1 Eigenvalue analysis

Consider preconditioning

$$\mathcal{Q}_\eta := (\eta I - \mathcal{L})^2 + \beta^2 I,$$

with a preconditioner $(\gamma I - \mathcal{L})^{-2}$ for some $\gamma \geq \eta$. The preconditioned operator takes the form

$$(\gamma I - \mathcal{L})^{-2}\mathcal{Q}_\eta = I - 2\frac{\gamma - \eta}{\gamma}(I - \tfrac{1}{\gamma}\mathcal{L})^{-1} + \frac{\beta^2 + (\gamma - \eta)^2}{\gamma^2}(I - \tfrac{1}{\gamma}\mathcal{L})^{-2}. \tag{1}$$ {{eq:gamma1}}

Suppose $\mathcal{L}$ is symmetric negative definite and, thus, has an orthogonal basis of eigenvectors, and consider the conditioning of (1). Assume that the eigenvalues of $(I - \frac{1}{\gamma}\mathcal{L})^{-1} \subset (0, 1)$, and are somewhat dense in this interval. This is to be expected for parabolic problems, where the eigenvalues of $-\mathcal{L}$ range from $\sim \delta t$ to $\sim \delta t/h^2$, which typically corresponds to $\sim (0, \infty)$ as $h, \delta t \to 0$.

Note that (1) is a quadratic polynomial in an SPD operator, and the eigenvalues of (1) are then a quadratic function $P(\lambda)$ of the eigenvalues $\{\lambda\}$ of $\mathcal{L}$, where

$$P(\lambda, \gamma) := \frac{\beta^2 + (\gamma - \eta)^2}{\gamma^2}\lambda^2 - 2\frac{\gamma - \eta}{\gamma}\lambda + 1. \tag{2}$$ {{eq:quadrati

Assume that we choose γ such that (1) is also SPD (choosing otherwise would be a poor choice in terms of conditioning). Then the condition number of (1) is given by

$$\text{cond}\left((\gamma I - \mathcal{L})^{-1}\mathcal{Q}_\eta\right) = \frac{\lambda_{\max}\left((\gamma I - \mathcal{L})^{-1}\mathcal{Q}_\eta\right)}{\lambda_{\min}\left((\gamma I - \mathcal{L})^{-1}\mathcal{Q}_\eta\right)}. \tag{3}$$ {{eq:cond2_0}

Again assuming that eigenvalues $\lambda \in \sigma(\mathcal{L})$ take on values $\lambda \in (0, 1)$, the condition number (18) can be expressed precisely as $h, \delta t \to 0$ via

$$\text{cond}\left((\gamma I - \mathcal{L})^{-1}\mathcal{Q}_\eta\right) = \frac{\max_{x\in(0,1)} P(x, \gamma)}{\min_{y\in(0,1)} P(y, \gamma)}. \tag{4}$$ {{eq:cond2_1}

With this closed form, it is natural to pose a minimization problem to find the optimal γ in terms of minimizing the condition number (3). We make the assumption that $\eta \leq \gamma \leq \eta^2 + \beta^2$, and consider the problem

$$\gamma_\times = \text{argmin}_{\gamma\geq\eta} \frac{\max_{x\in(0,1)} P(x, \gamma)}{\min_{y\in(0,1)} P(y, \gamma)}.$$

Note that $P(\lambda)$ (2) is a quadratic polynomial in λ, and thus its maximum over a closer interval $[0, 1]$ will be obtained at one of the endpoints,

$$P(0, \gamma) = 1, \qquad P(1, \gamma) = \frac{\eta^2 + \beta^2}{\gamma^2}.$$

For the maximum eigenvalue, this yields

$$\lambda_{\max} = \begin{cases} \frac{\eta^2+\beta^2}{\gamma^2} & \gamma < \sqrt{\eta^2 + \beta^2}, \\ 1 & \gamma \geq \sqrt{\eta^2 + \beta^2}. \end{cases} \tag{5}$$ {{eq:max0}}

The minimum eigenvalue will either be obtained at a critical point, or if there is no critical point in the interval $(0, 1)$, at the other endpoint than the maximum was obtained at. To consider the critical point, we differentiate (2) and obtain the root

$$\lambda_0 := \frac{\gamma(\gamma - \eta)}{\beta^2 + (\gamma - \eta)^2}. \tag{6}$$ {{eq:lambda_0

For $\gamma \geq \eta$, $\lambda_0 \geq 0$. To consider when $\lambda_0 \leq 1$, we can set it equal to one and rearrange for the equivalent condition

$$\gamma \leq \frac{\beta^2 + \eta^2}{\eta}. \tag{7}$$ {{eq:ass1}}

Assuming (7) holds, we have $\lambda_0 \in [0,1]$, and the minimum value of $P(\lambda,\gamma)$ in λ is achieved at λ_0,

$$\lambda_{\min} = \frac{\beta^2}{\beta^2 + (\gamma-\eta)^2}. \tag{8}$$ {{eq:min0}}

Combining (5), (7), and (8) yields

$$\operatorname{cond}\left((\gamma I - \mathcal{L})^{-1}\mathcal{Q}_\eta\right) = \begin{cases} \frac{(\eta^2+\beta^2)(\beta^2+(\gamma-\eta)^2)}{\beta^2\gamma^2} & \eta \le \gamma < \sqrt{\eta^2+\beta^2}, \\ \frac{\beta^2+(\gamma-\eta)^2}{\beta^2} & \sqrt{\eta^2+\beta^2} \le \gamma \le \frac{\eta^2+\beta^2}{\eta}. \end{cases} \tag{9}$$ {{eq:cases0}}

Here we have ended up at the result from The 2017 paper in (3.22) and (3.23), and they say both of the above equations are minimized at the interface

$$\gamma_\times := \sqrt{\eta^2+\beta^2}. \tag{10}$$ {{eq:gamma_op

2 Nonlinear/Schur complement

In the nonlinear setting we need to solve

$$\begin{bmatrix} \eta I - \widehat{\mathcal{L}} & \phi I \\ -\frac{\beta^2}{\phi} I & \eta I - \widehat{\mathcal{L}} \end{bmatrix}, \tag{11}$$ {{eq:block}}

with Schur complement of (11) given by

$$S := \eta I - \widehat{\mathcal{L}} + \beta^2 (\eta I - \widehat{\mathcal{L}})^{-1}. \tag{12}$$ {{eq:simpSchu

The initial idea is to consider a block lower triangular preconditioner for (11), given by

$$L_P := \begin{bmatrix} \eta I - \widehat{\mathcal{L}} & \mathbf{0} \\ -\frac{\beta^2}{\phi} I & \widehat{S} \end{bmatrix}^{-1}. \tag{13}$$ {{eq:Lprec}}

This raises the natural question as to how do we approximate S^{-1}? An easy first choice is to let $\widehat{S} := \eta I - \widehat{\mathcal{L}}$. Then the FOV analysis from the linear case immediately applies, and we know it is robust. Such an approach has the additional benefit of only requiring one preconditioner for both stages [OAK: This is true only for the simplified Newton case though; in the quasi-Newton algorithm, diagonal blocks in the 2×2 operator are different, so they don't use the same preconditioner anyway.]. Unfortunately, tests have also shown this choice to be suboptimal as the number of stages gets large, that is, convergence gets slower for higher order.

2.1 A factorization

In the linear setting, we were actually solving the equation

$$(\eta I - \widehat{\mathcal{L}})^2 + \beta^2 I,$$

which we found to be better (and scalably) preconditioned by $(kI - \widehat{\mathcal{L}})^{-2}$, for $k = \sqrt{\eta^2+\beta^2}$. How do we handle this with the Schur complement? One option is to factor S,

$$\begin{aligned} S &:= \Big((\eta I - \widehat{\mathcal{L}})^2 + \beta^2 I\Big)(\eta I - \widehat{\mathcal{L}})^{-1}, \\ \mapsto \qquad S^{-1} &= (\eta I - \widehat{\mathcal{L}})\Big((\eta I - \widehat{\mathcal{L}})^2 + \beta^2 I\Big)^{-1}, \end{aligned}$$

where we can then precondition the inverse term in S^{-1} exactly as we did in the linear setting. The downside here is we have introduced an additional solve, because now we must apply preconditioning to the (1,1)-block, followed by *two* preconditioning iterations to the Schur complement, as well as an additional matvec. That being said, for some of the linear advection-diffusion problems, the modified constant led to convergence $3-4\times$ faster, so it is possible this additional step of preconditioning is worth it.

Similarly, we can also suck the extra inverse out and solve it separately. Writing out the block LDU inverse of (11) we have

$$\begin{bmatrix} \eta I - \widehat{\mathcal{L}} & \phi I \\ -\frac{\beta^2}{\phi} I & \eta I - \widehat{\mathcal{L}} \end{bmatrix}^{-1} = \begin{bmatrix} I & -\phi(\eta I - \widehat{\mathcal{L}})^{-1} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} (\eta I - \widehat{\mathcal{L}})^{-1} & \mathbf{0} \\ \mathbf{0} & S^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \frac{\beta^2}{\phi}(\eta I - \widehat{\mathcal{L}})^{-1} & I \end{bmatrix}. \tag{14}$$ {{eq:ldu}}

In practice it is typically not advantageous to directly apply an LDU inverse, because when solving the Schur-complement inverse in an iterative fashion, each application of S requires computing an exact inverse of the (1,1)-block. However, with some algebra, we can rewrite (14) as

$$\begin{bmatrix} \eta I - \widehat{\mathcal{L}} & \phi I \\ -\frac{\beta^2}{\phi} I & \eta I - \widehat{\mathcal{L}} \end{bmatrix}^{-1} = \begin{bmatrix} (\eta I - \widehat{\mathcal{L}})^{-1} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} I & -\phi I \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \left((\eta I - \widehat{\mathcal{L}})^2 + \beta^2 I\right)^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \frac{\beta^2}{\phi} I & \eta I - \widehat{\mathcal{L}} \end{bmatrix}. \tag{15}$$ {{eq:ldu2}}

Here we have introduced an additional mat-vec by $\eta I - \widehat{\mathcal{L}}$, and otherwise separated the inverse into two separate pieces, $(\eta I - \widehat{\mathcal{L}})^{-1}$, which is a standard backward Euler step, and $\left((\eta I - \widehat{\mathcal{L}})^2 + \beta^2 I\right)^{-1}$, which is exactly the problem we solved in the linear setting, which we would precondition with two applications of $(kI - \widehat{\mathcal{L}})^{-1}$, for $k = \sqrt{\eta^2 + \beta^2}$. The nice thing about this problem and formulation is that although

2.2 A modified γ

Alternatively, suppose we precondition S with $(\gamma I - \widehat{\mathcal{L}})^{-1}$ for some $\gamma \neq \eta$? The preconditioned operator then takes the form

$$\begin{aligned} (\gamma I - \widehat{\mathcal{L}})^{-1} S &= (\gamma I - \widehat{\mathcal{L}})^{-1} \left[(\gamma I - \widehat{\mathcal{L}}) + (\eta - \gamma) I + \beta^2 (\eta I - \widehat{\mathcal{L}})^{-1} \right] \\ &= I - (\gamma - \eta)(\gamma I - \widehat{\mathcal{L}})^{-1} + \beta^2 (\gamma I - \widehat{\mathcal{L}})^{-1} (\eta I - \widehat{\mathcal{L}})^{-1} \\ &= I - \frac{\gamma - \eta}{\gamma} (I - \tfrac{1}{\gamma}\widehat{\mathcal{L}})^{-1} + \frac{\beta^2}{\gamma\eta} (I - \tfrac{1}{\gamma}\widehat{\mathcal{L}})^{-1} (I - \tfrac{1}{\eta}\widehat{\mathcal{L}})^{-1}. \end{aligned} \tag{16}$$ {{eq:gamma0}}

Suppose $-\mathcal{L}$ is SPD with a spectrum $\subset (0, \infty)$. Then the spectrum of (16) is given by

$$\mathcal{F}(\gamma, \lambda) := 1 - \frac{\gamma - \eta}{\gamma + \lambda} + \frac{\beta^2}{(\gamma + \lambda)(\eta + \lambda)}, \tag{17}$$ {{eq:eig_gamm

where $\lambda \in \sigma(-\mathcal{L})$. If we additionally choose γ such that (16) is also SPD (γ such that this does not hold would be a poor choice in terms of conditioning), the condition number of (16) is given by

$$\text{cond}\left((\gamma I - \widehat{\mathcal{L}})^{-1} S\right) = \frac{\lambda_{\max}\left((\gamma I - \widehat{\mathcal{L}})^{-1} S\right)}{\lambda_{\min}\left((\gamma I - \widehat{\mathcal{L}})^{-1} S\right)}. \tag{18}$$ {{eq:cond0}}

Again assuming that eigenvalues $\lambda \in \sigma\left(\widehat{\mathcal{L}}\right)$ take on values $\lambda \in (0, \infty)$, the condition number (18) can be expressed precisely as $h, \delta t \to 0$ via

$$\text{cond}\left((\gamma I - \widehat{\mathcal{L}})^{-1} S\right) = \frac{\max_{\lambda \in (0,\infty)} \mathcal{F}(\gamma, \lambda)}{\min_{\lambda \in (0,\infty)} \mathcal{F}(\gamma, \lambda)}. \tag{19}$$ {{eq:cond1}}

With this closed form, it is natural to pose a minimization problem to find the optimal γ in terms of minimizing the condition number (18). We make the assumption that $\eta \leq \gamma \leq \eta^2 + \beta^2$, and consider the problem

$$\gamma_* = \text{argmin}_{\gamma \geq \eta} \frac{\max_{\lambda \in (0,\infty)} \mathcal{F}(\gamma, \lambda)}{\min_{\lambda \in (0,\infty)} \mathcal{F}(\gamma, \lambda)}. \tag{20}$$ {{eq:gam_opt}}

Minima and maxima in λ may be obtained at one of the endpoints, $\lambda = 0$ or $\lambda \to \infty$, or at a critical point of (17) in λ. Taking the partial with respect to λ, we have

$$\frac{\partial \mathcal{F}}{\partial \lambda} = \frac{(\gamma - \eta)(\eta + \lambda)^2 - \beta^2(\gamma + \eta + 2\lambda)}{(\gamma + \lambda)^2 (\eta + \lambda)^2}. \tag{21}$$ {{eq:partial_

Noting that the denominator is nonnegative for $\eta, \gamma, \lambda > 0$, the critical points are obtained at zeros of the numerator in (21), which can be written as a quadratic polynomial in λ:

$$(\gamma - \eta)\lambda^2 - 2(\eta^2 + \beta^2 - \eta\gamma)\lambda + \gamma(\eta^2 - \beta^2) - \eta(\eta^2 + \beta^2) = 0.$$

Working through the algebra, for $\gamma > \eta$ the roots are given by

$$\lambda_\pm := \frac{\beta^2 + \eta^2 - \gamma\eta \pm \beta\sqrt{\eta^2 + \beta^2 + \gamma^2 - 2\gamma\eta}}{\gamma - \eta}. \tag{22}$$ {eq:roots}

For $\beta > 0$, it is straightforward to show that $\eta^2 + \beta^2 + \gamma^2 - 2\gamma\eta$ is a positive quadratic in γ with no real roots, which implies (22) defines two real roots.

Thus, we have four potential points at which a maximum or minimum in λ can be achieved, $\{0, \infty, \lambda_\pm\}$. Working through the algebra yields

$$\mathcal{F}(\gamma, \infty) = 1, \qquad \mathcal{F}(\gamma, \lambda_+) = \frac{2\beta}{\beta + \sqrt{\beta^2 + (\gamma - \eta)^2}},$$
$$\mathcal{F}(\gamma, 0) = 1 + \frac{\eta^2 + \beta^2 - \gamma^2}{\eta\gamma}, \qquad \mathcal{F}(\gamma, \lambda_-) = \frac{2\beta}{\beta - \sqrt{\beta^2 + (\gamma - \eta)^2}}.$$

In choosing γ, note that $\mathcal{F}(\gamma, \lambda_-) < 0$, which contradicts the assumption of positive definiteness. Thus, we must make an additional assumption that $\lambda_- \notin (0, \infty)$. From (22), this is equivalent to saying that

$$\begin{aligned}
&\beta^2 - \eta(\gamma - \eta) < \beta\sqrt{\beta^2 + (\gamma - \eta)^2}, \\
\iff \quad &\left(\beta^2 - \eta(\gamma - \eta)\right)^2 < \beta^2\left(\beta^2 + (\gamma - \eta)^2\right), \\
\iff \quad &(\gamma - \eta)\left[\beta^2(\gamma + \eta) - \eta^2(\gamma - \eta)\right] > 0, \\
\iff \quad &\frac{\beta^2}{\eta^2} > \frac{\gamma - \eta}{\gamma + \eta}.
\end{aligned}$$

Noting that $\frac{\gamma - \eta}{\gamma + \eta} < 1$, the above constraint clearly holds for $\beta > \eta$, which is the only regime in which we need a better constant anyways.

Assume now that $\beta > \eta$, in which case maxima and minima of $\mathcal{F}(\gamma, \lambda)$ in λ can be obtained at $\lambda \in \{0, \lambda_+, \infty\}$. Note for $\gamma < \sqrt{\eta^2 + \beta^2}$,

$$\max_{\lambda \in [0,\infty)} \mathcal{F}(\gamma, \lambda) = 1 + \frac{\eta^2 + \beta^2 - \gamma^2}{\eta\gamma} > 1,$$
$$\min_{\lambda \in [0,\infty)} \mathcal{F}(\gamma, \lambda) = \frac{2\beta}{\beta + \sqrt{\beta^2 + (\gamma - \eta)^2}},$$

while for $\gamma \geq \sqrt{\eta^2 + \beta^2}$,

$$\max_{\lambda \in [0,\infty)} \mathcal{F}(\gamma, \lambda) = 1,$$
$$\min_{\lambda \in [0,\infty)} \mathcal{F}(\gamma, \lambda) = \min\left\{\mathcal{F}(\gamma, 0), \mathcal{F}(\gamma, \lambda_+)\right\}.$$

Returning to (20), let us start with the case $\gamma \geq \sqrt{\eta^2 + \beta^2}$. We will do show by showing that both $\frac{1}{\mathcal{F}(\gamma,0)}$ and $\frac{1}{\mathcal{F}(\gamma,\lambda_+)}$ are minimized over $\gamma \geq \sqrt{\eta^2 + \beta^2}$ at $\gamma = \sqrt{\eta^2 + \beta^2}$ (because the maximum eigenvalue is 1). For $\mathcal{F}(\gamma, \lambda_+)$, taking the partial of $\frac{1}{\mathcal{F}(\gamma,\lambda_+)}$ with respect to γ yields

$$\frac{\gamma - \eta}{2\beta\sqrt{\beta^2 + (\gamma - \eta)^2}} > 0,$$

which implies the minimum is obtained at the beginning of the interval, in this case $\gamma = \sqrt{\eta^2 + \beta^2}$. Analogous derivations hold when evaluating at $\lambda = 0$, yielding the optimal $\gamma \geq \sqrt{\eta^2 + \beta^2}$ with respect to (20) given by $\gamma = \sqrt{\eta^2 + \beta^2}$.

For $\gamma < \sqrt{\eta^2 + \beta^2}$, we will also consider the derivative of (20) in γ to minimize, but without explicit construction. Consider γ_* as a product rule of $\lambda_{\max}(\gamma) \cdot \lambda_{\min}(\gamma)$, where

$$\lambda_{\max}(\gamma) := 1 + \frac{\eta^2 + \beta^2 - \gamma^2}{\eta\gamma}, \qquad \lambda_{\min}(\gamma) := \frac{\beta + \sqrt{\beta^2 + (\gamma - \eta)^2}}{2\beta}.$$

It is straightforward to verify that for all $\gamma \in (\eta, \sqrt{\eta^2+\beta^2})$, $\lambda_{\max}(\gamma) > 0$, $\lambda_{\min}(\gamma) > 0$, $\lambda'_{\min}(\gamma) > 0$, and $\lambda'_{\max}(\gamma) < 0$. The derivative of (20) is then given by

$$\mathcal{D}(\gamma) := \lambda_{\max}(\gamma)\lambda'_{\min}(\gamma) + \lambda'_{\max}(\gamma)\lambda_{\min}(\gamma),$$

and to show $\mathcal{D}(\gamma) < 0$ for $\gamma \in (\eta, \sqrt{\eta^2+\beta^2})$, it is sufficient to show that

$$-\lambda'_{\max}(\gamma)\lambda_{\min}(\gamma) > \lambda_{\max}(\gamma)\lambda'_{\min}(\gamma).$$

Plugging in, we want to show

$$\left(\frac{1}{\eta} + \frac{\eta^2+\beta^2}{\eta\gamma^2}\right)\left(\frac{\beta + \sqrt{\beta^2+(\gamma-\eta)^2}}{2\beta}\right) > \left(\frac{\gamma-\eta}{2\beta\sqrt{\beta^2+(\gamma-\eta)^2}}\right)\left(\frac{\eta\gamma+\eta^2+\beta^2-\gamma^2}{\eta\gamma}\right),$$

$$\left(\eta^2+\beta^2+\gamma^2\right)\left(\beta^2+(\gamma-\eta)^2+\beta\sqrt{\beta^2+(\gamma-\eta)^2}\right) > \gamma(\gamma-\eta)\left(\eta\gamma+\eta^2+\beta^2-\gamma^2\right).$$

Then note that for the first term on each side,

$$2\gamma(\gamma-\eta) = 2\gamma^2 - 2\gamma\eta < 2\gamma^2 \le \eta^2+\beta^2+\gamma^2.$$

For the second, first note that $\beta\sqrt{\beta^2+(\gamma-\eta)^2} < \beta^2$. Then we want to show that

$$\begin{aligned} 2\beta^2 + (\gamma-\eta)^2 &> \frac{1}{2}\left(\eta\gamma+\eta^2+\beta^2-\gamma^2\right), \\ 4\beta^2 + 2\gamma^2 + 2\eta^2 - 4\gamma\eta &> \eta\gamma+\eta^2+\beta^2-\gamma^2, \\ 3\beta^2 + 3\gamma^2 + \eta^2 &> 5\gamma\eta. \end{aligned}$$

Noting that $5\gamma\eta < 5\gamma^2$, it is sufficient to show that

$$\begin{aligned} 3\beta^2 + 3\gamma^2 + \eta^2 &> 5\gamma^2, \\ 3\beta^2 + \eta^2 &> 2\gamma^2. \end{aligned}$$

Finally, by assumption that $\gamma < \sqrt{\eta^2+\beta^2}$ and $\beta > \eta$, we have $2\gamma^2 < 2\eta^2 + 2\beta^2 < 3\beta^2 + \eta^2$.

Altogether, we have that for $\gamma \in (\eta, \sqrt{\eta^2+\beta^2})$,

$$\frac{\partial}{\partial\gamma}\left[\frac{\max_{\lambda\in[0,\infty)}\mathcal{F}(\gamma,\lambda)}{\min_{\lambda\in[0,\infty)}\mathcal{F}(\gamma,\lambda)}\right] < 0$$

meaning the optimal $\gamma \in (\eta, \sqrt{\eta^2+\beta^2})$ with respect to (20) is given by the maximum $\gamma = \sqrt{\eta^2+\beta^2}$. Plugging in, we can evaluate our resulting bound as

$$\operatorname{cond}\left((\gamma_* I - \widehat{\mathcal{L}})^{-1} S\right) = \frac{1}{2} + \frac{\sqrt{(\eta^2+\beta^2) - \eta\sqrt{\eta^2+\beta^2}}}{\sqrt{2}\beta}. \tag{23}$$ {{eq:cond_opt}}

This seems to small, need to check closer/make sure algebra is correct and the resulting condition numbers are reasonable.