ESAIM: M2AN 56 (2022) 177–211
https://doi.org/10.1051/m2an/2021086

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

GIBBS PHENOMENA FOR L^q-BEST APPROXIMATION IN FINITE ELEMENT SPACES *

Paul Houston[1], Sarah Roggendorf[2,*] and Kristoffer G. van der Zee[1]

Abstract. Recent developments in the context of minimum residual finite element methods are paving the way for designing quasi-optimal discretization methods in non-standard function spaces, such as L^q-type Sobolev spaces. For $q \to 1$, these methods have demonstrated huge potential in avoiding the notorious Gibbs phenomena, *i.e.*, the occurrence of spurious non-physical oscillations near thin layers and jump discontinuities. In this work we provide theoretical results that explain some of these numerical observations. In particular, we investigate the Gibbs phenomena for L^q-best approximations of discontinuities in finite element spaces with $1 \leq q < \infty$. We prove sufficient conditions on meshes in one and two dimensions such that over- and undershoots vanish in the limit $q \to 1$. Moreover, we include examples of meshes such that Gibbs phenomena remain present even for $q = 1$ and demonstrate that our results can be used to design meshes so as to eliminate the Gibbs phenomenon.

Mathematics Subject Classification. 65N30, 41A10.

Received December 22, 2020. Accepted December 16, 2021.

1. Introduction

This article investigates the Gibbs phenomenon in the context of the L^q-best approximation of discontinuous functions in finite element spaces by considering a few carefully selected cases that can be analysed in detail. The Gibbs phenomenon was originally discovered by Henry Wilbraham [37] and described by Willard Gibbs [9] in the context of approximating jump discontinuities by partial sums of Fourier series. It also occurs in the best approximation of functions either by a trigonometric polynomial in the L^1-metric [28] or spline functions in the L^2-metric [32]. The best approximation in finite element spaces consisting of piecewise polynomials is closely related to the last example. In [35], Saff and Tashev show that in one dimension the best approximation of a jump discontinuity by polygonal lines leads to Gibbs phenomena for all $1 < q < \infty$ but vanishes as $q \to 1$; this is the starting point of our investigation.

We consider several meshes in one and two dimensions and show that on certain meshes the over- and undershoots in the best approximation can be eliminated in the limit $q \to 1$. These results are extensions of [35]. However, there exist meshes in both one and two dimensions that do not satisfy this property. The aim

Keywords and phrases. Best approximation, Gibbs phenomenon, L^q, finite elements.

* *The research by KvdZ was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK under Grant EP/T005157/1.*

[1] School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham NG7 2RD, UK.
[2] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK.
*Corresponding author: sr957@cam.ac.uk

of this article is therefore to illustrate which properties the underlying mesh must satisfy to ensure that the oscillations vanish in the L^q-best approximation of discontinuous functions.

This study of L^q-best approximations in finite element spaces is motivated by approximating solutions to partial differential equations (PDEs) in subspaces of $L^1(\Omega)$. In [10], Guermond points out that there are only very few attempts at achieving this despite the fact that first-order PDEs and their non-linear generalizations have been extensively studied in $L^1(\Omega)$. The existing numerical methods which seek an approximation directly in $L^1(\Omega)$ include the ones outlined in the articles by Lavery [22–24], the reweighted least-squares method of Jiang [16, 17] and the methods outlined in the series of articles by Guermond *et al.* [10–14]. More recently, a novel approach to designing finite element methods in a very general Banach space setting has been introduced in [29] and applied to the advection-reaction equation [30] and to the convection-diffusion-reaction equation [15, 25, 26]. This approach is based on the so-called discontinuous Petrov–Galerkin methods, *cf.*, *e.g.*, [5], and extends the concept of optimal test norms and functions from Hilbert spaces to more general Banach spaces. At least in an abstract sense, this approach outlines how to design a numerical method that leads to a quasi-best approximation of the solution in a space of choice, provided the continuous problem is well-posed in a suitable sense. In practice, there are hurdles to overcome to design a practical method, but this is not the subject of this article. Nonetheless, it opens up a new approach to designing numerical methods that raises the question of which norms and spaces are favorable for the approximation of certain types of PDEs.

In the context of approximating solutions containing discontinuities and under resolved interior- and boundary layers, the numerical results for existing L^1-methods suggest such features can be approximated as sharply as a given mesh permits without exhibiting spurious over- or undershoots. This property clearly gives them an enormous advantage over traditional finite element methods yielding approximations in subspaces of $L^2(\Omega)$. Indeed, it is well-known that even seemingly simple examples such as the transport equation or convection-dominated diffusion equations require extra care in the design of the method, with the standard Galerkin finite element method being unstable, and alternative methods often requiring so-called *stabilization* and/or *shock-capturing* techniques, *cf.*, *e.g.*, [18–20, 34].

1.1. Notation

Throughout this article, we denote by $L^q(\Omega)$, $1 \leq q < \infty$, the Lebesgue space of q-integrable functions on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$; $L^\infty(\Omega)$ is the Lebesgue space of functions on Ω with finite essential supremum; and $W^{1,q}(\Omega)$, $1 \leq q \leq \infty$, is the Sobolev space of functions that are in $L^q(\Omega)$ such that their gradient is in $L^q(\Omega)^d$. Furthermore, $W_0^{1,q}(\Omega) \subset W^{1,q}(\Omega)$ is the subspace of all functions with zero trace on the boundary $\partial\Omega$. The corresponding norms are denoted by $\|\cdot\|_{L^q(\Omega)}$ and $\|\cdot\|_{W^{1,q}(\Omega)}$, respectively. For $q = 2$, we furthermore use the usual notation $H^1(\Omega) := W^{1,2}(\Omega)$ and $H_0^1 := W_0^{1,2}(\Omega)$. For $1 \leq q \leq \infty$, we write q' to denote the dual exponent such that $1/q + 1/q' = 1$. For any Banach space V, its dual space is denoted by V'. Furthermore, for $v \in V$ and $\varphi \in V'$, we have the duality pairing

$$\langle \varphi, v \rangle_{V',V} := \varphi(v).$$

The subdifferential of a function $f : V \to \mathbb{R}$ at a point $v \in V$ is denoted by $\partial f(v) \subset V'$.

1.2. Motivation

To motivate the best approximation problem we analyse in this article, we consider the following simple convection-diffusion problem: for $\varepsilon > 0$, find u such that

$$-\varepsilon u'' + u' = 0 \quad \text{in } (0,1), \qquad u(0) = 1, \qquad u(1) = 0. \tag{1.1}$$

The analytical solution to this problem is given by

$$u(x) = \frac{1 - e^{-\frac{1-x}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}};$$

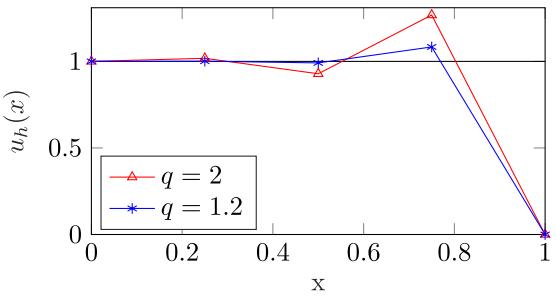


FIGURE 1. L^q-best approximation to $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$.

in particular, there is a boundary layer near $x = 1$ for small ε. In two dimensions, we consider a rather straightforward extension of the one-dimensional example: find u such that

$$\begin{aligned} -\varepsilon\Delta u + \partial_x u = 0 \quad \text{in } (0,1)^2, \quad u(0,\cdot) = 1, u(1,\cdot) = 0, \\ \partial_{\boldsymbol{n}} u = 0 \text{ if } y = 0 \text{ or } y = 1, \end{aligned} \tag{1.2}$$

where $\boldsymbol{n}$ denotes the unit outward normal vector on the boundary of the domain.

We seek an approximation of the analytical solution in a finite dimensional space that consists of continuous piecewise linear polynomials defined on a given mesh. If $\varepsilon \ll 1$, then the second-order term is completely dominated by the first-order term and away from the outflow boundary the solution is essentially given by the solution to the advection problem obtained by setting ε to zero. For the above problems this means that $u \approx 1$ away from the outflow boundary. Due to the Dirichlet boundary conditions, a boundary layer forms near the outflow boundary. If the diameter of the elements near the boundary layer is large compared with ε, the layer is fully contained within these elements and, in the above problems, $u \approx 1$ in the remainder of the domain. Numerically, this essentially means that we approximate the problems (1.1)/(1.2) with $\varepsilon = 0$ while still keeping the boundary conditions at both ends. Clearly, the analytical solution for the above problems with $\varepsilon = 0$ and the boundary conditions only imposed on the inflow part of the boundary is $u \equiv 1$. This motivates us to consider the best approximations of $u \equiv 1$ by linear finite element functions satisfying the boundary conditions given in (1.1) and (1.2), respectively.

Figure 1 shows the L^q-best approximation of $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$. We can see that in both cases over- and undershoots are present in the approximation, but that the magnitude of these oscillations is significantly smaller for $q = 1.2$. This example illustrates the phenomenon of reducing oscillations in the approximation as $q \to 1$ that we shall investigate in this article.

Before we delve into the precise analysis of the L^q-best approximation in more complex situations, let us look at the simplest example in order to gain some intuition why the over- and undershoots in the L^q-best approximation of discontinuities reduce as $q \to 1$. To this end, we consider an approximation u_h to $u \equiv 1$, where u_h is a piecewise linear function on a two-element mesh on $(0,1)$ satisfying $u_h(0) = 1$ and $u_h(1) = 0$. If h is the length of the second element, the free parameter in the approximation is $u_h(1-h) = 1 + \delta$. Three different choices for δ are shown in Figure 2. Clearly, the error with $\delta < 0$ is always larger than the error with $\delta = 0$. Hence, we can assume $\delta \geq 0$. Whether $\delta = 0$ or some $\delta > 0$ yields the better approximation is, however, less obvious. Roughly speaking, replacing $\delta = 0$ with a small $\delta > 0$ increases the overall area that contains an error while at the same time decreasing the area where the pointwise error is close to 1. Therefore, we can expect that in certain situations there exists a $\delta > 0$ that yields an approximation with a smaller error than $\delta = 0$. This argument clearly fails for $\delta \geq 1$, hence we would always expect $\delta < 1$.

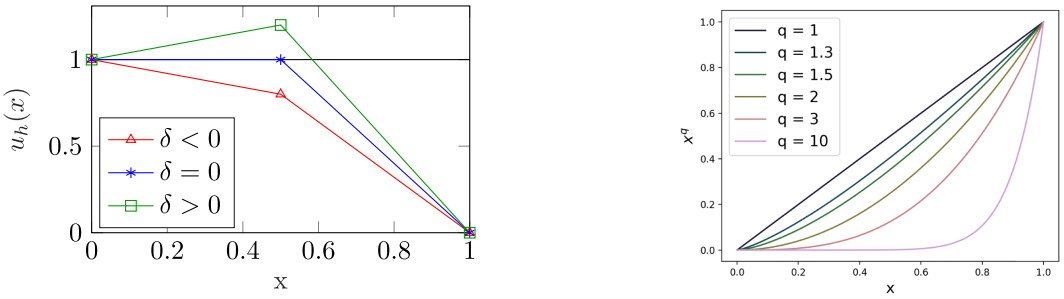


FIGURE 2. *Left*: possible approximations to $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of two elements such that $u_h(0.5) = 1+\delta$. *Right*: x^q for $x \in [0, 1]$ and several values of $q \geq 1$.

The error in the L^q-norm is determined by integrating $|u - u_h|^q$ over the whole interval. As $q \geq 1$ increases, areas containing larger pointwise errors are weighted increasingly more heavily compared to areas containing smaller pointwise errors. This can be seen by looking at the graph of x^q for different values of $q \geq 1$ as shown in Figure 2 on the right. As q increases the graph becomes flatter away from 1 such that the range of values for the pointwise error that contribute comparatively little to the L^q-error increases. At the same time the graph becomes steeper close to 1 implying that the contribution of pointwise errors close to 1 to the L^q-error increases. This suggests, that we can expect $\delta_2 > \delta_1$ if $u_{h,i} = 1 + \delta_i$, $i = 1, 2$, are the L^{q_i}-best approximations for some $1 \leq q_1 < q_2$. Indeed, we will later see that the overshoot is an increasing function in q. Moreover, for $q > 1$, $\delta = 0$ *never* yields an L^q-best approximation and for $q = 1$ only if the resulting area where $u_h < u$ is sufficiently small. Furthermore, the conjecture $\delta < 1$ will also be confirmed and a plot of the overshoot for large q suggests that $\delta \to 1$ as $q \to \infty$. In fact, it is easy to see that $\delta = 1$ yields an L^∞-best approximation. Indeed, any $\delta \in [-1, 1]$ yields an L^∞-best approximation since for $|\delta| \leq 1$ the maximal error never exceeds 1, but is always 1 independently of the precise value of δ due to the boundary condition at $x = 1$.

The observation that for larger q it is "better" to commit small errors in the entire domain than a very large error in one element, is similar to the observations made by Guermond *et al.* [14] in the context of residual minimization in L^1. The authors observe that the L^1-minimizer commits a large error in one cell and no error in all other cells in contrast to the L^2-minimizer that spreads the error over all cells. They furthermore observe that this corresponds to selecting a sparse residual vector in the discrete setting which reflects the sparsity property of discrete L^1-minimizers [6, 7].

1.3. Problem statement

We consider a subdivision Ω_h of the domain $\Omega = (0, 1)^d$, $d = 1, 2$ into n disjoint open simplicial elements (*i.e.*, subintervals when $d = 1$ and triangles when $d = 2$) κ_i, $i = 1, \ldots, n$, such that $\bar{\Omega} = \bigcup_{i=1}^n \bar{\kappa}_i$ and define U_h to be the standard finite element space consisting of continuous piecewise linear polynomials on the mesh Ω_h that are zero on the boundary. Let $u \in \bigcap_{1 \leq q \leq \tilde{q}} \mathrm{L}^q(\Omega)$ for some $\tilde{q} \in (1, \infty)$ and consider the following (constrained) best approximation problem:

$$u_h = \arg \min_{v_h \in U_h} \|u - v_h\|_{\mathrm{L}^q((0,1)^d)} \tag{1.3a}$$

subject to

$$\begin{aligned} u_h(0) &= u(0), \quad u_h(1) = g \quad \text{if } d = 1, \\ u_h(0, \cdot) &= u(0, \cdot), \quad u_h(1, \cdot) = g \quad \text{if } d = 2, \end{aligned} \tag{1.3b}$$

where either $\operatorname{sgn}(u-g) \equiv 1$ or $\operatorname{sgn}(u-g) \equiv -1$ on $\Gamma_1 = \{x = 1\}$. Note that the constraint can be removed by using a Dirichlet lift argument as commonly employed in the context of finite element methods such that indeed $u_h \in U_h$ as defined above, in particular, $u_h = 0$ on the boundary. This is possible since the conditions for the L^q-best approximation only depend on the difference $u - u_h$ which is not affected by the Dirichlet lift.

We usually assume u to be continuous and piecewise linear as well such that u and u_h satisfy different boundary conditions. In some cases we only consider the example $u \equiv 1$. In one dimension, we also consider the L^q-best approximation of the discontinuous function $u(x) = \operatorname{sgn}(x)$ on $(-1,1)$ by a continuous piecewise linear function u_h satisfying $-u_h(-1) = u_h(1) = 1$. We use this example to establish the link between our work and [35].

There is a related body of literature studying the L^2-projection onto finite element spaces, such as [2, 4, 8]. These works are mostly concerned with the stability of the projection operator in subspaces (*e.g.*, $L^q(\Omega)$, $W^{1,q}(\Omega)$, $H^1_0(\Omega)$).

1.4. Summary of results

The main result of this article consists of the precise analysis of specific cases that illustrate the behavior of L^q-best approximations of discontinuities by continuous piecewise linear polynomials on coarse meshes. The mesh configurations for these examples are chosen to be sufficiently simple such that an explicit solution of L^1/L^q-best-approximation problems is possible while at the same time allowing us to draw conclusions for more general meshes. In order to demonstrate the conclusions for more general situations, we use finite element techniques to numerically determine L^q-best approximations on more complex meshes. We employ an algorithm based on a regularization of the L^q-norm and Newton's method with line search described in Section 3.6 of [10]. In particular, we demonstrate that the over- and undershoots observed in L^q-best approximations for $1 < q < \infty$ decrease as $q \to 1$. Whether these oscillations disappear entirely depends on the mesh used to define the underlying finite dimensional approximation space. In one dimension, Gibbs phenomena can be eliminated on uniform meshes both for a boundary discontinuity and a jump discontinuity present in the interior of the domain. For non-uniform meshes it depends on the relative sizes of the elements. In two dimensions, we show that there exist uniform and structured meshes for which Gibbs phenomena are not eliminated. But, we also include examples of meshes in two dimensions on which the over- and undershoots vanish as $q \to 1$. Moreover, we establish sufficient conditions on meshes in one dimension and on certain classes of meshes in two dimensions that ensure that Gibbs phenomena can be eliminated. Additionally, we will illustrate that there exist infinitely many L^1-best approximations in certain cases which is due to the fact that $L^1(\Omega)$ is not strictly convex.

1.4.1. Boundary discontinuity in one dimension

The first case we consider is the approximation problem (1.3) with $d = 1$. The key result regarding this is a very general condition on the mesh for a general N-element mesh that guarantees the existence of an L^1-best approximation with no over- or undershoots.

Theorem 1.1 (A sufficient condition in one dimension)**.** *Let the mesh be given by a subdivision of the interval $(0,1)$ into $N \geq 2$ intervals (x_{i-1}, x_i), $i = 1, \dots, N$, with $0 = x_0 < x_2 < \cdots < x_{N-1} < x_N = 1$. The length h_i of the ith subinterval is given by $h_i = x_i - x_{i-1}$, $i = 1, \dots N$. Define*

$$\vartheta_N := 0, \tag{1.4a}$$

$$\vartheta_i^2 := \frac{1}{2}\left(1 - \left(2(1-\vartheta_{i+1})^2 - 1\right)\frac{h_{i+1}}{h_i}\right), i = N-1, \dots, 1, \tag{1.4b}$$

$$M := \max\left(\{0\} \cup \left\{i \in \{1, \dots, N-1\} \,:\, \vartheta_i \geq 1 - \frac{1}{\sqrt{2}}\right\}\right). \tag{1.4c}$$

Furthermore, denote by φ_i the continuous and piecewise linear function that satisfies $\varphi_i(x_j) = \delta_{ij}$. Let $U = \operatorname{span}\{\varphi_i \,:\, 0 \leq i \leq N\}$. If the mesh satisfies the condition

$$h_i \geq (2(1-\vartheta_{i+1})^2 - 1)h_{i+1}, \qquad \text{for} \quad i = M, M+1, \dots, N-1, \tag{1.5}$$

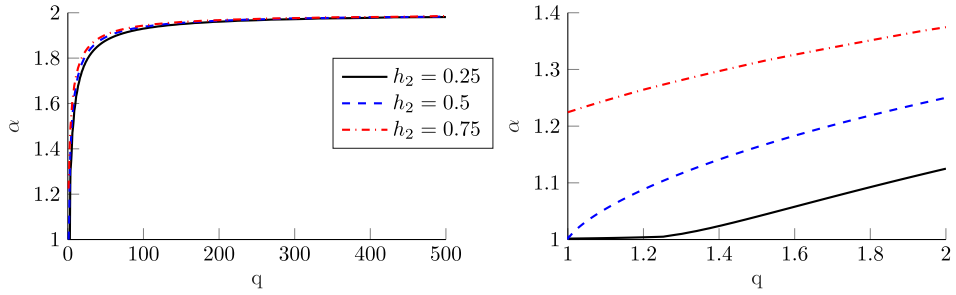


FIGURE 3. Values for α for different ranges of q and three different choices of h_2.

then there exists an L^1*-best approximation* $u_h \in U$ *of* $u \in U$ *subject to the constraint* $u_h(0) = u(0)$ *and* $u_h(1) = g \neq u(1)$ *with no over- or undershoots, i.e.,* $u_h(x_i) = u(x_i)$, $i = 1, \ldots, N-1$.

Note that condition (1.5) essentially states that elements cannot be too small compared to their neighboring element closer to the discontinuity. Furthermore, there are no conditions on the size of the elements contained in $(0, x_{M-1})$ if $M > 0$. Moreover, it is always possible to ensure $M > 0$ by selecting h_M sufficiently large in comparison to h_{M+1} such that $\vartheta_M > 1 - 1/\sqrt{2}$. For example, for $M = N-1$, we have

$$\vartheta_{N-1}^2 = \frac{1}{2}\left(1 - \frac{h_N}{h_{N-1}}\right) \geq \left(1 - \frac{1}{\sqrt{2}}\right)^2 \iff h_{N-1} \geq \frac{h_N}{2(\sqrt{2}-1)},$$

which is an explicit constraint for h_{N-1} if the mesh is to be designed such that $M = N-1$. Therefore, given x_{N-2} and x_N, one can place x_{N-1} such that this constraint on h_{N-1} is satisfied. This means that the mesh can be designed in such a way that it is allowed to be arbitrary away from the discontinuity without leading to oscillations. This observation is particularly useful if more than one discontinuity is to be approximated.

In the special case $N = 2$ and $u \equiv 1$, it is possible to fully analyse the L^q-best approximation for all $1 \leq q < \infty$. In this case condition (1.5) is equivalent to $h_2 \leq h_1 = 1 - h_2 \iff h_2 \leq 0.5$. If the condition is violated, *i.e.*, $h_2 > 0.5$, we will see that $u_h(x_1) = \sqrt{2h_2}$. For $q \neq 1$, we will see that $u_h(x_1) = \alpha > 1$ satisfying $0 = -(1-h_2)\alpha^2 q(\alpha-1)^{q-1} - h_2(\alpha q + 1)(\alpha-1)^q + h_2$. Figure 3 shows α for two different ranges of q and three different choices of h_2. The plot shows that $\alpha < 2$ for all $1 \leq q < \infty$ and that α decreases as $q \to 1$ for all three choices of h_2. Furthermore, we can see that the behavior as $q \to \infty$ is very similar for all choices of h_2, but that there are clear differences as $q \to 1$. For $h_2 = 0.25$ and $h_2 = 0.5$, α approaches 1 as $q \to 1$, hence the overshoot vanishes as $q \to 1$, whereas for $h_2 = 0.75$ it approaches $\sqrt{2h_2} \approx 1.2247$, hence the overshoot does not vanish. This is consistent with the results obtained for the L^1-best approximation.

Returning to the more general case, we will prove that in the simpler case that $h_N \leq \min_{i=1,\ldots,N-1} h_i$ the L^1-best approximation also contains no over- or undershoots. With very similar arguments it is easy to see that if $h_N > h_{N-1}$, but $h_{N-1} \leq h_i$ for all $i = 1, \ldots, N-2$, then every L^1-best approximation must contain over- or undershoots. Moreover, there exists an L^1-best approximation with overshoot only at the node x_{N-1} and no further over- or undershoots, *i.e.*, $u_h(x_i) = 1$ for $i = 1, \ldots, N-2$ and $u_h(x_{N-1}) = \sqrt{2h_N/(h_N + h_{N-1})}$. The value at $u_h(x_{N-1})$ follows from the case $N = 2$ and a rescaling of the interval. In Section 6.2 we include examples of two three-element meshes violating the sufficient condition in Theorem 1.1 such that one of the meshes satisfies (1.5), whereas the other mesh violates this condition as well. We will demonstrate that for the latter mesh the overshoot does indeed not vanish entirely as $q \to 1$.

1.4.2. Jump discontinuity in one dimension

The second L^q-best approximation problem we analyse is the best approximation of $u(x) = \mathrm{sgn}(x)$ on $(-1, 1)$ on a mesh consisting of exactly four elements that is symmetric with respect to $x = 0$. The main difference to

the result for $N = 2$ mentioned in the previous section is that there exists a whole family of best approximations if $q = 1$. For $q > 1$, we observe the same behavior as before.

Theorem 1.2 (L^q-best approximation of a jump discontinuity)**.** *Consider the mesh given by the subdivision of* $(-1, 1)$ *into the four intervals* $(-1, -h)$, $(-h, 0)$, $(0, h)$ *and* $(h, 1)$ *with* $h \in (0, 1)$. *For* $1 \le q < \infty$, *the* L^q*-best approximation of* $u = \operatorname{sgn}(x)$ *on* $(-1, 1)$ *by a continuous piecewise linear function* u_h *on the above mesh such that* $-u_h(-1) = u_h(1) = 1$ *can be characterized as follows.*

	$u_h(-h)$	$u_h(0)$	$u_h(h)$
$q > 1$	$-\alpha < -1$	0	$\alpha > 1$
$q = 1$, $h \le 0.5$	-1	$\beta \in [-1, 1]$	1
$q = 1$, $h > 0.5$	$-(1+\beta)\sqrt{2h} + \beta \le -1$	$\beta \in [-1, 1]$	$(1-\beta)\sqrt{2h} + \beta \ge 1$,

where α *satisfies* $0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h$ *and* $\beta \in [-1, 1]$ *is arbitrary. Furthermore, in the limit* $q \to 1$, *the* L^q*-best approximation converges pointwise to the* L^1*-best approximation with* $\beta = 0$.

Even though we only consider a very specific four-element mesh in Theorem 1.2, the results imply immediately that a condition analogous to Theorem 1.1 holds in this case as well. Furthermore, we will see that neither symmetry nor aligning the jump with an element boundary are essential for vanishing over- and undershoots in the limit $q \to 1$. The symmetric four-element mesh was selected for two reasons: firstly, the simplicity of the mesh allows us to explicitly determine *all* L^q-best approximations for $1 \le q < \infty$; secondly, the symmetry allows us to illustrate the non-uniqueness of the L^1-best approximation.

Note that the L^q-best approximation for $q > 1$ and the L^1-best approximation with $\beta = 0$ yields the same value for $u_h(h)$ as we obtained for $u_h(x_1)$ in the case $N = 2$ discussed in the previous section. However, while in the example in the previous section the L^1-best approximation is unique in the case $N = 2$, Theorem 1.2 characterizes a whole family of L^1-best approximation. This is possible since L^1 is not strictly convex and therefore minimizers are not necessarily unique. We recover uniqueness if we define the minimizer as the limit as $q \to 1$ of the L^q-minimizer. Moreover, it follows from the proof of Theorem 1.2 that the L^1-best approximation is unique if the subdivision of the interval is no longer symmetric, as we will see in Section 4.

In order to see how this result relates to the work in [35], it first has to be noted that there are two major differences between our investigation and [35]:

(1) The interval in [35] is subdivided into $2n$ subintervals of equal length. In contrast to this, we only consider the special case that $(-1, 1)$ is subdivided into 4 subintervals and instead allow the subdivision to be non-uniform but still symmetric with respect to the center of the interval.
(2) We consider bounded domains with fixed boundary conditions, which are relevant to finite element approximations, whereas the investigation in [35] considers the limit $n \to \infty$ for the interval $[-nh, nh]$ (ergo essentially an infinite domain) with no boundary conditions.

In [35] it is shown that for a uniform subdivision of the interval $[-nh, nh]$, the over- and undershoots disappear as $n \to \infty$ and $q \to 1$. The last point in Theorem 1.2 shows that, on a fixed mesh, we recover the result that the over- and undershoots disappear as $q \to 1$ for $h \le 0.5$, which includes the case of a uniform mesh. However, if $h > 0.5$, the over- and undershoots do not disappear as $q \to 1$.

1.4.3. Boundary discontinuity in two dimensions

The final theoretical results concern the solution to (1.3) with $d = 2$. We first consider the four meshes shown in Figure 4. Note that the discrete space U_h has only one degree of freedom on Mesh 1, corresponding to the value at the midpoint, and U_h has three degrees of freedom on the other meshes, corresponding to the values at the three nodes on the line $x = 0.5$.

We show that on all four meshes the L^q-best approximation of $u \equiv 1$ with $1 < q < \infty$ must contain over- or undershoots. For Mesh 1 we show that the L^1-best approximation of $u \equiv 1$ is unique and can be characterized by $u_h(0.5, 0.5) = \alpha$, where $\alpha > 1$ and $0 = 2\alpha^3 - 5\alpha + 2$. Hence, in this case, the overshoot

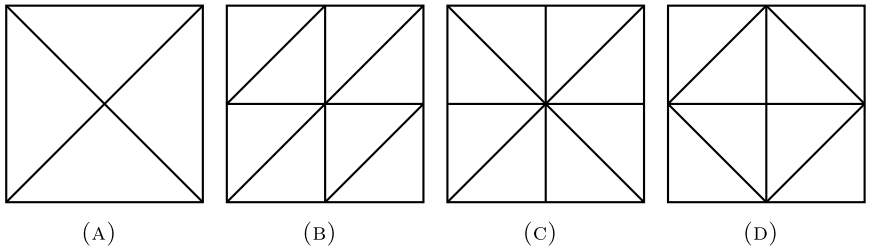


FIGURE 4. Four different meshes on $(0,1)^2$. (A) Mesh 1. (B) Mesh 2. (C) Mesh 3. (D) Mesh 4.

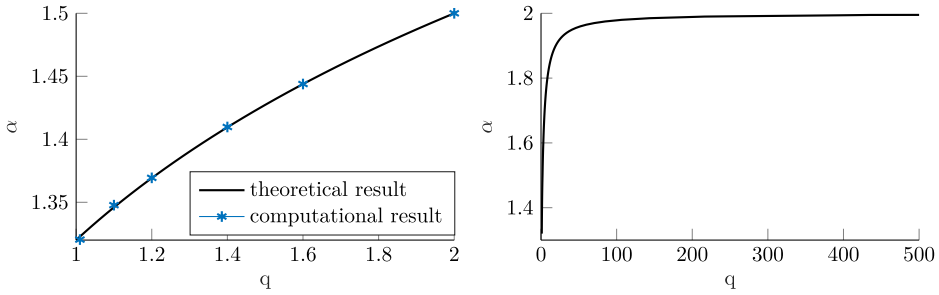


FIGURE 5. Values for α for different ranges of q on Mesh 1, *cf.*, Figure 4.

does not disappear in the limit $q \to 1$. For $q > 1$, we obtain $u_h(0.5, 0.5) = \alpha$, where $\alpha > 1$ and $0 = (\alpha - 1)^{q-1}\left[4\alpha^3 q + 4(1-q)\alpha^2 + (q-6)\alpha + 2\right] - \alpha(q+4) + 2$. Figure 5 shows the parameter α defining the L^q-best approximation on Mesh 1 for two different ranges of q. The plot shows that $\alpha < 2$ for all q and that α decreases as $q \to 1$, where it approaches 1.32. This is consistent with the conditions $\alpha > 1$ and $0 = 2\alpha^3 - 5\alpha + 2$ if $q = 1$. For Mesh 2 we show that any L^1-best approximation must contain over- or undershoots. One such best approximation is characterized by $u_h(0.5, 1) = u_h(0.5, 0.5) = 1$ and $u_h(0.5, 0) = \alpha$, where $\alpha > 1$ and $0 = -3\alpha^3 + 8\alpha - 4$. For Meshes 3 and 4 on the other hand, there exists an L^1-best approximation with no under- or overshoots. Moreover, we prove a necessary condition on general conforming meshes in two dimensions as well as a sufficient condition similar to Theorem 1.1 on a structured non-uniform mesh.

To confirm the theoretical results, we have also determined the L^q-best approximation numerically by implementing the best approximation problem as a variational problem using FEniCS [1]. The solution to the resulting non-linear system can be approximated using a Newton iteration if q is sufficiently close to 2. For $q < 2$, a regularization of the L^q-norm as introduced in [10] ensures that all terms in the Newton iteration are well-defined. We employ the algorithm described in Section 3.6 of [10] in order to determine the L^q-best approximations. This algorithm was originally developed for minimization problems of the form $u = \arg\min_{v \in E_h \subset E} \|Lv - f\|_F$, where $L : E \to F$ is a differential operator. In our case, the problem is much simpler, *i.e.*, $E = F = \mathrm{L}^q(\Omega)$ and L is the identity. The algorithm is based on a decreasing sequence of regularization parameters and Newton's method with line search. It should be noted that for q close to 1 a very high order quadrature rule is often necessary to ensure the accuracy of the algorithm. The left plot in Figure 5 shows numerically determined approximations of α for selected values of q which confirm the theoretical results.

We also include further numerical experiments in Section 6 illustrating that the observations remain the same if u is a more general smooth function and that the over- and undershoots cannot be eliminated by refining the mesh.

1.5. Outline of the paper

The remainder of this article is organized as follows: in Section 2 we describe a characterization of the L^q-best approximation of a function in a finite dimensional subspace that we will use to prove our theoretical results; Sections 3, 4 and 5 contain the proofs of the results described in Sections 1.4.1, 1.4.2 and 1.4.3, respectively. All results in the special cases $u \equiv 1$ and $u = \operatorname{sgn}(x)$ can also be found in [33]. We conclude with several numerical examples in Section 6 illustrating the effect of mesh refinement in one and two dimensions and showing the behavior of the L^q-best approximation as $q \to 1$ in one dimension, as well as on structured and unstructured meshes in two dimensions.

2. Characterization of best L^q-approximation

In this section we describe a characterization of best-approximations in Banach spaces and more specifically the Lebesgue spaces $L^q(\Omega)$, $1 \leq q < \infty$. This characterization will be used in the remainder of this article to determine the best L^q-approximation in specific cases.

If U is a Banach space and f a function $f : U \to \mathbb{R}$, the subdifferential $\partial f(u)$ of f at a point $u \in U$ is defined as the set

$$\partial f(u) := \{u' \in U' \,:\, f(w) - f(u) \geq \langle u', w - u\rangle_{U',U},\ \forall w \in U\}.$$

If f is convex and Gâteaux differentiable, the subdifferential is single valued and agrees with the Gâteaux derivative. We now quote the following theorem, *cf.*, Theorem 1.1 of [36].

Theorem 2.1 (Characterization of best approximation)**.** *Let U be a Banach space, $U_h \subset U$ a closed subspace and $u \in U$. The following statements are equivalent:*

(1) $u_h = \arg\min_{w_h \in U_h} \|u - w_h\|_U$.
(2) *There exists a functional $r' \in \partial(\|\cdot\|_U)(u - u_h)$ which annihilates U_h, i.e.,*

$$\langle r', w_h\rangle_{U',U} = 0 \qquad \text{for all } w_h \in U_h.$$

Remark 2.2. The subdifferential $\partial(\|\cdot\|_U)(\cdot)$ can be characterized as follows, *cf.*, *e.g.*, Chapter 1, Proposition 3.4 of [3]. For any $w \in U$,

$$\partial(\|\cdot\|_U)(w) := \begin{cases} \{w' \in U' \,:\, \langle w', w\rangle_{U',U} = \|w\|_U,\ \|w'\|_{U'} = 1\} & \text{if } w \neq 0, \\ \{w' \in U' \,:\, \|w'\|_{U'} = 1\} & \text{if } w = 0. \end{cases} \tag{2.1}$$

This characterization allows us to translate the above formulation of Theorem 2.1 directly into the formulation found in [36]. In [29] the same theorem is stated in terms of the so-called duality mapping, which can also be easily translated into the above formulation.

First, we will use Theorem 2.1 to characterize best approximations in subspaces of $L^q(\Omega)$, $1 < q < \infty$. To this end, we determine the subdifferential $\partial\big(\|\cdot\|_{L^q(\Omega)}\big)(w)$ for an arbitrary $w \in L^q(\Omega)$ and $1 < q < \infty$. Note that in this case the norm is Gâteaux differentiable; indeed, we can compute for $w \not\equiv 0$:

$$\begin{aligned} \partial\big(\|\cdot\|_{L^q(\Omega)}\big)(w)(v) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_\Omega |w + tv|^q \,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{q}}\Bigg|_{t=0} \\ &= \|w\|_{L^q(\Omega)}^{1-q} \int_\Omega \operatorname{sgn}(w)|w|^{q-1} v \,\mathrm{d}\boldsymbol{x}, \end{aligned}$$

where

$$\operatorname{sgn}(w(\boldsymbol{x})) = \begin{cases} -1 & \text{if } w(\boldsymbol{x}) < 0, \\ 1 & \text{if } w(\boldsymbol{x}) > 0, \\ 0 & \text{if } w(\boldsymbol{x}) = 0. \end{cases}$$

Hence, $\partial\big(\|\cdot\|_{\mathrm{L}^q(\Omega)}\big)(w) = \|w\|_{\mathrm{L}^q(\Omega)}^{1-q}\mathrm{sgn}(w)|w|^{q-1}$ by the canonical identification of an element in the dual space of $\mathrm{L}^q(\Omega)$ with a function in $\mathrm{L}^{q'}(\Omega)$, where $1 = 1/q + 1/q'$. The following corollary is an immediate consequence of this by setting $w = u - u_h$.

Corollary 2.3 (Characterization of L^q-best approximation)**.** *Let $U := \mathrm{L}^q(\Omega)$ and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^q-best approximation of u if and only if*

$$\int_\Omega \mathrm{sgn}(u-u_h)|u-u_h|^{q-1}v_h \,\mathrm{d}\boldsymbol{x} = 0 \qquad \forall v_h \in U_h. \tag{2.2}$$

Next we will use (2.1) to characterize best approximations in subspaces of $\mathrm{L}^1(\Omega)$. Note that in this case the subdifferential $\partial\big(\|\cdot\|_{\mathrm{L}^1(\Omega)}\big)(w)$ is in general not single valued for an arbitrary $w \in \mathrm{L}^1(\Omega)$. Since the dual space of $\mathrm{L}^1(\Omega)$ is isomorphic with $\mathrm{L}^\infty(\Omega)$, any $w' \in [\mathrm{L}^1(\Omega)]'$ can be identified with some $\psi \in \mathrm{L}^\infty(\Omega)$ such that $w'(v) = \int_\Omega \psi v \,\mathrm{d}\boldsymbol{x}$ for all $v \in \mathrm{L}^1(\Omega)$. From (2.1), we deduce that all $\psi \in \mathrm{L}^\infty(\Omega)$ that can be identified with an element $\partial\big(\|\cdot\|_{\mathrm{L}^1(\Omega)}\big)(w)$ are characterized by the following properties

(1) $\|\psi\|_{\mathrm{L}^\infty(\Omega)} = 1$.
(2) $\int_\Omega \psi w \mathrm{d}\boldsymbol{x} = \|w\|_{\mathrm{L}^1(\Omega)}$.

It is easy to see that any ψ such that $\psi = \mathrm{sgn}(w)$ if $w \neq 0$ and $|\psi| \leq 1$ almost everywhere satisfies the above conditions. Conversely, the first property implies $|\psi(\boldsymbol{x})| \leq 1$ almost everywhere and the second property implies that $\psi(\boldsymbol{x}) = 1$ almost everywhere on $\{u(\boldsymbol{x}) > 0\}$ and $\psi(\boldsymbol{x}) = -1$ almost everywhere on $\{u(\boldsymbol{x}) < 0\}$ since

$$\|w\|_{\mathrm{L}^1(\Omega)} = \int_\Omega |w| \,\mathrm{d}\boldsymbol{x} = \int_\Omega \psi w \,\mathrm{d}\boldsymbol{x} = \int_{\Omega\cap\{w(\boldsymbol{x})>0\}} \psi|w| \,\mathrm{d}\boldsymbol{x} - \int_{\Omega\cap\{w(\boldsymbol{x})<0\}} \psi|w| \,\mathrm{d}\boldsymbol{x}.$$

It is important to note, that the only condition on ψ on the set $\{w(\boldsymbol{x}) = 0\}$ is that $|\psi| \leq 1$ almost everywhere. The following corollary characterizing L^1-best approximations is a direct consequence of this by setting $w = u - u_h$.

Corollary 2.4 (Characterization of L^1-best approximation)**.** *Let $U := \mathrm{L}^q(\Omega)$ and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^1-best approximation of u if and only if there exists a function $\psi_0 \in \mathrm{L}^\infty(\Omega\cap\{u(\boldsymbol{x}) = u_h(\boldsymbol{x})\})$, $|\psi_0| \leq 1$, almost everywhere, such that for all $v_h \in U_h$ $0 = \int_\Omega \psi v_h \,\mathrm{d}\boldsymbol{x}$, where $\psi = \mathrm{sgn}(u - u_h)$ on $\{u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})\}$ and $\psi = \psi_0$ on $\{u(\boldsymbol{x}) = u_h(\boldsymbol{x})\}$.*

Note that in the case that u and u_h only agree on a set of measure zero, the choice of $\psi_0 \in [-1, 1]$ becomes irrelevant.

Remark 2.5 (Properties of the L^q-best-approximation operator)**.**

(1) If $q > 1$, the L^q-best approximation is always unique and hence the best-approximation operator is continuous, *cf.*, *e.g.*, Theorem 5.4 of [36].
(2) If $q = 1$, one does not in general have uniqueness of the L^1-best approximation operator. However, for $d = 1$, u continuous and $\Omega = [a, b] \subset \mathbb{R}$ the L^1-best approximation is unique if U_h is a spline space, *cf.*, [31].
(3) Uniqueness of the L^1-best approximation can be obtained by considering the so-called *natural* L^1-best approximation. Let $M_1(u)$ be the set of L^1-best approximations of $u \in U$. Then $u_h \in M_1(u)$ is called the natural L^1-best approximation if there exists $1 < \tilde{q} = \tilde{q}(u)$ such that

$$\|u - u_h\|_{\mathrm{L}^q(\Omega)} < \|u - w_h\|_{\mathrm{L}^q(\Omega)} \qquad \text{for all } 1 < q \leq \tilde{q} \text{ and all } u_h \neq w_h \in M_1(u).$$

In [21] it was proven that the natural L^1-best approximation exists, is unique and that the L^q-best approximation converges strongly in $\mathrm{L}^1(\Omega)$ to the natural L^1-best approximation in the limit $q \to 1$.

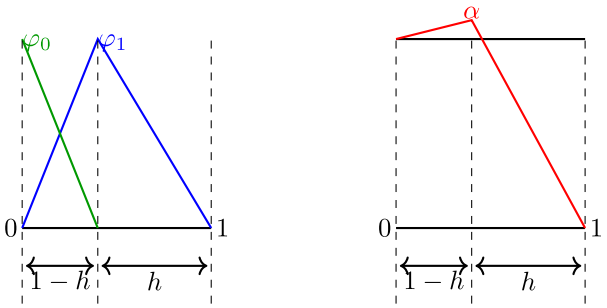


FIGURE 6. Proof of Lemmas 3.1 and 3.3. *Left*: hat functions φ_0 and φ_1. *Right*: approximation u_h with $\alpha > 1$.

3. Best approximation of a boundary discontinuity in one dimension

In this section we consider the best approximation problem (1.3) in one dimension and provide a proof of Theorem 1.1 in Section 3.3. Before we address the general case of an N-element mesh, we start with the special case $N = 2$. In Section 3.1 we determine the L^1-best approximation and in Section 3.2 the L^q-best approximation for $1 < q < \infty$.

3.1. L^1-best approximation

Lemma 3.1 (L^1-best approximation on a two-element mesh)**.** *Consider $\Omega = (0,1)$ and the mesh given by the two subintervals $(0, 1-h)$, $(1-h, 1)$. The L^1-best approximation of $u \equiv 1$ on $(0,1)$ by a continuous piecewise linear function u_h satisfying the boundary conditions $u_h(0) = 1$ and $u_h(1) = 0$ is uniquely determined by $u_h(1-h) = \alpha$, where $\alpha = 1$ if $h \le 0.5$ and $\alpha = \sqrt{2h}$ otherwise.*

Proof. We can write $u_h = \varphi_0 + \alpha\varphi_1$, where α is to be determined and

$$\varphi_0 = \begin{cases} \frac{(1-h)-x}{1-h} & \text{in } [0, 1-h], \\ 0 & \text{else,} \end{cases} \qquad \varphi_1 = \begin{cases} \dfrac{x}{1-h} & \text{in } [0, 1-h], \\ \dfrac{1-x}{h} & \text{in } [1-h, 1]. \end{cases}$$

Figure 6 shows the two functions φ_0 and φ_1 as well as an approximation u_h of $u \equiv 1$ with $\alpha > 1$. For consistency with Theorem 1.1, we could define the subspace U_h as the span of φ_1 and redefine $u = 1 - \varphi_0$ and $u_h = \alpha\varphi_1$. Note, however, that $u - u_h$ remains the same. The main consequence of this observation is, that the optimality conditions in Corollaries 2.3 and 2.4 do not have to be satisfied for $w_h = \varphi_0$ due to the boundary condition constraint.

We will now prove that the condition in Corollary 2.4 can only be satisfied if u_h is given as defined in Lemma 3.1. First note that $\alpha < 1$ can never be an L^1-best approximation. Indeed, in this case $u - u_h > 0$ almost everywhere and hence ψ is uniquely determined by $\operatorname{sgn}(u - u_h) = 1$. With $\varphi_1 > 0$ almost everywhere, we obtain $\int_0^1 \psi\varphi_1 \, dx > 0$. If $\alpha = 1$, then $u - u_h = 0$ in $[0, 1-h]$ and thus ψ is not uniquely determined in this subinterval. If $h \le 0.5$, we can choose $\psi_0 = h/(1-h)$ on $[0, 1-h]$ which satisfies $|\psi_0| \le 1$ if and only if $h \le 0.5$. A direct computation shows that for this choice of ψ_0 we have $\int_0^1 \psi\varphi_1 \, dx = 0$. On the other hand if $h > 0.5$, we use the fact that ψ_0 must satisfy $\psi_0 \ge -1$ almost everywhere, which implies $\int_0^1 \psi\varphi_1 \ge h/2 - (1-h)/2 > 0$ and hence $\alpha = 1$ cannot be an L^1-best approximation in this case. Finally, if $\alpha > 1$, the set $\{x \in (0,1) : u(x) = u_h(x)\}$ is a null set and hence ψ is uniquely determined by $\operatorname{sgn}(u - u_h)$. In this case, $u - u_h < 0$ in $(0, 1 - \vartheta h)$ and

$u - u_h > 0$ in $(1 - \vartheta h, 1)$, where $\vartheta = 1/\alpha \in (0, 1)$; we compute

$$\int_0^1 \operatorname{sgn}(u - u_h)\varphi_1 \mathrm{d}x = -\int_0^{1-\vartheta h} \varphi_1 \mathrm{d}x + \int_{1-\vartheta h}^1 \varphi_1 \mathrm{d}x = -\frac{1}{2} + \vartheta^2 h. \tag{3.1}$$

This integral becomes 0 for $\vartheta = 1/\sqrt{2h} \iff \alpha = \sqrt{2h}$. Note that this only yields an L^1-best approximation if $h > 1/2$. Indeed, if $h \leq 1/2$, then $\alpha = \sqrt{2h} \leq 1$, but we have assumed $\alpha > 1$ and have considered the case $\alpha \leq 1$ separately. □

Remark 3.2. Equation (3.1) shows that the optimality condition only depends on the point where u and u_h intersect in $[1 - h, 1)$. Therefore, the same argument can be applied to any u that is piecewise linear and approximated by a piecewise linear function u_h such that the boundary conditions are constraint to satisfy $u(0) = u_h(0)$ and $u(1) \neq u_h(1)$. The exact value of $u_h(1 - h)$ depends on u and $u_h(1)$, but ϑ does not. Hence, the ϑ determined in the proof of Lemma 3.1 can be used to determine $u(1 - h)$ in this more general setting as well.

3.2. L^q-best approximation

Lemma 3.3 (L^q-best approximation on a two-element mesh)**.** *Consider $\Omega = (0, 1)$ and the mesh given by the two subintervals $(0, 1 - h)$, $(1 - h, 1)$. The L^q-best approximation of $u \equiv 1$ on $(0, 1)$ by a continuous piecewise linear function u_h satisfying the boundary conditions $u_h(0) = 1$ and $u_h(1) = 0$ is given by $u_h = \varphi_0 + \alpha\varphi_1$, where $\alpha > 1$ and $0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q + h$. Furthermore, the L^q-best approximation converges pointwise to the L^1-best approximation in the limit $q \to 1$ and $\alpha(q)$ is increasing function in q.*

Proof. We use the same characterization of the function u_h and the basis functions φ_0 and φ_1 as in the proof of Lemma 3.1. This proof relies on the characterization of the L^q-best approximation given in Corollary 2.3. If $\alpha \leq 1$, we have $u - u_h \geq 0$ in $(0, 1 - h)$ and $u - u_h > 0$ in $(1 - h, 1)$. Thus,

$$\int_0^1 \operatorname{sgn}(u - u_h)|u - u_h|^{q-1}\varphi_1 \, \mathrm{d}x \geq \int_{1-h}^1 \operatorname{sgn}(u - u_h)|u - u_h|^{q-1}\varphi_1 \, \mathrm{d}x > 0,$$

hence $\alpha \leq 1$ is not possible. We can therefore assume $\alpha > 1$. In this case $u - u_h < 0$ in $(0, (\alpha-h)/\alpha)$ and $u - u_h > 0$ in $((\alpha-h)/\alpha, 1)$. A direct computation yields

$$\int_0^1 \operatorname{sgn}(u - u_h)|u - u_h|^{q-1}\varphi_1 \, \mathrm{d}x \quad = \frac{h - (1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q}{\alpha^2 q(q+1)}.$$

Hence, the L^q-best approximation can be determined by finding $\alpha_q > 1$ satisfying $f(\alpha_q, q) = 0$, where

$$f(\alpha, q) = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q + h.$$

Both existence and uniqueness of α_q are guaranteed since the L^q-best approximation always exists and is unique for $q > 1$. To see that $\alpha_q < 2$ for any $q \in (1, \infty)$, note that for fixed $q > 1$ and any $\alpha > 1$, f is strictly decreasing in α and $f(2, q) = -4(1-h)q - h(2q+1) + h < -4(1-h)q - 2hq < 0$. To see how α_q varies with respect to q, we first compute

$$\begin{aligned}\frac{\partial f}{\partial q}(\alpha, q) &= -(1-h)\alpha^2(\alpha-1)^{q-1} - h\alpha(\alpha-1)^q \\ &\quad - (1-h)\alpha^2 q(\alpha-1)^{q-1}\ln(\alpha-1) - h(\alpha q+1)(\alpha-1)^q \ln(\alpha-1) \\ &= \frac{f(\alpha, q) - h[1 - (\alpha-1)^q]}{q} + \ln(\alpha-1)(f(\alpha, q) - h).\end{aligned}$$

Using $f(\alpha_q, q) = 0$, we obtain

$$\frac{\partial f}{\partial q}(\alpha_q, q) = -h\left(\frac{[1-(\alpha-1)^q]}{q} + \ln(\alpha-1)\right) > 0,$$

since $g(x) = -(1-x^q)/q + \ln(x)$ satisfies $g'(x) < 0$ for $x \in (0,1)$ and $g(1) = g'(1) = 0$. By continuity of f and $\partial f/\partial_q$, there exists $\varepsilon > 0$ for any $q \in (1,\infty)$ such that for $\tilde{q} \in (q-\varepsilon, q+\varepsilon)$, $\partial f/\partial q(\alpha_q, \tilde{q}) > 0$. Hence, for $q_1 \in (q-\varepsilon, q)$ and $q_2 \in (q, q+\varepsilon)$, we obtain $f(\alpha_q, q_1) < 0 < f(\alpha_q, q_2)$. Recalling that f is strictly decreasing in α and that $f(2, q_i) < 0$, we have that $\alpha_{q_1} \in (1, \alpha_q)$ and $\alpha_{q_2} \in (\alpha_q, 2)$. Therefore, α_q is strictly increasing in q. Let $\{q_k\}_{k=0}^{\infty} \subset (1,\infty)$ be a decreasing sequence with $\lim_{k\to\infty} q_k = 1$. Then, $\{\alpha_{q_k}\}_{k=0}^{\infty}$ is a monotonically decreasing sequence in $[1,2]$ and hence converges to some limit $\tilde{\alpha}$. It remains to be shown that $\tilde{\alpha} = \alpha_1$. If $\tilde{\alpha} > 1$, $f(\alpha, q)$ is continuous on $[\tilde{\alpha}, 2] \times [1,\infty)$ and

$$0 = \lim_{k\to\infty} f(\alpha_{q_k}, q_k) = f(\tilde{\alpha}, 1) = -\tilde{\alpha}^2 + 2h.$$

This implies hat if $\tilde{\alpha} > 1$, h must satisfy $h > 0.5$ and $\tilde{\alpha} = \sqrt{2h}$. From Lemma 3.1 if follows that, if $h > 0.5$, we have $\alpha_1 = \sqrt{2h}$. Hence, $\alpha_1 = \tilde{\alpha}$ if $h \geq 0.5$. Otherwise, $h \leq 0.5$ and $\tilde{\alpha} = \alpha_1 = 1$, which concludes the proof. □

3.3. Sufficient conditions on general meshes

In this section we provide a proof of Theorem 1.1. To this end, let the mesh be given by a subdivision of the interval $(0,1)$ into $N \geq 2$ subintervals (x_{i-1}, x_i), $i = 1,\dots,N$, with $0 = x_0 < x_2 < \cdots < x_{N-1} < x_N = 1$. The length h_i of the ith subinterval is given by $h_i = x_i - x_{i-1}$, $i = 1,\dots N$. Furthermore, denote by φ_i the continuous and piecewise linear function that satisfies $\varphi_i(x_j) = \delta_{ij}$. Let $U = \operatorname{span}\{\varphi_i : 0 \leq i \leq N\}$. We show that the following conditions are sufficient for the existence of an L^1-best approximation $u_h \in U$ of $u \in U$ subject to the constraint $u_h(0) = u(0)$ and $u_h(1) = g \neq u(1)$ with no over- or undershoots, *i.e.*, $u_h(x_i) = u(x_i)$, $i = 1,\dots,N-1$:

$$h_i \geq \left(2(1-\vartheta_{i+1})^2 - 1\right)h_{i+1}, \qquad \text{for } i = M, M+1, \dots, N-1, \tag{3.2}$$

where

$$\begin{aligned}
\vartheta_N &:= 0,\\
\vartheta_i^2 &:= \frac{1}{2}\left(1 - \left(2(1-\vartheta_{i+1})^2 - 1\right)\frac{h_{i+1}}{h_i}\right), i = N-1,\dots,1,\\
M &:= \max\left(\{0\} \cup \left\{i \in \{1,\dots,N-1\} : \vartheta_i \geq 1 - \frac{1}{\sqrt{2}}\right\}\right).
\end{aligned}$$

Remark 3.4. (1) Note that it is sufficient to prove the result for the case $u_h(1) = u_h(0) = u(0) = 0$. Indeed, for $\tilde{u} = u - u_h(0)\varphi_0 - u_h(1)\varphi_N$ and $\tilde{u}_h = u_h - u_h(0)\varphi_0 - u_h(1)\varphi_N$, we have $u - u_h = \tilde{u} - \tilde{u}_h$ and $\tilde{u}(0) = \tilde{u}_h(0) = \tilde{u}_h(1) = 0$.

(2) The condition (3.2) was constructed as follows: Given ψ on $[x_i, 1]$, we can define $\widetilde{\psi}_0\big|_{[x_{i-1},x_i]} := -2\int_{x_i}^{x_{i+1}} \psi\varphi_i \,\mathrm{d}x/h_i$ such that $0 = \int_{x_{i-1}}^{x_i} \widetilde{\psi}_0\varphi_i \,\mathrm{d}x + \int_{x_i}^{x_{i+1}} \psi\varphi_i \,\mathrm{d}x$. Therefore, $\widetilde{\psi}_0$ would be a valid choice for ψ_0 in $[x_{i-1}, x_i]$ if $h_i \geq \left|2\int_{x_i}^{x_{i+1}} \psi\varphi_i \,\mathrm{d}x\right|$. In the next step, we replace $\widetilde{\psi}_0$ by another choice for ψ_0 that yields a weaker condition of this form for h_{i-1} than simply using $\widetilde{\psi}_0$. To achieve this, we split the interval (x_{i-1}, x_i)

into two parts, $(x_{i-1}, x_{i-1}+\vartheta_i h_i)$ and $(x_{i-1}+\vartheta_i h_i, x_i)$, and define $\hat{\psi}_0 = -\mathrm{sgn}\big(\widetilde{\psi}_0\big)$ in $(x_{i-1}, x_{i-1}+\vartheta_i h_i)$ and $\hat{\psi}_0 = \mathrm{sgn}\big(\widetilde{\psi}_0\big)$ in $(x_{i-1}+\vartheta_i h_i, x_i)$, where we choose $\vartheta_i \in (0,1)$ such that

$$\int_{x_{i-1}}^{x_i} \hat{\psi}_0 \varphi_i \,\mathrm{d}x = -\mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}}^{x_{i-1}+\vartheta_i h_i} \varphi_i \,\mathrm{d}x + \mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}+\vartheta_i h_i}^{x_i} \varphi_i \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \widetilde{\psi}_0 \varphi_i \,\mathrm{d}x.$$

It can then be shown that the following cases can occur:

(a) $\big|\widetilde{\psi}_0\big| = 1$ and $\vartheta_i = 0$.

(b) $\big|\widetilde{\psi}_0\big| < 1$ and $\left|\int_{x_{i-1}}^{x_i} \hat{\psi}_0 \varphi_{i-1} \,\mathrm{d}x\right| < \left|\int_{x_{i-1}}^{x_i} \widetilde{\psi}_0 \varphi_{i-1} \,\mathrm{d}x\right|$.

(c) $\big|\widetilde{\psi}_0\big| < 1$ and there exists $\alpha \in (0,1)$ such that

$$\begin{aligned} -\mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}}^{x_{i-1}+\vartheta_{i+1} h_i} \alpha \varphi_i \,\mathrm{d}x + \mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}+\vartheta_i h_i}^{x_i} \varphi_i \,\mathrm{d}x &= \int_{x_{i-1}}^{x_i} \widetilde{\psi}_0 \varphi_i \,\mathrm{d}x, \\ -\mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}}^{x_{i-1}+\vartheta_{i+1} h_i} \alpha \varphi_{i+1} \,\mathrm{d}x + \mathrm{sgn}\big(\widetilde{\psi}_0\big) \int_{x_{i-1}+\vartheta_i h_i}^{x_i} \varphi_{i+1} \,\mathrm{d}x &= 0. \end{aligned}$$

This construction yields the function ψ_α defined in the proof of Theorem 1.1 below with ϑ_i as defined above.

Proof of Theorem 1.1. Define u_h such that $u_h(1) = g$ and $u_h(x_i) = u(x_i)$ for all $i = 0, \ldots, N-1$. Note that then $\mathrm{sgn}(u - u_h) = 0$ in $[0, x_{N-1}]$ and hence ψ as in Corollary 2.4 is not uniquely determined in this subinterval. Therefore, we need to construct ψ_0 such that the conditions in Corollary 2.4 are satisfied. In $[x_{N-1}, x_N]$, we have $u - u_h \neq 0$. Without loss of generality, we can assume $\mathrm{sgn}(u - u_h) = 1$ in $[x_{N-1}, x_N]$. Indeed, for $\mathrm{sgn}(u - h_h) = -1$ the conclusion follows from using the same function ψ_0 with the opposite sign. For $\alpha \in (0,1]$, define $\psi_\alpha(x)$ as follows:

$$\psi_\alpha(x) = \begin{cases} (-1)^{N-i+1} & x \in (x_{i-1}, x_{i-1}+\vartheta_i h_i), \quad \text{for all } i = M+1, \ldots N, \\ (-1)^{N-i} & x \in (x_{i-1}+\vartheta_i h_i, x_i) \quad \text{for all } i = M+1, \ldots N, \\ (-1)^{N-M+1}\alpha & x \in (x_{M-1}, x_{M-1}+\tilde{\vartheta}_M h_M) \quad \text{if } M > 0, \\ (-1)^{N-M} & x \in (x_{M-1}+\tilde{\vartheta}_M h_M, x_M) \quad \text{if } M > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\tilde{\vartheta}_M^2 = \frac{2}{1+\alpha}\vartheta_M^2 = \frac{1}{\alpha+1}\left(1 - \big(2(1-\vartheta_{M+1})^2 - 1\big)\frac{h_{M+1}}{h_M}\right).$$

We claim that there exists $\tilde{\alpha} \in (0,1]$ such that $\int_0^1 \psi_{\tilde{\alpha}}(x)\varphi_i(x)\,\mathrm{d}x = 0$ for all $i = 1, \ldots, N-1$, if (3.2) is satisfied. The theorem follows from defining $\psi_0 := \psi_{\tilde{\alpha}}$ in $[0, x_{N-1}]$. Note that $\vartheta_N = 0$ implies $\psi_\alpha(x) = 1 = \mathrm{sgn}(u - u_h)$ in (x_{N-1}, x_N) and that $\|\psi_\alpha\|_{\mathrm{L}^\infty((0,1))} = 1$. It is easy to verify that $\int_0^1 \psi_\alpha \varphi_i \,\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} \psi_\alpha \varphi_i \,\mathrm{d}x = 0$ for all $i \neq M-1$. In particular, this is true for $i = M$ and any $\alpha \in (0,1]$. A direct computation shows, that $\int_{x_{M-1}}^{x_M} \psi_\alpha \varphi_{M-1} \,\mathrm{d}\boldsymbol{x} = 0$ if and only if α satisfies

$$\left(1 - \tilde{\vartheta}_M\right)^2 = \frac{\alpha}{1+\alpha} \Leftrightarrow \tilde{\vartheta}_M = \frac{1}{\sqrt{1+\alpha}}\left(\sqrt{1+\alpha} - \sqrt{\alpha}\right).$$

Hence, we need $\tilde{\alpha}$ such that

$$\left(\sqrt{1+\tilde{\alpha}} - \sqrt{\tilde{\alpha}}\right) = \sqrt{\left(1 - \left(2(1-\vartheta_{M+1})^2 - 1\right)\frac{h_{M+1}}{h_M}\right)} = \sqrt{2}\vartheta_M. \tag{3.3}$$

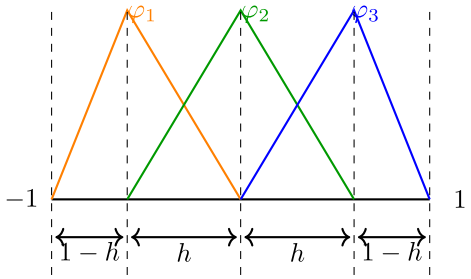


FIGURE 7. Proof of Theorem 1.2. Basis for U_h.

For $\alpha > 0$, $g(\alpha) = \sqrt{1+\alpha} - \sqrt{\alpha}$, is a strictly decreasing function of α and thus bijectively maps $(0,1]$ onto $[\sqrt{2}-1, 1)$. The equation (3.3) therefore has a unique solution $\tilde{\alpha} \in (0,1]$ if and only if $\sqrt{2}\vartheta_M \in [\sqrt{2}-1, 1) \iff \vartheta_M \in [1 - 1/\sqrt{2}, 1/\sqrt{2})$. By the definition of M, we have $\vartheta_M \geq 1 - 1/\sqrt{2}$ and $\vartheta_{M+1} < 1 - 1/\sqrt{2} \Rightarrow \vartheta_M < 1/\sqrt{2}$. □

Corollary 3.5 (A simple sufficient condition in one dimension)**.** *Let N, x_i, h_i, φ_i and U be defined as in Theorem 1.1. Then a sufficient condition for the existence of an L^1-best approximation $u_h \in U$ of $u \in U$ subject to the constraint $u_h(0) = u(0)$ and $u_h(1) = g \neq u(1)$ with no over- or undershoots is given by $h_N \leq \min_{i=1,\dots,N-1} h_i$.*

Proof. This can either be proven by showing that the condition in Theorem 1.1 is satisfied, or, by simply defining $\psi_0(x) \equiv \operatorname{sgn}(u(1) - u_h(1))(-1)^{j h_N / h_{N-j}}$ on (x_{N-j-1}, x_{N-j}) for $j = 1, \dots, N-1$. □

Remark 3.6. With very similar arguments, it is easy to see that if $h_N > h_{N-1}$, but $h_{N-1} \leq h_i$ for all $i = 1, \dots, N-2$, then every L^1-best approximation must contain over- or undershoots. Moreover, there exists an L^1-best approximation with overshoot only at the node x_{N-1} and no further over- or undershoots, *i.e.*, $u_h(x_i) = u(x_i)$ for $i = 1, \dots, N-2$ and, if additionally $u \equiv 1$, $u_h(x_{N-1}) = \sqrt{2h_N/(h_N + h_{N-1})}$. The value at $u_h(x_{N-1})$ follows from the proof in the case $N = 2$ and a rescaling of the interval.

4. Over- and undershoots at jump discontinuities

In this section we consider the L^q-best approximation of $u(x) = \operatorname{sgn}(x)$ in $(-1,1)$ as an example of a jump discontinuity in the interior of the domain and provide a proof of Theorem 1.2. We seek an L^q-best approximation of this function by a continuous piecewise linear function on the mesh consisting of $(-1,-h)$, $(-h,0)$, $(0,h)$ and $(h,1)$. We fix the boundary conditions at -1 and 1, *i.e.*, $u_h(1) = u(1) = 1$ and $u_h(-1) = u(-1) = -1$. The finite dimensional approximation space U_h is given by the span of the hat functions φ_i, $i = 1, 2, 3$, depicted in Figure 7.

We split the proof into three parts: in Section 4.1, we consider the case where the L^1-best approximation does not exhibit Gibbs phenomena, *i.e.*, the case when $h \leq 0.5$. We will also discuss at the end of Section 4.1, how this result implies a more general result similar to Theorem 1.1. In Section 4.2, we consider the case where the L^1-best approximation does exhibit Gibbs phenomena, *i.e.*, the case when $h > 0.5$; finally, in Section 4.3 we consider the L^q-best approximation for $1 < q < \infty$ and the limit as $q \to 1$.

4.1. L^1-best approximation without over- or undershoots

In this section we prove that if $h \leq 0.5$, a continuous piecewise linear function u_h on the mesh shown in Figure 7 such that $-u_h(-1) = u_h(1) = 1$ is an L^1-best approximation of $u(x) = \operatorname{sgn}(x)$ if and only if $u_h(0) = \beta$, with $\beta \in [-1,1]$ arbitrary, and $-u_h(-h) = u_h(h) = 1$. The approximation u_h is shown in Figures 8a–8c for $\beta = -1, 0, 1$.

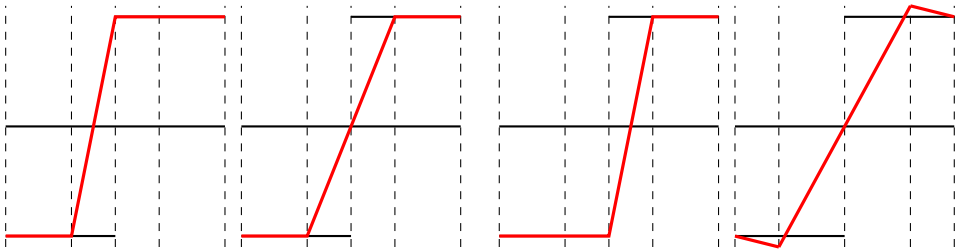


FIGURE 8. L^1-best approximation of a jump discontinuity, *cf.*, Theorem 1.2. (A) $u_h(0) = 1$. (B) $u_h(0) = 0$. (C) $u_h(0) = -1$. (D) $h > 1/2$.

Proof of Theorem 1.2 for $q = 1$ and $h \leq 0.5$. First note that u_h satisfying $u_h(-h) \neq -1$ cannot be an L^1-best approximation since in this case $\operatorname{sgn}(-1 - u(-h))(u - u_h) > 0$ in $(-1, -h + \delta)$, with $\delta > 0$. Thus, $\operatorname{sgn}(-1 - u(-h)) \int_{-1}^{1} \psi\varphi_1 \, \mathrm{d}x > \int_{-1}^{-h} \varphi_1 \, \mathrm{d}x - \int_{-h}^{0} \varphi_1 \, \mathrm{d}x \geq 0$ since $h \leq 0.5$. By an analogous argument, it also holds that u_h satisfying $u_h(h) \neq 1$ cannot be an L^1-best approximation. Then, since $u_h(\pm h) = \pm 1$, we have that $\beta \notin [-1, 1]$ implies $\psi = \operatorname{sgn}(u - u_h) = -\operatorname{sgn}(\beta)$ in $(-h, h)$ and hence $\int_{-1}^{1} \psi\varphi_2 \, \mathrm{d}x = -\operatorname{sgn}(\beta)h \neq 0$, which implies that u_h satisfying $u_h(0) \notin [-1, 1]$ cannot be an L^1-best approximation. In order to show that $-u_h(-h) = u_h(h) = 1$ and $u(0) = \beta$ is indeed an L^1-best approximation for any $\beta \in [-1, 1]$, we need to find ψ satisfying the conditions in Corollary 2.4. Since ψ is uniquely defined by $\operatorname{sgn}(u - u_h)$ whenever $u(x) \neq u_h(x)$, we only need to choose a suitable ψ_0 on $\{x \in [-1, 1] : u(x) = u_h(x)\}$ and verify that all conditions in Corollary 2.4 are satisfied. To this end, we choose $\psi_0 = h/(1-h)$ in $(-1, -h)$ and $\psi_0 = -h/(1-h)$ in $(h, 1)$. If $\beta = 1$, we furthermore choose $\psi_0 = 1$ in $(0, h)$. Analogously, we choose $\psi = -1$ in $(-h, 0)$ if $\beta = -1$. With these choices, we can easily verify that $\int_{-1}^{1} \psi\varphi_i \, \mathrm{d}x = 0$ for $i = 1, 2, 3$. □

Remark 4.1 (General meshes)**.**

(1) Note that we have shown that there is a whole family of L^1-best approximations with no over- or undershoots for this particular example if $h \leq 1/2$. The situation is quite different if we instead consider a non-symmetric subdivision of the interval $(-1, 1)$ into $(-1, -h_1)$, $(-h_1, 0)$, $(0, h_2)$ and $(h_2, 1)$ with $h_1 \neq h_2$. The integral involving φ_2 then implies that the case $-1 < u_h(0) < 1$ does not yield an L^1-best approximation; the case $u_h(0) = 1$ is an L^1-best approximation if and only if $h_1 < h_2 \leq 1/2$, and the case $u_h(0) = -1$ is an L^1-best approximation if and only if $h_2 < h_1 \leq 1/2$.

(2) It is by no means necessary that the jump discontinuity aligns with an element boundary. Indeed, consider the mesh $(-1, -\vartheta h)$, $(-\vartheta h, (1 - \vartheta)h)$, $((1 - \vartheta)h, 1)$ with $\vartheta \in (0, 1/2]$, *i.e.*, a three-element mesh, such that the middle element has length h and the jump is contained within the left half of this element. It can be verified, that $u_h(-\vartheta h) = -1$, $u_h((1 - \vartheta)h) = 1$ is an L^1-best approximation if and only if

$$h \leq \gamma(\vartheta) := \frac{1}{2 - \vartheta - 2\vartheta^2}.$$

Due to the symmetry of the problem, we obtain the condition $h \leq \gamma(1 - \vartheta)$, if $\vartheta \in [1/2, 1)$, *i.e.*, if the discontinuity is contained within the right half of the interval $(-\vartheta h, (1 - \vartheta)h)$. Note that $\gamma(\vartheta)$ is positive and monotonically increasing for $\vartheta \in (0, 1/2)$. Hence, we obtain the uniform bound $\gamma(\vartheta) \geq \gamma(0) = 1/2$. This yields a sufficient condition for the existence of an L^1-best approximation without over- or undershoots that is independent of the location of the jump discontinuity within the element. Moreover, $\gamma(\vartheta)$ is maximal if $\vartheta = 1/2$, *i.e.*, the mesh is symmetric with respect to 0. In this case the condition becomes $h \leq 1$. This in particular includes the case of a uniform mesh, *i.e.*, $h = 2/3$. We can also observe that the condition

for $\vartheta = 0$ and $\vartheta = 1$, *i.e.*, the discontinuity aligns with the element boundary, becomes $h \leq 1/2$ which is identical with the results in the proof of the first part of Theorem 1.2 above.

(3) Clearly, a sufficient condition for the existence of an L^1-best approximation on a general mesh similar to Theorem 1.1 can easily be derived by combining the results of Theorem 1.1, the above result and the first two points in this remark.

(4) Similar to Remark 3.2, the condition for the existence of an L^1-best approximation with no over- or undershoots is the same for any piecewise linear function u with a discontinuity at $x = 0$, but the magnitude of the overshoot does depend on u.

4.2. L^1-best approximation with over- and undershoots

In this section we prove that, if $h > 0.5$, a continuous piecewise linear function u_h on the mesh shown in Figure 7 such that $-u_h(-1) = u_h(1) = 1$ is an L^1-best approximation of $u(x) = \operatorname{sgn}(x)$ if and only if

$$u_h(-h) = \alpha := -\sqrt{2h} - \beta(\sqrt{2h} - 1) \tag{4.1a}$$

$$u_h(0) = \beta, \tag{4.1b}$$

$$u_h(h) = \gamma := \sqrt{2h} - \beta(\sqrt{2h} - 1), \tag{4.1c}$$

with $\beta \in [-1, 1]$ arbitrary.

Proof of Theorem 1.2 for $q = 1$ and $h > 0.5$. First note that if $u_h(0) = \beta = -1$, we require $u_h(-h) = \alpha = -1$, since otherwise $\operatorname{sgn}(u - u_h) \equiv \operatorname{sgn}(-1 - \alpha) \neq 0$ in $(-1, 0)$ and $\int_{-1}^{1} \psi\varphi_1 \,\mathrm{d}x = {}^{\operatorname{sgn}(-1-\alpha)}/_2 \neq 0$. Analogously, $\beta = 1 \Rightarrow u_h(h) = 1$. For any $\beta \neq -1$, the condition $\int_{-1}^{1} \psi\varphi_1 \,\mathrm{d}x = 0$ implies that α and β must satisfy (4.1a) which can be seen using analogous arguments to the proof of Lemma 3.1. Similarly, for any $\beta \neq 1$, we obtain – using the condition $\int_{-1}^{1} \psi\varphi_3 \,\mathrm{d}x = 0$ – that β and γ must satisfy (4.1c). It remains to be shown that $|\beta| \leq 1$ and that the optimality condition is also satisfied with $v_h = \varphi_2$. For the sake of contradiction, assume $\beta < -1$. We have already established that α, β and γ satisfy (4.1). Using this, we determine that u_h and u intersect at $x = \pm {}^{h}/_{\sqrt{2h}}$. Thus, we obtain

$$\int_{-1}^{1} \psi\varphi_2 \,\mathrm{d}x = h\Big(1 - 2(1 - {}^{1}/_{\sqrt{2h}})^2\Big) = 0 \iff (1 - {}^{1}/_{\sqrt{2h}})^2 = {}^{1}/_{2} \iff h = ({}^{1}/_{\sqrt{2}-1})^2 > 1.$$

This is a contradiction, since $h \in (0, 1)$. For $\beta > 1$, the sign of $u - u_h$ on $(-h, h)$ is exactly opposite compared to the case $\beta < -1$; therefore, it is easy to see that $\beta > 1$ also leads to a contradiction.

If on the other hand $\beta \in [-1, 1]$, we obtain, using the symmetry of φ_2,

$$\int_{-1}^{1} \psi\varphi_2 \,\mathrm{d}x = \underbrace{\int_{-h}^{-h/\sqrt{2h}} \varphi_2 \,\mathrm{d}x - \int_{h/\sqrt{2h}}^{h} \varphi_2 \,\mathrm{d}x}_{=0} + \underbrace{\int_{0}^{h/\sqrt{2h}} \varphi_2 \,\mathrm{d}x - \int_{-h/\sqrt{2h}}^{0} \varphi_2 \,\mathrm{d}x}_{=0} = 0.$$

Note that the above computation also applies to the cases $\beta = 1$ and $\beta = -1$ since it corresponds to a valid choice for ψ_0 in each case. □

4.3. L^q-best approximation

In this section, we prove the final part of Theorem 1.2. More precisely, we show that a continuous piecewise linear function u_h on the mesh shown in Figure 7 such that $-u_h(-1) = u_h(1) = 1$ is an L^q-best approximation of $u(x) = \operatorname{sgn}(x)$ for $1 < q < \infty$ if and only if $-u_h(-h) = u_h(h) = \alpha$ and $u_h = 0$, where α satisfies $0 = -(1 - h)\alpha^2 q(\alpha - 1)^{q-1} - h(\alpha q + 1)(\alpha - 1)^q + h$ and $\alpha > 1$. Furthermore, we show that in the limit $q \to 1$ the L^q-best approximation converges to the L^1-best approximation as defined in (4.1) with $\beta = 0$, for any $h \in (0, 1)$, *i.e.*, the corresponding L^1-best approximation is antisymmetric and satisfies $-u_h(-h) = u_h(h) = 1$ if $h \leq 0.5$ and $-u_h(-h) = u_h(h) = \sqrt{2h}$ otherwise.

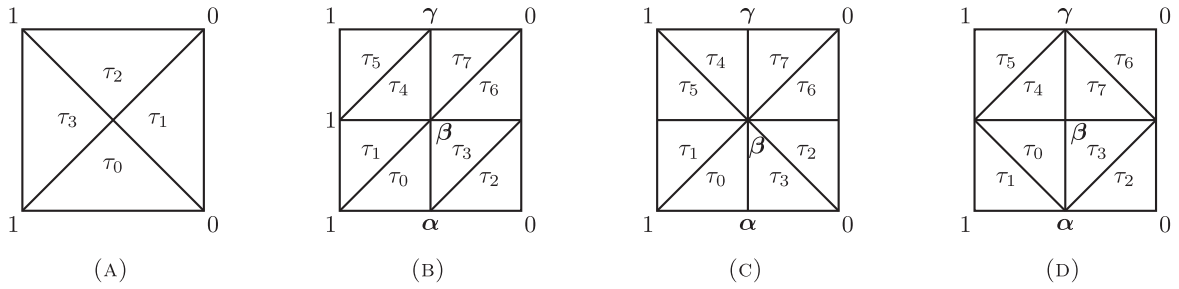


FIGURE 9. Four different meshes on $(0,1)^2$, *cf.*, Figure 4. (A) Mesh 1. (B) Mesh 2. (C) Mesh 3. (D) Mesh 4.

Proof of Theorem 1.2 for $q > 1$. We use the characterization of the L^q-best approximation in Corollary 2.3. Due to the uniqueness of the L^q-best approximation for $1 < q < \infty$ and the symmetry of the problem, we may assume that the L^q-best approximation is an odd function. This means that $u_h(0) = 0$ and $-u_h(-h) = u_h(h) = \alpha$ for some $\alpha \in \mathbb{R}$. It is easy to see that $\int_{-1}^{1} \operatorname{sgn}(u-u_h)|u-u_h|^{q-1}\varphi_2 \,\mathrm{d}x = 0$ for any choice of α and that $\int_{-1}^{1} \operatorname{sgn}(u-u_h)|u-u_h|^{q-1}\varphi_3 \,\mathrm{d}x = 0$ if and only if $\int_{-1}^{1} \operatorname{sgn}(u-u_h)|u-u_h|^{q-1}\varphi_1 \,\mathrm{d}x = 0$. To determine for which α the latter two integrals become zero, note that this is the same situation as in the example presented in Sections 3.1 and 3.2, only mirrored. Therefore, we again obtain that α satisfies $0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q + h$.

To show convergence to the L^1-best approximation with $\beta = 0$, first note that $u_h(0) = 0$ for all q and hence $\lim_{q\to 1} u_h(0) = 0$. From the proof of Lemma 3.3, we obtain that $\lim_{q\to 1} u(\pm h) = \pm 1$ if $h \le 0.5$ and $\lim_{q\to 1} u(\pm h) = \pm\sqrt{2h}$ otherwise. This is exactly the L^1-best approximation with $\beta = 1$. Therefore, in the limit we obtain the solution in Figure 8b if $h \le 1/2$. The corresponding L^1-best approximation for $h > 1/2$ is shown in Figure 8d. □

5. Best approximation of a boundary discontinuity in two dimensions

In this section we consider the best approximation problem (1.3) with $d = 2$. First, we consider the four meshes shown in Figure 9 and determine the best approximation of $u \equiv 1$ by a continuous function u_h that is a linear polynomial on each of the triangles and takes the following values in the four corners: $u_h(0,0) = u_h(0,1) = 1$ and $u_h(1,0) = u_h(1,1) = 0$. For all meshes except the first one, we additionally fix the boundary conditions $u_h(0,0.5) = 1$ and $u_h(1,0.5) = 0$.

The free parameter of the best approximation problem for the first mesh is $\alpha = u_h(0.5,0.5)$; there are three free parameters for each of the remaining meshes. For Meshes 2–4, we denote by v_1 the continuous piecewise linear function that is 1 at the node $(0.5,0)$ and 0 at all other nodes; by v_2 the continuous piecewise linear function that is 1 at $(0.5,0.5)$ and 0 at all other nodes; and by v_3 the continuous piecewise linear function that is 1 at $(0.5,1)$ and zero at all other nodes. The coefficients defining the solution u_h are denoted as follows

$$u(0.5,0) = \alpha, \qquad u(0.5,0.5) = \beta, \qquad u(0.5,1) = \gamma.$$

Furthermore, we prove a necessary condition on general conforming meshes and a condition similar to Theorem 1.1 on a structured but non-uniform mesh. The remainder of this section is organized as follows: in Section 5.1, we consider Mesh 1 with $q = 1$ and determine the unique L^1-best approximation; in Section 5.2 we continue with Mesh 1 and determine the L^q-best approximation for $1 < q < \infty$; in Section 5.3 we show that the L^q-best approximation contains over- or undershoots on all four meshes if $q > 1$ and consider the case $q = 1$ for Meshes 2, 3 and 4; in Section 5.4 we then consider more general meshes and prove a necessary and a sufficient condition for certain meshes.

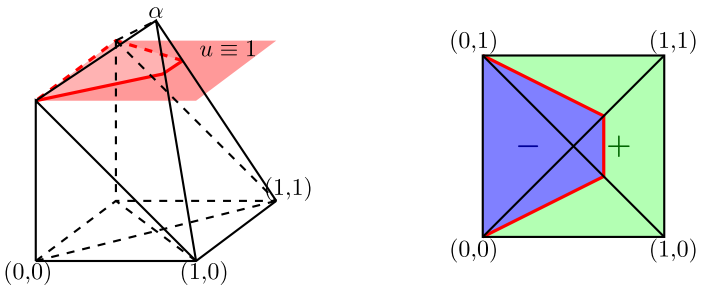


FIGURE 10. Proof of Lemma 5.1: the function u_h with $\alpha > 1$ and intersection with u is marked by red lines (*left*). The mesh with the area where $(u - u_h) < 0$ colored in blue and the area where $(u - u_h) > 0$ colored in green (*right*).

5.1. L^1-best approximation on Mesh 1

Lemma 5.1 (L^1-best approximation on Mesh 1)**.** *The L^1-best approximation of $u \equiv 1$ on Mesh 1 is unique and $u_h(0.5, 0.5) = \alpha$, where α satisfies $\alpha > 1$ and $0 = 2\alpha^3 - 5\alpha + 2$, hence $\alpha \approx 1.3200$.*

Proof. We again use the characterization of the L^1-best approximation in Corollary 2.4. The space U_h is the span of the continuous function v_h that is a linear polynomial on each element, zero at the boundary of the domain and 1 at the centroid $(0.5, 0.5)$. To see that α must satisfy $\alpha > 1$, note that if $\alpha \leq 1$, we have $u - u_h > 0$ in τ_i, $i = 0, 1, 2$ and $\psi \geq -1$ in τ_3. Hence, $\int_{(0,1)^2} \psi v_h \,\mathrm{d}\boldsymbol{x} \geq \sum_{i=0}^{2} \int_{\tau_i} v_h \,\mathrm{d}\boldsymbol{x} - \int_{\tau_3} v_h \,\mathrm{d}\boldsymbol{x} = 1/6 > 0$ and this cannot be an L^1-best approximation.

If $\alpha > 1$, ψ is uniquely determined by $\operatorname{sgn}(u - u_h)$ (see Figure 10) and a direct computation yields

$$\int_{(0,1)^2} \operatorname{sgn}(u - u_h) v_h \,\mathrm{d}\boldsymbol{x} = \sum_{i=0}^{3} \int_{\tau_i} \operatorname{sgn}(u - u_h) v_h \,\mathrm{d}\boldsymbol{x} \quad = -\frac{1}{6\alpha^3}\big(2\alpha^3 - 5\alpha + 2\big).$$

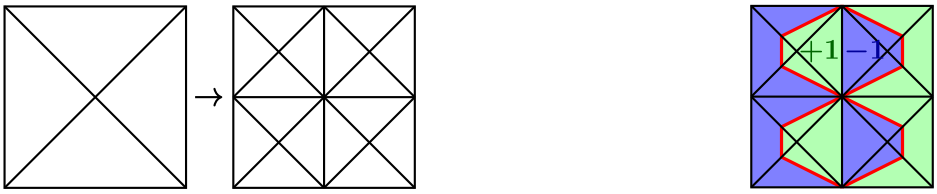
The polynomial $2\alpha^3 - 5\alpha + 2$ has three roots α_i, $i = 0, 1, 2$, where

$$\alpha_0 \approx -1.7623, \qquad \alpha_1 \approx 0.43232, \qquad \alpha_2 \approx 1.3200.$$

Only α_2 satisfies the condition $\alpha > 1$ and therefore $\alpha \approx 1.3200$ yields the only L^1-best approximation. □

Remark 5.2. As in one dimension, the argument that over-/undershoots must occur still holds true for any piecewise linear function such that $\operatorname{sgn}(u - u_h) \pm 1$ is constant on the boundary with the discontinuity, *i.e.*, the part of the boundary where u and u_h are constrained to satisfy different boundary conditions. If the difference $u - u_h$ is constant along this part of the boundary, we also obtain the same pattern for $\operatorname{sgn}(u - u_h)$ as in Figure 10 and from this $u_h(0.5, 0.5)$ can be determined. If, however, $u - u_h$ is not constant along the boundary with the discontinuity, the intersection of u and u_h in τ_1 is no longer parallel to the boundary and all computations have to be redone to determine the magnitude of the overshoot. Nevertheless, the decision whether over-/undershoots occur does not depend on u.

Remark 5.3 (Uniform refinement)**.** If the mesh is refined uniformly, keeping the same structure as shown in Figure 11a, it is easy to see that an L^1-best approximation is given by $u(x_i, y_j) = \alpha$, with α as specified in Section 5.1, if the node (x_i, y_j) is connected with the boundary $x = 1$, and $u(x_i, y_j) = 1$ at the remaining interior nodes. Indeed, in this case we can choose $\psi = \psi_0$ on the set $\{\boldsymbol{x} : u(\boldsymbol{x}) - u_h(\boldsymbol{x}) = 0\}$ as shown in Figure 11b. This shows that the overshoot in the L^1-best approximation remains constant under this type of mesh refinement.



(A) (B)

FIGURE 11. Uniform refinement of the mesh preserving the structure. (A) Remark 5.3: Refinement for which the overshoot in the L^1-best approximation remains constant. (B) Possible choice for sgn(0).

5.2. L^q-best approximation on Mesh 1

Lemma 5.4 (L^q-best approximation on Mesh 1)**.** *For $1 < q < \infty$, the L^q-best approximation of $u \equiv 1$ on Mesh 1 is determined by $u_h(0.5, 0.5) = \alpha$, where α satisfies $\alpha > 1$ and $0 = (\alpha - 1)^{q-1}\big[4\alpha^3 q + 4(1-q)\alpha^2 + (q-6)\alpha + 2\big] - \alpha(q+4) + 2$.*

Proof. If $\alpha \leq 1$, we have $\operatorname{sgn}(u - u_h)|u - u_h|^{q-1} > 0$ in τ_i, $i = 0, 1, 2$, and $\operatorname{sgn}(u - u_h)|u - u_h|^{q-1} \geq 0$ in τ_3. Therefore, we have $\int_{(0,1)^2} \operatorname{sgn}(u - u_h)|u - u_h|^{q-1} v_h \,\mathrm{d}\boldsymbol{x} > 0$, which implies that $\alpha \leq 1$ does not yield an L^q-best approximation for any $1 < q < \infty$. If $\alpha > 1$ a direct computation yields

$$\begin{aligned}\int_{(0,1)^2} \operatorname{sgn}(u - u_h) v_h \,\mathrm{d}\boldsymbol{x} &= \sum_{i=0}^{3} \int_{\tau_i} \operatorname{sgn}(u - u_h) v_h \,\mathrm{d}\boldsymbol{x} \\ &= -\frac{(\alpha-1)^{q-1}\big[4\alpha^3 q + 4(1-q)\alpha^2 + (q-6)\alpha + 2\big] - \alpha(q+4) + 2}{2\alpha^3 q(q+1)(q+2)}.\end{aligned}$$

The claim follows by observing that $(\alpha - 1)^{q-1} > 0$ and $2\alpha^3 q(q+1)(q+2) > 0$. □

5.3. L^q-best approximation on Meshes 2, 3 and 4

Lemma 5.5 (L^q-best approximation on Meshes 2, 3 and 4)**.**

(1) *If $q > 1$, the L^q-best approximation to (1.3) contains over- or undershoots on all three meshes.*
(2) *If $q = 1$, there exists a solution to (1.3) on Mesh 2 such that $u_h(0.5, 1) = u_h(0.5, 0.5) = 1$ and $u_h(0.5, 0) = \alpha$, where α satisfies $\alpha > 1$ and $0 = -3\alpha^3 + 8\alpha - 4$, hence $\alpha \approx 1.2723$. Furthermore, $u_h(0.5, 1) = u_h(0.5, 0.5) = u_h(0.5, 0) = 1$ does not define an L^1-best approximation.*
(3) *If $q = 1$, there exists a solution to (1.3) on Meshes 3 and 4 such that $u_h(0.5, 1) = u_h(0.5, 0.5) = u_h(0.5, 0) = 1$.*

Proof. To see that the first point of the lemma is true, note that if there are no over- or undershoots, *i.e.*, $u_h(0.5, 1) = u_h(0.5, 0.5) = u_h(0.5, 0) = 1$, we have $u - u_h = 0$ in $(0, 0.5) \times (0, 1)$ and $u - u_h > 0$ in $(0.5, 1) \times (0, 1)$ and hence

$$\int_{(0,1)^2} \operatorname{sgn}(u - u_h)|u - u_h|^{q-1} v_i \,\mathrm{d}\boldsymbol{x} > 0 \quad \text{for all } i = 1, 2, 3 \text{ and } 1 < q < \infty,$$

which contradicts Corollary 2.3. Furthermore, it is easy to see that for the second and third mesh, $\alpha = \beta = \gamma = 1$ is an L^1-best approximation by choosing $\psi = \psi_0 \equiv 1$ in $\{\boldsymbol{x} \in (0, 1)^2 : u(\boldsymbol{x}) = u_h(\boldsymbol{x})\}$. In this case, for each of the three nodes $(0.5, 0)$, $(0.5, 0.5)$ and $(0.5, 1)$, we have that $\psi = -1$ on exactly half of the connected elements and $\psi = 1$ on the remaining connected elements.

The second point is more interesting. Note that, if $\alpha = \beta = \gamma = 1$, we obtain $u - u_h > 0$ in τ_2 and τ_3. Hence, $\int_{(0,1)^2} \psi v_1 \,\mathrm{d}\boldsymbol{x} \geq \int_{\tau_2} v_1 \,\mathrm{d}\boldsymbol{x} + \int_{\tau_3} v_1 \,\mathrm{d}\boldsymbol{x} - \int_{\tau_0} v_1 \,\mathrm{d}\boldsymbol{x} > 0$. Hence, this does not yield an L^1-best approximation. If, on the other hand $\beta = 1$ and $\alpha > 1$, ψ is uniquely determined by $\operatorname{sgn}(u - u_h)$ in τ_i, $i = 0, 2, 3$. A direct computation yields

$$\int_{(0,1)^2} \psi v_1 \,\mathrm{d}\boldsymbol{x} = \frac{1}{24\alpha^3}\left(-3\alpha^3 + 8\alpha - 4\right).$$

Hence, $\alpha > 1$ has to satisfy the equation $0 = -3\alpha^3 + 8\alpha - 4$. The roots of the above polynomial are $\alpha_0 \approx -1.8414$, $\alpha_1 \approx 0.56913$ and $\alpha_2 \approx 1.2723$. Only the third root satisfies $\alpha > 1$. Next, we note that if $\beta = \gamma = 1$ implies that $u - u_h = 0$ in τ_i, $i = 1, 4, 5$ and $u - u_h > 0$ τ_i, $i = 6, 7$. It can easily be verified that choosing $\psi_0 \equiv -1$ in τ_4, $\psi_0 \equiv 0$ in τ_5 and

$$\psi \equiv -\frac{\int_{\tau_3} \operatorname{sgn}((u - u_h)(\tilde{\boldsymbol{x}}))v_2(\tilde{\boldsymbol{x}}) \,\mathrm{d}\tilde{\boldsymbol{x}}}{\int_{\tau_1} v_2(\tilde{\boldsymbol{x}}) \,\mathrm{d}\tilde{\boldsymbol{x}}} \in [-1, 1]$$

in τ_1 yields $\int_{(0,1)^2} \psi v_3 \,\mathrm{d}\boldsymbol{x} = \int_{(0,1)^2} \psi v_2 \,\mathrm{d}\boldsymbol{x} = 0$. Note that the last two conditions can be satisfied completely independently of the exact value of $\alpha \geq 1$. □

5.4. Towards general meshes in two dimensions

In this section, we generalize the result in Theorem 1.1 to a structured but non-uniform mesh in two dimensions (for the precise definition see Theorem 5.7). The main idea of the proof is in principle applicable to an arbitrary simplicial 2D mesh with no hanging nodes. However, certain steps become very complex in the general case and the benefit of formulating a sufficient condition for the existence of L^1-best approximations with no over- or undershoots based on our strategy is diminished significantly. Before we present the main result, we introduce a simple necessary condition on general meshes.

First, we introduce some useful notation that we use throughout this section. Denote by Γ_D the part of the boundary with Dirichlet-type boundary conditions. Let $\Gamma_1 \subset \Gamma_D$ be the boundary with the discontinuity (in our case $\Gamma_1 = \{(x, y) \in [0, 1]^2 : x = 1\}$). We will consider the elements in order of their distance to Γ_1 to formulate a sufficient condition on the size of the elements that only depends on elements that are closer to Γ_1. Consider the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ given by the vertices $\mathcal{V}$ and edges $\mathcal{E}$ in the mesh. Let $P(\boldsymbol{x}_i, \boldsymbol{x}_j)$ be a path in $\mathcal{G}$ connecting $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{V}$ and let $|P(\boldsymbol{x}_i, \boldsymbol{x}_j)|$ be the number of edges in the path. We define

$$\begin{aligned}
d(\boldsymbol{x}_i, \Gamma_1) &:= \min_{\boldsymbol{x}_j \in \Gamma_1} \min_{P(\boldsymbol{x}_i, \boldsymbol{x}_j)} |P(\boldsymbol{x}_i, \boldsymbol{x}_j)| \qquad \forall \boldsymbol{x}_i \in \mathcal{V}, \\
d(\tau, \Gamma_1) &:= \min_{\boldsymbol{x}_i \in \tau} d(\boldsymbol{x}_i, \Gamma_1) \qquad \forall \tau \in \mathcal{T}, \\
\mathcal{V}_k &:= \{\boldsymbol{x}_i \in \mathcal{V} : k = d(\boldsymbol{x}_i, \Gamma_1)\}, \\
\mathcal{T}_k &:= \{\tau \in \mathcal{T} : k = d(\tau, \Gamma_1)\}.
\end{aligned}$$

Furthermore, we denote by φ_i the continuous and piecewise linear function satisfying $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$.

5.4.1. A necessary condition on general meshes

We recall from Corollary 2.4 that u_h is an L^1-best approximation of u in U_h iff there exists $\psi \in L^\infty(\Omega)$ such that $\psi(\boldsymbol{x}) = \operatorname{sgn}(u - u_h)(\boldsymbol{x})$ if $u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})$ and $|\psi(\boldsymbol{x})| \leq 1$ otherwise such that

$$\int_\Omega \psi w \,\mathrm{d}\boldsymbol{x} = 0 \qquad \forall w \in U_h.$$

Furthermore, recall the notation $\psi_0 = \psi\big|_{\{u \neq u_h\}}$. Note that for a given u_h, ψ_0 is not necessarily unique.

Corollary 5.6 (Necessary condition on a general mesh in 2D)**.** *Let $U = \operatorname{span}\{\varphi_i : \boldsymbol{x}_i \in \mathcal{V}\}$. For $u \in U$, let $u_h \in U$ be the L^1-best approximation of u subject to the constraint $u = u_h$ on $\Gamma_D \setminus \Gamma_1$ and $u_h = g \neq u$ on Γ_1 with $\operatorname{sgn}(u-g) = \pm 1$ constant on Γ_1. If $u_h(\boldsymbol{x}_i) = u(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{V} \setminus \mathcal{V}_0$, the mesh satisfies the following condition for each $\boldsymbol{x}_i \in \mathcal{V}_1$:*

$$\sum_{\substack{\tau \in \mathcal{T}_1 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} A(\tau) \geq \sum_{\substack{\tau \in \mathcal{T}_0 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} A(\tau),$$

where $A(\tau)$ denotes the area of the element τ.

Proof. Since $u_h(\boldsymbol{x}_i) = u(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{V}_1 \cup \mathcal{V}_2$ and $u_h(\boldsymbol{x}_i) = g(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{V}_0$, we have that $\operatorname{sgn}(u - u_h) \equiv \pm 1$ in any $\bigcup_{\tau \in \mathcal{T}_0} \tau$ and $u - u_h \equiv 0$ in any $\tau \in \mathcal{T}_1$. Without loss of generality, we can assume $\operatorname{sgn}(u - u_h) \equiv 1$ in $\bigcup_{\tau \in \mathcal{T}_0} \tau$. Then, since u_h is an L^1-best approximation, there exists ψ_0 such that the condition in Corollary 2.4 is satisfied, *i.e.*,

$$\begin{aligned} 0 = \int_\Omega \psi \varphi_i \,\mathrm{d}\boldsymbol{x} &= \sum_{\substack{\tau \in \mathcal{T}_0 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} \int_\tau \varphi_i \,\mathrm{d}\boldsymbol{x} + \sum_{\substack{\tau \in \mathcal{T}_1 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} \int_\tau \psi_0 \varphi_i \,\mathrm{d}\boldsymbol{x} \\ &\geq \sum_{\substack{\tau \in \mathcal{T}_0 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} \int_\tau \varphi_i \,\mathrm{d}\boldsymbol{x} - \sum_{\substack{\tau \in \mathcal{T}_1 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} \int_\tau \varphi_i \,\mathrm{d}\boldsymbol{x} = \frac{1}{3} \left(\sum_{\substack{\tau \in \mathcal{T}_0 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} A(\tau) - \sum_{\substack{\tau \in \mathcal{T}_1 \\ \text{s.t. } \boldsymbol{x}_i \in \tau}} A(\tau) \right). \end{aligned}$$

Here we used $\psi_0 \geq -1$. □

5.4.2. A sufficient condition on a structured, non-uniform mesh

For a given ψ_0 on all $\tau \in \mathcal{T}_1$, we can clearly formulate a condition similar to Corollary 5.6 on the area of the triangles $\tau \in \mathcal{T}_2$. Rather than just depending on the area of the elements in $\mathcal{T}_1$, however, this condition depends on the choice for ψ_0 which is only unique in special cases. Therefore, proving that it is a *necessary* condition becomes very difficult and in the most general case perhaps even intractable. Instead, we will establish a condition that allows us to explicitly construct a valid choice for ψ_0 and thus is a *sufficient* rather than a *necessary* condition.

Theorem 5.7 (Sufficient condition on a structured, non-uniform mesh)**.** *Given $N, M \geq 1$, $h_1^x, \ldots h_N^x$ and $h_1^y, \ldots, h_M^y$, such that $\sum_{k=1}^N h_k^x = \sum_{k=1}^M h_k^y = 1$, we define the vertices*

$$(x_i, y_j) = \left(\sum_{k=1}^{i} h_k^x, \sum_{k=1}^{j} h_k^j \right), \qquad 0 \leq i \leq N, 0 \leq j \leq M.$$

We consider the mesh given by the elements τ_{ij}^1, τ_{ij}^2, $1 \leq i \leq N$ and $1 \leq j \leq M$ defined by

$$\tau_{ij}^1 = \begin{cases} \triangle((x_{i-1}, y_{j-1}), (x_i, y_{j-1}), (x_{i-1}, y_j)) & \text{if } i+j \text{ even}, \\ \triangle((x_{i-1}, y_{j-1}), (x_i, y_{j-1}), (x_i, y_j)) & \text{if } i+j \text{ odd}, \end{cases}$$

$$\tau_{ij}^2 = \begin{cases} \triangle((x_{i-1}, y_j), (x_i, y_{j-1}), (x_i, y_j)) & \text{if } i+j \text{ even}, \\ \triangle((x_{i-1}, y_{j-1}), (x_i, y_j), (x_{i-1}, y_j)) & \text{if } i+j \text{ odd}. \end{cases}$$

Let I_1^k and I_2^k, $1 \leq k \leq N$ be defined as

$$I_1^N = I_2^N = 1$$

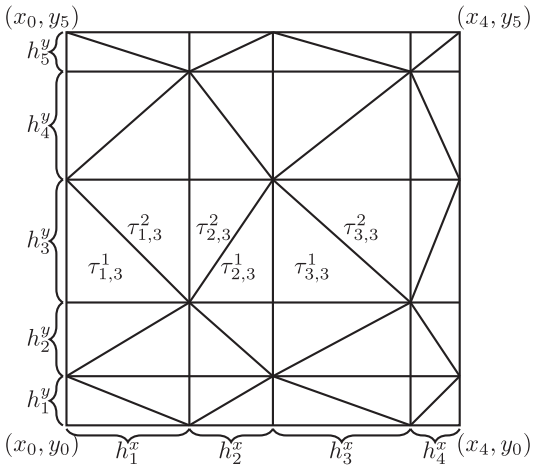


FIGURE 12. Example mesh as defined in Theorem 5.7 with $N = 4$, $M = 5$.

$$I_1^k = \begin{cases} \left(1 - 6\vartheta_{1,k}^2 + 4\vartheta_{1,k}^3\right) & \text{with } \vartheta_{1,k} = \sqrt[3]{\dfrac{1 - (I_2^{k_1} h_{k+1}^x)/h_k^x}{2}} \quad \text{if } \dfrac{I_2^{k+1} h_{k+1}^x}{h_k^x} > \dfrac{3}{4} \\ 0 & \text{otherwise.} \end{cases}$$

$$I_2^k = \begin{cases} \left(2\vartheta_{2,k}^3 - 1\right) & \text{with } \left(6\vartheta_{1,k}^2 - 4\vartheta_{1,k}^3 - 1\right) = \dfrac{I_1^{k+1} h_{k+1}^x}{h_k^x} \quad \text{if } \dfrac{I_1^{k+1} h_{k+1}^x}{h_k^x} > \dfrac{6}{\sqrt[3]{2}} - 3 \\ 0 & \text{otherwise.} \end{cases}$$

Denote by φ_{ij} the piecewise linear function satisfying $\varphi_{ij}(x_k, y_l) = \delta_{ik}\delta_{jl}$. Let $U = \operatorname{span}\{\varphi_{ij} : 0 \le i \le N, 0 \le j \le M\}$. For $u \in U$, let $u_h \in U$ be the L^1-best approximation of u subject to the constraint $u = u_h$ on $\Gamma_D \setminus \Gamma_1$ and $u_h = g \ne u$ on Γ_1 with $\operatorname{sgn}(u - g) = \pm 1$ constant on Γ_1. Then

$$h_k^x \ge \max\left(I_1^{k-1}, I_2^{k-1}\right) h_{k+1}^x, \qquad \text{for all } 1 \le k \le N - 1 \tag{5.1}$$

is a sufficient condition for the existence of an L^1-best approximation with no overshoots. In particular, this is independent of h_j^y for all $1 \le j \le M$.

An example of a mesh as defined in Theorem 5.7 is shown on Figure 12. The proof of Theorem 5.7 is conceptually a straight forward generalization of the proof of Theorem 1.1. To reduce the complexity of the proof, we use the specific structure of the mesh and instead of constructing ψ_0 such that $\int_\Omega \psi\varphi_{ij}\,\mathrm{d}\boldsymbol{x} = 0$ for all $(x_i, y_j) \in \mathcal{V} \setminus \Gamma_D$, we ensure that the following stronger conditions are satisfied for all $1 \le i \le N - 1$ and all $1 \le j \le M$:

$$\int_{\tau_{ij}^1} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x} = -\int_{\tau_{(i+1)j}^1} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x}, \quad \text{where } k \in \{j, j-1\} \quad \text{if } i + j \quad \text{odd, } k = j - 1 \quad \text{otherwise,} \tag{5.2a}$$

$$\int_{\tau_{ij}^2} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x} = -\int_{\tau_{(i+1)j}^2} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x}, \quad \text{where } k \in \{j, j-1\} \quad \text{if } i + j \quad \text{even, } k = j \quad \text{otherwise.} \tag{5.2b}$$

The advantage of this condition is that it is quasi-one-dimensional in the following sense: Given ψ on τ_{ij}^n for all $n = 1, 2$, $1 \le j \le M$ and $i > i_0$, we can construct ψ_0 in $\tau_{i_0,j_0}^{n_0}$ satisfying (5.2) with $(i, j, n) = (i_0, j_0, n_0)$ based only on ψ in $\tau_{i,j_0}^{n_0}$ with $i > i_0$ and independently of τ_{ij}^n where (i, j, n) either satisfies $j \ne j_0$ or $(i, j, n) = (i, j_0, n)$

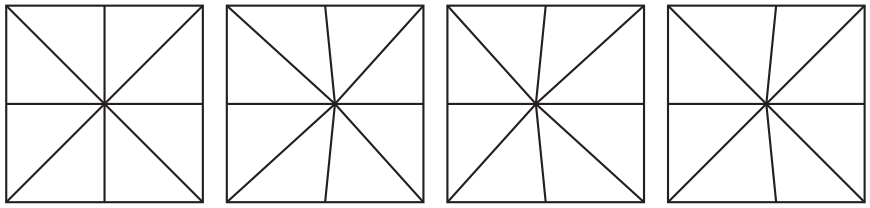


FIGURE 13. Perturbations of Mesh 3. (A) Mesh 3. (B) Mesh 3a. (C) Mesh 3b. (D) Mesh 3c.

with $n \neq n_0$. The construction of ψ_0 follows the construction explained in the second part of Remark 3.4 with the sole difference that instead of splitting an interval into two and assigning opposite signs to ψ_0 in each subinterval, we split a triangle into a triangle and a quadrilateral along a vertical line. The technical details of the proof are presented in Appendix A.

Remark 5.8. Note that a uniform refinement preserving the structure of Meshes 3 and 4, *cf.*, Figure 9, is a special case of Theorem 5.7, where the sufficient condition for the existence of an L^1-best approximation with no over- or undershoots is trivially satisfied.

5.4.3. Weakening the conditions on the structure of the mesh

The mesh in Theorem 5.7 is based on a non-uniform rectangular grid with one diagonal added per rectangle to obtain a simplicial mesh. We will now briefly consider some perturbations of Mesh 3 (*cf.*, Fig. 4) which are shown in Figure 13. The difference between the Meshes 3, 3a, 3b and 3c is that the nodes along the vertical line in the center have been moved slightly: In Mesh 3a the center node is now closer to the boundary $x = 1$; in Mesh 3b the center node is closer to the boundary $x = 0$; in Mesh 3c the nodes on the boundary $y = 0$ and $y = 1$ are closer to the boundary $x = 1$. It is easy to see that the necessary condition in Corollary 5.6 is satisfied if the minimum distance of any point along the line in the center to the boundary $x = 1$ is smaller than the minimum distance to the boundary $x = 0$. This is the case for Mesh 3a and 3c, but not 3b. It is also easy to see that we can define ψ_0 satisfying (5.2) similar to the proof of Lemma 3.1 if the necessary condition in Corollary 5.6 is satisfied. This shows that it is not necessary to require that the mesh is based on a rectangular grid. However, a condition similar to Theorem 5.7 becomes more complex if we impose fewer restrictions on the structure of the mesh. For example, the condition may become dependent on how the mesh varies in the y-direction. Furthermore, the graph given by the vertices and edges of the mesh may become much more complex. This may make it impossible to use the stronger condition (5.2), and thus it may increase the complexity of constructing $\widetilde{\psi}_0$. In order to write an algorithm to generate meshes that allow for L^1-best approximations without over- or undershoot, it will be necessary to find a balance between imposing as little structure as possible in order for it to be applicable to more general domains and discontinuities while imposing enough structure to keep the (computational) complexity of generating the mesh feasible.

6. NUMERICAL EXAMPLES

In this section we consider selected examples of meshes for which we have determined the solution of the best approximation problem (1.3) numerically by interpreting the condition (2.2) as a variational problem that can be implemented using standard finite element techniques and the algorithm described in Section 3.6 of [10] (*cf.*, Sect. 1.4). Here, we have used FEniCS [1] for the implementation. In Section 6.1 we illustrate that the overshoot in the L^q-best approximation does not vanish if the mesh is refined and that these observations even apply if u is a more general smooth function.

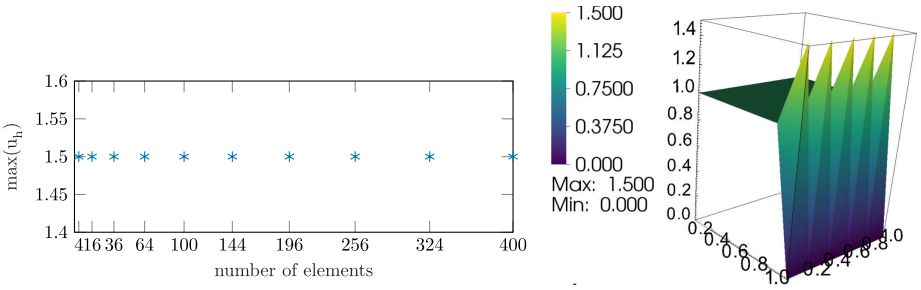


FIGURE 14. *Left*: $\max(u_h)$ for $q = 2$ and several refinements as shown in Figure 11. *Right*: u_h with 100 elements.

In Section 6.3 we illustrate that the L^q-approximation on the three meshes considered in Section 5.3 converges to the L^1-best approximation characterized in Lemma 5.5. Furthermore, we show how the understanding of these special cases can be applied to predict the behavior of the L^q-best approximation on a more general mesh.

6.1. Refinement of the mesh

6.1.1. Gibbs phenomenon on meshes in two dimensions

We start with Mesh 1 depicted in Figure 9 and the refinement shown in Figure 11 that preserves the structure of the mesh. We have already shown in Remark 5.3 that for $q = 1$ there exists an L^1-best approximation such that the overshoot remains constant as we refine the mesh.

Indeed, Figure 14 shows the maximum value of u_h for this example with $q = 2$ and for several refinements of the mesh, as well as the approximation u_h for a mesh with this structure consisting of 100 elements. We can clearly see that the maximum value remains constant under this type of refinement which suggests that the maximum overshoot also remains constant for $q \neq 1$, as well as in the limit $q \to 1$.

6.1.2. Gibbs phenomenon on meshes in one dimension

Next, we consider a one-dimensional example such that u is not piecewise linear and compute the L^q-best approximation numerically. Let $u(x) = 1 + 0.1\sin(2\pi x)$ on $(0, 1)$ and consider the L^q-best approximation u_h with $u_h(0) = 1$ and $u_h(1) = 0$ on four different grids: two uniform grids with 5 and 100 elements, respectively, and two meshes where all elements are the same size except the last one which is twice the size of the others. Again we consider a mesh with 5 elements and one with 100 elements. Note that the latter two meshes violate the conditions in Theorem 1.1, but satisfy the condition in Remark 3.6. We therefore expect the overshoot to vanish as $q \to 1$ in the first two cases and to decrease but still be present in the last two. Remark 3.6 and the observations for the previous example suggest that for $u \equiv 1$, we could expect the overshoot to be the same both when 5 and 100 elements are employed on both the uniform and the non-uniform meshes.

Figure 15 shows the maximum error at the nodes in all four cases for several values of q. We observe that the overshoot indeed decreases as $q \to 1$. Furthermore, we see that the overshoot is very similar for the coarse and fine meshes in both cases which confirms that the overshoot does not disappear under mesh refinement. However, the overshoot is not identical for 5 and for 100 elements in both cases which can be attributed to the fact that u is not constant. Furthermore, note that the overshoot for the non-uniform mesh is consistently larger than for the uniform mesh, which suggests that it does not disappear entirely as $q \to 1$. Note that on the non-uniform mesh when $u \equiv 1$ and $q = 1$, the overshoot would be $2\sqrt{3}/3 - 1 \approx 0.15$; see Remark 3.6.

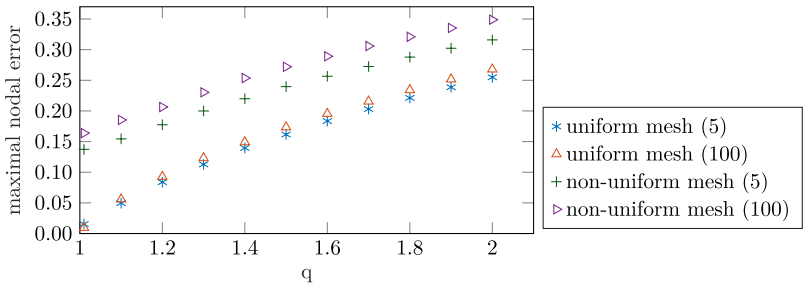


FIGURE 15. Maximal nodal error in the L^q-best approximation of $u(x) = 1 + 0.1\sin(2\pi x)$ with $u_h(1) = 0$ for different values of q and four different meshes. Two of the meshes are uniform meshes consisting of 5 and 100 elements, respectively. The remaining two meshes satisfy the condition in Remark 3.6 and again consist of 5 and 100 elements, respectively.

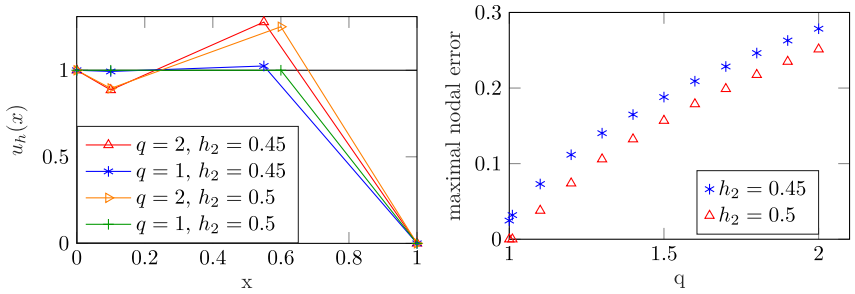


FIGURE 16. L^q-best approximations on two three-element meshes on $(0, 1)$ with $h_1 = 0.1$ and two different choices for h_2. *Left*: L^q-best approximation with $q = 2$ and $q = 1$. *Right*: maximal nodal error for several values of q.

6.2. (Vanishing) Overshoot in one dimension

To illustrate the graded mesh condition in Theorem 1.1, we consider two three-element meshes on $(0, 1)$. For the first one we choose $h_1 = 0.1$ and $h_2 = h_3 = 0.45$, *i.e.*, the mesh consists of the subintervals $(0, 0.1)$, $(0.1, 0.55)$ and $(0.55, 1)$. For the second one we choose $h_1 = 0.1$, $h_2 = 0.5$ and $h_3 = 0.4$, *i.e.*, the mesh consisting of the subintervals $(0, 0.1)$, $(0.1, 0.6)$ and $(0.6, 1)$. We will check the condition (1.5) for both meshes; indeed, we will see that for the first mesh the condition is violated, but it is satisfied for the second mesh. In the latter case, we therefore know that there exists an L^1-best approximation without over- or undershoots. In the former case, it is *a priori* unknown whether such an L^1-best approximation exists, since it is an open problem whether (1.5) is also a necessary condition.

In the first case, we obtain from (1.4b) that $\vartheta_3 = \vartheta_2 = 0$ yielding the following sufficient conditions for the existence of an L^1-best approximation without over- or undershoots: $h_2 \geq h_3$ and $h_1 \geq h_2$. The second condition is violated. In fact, it is easy to show that, if $h_2 = h_3$, the condition $h_1 \geq h_2$ is necessary for the existence of an L^1-best approximation without over- or undershoots. Moreover, one can show that the L^1-best approximation is unique in this case by solving the optimality condition in Corollary 2.4 for the points where u and u_h intersect. The intersection points uniquely determine $u_h(0.1) \approx 0.9931$ and $u_h(0.55) \approx 1.0247$. For brevity, the details are omitted here.

For the second mesh, we again have $\vartheta_3 = 0$ and (1.5) with $i = 2$ becomes $h_2 \geq h_3$, which holds for $h_2 = 0.5$ and $h_3 = 0.4$. If $i = 2$, we obtain from (1.4b) that $\vartheta_2^2 = 0.1$. Hence, $\vartheta_2 > 1 - 1/\sqrt{2}$ and there is no condition on h_1 according to Theorem 1.1. Therefore, there exists an L^1-best approximation without over- or undershoots.

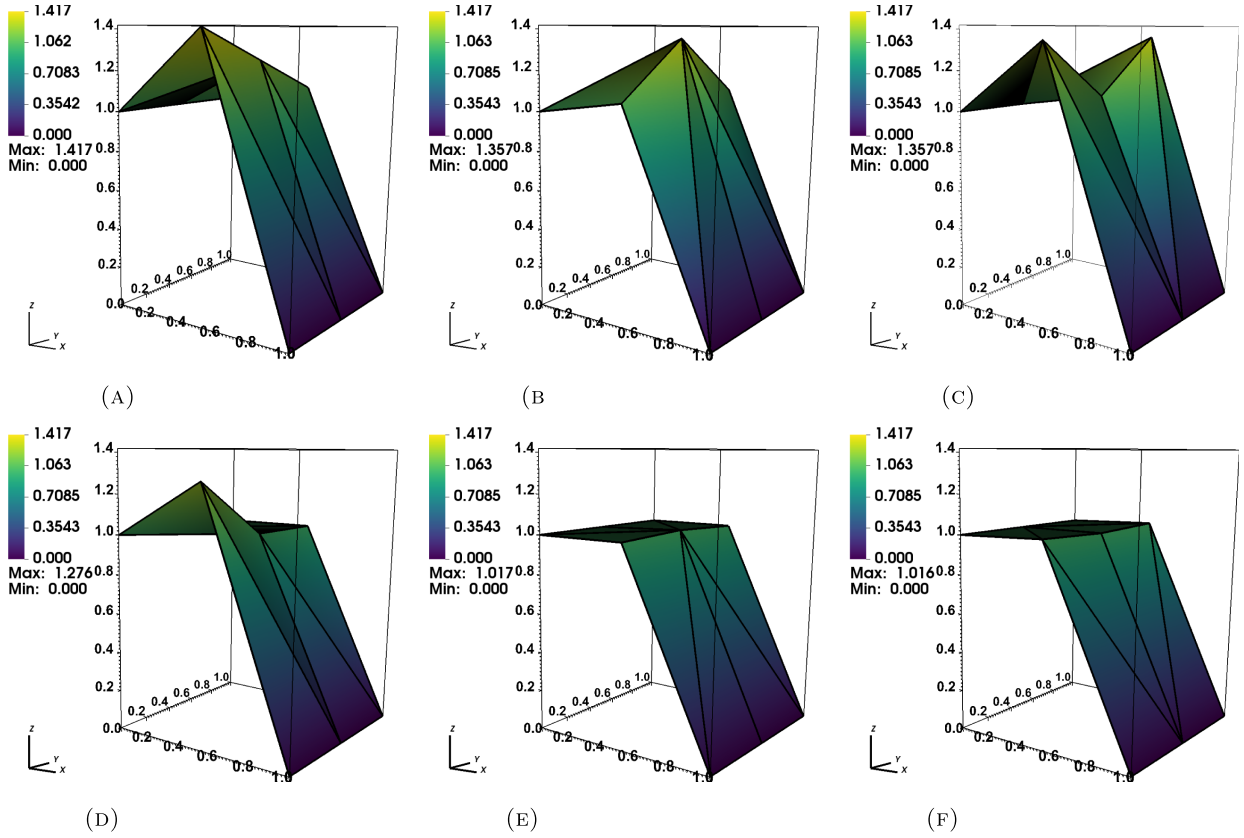


FIGURE 17. L^q-best approximation of a boundary discontinuity in two dimensions on Meshes 2, 3 and 4, *cf.*, Figures 4 and 9. (A) Mesh 2, $q = 2$. (B) Mesh 3, $q = 2$. (C) Mesh 4, $q = 2$. (D) Mesh 2, $q = 1.01$. (E) Mesh 3, $q = 1.01$. (F) Mesh 4, $q = 1.01$.

Figure 16 shows the L^q-best approximation on both meshes for $q = 2$ and $q = 1$ on the left and the maximal nodal error on both meshes for several values of q on the right. The approximations for $q > 1$ were again obtained using the implementation of the best approximation problem in FEniCS. We can clearly see, that the maximal overshoot is always larger on the first mesh. In both cases it decreases as $q \to 1$, but the overshoot only vanishes completely on the second mesh. However, even on the first mesh the maximal overshoot is very small for $q = 1$. Note that, if h_2 and h_1 as chosen for the first mesh were swapped, the maximal overshoot for $q = 1$ would be $u_h(0.55) - 1 = 0.2792$ according to Remark 3.6 and thus significantly larger than the overshoot we can observe. This shows that the effect of an element being too small and causing the L^1-best approximation to contain over- and undershoots is much weaker away from the discontinuity than near the discontinuity.

6.3. (Vanishing) Overshoot in two dimensions

6.3.1. Overshoot on Meshes 2, 3 and 4 from Section 5.3

Figure 17 shows the best approximations for $q = 2$ and $q = 1.01$ for three of the meshes we have considered in Section 5.3. Even just a comparison of these two cases for each of the meshes illustrates clearly how the overshoot gradually vanishes on Meshes 3 and 4. On Mesh 2, the overshoot vanishes away from the boundary $y = 0$; this is consistent with the L^1-best approximation described above that only exhibits an overshoot at the node $(0.5, 0)$ and no overshoot at all other nodes.

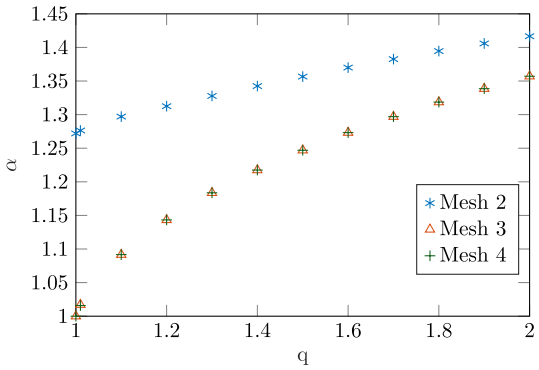
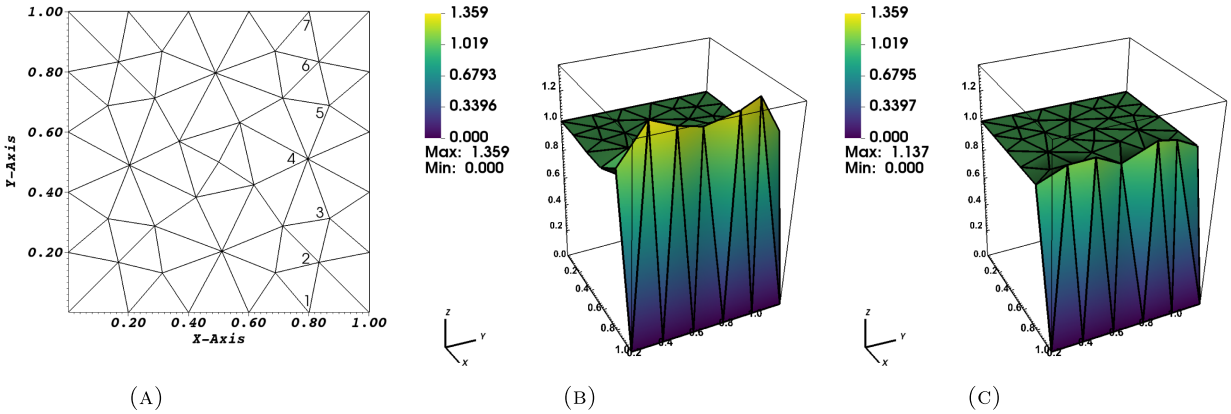


FIGURE 18. Maximum overshoot on Meshes 2, 3 and 4 (*cf.*, Figs. 4 and 9) for different values of q.



(A) (B) (C)

FIGURE 19. L^q-best approximation on an unstructured mesh. (A) Mesh A. (B) $q = 2$. (C) $q = 1.01$.

Figure 18 shows the maximum overshoot for all three meshes for different values of q. The overshoot for $q = 1$ is taken from the theoretically determined L^1-best approximations discussed in Section 5.3. All remaining values have been determined numerically with an implementation in FEniCS [1]. The plot shows that for the third and fourth mesh, the overshoot indeed disappears as $q \to 1$, whereas for the second mesh it decreases but does not vanish.

6.3.2. Overshoot on unstructured meshes

As a final example, we consider the unstructured mesh shown on the left in Figure 19. The interior nodes connected to the boundary are labelled $1, 2, \ldots, 7$. We first check the necessary condition in Corollary 5.6. At nodes $1, 4$ and 7 the condition is satisfied, whereas it is not satisfied at nodes $2, 3, 5$ and 6. From this we immediately obtain that the overshoot cannot vanish at the nodes $2, 3, 5$ and 6 as $q \to 1$. Even though Theorem 5.7 is not directly applicable in this case, we observe that the relative sizes of the triangles close to nodes 1, 4 and 7 suggest that it is likely that ψ_0 can be constructed similarly to Theorem 5.7 in the triangles connected to nodes 1, 4 and 7 following the same principle as the construction in the second part of Remark 3.4. Hence, we expect the overshoot to vanish at these nodes. Figure 19 shows the L^q-best approximation on the unstructured mesh for $q = 2$ and $q = 1.01$ in the center and on the right, respectively. Here, we clearly observe that the approximation for $q = 2$ exhibits overshoots at all nodes connected to the boundary $x = 1$ with larger

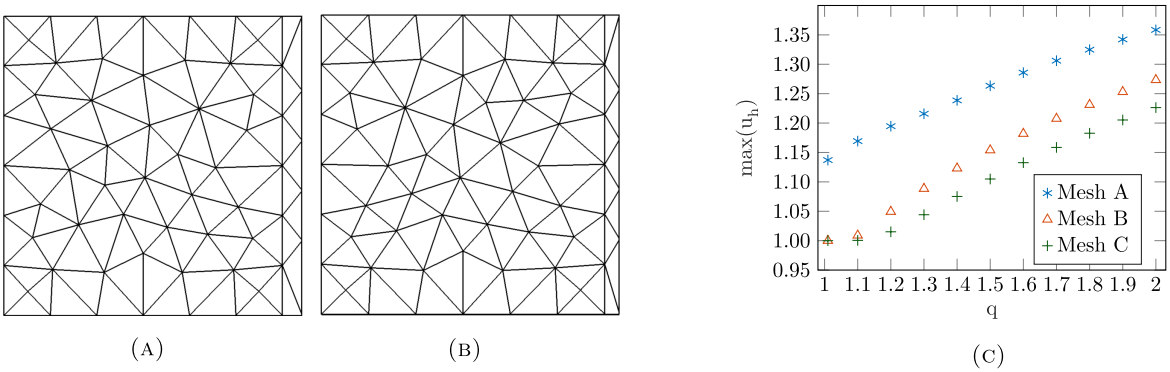


FIGURE 20. Overshoot in the L^q-best approximation on unstructured meshes. See Figure 19 for Mesh A. (A) Mesh B. (B) Mesh C. (C) $\max(u_h)$.

overshoots at the nodes $2, 3, 5$ and 6. At these nodes the overshoot is reduced but still clearly visible for $q = 1.01$. On the other hand at the nodes $1, 4$ and 7 the overshoot has nearly vanished for $q = 1.01$.

Figure 20 shows two further unstructured meshes which have been designed to satisfy the necessary condition in Corollary 5.6. The difference between the two meshes is that the distance between the boundary $x = 1$ and the vertical line containing all nodes connected to this boundary is smaller in Mesh C than in Mesh B. Figure 20c shows the maximum value of u_h for different q and Meshes A, B and C. This illustrates that the overshoot decreases on all three meshes as $q \to 1$. The overshoot on Mesh C is always smaller than the overshoot on the other two meshes and the overshoot on Mesh A is always larger than on the other two meshes. This illustrates that if the area of the elements connected to the boundary is decreased in comparison the area of the remaining elements, then the overshoot is reduced for any q and decreases more rapidly as $q \to 1$. This is consistent with the theoretical results in one dimension illustrated at the start of this article in Figure 3.

Remark 6.1. The mesh in the final example, as well as the mesh conditions in Theorems 1.1 and 5.7, have a certain similarity with so-called *Shishkin* meshes or other types of layer-adapted meshes (*cf.*, [27, 34]). It is important to note that the condition in our case is much weaker in the sense that even certain uniform meshes are sufficient, whereas a certain grading is necessary for numerical methods based on layer-adapted meshes. Nonetheless, the difference between Meshes B and C in Figure 20 as well as the difference seen for different h in Figure 3 illustrate that introducing a grading may improve how fast the over- and undershoots vanish in the limit $q \to 1$. Therefore, one can benefit from including ideas based on layer-adapted meshes in order to generate meshes that allow for L^q-best approximation with negligible overshoot for $q = 1 + \delta > 1$.

7. CONCLUSIONS

In this article, we have investigated Gibbs phenomena in the L^q-best approximation of discontinuities within finite element spaces. Using selected cases, we have proven that the Gibbs phenomenon can be eliminated as $q \to 1$ on certain meshes. However, we have seen that there exist non-uniform meshes in one dimension that lead to Gibbs phenomena even if $q = 1$. In two dimensions, even some uniform meshes lead to Gibbs phenomena if $q = 1$. Nonetheless, the magnitude of the oscillations decreases as $q \to 1$ on all meshes.

The computational examples presented in this article confirm the theoretical results. Moreover, we have seen that similar observations can be made for more general cases. Furthermore, we have demonstrated that the Gibbs phenomenon cannot be eliminated on certain meshes under mesh refinement that preserves certain properties of the mesh. For the final computational example, we have been able to establish a link between the structure of the mesh near the discontinuity and the magnitude of the overshoot at the nodes. This observation suggests

that the oscillations can be eliminated in the limit as q tends to 1 if the mesh structure near the discontinuity is suitably adjusted. Indeed, this has been used to design meshes for the non-linear Petrov–Galerkin method for the convection-diffusion-reaction equation presented in [15].

Appendix A. Proof of Theorem 5.7

To prove Theorem 5.7, we will construct ψ_0 iteratively, where ψ_0 in $\tau \in \mathcal{T}_k$ depends on ψ_0 in $\mathcal{T}_{k-1}$. As in Corollary 5.6, ψ_0 on $\tau \in \mathcal{T}_1$ depends on the area of the connected elements in $\mathcal{T}_0$. The proof relies on the following Proposition, which holds on any conforming simplicial two-dimensional mesh regardless of the structure of the mesh and will be proven in Section A.2.

Proposition A.1 (Improved ψ_0)**.** *Let u, u_h be piecewise linear function on a conforming, simplicial mesh such that $u = u_h$ on $\Gamma_D \setminus \Gamma_1$ and $u_h = g \neq u$ on Γ_1 with $\operatorname{sgn}(u-g) = \pm 1$ constant. Denote by φ_i the piecewise linear function satisfying $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$ for any $\boldsymbol{x}_j \in \mathcal{V}$, where $\mathcal{V}$ are the vertices of the mesh. Let $\psi \in \mathrm{L}^\infty(\Omega)$ be defined on $\bigcup_{\tau\in\mathcal{T}_{k-1}} \tau \cup \left(\bigcup_{\tau\in\mathcal{T}_k} \tau \cap \{\boldsymbol{x} \in \Omega : u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})\}\right)$ such that $\psi = \operatorname{sgn}(u-u_h)$ on $\bigcup_{\tau\in\mathcal{T}_{k-1}\cup\mathcal{T}_k} \tau \cap \{\boldsymbol{x} \in \Omega : u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})\}$ and $|\psi| \leq 1$ everywhere. Assume there exists $\widetilde{\psi}_0$ on $\left(\bigcup_{\tau\in\mathcal{T}_k} \tau\right) \setminus \{\boldsymbol{x} \in \Omega : u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})\}$ such that $\widetilde{\psi}_0\big|_\tau \in [-1,1]$ is a constant in each $\tau \in \mathcal{T}_k$ with $u \equiv u_h$ in τ and, for all $\boldsymbol{x}_i \in \mathcal{V}_k$,*

$$0 = \sum_{\substack{\tau\in\mathcal{T}_{k-1}\\ \text{s.t. } \boldsymbol{x}_i\in\tau}} \int_\tau \psi\varphi_i \,\mathrm{d}\boldsymbol{x} + \sum_{\substack{\tau\in\mathcal{T}_k\\ \text{s.t. } \boldsymbol{x}_i in\tau}} \left(\int_{\tau\cap\{\boldsymbol{x}\in\Omega\,:\,u(\boldsymbol{x})\neq u_h(\boldsymbol{x})\}} \psi\varphi_i \,\mathrm{d}\boldsymbol{x} + \int_{\tau\setminus\{\boldsymbol{x}\in\Omega\,:\,u(\boldsymbol{x})\neq u_h(\boldsymbol{x})\}} \widetilde{\psi}_0\varphi_i \,\mathrm{d}\boldsymbol{x} \right). \tag{A.1}$$

Then there exists ψ_0 on $\left(\bigcup_{\tau\in\mathcal{T}_k} \tau\right) \setminus \{\boldsymbol{x} \in \Omega : u(\boldsymbol{x}) \neq u_h(\boldsymbol{x})\}$ satisfying

$$\int_\tau \psi_0\varphi_i \,\mathrm{d}\boldsymbol{x} = \int_\tau \widetilde{\psi}_0\varphi_i \,\mathrm{d}\boldsymbol{x} \qquad \text{for all } \tau \in \mathcal{T}_k \quad \text{such that } (u-u_h)\big|_\tau \equiv 0 \quad \text{and all } \boldsymbol{x}_i \in \tau \cap \mathcal{V}_k. \tag{A.2}$$

Furthermore, for all $\tau \in \mathcal{T}_k$ such that $(u-u_h)\big|_\tau \equiv 0$ and all $\boldsymbol{x}_i \in \tau \cap \mathcal{V}_{k+1}$,

$$\int_\tau \psi_0\varphi_i \,\mathrm{d}\boldsymbol{x} = \begin{cases} 0 & \text{if } \left(|\tau\cap\mathcal{V}_k| = 2 \quad \text{and } \left|\widetilde{\psi}_0\big|_\tau\right| \leq 3/4\right) \\ & \text{or } \left(|\tau\cap\mathcal{V}_k| = 1 \quad \text{and } \left|\widetilde{\psi}_0\big|_\tau\right| \leq 6/\sqrt[3]{2} - 3\right) \\ \dfrac{\operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right) A(\tau)}{3}(1 - 6\vartheta_1^2 + 4\vartheta_1^3) & \text{if } |\tau\cap\mathcal{V}_k| = 2 \quad \text{and } \left|\widetilde{\psi}_0\big|_\tau\right| > 3/4, \\ \dfrac{\operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right) A(\tau)}{3}(2\vartheta_2^3 - 1) & \text{if } |\tau\cap\mathcal{V}_k| = 1 \quad \text{and } \left|\widetilde{\psi}_0\big|_\tau\right| > 6/\sqrt[3]{2} - 3 \\ \dfrac{\widetilde{\psi}_0\big|_\tau A(\tau)}{3} & \text{otherwise,} \end{cases} \tag{A.3}$$

where ϑ_1 and ϑ_2 are given by $\vartheta_1 = \sqrt[3]{\left(1-\left|\psi_0\big|_\tau\right|\right)/2} \in (0,1)$ and $6\vartheta_2^2 - 4\vartheta_2^3 - 1 = \left|\psi_0\big|_\tau\right|$. Furthermore, for any $\tau \in \mathcal{T}_k$ such that $\left|\psi_0\big|_\tau\right| \in (0,1)$ and for any $\boldsymbol{x}_i \in \mathcal{V}_{k+1} \cap \tau$,

$$\left| \int_\tau \psi_0\varphi_i \,\mathrm{d}\boldsymbol{x} \right| < \left| \int_\tau \widetilde{\psi}_0\varphi_i \,\mathrm{d}\boldsymbol{x} \right|. \tag{A.4}$$

Note that equation (A.1) states that ψ as in Proposition A.1, where additionally $\psi := \widetilde{\psi}_0$ on $\tau \in \mathcal{T}_k$ with $(u-u_h)\big|_\tau \equiv 0$, satisfies the condition in Corollary 2.4 with regard to the subspace of U_h spanned by the basis functions φ_i associated with the vertices $\boldsymbol{x}_i \in \mathcal{V}_k$. Equation (A.2) ensures that the same holds true after replacing $\widetilde{\psi}_0$ with ψ_0. The identity (A.4) shows in what sense replacing $\widetilde{\psi}_0$ by ψ_0 is an improvement. We will see that (A.4) directly follows from the expression for the integral in (A.3)

A.1. Proof of Theorem 5.7

Proof. This is mainly an application of Proposition A.1. First note that if u_h is an L^1-best approximation of u such that $u_h(x_i, y_j) = u_h(x_i, y_j)$ for $0 \le i \le N-1$ and $0 \le j \le M$, we have that $\psi = \operatorname{sgn}(u - u_h) \equiv \pm 1$ on $\bigcup_{1\le j\le M}(\tau^1_{0,j} \cup \tau^2_{0,j})$. As noted in Section 5.4.2, the condition in Corollary 2.4 is satisfied if for all $1 \le i \le N-1$ and all $1 \le j \le M$ the following conditions are satisfied (*cf.*, (5.2)):

$$\int_{\tau^1_{ij}} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x} = -\int_{\tau^1_{(i+1)j}} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x}, \quad \text{where } k \in \{j, j-1\} \quad \text{if } i+j \quad \text{odd, } k = j-1 \quad \text{otherwise,} \tag{A.5a}$$

$$\int_{\tau^2_{ij}} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x} = -\int_{\tau^2_{(i+1)j}} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x}, \quad \text{where } k \in \{j, j-1\} \quad \text{if } i+j \quad \text{even, } k = j \quad \text{otherwise.} \tag{A.5b}$$

Given ψ on $\bigcup_{\substack{i+1\le l\le N\\ 1\le j\le M}}\left(\tau^1_{lj} \cup \tau^2_{lj}\right)$, we can therefore define

$$\widetilde{\psi}_0\big|_{\tau^n_{ij}} := -\frac{3}{A(\tau_{ij})}\int_{\tau^n_{(i+1)j}} \psi\varphi_{ik}\,\mathrm{d}\boldsymbol{x}, \quad \text{where } k \in \{j, j-1\} \quad \text{such that } (\boldsymbol{x}_i, y_k) \in \tau_{ij} \cap \tau_{(i+1),j}. \tag{A.6}$$

If $\left|\widetilde{\psi}_0\right| \le 1$, we can then apply Proposition A.1 to define $\psi = \psi_0$ on $\bigcup_{1\le j\le M}\tau_{ij}$, yielding

$$\int_{\tau^n_{ij}} \psi\varphi_{(i-1)k}\,\mathrm{d}\boldsymbol{x} = \begin{cases} \dfrac{\operatorname{sgn}(u-g)(-1)^{N-i}A(\tau^n_{ij})}{3} I^i_1 & \text{if } \left|\{(\boldsymbol{x}_i, y_j), (\boldsymbol{x}_i, y_{j-1})\} \cap \tau^n_{ij}\right| = 2 \\ \dfrac{\operatorname{sgn}(u-g)(-1)^{N-i}A(\tau^n_{ij})}{3} I^i_2 & \text{if } \left|\{(\boldsymbol{x}_i, y_j), (\boldsymbol{x}_i, y_{j-1})\} \cap \tau^n_{ij}\right| = 1. \end{cases} \tag{A.7}$$

We have that (A.7) holds for $i = N$ since $\psi \equiv \operatorname{sgn}(u-g)$ – which was assumed to be constant on Γ_1 – and $I^N_1 = I^N_2 = 1$. Hence, by induction (A.7), holds for all $1 \le i \le N$ provided $\left|\widetilde{\psi}_0\right| \le 1$ in each step. Note that if both $(x_i, y_{j-1}) \in \tau^n_{ij} \cap \tau^n_{(i+1),j}$ and $(x_i, y_j) \in \tau^n_{ij} \cap \tau_{(i+1),j}$ in (A.6), either choice for k yields the same result due to Proposition A.1. To finish the proof, it remains to be shown that (5.1) guarantees $\left|\widetilde{\psi}_0\right| \le 1$ in each step so that Proposition A.1 can be applied. To see this, we combine (A.6) and (A.7) and obtain

$$\left|\widetilde{\psi}_0\big|_{\tau^n_{ij}}\right| = \begin{cases} \dfrac{A(\tau_{(i+1),j})}{A(\tau_{ij})} I^{i-1}_1 & \text{if } \left|\{(x_{i+1}, y_j), (x_{i+1}, y_{j-1})\} \cap \tau^n_{(i+1),j}\right| = 2, \\ \dfrac{A(\tau_{(i+1),j})}{A(\tau_{ij})} I^{i-1}_2 & \text{if } \left|\{(x_{i+1}, y_j), (x_{i+1}, y_{j-1})\} \cap \tau^n_{(i+1),j}\right| = 1. \end{cases} \tag{A.8}$$

Noticing that $A(\tau_{(i+1),j})/A(\tau_{ij}) = h^x_{i+1}h^y_j/h^x_i h^y_j = h^x_{i+1}/h^x_i$ finishes the proof. □

A.2. Proof of Proposition A.1

Proof. Let τ be a triangle with vertices, edges and angles labelled as in Figure A.1. Denote by φ_X, φ_Y and φ_Z the hat functions that are one at X, Y and Z, respectively, and zero at all other vertices. First note that, since $\widetilde{\psi}_0$ is constant on τ,

$$\int_\tau \widetilde{\psi}_0\varphi_X\,\mathrm{d}\boldsymbol{x} = \int_\tau \widetilde{\psi}_0\varphi_Y\,\mathrm{d}\boldsymbol{x} = \int_\tau \widetilde{\psi}_0\varphi_Z\,\mathrm{d}\boldsymbol{x} = \widetilde{\psi}_0\big|_\tau \frac{A(\tau)}{3}. \tag{A.9}$$

We start our proof with the case $|\tau \cap \mathcal{V}_k| = 2$. Assume that $Z \in \mathcal{V}_{k+1}$ and $X, Y \in \mathcal{V}_k$. In this case we add a line parallel to XY intersecting XZ in X' and YZ in Y' such that

$$\frac{|X'Z|}{|XZ|} = \frac{|Y'Z|}{|YZ|} = \vartheta_1 \in (0, 1). \tag{A.10}$$

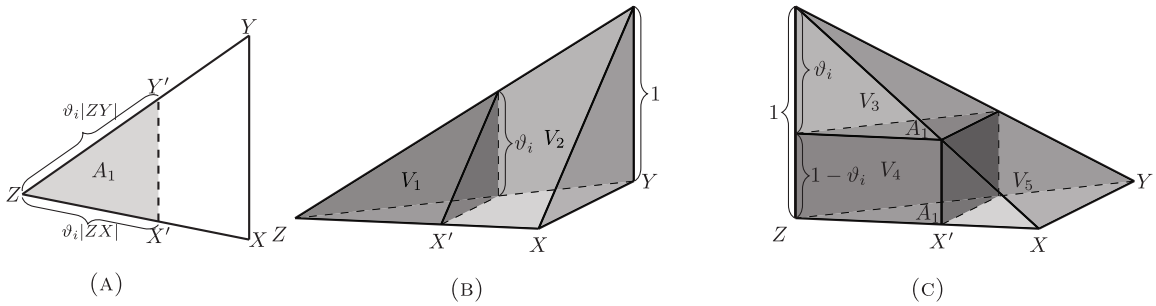


FIGURE A.1. Visualizations used for the construction of ψ_0 in the proof of Proposition A.1. (A) Subdivision of τ. (B) Subdivision of $\int_\tau \varphi_Y \,\mathrm{d}\boldsymbol{x}$ into V_1 and V_2. (C) Subdivision of $\int_\tau \varphi_Z \,\mathrm{d}\boldsymbol{x}$ into V_3, V_4, V_5.

For $\vartheta_1 \in (0,1)$ and $s \in [0,1]$ we define

$$\psi_0 = \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right) \qquad \text{in } \square(X',X,Y,Y'), \qquad\qquad \psi_0 = -\operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)s \qquad \text{in } \triangle(X',Y',Z), \tag{A.11}$$

where $\square(X',X,Y,Y')$ denotes the quadrilateral defined by the vertices X',X,Y,Y' and $\triangle(X',Y',Z)$ denotes the triangle defined by the vertices X',Y',Z.

We will now show, that for any $s \in [0,1]$, there exists $\vartheta_1(s)$ such that ψ_0 as defined above satisfies (A.2). Then, we will select s such that ψ_0 additionally satisfies (A.3). We first note that since $X'Y'$ is parallel to XY, $\int_\tau \psi_0\varphi_X \,\mathrm{d}\boldsymbol{x} = \int_\tau \psi_0\varphi_Y \,\mathrm{d}\boldsymbol{x}$. Furthermore, the area A_1 of $\triangle(X',Y',Z)$ is given by $A_1 = \vartheta_1^2 A(\tau)$, where $A(\tau)$ denotes the area of τ. In order to compute the integral $\int_\tau \psi_0\varphi_Y \,\mathrm{d}\boldsymbol{x}$, we split the volume of the pyramid formed by φ_Y on τ into two parts V_1 and V_2 as shown in Figure A.1. Note that $V_1 = {}^{\vartheta_1 A_1}\!/_3 = {}^{\vartheta_1^3 A(\tau)}\!/_3$. Hence,

$$\begin{aligned}\int_\tau \psi_0\varphi_X \,\mathrm{d}\boldsymbol{x} = \int_\tau \psi_0\varphi_Y \,\mathrm{d}\boldsymbol{x} &= \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\left(\int_{\square(X',X,Y,Y')} \varphi_Y \,\mathrm{d}\boldsymbol{x} - s\int_{\triangle(X',Y',Z)} \varphi_Y \,\mathrm{d}\boldsymbol{x}\right)\\ &= \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\left(\int_\tau \varphi_Y \,\mathrm{d}\boldsymbol{x} - V_1 - sV_1\right) = \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\frac{1-(1+s)\vartheta_1^3}{3}A(\tau).\end{aligned} \tag{A.12}$$

Thus, (A.2) is satisfied iff

$$\vartheta_1(s) = \sqrt[3]{\frac{1-\left|\widetilde{\psi}_0\big|_\tau\right|}{1+s}} \in (0,1). \tag{A.13}$$

Note that for $\left|\widetilde{\psi}_0\big|_\tau\right| = 1$, $\vartheta_1(s) = 0$ for all s, $\psi_0 = \widetilde{\psi}_0$ on τ and (A.2)–(A.4) are trivially satisfied. We therefore only need to consider $\left|\widetilde{\psi}_0\big|_\tau\right| < 1$. In order to compute $\int_\tau \psi_0\varphi_Z \,\mathrm{d}\boldsymbol{x}$, we split the volume of the pyramid formed by φ_Z on τ into three parts V_3, V_4 and V_5 as shown in Figure 20b. We observe that $V_3 = {}^{\vartheta_1 A_1}\!/_3 = {}^{\vartheta_1^3 A(\tau)}\!/_3$ and $V_4 = (1-\vartheta_1)\vartheta_1^2 A(\tau)$. Hence,

$$\begin{aligned}\int_\tau \psi_0\varphi_Z \,\mathrm{d}\boldsymbol{x} &= \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\left(\int_{\square(X',X,Y,Y')} \varphi_Z \,\mathrm{d}\boldsymbol{x} - s\int_{\triangle(X',Y',Z)} \varphi_Z \,\mathrm{d}\boldsymbol{x}\right)\\ &= \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\left(\int_\tau \varphi_Z \,\mathrm{d}\boldsymbol{x} - (1+s)(V_3+V_4)\right) = \operatorname{sgn}\left(\widetilde{\psi}_0\big|_\tau\right)\frac{1-(1+s)(3(1-\vartheta_1)\vartheta_1^2+\vartheta_1^3)}{3}A(\tau).\end{aligned} \tag{A.14}$$

Next consider $f(s) = 1 + 2(1+s)\vartheta_1(s)^3 - 3(1+s)\vartheta_1(s)^2$ which is a continuous function in s. We observe for $\left(1 - \left|\widetilde{\psi}_0|_\tau\right|\right) \in (0,1)$,

$$f(0) = 1 + 2\left(1 - \left|\widetilde{\psi}_0|_\tau\right|\right) - 3\sqrt[3]{1 - \left|\widetilde{\psi}_0|_\tau\right|}^2 \geq 3\left[\left(1 - \left|\widetilde{\psi}_0|_\tau\right|\right) - \sqrt[3]{\left(1 - \left|\widetilde{\psi}_0|_\tau\right|\right)}\right] > 0. \tag{A.15}$$

Furthermore, we have for $\vartheta_1(1) \in (0,1)$,

$$0 \geq f(1) = 1 + 4\vartheta_1(1)^3 - 6\vartheta_1(1)^2 \iff 0 \geq (2\vartheta_1(1) - 1)\underbrace{(2\vartheta_1(1)^2 - 2\vartheta_1(1) - 1)}_{<0}$$

$$\iff \vartheta_1(1) \geq \frac{1}{2} \iff \left|\widetilde{\psi}_0|_\tau\right| \leq \frac{3}{4}.$$

Then, due to the continuity of f, there exists $s \in (0,1]$ such that $f(s) = 0$ if $\left|\widetilde{\psi}_0|_\tau\right| \leq 3/4$ and (A.2)–(A.4) are satisfied for this choice of s. Otherwise, equation (A.3) follows from (A.13) and (A.14) with $s = 1$ and only (A.4) remains to be shown. This follows from $\vartheta_1(1)^3 < \vartheta_1(1)^2 \Rightarrow f(1) = 1 + 4\vartheta_1(1)^3 - 6\vartheta_1(1)^2 < 1 - 2\vartheta_1(1)^3 = \left|\widetilde{\psi}_0|_\tau\right|$ and $f(1) > 0$. Here we used (A.9), (A.12) and (A.13).

Let us continue with the case $|\tau \cap \mathcal{V}_k| = 1$. We again consider the triangle in Figure A.1. This time we assume $Z \in \mathcal{V}_k$ and $X, Y \in \mathcal{V}_{k+1}$. As in the previous case, we add a line parallel to XY intersecting XZ in X' and YZ in Y' such that $|X'Z|/|XZ| = |Y'Z|/|YZ| = \vartheta_2 \in (0,1)$. For $\vartheta_2 \in (0,1)$ and $s \in [0,1]$ we define

$$\psi_0 = -\operatorname{sgn}\left(\widetilde{\psi}_0|_\tau\right)s \qquad \text{in } \square(X', X, Y, Y'), \qquad\qquad \psi_0 = \operatorname{sgn}\left(\widetilde{\psi}_0|_\tau\right) \qquad \text{in } \triangle(X', Y', Z). \tag{A.16}$$

We now prove that there exists $s \in [0,1]$ and $\vartheta_2(s)$ such that (A.2)–(A.4) are satisfied. Note that here the sign and scaling factors of ψ_0 are swapped compared to the previous case. By similar considerations as before, we obtain

$$\begin{aligned}\int_\tau \psi_0 \varphi_X \,\mathrm{d}\boldsymbol{x} = \int_\tau \psi_0 \varphi_Y \,\mathrm{d}\boldsymbol{x} &= \operatorname{sgn}\left(\widetilde{\psi}_0|_\tau\right)\frac{(1+s)\vartheta_2^3 - s}{3}A(\tau)\\ \int_\tau \psi_0 \varphi_Z \,\mathrm{d}\boldsymbol{x} &= \operatorname{sgn}\left(\widetilde{\psi}_0|_\tau\right)\frac{(1+s)(3(1-\vartheta_2)\vartheta_2^2 + \vartheta_2^3) - s}{3}A(\tau).\end{aligned} \tag{A.17}$$

Therefore, equation (A.2) is satisfied iff

$$(1+s)(3\vartheta_2^2(s) - 2\vartheta_2^3(s)) - s = \left|\widetilde{\psi}_0\right| \iff 3\vartheta_2^2(s) - 2\vartheta_2^3(s) = \frac{\left|\widetilde{\psi}_0\right| + s}{1+s} \in [0,1]. \tag{A.18}$$

Note that $g(\vartheta_2) = 3\vartheta_2^2 - 2\vartheta_2^3$ is increasing from 0 to 1 with $g(0) = 0$ and $g(1) = 1$ and hence bijective from $[0,1]$ to $[0,1]$. Thus, there exists $\vartheta_2(s)$ satisfying the above equation for any $s \in [0,1]$. Next consider $h(s) = (1+s)\vartheta_2^3(s) - s = 0$ which is a continuous function in s. Clearly, $h(0) \geq 0$. Furthermore, using (A.18) and the monotonicity of $g(\vartheta_2)$, $h(1) \leq 0$ iff

$$\vartheta_2(1) \leq \frac{1}{\sqrt[3]{2}} \iff \frac{3}{\sqrt[3]{4}} - 1 = g\left(\frac{1}{\sqrt[3]{2}}\right) \geq g(\vartheta_2(1)) = \frac{\left|\widetilde{\psi}_0|_\tau\right| + 1}{2} \iff \left|\widetilde{\psi}_0|_\tau\right| \leq \frac{6}{\sqrt[3]{4}} - 3. \tag{A.19}$$

Due to the continuity of h, there exists $s \in [0,1]$ such that $h(s) = 0$ if $\left|\widetilde{\psi}_0|_\tau\right| \leq 6/\sqrt[3]{4} - 3$ and (A.2)–(A.4) are satisfied for this choice of s. Otherwise, equation (A.3) follows from (A.17) and (A.18) with $s = 1$ and only (A.4) remains to be shown. This follows from $\vartheta_2(1)^3 < \vartheta_2(1)^2 \Rightarrow 2\vartheta_2(1)^3 - 1 < 6\vartheta_2(1)^2 - 4\vartheta_2(1)^3 - 1 = \left|\widetilde{\psi}_0|_\tau\right|$. □

Acknowledgements. The authors would like to thank the anonymous reviewers for their helpful comments and suggestions one of which led to the addition of Theorem 5.7.

References

[1] M.S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M.E. Rognes and G.N. Wells, The FEniCS project version 1.5. *Arch. Numer. Softw.* **3** (2015) 100.

[2] R.E. Bank and H. Yserentant, On the H^1-stability of the L_2-projection onto finite element spaces. *Numer. Math.* **126** (2014) 361–381.

[3] I. Cioranescu, Geometry of Banach spaces, duality mappings and nonlinear problems. In: Mathematics and its Applications. Vol. 62, Kluwer Academic Publishers Group, Dordrecht (1990).

[4] M. Crouzeix and V. Thomée, The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces. *Math. Comp.* **48** (1987) 521–532.

[5] L.F. Demkowicz and J. Gopalakrishnan, An overview of the discontinuous Petrov Galerkin method. In: Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations. Vol. 157 of *IMA Vol. Math. Appl.* Springer, Cham (2014) 149–180.

[6] D.L. Donoho, For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** (2006) 907–934.

[7] D.L. Donoho, For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59** (2006) 797–829.

[8] J. Douglas, Jr., T. Dupont and L. Wahlbin, Optimal L_∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems. *Math. Comp.* **29** (1975) 475–483.

[9] J.W. Gibbs, Fourier's series. *Nature* **59** (1899) 606.

[10] J.-L. Guermond, A finite element technique for solving first-order PDEs in Lp. *SIAM J. Numer. Anal.* **42** (2004) 714–737.

[11] J.-L. Guermond and B. Popov, Linear advection with ill-posed boundary conditions via L^1-minimization. *Int. J. Numer. Anal. Model.* **4** (2007) 39–47.

[12] J.-L. Guermond and B. Popov, L^1-approximation of stationary Hamilton-Jacobi equations. *SIAM J. Numer. Anal.* **47** (2008/2009) 339–362.

[13] J.-L. Guermond and B. Popov, An optimal L^1-minimization algorithm for stationary Hamilton-Jacobi equations. *Commun. Math. Sci.* **7** (2009) 211–238.

[14] J.-L. Guermond, F. Marpeau and B. Popov, A fast algorithm for solving first-order PDEs by L^1-minimization. *Commun. Math. Sci.* **6** (2008) 199–216.

[15] P. Houston, S. Roggendorf and K.G. van der Zee, Eliminating Gibbs phenomena: a non-linear Petrov–Galerkin method for the convection-diffusion-reaction equation. *Comput. Math. App.* **80** (2020) 851–873.

[16] B.-N. Jiang, Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method. *J. Comput. Phys.* **105** (1993) 108–121.

[17] B.-N. Jiang, The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics. Springer Science & Business Media (1998).

[18] V. John and P. Knobloch, On the performance of SOLD methods for convection-diffusion problems with interior layers. *Int. J. Comput. Sci. Math.* **1** (2007) 245–258.

[19] V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part I – a review. *Comput. Methods Appl. Mech. Eng.* **196** (2007) 2197–2215.

[20] V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part II – analysis for P1 and Q1 finite elements. *Comput. Methods Appl. Mech. Eng.* **197** (2008) 1997–2014.

[21] D. Landers and L. Rogge, Natural choice of l1-approximants. *J. Approx. Theory* **33** (1981) 268–280.

[22] J.E. Lavery, Nonoscillatory solution of the steady-state inviscid Burgers' equation by mathematical programming. *J. Comput. Phys.* **79** (1988) 436–448.

[23] J.E. Lavery, Solution of steady-state one-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.* **26** (1989) 1081–1089.

[24] J.E. Lavery, Solution of steady-state, two-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.* **28** (1991) 141–155.

[25] J. Li and L. Demkowicz, An L^p-DPG method for the convectiondiffusion problem. *Comput. Math. Appl.* **95** (2021) 172–185.

[26] J. Li and L. Demkowicz, An L^p-DPG Method with Application to 2D Convection-Diffusion Problems. Oden Institute REPORT 202106 (2021).

[27] J.J.H. Miller, Fitted Numerical Methods for Singular Perturbation Problems Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions/J.J.H. Miller, E. O'Riordan and G.I. Shishkin, revised edition. World Scientific, Singapore; River Edge, NJ (2012) (eng).

[28] E. Moskona, P. Petrushev and E.B. Saff, The gibbs phenomenon for best L1-trigonometric polynomial approximation. *Constr. Approx.* **11** (1995) 391–416.

[29] I. Muga and K.G. van der Zee, Discretization of linear problems in banach spaces: residual minimization, nonlinear Petrov-Galerkin, and monotone mixed methods. *SIAM J. Numer. Anal.* **58** (2020) 3406–3426.

[30] I. Muga, M.J.W. Tyler and K.G. van der Zee, The discrete-dual minimal-residual method (DDMRes) for weak advection-reaction problems in Banach spaces. *Comput. Methods Appl. Math.* **19** (2019) 557–579.
[31] G. Nürnberger, Approximation by Spline Functions. Springer-Verlag Berlin Heidelberg (1989).
[32] F.B. Richards, A Gibbs phenomenon for spline functions. *J. Approx. Theory* **66** (1991) 334–351.
[33] S. Roggendorf, *Eliminating the Gibbs phenomenon: the non-linear Petrov–Galerkin method for the convection-diffusion-reaction equation*. Ph.D. thesis, School of Mathematical Sciences, The University of Nottingham (2019).
[34] H.-G. Roos, M. Stynes and L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations, 2nd edition. Vol. 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin (2008). Convection-diffusion-reaction and flow problems.
[35] E.B. Saff and S. Tashev, Gibbs phenomenon for best L_p approximation by polygonal lines. *East J. Approx.* **5** (1999) 235–251.
[36] I. Singer, Best approximation in normed linear spaces by elements of linear subspaces. Translated by R. Georgescu. Vol. 171 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg (1970).
[37] H. Wilbraham, On a certain periodic function. *Cambridge Dublin Math. J.* **3** (1848) 1848.