Randomized Nyström Preconditioning with RPChloesky

Aaron Dinesh

December 20, 2024

A project submitted in fulfilment
of the requirements for MATH-403

Declaration

I hereby declare that this work is fully my own.

Signed: Aaron Dinesh Date: December 20, 2024

1.1 Question 1

Using the result from L6S76, show that the approximation $\hat{A}^{(k)}$ returned after k steps of RPChloesky satisfies

$$\mathbb{E}\left[\|A - \hat{A}^{(k)}\|_2\right] \leq 3 \cdot \mathrm{sr}_p(A) \cdot \lambda_p$$

for $k \geq (p-1)(\frac{1}{2} + \log(\frac{\eta^{-1}}{2}))$ with $\mathrm{sr}_p(A)$ defined in [1].

Using the definition of $\mathrm{sr}_p(A)$ from [1] we can see that:

$$\begin{aligned}
\mathrm{sr}_p(A) \cdot \lambda_p &= \left[\lambda_p^{-1} \sum_{j>p}^{n} \lambda_j\right] \lambda_p \\
&= \sum_{j \geq p}^{n} \lambda_j \\
&= \mathrm{trace}(A - \mathcal{T}_{p-1}(A))
\end{aligned}$$

Where $\mathcal{T}_r(A)$ denotes the best rank-r approximation of A. We then also note that:

$$\mathbb{E}\left[\|A - \hat{A}^{(k)}\|\right] \leq \mathbb{E}\left[\mathrm{trace}(A - \hat{A}^{(k)})\right]$$

Then we can use the theorem from L6S76 to begin our proof:

$$\begin{aligned}
\mathbb{E}\left[\|A - \hat{A}^{(k)}\|\right] &\leq \mathbb{E}\left[\mathrm{trace}(A - \hat{A}^{(k)})\|\right] \\
&\leq (1+\epsilon)\mathrm{trace}(A - \mathcal{T}_{p-1}(A))
\end{aligned}$$

To complete the proof we let $\epsilon = 2$, and then the equation above becomes:

$$\mathbb{E}\left[\|A - \hat{A}^{(k)}\|\right] \leq 3 \cdot \mathrm{trace}(A - \mathcal{T}_{p-1}(A))$$

According to the theorem in L6S76, for the above bound to hold we need to choose:

$$\begin{aligned}
k &\geq \frac{r}{\epsilon} + r\log(\frac{1}{\epsilon\eta}) \\
&= \frac{p-1}{2} + (p-1)\log(\frac{\eta^{-1}}{2}) \\
&= (p-1)(\frac{1}{2} + \log(\frac{\eta^{-1}}{2}))
\end{aligned}$$

where $\eta = \mathrm{trace}(A - \mathcal{T}_{p-1}(A))/\mathrm{trace}(A)$. Thus completing the proof.

1.2 Question 2

By mimicking the proof of Theorem 5.1 in [1], derive a sensible upper bound on:

$$\mathbb{E}\left[\kappa_2(P^{-\frac{1}{2}} A_\mu P^{-\frac{1}{2}})\right]$$

where P is constructed as described in equations (1.3) of [1], with $\hat{A}_{nys}$ replaced by $\hat{A}^{(k)}$ for a suitable value for k. Explain what this bound means in terms of the quality of the preconditioner.

First we start by notcing that:

$$A_\mu = A + \mu I = \hat{A}^{(k)} + \mu I + \underbrace{A - \hat{A}^{(k)}}_{E}$$

This then allows us to rewite $S = P^{-\frac{1}{2}} A_\mu P^{-\frac{1}{2}}$ as:

$$\begin{aligned} P^{-\frac{1}{2}} A_\mu P^{-\frac{1}{2}} &= P^{-\frac{1}{1}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}} + P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}} \\ \text{with } P &= \frac{1}{\lambda_{p-1} + \mu} U(\Lambda + \mu I)U^\top + (I - UU^\top) \end{aligned}$$

Since P is PSD it has a well defined sqaure root. Using the formula given by Frangella for P^{-1} we can get:

$$P^{-\frac{1}{2}} = U(\frac{\Lambda + \mu I}{\lambda_{p-1}} + \mu)^{-\frac{1}{2}} U^\top + (I - UU^\top)$$

By Weyl's inequality we can bound the largest eigenvalue of S as:

$$\lambda_1(S) \leq \lambda_1(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) + \lambda_1(P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}})$$

Let's consider the first half of the equation. We can substitute our expression for $P^{-\frac{1}{2}}$ and compute the eigenvalue decomposition of $\hat{A}^{(k)}$ to get:

$$\begin{aligned} &\left[U \left(\frac{\Lambda + \mu I}{\lambda_{p-1} + \mu} \right)^{-\frac{1}{2}} U^\top + (I - UU^\top) \right] \cdot \\ &\left[U(\Lambda + \mu I)U^\top + (\mu I)(I - UU^\top) \right] \cdot \\ &\left[U \left(\frac{\Lambda + \mu I}{\lambda_{p-1} + \mu} \right)^{-\frac{1}{2}} U^\top + (I - UU^\top) \right] \end{aligned}$$

We have to split $\hat{A}^{(k)} + \mu I$ into the component that lies in the subspace spanned by U and into the space spanned by $U_\perp$, this is done by projecting μI using the projector $I - UU^\top$. Now if we only look at the subspace spanned by U we can see that the largest eigenvalue will be $\lambda_{p-1} + \mu$:

$$U(\frac{\lambda_{p-1} + \mu}{\Lambda + \mu I})^{\frac{1}{2}} [\Lambda + \mu I] (\frac{\lambda_{p-1} + \mu}{\Lambda + \mu I})^{\frac{1}{2}} U^\top$$

The largest eigenvalue in the space spanned by $U_\perp$ is μ. So overall the largest eigenvalue achieved on both these space is $\lambda_{p-1} + \mu$. Now we have to find an expression for the second part of Weyl's inequality above.

$$\begin{aligned} &\lambda_1(P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}}) = \lambda_1(P^{-1}E) \leq \lambda_1(P^{-1}) \|E\|_2 \\ &\qquad\qquad \text{with } P^{-1} = (\lambda_{p-1} + \mu)U(\Lambda + \mu I)^{-1} U^\top + (I - UU^\top) \end{aligned}$$

To find $\lambda_1(P^{-1})$ we perform a smilar arguement as before. On the subspace spanned by U the largest eigenvalue is $(\lambda_{p-1}+\mu)$ $(\lambda_{p-1}+\mu)=1$ on the subspace spanned by $U_\perp$ we see that the largest eigenvalue is also 1. So the largest eigenvalue attained on both these subspaces is 1, hence $\lambda_1(P^{-1})=1$ and so we can say that $\lambda_1(P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}})=\|E\|_2$. So we arrive at the bound:

$$\lambda_1(S)=\lambda_{p-1}+\mu+\|E\|_2$$

Now we need to bound the minimum eigenvalues of S. Once again we can use Weyl's inequality for this:

$$\lambda_n(S)\geq\lambda_n(P^{-\frac{1}{2}}(\hat{A}^{(K)}+\mu I)P^{-\frac{1}{2}})+\lambda_n(P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}})$$

The smallest eigenvalue of $P^{-\frac{1}{2}}(E)P^{-\frac{1}{2}}$ is 0 since the rank-$(p-1)$ approximation we have is rank deficient and so the smallest eigenvalue is 0. Once again analyse the first term by looking on the subspace spanned by U and the subspace spanned by $U_\perp$

$$U(\frac{\lambda_{p-1}+\mu}{\Lambda+\mu I})^{\frac{1}{2}}[\Lambda+\mu I](\frac{\lambda_{p-1}+\mu}{\Lambda+\mu I})^{\frac{1}{2}}U^\top$$

Since the approximations of RPChloesky are PSD, the minimum eigenvalue on the subspace spanned by U is μ. In the subpace spanned by $U_\perp$ we have $\mu I(I-UU^\top)$ of which the minimum eigenvalue is μ. Hence the smallest eigenvalue overall is just μ and so $\lambda_n(S)\geq\mu$. Now we can finally bound the condition number as:

$$\begin{aligned}\kappa_2(P^{-\frac{1}{2}}A_\mu P^{-\frac{1}{2}})&\leq\frac{\mu+\lambda_{p-1}+\|A-\hat{A}^{(k)}\|_2}{\mu}\\&=\frac{1}{\mu}\left[\mu+\lambda_{p-1}+\|A-\hat{A}^{(k)}\|_2\right]\end{aligned}$$

Then by using the linearity of the expectation operator we can say:

$$\begin{aligned}\mathbb{E}\left[\kappa_2(P^{-\frac{1}{2}}A_\mu P^{-\frac{1}{2}})\right]&\leq\mathbb{E}\left(\frac{\mu+\lambda_{p-1}+\|A-\hat{A}^{(k)}\|_2}{\mu}\right)\\&=\frac{1}{\mu}\left(\mu+\lambda_{p-1}+\mathbb{E}\left[\|A-\hat{A}^{(k)}\|_2\right]\right)\\&\leq\frac{1}{\mu}\left[\mu+\lambda_{p-1}+3\cdot\mathrm{trace}(A-\mathcal{T}_{p-1}(A))\right]\end{aligned}$$

for $k\geq(p-1)(\frac{1}{2}+\log(\frac{\eta^{-1}}{2}))$, where $\eta=\mathrm{trace}(A-\mathcal{T}_{p-1}(A))/\mathrm{trace}(A)$.
While this bound is good, we can get a simpler bound if we make some assumptions on $p-1$. We can leverage Lemma 2.1 Item 4 and Lemma 5.4 Item 1 in [1]. Since these items are agnostic of the approximation algorithm used, we can use them without proof. If we choose $\gamma>0$ and if we choose $p-1\geq(1+\gamma^{-1})\,\mathrm{deff}(\mu)$ then $\lambda_{p-1}\leq\gamma\mu$. So if we choose $\gamma=1$ then we get:

$$\frac{1}{\mu}\left[\mu+\lambda_{p-1}+3\cdot\mathrm{trace}(A-\mathcal{T}_{p-1}(A))\right]=2+\frac{3}{\mu}\cdot\mathrm{trace}(A-\mathcal{T}_{p-1}(A))$$

Without preconditioning the condition number of A_μ would be $\lambda_1(A_\mu)/\lambda_n(A_\mu)$ which can be large if the smallest eigenvalue of A_μ is small. However if we use RPChloesky, obtain a sufficently good approximation of A and use this preconditioning then the condition number will be $\leq 1 + \frac{3}{\mu} \cdot \text{trace}(A - \mathcal{T}_{p-1}(A))$ which could even be upperbounded by 1 since we already obtain a good approximation of A by using RPChloesky. However this does assume a quick decay of the singular values, which is the case for most scietific applications. Diaz et al. analyse a specific case of using RPChloesky and preconditioning in quantum chemistry in section (4.1) of their paper []. They show how RPChloesky and preconditioning can reduce the relative residual in a ridge regression task to an order of 10^{-2} in 100 iterations, highlighting how good the method is. Also iterative solvers like the Conjugate Gradient method is proportional to the sqaure root of the condition number so achieveing a low condition number through preconditioning would be ideal for these solvers.

1.3 Question 3

The proof of Proposition 2.2 from [1] on the quality of the Nyström approximation (with Gaussian random sketches) uses a squared Chevet bound. Provide a detailed proof of this bound (see Section B.2 in [1]) in your own words. Include all missing details (such as verifying the conditions of Slepian's lemma).

We first begin by defining two vector sets:

$$U = \{S^\top a : \|a\|_2 = 1\} \subset \mathbb{R}^m$$
$$V = \{Tb : \|b\|_2 = 1\} \subset \mathbb{R}^n$$

Where $S \in \mathbb{R}^{r \times m}$ and $T \in \mathbb{R}^{n \times s}$ are fixed matices and $a \in \mathbb{R}^r$ and $b \in \mathbb{R}^s$ are vectors living on their respective ℓ_2-normball. Now from these sets we choose two vectors $u \in U$ and $v \in V$ and then we consider the Gaussian process:

$$Y_{uv} = \langle u, Gv \rangle + \|S\|_2 \|v\|_2 \gamma$$
$$X_{uv} = \|S\| \langle h, v \rangle + \|v\| \langle g, u \rangle$$

Where $G \in \mathbb{R}^{m \times n}$ is a $(0,1)$-Gaussian random matrix, g, h are $\mathbb{R}^m$ and $\mathbb{R}^n$ $(0,1)$-Gaussian random vectors and $\gamma \sim \mathcal{N}(0,1)$. We also assume that $G, g, h,$ and γ are all independant.

Our first step is to analze the conditions regarding Slepian's Lemma. The two ocnditons of Slepian's Lemma are:

$$\mathbb{E}\left[X_{u_1,v_1} X_{u_2,v_2}\right] \leq \mathbb{E}\left[Y_{u_1,v_1} Y_{u_2,v_2}\right] \text{ for } u_1 \neq u_2 \text{ and } v_1 \neq v_2$$
$$\mathbb{E}\left[X_{u,v} X_{u,v}\right] \leq \mathbb{E}\left[Y_{u,v} Y_{u,v}\right]$$

Let's first analyze the autocorrelation terms. For convenience sake we will abreviate $X_{u,v} = X_i$ similarly for $Y_{u,v}$

$$\begin{aligned} X_i^2 &= [\|S\| \langle h, v \rangle + \|v\| \langle g, u \rangle]^2 \\ &= \|S\|^2 \langle h, v \rangle^2 + 2\|S\| \langle h, v \rangle \|v\| \langle g, u \rangle + \|v\|^2 \langle g, u \rangle^2 \\ \mathbb{E}\left[X_i^2\right] &= \|S\|^2 \mathbb{E}\left[\langle h, v \rangle^2\right] + 2\|S\| \|v\| \mathbb{E}\left[\langle h, v \rangle \langle g, u \rangle\right] + \|v\|^2 \mathbb{E}\left[\langle g, u \rangle^2\right] \end{aligned}$$

We will analyse this equation term by term and use the fact that $E[X] = \mathrm{Var}[X] + \mathbb{E}[X]^2$. We will also use the fact that if $x \sim \mathcal{N}(\mu, \Sigma_n)$ then for any fixed vector a we have that $ax \sim \mathcal{N}(a\mu, a\Sigma_n a^\top)$. Analysing the first term we see:

$$\mathbb{E}[\langle h, v\rangle] = \|v\|^2 + 0^2$$

The third term can be analyzed in much the same way:

$$\mathbb{E}[\langle g, u\rangle] = \|u\|^2 + 0^2$$

The second term can by analysed by first noting that h and g are independant as so:

$$\begin{aligned}\mathbb{E}[2\|S\|\|v\|\langle h, v\rangle\langle g, u\rangle] &= 2\|S\|\|v\|\mathbb{E}[\langle h, v\rangle]\,\mathbb{E}[\langle g, u\rangle] \\ &= 0\end{aligned}$$

And so we are left with:

$$\mathbb{E}[X_i] = \|S\|^2\|v\|^2 + \|v\|^2\|u\|^2$$

Y_i will be analysed in the same way:

$$\begin{aligned}Y_i^2 &= \langle u, Gv\rangle^2 + 2\|S\|\|v\|\gamma\langle u, Gv\rangle^2 + \|S\|^2\|v\|^2\gamma^2 \\ \mathbb{E}\left[Y_i^2\right] &= \mathbb{E}\left[\langle u, Gv\rangle^2\right] + 2\|S\|\|v\|\mathbb{E}[\gamma\langle u, Gv\rangle] + \|S\|^2\|v\|^2\mathbb{E}\left[\gamma^2\right]\end{aligned}$$

The first and third terms are analysed in much the same way as before:

$$\mathbb{E}\left[\|S\|^2\|v\|^2\gamma^2\right] = \|S\|^2\|v\|^2\mathbb{E}\left[\langle u, Gv\rangle^2\right] = \|u\|^2\|v\|^2$$

The second term can be analyzed in the same way as before:

$$\begin{aligned}\mathbb{E}[2\|S\|\|v\|\gamma\langle u, Gv\rangle] &= 2\|S\|\|v\|\mathbb{E}[\gamma\langle u, Gv\rangle] \\ &= 2\|S\|\|v\|\mathbb{E}[\gamma]\,\mathbb{E}[\langle u, Gv\rangle] \\ &= 0\end{aligned}$$

And so we are left with:

$$\mathbb{E}\left[Y_i^2\right] = \|S\|^2\|v\|^2 + \|u\|^2\|v\|^2$$

Comparing $\mathbb{E}[X_i^2]$ with $\mathbb{E}[Y_i^2]$ we see that the second condition of Slepian's lemma is satisfied. Now we will look at the first condition. We begin by letting:

$$\begin{aligned}X_1 &= \|S\|\langle h, v_1\rangle + \|v_1\|\langle g, u_1\rangle \\ X_2 &= \|S\|\langle h, v_2\rangle + \|v_2\|\langle g, u_2\rangle\end{aligned}$$

We can multiply and expand out the terms to get:

$$\begin{aligned}X_1X_2 &= \|S\|^2\langle h, v_1\rangle\langle h, v_2\rangle + \|S\|\|v_2\|\langle h, v_1\rangle\langle g, u_2\rangle \\ &= +\|S\|\|v_1\|\langle h, v_2\rangle\langle g, u_1\rangle + \|v_1\|\|v_2\|\langle g, u_1\rangle\langle g, u_2\rangle \\ \mathbb{E}[X_1X_2] &= \|S\|^2\mathbb{E}[\langle h, v_1\rangle\langle h, v_2\rangle] + \|S\|\|v_2\|\mathbb{E}[\langle h, v_1\rangle\langle g, u_2\rangle] \\ &= +\|S\|\|v_1\|\mathbb{E}[\langle h, v_2\rangle\langle g, u_1\rangle] + \|v_1\|\|v_2\|\mathbb{E}[\langle g, u_1\rangle\langle g, u_2\rangle]\end{aligned}$$

Once again we can analyse this term by term. Looking at the first term we have:

$$\mathbb{E}\left[\|S\|^2\langle h, v_1\rangle\langle h, v_2\rangle\right] = \|S\|^2\mathbb{E}\left[\langle h, v_1\rangle\langle h, v_2\rangle\right] \quad = \|S\|^2\mathbb{E}\left[\underbrace{\langle h, h\rangle}_{\mathbb{E}[\cdot]=0} + \underbrace{\langle h, v_1\rangle}_{\mathbb{E}[\cdot]=0} + \underbrace{\langle h, v_2\rangle}_{\mathbb{E}[\cdot]=0} + \underbrace{\langle v_1, v_2\rangle}_{\mathbb{E}[\cdot]\neq 0}\right]$$

By the properties of multiplication with standard gaussian random vectors the above equation simplifies to:

$$\mathbb{E}\left[\|S\|^2\langle h, v_1\rangle\langle h, v_2\rangle\right] = \|S\|^2\langle v_1, v_2\rangle$$

Now we look at the second and third term, and by independence and the properties of multiplication with standard gaussian random vectors we have:

$$\begin{aligned}\mathbb{E}\left[\|S\|\|v_2\|\langle g, u_2\rangle\langle h, u_1\rangle\right] = 0\\ \mathbb{E}\left[\|S\|\|v_1\|\langle g, u_1\rangle\langle h, u_2\rangle\right] = 0\end{aligned}$$

Regarding the fourth term, once it is expanded all the terms multiplied with the gaussian vector in expectation will be zero so the only term left will be:

$$\begin{aligned}\mathbb{E}\left[\|v_1\|\|v_2\|\langle g, u_1\rangle\langle g, u_2\rangle\right] &= \|v_1\|\|v_2\|\mathbb{E}\left[\langle u_1, u_2\rangle\right]\\ &= \|v_1\|\|v_2\|\langle u_1, u_2\rangle\end{aligned}$$

So finally we are left with:

$$\mathbb{E}\left[X_1X_2\right] = \|S\|^2\langle v_1, v_2\rangle + \|v_1\|\|v_2\|\langle u_1, u_2\rangle$$

Now we will perform the same operations for Y_1Y_2. Letting:

$$\begin{aligned}Y_1 &= \langle u_1, Gv_1\rangle + \|S\|\|v_1\|\gamma\\ Y_2 &= \langle u_2, Gv_2\rangle + \|S\|\|v_2\|\gamma\end{aligned}$$

We can multiply and expand out the terms to get:

$$\begin{aligned}Y_1Y_2 =&\langle u_1, Gv_1\rangle\langle u_2, Gv_2\rangle + \|S\|\|v_1\|\gamma\langle u_2, Gv_2\rangle\\ &+ \|S\|\|v_2\|\gamma\langle u_1, Gv_1\rangle + \|S\|^2\|v_1\|\|v_2\|\gamma^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[Y_1Y_2\right] =&\mathbb{E}\left[\langle u_1, Gv_1\rangle\langle u_2, Gv_2\rangle\right] + \|S\|\|v_1\|\mathbb{E}\left[\gamma\langle u_2, Gv_2\rangle\right]\\ &+ \|S\|\|v_2\|\mathbb{E}\left[\gamma\langle u_1, Gv_1\rangle\right] + \|S\|^2\|v_1\|\|v_2\|\mathbb{E}\left[\gamma^2\right]\end{aligned}$$

As before, due to the properties of independence and multiplication with standard gaussian random vectors, the second and third terms in expectaion equal 0:

$$\|S\|\|v_1\|\mathbb{E}\left[\gamma\langle u_2, Gv_2\rangle\right] = 0\|S\|\|v_2\|\mathbb{E}\left[\gamma\langle u_1, Gv_1\rangle\right] = 0$$

The fourth term in expectation is:

$$\begin{aligned}\|S\|^2\|v_1\|\|v_2\|\mathbb{E}\left[\gamma^2\right] &= \|S\|^2\|v_1\|\|v_2\|\left[\mathrm{Var}\left(\gamma\right) + \mathbb{E}\left[\gamma\right]\right]\\ &= \|S\|^2\|v_1\|\|v_2\|\end{aligned}$$

The first term needs some careful consideration. We fist see that the dot product can be written as:

$$\langle u_1, Gv_1 \rangle = \sum_{i,j} u_i G_{ij} v_j$$

Next we begin by noting that $\mathbb{E}\left[G_{ij}G_{kl}\right] = 0 \ \forall \ i \neq k$ and $j \neq l$. So the first term, in expectation reduces to:

$$\mathbb{E}\left[\langle u_1, Gv_1 \rangle \langle u_2, Gv_2 \rangle\right] = \langle u_1, u_2 \rangle \langle v_1, v_2 \rangle$$

Putting this all together we get:

$$\mathbb{E}\left[Y_1 Y_2\right] = \langle u_1, u_2 \rangle \langle v_1, v_2 \rangle + \|S\|^2 \|v_1\| \|v_2\|$$

We can now use a double application of the Cauchy-Schwarz inequality (one on $\mathbb{E}\left[X_1 X_2\right]$ and another on $\mathbb{E}\left[Y_1 Y_2\right]$) to see that $\mathbb{E}\left[X_1 X_2\right] \leq \mathbb{E}\left[Y_1 Y_2\right]$. Thereby fulfilling the second condition for Slepian's lemma. Thereby can also conclude that:

$$\mathbb{P}\left(\max_{u,v} Y_{uv} > t\right) \leq \mathbb{P}\left(\max_{u,v} X_{uv} > t\right)$$

For convenience we will also introduce the notation $X_+ = \max\{X, 0\}$. We will now begin to prove the sqaured Chevet bound. We start off by stating:

$$\mathbb{E}\left[\max_{u,v} (Y_{uv})_+^2\right] = \mathbb{E}\left[\max_{\|a\|=1,\ \|b\|=1} \left(\left[\langle S^\top a, GTb \rangle + \|S\| \|Tb\| \gamma\right]_+^2\right)\right]$$

Next the paper applies Jensen's inequality. However we should first make sure that Jensen's inequality is valid for this function. Let's first breakdown the function inside the expectation as a composition of functions:

$$\mathbb{E}\left[\underbrace{\max_{\|a\|=1, \|b\|=1} \left\{\left(\langle S^\top a, GTb \rangle + \|S\| \|Tb\| \gamma\right)_+^2\right\}}_{g(x)}\right]$$

It is evident that $g(x)$ is composed of $f(x) = x_+^2$ and $k(x) = \langle S^\top a, GTb \rangle + \|S\| \|Tb\| \gamma$ The former is a convex function and the latter is a linear function (which is both convex and concave). Their composition is also a convex function, and taking the max leaves it as a convex function. Thus we can apply Jensen's inequality to in this context. In this next step we integrate out the terms with γ and we are left with:

$$\mathbb{E}\left[\max_{\|a\|=1, \|b\|=1} \left\{\left(\langle S^\top a, GTb \rangle + \|S\| \|Tb\| \gamma\right)_+^2\right\}\right] = \mathbb{E}_G\left[\max_{\|a\|=1, \|b\|=1} (\langle S^\top a, GTb \rangle^2)_+\right]$$

Now we can see that the dot product is just the 2->2 matrix operator norm. Thus we can write:

$$\mathbb{E}_G\left[\max_{\|a\|=1, \|b\|=1} (\langle S^\top a, GTb \rangle^2)_+\right] = \mathbb{E}_G\left[\|SGT\|^2\right]$$

So we can see that $\mathbb{E}\left[\max_{u,v}(Y_{uv})_+^2\right]$ acts as a majorizer of $\mathbb{E}_G\left[\|SGT\|^2\right]$. Now we will perform the same calculation with X_{uv}. We can see that:

$$\begin{aligned}\mathbb{E}\left[\max\left\{(X_{uv})_+^2\right\}\right] &\leq \mathbb{E}\left[\max\left\{X_{uv}^2\right\}\right] \\ &= \mathbb{E}\left[\max_{\|a\|=1,\|b\|=1}\left\{(\|S\|\langle h, Tb\rangle + \|Tb\|\langle g, S^\top a\rangle)^2\right\}\right]\end{aligned}$$

We can expand out the terms and use Cauchy-Schwarz to turn the dot prodcts into a product of norms:

$$\left(\|S\|\langle h, Tb\rangle + \|Tb\|\langle g, S^\top a\rangle\right)^2 = \|S\|^2\langle h, Tb\rangle^2 + 2\|S\|\|Tb\|\langle h, Tb\rangle\langle g, S^\top a\rangle + \|Tb\|^2\langle g, S^\top a\rangle^2$$

Let's analyse this term by term. We will make extensive use of the Cauchy-Schwarz inequality. The first term can be bounded as:

$$\begin{aligned}\|S\|^2\langle h, Tb\rangle^2 &= \|S\|^2\langle T^\top h, b\rangle \\ &\leq \|S\|^2\|T^\top h\|^2\|b\|^2 \\ &= \|S\|^2\|T^\top h\|^2\end{aligned}$$

The second term is bounded as:

$$\begin{aligned}2\|S\|\|Tb\|\langle h, Tb\rangle\langle g, S^\top a\rangle &= 2\|S\|\|Tb\|\langle T^\top h, b\rangle\langle Sg, a\rangle \\ &\leq 2\|S\|\|Tb\|\|T^\top h\|\|b\|\|Sg\|\|a\| \\ &= 2\|S\|\|Tb\|\|T^\top h\|\|Sg\|\end{aligned}$$

The third term is bounded as:

$$\begin{aligned}\|Tb\|^2\langle g, S^\top a\rangle^2 &= \|Tb\|^2\langle Sg, a\rangle^2 \\ &\leq \|Tb\|^2\|Sg\|^2\|a\|^2 \\ &= \|Tb\|^2\|Sg\|^2\end{aligned}$$

So we end up with:

$$\begin{aligned}\mathbb{E}\left[\max_{\|a\|=1,\|b\|=1}\left\{(\|S\|\langle h, Tb\rangle + \|Tb\|\langle g, S^\top a\rangle)^2\right\}\right] &\leq \mathbb{E}\left[\|S\|^2\|T^\top h\|^2 + 2\|S\|\|Tb\|\|T^\top h\|\|Sg\|\right] \\ &\quad + \mathbb{E}\left[\|Tb\|^2\|Sg\|^2\right]\end{aligned}$$

Since h and g are independant and they are standard normal vectors, we have the equality:

$$\begin{aligned}\mathbb{E}\left[\|T^\top h\|^2\right] &= \|T\|_F^2 \\ \mathbb{E}\left[\|Sg\|^2\right] &= \|S\|_F^2\end{aligned}$$

Next we need to make use of Hölder's inequality for expectations which (in our particular case of the ℓ_2 norm) can be written as:

$$\mathbb{E}\left[\|XY\|\right] \leq \sqrt{\mathbb{E}\left[\|X\|^2\right]}\sqrt{\mathbb{E}\left[\|Y\|^2\right]}$$

Using this version of Hölder's inequality we can bound the middle term as:

$$\begin{aligned}2\|S\|\|Tb\|\mathbb{E}\left[\|T^\top h\|\|Sg\|\right] &\leq 2\|S\|\|Tb\|\sqrt{\mathbb{E}\left[\|T^\top h\|^2\right]}\sqrt{\mathbb{E}\left[\|Sg\|^2\right]} \\ &= 2\|S\|\|Tb\|\sqrt{\|T\|_F^2}\sqrt{\|S\|_F^2} \\ &= 2\|S\|\|Tb\|\|T\|_F\|S\|_F\end{aligned}$$

Using the two factrs above we can bound $\mathbb{E}\left[\|S\|^2\|T^\top h\|^2 + 2\|S\|\|Tb\|\|T^\top h\|\|Sg\| + \|Tb\|^2\|Sg\|^2\right]$ as:

$$\begin{aligned}\mathbb{E}\left[\|S\|^2\|T^\top h\|^2 + 2\|S\|\|Tb\|\|T^\top h\|\|Sg\| + \|Tb\|^2\|Sg\|^2\right] \leq \|S\|^2\|T\|_F^2 + 2\|S\|\|T\|\|T\|_F\|S\|_F \\ + \|Tb\|^2\|S\|_F^2\end{aligned}$$

As we can see this is a perfect square. After factorizing we are left with the following bound:

$$\mathbb{E}\left[\max\left\{(X_{uv})_+^2\right\}\right] \leq (\|S\|\|T\|_F + \|T\|\|S\|_F)^2$$

Next we use Corolary 3.12 on p.75 of [3] and some relations from probability to finish the question. The calculations of which are listed below. Recall that for a random variable Z we have:

$$\mathbb{E}\left[Z^2\right] = \int_0^\infty 2x\mathbb{P}(Z > x)dx$$

This then allows us to say:

$$\mathbb{E}\left[\|SGT\|^2\right] \leq \mathbb{E}\left[\max_{u,v}\left\{(Y_{uv})_+^2\right\}\right] = \int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(Y_{uv})_+\right\} > t\right)dt$$

Since everything here is positive we can drop $(\cdot)_+$ and apply Corolary 3.12:

$$\begin{aligned}\int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(Y_{uv})_+\right\} > t\right)dt &= \int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(Y_{uv})\right\} > t\right)dt \\ &\leq \int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(X_{uv})\right\} > t\right)dt\end{aligned}$$

Again we can bring in the $(\cdot)_+$ without changing this integral:

$$\begin{aligned}\int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(X_{uv})\right\} > t\right)dt &= \int_0^\infty 2t\mathbb{P}\left(\max_{u,v}\left\{(X_{uv})_+\right\} > t\right)dt \\ &= 2 * \mathbb{E}\left[\max_{u,v}\left\{(X_{uv})\right\}\right] \\ &\leq 2 * (\|S\|\|T\|_F + \|T\|\|S\|_F)^2\end{aligned}$$

And so we have complated the proof showing that:

$$\mathbb{E}\left[\|SGT\|^2\right] \leq 2\left(\|S\|\|T\|_F + \|T\|\|S\|_F\right)^2$$

1.4 Question 4

The results of this section can be found by runnning the MATLAB script.

Bibliography

[1] Z. Frangella, J. A. Tropp, and M. Udell, "Randomized nyström preconditioning," 2021. [Online]. Available: https://arxiv.org/abs/2110.02820

[2] M. Díaz, E. N. Epperly, Z. Frangella, J. A. Tropp, and R. J. Webber, "Robust, randomized preconditioning for kernel ridge regression," 2024. [Online]. Available: https://arxiv.org/abs/2304.12465

[3] M. Ledoux, Probability in Banach spaces. Berlin: Springer, 2011. [Online]. Available: https://openlibrary.org/books/OL25189609M/Probability_in_Banach_spaces