Research Article

Carsten Carstensen*

A Note on the Quasi-Best Approximation Constant

https://doi.org/10.1515/cmam-2025-2025
Received February 11, 2025; revised April 9, 2025; accepted May 7, 2025

Abstract: The solution operator in a Petrov Galerkin scheme in Hilbert spaces is an oblique projection with quasi-best approximation property. The latter estimate involves a multiplicative constant and the best-possible of those is the target of the note: We present a new direct proof of the formula of the quasi-best approximation constant and avoid the direct application of the Kato lemma. In fact, our characterisation leads to another proof of the Kato oblique projection lemma. The abstract result in Hilbert spaces is embedded in the setting of the best-approximation of conforming Petrov Galerkin schemes with a rich history that eventually led to the Tantardini–Veeser formula. A final application discusses the classical nonconforming schemes with a smoother in this framework.

Keywords: Quasi-Best Approximation Constant, Petrov–Galerkin, Kato Lemma, Tantardini–Veeser Formula

MSC 2020: 65N12, 65N15, 65N30

1 Quasi-Best Approximation by an Oblique Projection in a Hilbert Space

This section presents a new short direct proof that the best possible constant in the quasi-best approximation by an oblique projection is its norm.

Lemma 1 (Constant in Quasi-Best Approximation). *Let $P \in L(H) \setminus \{0, 1\}$ be an oblique projection in the (real or complex) Hilbert space H different from zero and identity. Then the H-orthogonal projection Π_S onto the closed range $S = \mathcal{R}(P)$ of P satisfies*

$$\|P\|_{L(H)} = \sup_{t \in H \setminus S} \frac{\|t - Pt\|}{\|t - \Pi_S t\|}. \tag{1.1}$$

Proof. Since $S := \mathcal{R}(P) = \mathcal{N}(1 - P)$ is the kernel of $1 - P$ and therefore closed, the H-orthogonal projection $\Pi_S \in L(H)$ onto the closed range $S = \mathcal{R}(P)$ of P is well defined. Since $Ps = s$ for any $s \in S \neq \{0\}$ and P is linear, the operator norm reads

$$1 \le \|P\|_{L(H)} = \sup_{\substack{s \in S\\ t \in H\\ s+t \neq 0}} \frac{\|s + Pt\|}{\|s + t\|} = \sup_{\substack{t \in H\\ r \in S\\ r \neq Pt - t}} \frac{\|r\|}{\|r + t - Pt\|} = \sup_{t \in H \setminus S}\, \sup_{r \in S} \frac{\|r\|}{\|r + t - Pt\|}$$

with an elementary transformation $r = s + Pt \in S$ and a careful exclusion of $t \in S$ (then last quotient is one) in the last steps. Given any $t \in H \setminus S$ and $r \in S$ with norm $\varrho := \|r\| \ge 0$, the term $\|r + t - Pt\|^2 = \varrho^2 + 2\Re\langle r, \Pi_S t - Pt\rangle_H + \|t - Pt\|^2$ has a minimum $\varrho^2 - 2\varrho\|\Pi_S t - Pt\| + \|t - Pt\|^2$ amongst all $r \in S$ with fixed norm $\|r\| = \varrho$ (attained at $r = -\varrho(\Pi_S t - Pt)/\|\Pi_S t - Pt\|$ for $\|\Pi_S t - Pt\| > 0$). Consequently,

$$\|P\|^2_{L(H)} = \sup_{t \in H \setminus S}\, \sup_{\varrho \ge 0} \frac{\varrho^2}{\varrho^2 - 2\varrho\|\Pi_S t - Pt\| + \|t - Pt\|^2} = \sup_{t \in H \setminus S} \frac{\|t - Pt\|^2}{\|t - Pt\|^2 - \|\Pi_S t - Pt\|^2}$$

***Corresponding author: Carsten Carstensen,** Department of Mathematics, Humboldt-Universität zu Berlin, Berlin, Germany, e-mail: cc@math.hu-berlin.de

with a straightforward computation of the minimum $1 - 2\|\Pi_S t - Pt\|^2/\|t - Pt\|^2$ of $\|t - Pt\|^2/\varrho^2 - 2\|\Pi_S t - Pt\|/\varrho + 1$ amongst all $\varrho > 0$ in the last step. This and the Pythagoras identity $\|t - Pt\|^2 = \|t - \Pi_S t\|^2 + \|\Pi_S t - Pt\|^2$ conclude the proof. □

A typical proof of the quasi-best approximation (1.1) is based on the Kato lemma $\|P\|_{L(H)} = \|1 - P\|_{L(H)}$. We refer to [3, Section 26.3.4] for details, references, and historical remarks. The direct proof of (1.1) does not use explicitly the Kato lemma, but immediately implies it. This is yet another proof amongst a longer list [4].

Corollary 2 (Kato Oblique Projection Lemma). *Under the hypothesis of Lemma 1,* $\|P\|_{L(H)} = \|1 - P\|_{L(H)}$.

Proof. Since $\|t - Pt\| = \|(1 - P)(t - \Pi_S t)\| \le \|1 - P\|_{L(H)}\|t - \Pi_S t\|$ for all $t \in H$, Lemma 1 provides that $\|P\|_{L(H)} \le \|1 - P\|_{L(H)}$. One may interchange the roles of P and $1 - P$ to understand that the reverse inequality $\|1 - P\|_{L(H)} \le \|P\|_{L(H)}$ also follows. □

2 Oblique Projections from Petrov–Galerkin Methods

Petrov–Galerkin schemes give rise an oblique projection $P \in L(H) \setminus \{0, 1\}$ in the (real or complex) Hilbert space H different from zero and identity, but the converse holds as well. This motivates the relevance of quasi-best approximation (1.1).

Suppose throughout this section that the oblique projection $P \in L(H) \setminus \{0, 1\}$ has finite range, i.e., the range $S := \mathcal{R}(P) := \{Px : x \in H\}$ is finite-dimensional, say $N := \dim S \in \mathbb{N}$. (The orthogonality notation $\perp$ and the orthogonal complement $U^\perp$ of a linear subspace U is understood with respect to the scalar product $\langle \bullet, \bullet \rangle_H$ in the Hilbert space H.)

The smallest singular value σ of (2.2) below leads to an alternative characterisation of the quasi-best approximation constant $\|P\| = \sigma^{-1}$.

Lemma 3 (Any Oblique Projection Stems from Petrov–Galerkin Scheme). *There exists some N-dimensional subspace $T = \mathcal{N}(1 - P^*)$ of H such that $P \in L(H)$ is characterised by*

$$Px \in S \quad \text{and} \quad x - Px \perp T \quad \text{for all } x \in H. \tag{2.1}$$

The characterisation (2.1) *is unique in that, given any $x \in H$, there is a unique Px with* (2.1). *Moreover, the restricted orthogonal and oblique projection $\Pi_S|_T : T \to S$ and $P|_T : T \to S$ are isomorphisms, $S \cap T^\perp = \{0\} = T \cap S^\perp$ are trivial, and the discrete* inf-sup *constant $\sigma > 0$ is positive and equal to the reciprocal operator norm*

$$\sigma := \inf_{\substack{s \in S \\ \|s\|=1}} \sup_{\substack{t \in T \\ \|t\|=1}} \Re\langle s, t\rangle_H = \inf_{\substack{t \in T \\ \|t\|=1}} \sup_{\substack{s \in S \\ \|s\|=1}} \Re\langle s, t\rangle_H = \|P\|_{L(H)}^{-1} > 0. \tag{2.2}$$

Proof. This is certainly known to the experts and hence we solely sketch the arguments. The finite range of P makes P a compact operator with eigenvalue 1 and eigenspace S (recall idempotence $P^2 = P$ for a projection and infer $Ps = s$ exactly for $s \in S$). The Hilbert space adjoint $P^* \in L(H)$ is also compact and has the eigenvalue 1 with an eigenspace T of the finite dimension $N = \dim T$ equal to that of P. Duality implies $x - Px \perp T$ for all $x \in H$ because $\langle x - Px, t\rangle_H = \langle x, t\rangle_H - \langle x, P^* t\rangle_H = 0$ vanishes for any eigenvector $t \in T$ of P^* to the eigenvalue 1. This establishes (2.1).

The further assertions follow from this. For instance, given any $t \in T$ with $\|t\| = 1$, $x = t$ implies in (2.1) that $1 = \|t\|^2 = \langle Pt, t\rangle_H = \langle Pt, \Pi_S t\rangle_H$ (by $Pt \in S$), whence $Pt \ne 0 \ne \Pi_S t$. Consequently, the restrictions $P|_T : T \to S$ and $\Pi_S|_T : T \to S$ are isomorphisms. In particular, given any $s \in S$ with $\|s\| = 1$ there exists some $t \in T \setminus \{0\}$ with $s = \Pi_S t$ and so $\langle s, t\rangle_H = \langle s, \Pi_S t\rangle_H = \|s\|^2 = 1$; in other words $s \perp T$ and $s \in S$ implies $s = 0$, i.e., $S \cap T^\perp = \{0\}$. The discrete inf-sup constant $\sigma > 0$ is attained in equation (2.2) in that there exists $s \in S$ with $\|s\| = 1$ and $\sigma = \sup_{t \in T\setminus\{0\}} \Re\langle s, t\rangle_H/\|t\|$. The aforementioned calculation with $s = \Pi_S t$ and $t \in T \setminus \{0\}$ provides the positive lower bound $\|t\|^{-1} \le \sigma$. The second equality asserts that the singular values of a matrix and its complex transpose coincide. Using the second identity for the smallest singular value σ, we infer

$$\sigma = \inf_{t \in T\setminus\{0\}} \sup_{s \in S\setminus\{0\}} \frac{\Re\langle s, \Pi_S t\rangle_H}{\|t\|\,\|s\|} = \inf_{t \in T\setminus\{0\}} \frac{\|\Pi_S t\|}{\|t\|} \tag{2.3}$$

with a Cauchy inequality (and the discussion of the equality sign therein) in the last step. Consequently,

$$\sigma^{-1} = \sup_{t\in T\setminus\{0\}} \frac{\|t\|}{\|\Pi_S t\|} = \sup_{t\in T\setminus\{0\}} \sup_{x\in H\setminus\{0\}} \frac{\mathbb{R}\langle x,t\rangle_H}{\|x\|\|\Pi_S t\|} = \sup_{t\in T\setminus\{0\}} \sup_{x\in H\setminus\{0\}} \frac{\mathbb{R}\langle Px,\Pi_S t\rangle_H}{\|x\|\|\Pi_S t\|} \tag{2.4}$$

with $\langle x,t\rangle_H = \langle Px,t\rangle_H = \langle Px,\Pi_S t\rangle_H$ (by definition of P resp. Π_S) in the last step. An interchange of the two suprema in the last expression and another Cauchy inequality (with possibly $Px = \Pi_S t$) reveals equality to $\sup_{x\in H\setminus\{0\}} \|Px\|/\|x\| = \|P\|$. □

3 Tantardini–Veeser formula for Petrov–Galerkin

The typical application of the above results in a Hilbert space concern the Petrov–Galerkin schemes for Hilbert spaces $X_h \subset X$ and $Y_h \subset Y$ with finite dimension $N := \dim X_h = \dim Y_h \in \mathbb{N}$ and a bilinear form $b : X\times Y \to \mathbb{K}$ (for the underlying complex or real field $\mathbb{K}$). The main assumption on the discrete subspaces X_h and Y_h of the same finite dimension N is the discrete inf-sup constant $\beta_h > 0$, where

$$\beta_h := \inf_{x_h\in X_h\setminus\{0\}} \sup_{y_h\in Y_h\setminus\{0\}} \frac{\mathbb{R} b(x_h,y_h)}{\|x_h\|_X\,\|y_h\|_Y}. \tag{3.1}$$

Given the situation of a positive β_h, textbook analysis defines an operator $P \in L(X)$ by

$$Px \in X_h \quad \text{satisfying} \quad b(x - Px, y_h) = 0 \quad \text{for all } y_h \in Y_h. \tag{3.2}$$

It turns out that the discrete problem with a $N\times N$ stiffness matrix, that represents the discrete bilinear form $b|_{X_h\times Y_h}$, is regular and Px is uniquely determined and P is a linear operator and a projection. From now on we write $H := X$ and $S := X_h = \mathcal{R}(P)$ to compare with the results from the previous sections. The quasi-best approximation in textbooks, however, utilises the immediate estimate

$$\|P\|_{L(H)} \le \beta_h^{-1}\|b\|$$

with the continuous bound $\|b\|$ of the bilinear form b. This and the Kato lemma provide the quasi-best approximation with a multiplicative (possibly suboptimal) constant $\beta_h^{-1}\|b\|$.

In comparison, the finer Tantardini–Veeser characterisation from [5] reads (1.1) with

$$\|P\|_{L(H)} = \sup_{y_h\in Y_h\setminus\{0\}} \frac{\|b(\bullet,y_h)\|_{X^*}}{\|b(\bullet,y_h)\|_{X_h^*}} \tag{3.3}$$

for the dual norms defined, for the subspace $U = X_h$ and $U = X = H$, by

$$\|b(\bullet,y_h)\|_{U^*} := \sup_{u\in U\setminus\{0\}} \frac{\mathbb{R} b(u,y_h)}{\|u\|}.$$

Corollary 4 (Tantardini–Veeser). *The identity* (3.3) *is a rewriting of* (2.3).

Proof. Let $R : H \to H^*$ denote the Riesz isomorphism in the Hilbert space $H = X$ and let the linear operator $B_{2h} : Y_h \to H^*, y_h \mapsto b(\bullet,y_h)$ be associated to the reduced bilinear form $b|_{H\times Y_h}$. Given $y_h \in Y_h$, set $t := R^{-1}B_{2h}y_h \in T := \mathcal{R}(R^{-1}B_{2h}) \subset H$ to define a linear surjection $R^{-1}B_{2h} : Y_h \to T$. Since $N = \dim X_h = \dim Y_h \in \mathbb{N}$ and $\beta_h > 0$, $R^{-1}B_{2h} : Y_h \to T$ is injective, whence an isomorphism. We deduce $N = \dim S = \dim T$ (recall $S := X_h \subset H = X$) and (2.1): Given any $(x,t) \in H\times T$ and $t = R^{-1}B_{2h}y_h$ for some $y_h \in Y_h$, the definition of P for the Petrov–Galerkin scheme (3.2) implies

$$\langle x - Px, t\rangle_H = (B_{2h}y_h)(x - Px) = b(x - Px, y_h) = 0.$$

Lemma 1 provides quasi-best approximation with the constant $\|P\|_{L(H)} = \sigma^{-1}$ and the proof of Lemma 3 implies (2.3). The isomorphism $R^{-1}B_{2h} : Y_h \to T$ rewrites any $t = R^{-1}B_{2h}y_h \in T\setminus\{0\}$ in terms of $y_h \in Y_h\setminus\{0\}$ and vice versa. It follows

$$\|t\| = \|B_{2h}y_h\|_{X^*} = \|b(\bullet,y_h)\|_{X^*},$$

$$\|\Pi_S t\| = \sup_{s\in S\setminus\{0\}} \frac{\mathbb{R}\langle t,s\rangle_H}{\|s\|} = \sup_{s\in S\setminus\{0\}} \frac{\mathbb{R} b(s,y_h)}{\|s\|} = \|b(\bullet,y_h)\|_{X_h^*}$$

with $\langle s, t\rangle_H = (B_{2h}y_h)(s) = b(s, y_h)$ in the second last step. Consequently, the substitution of $\sigma^{-1} = \|P\|_{L(H)}$ in (2.3) (cf. also the first identity in (2.4)) is a rewriting of (3.3). □

4 Nonconforming FEM with Smoother

The above examples were all conforming in that $X_h = S \subset X = H$ and so there are at least two things very different with nonconforming schemes. First we need to define a new Hilbert space $H := V + V_h$ out of the Hilbert space (V, a) on the continuous level and (V_h, a_h) on the discrete one. Second there is no quasi-best approximation in general and we need a smoother J as a game changer [6].

The abstract conditions of this section are satisfied for either the classical lowest-order nonconforming finite element schemes, namely the Crouzeix–Raviart finite element scheme, where $V = H^1_0(\Omega)$ and a is the energy scalar product in H^1, or the Morley finite element scheme with $V = H^2_0(\Omega)$ and the energy scalar product a in H^2.

Let (V, a) and (V_h, a_h) denote two Hilbert spaces and suppose that V and V_h are composed of Lebesgue functions over some domain Ω so that the sum $H := V + V_h$ is well defined as a linear space. Suppose moreover that there is a scalar product $\langle \bullet, \bullet\rangle_H$ on H that makes H a Hilbert space. The point is that this scalar product is equal to $a = \langle \bullet, \bullet\rangle_H | V \times V$ for continuous and equal to $a_h = \langle \bullet, \bullet\rangle_H | V_h \times V_h$ for nonconforming functions. The aforementioned classical examples are covered by [2, Remark 2.7] and $\langle \bullet, \bullet\rangle_H$ is the piecewise version of the energy scalar products and this semi-scalar product leads in fact to a scalar product $\langle \bullet, \bullet\rangle_H$. From now on we abbreviate $S := V_h$ as a closed subspace of H of finite dimension $N \in \mathbb{N}$.

The discrete schemes comes with an interpolation operator $I : V \to S$ with various remarkable properties. We utilise the orthogonality that makes I a best-approximation and we can, in fact, define $I \in L(H)$ by

$$Ix \in S \quad \text{and} \quad \langle x - Ix, s\rangle_H = 0 \quad \text{for all } (s, x) \in S \times H. \tag{4.1}$$

This orthogonality makes the interpolation operator $I = \Pi_S$ equal to the orthogonal projection Π_S of Lemma 1. The point is that the applications allow for a local definition of I and this miracle allows for the design of a right-inverse $J \in L(V_h; V)$ of $I \in L(V; V_h)$; cf. [1, 2, 6] for details or references on the design by local averaging and bubble-function corrections. In other words the linear operator $J : V_h \to V$ satisfies $IJ = 1$ in S, i.e.,

$$I(Js) = s \quad \text{for all } s \in S. \tag{4.2}$$

The discrete scheme then modifies the right-hand side as follows: Given any $F = a(u, \bullet) \in H^*$ with exact solution $u \in V$ (the Riesz representation of F in (V, a)),

$$\text{let} \quad Pu \in S \quad \text{solve} \quad a_h(Pu, r) = F(Jr) \quad \text{for any} \quad r \in S \tag{4.3}$$

(i.e., Pu is the Riesz representation of $F \circ J \in V_h^*$ in (V_h, a_h)). Since $\langle u, Jr\rangle_H = a(u, Jr) = F(Jr) = a_h(Pu, r) = \langle Pu, r\rangle_H$ and $\langle t, Pu\rangle_H = \langle It, Pu\rangle_H$ holds for $t = Jr \in T := J(S) = \mathcal{R}(J)$ (by (4.1)–(4.2) for $x = Jr = t$ and $r, Pu \in S$), we infer $\langle u - Pu, t\rangle_H = 0$ for all $(u, t) \in V \times T$. This defines $P \in L(V; S)$ with quasi-best approximation [2, 6]

$$\|u - Pu\| \le C_{\mathrm{qo}} \|u - \Pi_S u\| \quad \text{for all } u \in V. \tag{4.4}$$

Lemma 5 ([2, 6]). *The nonconforming scheme* (4.3) *allows for quasi-best approximation* (4.4) *with best-possible constant $C_{qo} = \|J\|_{L(V_h;V)}$, the operator norm of J in $L(V_h; V)$.*

Proof. The above arguments on the nonconforming scheme (4.3) lead to an operator $P : V \to S$, which we extend to $P \in L(H)$ by

$$Px \in S \quad \text{satisfies} \quad \langle x - Px, t\rangle_H = 0 \quad \text{for all } (x, t) \in H \times T$$

for the subspaces $S = V_h \subset H$ and $T := J(S) = \mathcal{R}(J) \subset H$ of the same finite dimension and $\sigma > 0$. This is (3.2) and we obtain quasi-best approximation (1.1) with best possible constant $\|P\| = \sigma^{-1}$. Another representation by $\|J\|_{L(V_h;V)}$ is derived in the following. Given any $t = Jr \in T$ with $r \in S$, recall that $r = IJr = It$ (by (4.2)) and $It = \Pi_S t$ (by $I = \Pi_S$ from (4.1)); whence $\Pi_S t = r$. This and $t = Jr \in T$ for any $r \in S \setminus \{0\}$ reveal in (2.3) that

$$\|P\|_{L(H)} = \sigma^{-1} = \sup_{t \in T\setminus\{0\}} \frac{\|t\|}{\|\Pi_S t\|} = \sup_{s \in S\setminus\{0\}} \frac{\|Js\|}{\|\Pi_S Js\|} = \sup_{s \in S\setminus\{0\}} \frac{\|Js\|}{\|s\|} = \|J\|_{L(V_h;V)}.$$

Therefore the multiplicative constant $\|J\|_{L(V_h;V)}$ is best possible in the inequality

$$\|x - Px\| \leq \|J\|_{L(V_h;V)} \|x - \Pi_S x\| \quad \text{for all } x \in H. \tag{4.5}$$

The estimate (4.5) implies (4.4) for C_{qo} replaced by $C_{\mathrm{qo}} = \|J\|_{L(V_h;V)}$; but, at this point, there remains a little twist: Since (4.4) merely asks for $u \in V$ (and not for *all* $x \in H$ as in (4.5)), this additional restriction could, in principle, lead to a better constant. The proof that this is impossible considers any positive $\kappa < 1$ so that $\kappa\|J\|_{L(V_h;V)}$ cannot replace the optimal factor $\|J\|_{L(V_h;V)} > 0$ in (4.5). Hence there exists some $x \in H \setminus S$ with

$$\kappa\|J\|_{L(V_h;V)} \|x - \Pi_S x\| \leq \|x - Px\|.$$

Since $x = v + s$ for some $v \in V$ and $s \in S = V_h$ with $Ps = s = \Pi_S s$, we deduce $x - \Pi_S x = v - \Pi_S v \neq 0$ and $x - Px = v - Pv$; whence

$$0 < \kappa\|J\|_{L(V_h;V)} \|v - \Pi_S v\| \leq \|v - Pv\| \leq C_{\mathrm{qo}} \|v - \Pi_S v\|$$

follows with (4.4) in the last step. We infer first $0 < \kappa\|J\|_{L(V_h;V)} \leq C_{\mathrm{qo}}$ for all positive $\kappa < 1$ and second $\|J\|_{L(V_h;V)} \leq C_{\mathrm{qo}}$. Recall the converse inequality from above to conclude the proof that $\|J\|_{L(V_h;V)} = C_{\mathrm{qo}}$ is best possible in (4.4). □

Acknowledgment: It is my pleasure to thank Alexandre Ern for stimulating discussions on the topic that led to this paper. A Gordon Preston sabbatical fellowship at Monash University, Melbourne, Australia, and the hospitality at its school of mathematics, where this work was finalised, is thankfully acknowledged.

References

[1] C. Carstensen, D. Gallistl and M. Schedensack, Adaptive nonconforming Crouzeix–Raviart FEM for eigenvalue problems, *Math. Comp.* **84** (2015), no. 293, 1061–1087.

[2] C. Carstensen and N. Nataraj, A priori and a posteriori error analysis of the Crouzeix–Raviart and Morley FEM with original and modified right-hand sides, *Comput. Methods Appl. Math.* **21** (2021), no. 2, 289–315.

[3] A. Ern and J.-L. Guermond, *Finite Elements II—Galerkin Approximation, Elliptic and Mixed PDEs*, Texts Appl. Math. 73, Springer, Cham, 2021.

[4] D. B. Szyld, The many proofs of an identity on the norm of oblique projections, *Numer. Algorithms* **42** (2006), no. 3–4, 309–323.

[5] F. Tantardini and A. Veeser, The L^2-projection and quasi-optimality of Galerkin methods for parabolic equations, *SIAM J. Numer. Anal.* **54** (2016), no. 1, 317–340.

[6] A. Veeser and P. Zanotti, Quasi-optimal nonconforming methods for symmetric elliptic problems. I—Abstract theory, *SIAM J. Numer. Anal.* **56** (2018), no. 3, 1621–1642.