1 Algorithms for solving linear systems of equations

One such application of Phase Estimation (Section **??**) is with respect to solving linear systems of equations. This is the so-called HHL algorithm [**?**].

The general problem statement of a linear system is if we are given matrix A and unit vector $\vec{b}$, then find $\vec{x}$ satisfying, $A\vec{x} = \vec{b}$.

However, assume that instead of solving for x itself, we instead solve for an expectation value $x^T M x$ for some linear operator M. Hence, one can show that our algorithm has a runtime bound of $O(\log(N)\kappa^2)$, if we can further assume that the linear system is sparse and has a low condition number κ.

So, assume that A in our linear system is an $N \times N$ Hermitian matrix. Notice that this is an "unrestrictive" constraint on A because we can always take non-Hermitian matrix A' and linear system $A'\vec{x} = \vec{b}$ and instead solve $\begin{bmatrix} 0 & A' \\ A'^{\dagger} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$. Hence, we we will assume that A is Hermitian from here on.

Recall that because A is hermitian $\Rightarrow$ we can perform quantum phase estimation using e^{-iAt} as the unitary transformation. This can be done efficiently if A is sparse.

So, we first prepare $|b\rangle$ (the representation of $\vec{b}$). We assume that this can be done efficiently or that $|b\rangle$ is supplied as an input.

Denote by $|\psi_j\rangle$ the eigenvectors of A with associated eigenvalues λ_j. Hence, we can express $|b\rangle$ as $|b\rangle = \sum_j \beta_j |\psi_j\rangle$. So, we initialize a first register to state $\sum_j \beta_j |\psi_j\rangle$ and second register to state $|0\rangle$. After applying phase estimation, we then have the joint state $\sum_j \beta_j |\psi_j\rangle \left|\widetilde{\lambda}_j\right\rangle$, where $\widetilde{\lambda}_j$ is an approximation of λ_j. We'll assume that this approximation is perfect from here on.

Next we add an ancilla qubit and perform a rotation conditional on the first register while now holds $|\lambda_j\rangle$. The rotation transforms the system to

$$\sum_j \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1 - \frac{C^2}{\lambda_j^2}} |0\rangle + \frac{C}{\lambda_j} |1\rangle \right)$$

for some small constant $C \in \mathbb{R}$ that is $O(1/\kappa)$.

Hence, we can undo phase estimation to restore the second register to $|0\rangle$.

Now, if we measure the ancillary qubit in the computational basis, we'll evidently collapse the state to $|1\rangle$ with some probability. We'd then have

$$\sum_j \frac{C}{\lambda_j} \beta_j |\psi_j\rangle |\lambda_j\rangle |1\rangle = C(A^{-1} |b\rangle)$$

In particular, the probability of getting this result is

$$
\begin{aligned}
p(-1) &= \left(\sum_j \beta_j \langle\psi_j| \langle\lambda_j| \left(\sqrt{1-\frac{C^2}{\lambda_j^2}} \langle 0| + \frac{C}{\lambda_j} \langle 1|\right)\right) |1\rangle \langle 1| \left(\sum_j \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1-\frac{C^2}{\lambda_j^2}} |0\rangle + \frac{C}{\lambda_j} |1\rangle\right)\right) \\
&= \sum_j \beta_j \langle\psi_j| \langle\lambda_j| \left(\sqrt{1-\frac{C^2}{\lambda_j^2}} \langle 0| + \frac{C}{\lambda_j} \langle 1|\right) |1\rangle \langle 1| \beta_j |\psi_j\rangle |\lambda_j\rangle \left(\sqrt{1-\frac{C^2}{\lambda_j^2}} |0\rangle + \frac{C}{\lambda_j} |1\rangle\right) \\
&= \sum_j \beta_j \langle\psi_j| \langle\lambda_j| \frac{C}{\lambda_j} \langle 1|1\rangle \langle 1| \beta_j |\psi_j\rangle |\lambda_j\rangle \frac{C}{\lambda_j} |1\rangle \\
&= \sum_j \beta_j^2 \frac{C^2}{\lambda_j^2} \\
&= \|A^{-1} |b\rangle \|^2 C^2 = O(1/\kappa^4)
\end{aligned}
$$

Finally, we can make a measurement M whose expectation value $\langle x| M |x\rangle$ corresponds to the feature of x we wish to evaluate.

2 Supervised learning with quantum enhanced feature spaces

2.1 Prelude

We are given data from a training set T and a test set S of a subset $\Omega \subset \mathbb{R}^d$. We assume that S and T are drawn from the same input space X. Furthermore, there exists output space $Y = \{-1, +1\}$ and a distribution D on $X \times Y$.

Now, suppose we have a labelling $m : T \cup S \to Y$. Our goal is to use this information to find some approximation function $\tilde{f} : X \to Y$ that minimizes estimation error for function class F. In other words, let true risk for function f be defined as

$$R^{true}(f) = P_{X,Y \sim D}(f(X) \neq Y)$$

Then, estimation error is the difference in true risk between $\tilde{f}$ and optimal choice $f^* = \inf_{f \in F} R^{true}(f)$.

One classical method is using so-called Support Vector Machines (SVM), which construct a separating hyperplane such that the distance to the nearest training observation (minimum margin) is maximized. Much of the popularity of SVMs can be attributed to its association with the "kernel trick" which maps the data to a higher dimensional space so that it is separable or approximately separable.

Here, we suppose that the data is given classically and we seek to show that, in some cases, we can obtain a quantum advantage by either generating the separating hyperplane in quantum feature space or simply estimating the kernel function.

2.2 Feature Map

Consider the feature vector kernel $K(x,z) = |\langle \Phi(x)|\Phi(z)\rangle|^2$

3 Singular Value Transformation using Length-Square Sampling Methods

3.1 Stochastic Regression

3.1.1 Definitions and Assumptions

Let $b \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m\times n}$ s.t. $\|A\| \leq 1$ where $\|\cdot\|$ signifies the operator norm (or spectral norm). Furthermore, require that $\text{rank}(A) = k$ and $\|A^+\| \leq \kappa$ where A^+ is the pseudoinverse of A. Hence, observe that $\|A\| \leq 1$ is equivalent to A having maximum singular value 1[1]. Similarly, A^+ has inverted singular values from A and so $\|A^+\|$ is equal to the reciprocal of the minimum nonzero singular value. Therefore, the condition number of A is given by $\|A\|\|A^+\| \leq \kappa$.

So, define x to be the least-squares solution to the linear system $Ax = b$ i.e. $x = A^+b$. Then, in terms of these definitions, we define two primary goals:

1. Query a vector $\tilde{x}$ s.t. $\|\tilde{x} - x\| \leq \epsilon\|x\|$
2. Sample from a distribution that approximates $\frac{|x_j|^2}{\|x\|^2}$ within total variation distance (Theorem 4.7) 2ϵ.

In order to do this, we simply assume that we have length-square sampling access to A. In other words, we are able to sample row indices of A from the distribution $\frac{\|A_{(i,\cdot)}\|^2}{\|A\|_F^2}$

3.1.2 Sequence of Approximations

First, we'll summarize the sequence of approximations that we'll perform using length-squared sampling techniques. We'll describe these steps in depth in the following sections.

Of course, we know that the least squares solution of the linear system is given by the orthogonal projection

$$(A^\dagger A)^+ A^\dagger = A^+ b$$

So, we first approximate $A^\dagger A$ by $R^\dagger R$ where $R \in \mathbb{C}^{r\times n}$, $r \ll m$ is constructed from length-square sampling r rows of A. Now, denote the spectral decomposition

[1] To see this, simply consider Spectral Theorem applied to Hermitian matrix $A^\dagger A$

$$A^\dagger A \approx R^\dagger R = \sum_{l=1}^{k} \frac{1}{\sigma_l^2} \left|v^{(l)}\right\rangle \left\langle v^{(l)}\right|$$

where of course σ_i and $\left|v^{(i)}\right\rangle \in \mathbb{C}^n$ are the singular values and right singular vectors of R, respectively.

We see that computing these right singular vectors of R can still be computationally prohibitive given the dimension n. Hence, we can use length-square sampling again, this time on the columns of R to give a matrix $C \in \mathbb{C}^{r \times c}$, $c \ll n$. Now, the left singular vectors of C which we denote as $\left|w^{(i)}\right\rangle \in \mathbb{C}^r$ can be efficiently computed via standard SVD methods. So,

$$RR^\dagger \approx CC^\dagger = \sum_{l=1}^{k} \frac{1}{\sigma_l^2} \left|w^{(l)}\right\rangle \left\langle w^{(l)}\right|$$

We can then show that ()

$$\left|\tilde{v}^{(i)}\right\rangle := R^\dagger \left|w^{(l)}\right\rangle / \tilde{\sigma}_l \tag{1}$$

provides a good approximation of $\left|v^{(i)}\right\rangle$. Note that $\tilde{\sigma}_l$ are the singular values of C which then approximate the singular values of R which similarly approximate the singular values of A. This follows from $A^\dagger A \approx R^\dagger R$ and $RR^\dagger \approx CC^\dagger$ by the Hoffman–Wielandt inequality detailed in Lemma 2.7 of [**?**] and stated without proof below.

Lemma 3.1. *Hoffman–Wielandt inequality*

If P, Q are two real, symmetric $n \times n$ matrices and $\lambda_1, \cdots \lambda_n$ denote eigenvalues in non-decreasing order, then

$$\sum_{t=1}^{n} (\lambda_t(P) - \lambda_t(Q))^2 \leq \|P - Q\|_F^2$$

At this point, it seems like we haven't made much progress since computing $R^\dagger \left|w^{(l)}\right\rangle$ is still expensive. However, it turns out that all we need to enable query access to $\tilde{x}$ is the ability to efficiently estimate the trace inner product $\mathrm{tr}\left(U^\dagger V\right)$ where U and V are operators such that U can be the length-square sampled and V can be queried. To see this, we write our solution, $\tilde{x}$, in terms of the approximations thus far

$$
\begin{aligned}
\tilde{x} &\approx A^{+} \left|b\right\rangle \\
&\approx (R^{\dagger}R)^{+}A^{\dagger} \left|b\right\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \left\langle\tilde{v}^{(l)}\right| A^{\dagger} \left|b\right\rangle
\end{aligned}
$$

Hence, define $U := A$, $V := \left|b\right\rangle \left\langle\tilde{v}^{(l)}\right|$ in which case

$$
\begin{aligned}
\operatorname{tr}\Big(U^{\dagger}V\Big) &= \operatorname{tr}\Big(A^{\dagger} \left|b\right\rangle \left\langle\tilde{v}^{(l)}\right|\Big) \\
&= \operatorname{tr}\Big(\left\langle\tilde{v}^{(l)}\right| A^{\dagger} \left|b\right\rangle\Big) \\
&= \left\langle\tilde{v}^{(l)}\right| A^{\dagger} \left|b\right\rangle
\end{aligned}
$$

since $\left\langle\tilde{v}^{(l)}\right| A^{\dagger} \left|b\right\rangle$ is a scalar. Therefore, say that

$$
\tilde{\lambda}_l \approx \operatorname{tr}\Big(A^{\dagger} \left|b\right\rangle \left\langle\tilde{v}^{(l)}\right|\Big)
$$

and assume that we can compute and memoize these scalars $\tilde{\lambda}_i$ efficiently. In which case,

$$
\tilde{x} \approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left|\tilde{v}^{(l)}\right\rangle \tilde{\lambda}_l
$$

Recalling the definition of $\left|\tilde{v}^{(i)}\right\rangle$ (1),

$$
\begin{aligned}
&= \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} R^{\dagger} \left|w^{(l)}\right\rangle \tilde{\lambda}_l \\
&= R^{\dagger} \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left|w^{(l)}\right\rangle \tilde{\lambda}_l
\end{aligned}
$$

and so defining $z := \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left|w^{(l)}\right\rangle \tilde{\lambda}_l$,

$$
= R^{\dagger} z
$$

We see that we can compute z efficiently (and memoize it for future queries) because it is a k-linear combination of left singular vectors in $\mathbb{C}^r$. So, say that we wish to query an element $\tilde{x}_j$. We can simply query column $R_{\cdot,j} \in \mathbb{C}^r$ (or equivalently row $R^\dagger_{j,\cdot}$) and compute $R_{\cdot,j} \cdot z$. Hence, we've achieved our first goal.

In order to achieve our second goal, enabling sample access to a distribution that approximates $\frac{|x_j|^2}{\|x\|^2}$, we require one more trick: rejection sampling which we detail in Section ().

All in all, we've performed the chain of approximations,

$$\begin{aligned}
|x\rangle &= A^+ |b\rangle = (A^\dagger A)^+ A^\dagger |b\rangle \\
&\approx (R^\dagger R)^+ A^\dagger |b\rangle = \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left| v^{(l)} \right\rangle \left\langle v^{(l)} \right| A^\dagger |b\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left| \tilde{v}^{(l)} \right\rangle \left\langle \tilde{v}^{(l)} \right| A^\dagger |b\rangle \\
&\approx \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^2} \left| \tilde{v}^{(l)} \right\rangle \tilde{\lambda}_l = R^\dagger \sum_{l=1}^{k} \frac{1}{\tilde{\sigma}_l^3} \left| w^{(l)} \right\rangle \tilde{\lambda}_l = R^\dagger z
\end{aligned}$$

Now that we've sketched the steps of this process, we detail each approximation and show that we can achieve the claimed correctness and complexity bounds.

3.1.3 Computing Approximate Singular Vectors

As described above, we begin by length-square sampling the original matrix $A \in \mathbb{C}^{m \times n}$. So, pick row index i with probability

$$p_i = \frac{\|A_{(i,\cdot)}\|^2}{\|A\|_F^2}$$

and output random row $Y = A_{(i,\cdot)}/\sqrt{p_i} = \frac{A_{(i,\cdot)}}{\|A_{(i,\cdot)}\|} \|A\|_F$. After sampling r rows, we implicitly define matrix R to be the concatenation of the outputted random rows. Furthermore, we rescale R so that $E[R^\dagger R] = A^\dagger A$, which requires normalizing by $\sqrt{r}$. Therefore,

$$R = \frac{1}{\sqrt{r}} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_r \end{bmatrix} \in \mathbb{C}^{r \times n}$$

4 Appendix

Definition 4.1. *Pauli Matrices*

$$\sigma_x = X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$\sigma_y = Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$
$$\sigma_z = Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Definition 4.2. *Bell States*

$$\frac{|00\rangle + |11\rangle}{\sqrt{2}}$$
$$\frac{|00\rangle - |11\rangle}{\sqrt{2}}$$
$$\frac{|10\rangle + |01\rangle}{\sqrt{2}}$$
$$\frac{|01\rangle - |10\rangle}{\sqrt{2}}$$

Definition 4.3. *Positive Operators*

Let A be a bounded[2] linear operator on complex Hilbert space $\mathcal{H}$. The following conditions are equivalent to A being positive

1. *$A = S^\dagger S$ for some bounded operator S on $\mathcal{H}$*
2. *A is hermitian and $\langle x| A |x\rangle \geq 0$ for every $|x\rangle \in \mathcal{H}$*
3. *the spectrum of A is non-negative*

Definition 4.4. *Trace of an Operator*

Let $\{|i\rangle\}$ be an orthonormal basis for A and so

$$\begin{aligned} \text{tr}(A) &= \sum_i A_{ii} \\ &= \sum_i \langle i| A |i\rangle \end{aligned}$$

[2] $\|Av\| \leq M\|v\|$ for some $M > 0$ and all $v \in \mathcal{H}$

Hence, if we extend $|\psi\rangle$ to the orthonormal basis $\{|i\rangle\}$ which includes $|\psi\rangle$ as the first element (for example via the Gram-Schmidt procedure) then

$$\begin{aligned}\operatorname{tr}(A|\psi\rangle\langle\psi|) &= \sum_i \langle i| A |\psi\rangle \langle\psi|i\rangle \\ &= \langle\psi| A |\psi\rangle\end{aligned}$$

by orthonormality.

Theorem 4.5. *Spectral Theorem Suppose A is a compact*[3] *hermitian operator (compactness ensures A has eigenvectors) on complex Hilbert space $\mathcal{H}$. Hence, there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of A. Each eigenvalue is in $\mathbb{R}$.*

Lemma 4.6. *Let A, B, C by symmetric $d \times d$ matrices satisfying $A \succeq 0$ and $B \preceq C$. Hence, $\operatorname{Tr}(AB) \leq \operatorname{Tr}(AC)$*

Proof. Write A in its spectral decomposition $A = \sum \lambda_i |i\rangle\langle i|$, invoking Spectral Theorem (4.5). Hence,

$$\begin{aligned}\operatorname{Tr}(AB) &= \operatorname{Tr}\Big(\sum \lambda_i |i\rangle\langle i| B\Big) \\ &= \sum \lambda_i \operatorname{Tr}(|i\rangle\langle i| B) && \text{(linearity of trace)} \\ &= \sum \lambda_i \operatorname{Tr}(\langle i| B |i\rangle) && \text{(cyclic property of trace)} \\ &\leq \sum \lambda_i \operatorname{Tr}(\langle i| C |i\rangle) \\ &= \sum \lambda_i \operatorname{Tr}(|i\rangle\langle i| C) = \operatorname{Tr}\Big(\sum \lambda_i |i\rangle\langle i| C\Big) = \operatorname{Tr}(AC)\end{aligned}$$

□

Corollary 4.6.1. *If $A, B \succeq 0$, then $\operatorname{Tr}(AB) \leq \|B\|_2 \operatorname{Tr}(A)$*

Proof. Note that the singular values of B coincide with the eigenvalues of B since $B^\dagger B = B^2$ and $B \succeq 0 \Rightarrow \lambda_i(B) \geq 0$, $\forall i$. So, let $C = \|B\|_2 I$ which then trivially satisfies $\lambda_i(C) = \lambda_{\max}(B)$, $\forall i$ since C is the diagonal matrix with diagonal values all equal to $\lambda_{\max}(B)$. Therefore, $B \preceq C$. So, we can simply apply 4.6 above,

$$\begin{aligned}\operatorname{Tr}(AB) &\leq \operatorname{Tr}(AC) \\ &= \operatorname{Tr}\left(A\|B\|_2 I\right) \\ &= \|B\|_2 \operatorname{Tr}(A)\end{aligned}$$

□

[3] the image under A acting on any bounded subset of $\mathcal{H}$ is a compact subset of $\mathcal{H}$

Definition 4.7. *Total Variation Distance.*

Let P and Q be distinct probability measures on a σ-algebra $\mathcal{F}$ of subsets of the sample space Ω. Then, the total variation distance is given by

$$\delta(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

Lemma 4.8. *Hoeffding–Chernoff Inequality*

Let $X_1, X_2, \cdots, X_s$ be i.i.d real random variables. For any positive, real numbers a, t we have that, from Markov's inequality,

$$\begin{aligned} \Pr\left(\sum_{i=1}^{s} X_i \geq a\right) &\leq e^{-ta} E\left[\prod_{i=1}^{s} e^{tX_i}\right] \\ &= e^{-ta} \prod_{i=1}^{s} E\left[e^{tX_i}\right] \end{aligned}$$

by independence. □

Theorem 4.9. *Hoeffding–Chernoff Inequality for matrix-valued random variables [**?**]*

Let X be a random variable taking values which are real symmetric $d \times d$ matrices. Suppose $X_1, X_2, \cdots, X_s$ are i.i.d. draws of X. For any positive real numbers a, t, we have

$$\Pr\left(\lambda_{\max}\Big(\sum_{i=1}^{s} X_i\Big) \geq a\right) \leq de^{-ta} \|E[e^{tX}]\|_2^s \tag{2}$$

$$\Pr\left(\Big\|\Big(\sum_{i=1}^{s} X_i\Big)\Big\|_2 \geq a\right) \leq de^{-ta}(\|E[e^{tX}]\|_2^s + \|E[e^{-tX}]\|_2^s) \tag{3}$$

where $\lambda_{\max}$ is the largest eigenvalue.

Proof. First, we can show that (2) ⇒ (3). By definition of the 2-norm of a matrix,

$$\|\sum_i X_i\|_2 = \max\left(\lambda_{\max}\Big(\sum_i X_i\Big), \lambda_{\max}\Big(\sum_i (-X_i)\Big)\right)$$

since it is the square root of the maximum eigenvalue of $(\sum_i X_i^T) \sum_i X_i = (\sum_i X_i) \sum_i X_i$ and hence, equivalently, the maximum absolute value of an eigenvalue of X_i. Therefore, we can simply apply (2) to both X_i and $-X_i$ and we get (3).

So, we can focus our attention on (3). Let $S = \sum_i^s X_i$. Hence,

$$\lambda_{\max}(S) \geq a \Leftrightarrow \lambda_{\max}(tS) \geq ta$$

Furthermore, by considering the power series definition of the exponential,

$$\begin{aligned} \Leftrightarrow \lambda_{\max}(e^{tS}) &\geq e^{ta} \\ \Rightarrow \mathrm{Tr}\big(e^{tS}\big) &\geq e^{ta} \end{aligned}$$

since the trace is the sum of the matrix's eigenvalues. Since $\mathrm{Tr}\big(e^{tS}\big) \geq 0$, we can apply Markov's inequality

$$\Pr\big(\mathrm{Tr}\big(e^{tS}\big) \geq e^{ta}\big) \leq \frac{E[\mathrm{Tr}\big(e^{tS}\big)]}{e^{ta}}$$

Now, we use the following lemma

Lemma 4.10. *Golden-Thompson Inequality*
If A and B are Hermitian matrices, then

$$\mathrm{Tr}\big(e^{A+B}\big) \leq \mathrm{Tr}\big(e^{A}e^{B}\big)$$

□

Hence, we can let $A = t(\sum_i^{s-1} X_i)$ and $B = tX_s$. Then,

$$E_X\Big[\,\mathrm{Tr}\big(e^{tS}\big)\Big] \leq E_X\Big[\,\mathrm{Tr}\Big(e^{t\big(\sum_i^{s-1} X_i\big)}e^{tX_s}\Big)\Big]$$

Since the expectation operator commutes with the summation of the trace by linearity of trace,

$$\begin{aligned} &= \mathrm{Tr}\left(E_X\left[e^{t\big(\sum_i^{s-1} X_i\big)}e^{tX_s}\right]\right) \\ &= \mathrm{Tr}\left(E_{X_1,X_2,\cdots,X_{s-1}}\left[e^{t\big(\sum_i^{s-1} X_i\big)}\right]E_{X_s}\left[e^{tX_s}\right]\right) \qquad \text{(by independence)} \end{aligned}$$

Now, we can apply Corollary (4.6.1), which gives

$$\begin{aligned} &\leq \mathrm{Tr}\left(E_{X_1,X_2,\cdots,X_{s-1}}\left[e^{t\big(\sum_i^{s-1} X_i\big)}\right]\right)\left\|E_{X_s}\left[e^{tX_s}\right]\right\|_2 \\ &= \mathrm{Tr}\left(E_X\left[e^{t\big(\sum_i^{s-1} X_i\big)}\right]\right)\left\|E_X\left[e^{tX}\right]\right\|_2 \\ &= E_X\left[\,\mathrm{Tr}\left(e^{t\big(\sum_i^{s-1} X_i\big)}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2 \end{aligned}$$

So we can repeat this process iteratively, peeling an X_i each time from the left term. For clarity, the next step gives,

$$\begin{aligned} E_X\left[\operatorname{Tr}\left(e^{t\left(\sum_i^{s-1} X_i\right)}\right)\right] &\leq E_X\left[\operatorname{Tr}\left(e^{t\left(\sum_i^{s-2} X_i\right)} e^{tX_{s-1}}\right)\right] \\ &\leq E_X\left[\operatorname{Tr}\left(e^{t\left(\sum_i^{s-2} X_i\right)}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2 \qquad \text{(applying (4.6.1) again)} \end{aligned}$$

Therefore, after peeling all terms but the last X_i, we have

$$E_X\left[\operatorname{Tr}\left(e^{tS}\right)\right] \leq E_X\left[\operatorname{Tr}\left(e^{tX}\right)\right]\left\|E_X\left[e^{tX}\right]\right\|_2^{s-1}$$

Hence, since the trace is the sum of eigenvalues, $\operatorname{Tr}\left(e^{tX}\right) \leq d\lambda_{\max}(e^{tX})$ i.e. the worst case of all d eigenvalues being the max

$$\leq d\left\|E_X\left[e^{tX}\right]\right\|_2^{s}$$

as desired. □