Part XI, Chapter 50

Mixed finite element approximation

This chapter is concerned with the approximation of the model problem analyzed in Chapter 49. We focus on the Galerkin approximation in the conforming setting. We establish necessary and sufficient conditions for well-posedness, and we derive error bounds in terms of the best-approximation error. Then we consider the algebraic viewpoint, and we discuss augmented Lagrangian methods in the context of saddle point problems. Finally, we examine iterative solvers, including Uzawa iterations and Krylov subspace methods.

50.1 Conforming Galerkin approximation

Let V and Q be two reflexive (real) Banach spaces. Let a and b be two bounded bilinear forms on $V\times V$ and on $V\times Q$ respectively. Let $f \in V'$ and let $g \in Q'$. We consider the following model problem:

$$\begin{cases} \text{Find } u \in V \text{ and } p \in Q \text{ such that} \\ a(u,w) + b(w,p) = f(w), \quad \forall w \in V, \\ b(u,q) = g(q), \quad \forall q \in Q. \end{cases} \tag{50.1}$$

We introduce the associated operators $A \in \mathcal{L}(V;V')$ and $B \in \mathcal{L}(V;Q')$ such that $a(v,w) := \langle A(v), w\rangle_{V',V}$ and $b(v,q) := \langle B(v), q\rangle_{Q',Q}$. We assume that (50.1) is well-posed. Owing to Theorem 49.13, this means that the bilinear form a satisfies the conditions (49.36) (implying the inf-sup condition $\inf_{v\in\ker(B)} \sup_{w\in\ker(B)} \frac{|a(v,w)|}{\|v\|_V\|w\|_V} =: \alpha > 0$) and that the bilinear form b satisfies the inf-sup condition (49.37), i.e., $\inf_{q\in Q} \sup_{v\in V} \frac{|b(v,q)|}{\|v\|_V\|q\|_Q} =: \beta > 0$.

A conforming Galerkin approximation of (50.1) is obtained by considering two finite-dimensional subspaces $V_h \subset V$, $Q_h \subset Q$. The discrete problem is

$$\begin{cases} \text{Find } u_h \in V_h \text{ and } p_h \in Q_h \text{ such that} \\ a(u_h, w_h) + b(w_h, p_h) = f(w_h), & \forall w_h \in V_h, \\ b(u_h, q_h) = g(q_h), & \forall q_h \in Q_h. \end{cases} \tag{50.2}$$

50.1.1 Well-posedness

Let $B_h : V_h \to Q_h'$ be the discrete counterpart of the operator $B : V \to Q'$, that is, $\langle B_h(v_h), q_h\rangle_{Q_h',Q_h} := \langle B(v_h), q_h\rangle_{Q',Q} = b(v_h, q_h)$ for all $(v_h, q_h) \in V_h \times Q_h$. The null space of B_h is such that

$$\ker(B_h) = \{v_h \in V_h \mid \forall q_h \in Q_h,\, b(v_h, q_h) = 0\}. \tag{50.3}$$

One important aspect of the discretization is that the surjectivity of B does not imply that of B_h. One rare occasion where this is nevertheless the case is when $B^*(Q_h) \subset V_h$, i.e., $B_h^* = B^*_{|Q_h}$. This exceptional situation is illustrated in Exercise 49.6. Note also that in general, $\ker(B_h)$ is not necessarily a subspace of $\ker(B)$.

Proposition 50.1 (Well-posedness). (50.2) *is well-posed if and only if*

$$\inf_{v_h \in \ker(B_h)} \sup_{w_h \in \ker(B_h)} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_V} := \alpha_h > 0, \tag{50.4a}$$

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{|b(v_h, q_h)|}{\|v_h\|_V \|q_h\|_Q} := \beta_h > 0. \tag{50.4b}$$

Proof. Apply Theorem 49.13 and use the fact that (50.4a) implies both conditions in (49.36) since V_h is finite-dimensional; see Remark 26.7. □

The condition (50.4a) holds true for all conforming subspaces V_h and Q_h if a is V-coercive (the coercivity of a on $\ker(B)$ may not be sufficient since it may happen that $\ker(B_h) \not\subset \ker(B)$). Note that verifying the inf-sup condition for a on $V_h \times V_h$ is not sufficient to prove (50.4a) (think of an invertible matrix having a square diagonal sub-block that is not invertible; see Exercise 50.1). Furthermore, the condition (50.4b) is equivalent to B_h being surjective, which is also equivalent to B_h^* being injective since the setting is finite-dimensional. In practice, it is important that both (50.4a) and (50.4b) hold true uniformly w.r.t. $h \in \mathcal{H}$, i.e., $\inf_{h\in\mathcal{H}} \alpha_h =: \alpha_0 > 0$ and $\inf_{h\in\mathcal{H}} \beta_h =: \beta_0 > 0$.

50.1.2 Error analysis

Our goal is to estimate the approximation errors $(u - u_h)$ and $(p - p_h)$ in terms of the best-approximation error on u by a discrete field in V_h and the best-approximation error on p by a discrete function in Q_h. Céa's lemma (Lemma 26.13) could be applied to the bilinear form $t((v, q), (w, r)) := a(v, w) + b(w, q) + b(v, r)$ introduced in §49.4.2 (see Exercise 50.2). But here, we present a more specific analysis distinguishing the errors on u and on p.

We say that $\Pi_h \in \mathcal{L}(V;V_h)$ is a Fortin operator for the bilinear form b if $b(\Pi_h(v)-v,q_h)=0$ for all $q_h\in Q_h$ and all $v\in V$. (We do not assume V_h to be pointwise invariant under Π_h.) This class of operators is investigated in §26.2.3. In particular, Lemma 26.9 shows that the inf-sup condition (50.4b) implies the existence of a Fortin operator with $\|\Pi_h\|_{\mathcal{L}(V;V_h)} \le \frac{\|b\|}{\beta_h}$.

Lemma 50.2 (Error estimate). *Let (u,p) solve (50.1). Assume (50.4) and let (u_h,p_h) solve the discrete problem (50.2). Let $\Pi_h\in\mathcal{L}(V;V_h)$ be any Fortin operator. The following error estimates hold true:*

$$\|u-u_h\|_V \le c_{1h} \inf_{v_h\in\Pi_h(u)+\ker(B_h)} \|u-v_h\|_V + c_{2h} \inf_{q_h\in Q_h}\|p-q_h\|_Q, \quad (50.5a)$$

$$\|p-p_h\|_Q \le c_{3h} \inf_{v_h\in\Pi_h(u)+\ker(B_h)} \|u-v_h\|_V + c_{4h} \inf_{q_h\in Q_h}\|p-q_h\|_Q, \quad (50.5b)$$

with $c_{1h} := (1+\frac{\|a\|}{\alpha_h})$, $c_{2h} := \frac{\|b\|}{\alpha_h}$ if $\ker(B_h)\not\subset\ker(B)$ and $c_{2h}:=0$ otherwise, $c_{3h} := c_{1h}\frac{\|a\|}{\beta_h}$, and $c_{4h} := 1+\frac{\|b\|}{\beta_h}+c_{2h}\frac{\|a\|}{\beta_h}$.

Proof. (1) Estimate on $(u-u_h)$. Let $v_h\in\Pi_h(u)+\ker(B_h)$, i.e., $v_h := \Pi_h(u)+\gamma_h$ with $\gamma_h\in\ker(B_h)$. Then $u_h-v_h\in\ker(B_h)$ since we have

$$b(u_h-v_h,q_h) = b(u_h-\Pi_h(u),q_h)+b(\gamma_h,q_h) = b(u_h-u,q_h) = 0,$$

for all $q_h\in Q_h$, where we used the Galerkin orthogonality property for the second equation in (50.2). Owing the inf-sup condition (50.4a), we infer that

$$\begin{aligned}\alpha_h\|u_h-v_h\|_V &\le \sup_{y_h\in\ker(B_h)} \frac{|a(u_h-v_h,y_h)|}{\|y_h\|_V}\\ &= \sup_{y_h\in\ker(B_h)} \frac{|b(y_h,p-p_h)+a(u-v_h,y_h)|}{\|y_h\|_V},\end{aligned}$$

where the equality follows from the Galerkin orthogonality property for the first equation in (50.2), i.e., we have $a(u-u_h,y_h)+b(y_h,p-p_h)=0$ for all $y_h\in V_h$. If $\ker(B_h)\subset\ker(B)$, then $b(y_h,p-p_h)=0$ for all $y_h\in\ker(B_h)$, yielding

$$\alpha_h\|u_h-v_h\|_V \le \|a\|\,\|u-v_h\|_V.$$

In the general case, we have $b(y_h,p_h)=0=b(y_h,q_h)$ for all $q_h\in Q_h$, since y_h is in $\ker(B_h)$. This implies that

$$\alpha_h\|u_h-v_h\|_V \le \|a\|\,\|u-v_h\|_V + \|b\|\,\|p-q_h\|_Q.$$

Hence, both cases are summarized by the following estimate:

$$\|u_h-v_h\|_V \le \frac{\|a\|}{\alpha_h}\|u-v_h\|_V + c_{2h}\|p-q_h\|_Q,$$

with c_{2h} as in the assertion. Using the triangle inequality and taking the infimum over $v_h \in \Pi_h(u) + \ker(B_h)$ and over $q_h \in Q_h$ leads to (50.5a).
(2) Estimate on $(p - p_h)$. Using again the Galerkin orthogonality property for the first equation in (50.2), we have

$$b(v_h, q_h - p_h) = a(u_h - u, v_h) + b(v_h, q_h - p), \qquad \forall (v_h, q_h) \in V_h \times Q_h.$$

Combined with the inf-sup condition (50.4b), this implies that

$$\begin{aligned}\beta_h \|q_h - p_h\|_Q &\le \sup_{v_h \in V_h} \frac{|b(v_h, q_h - p_h)|}{\|v_h\|_V} \\ &\le \|a\| \, \|u - u_h\|_V + \|b\| \, \|p - q_h\|_Q.\end{aligned}$$

The bound (50.5b) follows from the triangle inequality, the bound on $(u - u_h)$, and by taking the infimum over $q_h \in Q_h$. □

The estimate on $(u - u_h)$ involves the best-approximation error on u by a member of the affine subspace $\Pi_h(u) + \ker(B_h)$. This error may not be easy to estimate in practice, and it is sometimes preferable to bound it by the best-approximation error on u by a member of V_h since $\Pi_h(u) + \ker(B_h) \subset V_h$. Of course, the best-approximation error in $\Pi_h(u) + \ker(B_h)$ is larger than the best-approximation error in V_h. The following lemma quantifies the discrepancy. (Recall that (50.4b) is equivalent to the existence of Fortin operators.)

Lemma 50.3 (Best-approximation in V_h). *Assume* (50.4b). *The following holds true for all $u \in V$ and any Fortin operator $\Pi_h \in \mathcal{L}(V; V_h)$:*

$$\inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V \le (1 + \|\Pi_h\|_{\mathcal{L}(V;V_h)}) \inf_{y_h \in V_h} \|u - y_h\|_V. \quad (50.6)$$

Proof. Let $u \in V$. Let $y_h \in V_h$ and set $z_h := \Pi_h(u - y_h)$. Then $y_h + z_h = \Pi_h(u) + y_h - \Pi_h(y_h) \in \Pi_h(u) + \ker(B_h)$ since $b(y_h - \Pi_h(y_h), q_h) = 0$ for all $q_h \in Q_h$. This implies that

$$\begin{aligned}\inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V &\le \|u - (y_h + z_h)\|_V \le \|u - y_h\|_V + \|z_h\|_V \\ &\le (1 + \|\Pi_h\|_{\mathcal{L}(V;V_h)}) \|u - y_h\|_V,\end{aligned}$$

and we conclude by taking the infimum over $y_h \in V_h$. □

Remark 50.4 ($g = 0$). If $g = 0$, then $\Pi_h(u) \in \ker(B_h)$, and the infimum in (50.5) and (50.6) reduces to $v_h \in \ker(B_h)$. □

Corollary 50.5 (Error estimate). *Let (u, p) solve* (50.1). *Assume* (50.4) *and let (u_h, p_h) solve the discrete problem* (50.2). *The following error estimates hold true:*

$$\|u-u_h\|_V \le c'_{1h} \inf_{v_h\in V_h} \|u-v_h\|_V + c_{2h} \inf_{q_h\in Q_h} \|p-q_h\|_Q, \tag{50.7a}$$

$$\|p-p_h\|_Q \le c'_{3h} \inf_{v_h\in V_h} \|u-v_h\|_V + c_{4h} \inf_{q_h\in Q_h} \|p-q_h\|_Q, \tag{50.7b}$$

with $c'_{1h} := (1+\frac{\|a\|}{\alpha_h})(1+\|\Pi_h\|_{\mathcal{L}(V;V_h)})$ *for every Fortin operator* $\Pi_h \in \mathcal{L}(V;V_h)$, $c'_{3h} := c'_{1h}\frac{\|a\|}{\beta_h}$, *and* c_{2h}, c_{4h} *are as in Lemma* 50.2.

Proof. Combine Lemma 50.2 with Lemma 50.3. □

Remark 50.6 (c'_{1h}). Lemma 26.9 asserts the existence of a Fortin operator with $\|\Pi_h\|_{\mathcal{L}(V;V_h)} \le \frac{\|b\|}{\beta_h}$. Hence, the upper bound $c'_{1h} \le (1+\frac{\|a\|}{\alpha_h})(1+\frac{\|b\|}{\beta_h})$ always holds true. However, the estimate $\|\Pi_h\|_{\mathcal{L}(V;V_h)} \le \frac{\|b\|}{\beta_h}$ can be pessimistic. For instance, for the Stokes equations in elongated domains, the boundedness constant of the bilinear form $b(\boldsymbol{v},p) = (\nabla\cdot\boldsymbol{v},p)_{L^2(D)}$ on $\boldsymbol{H}^1_0(D)\times L^2(D)$ is $\|b\| = 1$, and the inf-sup constant β_h can be shown to be very small (see Chizhonkov and Olshanskii [119], Dobrowolski [170]), whereas for some of these domains it is possible to construct a Fortin operator with norm of order unity (see Mardal et al. [294], Linke et al. [284]). □

Remark 50.7 ($\ker(B_h)$). We refer the reader to Theorem 51.16 for an example of error estimate exploiting the approximation properties in $\ker(B_h)$ in the context of Darcy's equations. □

Remark 50.8 (c_{2h}). The constant c_{2h} vanishes whenever $\ker(B_h)\subset\ker(B)$. Using a discrete pair (V_h,Q_h) that guarantees that $\ker(B_h)\subset\ker(B)$ may be interesting when the best approximation error on p is (much) larger than that on u. A simple example where this occurs is when $f = B^*(p)$ for some $p\in Q$ and $g = 0$, so that the solution to (50.1) is $(0,p)$. If $\ker(B_h)\subset\ker(B)$, the estimate (50.7a) implies that $u_h = u = 0$. But if $\ker(B_h)\not\subset\ker(B)$, then u_h is generally nonzero and grows linearly with the size of p, which is not a desirable property. More generally, the well-posedness of (50.1) with $g := 0$ implies the abstract *Helmholtz decomposition* $V' = Y_0\oplus Y_1$ with $Y_0 := A(\ker(B))$ and $Y_1 = \mathrm{im}(B^*)$. Whenever the component of f in Y_1 is much larger than that in Y_0, the best-approximation error on p dominates the approximation error on u unless the discretization satisfies $\ker(B_h)\subset\ker(B)$. See also Remark 53.22 for further insight in the context of the Stokes equations. □

Remark 50.9 (Stabilization). It is possible to approximate (50.1) using discrete spaces V_h and Q_h that violate the inf-sup condition (50.4b) by replacing the bilinear forms a and b by some stabilized versions a_h and b_h; see Chapters 62 and 63. □

We now establish an error estimate on u in a norm that is weaker than that in V. We do so by using a duality argument in the spirit of the Aubin–Nitsche lemma (Lemma 32.11).

Definition 50.10 (Smoothing property). *The problem* (50.1) *is said to have a* smoothing property *if there is a Hilbert space $H \hookrightarrow V$ with inner product $(\cdot,\cdot)_H$, two Banach spaces $Y \hookrightarrow V$ and $N \hookrightarrow Q$, and a constant c_{smo} such that the following adjoint problem:*

$$\begin{cases} \text{Find } \varphi(g) \in V \text{ and } \vartheta(g) \in Q \text{ such that} \\ a(v,\varphi(g)) + b(v,\vartheta(g)) = (g,v)_H, & \forall v \in V, \\ b(\varphi(g),q) = 0, & \forall q \in Q, \end{cases}$$

has a unique solution for all $g \in H$ and satisfies the a priori estimate $\|\varphi(g)\|_Y + \|\vartheta(g)\|_N \le c_{\mathrm{smo}}\|g\|_H$.

In addition to the smoothing property, we assume that the spaces H, Y, and N satisfy an additional approximation property, i.e., there are $s > 0$ and c such that the following holds true for all $(v,q) \in Y \times N$ and all $h \in \mathcal{H}$:

$$\inf_{(v_h,q_h)\in V_h\times Q_h} (\|v - v_h\|_V + \|q - q_h\|_Q) \le c\, h^s (\|v\|_Y + \|q\|_N). \tag{50.8}$$

Lemma 50.11 (Improved error estimate in weaker norm). *Let (u,p) solve* (50.1)*. Assume* (50.4) *and let (u_h,p_h) solve the discrete problem* (50.2)*. Assume that* (50.1) *has a smoothing property and that* (50.8) *holds true. Then we have*

$$\|u - u_h\|_H \le c\, h^s (\|u - u_h\|_V + \|p - p_h\|_Q),$$

where c is independent of (u,p), (u_h,p_h) and $h \in \mathcal{H}$.

Proof. Set $V := V \times Q$, $Z := Y \times N$, and $L := H \times Q$, each equipped with the product norm. Define the symmetric positive bilinear form $l((v,q),(w,r)) := (v,w)_H$ and the seminorm $|(v,q)|_L := \|v\|_H$. Apply Lemma 32.11 in the conforming setting with the bilinear form $t((u,p),(v,q)) := a(u,v) + b(v,p) + b(u,q)$ to conclude. □

50.2 Algebraic viewpoint

In this section, we study the linear system associated with the discrete problem (50.2) assuming that the well-posedness conditions (50.4a)-(50.4b) are satisfied. We also assume that the bilinear form a satisfies an inf-sup condition on $V_h \times V_h$. For simplicity, we consider real vector spaces.

50.2.1 The coupled linear system

Let $N := \dim(V_h)$ and $M := \dim(Q_h)$. Let $\{\varphi_i\}_{i\in\{1:N\}}$ be a basis for V_h and let $\{\psi_k\}_{k\in\{1:M\}}$ be a basis for Q_h. Recall that these bases consist of global shape functions when V_h and Q_h are finite element spaces. Proceeding as in §28.1.1, for every column vectors $\mathsf{U} = (\mathsf{U}_1,\dots,\mathsf{U}_N)^\mathsf{T}$ in $\mathbb{R}^N$

and $\mathsf{P} = (\mathsf{P}_1, \dots, \mathsf{P}_M)^\mathsf{T}$ in $\mathbb{R}^M$, we define the functions $\mathsf{R}_\varphi(\mathsf{U}) \in V_h$ and $\mathsf{R}_\psi(\mathsf{P}) \in Q_h$ by $\mathsf{R}_\varphi(\mathsf{U}) := \sum_{i\in\{1:N\}} \mathsf{U}_i\varphi_i$ and $\mathsf{R}_\psi(\mathsf{P}) := \sum_{k\in\{1:M\}} \mathsf{P}_k\psi_k$. The correspondences between $\mathsf{R}_\varphi(\mathsf{U})$ and U and between $\mathsf{R}_\psi(\mathsf{P})$ and P are one-to-one since $\{\varphi_i\}_{i\in\{1:N\}}$ and $\{\psi_k\}_{k\in\{1:M\}}$ are bases.

Inserting the expansions of $\mathsf{R}_\varphi(\mathsf{U})$ and $\mathsf{R}_\psi(\mathsf{P})$ into (50.2) and choosing the basis functions of V_h and Q_h to test (50.2), we obtain the linear system

$$\mathcal{C}\begin{pmatrix}\mathsf{U}\\ \mathsf{P}\end{pmatrix} = \begin{pmatrix}\mathsf{F}\\ \mathsf{G}\end{pmatrix}, \qquad \mathcal{C} := \begin{pmatrix}\mathcal{A} & \mathcal{B}^\mathsf{T}\\ \mathcal{B} & \mathbb{O}\end{pmatrix}, \tag{50.9}$$

where the matrices $\mathcal{A} \in \mathbb{R}^{N\times N}$ and $\mathcal{B} \in \mathbb{R}^{M\times N}$ are such that $\mathcal{A}_{ij} := a(\varphi_j, \varphi_i)$ and $\mathcal{B}_{ki} := b(\varphi_i, \psi_k)$ for all $i \in \{1:N\}$ and all $k \in \{1:M\}$, $\mathbb{O}$ is the zero matrix in $\mathbb{R}^{M\times M}$, and the vectors $\mathsf{F} \in \mathbb{R}^N$ and $\mathsf{G} \in \mathbb{R}^M$ are such that $\mathsf{F}_i = f(\varphi_i)$ and $\mathsf{G}_k = g(\psi_k)$ for all $i, j \in \{1:N\}$ and all $k \in \{1:M\}$.

The matrix $\mathcal{C}$ is invertible since (50.2) is well-posed owing to (50.4a)-(50.4b). Notice also that (50.4b) implies that $\mathcal{B}^\mathsf{T}$ has full column rank and $\mathcal{B}$ has full row rank (these ranks are equal to M). Moreover, $\mathcal{A}$ is invertible since we additionally assumed that a satisfies an inf-sup condition on $V_h\times V_h$. Algebraic counterparts of the boundedness and inf-sup conditions on the bilinear forms a and b can be established by using the dual norm

$$\|\mathsf{U}\|_{\ell^2_\varphi} := \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{\mathsf{U}^\mathsf{T}\mathsf{Y}}{\|\mathsf{R}_\varphi(\mathsf{Y})\|_V}, \qquad \forall \mathsf{U} \in \mathbb{R}^N. \tag{50.10}$$

Proposition 50.12 (Norm equivalence). *The following holds true:*

$$\alpha_h \|\mathsf{R}_\varphi(\mathsf{U})\|_V \le \|\mathcal{A}\mathsf{U}\|_{\ell^2_\varphi} \le \|a\| \|\mathsf{R}_\varphi(\mathsf{U})\|_V, \qquad \forall \mathsf{U} \in \mathbb{R}^N, \tag{50.11a}$$

$$\beta_h \|\mathsf{R}_\psi(\mathsf{P})\|_Q \le \|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2_\varphi} \le \|b\| \|\mathsf{R}_\psi(\mathsf{P})\|_Q, \qquad \forall \mathsf{P} \in \mathbb{R}^M. \tag{50.11b}$$

Proof. See Exercise 50.4. □

50.2.2 Schur complement

Since the matrix $\mathcal{A}$ is invertible, the vector U can be eliminated from the linear system (50.9) yielding

$$\mathcal{S}\mathsf{P} = \mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G}, \qquad \mathcal{S} := \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}. \tag{50.12}$$

Once P is known, U is obtained by solving $\mathcal{A}\mathsf{U} = F - \mathcal{B}^\mathsf{T}\mathsf{P}$. The matrix $\mathcal{S} \in \mathbb{R}^{M\times M}$ (up to a sign convention) is called *Schur complement* of $\mathcal{A}$; see §49.2.1 for the infinite-dimensional counterpart. Notice that the matrix $\mathcal{S}$ is invertible (if $\mathcal{S}\mathsf{P} = 0$, setting $\mathsf{U} := -\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P}$, we infer that $\mathcal{C}(\mathsf{U}, \mathsf{P})^\mathsf{T} = (0, 0)^\mathsf{T}$, and $\mathcal{C}$ being invertible, this implies that $\mathsf{U} = 0$ and $\mathsf{P} = 0$).

Additional properties of the Schur complement matrix $\mathcal{S}$ are available when the bilinear form a is symmetric and coercive, since in this case the matrix $\mathcal{A}$ is symmetric positive definite.

Proposition 50.13 (Symmetry and positivity of $\mathcal{S}$). *If $\mathcal{A}$ is symmetric positive definite, so is $\mathcal{S}$.*

Proof. The definition of $\mathcal{S}$ implies that $\mathcal{S}^\mathsf{T} = \mathcal{B}(\mathcal{A}^{-1})^\mathsf{T}\mathcal{B}^\mathsf{T}$, but $(\mathcal{A}^{-1})^\mathsf{T} = (\mathcal{A}^\mathsf{T})^{-1}$. Hence, $\mathcal{S}$ is symmetric if $\mathcal{A}$ is symmetric. Let now $\mathsf{P} \in \mathbb{R}^M$. Then $\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P} = \mathsf{P}^\mathsf{T}\mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P} = (\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P} \geq 0$. This proves that $\mathcal{S}$ is positive semidefinite. Moreover, $\mathcal{S}\mathsf{P} = 0$ implies that $\mathcal{B}^\mathsf{T}\mathsf{P} = 0$, so that $\mathsf{P} = 0$ since $\mathcal{B}^\mathsf{T}$ has full column rank. Hence, $\mathcal{S}$ is positive definite. □

Note that even if $\mathcal{A}$ is symmetric positive definite, the matrix $\mathcal{C}$ is symmetric but indefinite. Observing that

$$\mathcal{C} = \begin{pmatrix} \mathcal{I}_N & \mathbb{O}_{N,M} \\ \mathcal{B}\mathcal{A}^{-1} & \mathcal{I}_M \end{pmatrix} \begin{pmatrix} \mathcal{A} & \mathbb{O}_{N,M} \\ \mathbb{O}_{M,N} & -\mathcal{S} \end{pmatrix} \begin{pmatrix} \mathcal{I}_N & \mathcal{A}^{-1}\mathcal{B}^\mathsf{T} \\ \mathbb{O}_{N,M} & \mathcal{I}_M \end{pmatrix},$$

we infer from the Sylvester Law of Inertia (stating that two symmetric matrices $\mathcal{C}$ and $\mathcal{C}'$ satisfying $\mathcal{C} = \mathcal{P}\mathcal{C}'\mathcal{P}^\mathsf{T}$ with $\mathcal{P}$ invertible have the same number of positive, zero, and negative eigenvalues; see Golub and van Loan [218, p. 403]) that $\mathcal{C}$ has N positive eigenvalues and M negative ones. Upper and lower bounds on the clusters of positive and negative eigenvalues of $\mathcal{C}$ are derived in Rusten and Winther [338], Wathen and Silvester [390]. In practice, the matrix $\mathcal{C}$ can be very poorly conditioned. We return to this issue in §50.3.2. Note that changing the lower-left block of $\mathcal{C}$ into $-\mathcal{B}$ produces a positive semidefinite, but nonsymmetric, matrix.

Let us now examine more closely the eigenvalues of $\mathcal{S}$ (see Verfürth [375]). To this purpose, let $\mathcal{M}_Q \in \mathbb{R}^{M\times M}$ be the matrix with entries $\mathcal{M}_{Q,kl} := (\psi_k, \psi_l)_Q$ for all $k, l \in \{1{:}M\}$. The matrix $\mathcal{M}_Q$ is symmetric by construction, and the identity $\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} = (\mathsf{R}_\psi(\mathsf{P}), \mathsf{R}_\psi(\mathsf{P}))_Q = \|\mathsf{R}_\psi(\mathsf{P})\|_Q^2$ for all $\mathsf{P} \in \mathbb{R}^M$ shows that $\mathcal{M}_Q$ is positive definite. Since Q is the L^2-space in many applications, the matrix $\mathcal{M}_Q$ is called mass matrix (see §28.1.1). Let $\mu_{\min}$ and $\mu_{\max}$ be the lowest and largest eigenvalues of $\mathcal{M}_Q$. Recall from §28.2.1 that the (Euclidean) condition number $\kappa(\mathcal{Z})$ of a symmetric invertible matrix $\mathcal{Z}$ is the ratio of the largest to the smallest eigenvalues of $\mathcal{Z}$ in absolute value.

Proposition 50.14 (Spectrum of $\mathcal{S}$). *Assume that the bilinear form a is symmetric and coercive on V_h with constant α_h and that the inf-sup condition* (50.4b) *for b is satisfied with constant β_h. Then the matrices $\mathcal{S}$ and $\mathcal{M}_Q$ are spectrally equivalent, i.e., the following holds true for all $\mathsf{P} \in \mathbb{R}^M$:*

$$\frac{\beta_h^2}{\|a\|} \leq \frac{\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P}}{\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P}} \leq \frac{\|b\|^2}{\alpha_h}. \tag{50.13}$$

Moreover, $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}) \subset \left[\frac{\beta_h^2}{\|a\|}, \frac{\|b\|^2}{\alpha_h}\right]$, and $\sigma(\mathcal{S}) \subset \left[\mu_{\min}\frac{\beta_h^2}{\|a\|}, \mu_{\max}\frac{\|b\|^2}{\alpha_h}\right]$, which implies that $\kappa(\mathcal{M}_Q^{-1}\mathcal{S}) \leq \frac{\|a\|}{\alpha_h}\left(\frac{\|b\|}{\beta_h}\right)^2$ and $\kappa(\mathcal{S}) \leq \frac{\|a\|}{\alpha_h}\left(\frac{\|b\|}{\beta_h}\right)^2 \kappa(\mathcal{M}_Q)$.

Proof. (1) Proof of (50.13). For all $\mathsf{P} \in \mathbb{R}^M$, we observe that

$$\sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathsf{Y}}{(\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y})^{\frac12}} = \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathcal{A}^{-\frac12}\mathsf{Y}}{\|\mathsf{Y}\|_{\ell^2(\mathbb{R}^N)}} = \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{A}^{-\frac12}\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathsf{Y}}{\|\mathsf{Y}\|_{\ell^2(\mathbb{R}^N)}}$$
$$= \|\mathcal{A}^{-\frac12}\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2(\mathbb{R}^N)} = (\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P})^{\frac12},$$

since $\mathcal{A}$ is symmetric positive definite. Observing that $\frac{1}{\|a\|} \le \frac{\|\mathsf{R}_\varphi(\mathsf{Y})\|_\mathsf{Y}^2}{\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y}} \le \frac{1}{\alpha_h}$ for all $\mathsf{Y} \in \mathbb{R}^N$, we infer that

$$\frac{1}{\|a\|}\|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2_\varphi}^2 \le \mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P} \le \frac{1}{\alpha_h}\|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2_\varphi}^2.$$

Finally, (50.13) follows from (50.11b) using $\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} = \|\mathsf{R}_\psi(\mathsf{P})\|_Q^2$.
(2) The spectrum and condition number for $\mathcal{M}_Q^{-1}\mathcal{S}$ readily follow from (50.13), and the results for $\mathcal{S}$ follow from the fact that $\mu_{\min}\|\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}^2 \le \mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} \le \mu_{\max}\|\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}^2$ for all $\mathsf{P} \in \mathbb{R}^M$. □

50.2.3 Augmented Lagrangian for saddle point problems

Assume that the matrix $\mathcal{A}$ is symmetric positive definite and that $\mathcal{B}^\mathsf{T}$ has full column rank. Referring to §49.3.2 for the infinite-dimensional setting we infer that the pair (U,P) solves the linear system (50.9) iff it is a saddle point of the Lagrangian

$$\mathcal{L}(\mathsf{Y},\mathsf{R}) := \frac12\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y} - \mathsf{F}^\mathsf{T}\mathsf{Y} + \mathsf{R}^\mathsf{T}(\mathcal{B}\mathsf{Y} - \mathsf{G}). \tag{50.14}$$

Recall that

$$\inf_{\mathsf{Y}\in\mathbb{R}^N}\sup_{\mathsf{R}\in\mathbb{R}^M} \mathcal{L}(\mathsf{Y},\mathsf{R}) = \mathcal{L}(\mathsf{U},\mathsf{P}) = \sup_{\mathsf{R}\in\mathbb{R}^M}\inf_{\mathsf{Y}\in\mathbb{R}^N} \mathcal{L}(\mathsf{Y},\mathsf{R}). \tag{50.15}$$

The optimization problem on the left-hand side of (50.15) amounts to minimizing the convex energy functional $\mathfrak{E}(\mathsf{Y}) := \frac12\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y} - \mathsf{F}^\mathsf{T}\mathsf{Y}$ over the affine subspace $\{\mathsf{Y} \in \mathbb{R}^N \mid B\mathsf{Y} = \mathsf{G}\}$ since $\sup_{\mathsf{R}\in\mathbb{R}^M}\mathcal{L}(\mathsf{Y},\mathsf{R}) = \infty$ if $B\mathsf{Y} \neq \mathsf{G}$. Consider now the optimization problem on the right-hand side of (50.15). The minimization of $\mathcal{L}(\mathsf{Y},\mathsf{R})$ over $\mathsf{Y} \in \mathbb{R}^N$ leads to the optimal solution $\mathsf{Y}_* := \mathcal{A}^{-1}(\mathsf{F} - \mathcal{B}^\mathsf{T}\mathsf{R})$, and we are left with maximizing the following concave functional over $\mathbb{R}^M$:

$$\mathsf{R} \mapsto \mathcal{L}(\mathsf{Y}_*,\mathsf{R}) = -\frac12\mathsf{R}^\mathsf{T}\mathcal{S}\mathsf{R} + (\mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G})^\mathsf{T}\mathsf{R} - \frac12\mathsf{F}^\mathsf{T}\mathcal{A}^{-1}\mathsf{F},$$

where $\mathcal{S} := \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}$. The optimal solution to this maximization problem solves $\mathcal{S}\mathsf{R} = \mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G}$, i.e., we recover the Schur complement system (50.12).

The main idea of augmented Lagrangian methods (see Fortin and Glowinski [203]) is to add to the Lagrangian a least-squares penalty on the constraint. Specifically, letting $\rho > 0$ be a real parameter and recalling the mass matrix $\mathcal{M}_Q \in \mathbb{R}^{M\times M}$, the *augmented Lagrangian* is defined as

$$\mathcal{L}_\rho(\mathsf{Y},\mathsf{R}) := \mathcal{L}(\mathsf{Y},\mathsf{R}) + \frac{\rho}{2}(\mathcal{B}\mathsf{Y}-\mathsf{G})^\mathsf{T}\mathcal{M}_Q^{-1}(\mathcal{B}\mathsf{Y}-\mathsf{G}).$$

Since we also have $B\mathsf{U} = \mathsf{G}$, the solution to (50.9) is also the unique saddle point of the augmented Lagrangian $\mathcal{L}_\rho$, i.e., (U,P) can be found by solving the following linear system:

$$\begin{pmatrix}\mathcal{A}_\rho & \mathcal{B}^\mathsf{T}\\ \mathcal{B} & \mathbb{O}\end{pmatrix}\begin{pmatrix}\mathsf{U}\\ \mathsf{P}\end{pmatrix} = \begin{pmatrix}\mathsf{F}_\rho\\ \mathsf{G}\end{pmatrix}, \quad \begin{aligned}\mathcal{A}_\rho &:= \mathcal{A} + \rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B},\\ \mathsf{F}_\rho &:= \mathsf{F} + \rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathsf{G}.\end{aligned} \tag{50.16}$$

The augmented Schur complement is defined as $\mathcal{S}_\rho := \mathcal{B}\mathcal{A}_\rho^{-1}\mathcal{B}^\mathsf{T}$. Recall that $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}) \subset [s_\flat, s_\sharp]$, with $s_\flat := \frac{\beta_h^2}{\|a\|}$ and $s_\sharp := \frac{\|b\|^2}{\alpha_h}$.

Proposition 50.15 (Spectrum of $\mathcal{S}_\rho$). *The following holds true:*

$$\mathcal{S}_\rho^{-1} = \rho\mathcal{M}_Q^{-1} + \mathcal{S}^{-1}, \tag{50.17}$$

and $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}_\rho) \subset [(\rho + s_\flat^{-1})^{-1}, (\rho + s_\sharp^{-1})^{-1}]$ *and* $\kappa(\mathcal{M}_Q^{-1}\mathcal{S}_\rho) \le \frac{\rho+s_\flat^{-1}}{\rho+s_\sharp^{-1}}$.

Proof. See Exercise 50.5 for the proof of (50.17). The properties on the spectrum and the condition number of $\mathcal{S}_\rho$ follow readily. □

Remark 50.16 (Value of ρ). Proposition 50.15 shows that taking $\rho \gg 1$ improves the condition number of the Schur complement $\mathcal{S}_\rho$. A large value of ρ however deteriorates the conditioning of the matrix $\mathcal{A}_\rho$ which makes it more difficult to invert iteratively. In practice, it is necessary to strike a balance between these two criteria. □

Remark 50.17 (Hilbert setting). The notion of augmented Lagrangian can be extended to the infinite-dimensional setting. The mass matrix $\mathcal{M}_Q$ is then replaced by the Riesz–Fréchet isomorphism $J_Q : Q \to Q'$. □

The augmented Lagrangian technique is in general preferable to the following unconstrained penalty method:

$$\begin{pmatrix}\mathcal{A} & \mathcal{B}^\mathsf{T}\\ \mathcal{B} & -\epsilon\mathcal{M}_Q\end{pmatrix}\begin{pmatrix}\mathsf{U}_\epsilon\\ \mathsf{P}_\epsilon\end{pmatrix} = \begin{pmatrix}\mathsf{F}\\ \mathsf{G}\end{pmatrix}, \tag{50.18}$$

where $\epsilon > 0$ is a small parameter. This technique is often referred to as artificial compressibility in the fluid mechanics literature. Eliminating P_ϵ from the first equation yields

$$\mathcal{A}_{\frac{1}{\epsilon}}\mathsf{U}_\epsilon = \mathsf{F} + \epsilon^{-1}\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathsf{G}. \tag{50.19}$$

The advantage of (50.19) with respect to (50.9) (or to (50.16)) is that the system matrix (i.e., $\mathcal{A}_{\frac{1}{\epsilon}}$) is symmetric positive definite. The solution $(\mathsf{U}_\epsilon, \mathsf{P}_\epsilon)$ however differs from (U, P). In particular, U fails to satisfy the constraint $\mathcal{B}\mathsf{U} = \mathsf{G}$, although the difference $(\mathsf{U} - \mathsf{U}_\epsilon, \mathsf{P} - \mathsf{P}_\epsilon)$ tends to zero as $\epsilon \to 0$. Unfortunately, taking $\epsilon \ll 1$ makes the linear system (50.19) ill-conditioned.

Proposition 50.18 (Penalty). *Let $\epsilon > 0$. Let (U, P) solve (50.9) and $(\mathsf{U}_\epsilon, \mathsf{P}_\epsilon)$ solve (50.18). The following holds true:*

$$\tfrac{\alpha_h \beta_h}{\|a\|} \|\mathsf{R}_\varphi(\mathsf{U} - \mathsf{U}_\epsilon)\|_V + \tfrac{\alpha_h \beta_h^2}{\|a\|^2} \|\mathsf{R}_\psi(\mathsf{P} - \mathsf{P}_\epsilon)\|_Q \le \epsilon \|\mathsf{R}_\psi(\mathsf{P})\|_Q. \tag{50.20}$$

Proof. See Exercise 50.6. □

50.3 Iterative solvers

In this section, we discuss iterative solvers for the linear system (50.9) (or its augmented Lagrangian version (50.16)). First, we discuss the Uzawa algorithm as an example of a technique based on stationary iterations. Then, we present more efficient techniques based on preconditioned Krylov subspaces. We assume that the matrix $\mathcal{A}$ is invertible and that the matrix $\mathcal{B}^\mathsf{T}$ has full column rank.

50.3.1 Uzawa algorithm

The Uzawa algorithm is an iterative method where U and P are updated one after the other. Given $\mathsf{P}_0 \in \mathbb{R}^M$ and a parameter $\eta > 0$, the algorithm consists of constructing the iterates $(\mathsf{U}_m, \mathsf{P}_m)$ for $m = 1, 2, \ldots$ as follows:

$$\mathcal{A}\mathsf{U}_m = \mathsf{F} - \mathcal{B}^\mathsf{T}\mathsf{P}_{m-1}, \tag{50.21a}$$

$$\mathcal{M}_Q\mathsf{P}_m = \mathcal{M}_Q\mathsf{P}_{m-1} + \eta(\mathcal{B}\mathsf{U}_m - \mathsf{G}). \tag{50.21b}$$

This makes sense since $\mathcal{A}$ and $\mathcal{M}_Q$ are invertible. Eliminating U_m gives

$$\mathcal{M}_Q\mathsf{P}_m = \mathcal{M}_Q\mathsf{P}_{m-1} - \eta(\mathcal{S}\mathsf{P}_{m-1} - \mathcal{B}\mathcal{A}^{-1}\mathsf{F} + \mathsf{G}). \tag{50.22}$$

In other words, the Uzawa algorithm is equivalent to the Richardson iteration applied to the linear system (50.12) left-preconditioned by the mass matrix $\mathcal{M}_Q$. (Recall that for a generic linear system $\mathcal{Z}\mathsf{X} = \mathsf{Y}$, the Richardson iteration reads $\mathsf{X}_m = \mathsf{X}_{m-1} + \eta(\mathcal{Z}\mathsf{X}_{m-1} - \mathsf{Y})$.) If $\mathcal{A}$ is symmetric positive definite, we can use the bounds on the spectrum of $\mathcal{M}_Q^{-1}\mathcal{S}$ from Proposition 50.14, that is, $s_\flat := \frac{\beta_h^2}{\|a\|} \le \mathcal{M}_Q^{-1}\mathcal{S} \le \frac{\|b\|^2}{\alpha_h} := s_\sharp$ in the sense of quadratic forms. We then infer that the Richardson iteration (50.22) converges geometrically provided we take $0 < \eta < \frac{2}{s_\sharp}$, and the error reduction factor is maximized

by taking the optimal value $\eta_{\mathrm{opt}} := \frac{2}{s_\flat + s_\sharp}$; see Saad [339, p. 106]. It is often easier to estimate $s_\sharp$ than $s_\flat$ since β_h is more difficult to estimate than α_h.

Remark 50.19 (Implementation). The matrices $\mathcal{A}$ and $\mathcal{B}$ are sparse (see §29.1), but $\mathcal{S}$ is a dense matrix owing to the presence of $\mathcal{A}^{-1}$ in the definition of $\mathcal{S}$. Since precomputing $\mathcal{A}^{-1}$ is generally too expensive, an inner iteration has to be employed to compute the action of $\mathcal{A}^{-1}$ on vectors in $\mathbb{R}^N$. The matrix $\mathcal{B}$ can be assembled and stored once and for all, or its action on a given vector in $\mathbb{R}^M$ can be computed on the fly whenever needed. In practice, one must often find a compromise between many (often conflicting) criteria: the memory space available; the number of times the linear system has to be solved; the ratio between the speed to access memory and the speed to perform floating point operations; parallelization; etc. □

Remark 50.20 (Variants). The mass matrix $\mathcal{M}_Q$ can be replaced by the identity matrix $\mathcal{I}_M$ in (50.21b). The advantage is that this avoids computing the inverse of the mass matrix (although this matrix is generally easy to invert since it is well-conditioned). The drawback is that the choice of the relaxation parameter η now depends on the spectrum of the unpreconditioned Schur complement matrix, which requires some information on the (mesh-dependent) spectrum of $\mathcal{M}_Q$. Another variant is to consider an approximate inverse of $\mathcal{A}$ that is easy to compute, say $\mathcal{H}$, and to replace (50.21a) by $\mathsf{U}_m = \mathsf{U}_{m-1} + \mathcal{H}(\mathsf{F} - \mathcal{A}\mathsf{U}_{m-1} - \mathcal{B}^\mathsf{T}\mathsf{P}_{m-1})$ leading to an inexact Uzawa algorithm; see Bacuta [41] for a convergence analysis. □

50.3.2 Krylov subspace methods

Krylov subspace methods for solving (preconditioned) linear systems of the form (50.9) or variations thereof constitute an active area of research. In this section, we sketch a few important ideas and refer to Benzi et al. [52, §9] for a broader treatment and to Elman et al. [185, Chap. 6&8], Turek [366] for applications to fluid mechanics.

In the context of saddle point problems, the matrix $\mathcal{C}$ in (50.9) is symmetric, but indefinite (recall that the matrix $\mathcal{A}$ is symmetric positive definite by assumption). In this case, MINRES is a method of choice to solve (50.9); see [185, p. 289]. The attractive feature of MINRES is that it achieves an optimality property on the residual while employing only short-term recurrences. Specifically, at step $m \geq 1$, the iterate $\mathsf{X}_m \in \mathbb{R}^{N+M}$ with residual $\mathsf{R}_m := (\mathsf{F}, \mathsf{G})^\mathsf{T} - \mathcal{C}\mathsf{X}_m$ satisfies the following optimality property (compare with Proposition 28.20 for the conjugate gradient method applied to symmetric positive definite linear systems):

$$\|\mathsf{R}_m\|_{\ell^2(\mathbb{R}^{N+M})} = \min_{\mathsf{Y} \in \mathsf{U}_0 + K_m} \|(\mathsf{F}, \mathsf{G})^\mathsf{T} - \mathcal{C}\mathsf{Y}\|_{\ell^2(\mathbb{R}^{N+M})}, \tag{50.23}$$

with the Krylov subspace $K_m := \mathrm{span}\{\mathsf{R}_0, \mathcal{C}\mathsf{R}_0, \dots, \mathcal{C}^{m-1}\mathsf{R}_0\}$. The convergence rate of MINRES depends on the spectrum of $\mathcal{C}$. More precisely, defining

$\tilde{c}_m := \min_{p\in\mathbb{P}_m, p(0)=1} \max_{\lambda\in\sigma(\mathcal{C})} |p(\lambda)|$, one can prove that $\|\mathsf{R}_m\|_{\ell^2(\mathbb{R}^{N+M})} \le \tilde{c}_m \|\mathsf{R}_0\|_{\ell^2(\mathbb{R}^{N+M})}$ (this bound is sharp). The constant $\tilde{c}_m$ can be estimated under the assumption that $\sigma(\mathcal{C}) \subset [-a,-b]\cup[c,d]$ with positive real numbers a,b,c,d such that the two intervals have the same length (i.e., $d-c=a-b$). One can show that (see Greenbaum [221, Chap. 3])

$$\|\mathsf{R}_{2m}\|_{\ell^2(\mathbb{R}^{N+M})} \le 2\left(\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)^m \|\mathsf{R}_0\|_{\ell^2(\mathbb{R}^{N+M})}. \tag{50.24}$$

The minimization property of Minres implies that $\|\mathsf{R}_{2m+1}\|_{\ell^2(\mathbb{R}^{N+M})} \le \|\mathsf{R}_{2m}\|_{\ell^2(\mathbb{R}^{N+M})}$, but it is possible that no reduction of the norm of the residual occurs in every other step, leading to a staircasing behavior of the iterates. A comparison of (28.23) with (50.24) shows that Minres requires twice as many iterations as the Conjugate Gradient to reach a given threshold for a symmetric positive definite matrix with condition number κ^2. Hence, solving linear systems like (50.9) is a significant computational challenge, and preconditioning is essential. Before addressing this question let us observe that Minres is bound to fail if $\mathcal{A}$ is not symmetric, since symmetry is essential for Minres to work properly. This happens, for instance, in fluid mechanics when solving the Oseen or (linearized) Navier–Stokes equations. One alternative is to use the Gmres method which retains an optimality property over the Krylov subspace at the price of storing a complete basis thereof; see Saad [339, §6.5] for a thorough description.

Preconditioning is a very important ingredient of Krylov subspace methods, especially for linear systems of the form (50.9). Here, we only discuss block preconditioners and refer the reader to [52, §10] and references therein for further insight into preconditioned Krylov methods. Block diagonal and triangular preconditioners are, respectively, of the form

$$\mathcal{P}_{\mathrm{d}} := \begin{pmatrix} \hat{\mathcal{A}} & \mathbb{O} \\ \mathbb{O} & \hat{\mathcal{S}} \end{pmatrix}, \qquad \mathcal{P}_{\mathrm{t}} := \begin{pmatrix} \hat{\mathcal{A}} & \mathcal{B}^{\mathsf{T}} \\ \mathbb{O} & \hat{\mathcal{S}} \end{pmatrix}, \tag{50.25}$$

where $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ are easy-to-invert approximations of $\mathcal{A}$ and $\mathcal{S}$. In the ideal case where $\hat{\mathcal{A}} := \mathcal{A}$ and $\hat{\mathcal{S}} := \mathcal{S}$, a direct calculation shows that the left-preconditioned matrices $\mathcal{P}_{\mathrm{d}}^{-1}\mathcal{C}$ and $\mathcal{P}_{\mathrm{t}}^{-1}\mathcal{C}$ are zeroes of the polynomials $p_{\mathrm{d}}(\lambda) := (\lambda-1)(\lambda-\frac{1\pm\sqrt{5}}{2})$ and $p_{\mathrm{t}}(\lambda) := \lambda^2-1$, respectively (see Kuznetsov [272], Murphy et al. [309]; see also (49.22) and Theorem 49.6), implying convergence in at most three (resp., two) steps for every preconditioned Krylov subspace method. The block triangular preconditioner $\mathcal{P}_{\mathrm{t}}$ breaks the symmetry of the system even if $\mathcal{A}$ is symmetric, but this preconditioner is still quite effective in many cases, particularly for Oseen and Navier–Stokes flows (where the convective term breaks the symmetry of $\mathcal{A}$ anyway). Note also that the costs of the two preconditioners in (50.25) are essentially identical since the cost of the additional multiplication by $\mathcal{B}^{\mathsf{T}}$ is often marginal.

Effective choices for $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ are often driven by the application at hand. For Darcy's and Maxwell's equations (see Examples in §49.1.1 and §49.1.3), $\mathcal{A}$ represents a zeroth-order differential operator (multiplication by a material property), and choosing a diagonal lumping for $\hat{\mathcal{A}}$ together with some multilevel technique for $\hat{\mathcal{S}}$ often works well if the material coefficients are smooth (see, e.g., Perugia and Simoncini [324] for magnetostatics problems). For the Stokes equations (see §49.1.2), $\mathcal{A}$ represents a second-order differential operator, and the preconditioner $\hat{\mathcal{A}}$ is typically based on some multilevel technique. The mass matrix associated with p can be used for $\hat{\mathcal{S}}$ and a detailed eigenvalue analysis of the resulting block-diagonal preconditioned system can be found in Silvester and Wathen [347]. The approximation of the Schur complement becomes more delicate in the unsteady case and in the presence of convection. Preconditioners devised from the structure of the steady Navier–Stokes equations can be found in Elman et al. [185, Chap. 8] and the references therein. Furthermore, an attractive idea for transient and high-Reynolds number flows is to consider a block triangular preconditioner based on the augmented Lagrangian formulation (50.16) for the $(1,1)$-block, together with the (scaled) mass matrix for the $(2,2)$-block (thereby avoiding to consider the Schur complement); see Benzi and Olshanskii [51], Benzi et al. [53].

Exercises

Exercise 50.1 (Algebraic setting). Let $\mathcal{A} := \left(\begin{smallmatrix} 1 & \sqrt{2} \\ \sqrt{2} & 0 \end{smallmatrix}\right)$ and $\mathcal{B} := (1,0)^\mathsf{T}$. Show that

$$\inf_{\mathsf{V}\in\ker(\mathcal{B})}\sup_{\mathsf{W}\in\ker(\mathcal{B})}\frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}} < \inf_{\mathsf{V}\in\mathbb{R}^2}\sup_{\mathsf{W}\in\mathbb{R}^2}\frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}}.$$

(*Hint*: one number is equal to 0 and the other is equal to 1.)

Exercise 50.2 (Saddle point problem). Let V, Q be Hilbert spaces and let a be a symmetric, coercive, bilinear form. Consider the discrete problem (50.2) and the bilinear form $t(y,z) := a(v,w) + b(w,q) + b(v,r)$ for all $y := (v,q), z := (w,r) \in X := V\times Q$. Let $X_h := V_h\times Q_h$ and consider the linear map $P_h \in \mathcal{L}(X;X_h)$ such that for all $x \in X$, $P_h(x) \in X_h$ is the unique solution of $t(P_h(x), y_h) = t(x, y_h)$ for all $y_h \in X_h$. Equip X and X_h with the norm $\|(v,q)\|_{\tilde{X}} := (\|v\|_a^2 + \|q\|_Q^2)^{\frac12}$ with $\|v\|_a^2 := a(v,v)$. (i) Prove that $\|P_h\|_{\mathcal{L}(X;X)} \le \tilde{c}_h := \frac{(4\frac{\|b\|^2}{\alpha}+1)^{\frac12}+1}{\frac{\beta_h^2}{\|a\|}(4\frac{\|b\|^2}{\alpha}+1)^{\frac12}-1}$. (*Hint*: use Proposition 49.8.) (ii) Prove that $\|u-u_h\|_a^2 + \|p-p_h\|_Q^2 \le \tilde{c}_h^2(\inf_{v_h\in V_h}\|u-u_h\|_a^2 + \inf_{q_h\in Q_h}\|p-q_h\|_Q^2)$. (*Hint*: see the proof of Theorem 5.14.)

Exercise 50.3 (Error estimate). (i) Prove directly the estimate (50.7a) with c'_{1h} replaced by $c''_{1h} := (1+\frac{\|a\|}{\alpha_h})(1+\frac{\|b\|}{\beta_h})$. (*Hint*: consider $z_h \in V_h$ s.t. $B_h(z_h) := B_h(u_h - v_h)$ with $v_h \in V_h$ arbitrary.) (ii) Assume that V is a Hilbert space, $\ker(B_h) \subset \ker(B)$, and $g := 0$. Prove that $\|u-u_h\|_V \le \frac{\|a\|}{\alpha_h} \inf_{v_h\in\ker(B_h)} \|u-v_h\|_V$.

Exercise 50.4 (Bound on $\mathcal{A}$ and $\mathcal{B}$). (i) Prove Proposition 50.12. (*Hint*: observe that $(\mathcal{A}\mathsf{U})^\mathsf{T}\mathsf{Y} = a(\mathsf{R}_\varphi(\mathsf{U}),\mathsf{R}_\varphi(\mathsf{Y}))$.) (ii) Let $\mathcal{J}_V \in \mathbb{R}^{N\times N}$ be the symmetric positive definite matrix with entries $\mathcal{J}_{V,ij} := (\varphi_i,\varphi_j)_X$ for all $i,j\in\{1{:}N\}$. Let $\|\cdot\|_{\ell^2(\mathbb{R}^N)}$ denote the Euclidean norm in $\mathbb{R}^N$. Verify that $\|\mathsf{R}_\varphi(\mathsf{U})\|_V = \|\mathcal{J}_V^{\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ and $\|\mathsf{U}\|_{\ell^2(\mathbb{R}^N)} = \|\mathcal{J}_V^{-\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ for all $\mathsf{U}\in\mathbb{R}^N$.

Exercise 50.5 ($\mathcal{S}_\rho$). The goal is to prove the identity (50.17). (i) Verify that $\mathcal{A}_\rho^{-1} = \mathcal{A}^{-1} - \rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}$. (*Hint*: multiply the right-hand side by $\mathcal{A}_\rho$ and develop the product.) (ii) Infer that $\mathcal{S}_\rho = \mathcal{S} - \rho\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S}$. (iii) Conclude. (*Hint*: multiply the right-hand side by $\rho\mathcal{M}_Q^{-1}+\mathcal{S}^{-1}$.)

Exercise 50.6 (Penalty). (i) Prove Proposition 50.18. (*Hint*: verify that $\mathcal{C}(\mathsf{U}-\mathsf{U}_\epsilon,\mathsf{P}-\mathsf{P}_\epsilon)^\mathsf{T} = (0,-\epsilon\mathcal{M}_Q\mathsf{P}_\epsilon)^\mathsf{T}$ and use Proposition 50.12.) (ii) Replace the mass matrix $\mathcal{M}_Q$ by the identity matrix $\mathcal{I}_M$ times a positive coefficient λ in (50.18). Does the method still converge? Is there any interest of doing so? Can you think of another choice?

Exercise 50.7 (Inexact Minres and DPG). Let V,Y be Hilbert spaces and $B\in\mathcal{L}(V;Y')$ be s.t. $\beta\|v\|_V \le \|B(v)\|_{Y'} \le \|b\|\|v\|_V$ for all $v\in V$ with $0<\beta\le\|b\|<\infty$. Set $b(v,y) := \langle B(v),y\rangle_{Y',Y}$. Let $f\in Y'$. Let $J_Y : Y\to Y'$ denote the isometric Riesz–Fréchet isomorphism. (i) Show that the Minres problem $\min_{v\in V}\|f-B(v)\|_{Y'}$ has a unique solution $u\in V$. (*Hint*: introduce the sesquilinear form $a(v,w) := \langle B(v), J_Y^{-1}(B(w))\rangle_{Y',Y}$ and invoke the Lax–Milgram Lemma.) (ii) Let $\{V_h\subset V\}_{h\in\mathcal{H}}$ and $\{Y_h\subset Y\}_{h\in\mathcal{H}}$ be sequences of subspaces approximating V and Y, respectively. Assume that there is $\beta_0>0$ s.t. for all $h\in\mathcal{H}$,

$$\inf_{v_h\in V_h}\sup_{y_h\in Y_h}\frac{|b(v_h,y_h)|}{\|v_h\|_V\|y_h\|_Y} \ge \beta_0. \tag{50.26}$$

Let $I_h : Y_h\to Y$ be the canonical injection and $I_h^* : Y'\to Y_h'$. Show that the inexact Minres problem $\min_{v_h\in V_h}\|I_h^*(f-B(v_h))\|_{Y_h'}$ has a unique solution $u_h\in V_h$. (*Hint*: introduce the residual representative $r_h := J_{Y_h}^{-1}I_h^*(f-B(u_h)) \in V_h$ and show that the pair $(u_h,r_h)\in V_h\times Y_h$ solves a saddle point problem.) (iii) Show that the residual representative $r_h\in Y_h$ is the unique solution of the following constrained minimization problem: $\min_{z_h\in Y_h\cap(I_h^*(B(V_h)))^\perp}\frac12\|z_h\|_Y^2 - \langle I_h^*(f),z_h\rangle_{Y_h',Y_h}$. (*Hint*: see Proposition 49.11.) (iv) Assume now that $f\in\mathrm{im}(B)$ so that $B(u)=f$. Prove that there is c s.t. $\|u-u_h\|_V \le c\inf_{w_h\in V_h}\|u-w_h\|_V$ for all $h\in\mathcal{H}$. (*Hint*: use a Fortin operator.) *Note*: since $\beta\|v_h\|_V\le\|B(v_h)\|_{Y'}$ for all $v_h\in V_h$, it is natural to expect that the inf-sup condition (50.26) is satisfied if the subspace $Y_h\subset Y$ is chosen rich

enough. The inexact residual minimization in a discrete dual norm is at the heart of the discontinuous Petrov–Galerkin (dPG) method; see Demkowicz and Gopalakrishnan [158], Gopalakrishnan and Qiu [219], Carstensen et al. [111]. The extension to reflexive Banach spaces is studied in Muga and van der Zee [308].