# Randomized Nyström Preconditioning with RPCholesky:

a) Let us analyze the approximation $\hat{A}^{(k)}$ returned after $k$ steps of the RPCholesky Algorithm. Note that throughout this paper we use notation $\|.\|_2 = \|.\|$ alternatively to denote the spectral norm.
We define

$$\hat{A}^{(k)} = A(:, I)A(I, I)^{-1}A(I, :)$$

With $I = \{s_1, ..., s_k\}$ the output of RPCholesky, note that the random pivots $s_i$ are not independent for different $i$.
We also define the p-stable rank, which reflects decay in the tail eigenvalues as:

$$sr_p(A) = \lambda_p^{-1} \sum_{j=p}^{n} \lambda_j.$$

Notice that given $\lambda_p \neq 0$, $1 \leqslant sr_p(A) \leqslant (n - p)$.
We have,

$$sr_p(A)\lambda_p = \sum_{j=p}^{n} \lambda_j = \operatorname{tr}(A - \mathcal{T}_{p-1}(A)),$$

where $\mathcal{T}_p(A)$ denotes the best rank $p$ approximation of $A$.
From L6S76 we have that if

$$k \geqslant \frac{r}{\varepsilon} + r \log\left(\frac{1}{\varepsilon\eta}\right)$$

where $\eta := \frac{\operatorname{tr}(A-\mathcal{T}_r(A))}{\operatorname{tr}(A)}$, then

$$\mathbb{E}[\operatorname{tr}(A - \hat{A}^{(k)})] \leqslant (1 + \varepsilon) \operatorname{tr}(A - \mathcal{T}_r(A)) = (1 + \varepsilon)sr_{r+1}(A)\lambda_{r+1}.$$

Thus, taking $r + 1 = p$ and $\varepsilon = 2$ we get that if

$$k \geqslant (p - 1)(\frac{1}{2} + \log\left(\frac{1}{2\eta}\right)),$$

then,

$$\mathbb{E}[\operatorname{tr}(A - \hat{A}^{(k)})] \leqslant 3sr_p(A)\lambda_p$$

which is the desired result. □

b) We consider that we are solving the following regularized linear system:

$$(A + \mu I)x = b \quad \mu \geqslant 0.$$

We denote the regularized matrix as $A_\mu = (A + \mu I)$, the regularizer has the effect of adding a value of $\mu$ to the eigenvalues of $A$, pushing them away from $0$ and alleviating

some potential ill-conditioning problems.
Given the eigenvalue decomposition $\hat{A}^k = U\hat{\Lambda}U^\top$ of the RPCholesky approximation, we define the following preconditioner

$$P = \frac{1}{\hat{\lambda}_k + \mu} U(\hat{\Lambda} + \mu I)U^\top + (I - UU^\top).$$

where $\hat{\lambda}_k$ is the $k$th largest eigenvalue of $\hat{A}^{(k)}$ (the smallest non-zero eigenvalue of $\hat{A}^{(k)}$.
We recall the condition number is defined as

$$\kappa_2 = \|A\|_2 \left\|A^{-1}\right\|_2 = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}.$$

We also define the effective dimension of $A_\mu$ as

$$d_{eff}(\mu) = \mathrm{tr}(AA_\mu^{-1}) = \sum_{j=1}^{n} \frac{\lambda_j(A)}{\lambda_j(A) + \mu}$$

this quantity can be viewed as a smoothed count of eigenvalues significantly larger than $\mu$. It is often much smaller than the nominal dimension $n$, since a large number of real world matrices exhibit strong spectral decay.
Let us now find a deterministic bound for the condition number of a rank k RPCholesky Preconditioner.
We have

$$\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) = \frac{\lambda_1(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}})}{\lambda_n(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}})}$$

since $P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}$ is psd.
Note that we can decompose $A_\mu = \hat{A}^{(k)} + I\mu + A - \hat{A}^{(k)}$ which gives:

$$P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}} = P^{-\frac{1}{2}}(\hat{A}^{(k)} + I\mu)P^{-\frac{1}{2}} + P^{-\frac{1}{2}}(A - \hat{A}^{(k)})P^{-\frac{1}{2}}$$

Note that since $A - \hat{A}^{(k)}$ is psd, $P^{-\frac{1}{2}}(A - \hat{A}^{(k)})P^{-\frac{1}{2}}$ is also psd.
Let us bound the maximum eigenvalue. Weyl's inequalities implies that

$$\lambda_1(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant \lambda_1(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) + \lambda_1(P^{-\frac{1}{2}}(A - \hat{A}^{(k)})P^{-\frac{1}{2}})$$

Note that,

$$\hat{A}^{(k)} + \mu I = U\hat{\Lambda}U^\top + \mu I = U((\hat{\Lambda} + \mu I)U^\top + \mu(I - UU^\top))$$

We then have

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

$$= (\hat{\lambda}_k + \mu)P^{-\frac{1}{2}}(P - \frac{\lambda_k}{\lambda_k + \mu}(I - UU^\top))P^{-\frac{1}{2}} = (\hat{\lambda}_k + \mu)I - \lambda_k P^{-\frac{1}{2}}(I - UU^\top)P^{-\frac{1}{2}}$$

We have $P^{-\frac{1}{2}}(I - UU^\top)P^{-\frac{1}{2}}$ is psd, and since $(I - UU^\top)$ is an orthogonal projector on $U_\perp$, it is equal to zero on the subspace spanned by $U$, and equal to the identity on the span of $U_\perp$. hence,

$$\lambda_1(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) = \lambda_1((\hat{\lambda}_k + \mu)I - \lambda_k P^{-\frac{1}{2}}(I - UU^\top)P^{-\frac{1}{2}}) = \hat{\lambda}_k + \mu.$$

We also notice (for later)

$$\lambda_n(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) = \lambda_n((\hat{\lambda}_k + \mu)I - \lambda_k P^{-\frac{1}{2}}(I - UU^\top)P^{-\frac{1}{2}}) = \hat{\mu}$$

Setting $k < n$ we have that the largest eigenvalue of $P^{-1}$ is equal to the smallest eigenvalue of $P$.
On the subspace spanned by $U$, $(I - UU^\top) = 0$, hence

$$P_U = U(\frac{\hat{\Lambda} + \mu}{\hat{\lambda}_k + \mu})U^\top$$

On the span of $U$ this matrix has eigenvalues of $1$ or greater, with $1$ achieved at the $k$th eigenvalue.
On the span of $U_\perp$, $(I - UU^\top)$ acts as the identity and thus has eigenvalues $1$. Hence the smallest eigenvalue of $P$ is one, and

$$\lambda_1(P^{-1}) = 1.$$

Using this we have:

$$\lambda_1(P^{-\frac{1}{2}}(A - \hat{A}^{(k)})P^{-\frac{1}{2}}) = \lambda_1(P^{-1}(A - \hat{A}^{(k)})) \leqslant \lambda_1(P^{-1})\lambda_1(A - \hat{A}^{(k)}) = \left\|A - \hat{A}^{(k)}\right\|_2$$

Hence, bringing everything together we have

$$\lambda_1(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant \hat{\lambda}_k + \mu + \left\|A - \hat{A}^{(k)}\right\|_2.$$

Let us now bound the smallest eigenvalue. We assume $\mu > 0$. Applying Weyl's inequality to our decomposition we get

$$\lambda_n(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \geqslant \lambda_n(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) + \lambda_n(P^{-\frac{1}{2}}(A - \hat{A}^{(k)})P^{-\frac{1}{2}})$$

$$\geqslant \lambda_n(P^{-\frac{1}{2}}(\hat{A}^{(k)} + \mu I)P^{-\frac{1}{2}}) = \mu$$

. Hence putting both bounds together we get:

$$\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant \frac{\hat{\lambda}_k + \mu + \left\|A - \hat{A}^{(k)}\right\|_2}{\mu}. \tag{1}$$

To simplify this bound further we must first prove some lemmas.
Let us show that $\hat{A}^k$ can be decomposed as $\hat{A}^k = A^{1/2}\Pi A^{1/2}$, where $\Pi$ is an orthogonal projector. Note that since $A$ is SPSD, it has a well defined square root.
We define

$$V = A^{\frac{1}{2}}(:, I).$$

We then have

$$\begin{cases} A(:, I) = A^{\frac{1}{2}}V \\ A(I, I)^{-1} = (V^\top V)^{-1} \\ A(I, :) = V^\top A^{\frac{1}{2}} \end{cases}$$

Hence,

$$\hat{A}^{(k)} = A^{\frac{1}{2}}V(V^\top V)^{-1}V^\top A^{\frac{1}{2}}$$

Define

$$W := V(V^\top V)^{-\frac{1}{2}}$$

then,

$$\hat{A}^{(k)} = A^{\frac{1}{2}}WW^\top A^{\frac{1}{2}}$$

We define $\Pi := WW^\top$.
We see trivially that $\Pi^T = \Pi$, so $\Pi$ is symmetric.
Let us check that $\Pi$ is idempotent:

$$W^\top W = (V^\top V)^{-\frac{1}{2}}V^\top V(V^\top V)^{-\frac{1}{2}} = I \Rightarrow$$

$$\Pi^2 = W^\top(WW^\top)W = W^\top W = \Pi.$$

Hence $\Pi$ is an orthogonal projector.
We then have for $1 \leqslant i \leqslant n$

$$\lambda_i(\hat{A}^{(k)}) = \lambda_i(A^{\frac{1}{2}}\Pi A^{\frac{1}{2}}) \leqslant \lambda_i(\Pi)\lambda_i(A) = \lambda_i(A).$$

We can thus reduce **??** to

$$\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant \frac{\lambda_k + \mu + \left\|A - \hat{A}^{(k)}\right\|_2}{\mu}. \tag{2}$$

We then use lemma $5.4$ item $1$ in reference 2. The proof is the same in our case as in the paper, as it does not depend on whether we use Nystrom or RPCholesky (does not depend on $\hat{A}^{(k)}$). We restate the proof of item $1$ and $2$ here varbatim:
Fix a parameter $\gamma \geq 1$, and set $j_\star = \max\{1 \leq j \leq n : \lambda_j > \gamma\mu\}$.
We can bound the effective dimension below by the following mechanism.

$$d_{\text{eff}}(\mu) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu} \geq \sum_{j=1}^{j_\star} \frac{\lambda_j}{\lambda_j + \mu} \geq j_\star \cdot \frac{\lambda_{j_\star}}{\lambda_{j_\star} + \mu}.$$

We have used the fact that $t \mapsto t/(1+t)$ is increasing for $t \geq 0$, Solving for $j_\star$, we determine that

$$j_\star \leq (1 + \mu/\lambda_{j_\star}) d_{\text{eff}}(\mu) < (1 + \gamma^{-1}) d_{\text{eff}}(\mu).$$

The last inequality depends on the definition of $j_\star$. This is the required result.
Item 2 follows from a short calculation:

$$\begin{aligned}
\frac{1}{k}\sum_{j>k}\lambda_j &= \frac{\lambda_k+\mu}{k}\sum_{j>k}\frac{\lambda_j}{\lambda_k+\mu} \leq \frac{\lambda_k+\mu}{k}\sum_{j>k}\frac{\lambda_j}{\lambda_j+\mu} \\
&= \frac{\lambda_k+\mu}{k}\left(d_{\text{eff}}(\mu) - \sum_{j=1}^{k}\frac{\lambda_j}{\lambda_j+\mu}\right) \leq \frac{\lambda_k+\mu}{k}\left(d_{\text{eff}}(\mu) - \frac{k\lambda_k}{\lambda_k+\mu}\right) \\
&= \frac{\mu\, d_{\text{eff}}(\mu)}{k} + \lambda_k\left(\frac{d_{\text{eff}}(\mu)}{k} - 1\right) \leq \frac{\mu\, d_{\text{eff}}(\mu)}{k}.
\end{aligned}$$

The last inequality depends on the assumption that $k \geq d_{\text{eff}}(\mu)$.

Choosing $\gamma = 1$, we get that if $j \geqslant 2d_{\text{eff}}(\mu)$, then $\lambda_j \leqslant \mu$. Choosing a sketch size $k \geqslant 2d_{\text{eff}}(\mu)$ we thus get $\lambda_k \leqslant \mu$, and we can further simplify **??** into

$$\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant \frac{2\mu + \left\|A - \hat{A}^{(k)}\right\|_2}{\mu} = 2 + \frac{\left\|A - \hat{A}^{(k)}\right\|_2}{\mu}. \tag{3}$$

We now want to choose $k, p$ such that

$$\begin{cases} k \geqslant (p-1)(\frac{1}{2} + \log\left(\frac{1}{2\eta}\right)) \\ k \geqslant 2d_{\text{eff}}(\mu) \end{cases}$$

Note that

$$\frac{1}{\eta} = \frac{\operatorname{tr}(A)}{sr_p(A)\lambda_p} = \frac{\sum_{i=1}^{n}\lambda_i}{\sum_{i=p}^{n}\lambda_i} = 1 + \frac{\sum_{i=1}^{p-1}\lambda_i}{\sum_{i=p}^{n}\lambda_i} \leqslant 1 + \frac{(p-1)}{(n-p)}\kappa_2(A)$$

Using the fact that $\log(1+x) \leqslant x$ we can reduce the first criterion to

$$k \geqslant (p-1)(\frac{1}{2} - \log(2) + \frac{\sum_{i=1}^{p-1}\lambda_i}{\sum_{i=p}^{n}\lambda_i}))$$

or if we do not have access to all the eigenvalues of $A$:

$$k \geqslant (p-1)(\frac{1}{2} - \log(2) + \frac{(p-1)}{(n-p)}\kappa_2(A)).$$

Assuming that these conditions on $k$ and $p$ are met, taking the expectation of **??** and using our result from part a) we get:

$$\mathbb{E}\,\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant 2 + \frac{3sr_p(A)\lambda_p}{\mu} = 2 + \frac{3\sum_{j>(p-1)}\lambda_j}{\mu}$$

From lemma 5.4 item 2 (with $k$ in the lemma equal to $(p-1)$), given $(p-1) \geqslant d_{eff}(\mu)$ we have

$$\sum_{j>(p-1)} \lambda_j \leqslant \mu d_{eff}(\mu)$$

hence

$$\mathbb{E}\,\kappa_2(P^{-\frac{1}{2}}(A_\mu)P^{-\frac{1}{2}}) \leqslant 2 + 3d_{eff}(\mu) \leqslant 2 + 6k$$

where we have chosen $p = d_{eff}(\mu) + 1$ and $k = 2(p-1)$. Note that this bound only holds if $\frac{1}{\eta} \leqslant 2e^{\frac{3}{2}} \Leftrightarrow 2\eta \geqslant e^{-\frac{3}{2}}$ or if $\frac{d_{eff}(\mu)}{(n-1-d_{eff}(\mu))}\kappa_2(A) \leqslant e^{-\frac{3}{2}} - 1$.
This bound shows that if the effective dimension of $\mu$ is low, and we choose a rank k approximation with $k$ proportional to it, then the RPCholesky preconditioner will be very effective, and on average will be well conditioned. This means that the resulting matrix can then be fed into condition number dependent algorithms such as the conjugate gradient method.

c) Let us now prove the square chevet bound. Here is the statement.
Fix matrices $S \in \mathbb{R}^{r\times m}$ and $T \in \mathbb{R}^{n\times s}$ and let $G \in \mathbb{R}^{m\times n}$ be a standard Gaussian matrix. Then

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

We first define

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

$U$ represents the projection of the unit sphere of $\mathbb{R}^r$ onto $\mathbb{R}^m$ by $S^T$, while $V$ is the projection of the unit sphere of $\mathbb{R}^s$ onto $\mathbb{R}^n$.
We then take $u \in U$ and $v \in V$ and consider the following Gaussian Processes:

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

and

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

where:

- $G \in \mathbb{R}^{m\times n}$ is a Gaussian random matrix,
- $g, h$ are Gaussian random vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively,
- $\gamma$ is $N(0,1)$ in $\mathbb{R}$.
- and $G, g, h$ and $\gamma$ are all independent.

We recall from its definition that $\|v\|_2 = \|Tb\|_2$ for some $b$ with $\|b\|_2 = 1$. We note that both processes are centered processes around zero.

Let us now state Slepian's lemma:
Given $X, Y$ Gaussian in $R^N$ s.t

$$\begin{cases} \mathbb{E}(X_i^2) = \mathbb{E}(Y_i^2) \quad \forall i \\ \mathbb{E}(X_i X_j) \leqslant \mathbb{E}(Y_i Y_j) \quad \forall i \neq j \end{cases}$$

then for all real numbers $\lambda_i$, $i \leqslant N$:

$$\mathbb{P}\left\{\cup_{i=1}^N (Y_i > \lambda_i)\right\} \leqslant \mathbb{P}\left\{\cup_{i=1}^N (X_i > \lambda_i)\right\}$$

Let us now prove that our two processes satisfy the conditions of Slepian's lemma. The idea behind these conditions is that we want $Y$ to be "more correlated" than $X$ so that it's "harder", or less likely, for $Y$ to exceed the given threshold on all coordinates at the same time.
Let us prove that the first condition holds. We have

$$Y_{uv}^2 = \langle u, Gv\rangle^2 + \langle u, Gv\rangle \|S\|_2 \|v\|_2 \gamma + \|S\|_2^2 \|v\|_2^2 \gamma^2$$

Let us analyze the expectation term by term.
Since all entries of $G$ are independent Gaussian variables, $Gv$ is a Gaussian vector in $R^m$ with covariance $\|v\|_2^2 I_m$, thus:

$$\mathbb{E}\,\langle u, Gv\rangle^2 = \|u\|_2^2 \|v\|_2^2$$

then similarly since $\gamma$ is a standard Gaussian random variable, $\mathbb{E}\,\gamma^2 = 1$ and thus:

$$\mathbb{E}\,(\|S\|_2^2 \|v\|_2^2 \gamma^2) = \|S\|_2^2 \|v\|_2^2\, \mathbb{E}\,\gamma^2 = \|S\|_2^2 \|v\|_2^2$$

and since $\mathbb{E}(\gamma) = 0$, $\mathbb{E}\,\langle u, Gv\rangle = 0$, and the two are independent:

$$\mathbb{E}\,(\langle u, Gv\rangle \|S\|_2 \|v\|_2 \gamma) = \|S\|_2 \|v\|_2\, \mathbb{E}(\langle u, Gv\rangle)\, \mathbb{E}\,\gamma = 0$$

hence,

$$\mathbb{E}\,Y_{uv}^2 = \|u\|_2^2 \|v\|_2^2 + \|S\|_2^2 \|v\|_2^2$$

Then,

$$X_{uv}^2 = \|S\|_2^2 \langle h, v\rangle^2 + \|v\|_2^2 \langle g, u\rangle^2 + \|S\|_2 \|v\|_2 \langle h, v\rangle \langle g, u\rangle$$

And by the same reasoning:

$$\begin{cases} \mathbb{E}(\|S\|_2^2 \langle h, v\rangle^2) = \|S\|_2^2 \|v\|_2^2 \\ \mathbb{E}(\|v\|_2^2 \langle g, u\rangle^2) = \|v\|_2^2 \|u\|_2^2 \\ \mathbb{E}(\|S\|_2 \|v\|_2 \langle h, v\rangle \langle g, u\rangle) = 0 \end{cases}$$

thus,

$$\mathbb{E}\,X_{uv}^2 = \|u\|_2^2 \|v\|_2^2 + \|S\|_2^2 \|v\|_2^2 = \mathbb{E}\,Y_{uv}^2$$

which proves the first condition holds.
Let us now prove the second condition.
Expanding the $X$ variance term for some $(u, v) \neq (u', v')$ gives

$$X_{u,v}X_{u',v'} = (\|S\|_2\langle h, v\rangle + \|v\|_2\langle g, u\rangle)(\|S\|_2\langle h, v'\rangle + \|v'\|_2\langle g, u'\rangle)$$

When taking the expectation, terms with two independent Gaussians (cross terms) will vanish, while terms with a Gaussian squared (square terms) will remain, in the same manner as when we proved the previous condition.
Indeed,

$$\mathbb{E}(X_{u,v}X_{u',v'}) = v \cdot v'\|S\|_2^2 + v \cdot v'(u \cdot u') = v \cdot v'(\|S\|_2^2 + u \cdot u')$$

and similarly

$$Y_{u,v}Y_{u',v'} = (\langle u, Gv\rangle + \|S\|_2\|v\|_2\gamma)(\langle u', Gv'\rangle + \|S\|_2\|v'\|_2\gamma) \Rightarrow$$

$$\mathbb{E}(Y_{u,v}Y_{u',v'}) = u \cdot u'(v \cdot v') + \|S\|_2^2\|v\|_2\|v'\|_2$$

Then, by Cauchy Schwartz:

$$(v \cdot v') \leqslant \|v\|_2\|v'\|_2$$

thus,

$$\mathbb{E}(X_{u,v}X_{u',v'}) \leqslant \mathbb{E}(Y_{u,v}Y_{u',v'})$$

which proves that the second condition of the Slepian's lemma holds.
We can thus apply it and get

$$\mathbb{P}\left\{\cup_{u,v}Y_{uv} > t\right\} \leq \mathbb{P}\left\{\cup_{u,v}X_{uv} > t\right\} \Rightarrow$$

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

Throughout the proof, we will use notation $X_+ = \max X, 0$.
Let us begin to analyse the term $\mathbb{E}\,\|SGT\|_2^2$.
We first have, from the scalar product characterization of the spectral norm

$$\|SGT\|_2 = \max_{\|a\|_2=1,\|b\|_2=1} \langle S^\top a, GTb\rangle \Rightarrow$$

$$\|SGT\|_2^2 = \max_{\|a\|_2=1,\|b\|_2=1} \langle S^\top a, GTb\rangle^2$$

We define the function

$$f(\gamma) = \max_{u,v}(\langle u, Gv\rangle + \|S\|_2\|v\|_2\gamma)_+^2$$

Notice that

- $h(x) = (x)_+^2$ is convex in $x$
- $g(\gamma) = \langle u, Gv\rangle + \|S\|_2\|v\|_2\gamma$ is affine in $\gamma$.
- the maximum of a set of convex functions is also convex.

Hence, $f(\gamma) = \max_{u,v} h \circ g(\gamma)$ is convex in $\gamma$.
We can then apply Jensen's inequality to get:

$$\mathbb{E}_\gamma(f(\gamma)) \geqslant f(\mathbb{E}\,\gamma) = f(0) = \max_{u,v}(\langle u, Gv\rangle)_+^2$$

Then, using the definitions of $U$ and $V$:

$$\max_{u,v}(\langle u, Gv\rangle)_+^2 = \max_{\|a\|_2=1,\|b\|_2=1}(\langle S^\top a, GTb\rangle)_+^2 = \|SGT\|_2^2$$

And using the law of total expectation:

$$\mathbb{E}(\max_{u,v}(Y_{uv})_+^2) = \mathbb{E}_G(\mathbb{E}_\gamma(f(\gamma)|G)) \geqslant \mathbb{E}_G(\|SGT\|_2^2).$$

We now analyze $X_{uv}$, we first see that

$$\mathbb{E}\max_{u,v}(X_{uv})_+^2 \leq \mathbb{E}\max_{u,v} X_{uv}^2 = \mathbb{E}\max_{\|a\|=1,\|b\|=1}\left(\|S\|\langle h, Tb\rangle + \|Tb\|\langle g, S^Ta\rangle\right)^2,$$

we can drop the positive part since $\forall x \in \mathbb{R}$, $(x)_+^2 \leqslant x^2$.
Expanding the quadratic we get

$$\mathbb{E}\max_{\|a\|=1,\|b\|=1}\left(\|S\|^2\langle h, Tb\rangle^2 + \|Tb\|^2\langle g, S^Ta\rangle^2 + 2\|S\|\|Tb\|\langle h, Tb\rangle\langle g, S^Ta\rangle\right) \tag{4}$$

Cauchy Schwartz states that for vectors $u, v$:

$$|\langle u, v\rangle| \leqslant \|u\|_2\|v\|_2$$

hence

$$\begin{cases} \|S\|^2\langle h, Tb\rangle^2 = \|S\|^2\langle T^\top h, b\rangle^2 \leqslant \|S\|^2\left\|T^\top h\right\|^2\|b\|^2 = \|S\|^2\left\|T^\top h\right\|^2 \\ \|Tb\|^2\langle g, S^Ta\rangle^2 = \|Tb\|^2\langle Sg, a\rangle^2 \leqslant \|Tb\|^2\|Sg\|^2\|a\|^2 \leqslant \|T\|^2\|Sg\|^2 \\ \|S\|\|Tb\|\langle h, Tb\rangle\langle g, S^Ta\rangle \leqslant \|S\|\|T\|\left\|T^\top h\right\|\|Sg\| \end{cases}$$

So we bound **??** by

$$\leqslant \mathbb{E}\left(\|S\|^2\|T^Th\|^2 + 2\|S\|\|T\|\|T^Th\|\|Sg\| + \|T\|^2\|Sg\|^2\right) \tag{5}$$

Then, since $h, g$ are standard Gaussian vectors:

$$\mathbb{E}\left(\|S\|^2\|T^\top h\|^2 + \|T\|^2\|Sg\|^2\right) = \|S\|^2\|T\|_F^2 + \|T\|^2\|S\|_F^2$$

Holder's inequality for $p = 2, q = 2$ states that for two random variables $X, Y$,

$$\mathbb{E}(\|XY\|) \leqslant \sqrt{\mathbb{E}\,X^2\,\mathbb{E}\,Y^2}.$$

Setting $X = T^\top h$ and $Y = Sg$ we get

$$\mathbb{E}(\|T^\top h\|\|Sg\|) \leqslant \sqrt{\mathbb{E}\,\|T^\top h\|^2\,\mathbb{E}\,\|Sg\|^2} = \|T\|_{\mathrm{F}}\|S\|_{\mathrm{F}}.$$

Putting everything together, we can bound **??** by

$$\begin{aligned}&\leqslant \|S\|^2\|T\|_F^2 + 2\|S\|\|T\|\|S\|_F\|T\|_F + \|T\|^2\|S\|_F^2\\ &= (\|S\|\|T\|_F + \|T\|\|S\|_F)^2.\end{aligned}$$

We can now prove the whole square Chevet bound by assembling our current bounds and using integration by parts.
Recall that for a random variable X,

$$\mathbb{E}\,X = \int_0^\infty \mathbb{P}\,(X > t)\,dt$$

then,

$$\mathbb{E}\,X^2 = \int_0^\infty \mathbb{P}\left(X^2 > t\right)dt = \int_0^\infty \mathbb{P}\left(X > \sqrt{t}\right)dt.$$

using the substitution $u = \sqrt{t} \Leftrightarrow t = u^2$, $dt = 2udu$ this becomes

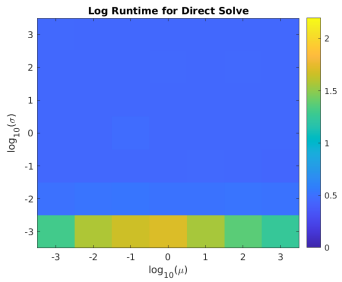
$$\int_0^\infty \mathbb{P}\left(X > \sqrt{t}\right)dt = \int_0^\infty 2u\mathbb{P}\,(X > u)\,du$$

applying this for $X = \max_{u,v}(Y_{uv})_+$ we finally get

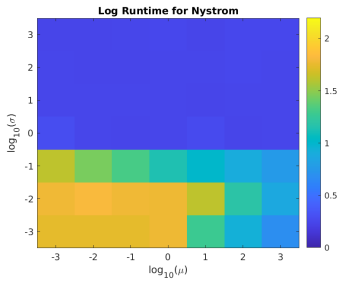
$$\begin{aligned}\mathbb{E}_G\,\|SGT\|^2 \leq \mathbb{E}\max_{u,v}(Y_{uv})_+^2 &= \int_0^\infty \mathbb{P}\left(\max_{u,v}(Y_{uv})_+^2 > t\right)dt\\ &= 2\int_0^\infty t\mathbb{P}\left(\max_{u,v}(Y_{uv})_+ > t\right)dt\\ &= 2\int_0^\infty t\mathbb{P}\left(\max_{u,v}Y_{uv} > t\right)dt \leq 2\int_0^\infty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

which proves the square Chevet bound. □

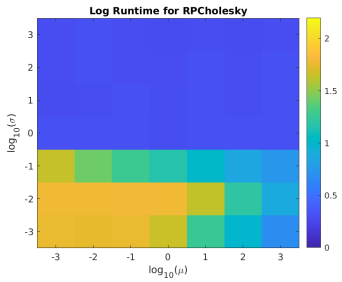
d) We now implement the Nystrom and RPCholesky preconditioners on a gaussian random matrix sampled from $[0, 1]$ with parameter $\sigma > 0$. We first plot a heatmap of the runtime for both algorithms with a direct solve using the backslash operator as comparaison.



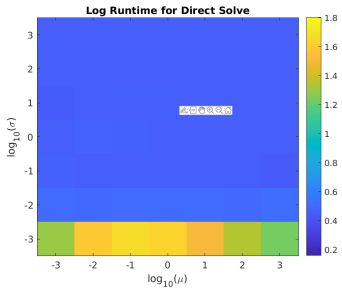

(a) Subfigure 1



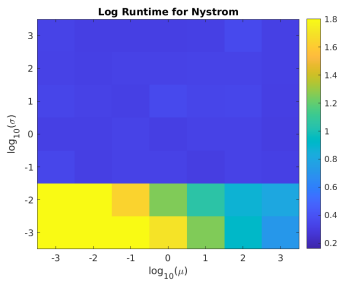

(b) Subfigure 2



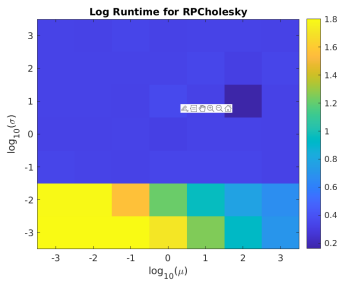

(c) Subfigure 3



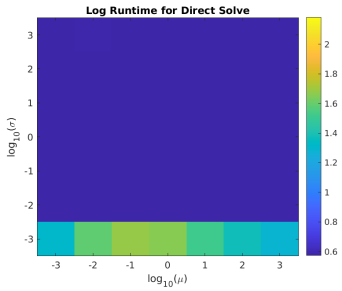

(d) Subfigure 4



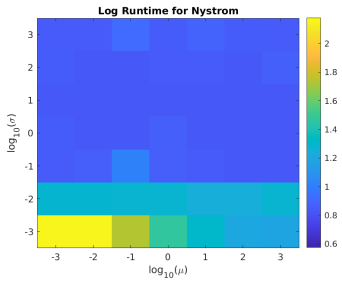

(e) Subfigure 5



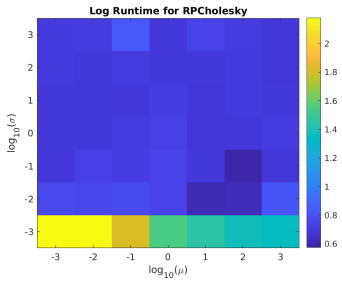

(f) Subfigure 6




(g) Subfigure 7




(h) Subfigure 8




(i) Subfigure 9

Figure 1: A 3x3 grid of subfigures.