $\textcircled{A}$ using Jensen's inequality, and use Inequality 3.31 in Inequality 3.30 to get

$$
\mathbf{E}\,\|\!|X^{(t)} - \widetilde{X}^{(t)}|\!\| \leq 2\sqrt{K+n}\sqrt{\textcircled{G}} + \textcircled{G}. \tag{3.32}
$$

Note that to bound $\textcircled{G}$, we only need to take care of the second term in Inequality 3.31, because the first term is already fixed, and the remaining can be fixed by appropriate choices of δ_{exp}. We choose the step size to be

$$
\eta = \varepsilon^2\frac{1}{8\times 10^4(\log(n/\varepsilon))^{11}}. \tag{3.33}
$$

Substituting this in Inequality 3.31 gives

$$\textcircled{G} \le \frac{K\varepsilon^2}{10^4 (\log n)^{20}} + \frac{K\varepsilon^2}{6\times 10^5 (\log(n/\varepsilon))^{21}} + \frac{K\varepsilon^2 n\delta_{\exp}}{2500 (\log(n/\varepsilon))^{12}} + \frac{K\varepsilon^2 n^2 \left(\sqrt{2}(\varepsilon/n)^{400} + 2\delta_{\exp}\right)}{4\times 10^5 \times (\log(n/\varepsilon))^{12}}.$$

Plugging this back into Lemma 3.32 with the value of $\delta_{\exp}$ from Definition 4 gives:

$$\begin{aligned}
\textcircled{G} &\le \frac{K\varepsilon^2}{10^4 (\log n)^{20}} + \frac{K\varepsilon^2}{6\times 10^5 (\log n)^{21}} + \frac{2K\varepsilon^{403}}{(\log(n/\varepsilon))^{12} n^{389}} + \frac{3K\varepsilon^{402}}{41 (\log(n/\varepsilon))^{12} n^{388}} \\
&\le K\varepsilon^2 \left(\frac{1}{10^4 (\log n)^{20}} + \frac{1}{6\times 10^5 (\log(n/\varepsilon))^{21}} + \frac{2\varepsilon^{401}}{(\log(n/\varepsilon))^{12} n^{389}} + \frac{3\varepsilon^{402}}{41 n^{388} (\log(n/\varepsilon))^{12}} \right) \\
&\le K\varepsilon^2 \left(\frac{1}{5\times 10^3 (\log n)^{20}} + \frac{6\varepsilon^{401}}{(\log n)^{20} n^{380}} \right) \\
&\le \frac{K\varepsilon^2}{4999 (\log n)^{20}}
\end{aligned}$$

Plugging this back into Inequality 3.32 and using $K = 40n(\log n)^{10}$ gives $\mathbf{E}\, |||X^{(t)} - \widetilde{X}^{(t)}||| \le 1.132 n\varepsilon$. Since Algorithm 1 only uses the diagonal entries of $\widetilde{X}^{(t)}$ at any iteration t, we can assume the off-diagonal entries exactly equal those in $X^{(t)}$. Therefore $\widetilde{X}^{(t)} - X^{(t)}$ is a diagonal matrix. For a diagonal matrix A, we can see that $|||A||| = \|A\|_{\text{nuc}}$. Therefore, we have $\mathbf{E}\, \|X^{(t)} - \widetilde{X}^{(t)}\|_{\text{nuc}} \le 1.132 n\varepsilon$. ∎

3.6. The Expanded Domain Trick for Projection

The goal of this section is two-fold: first, we show that if the trace constraint is inactive, the projection step is simple and requires no trace normalization; second, we prove that the trace constraint remains inactive throughout the run of our algorithm. We remark that this is also the lemma where we choose the optimal number of iterations in the outer loop of Algorithm 1.

Lemma 42 *Consider the mirror map $\Phi(X) = X \bullet \log X - \operatorname{Tr} X$ over the domain $\{X : X \succeq 0, \operatorname{Tr} X \le K\}$. Assuming that the trace inequality is never active, we have that $\exp Y = \operatorname{argmin}_{X\succeq 0, \operatorname{Tr} X \le K} \Phi(X) - Y \bullet X$.*

Proof We wish to solve

$$\min X \bullet \log X - \operatorname{Tr} X - X \bullet Y, \text{ subject to } X \succeq 0, \operatorname{Tr} X \le K. \tag{3.34}$$

By diagonalizing X as $X = U\Lambda U^\top$ and Y as $Y = V\Sigma V^\top$, we can rewrite this problem as

$$\min \sum_{i=1}^n \lambda_i \log \lambda_i - \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \lambda_i \widetilde{y}_i, \text{ subject to } \lambda_i \ge 0, \sum_{i=1}^n \lambda_i \le K, \tag{3.35}$$

where $\widetilde{y}_i$ is the i'th diagonal entry of the matrix $U^\top Y U$. The Lagrangian is given by $\mathcal{L}(\lambda_i, \nu) = \sum_{i=1}^n \lambda_i \log \lambda_i - \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \lambda_i \widetilde{y}_i + \nu \left(\sum_{i=1}^n \lambda_i - K\right)$. Setting the gradient to zero gives $\nabla_\Lambda \mathcal{L} = \mathbf{1} + \log\lambda^* - \mathbf{1} - \widetilde{y} + \nu\mathbf{1} = 0$, which gives $\lambda_i^* = \exp(\widetilde{y}_i - \nu)$ for all i. Since we assumed that the trace constraint is *not* active, it means, by complementary slackness, $\nu = 0$ (note that this assumption is justified because we prove it in Lemma 10). This gives $\lambda_i^* = \exp(\widetilde{y}_i)$ which translates to $X^* = \exp(Y)$, as claimed. ■

Before we start the second proof, we need the following result.

Lemma 43 *Fix a norm $\|\cdot\|$. Given an α-strongly convex mirror map $\Phi : \mathcal{D} \to \mathbb{R}$, a convex, G-Lipschitz objective $f : \mathcal{X} \to \mathbb{R}$, the diameter of $\mathcal{X} \cap \mathcal{D}$ denoted by $D \overset{\text{def}}{=} \sup_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(X) - \inf_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x)$, step size η, and parameter δ' where $\mathbf{E}\left\|x^{(t)} - \widetilde{x}^{(t)}\right\| \le \delta'$, running mirror descent for T iterations gives iterates $\{\widetilde{x}^{(t)}\}_{t=1}^T$ that satisfy the inequality*

$$f\left(\frac{1}{T-1}\sum_{t=1}^{T-1}\widetilde{x}^{(t)}\right) - f(x^*) \le \frac{\eta G^2}{2\alpha} + \frac{1}{\eta(T-1)}\left(D_\Phi(x^*, \widetilde{x}^{(1)}) - D_\Phi(x^*, \widetilde{x}^{(T)})\right) + \delta' G.$$

This can be derived the same way as Theorem 4.2 in Bubeck et al. (2015), by incorporating the error in iterate, just as we did in the proof of Theorem 1.

Lemma 10 *With the choice of parameters in Algorithm 1, the iterate $\widetilde{X}^{(t)}$ at any iteration t satisfies* $\operatorname{Tr}\widetilde{X}^{(t)} < K$ *for* $K = 40n(\log n)^{10}$.

Proof We prove this by induction on the iteration count.

Induction Hypothesis. We assume that for any iteration t, the primal iterate is not too far from the optimal point, satisfying $|||\widetilde{X}^{(t)} - X^*||| \le 38n(\log n)^{10}$.

Base Case. Since $Y^{(1)} = 0$, the primal iterate $\widetilde{X}^{(1)} = I$. We also know that the optimal point satisfies $\operatorname{Tr} X^* = n$. Therefore, $|||\widetilde{X}^{(1)} - X^*||| \le 2n \le 38n(\log n)^{10}$. The hypothesis is thus true for the base case, $t = 1$.

Induction. Suppose that the hypothesis is true for some $t = t'$. We prove that this would make it true for $t = t' + 1$ as well. Our technique is to first prove a weak bound for $|||\widetilde{X}^{(t)} - X^*|||$ using triangle inequality of norms; then we boost our bound (and obtain the stronger guarantee of the induction hypothesis) by invoking strong convexity of Bregman Divergence. We now show the details.

$$\begin{aligned}
|||\widetilde{X}^{(t'+1)} - X^*||| &\le |||\widetilde{X}^{(t'+1)} - \widetilde{X}^{(t')}||| + |||\widetilde{X}^{(t')} - X^*||| \\
&\le \underbrace{\left\|\widetilde{X}^{(t'+1)} - \widetilde{X}^{(t')}\right\|_{\text{nuc}}}_{\text{Inequality 1.6}} + \underbrace{|||\widetilde{X}^{(t')} - X^*|||}_{\text{induction hypothesis}} . \\
&\le \underbrace{\frac{2\eta G}{\alpha}}_{\text{Ⓐ}} + 38n(\log n)^{10} .
\end{aligned} \tag{3.36}$$

The first step here used the fact that $|||M||| \le \|M\|_{\text{nuc}}$(We can show this by Hölder's Inequality, $\langle X, Y\rangle \le \|Y\|_{\text{op}}\|X\|_{\text{nuc}}$. Select $Y = \mathbf{diag}(\operatorname{sgn}(\mathbf{diag}\, X))$, that is, Y is a diagonal matrix with $Y_{ii} = \operatorname{sgn}(X_{ii})$). We can plug in parameters of the mirror map and the step size, as displayed in Table 1, to obtain:

$$\text{Ⓐ} = 2 \cdot \frac{\varepsilon^2}{80000(\log(n/\varepsilon))^{11}} \cdot 2 \cdot 4(40n(\log n)^{10}) \le \frac{n\varepsilon^2}{125}.$$

Plugging this back into Equation 3.36 while using $\varepsilon < 1/2$ and $K = 40n(\log n)^{10}$ gives $|||\widetilde{X}^{(t'+1)} - X^*||| \le \frac{n\varepsilon^2}{125} + 38n\left(\log n\right)^{10}$, which implies that $\operatorname{Tr}\left(\widetilde{X}^{(t')}\right) < (n(\varepsilon^2/125 + 38(\log n)^{10}) + n) < 40n(\log n)^{10} = K$, which says that the trace constraint on the iterates is not active on the first t' iterations.

Since the trace constraint is not active on the first t' iterations, the projection step does not require a normalization. This implies that Algorithm 3 now is identical to Approximate Mirror Descent with this mirror map and objective. We now recall Lemma 43 for $T = t'+1$:

$$f\left(\frac{1}{t'}\sum_{t=1}^{t'}\widetilde{X}^{(t)}\right) - f\left(X^*\right) \le \frac{\eta G^2}{2\alpha} + \frac{1}{\eta t'}(D_\Phi(X^*, \widetilde{X}^{(1)}) - D_\Phi(X^*, \widetilde{X}^{(t'+1)})) + \delta' G.$$

Multiplying throughout by $\eta t'$ and rearranging the terms gives

$$D_\Phi(X^*, \widetilde{X}^{(t'+1)}) \le \frac{\eta^2 G^2 t'}{2\alpha} + D_\Phi(X^*, \widetilde{X}^{(1)}) - \eta t' \underbrace{\left(f\left(\frac{1}{t'}\sum_{t=1}^{t'}\widetilde{X}^{(t)}\right) - f\left(X^*\right)\right)}_{\text{positive}} + \eta t' \delta' G \tag{3.37}$$

Since Φ is α-strongly convex in the nuclear norm, we have $D_\Phi(X^*, \widetilde{X}) \ge \frac{\alpha}{2}\|X^* - \widetilde{X}\|_{\text{nuc}}^2$. Since this is at least $\frac{\alpha}{2}|||X^* - \widetilde{X}|||^2$. Chaining this with Inequality 3.37 gives

$$|||\widetilde{X}^{(t'+1)} - X^*|||^2 \le \underbrace{\frac{\eta^2 G^2 t'}{\alpha^2}}_{\textcircled{B}} + \underbrace{\frac{2D_\Phi(X^*, \widetilde{X}^{(1)})}{\alpha}}_{\textcircled{C}} + \underbrace{\frac{2}{\alpha}\eta t' \delta' G}_{\textcircled{D}}, \tag{3.38}$$

We now bound each of the terms on the right-hand side. We remark that this is actually where we choose the appropriate value of T_{outer}.

$$\begin{aligned}
\textcircled{B} &= \frac{\eta^2 G^2 T_{\text{inner}} T_{\text{outer}}}{\alpha^2} \\
&= \frac{\varepsilon^4}{64 \times 10^8 \left(\log(n/\varepsilon)\right)^{22}} \cdot 4 \cdot \frac{1}{\varepsilon^2} \cdot \frac{1}{\varepsilon} 24 \times 10^5 \left(\log(n/\varepsilon)\right)^{11} \log n \cdot 16 \left(40n \left(\log n\right)^{10}\right)^2 \\
&\le 40\varepsilon n^2 \left(\log n\right)^{10}
\end{aligned}$$

To bound the second term $\textcircled{C} = \frac{2D_\Phi(\widetilde{X}^{(1)}, X^*)}{\alpha}$, we need to compute $D_\Phi(\widetilde{X}^{(1)}, X^*)$. Recall that $\widetilde{X}^{(1)} = I$ by our algorithm. Therefore, $\Phi(\widetilde{X}^{(1)}) = -n$ and $\nabla\Phi(\widetilde{X}^{(1)}) = 0$. Applying Hölder's inequality gives $\Phi(X^*) \le \operatorname{Tr} X^* \log \|X^*\|_{\text{op}} \le n \log n$. Therefore $D_\Phi(X^*, \widetilde{X}^{(1)}) \le n \log n$. Now we go back to the quantity we were trying to bound:

$$\textcircled{C} \le 2 \cdot n \log n \cdot 4(40n(\log n)^{10}) \le 320n^2 \left(\log n\right)^{11}.$$

Finally, the last term is:

$$\textcircled{D} = \frac{2}{\alpha}\eta T_{\text{inner}} T_{\text{outer}} \delta' G \le 2 \cdot 4K \cdot \frac{30 \log n}{\varepsilon} \cdot 1.132 n\varepsilon \cdot 2 = 21735 n^2 \left(\log n\right)^{11}$$

Summing these terms and plugging back into Inequality 3.38 gives

$$\begin{aligned}\left\|\kern-0.1em\left|\widetilde{X}^{(t'+1)} - X^*\right|\kern-0.1em\right\|^2 &\leq n^2(40\varepsilon(\log n)^{10} + 320(\log n)^{11} + 21735(\log n)^{11}).\\ &< n^2(0.77(\log n)^{20} + 17(\log n)^{20} + 1150(\log n)^{20})\\ &\leq 1168n^2\,(\log n)^{20} \leq 35n\,(\log n)^{10}\,,\end{aligned}$$

which completes the induction. Therefore we have $|||\widetilde{X}^{(t)} - X^*||| \leq 38n\,(\log n)^{10}$ for all t. Since $\operatorname{Tr} X^* = n$, this proves $\operatorname{Tr} \widetilde{X}^{(t)} < 40n\,(\log n)^{10} = K$. ■

3.7. Error bound

Finally, we put together all the parameters derived above to obtain our claimed error bound.

Lemma 44 *Running Algorithm 1 gives an output for (1.2) that has an error bound of $K\varepsilon$.*

Our algorithm is in the framework of approximate lazy mirror descent, with error bound given by Theorem 1, restated below.

Theorem 1 (Convergence of Lazy Mirror Descent) *Fix a norm $\|\cdot\|$. Given an α-strongly convex mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ and a convex, G-Lipschitz objective $f : \mathcal{X} \to \mathbb{R}$, run Algorithm 3 with step size η and $\mathbf{E}\,\|x^{(t)} - \widetilde{x}^{(t)}\| \leq \delta$. Let $D \overset{\text{def}}{=} \sup_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x) - \inf_{x\in\mathcal{X}\cap\mathcal{D}} \Phi(x)$. Then, Algorithm 3 after T iterations returns $\widetilde{x}^{t^*}$, satisfying*

$$\mathbf{E}\, f(\widetilde{x}^{(t^*)}) - f(x^*) \leq \frac{D}{T\eta} + \frac{2\eta G^2}{\alpha} + \delta G. \tag{1.7}$$

Proof Our proof involves plugging in the values of the parameters (from Table 1) in the above bound. Since we assume $n \geq 4$, we use $\log n \leq \sqrt{n}$ in one of the calculations below.

$$\begin{aligned}\frac{D}{T\eta} &= K\varepsilon\frac{\log K}{30\log n} \leq K\varepsilon\frac{\log 40 + 6\log n}{30\log n} \leq 0.29K\varepsilon.\\ \frac{2\eta G^2}{\alpha} &= \frac{32\varepsilon^2 K}{8\times 10^4\,(\log n)^{11}} = \frac{K\varepsilon}{2500\,(\log n)^{11}} \leq 2\times 10^{-5}K\varepsilon\\ \delta G &= 1.132n\varepsilon \leq \frac{K\varepsilon}{35\,(\log n)^{10}} \leq 11\times 10^{-4}K\varepsilon\end{aligned}$$

Summing these quantities gives the upper bound on the error to be εK, as claimed. ■

4. General Technical Results

Lemma 45 *Given $a, b \in \mathbb{R}^n$, we have that $\mathbf{E}_{\zeta\sim\mathcal{N}(0,I)}\left((\zeta^T a)^2(\zeta^T b)^2\right) \leq 3\|a\|_2^2\|b\|_2^2$.*

Proof By Cauchy-Schwarz inequality, the functions f_1 and f_2 satisfy $\mathbf{E}_{\zeta\sim\mathcal{N}(0,I)}(f_1(\zeta)\,f_2(\zeta)) \leq \sqrt{\mathbf{E}_\zeta(f_1(\zeta))^2\,\mathbf{E}_\zeta(f_2(\zeta))^2}$. Choose $f_1(\zeta) = (\zeta^T a)^2$ and $f_2(\zeta) = (\zeta^T b)^2$. Since $\zeta \sim \mathcal{N}(0, I)$ and all the coordinates of ζ are independent, $\mathbf{Var}(\zeta^T a) = \sum_{i=1}^n \mathbf{Var}(\zeta_i a_i) = \sum_{i=1}^n a_i^2 = \|a\|_2^2$. Therefore $\zeta^T a \sim \mathcal{N}(0, \|a\|_2^2)$. For $X \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbf{E}\, X^4 = 3\sigma^4$. Applying this to $\zeta^T a$ and $\zeta^T b$ proves the desired inequality. ■