



# Solving Minimal Residual Methods in $W^{-1,p'}$ with Large Exponents $p$

Johannes Storn<sup>1</sup>

Received: 14 July 2023 / Revised: 3 January 2024 / Accepted: 17 February 2024 /  
Published online: 21 March 2024  
© The Author(s) 2024

## Abstract

We introduce a numerical scheme that approximates solutions to linear PDE's by minimizing a residual in the  $W^{-1,p'}(\Omega)$  norm with exponents  $p > 2$ . The resulting problem is solved by regularized Kačanov iterations, allowing to compute the solution to the non-linear minimization problem even for large exponents  $p \gg 2$ . Such large exponents remedy instabilities of finite element methods for problems like convection-dominated diffusion.

**Keywords** Minimization in  $W^{-1,p'}$  · Minimal residual method · Convection-dominated diffusion · Relaxed Kačanov iteration

**Mathematics Subject Classification** 65N12 · 65N15 · 65N20 · 49M29 · 49M20

## 1 Introduction

Starting with the work of Guermond [9], recent papers like [11, 12, 15, 16, 18] have approximated linear PDE's by minimal residual methods in Banach spaces. The reasons for using Banach spaces like  $L^p(\Omega)$  or  $W_0^{1,p}(\Omega)$  rather than Hilbert spaces like  $L^2(\Omega)$  or  $W_0^{1,2}(\Omega)$  are manifold. For example, rough data might lead to solutions that are not in  $L^2(\Omega)$ , see [11, 16]. Furthermore, minimizing in  $L^1(\Omega)$  seems to allow for computations of viscosity solutions, see [9, Sec. 4.6]. Moreover, classical finite element methods lead for problems like singular perturbed problems or convection-dominated diffusion to non-physical oscillations that can be overcome by the use of minimal residual methods in spaces like  $L^1(\Omega)$ , see for example [12]. Unfortunately, the resulting numerical schemes are non-linear minimization problems which are difficult to solve. We overcome this downside for a minimal residual method in  $W^{-1,p'}(\Omega)$  with  $p > 2$  by introducing a regularized Kačanov scheme that converges even for large exponents  $p \gg 2$  towards the exact discrete minimizer. More precisely, we do the following.

---

✉ Johannes Storn  
jstorn@math.uni-bielefeld.de

<sup>1</sup> Faculty of Mathematics and Computer Science, Institute of Mathematics, Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany

Suppose we have some linear operator  $B: W_0^{1,p'}(\Omega) \rightarrow W^{-1,p'}(\Omega)$  that maps the Sobolev space  $W_0^{1,p}(\Omega)$  equipped with homogeneous Dirichlet boundary data onto the dual space  $W^{-1,p'}(\Omega) := (W_0^{1,p}(\Omega))^*$  with  $1/p + 1/p' = 1$ . Given a right-hand side  $F \in W^{-1,p'}(\Omega)$  and discrete subspaces  $U_h \subset W_0^{1,p'}(\Omega)$  and  $V_h \subset W_0^{1,p}(\Omega)$ , we approximate the solution  $u \in W_0^{1,p'}(\Omega)$  to  $Bu = F$  by a minimizer

$$u_h \in \operatorname{argmin}_{u_h \in U_h} \|Bu_h - F\|_{V_h^*} \text{ with } \|Bu_h - F\|_{V_h^*} = \sup_{v_h \in V_h \setminus \{0\}} \frac{Bu_h(v_h) - F(v_h)}{\|\nabla v_h\|_{L^p(\Omega)}}. \quad (1)$$

For  $p = 2$  the computation of the minimizer in (1) has been discussed in [17, Sec. 3.2]. For  $p > 2$  we modify the saddle point problem therein by introducing a computable weight  $\sigma_n^\zeta$  with values within some relaxation interval  $\zeta = [\zeta_-, \zeta_+] \subset (0, \infty)$  in the sense that  $\sigma_n^\zeta(x) \in \zeta$  for almost all  $x \in \Omega$ . The resulting scheme introduced in (19) seeks  $\psi_{h,n+1} \in V_h$  and  $u_{h,n+1} \in U_h$  with

$$\begin{aligned} \int_{\Omega} (\sigma_n^\zeta)^{2-p'} \nabla \psi_{h,n+1} \cdot \nabla v_h \, dx + Bu_{h,n+1}(v_h) &= F(v_h) \quad \text{for all } v_h \in V_h, \\ Bu_h(\psi_{h,n+1}) &= 0 \quad \text{for all } u_h \in U_h. \end{aligned} \quad (2)$$

Solving this problem allows us to update the weight and to proceed inductively.

To verify the convergence of the iterative scheme, we introduce in Sect. 2 equivalent formulations of the problem in (1) using duality. Since the resulting problems share similarities with the  $p$ -Laplace problem, we can exploit recent ideas for the  $p$ -Laplace operator from [1, 5]. In particular, we introduce a regularization of the dual problem via a relaxation interval  $\zeta = [\zeta_-, \zeta_+]$  and show convergence of the minimizers of the regularized problem towards the exact minimizer as  $\zeta_- \rightarrow 0$  and  $\zeta_+ \rightarrow \infty$  in Sect. 3. We verify the convergence of the Kačanov iterations towards the minimizers of the regularized dual problems in Sect. 4. Additionally, we use duality again to rewrite the Kačanov iterations as a primal problem, leading to the scheme in (2). We conclude our analysis with a study of a priori and a posteriori error estimates in Sect. 5 and suggest an adaptive scheme in Sect. 6. Finally, we study numerically the beneficial properties of the scheme and discuss strategies to solve challenging problems like convection-dominated diffusion with vanishing viscosity in Sect. 7.

## 2 Primal and Dual Formulation

Before we discuss the problem in (1), let us introduce some notation:

- The operator  $B: W_0^{1,p'}(\Omega) \rightarrow W^{-1,p'}(\Omega)$  with exponent  $p > 2$  is a bounded linear mapping, defining a bilinear form  $b(u, v) := Bu(v)$  for all  $u \in W_0^{1,p'}(\Omega)$  and  $v \in W_0^{1,p}(\Omega)$ . Moreover, let  $F \in W^{-1,p'}(\Omega)$  be some given data.
- Given a regular triangulation  $\mathcal{T}$  of the bounded Lipschitz domain  $\Omega \subset \mathbb{R}^d$ , set for all  $\ell \in \mathbb{N}_0$  the space of piece-wise polynomials  $\mathcal{L}_\ell^0(\mathcal{T}) := \{w \in L^2(\Omega): w|_T \text{ is a polynomial of maximal degree } \ell \text{ for all } T \in \mathcal{T}\}$  and set for some fixed degrees  $k, \delta \in \mathbb{N}$  the Lagrange finite element spaces

$$\begin{aligned} U_h &:= \mathcal{L}_{k,0}^1(\mathcal{T}) := \mathcal{L}_k^0(\mathcal{T}) \cap W_0^{1,p'}(\Omega), \\ V_h &:= \mathcal{L}_{k+\delta,0}^1(\mathcal{T}) := \mathcal{L}_{k+\delta}^0(\mathcal{T}) \cap W_0^{1,p}(\Omega). \end{aligned}$$

- Set for all  $G \in W^{-1,p'}(\Omega)$  the discrete dual seminorm

$$\|G\|_{V_h^*} := \sup_{v_h \in V_h \setminus \{0\}} \frac{G(v_h)}{\|\nabla v_h\|_{L^p(\Omega)}}.$$

- Set the subspace  $(BU_h)^\perp := \{v_h \in V_h : b(u_h, v_h) = 0 \text{ for all } u_h \in U_h\} \subset V_h$ .

We can characterize the solution  $u_h \in U_h$  to (1) via the saddle point problem: Seek  $\psi_h \in V_h$  and  $u_h \in U_h$  such that

$$\begin{aligned} \int_{\Omega} |\nabla \psi_h|^{p-2} \nabla \psi_h \cdot \nabla v_h \, dx + b(u_h, v_h) &= F(v_h) \quad \text{for all } v_h \in V_h, \\ b(u_h, \psi_h) &= 0 \quad \text{for all } u_h \in U_h. \end{aligned} \quad (3)$$

A further related problems seeks the minimizer

$$\psi_h = \operatorname{argmin}_{v_h \in (BU_h)^\perp} \frac{1}{p} \int_{\Omega} |\nabla v_h|^p \, dx - F(v_h). \quad (4)$$

**Lemma 1** (Existence and equivalent characterization)

- There exists a unique solution  $\psi_h \in (BU_h)^\perp$  to the minimization problem in (4).
- There exists a solution  $u_h \in U_h$  to the minimization problem in (1). The solution  $u_h$  is unique up to the kernel  $\ker B|_{U_h} := \{u_h \in U_h : Bu_h = 0 \text{ in } V_h^*\}$ .
- The pair  $(\psi_h, u_h) \in V_h \times U_h$  solves (3) if and only if  $u_h \in U_h$  solves (1) and  $\psi_h \in (BU_h)^\perp$  solves (4).

This lemma is shown in [18, Thm. 4.1] within an abstract framework. We give a direct proof utilizing the following statement.

**Lemma 2** (Duality mapping) Let  $G \in V_h^*$ .

- There exists a unique solution  $R(G) \in V_h$  to the problem

$$\int_{\Omega} |\nabla R(G)|^{p-2} \nabla R(G) \cdot \nabla v_h \, dx = G(v_h) \quad \text{for all } v_h \in V_h.$$

- If  $G \neq 0$ , the function  $\|\nabla R(G)\|_{L^p(\Omega)}^{-1} R(G)$  is the unique normed function that attains the supremum in the definition of the  $V_h^*$  norm of  $G$  in the sense that any  $\Theta_h \in V_h$  with  $\|\nabla \Theta_h\|_{L^p(\Omega)} = 1$  satisfies

$$G(\Theta_h) = \|G\|_{V_h^*} \quad \text{if and only if} \quad \Theta_h = \|\nabla R(G)\|_{L^p(\Omega)}^{-1} R(G).$$

**Proof** Let  $G \in V_h^*$ . The direct method in calculus of variations yields the existence of unique minimizers  $R(G) \in V_h$  with

$$R(G) = \operatorname{argmin}_{v_h \in V_h} \frac{1}{p} \int_{\Omega} |\nabla v_h|^p \, dx - G(v_h).$$

Differentiation shows that this existence result is equivalent to the statement in (a).

Let  $G \neq 0$ . Hölder's inequality and testing with  $v_h = R(G)$  shows that

$$\|G\|_{V_h^*} = \sup_{v_h \in V_h \setminus \{0\}} \frac{\int_{\Omega} |\nabla R(G)|^{p-2} \nabla R(G) \cdot \nabla v_h \, dx}{\|\nabla v_h\|_{L^p(\Omega)}} = \|\nabla R(G)\|_{L^p(\Omega)}^{p-1}.$$

This yields  $G(\|\nabla R(G)\|_{L^p(\Omega)}^{-1} R(G)) = \|G\|_{V_h^*}$ . Let  $\Theta_h \in V_h$  with  $\|\nabla \Theta_h\|_{L^p(\Omega)} = 1$  be a further function that attains the supremum in the sense that  $G(\Theta_h) = \|G\|_{V_h^*}$ . The linearity of  $G$  implies

$$\sup_{v_h \in V_h \setminus \{0\}} \frac{G(v_h)}{\|\nabla v_h\|_{L^p(\Omega)}} = \|G\|_{V_h^*} = G\left(\frac{1}{2}\|\nabla R(G)\|_{L^p(\Omega)}^{-1} R(G) + \frac{1}{2}\Theta_h\right),$$

which yields in particular that  $1 \leq \frac{1}{2}\|\nabla R(G)\|_{L^p(\Omega)}^{-1} \nabla R(G) + \frac{1}{2}\nabla \Theta_h\|_{L^p(\Omega)}$ . This estimate and the triangle inequality shows that

$$\begin{aligned} 2 &\leq \left\| \|\nabla R(G)\|_{L^p(\Omega)}^{-1} \nabla R(G) + \nabla \Theta_h \right\|_{L^p(\Omega)} \\ &\leq \left\| \|\nabla R(G)\|_{L^p(\Omega)}^{-1} \nabla R(G) \right\|_{L^p(\Omega)} + \|\nabla \Theta_h\|_{L^p(\Omega)} = 2. \end{aligned}$$

Since  $W_0^{1,p}(\Omega)$  is a strictly convex space [10], this identity yields

$$\Theta_h = \|\nabla R(G)\|_{L^p(\Omega)}^{-1} R(G).$$

□

**Proof of Lemma 1** *Step 1 (Proof of (a) and (b)).* The direct method in calculus of variations yields the existence of unique minimizers  $\psi_h \in (BU_h)^\perp$  of the strictly convex energy in (4), that is, it verifies (a). Similarly, we conclude the existence of a unique minimizer

$$Bu_h = \operatorname{argmin}_{\{Bu_h : u_h \in U_h\}} \|Bu_h - F\|_{V_h^*}.$$

This yields the existence of a minimizer  $u_h \in U_h$  to the problem (1) and shows (b).

*Step 2 (Proof of (c), trivial case).* Let  $u_h \in U_h$  and  $\psi_h \in (BU_h)^\perp$  satisfy (1) and (4). If  $Bu_h = F$  in  $V_h^*$ , the problem in (3) is satisfied with  $\psi_h = 0$  and vice versa.

*Step 3 (Proof of “ $\Leftarrow$ ” in (c)).* Let  $u_h \in U_h$  satisfy (1) and let  $\psi_h \in (BU_h)^\perp$  satisfy (4) with  $Bu_h \neq F$  in  $V_h^*$ . Since  $\{Bu_h : u_h \in U_h\}$  is a closed subspace of  $V_h^*$ , a consequence of the Hahn-Banach theorem (see for example [21, Prop. 3]) yields the existence of a function  $\Theta_h \in V_h$  with  $\|\nabla \Theta_h\|_{L^p(\Omega)} = 1$ ,

$$(F - Bu_h)(\Theta_h) = \|Bu_h - F\|_{V_h^*}, \quad \text{and} \quad Bu_h(\Theta_h) = 0 \quad \text{for all } u_h \in U_h. \quad (5)$$

Lemma 2 characterizes the function  $\Theta_h \in V_h$  due to the first identity in (5) as  $\Theta_h = \|\phi_h\|_{L^p(\Omega)}^{-1} \phi_h$ , where  $\phi_h \in V_h$  solves the problem

$$\int_{\Omega} |\nabla \phi_h|^{p-2} \nabla \phi_h \cdot \nabla v_h \, dx = (F - Bu_h)(v_h) \quad \text{for all } v_h \in V_h. \quad (6)$$

In particular, the function  $\phi_h$  solves

$$\int_{\Omega} |\nabla \phi_h|^{p-2} \nabla \phi_h \cdot \nabla v_h \, dx = F(v_h) \quad \text{for all } v_h \in (BU_h)^\perp.$$

Since this characterizes the minimizer in (4) and  $\phi_h \in (BU_h)^\perp$  due to the second identity in (5), we have  $\phi_h = \psi_h$ . Hence, the functions  $(u_h, \psi_h) \in U_h \times V_h$  solve (3).

*Step 4 (Proof of “ $\Rightarrow$ ” in (c)).* If there exists a solution  $(u_h, \psi_h) \in U_h \times V_h$  to (3), the function  $\psi_h \in V_h$  is an element in  $(BU_h)^\perp$  and satisfies in particular

$$\int_{\Omega} |\nabla \psi_h|^{p-2} \nabla \psi_h \cdot \nabla v_h \, dx = F(v_h) \quad \text{for all } v_h \in (BU_h)^\perp. \quad (7)$$

This identity characterizes the unique (Step 1) solution to (4), that is,  $\psi_h$  must be the minimizer in (4). The solution  $u_h \in U_h$  to (3) is characterized via the identity

$$b(u_h, v_h) = F(v_h) - \int_{\Omega} |\nabla \psi_h|^{p-2} \nabla \psi_h \cdot v_h \, dx \quad \text{for all } v_h \in V_h. \quad (8)$$

Since the right-hand side equals zero for all  $v_h \in (BU_h)^\perp$  due to (7), it is in the range of the operator  $B: U_h \rightarrow V_h^*$ , that is, there exist a unique solution  $u_h \in U_h / \ker B|_{U_h}$  to (8). We know from Step 3 that the solution to (1) solves the problem in (8) as well. The uniqueness of these solutions up to the kernel  $\ker B|_{U_h}$  (Step 1) implies that they must coincide.  $\square$

The minimization problem in (4) shares similarities with the  $p$ -Laplace problem, which can be solved by the regularized Kačanov scheme introduced in [5]. Unfortunately, this schemes converges only for  $p \leq 2$ . We remedy this downside as in [1] by the use of duality. The dual problem of (4) involves the affine space

$$\Sigma := \left\{ \tau \in L^{p'}(\Omega; \mathbb{R}^d) : \int_{\Omega} \nabla v_h \cdot \tau \, dx = F(v_h) \text{ for all } v_h \in (BU_h)^\perp \right\}. \quad (9)$$

It seeks the minimizer to the problem

$$\sigma = \operatorname{argmin}_{\tau \in \Sigma} \frac{1}{p'} \int_{\Omega} |\tau|^{p'} \, dx. \quad (10)$$

Let us show the equivalence of the problems in (4) and (10). The solution to (4) is characterized via the Euler-Lagrange equation as unique solution  $\psi_h \in (BU_h)^\perp$  to

$$\int_{\Omega} |\nabla \psi_h|^{p-2} \nabla \psi_h \cdot \nabla v_h \, dx = F(v_h) \quad \text{for all } v_h \in (BU_h)^\perp. \quad (11)$$

The solution  $\sigma \in L^{p'}(\Omega; \mathbb{R}^d)$  to (10) solves with unique function  $\phi_h \in (BU_h)^\perp$  the saddle point problem

$$\begin{aligned} \int_{\Omega} |\sigma|^{p'-2} \sigma \cdot \tau \, dx - \int_{\Omega} \nabla \phi_h \cdot \tau \, dx &= 0 & \text{for all } \tau \in L^{p'}(\Omega; \mathbb{R}^d), \\ - \int_{\Omega} \nabla v_h \cdot \sigma \, dx &= -F(v_h) & \text{for all } v_h \in (BU_h)^\perp. \end{aligned} \quad (12)$$

**Lemma 3** (Duality) *The solutions to (11) and (12) are related via the identities*

$$\sigma = |\nabla \psi_h|^{p-2} \nabla \psi_h, \quad \nabla \psi_h = |\sigma|^{p'-2} \sigma, \quad \text{and} \quad \phi_h = \psi_h. \quad (13)$$

Furthermore, the minimal energies satisfy

$$\frac{1}{p} \int_{\Omega} |\nabla \psi_h|^p \, dx - F(\psi_h) = -\frac{1}{p'} \int_{\Omega} |\sigma|^{p'} \, dx. \quad (14)$$

**Proof** Let  $\psi_h \in (BU_h)^\perp$  solve (4) and define the functions

$$\sigma := |\nabla \psi_h|^{p-2} \nabla \psi_h \in L^{p'}(\Omega; \mathbb{R}^d) \quad \text{and} \quad \phi_h := \psi_h \in (BU_h)^\perp.$$

Direct calculations show that these functions solve the saddle point problem in (12). Since the solution to (12) is unique (due to the uniqueness of the minimizer  $\sigma$  and the fact that the first line in (12) uniquely determines  $\phi_h$  via the identity  $|\sigma|^{p'-2} \sigma = \nabla \phi_h$ ), we obtain the

equivalence stated in (13). Since  $1/p + 1/p' = 1$  implies with (13) that  $|\nabla \psi_h|^p = |\sigma|^{p'}$ , the identity in (11) yields

$$\frac{1}{p} \int_{\Omega} |\nabla \psi_h|^p dx - F(\psi_h) = \left( \frac{1}{p} - 1 \right) \int_{\Omega} |\nabla \psi_h|^p dx = -\frac{1}{p'} \int_{\Omega} |\sigma|^{p'} dx.$$

This shows (14) and concludes the proof.  $\square$

We want to solve the non-linear problem in (12) via the iterative scheme

$$\begin{aligned} \int_{\Omega} |\sigma_n|^{p'-2} \sigma_{n+1} \cdot \tau dx - \int_{\Omega} \nabla \phi_{h,n+1} \cdot \tau dx &= 0 & \text{for all } \tau \in L^{p'}(\Omega; \mathbb{R}^d), \\ - \int_{\Omega} \nabla v_h \cdot \sigma_{n+1} dx &= -F(v_h) & \text{for all } v_h \in (BU_h)^{\perp}. \end{aligned}$$

However, the resulting problems are in general not well posed since  $\sigma_n$  might degenerate. We thus introduce the following regularization.

### 3 Regularization

Following [5] and [1], we define for any relaxation interval  $\zeta = [\zeta_-, \zeta_+] \subset (0, \infty)$  and all  $t \geq 0$  the integrant

$$\kappa_{\zeta}^*(t) := \begin{cases} \frac{1}{2} \zeta_-^{p'-2} t^2 + \left( \frac{1}{p'} - \frac{1}{2} \right) \zeta_-^{p'} & \text{for } t \leq \zeta_-, \\ \frac{1}{p'} t^{p'} & \text{for } \zeta_- \leq t \leq \zeta_+, \\ \frac{1}{2} \zeta_+^{p'-2} t^2 + \left( \frac{1}{p'} - \frac{1}{2} \right) \zeta_+^{p'} & \text{for } \zeta_+ \leq t. \end{cases}$$

We furthermore define for all  $\tau \in L^{p'}(\Omega; \mathbb{R}^d)$  the energies

$$\mathcal{J}_{\zeta}^*(\tau) := \int_{\Omega} \kappa_{\zeta}^*(|\tau|) dx \quad \text{and} \quad \mathcal{J}^*(\tau) := \frac{1}{p'} \int_{\Omega} |\tau|^{p'} dx.$$

Notice that the regularized energy  $\mathcal{J}_{\zeta}^*(\tau_h)$  equals infinity if  $\tau \in L^{p'}(\Omega; \mathbb{R}^d) \setminus L^2(\Omega; \mathbb{R}^d)$ . Furthermore, the relaxed energy is monotone with respect to the relaxation interval in the sense that all  $\tau \in L^{p'}(\Omega; \mathbb{R}^d)$  and relaxation intervals  $\zeta^2 = [\zeta_-^2, \zeta_+^2] \subset \zeta^1 = [\zeta_-^1, \zeta_+^1] \subset (0, \infty)$  satisfy

$$\mathcal{J}^*(\tau) \leq \mathcal{J}_{\zeta_1}^*(\tau) \leq \mathcal{J}_{\zeta_2}^*(\tau).$$

The direct method in calculus of variations verifies the existence of a unique minimizer  $\sigma_{\zeta}$  of  $\mathcal{J}_{\zeta}^*$  in  $\Sigma$  in the sense that

$$\sigma_{\zeta} = \operatorname{argmin}_{\tau \in \Sigma} \mathcal{J}_{\zeta}^*(\tau). \quad (15)$$

In the following we investigate the convergence of  $\sigma_{\zeta}$  towards the minimizer  $\sigma \in \Sigma$  in (10). Rather than investigating convergence in the  $L^p(\Omega)$  norm, we investigate the convergence of the energies. This energy difference leads to the following bound.

**Lemma 4** (Notion of distance) *Let  $\sigma \in \Sigma$  be the minimizer in (10) and let  $\tau \in \Sigma$ . Then we have*

$$\| |\sigma| + |\sigma - \tau| \|_{L^{p'}(\Omega)}^{p'-2} \| \sigma - \tau \|_{L^{p'}(\Omega)}^2 \lesssim \mathcal{J}^*(\tau) - \mathcal{J}^*(\sigma) \lesssim \| \sigma - \tau \|_{L^{p'}(\Omega)}^{p'}.$$

Furthermore, we have the lower bound

$$\int_{\Omega} (|\tau| + |\sigma - \tau|)^{\frac{p'-2}{2}} \|\sigma - \tau\|_{L^{p'}(\Omega)}^2 \lesssim \mathcal{J}^*(\tau) - \mathcal{J}^*(\sigma).$$

The hidden constants depend on  $p$  but are independent of the solution  $\sigma$ .

**Proof** Since this result is well-known in the context of the  $p$ -Laplacian, let us briefly summarize its derivation. Let  $\sigma$  and  $\tau$  be as in the lemma. Since  $(\mathcal{J}^*)'(\sigma)(\tau - \sigma) = 0$  due to the minimization property of  $\sigma$ , the convexity of  $\mathcal{J}^*$  yields

$$\begin{aligned} \mathcal{J}^*(\tau) - \mathcal{J}^*(\sigma) &\leq (\mathcal{J}^*)'(\tau)(\tau - \sigma) = ((\mathcal{J}^*)'(\tau) - (\mathcal{J}^*)'(\sigma))(\tau - \sigma) \\ &= \int_{\Omega} (|\tau|^{p'-2}\tau - |\sigma|^{p'-2}\sigma) \cdot (\tau - \sigma) \, dx. \end{aligned}$$

Further arguments for the integrand as for example shown in [5, Lem. 42] lead to the lower bound

$$\int_{\Omega} (|\tau|^{p'-2}\tau - |\sigma|^{p'-2}\sigma) \cdot (\tau - \sigma) \, dx \lesssim \mathcal{J}^*(\tau) - \mathcal{J}^*(\sigma).$$

Additionally, the equivalence  $(|P|^{p'-2}P - |Q|^{p'-2}Q) \cdot (P - Q) \approx (|Q| + |P - Q|)^{p'-2}|P - Q|^2$  for all  $P, Q \in \mathbb{R}^d$  as shown in [5, Lem. 39] implies

$$\int_{\Omega} (|\tau|^{p'-2}\tau - |\sigma|^{p'-2}\sigma) \cdot (\tau - \sigma) \, dx \approx \int_{\Omega} (|\sigma| + |\sigma - \tau|)^{p'-2} |\sigma - \tau|^2 \, dx. \quad (16)$$

These observations lead to the upper bound in the lemma. The lower bound follows from Hölder's reverse inequality

$$\begin{aligned} &\left( \int_{\Omega} (|\sigma| + |\sigma - \tau|)^{p'-2} \, dx \right)^{1-q} \left( \int_{\Omega} |\sigma - \tau|^{\frac{2}{q}} \, dx \right)^q \\ &\leq \int_{\Omega} (|\sigma| + |\sigma - \tau|)^{p'-2} |\sigma - \tau|^2 \, dx \quad \text{with } q := \frac{2}{p'}. \end{aligned}$$

Exchanging the role of  $\sigma$  and  $\tau$  in (16) leads to the alternative lower bound.  $\square$

We have the following convergence result for the energy differences.

**Proposition 5** (Convergence in  $\zeta$ ) *Let  $\psi_h \in V_h$  denote the solution to (4) and let  $\sigma$  and  $\sigma_{\zeta}$  denote the minimizers in (10) and (15), respectively. Their energy difference is bounded for all relaxation intervals  $\zeta = [\zeta_-, \zeta_+] \subset (0, \infty)$  and  $r > 2$  by*

$$\begin{aligned} \mathcal{J}^*(\sigma_{\zeta}) - \mathcal{J}^*(\sigma) &\leq \mathcal{J}_{\zeta}^*(\sigma_{\zeta}) - \mathcal{J}^*(\sigma) \leq \frac{|\Omega|}{p'} \zeta_-^{p'} + \frac{1}{p'} \zeta_+^{-(r-p')} \|\sigma\|_{L^r(\Omega)}^r \\ &= \frac{|\Omega|}{p'} \zeta_-^{p'} + \frac{1}{p'} \zeta_+^{-(r-p')} \|\nabla \psi_h\|_{L^{r(p-1)}(\Omega)}^{r(p-1)}. \end{aligned}$$

**Proof** This first two inequalities follow as in [1, Thm. 3.1]. Since  $|\nabla \psi_h|^p = |\sigma|^{p'}$  due to Lemma 3, the equality then follows from the identity

$$\int_{\Omega} |\sigma|^r \, dx = \int_{\Omega} |\sigma|^{p' \frac{r}{p'}} = \int_{\Omega} |\nabla \psi_h|^r \frac{p'}{p'} \, dx = \int_{\Omega} |\nabla \psi_h|^{r(p-1)} \, dx.$$

$\square$

**Remark 6** (*Regularity*) The convergence result in Proposition 5 assumes the regularity property  $\psi_h \in W_0^{1,r(p-1)}(\Omega)$ . Such a result is indeed true for all  $r \leq \infty$ , since  $\psi_h \in V_h = \mathcal{L}_{k+\delta,0}^1(T)$  is a function in a finite dimensional space. However, the norm might increase as the mesh is refined. In practical computations this issue does not seem to cause problems, since we can control the impact of the regularization by comparing the energies  $\mathcal{J}_\zeta^*(\tau)$  and  $\mathcal{J}^*(\tau)$ , cf. Sect. 6, and our numerical experiments in Sect. 7 do not indicate a significantly decreased rate of convergence.

## 4 Relaxed Kačanov Scheme

In this section we introduce our iterative scheme that converges towards the minimizer  $\sigma_\zeta$  in (15) with relaxation interval  $\zeta = [\zeta_-, \zeta_+] \subset (0, \infty)$ . Set  $b \vee c := \max\{b, c\}$  and  $b \wedge c := \min\{b, c\}$  for all  $b, c \in \mathbb{R}$ . Given some initial value  $\sigma_0 \in L^{p'}(\Omega; \mathbb{R}^d)$ , we compute iteratively for any  $n \in \mathbb{N}_0$  the solution  $\sigma_{n+1} \in L^{p'}(\Omega; \mathbb{R}^d)$  and  $\psi_{h,n+1} \in (BU_h)^\perp$  satisfying for all  $\tau \in L^{p'}(\Omega; \mathbb{R}^d)$  and  $v_h \in (BU_h)^\perp$

$$\begin{aligned} \int_{\Omega} (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{p'-2} \sigma_{n+1} \cdot \tau \, dx - \int_{\Omega} \nabla \psi_{h,n+1} \cdot \tau \, dx &= 0, \\ - \int_{\Omega} \nabla v_h \cdot \sigma_{n+1} \, dx &= -F(v_h). \end{aligned} \quad (17)$$

The following proposition shows convergence of the solutions  $\sigma_n$  towards the minimizer  $\sigma_\zeta$  in (15).

**Proposition 7** (Convergence) *We have with constant  $\rho \lesssim (\zeta_-/\zeta_+)^{2-p'}$  the linear convergence result*

$$\mathcal{J}_\zeta^*(\sigma_{n+1}) - \mathcal{J}_\zeta^*(\sigma_\zeta) \leq (1 - \rho)^n (\mathcal{J}_\zeta^*(\sigma_0) - \mathcal{J}_\zeta^*(\sigma_\zeta)) \quad \text{for all } n \in \mathbb{N}.$$

**Proof** This result follows as in [1, Sec. 4].  $\square$

To solve the problem in (17), we utilize duality to obtain a primal problem which seeks  $\psi_{h,n+1} \in (BU_h)^\perp$  such that for all  $v_h \in (BU_h)^\perp$

$$\int_{\Omega} (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1} \cdot \nabla v_h \, dx = F(v_h). \quad (18)$$

The corresponding saddle point problem seeks  $\psi_{h,n+1} \in V_h$  and  $u_{h,n+1} \in U_h$  with

$$\begin{aligned} \int_{\Omega} (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1} \cdot \nabla v_h \, dx + b(u_{h,n+1}, v_h) &= F(v_h) \quad \text{for all } v_h \in V_h, \\ b(u_h, \psi_{h,n+1}) &= 0 \quad \text{for all } u_h \in U_h. \end{aligned}$$

**Proposition 8** (Equivalence) *The solution  $\psi_{h,n+1} \in (BU_h)^\perp$  to (18) and  $\sigma_{n+1} \in \Sigma$  to (17) are related via the identity*

$$\sigma_{n+1} = (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1}.$$

**Proof** The same arguments as in the proof of Lemma 3 yield the proposition.  $\square$

The problem in (18) can be solved iteratively. More precisely, we compute with given initial value  $\sigma_0 \in L^{p'}(\Omega; \mathbb{R}^d)$  for all  $n \in \mathbb{N}_0$  iteratively the solution  $\psi_{h,n+1} \in (BU_h)^\perp$  to the



problem in (18) and define  $\sigma_{n+1} := (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1}$ . Notice that the problem in (18) can be equivalently reformulated as the saddle point problem in (2), that is, we seek  $\psi_{h,n+1} \in V_h$  and  $u_{h,n+1} \in U_h$  such that

$$\begin{aligned} \int_{\Omega} (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1} \cdot \nabla v_h \, dx + b(u_{h,n+1}, v_h) &= F(v_h) \quad \text{for all } v_h \in V_h, \\ b(u_h, \psi_{h,n+1}) &= 0 \quad \text{for all } u_h \in U_h. \end{aligned} \quad (19)$$

Due to Proposition 8 the resulting functions  $\sigma_{n+1}$  are the same as the functions obtained by (17) and thus the sequence converges according to Proposition 7 towards the exact discrete minimizer  $\sigma_{\zeta}$ .

**Remark 9** (*Finer mesh*) The contribution  $\sigma_{n+1} = (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1}$  is in general not a polynomial. In the experiments shown below we thus simplified our computations by using the  $L^2(\Omega)$  orthogonal projection  $\Pi_{k+\delta-1} \sigma_{n+1}$  onto the space of piece-wise polynomials of maximal degree  $k + \delta - 1$ . Alternatively, it is possible to use instead of a higher polynomial degree  $k + \delta$  the space  $V_h = \mathcal{L}_{1,0}^1(\mathcal{T}^+)$  with a sufficiently fine refinement of  $\mathcal{T}^+$  of  $\mathcal{T}$ . In this case the contribution  $(\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1}$  is a piece-wise constant polynomial with respect to the underlying triangulation  $\mathcal{T}^+$  which allows for the exact evaluation of  $\sigma_{n+1}$ .

## 5 Error Control

The a priori and a posteriori error control for minimal residual methods is well established, see for example [3, 4, 17–19]. Let us briefly adapt the proofs therein to our situation. We assume that

- (a) there exists a unique solution  $u \in W_0^{1,p'}(\Omega)$  with  $Bu = F$  in  $W^{-1,p'}(\Omega)$  and
- (b) there exist a Fortin operator  $\Pi: W_0^{1,p}(\Omega) \rightarrow V_h$  with continuity constant  $\|\Pi\| < \infty$  in the sense that for all  $u_h \in U_h$  and  $v \in W_0^{1,p}(\Omega)$

$$b(u_h, v - \Pi v) = 0 \quad \text{and} \quad \|\nabla \Pi v\|_{L^p(\Omega)} \leq \|\Pi\| \|\nabla v\|_{L^p(\Omega)}. \quad (20)$$

**Proposition 10** (Error control) *Suppose that the assumptions in (a) and (b) are true. Then the solution  $u$  to  $Bu = F$  in  $W^{-1,p'}(\Omega)$  and  $u_h \in U_h$  to (1) satisfy*

$$\|Bu - Bu_h\|_{W^{-1,p'}(\Omega)} \leq (1 + 2 \|\Pi\|) \min_{u_h \in U_h} \|Bu - Bu_h\|_{W^{-1,p'}(\Omega)}. \quad (21)$$

Moreover, with oscillation  $\text{osc}(F) := \sup_{v \in W^{1,p}(\Omega) \setminus \{0\}} F(v - \Pi v) / \|\nabla v\|_{L^p(\Omega)}$  we have for any  $u_h \in U_h$  the a posteriori error estimate

$$\|Bu - Bu_h\|_{W^{-1,p'}(\Omega)} \leq \|\Pi\| \|Bu - Bu_h\|_{V_h^*} + \text{osc}(F). \quad (22)$$

**Proof** Let  $u_h \in U_h$ . Any  $v \in W_0^{1,p}(\Omega)$  with  $\|\nabla v\|_{L^p(\Omega)} = 1$  satisfies

$$b(u - u_h, v) = b(u - u_h, \Pi v) + b(u - u_h, v - \Pi v) \leq \|\Pi\| \|Bu - Bu_h\|_{V_h^*} + \text{osc}(F).$$

This proves the a posteriori estimate in (22). To obtain the a priori estimate in (21), we use the minimization property in (1), that is,

$$\|Bu - Bu_h\|_{V_h^*} = \min_{u_h \in U_h} \|Bu - Bu_h\|_{V_h^*} \leq \min_{u_h \in U_h} \|Bu - Bu_h\|_{W^{-1,p'}(\Omega)}.$$

Moreover, the oscillation satisfies

$$\begin{aligned} \text{osc}(F) &= \sup_{v \in W^{1,p}(\Omega) \setminus \{0\}} \frac{F(v - \Pi v)}{\|\nabla v\|_{L^p(\Omega)}} = \min_{u_h \in U_h} \sup_{v \in W^{1,p}(\Omega) \setminus \{0\}} \frac{(Bu - Bu_h)(v - \Pi v)}{\|\nabla v\|_{L^p(\Omega)}} \\ &\leq (1 + \|\Pi\|) \|Bu - Bu_h\|_{W^{-1,p'}(\Omega)}. \end{aligned}$$

Combining these estimates with (22) concludes the proof of (21).  $\square$

**Corollary 11** (A posteriori for exact solution) *Let  $(\psi_h, u_h) \in V_h \times U_h$  solve (3) and assume that the assumptions in (a) and (b) are satisfied. Moreover, let  $\sigma \in \Sigma$  denote the solution to (10). Then we have the a posteriori error estimate*

$$\|Bu - Bu_h\|_{W^{-1,p'}(\Omega)} \lesssim \|\nabla \psi_h\|_{L^p(\Omega)}^{p-1} + \text{osc}(F) = \|\sigma\|_{L^{p'}(\Omega)} + \text{osc}(F). \quad (23)$$

**Proof** Hölder's inequality, the first equation in (3), and the identity in (13) show

$$\|Bu - Bu_h\|_{V_h^*} = \sup_{v_h \in V_h \setminus \{0\}} \frac{\int_{\Omega} |\nabla \psi_h|^{p-2} \nabla \psi_h \cdot \nabla v_h \, dx}{\|\nabla v_h\|_{L^p(\Omega)}} = \|\nabla \psi_h\|_{L^p(\Omega)}^{p-1} = \|\sigma\|_{L^{p'}(\Omega)}.$$

Using this identity in the a posteriori estimate in (22) leads to the upper bound in (23). Equivalence follows from the upper bound for the oscillation

$$\begin{aligned} \text{osc}(F) &= \sup_{v \in W_0^{1,p}(\Omega)} \frac{F(v - \Pi v)}{\|\nabla v\|_{L^p(\Omega)}} = \sup_{v \in W_0^{1,p}(\Omega)} \frac{b(u - u_h, v - \Pi v)}{\|\nabla v\|_{L^p(\Omega)}} \\ &\leq (1 + \|\Pi\|) \|Bu - Bu_h\|_{W^{-1,p'}(\Omega)}. \end{aligned}$$

$\square$

We conclude this section with a discussion of the following additional assumption:

- (c) The operator  $B: W_0^{1,p'}(\Omega) \rightarrow W^{-1,p'}(\Omega)$  is bounded from above and below in the sense that

$$\|\nabla u\|_{L^{p'}(\Omega)} \lesssim \|Bu\|_{W^{-1,p'}(\Omega)} \quad \text{for all } u \in W_0^{1,p'}(\Omega).$$

Under this additional assumption the error estimates in Proposition 10 and Corollary 11 allow for any estimate of the more natural error quantity  $\|\nabla u - \nabla u_h\|_{L^{p'}(\Omega)}$  due to the equivalence

$$\|\nabla u - \nabla u_h\|_{L^{p'}(\Omega)} \lesssim \|Bu - Bu_h\|_{W^{-1,p'}(\Omega)}.$$

The assumption in (a) seems to be natural. The assumption in (b) can in many situations be achieved by choosing sufficiently large polynomial degrees  $k + \delta$  for the test space  $V_h = \mathcal{L}_{k+\delta}^1(\mathcal{T})$  as for example investigated in [17, Sec. 4]. The assumption in (c) has been investigated in [11] but seems to be rather restrictive. Indeed, there exist counterexamples for the Laplace problem  $Bu = \int_{\Omega} \nabla u \cdot \nabla \cdot dx$  for exponents  $p > 4$  and non-smooth non-convex domains  $\Omega$  as shown in [13]. Notice that even in cases where (c) is satisfied, the Galerkin scheme investigated in [11] requires stability of the  $W_0^{1,2}(\Omega)$ -projection in  $W_0^{1,p}(\Omega)$ . Such stability results are known for uniform and mildly graded meshes [6, 7], but are an open problem for adaptively refined meshes. Our minimal residual method circumvents this problem by suitable designs of Fortin operators in (b).

## 6 Adaptive Scheme

As pointed out in Corollary 11, the minimizer  $\sigma$  with (10) allows us to drive an adaptive mesh refinement scheme. However, our iterative scheme does not compute the exact solution  $\sigma$ . We thus introduce an adaptive scheme that additionally takes the distance of the current iterate  $\sigma_n$  to  $\sigma$  into account. The error indicator that indicate errors caused by

- (a) the upper interval bound  $\zeta_+$  reads  $\eta_{\zeta_+}^2(\sigma_n) := \mathcal{J}_{\zeta}^*(\sigma_n) - \mathcal{J}_{[\zeta_-, \infty)}^*(\sigma_n)$ ,
- (b) the lower interval bound  $\zeta_-$  reads  $\eta_{\zeta_-}^2(\sigma_n) := \mathcal{J}_{\zeta}^*(\sigma_n) - \mathcal{J}_{[0, \zeta_+]}^*(\sigma_n)$ ,
- (c) the error due to the fixed-point iteration reads

$$\eta_{\text{Ka}\check{\text{c}}, \zeta}^2(\sigma_n) := \left( \frac{\zeta_+}{\zeta_-} \right)^{2-p'} (\mathcal{J}_{\zeta}^*(\sigma_n) - \mathcal{J}_{\zeta}^*(\sigma_{n+1})),$$

- (d) the error due to the discretization reads

$$\eta_h^{p'} := \sum_{T \in \mathcal{T}} \eta_h^{p'}(T) \quad \text{with} \quad \eta_h^{p'}(T) := \|\sigma_n\|_{L^{p'}(T)}^{p'}.$$

The indicators in (a) and (b) provide some information on the impact of the relaxation interval  $\zeta$  on the current iterate. The indicator in (c) is motivated by the convergence result in Proposition 7. The error indicator in (d) is motivated by the a posteriori error estimate for  $\sigma$  in Corollary 11. Notice that  $\sigma_n$  is indeed a good approximation of  $\sigma$  if  $\|\sigma - \sigma_n\|_{L^{p'}(\Omega)} \ll \|\sigma\|_{L^{p'}(\Omega)}$ , which can be seen by the triangle inequality

$$|\|\sigma_n\|_{L^{p'}(\Omega)} - \|\sigma - \sigma_n\|_{L^{p'}(\Omega)}| \leq \|\sigma\|_{L^{p'}(\Omega)} \leq \|\sigma_n\|_{L^{p'}(\Omega)} + \|\sigma - \sigma_n\|_{L^{p'}(\Omega)}.$$

Lemma 4 states that

$$\|\sigma - \sigma_n\|_{L^{p'}(\Omega)}^2 \lesssim \|\sigma_n\| + \|\sigma - \sigma_n\|_{L^{p'}(\Omega)}^{2-p'} (\mathcal{J}(\sigma_n) - \mathcal{J}(\sigma)).$$

Hence, the estimate  $\|\sigma - \sigma_n\|_{L^{p'}(\Omega)} \ll \|\sigma\|_{L^{p'}(\Omega)}$  follows from an estimate like

$$\mathcal{J}(\sigma_n) - \mathcal{J}(\sigma) \ll \|\sigma_n\| + \|\sigma - \sigma_n\|_{L^{p'}(\Omega)}^{p'-2} \|\sigma_n\|_{L^{p'}(\Omega)}^2 \leq \|\sigma_n\|_{L^{p'}(\Omega)}^{p'} = \eta_h^{p'}.$$

This motivates the following refinement strategy with some small weight  $w > 0$ :

- (a) If  $\eta_{\zeta_+}^2(\sigma_n) + \eta_{\zeta_-}^2(\sigma_n) + \eta_{\text{Ka}\check{\text{c}}, \zeta}^2(\sigma_n) \leq w \eta_h^{p'}$ , refine the mesh adaptively with the local error contributions  $\eta_h^{p'}(T)$  as refinement indicator.
- (b) Otherwise, if  $\max\{\eta_{\zeta_-}^2(\sigma_n), \eta_{\text{Ka}\check{\text{c}}, \zeta}^2(\sigma_n)\} \leq \eta_{\zeta_+}^2(\sigma_n)$ , increase  $\zeta_+$ .
- (c) Otherwise, if  $\max\{\eta_{\zeta_+}^2(\sigma_n), \eta_{\text{Ka}\check{\text{c}}, \zeta}^2(\sigma_n)\} \leq \eta_{\zeta_-}^2(\sigma_n)$ , decrease  $\zeta_-$ .

Then we perform another Kačanov iteration and continue with the evaluation of the resulting error indicators. This leads to an adaptive loop which solves (19) and then might refine the mesh or adapt  $\zeta$  according to (a)–(c). Then the loop resets and proceeds with solving (19) again.

**Remark 12** (Primal-dual error estimator) In [1, Sec. 6.2] we use the dual problem with energy  $\mathcal{J}_{\zeta}$  of the minimization problem in (10) to define the estimator

$$\eta_{\text{Ka}\check{\text{c}}, \zeta, \text{Dual}}^2(\sigma_n) := \mathcal{J}_{\zeta}(\psi_{h,n}) + \mathcal{J}_{\zeta}^*(\sigma_n). \quad (24)$$

This error estimator is a guaranteed upper bound for the error

$$\mathcal{J}_{\zeta}^*(\sigma_n) - \mathcal{J}_{\zeta}^*(\sigma_{\zeta}) \leq \eta_{\text{Ka}\check{\text{c}}, \zeta, \text{Dual}}^2(\sigma_n).$$

However, in [1] we focused on a lowest-order scheme in the sense that  $V_h = \mathcal{L}_{1,0}^1(\Omega)$ , which allows for accurate evaluations of  $\sigma_{n+1} = (\zeta_- \vee |\sigma_n| \wedge \zeta_+)^{2-p'} \nabla \psi_{h,n+1}$  with piecewise constant function  $\sigma_n$ . Since in this paper's minimal residual method the space  $V_h$  is of higher polynomial degree, the evaluation of  $\sigma_{n+1}$  becomes more intricate, cf. Remark 9. Our numerical experiments indicate that this challenge does not impact the convergence of the Kačanov scheme, but it causes difficulties when evaluating the duality gap in (24). We thus use the alternative indicator in (c). Replacing the test space  $V_h$  of higher polynomial degree by a test space  $V_h = \mathcal{L}_{1,0}^1(\mathcal{T}^+)$  with finer mesh  $\mathcal{T}^+ \geq \mathcal{T}$  as discussed in Remark 9 circumvents these issues and thus allows us to use of the primal-dual error estimator in (24).

**Remark 13** (*Cheaper approaches*) The adaptive loop suggested in this section is much more costly than the adaptive scheme with linear minimal residual methods for  $p = 2$ . This might be a price we have to pay the solve challenging PDE's. On the other hand, for less challenging problems, we might use cheaper versions of the suggested scheme. For example, the scheme performed well in our experiments with fixed relaxation interval and a fixed small number of Kačanov iterations after each mesh refinement as for example done in Sect. 7.3. Alternatively, one might use our scheme only in the last step of an adaptive finite element loop to smoothen oscillations.

## 7 Applications

We conclude this paper with an application of our algorithm to convection-diffusion problems. Given a bounded Lipschitz domain  $\Omega \subset \mathbb{R}^d$ , a diffusion coefficient  $\varepsilon > 0$ , an incompressible advection field  $\beta \in L^\infty(\Omega; \mathbb{R}^d)$ , a function  $c \in L^\infty(\Omega)$ , and a right-hand side  $f \in L^2(\Omega)$ , this problems seeks  $u \in W_0^{1,2}(\Omega)$  with

$$-\operatorname{div}(\varepsilon \nabla u - \beta u) + cu = f. \quad (25)$$

Set for all  $u, v \in W_0^{1,2}(\Omega)$  the functional  $F(v) := \int_\Omega f v \, dx$  and the bilinear form

$$b(u, v) := \int_\Omega \varepsilon \nabla u \cdot \nabla v \, dx - \int_\Omega u \beta \cdot \nabla v \, dx + \int_\Omega cuv \, dx.$$

The variational formulation of (25) seeks the solution  $u \in W_0^{1,2}(\Omega)$  to

$$b(u, v) = F(v) \quad \text{for all } v \in W_0^{1,2}(\Omega). \quad (26)$$

This formulation allows for the application of our minimal residual method. We therefore discretize the spaces  $W_0^{1,p'}(\Omega)$  and  $W_0^{1,p}(\Omega)$  with  $p := 100$  by

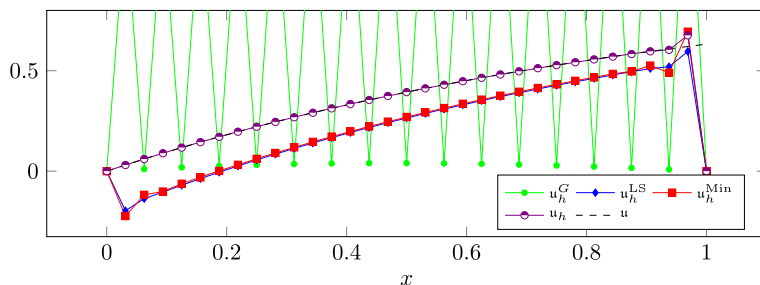
$$U_h := \mathcal{L}_{1,0}^1(\mathcal{T}) \quad \text{and} \quad V_h := \mathcal{L}_{2,0}^1(\mathcal{T}).$$

Suitable Fortin operators (20), which might in fact require higher polynomial degrees in  $V_h$ , are discussed in [17, Sec. 4]. To compare the results, we apply the following alternative schemes:

- (a) Our first alternative numerical scheme is the classical Galerkin FEM. It seeks the solution  $u_h^G \in U_h$  to the problem

$$b(u_h^G, w_h) = F(w_h) \quad \text{for all } w_h \in U_h.$$

Adaptive mesh refinements are driven by the standard residual error estimator investigated for example in [20, Sec. 1.2].



**Fig. 1** Approximations of the viscosity solution  $u$  to (27)

- (b) The second alternative is the classical first-order system least squares method [2] with Raviart-Thomas space  $RT_0(T) := \{q \in H(\operatorname{div}, \Omega) : \text{for all } T \in \mathcal{T} \text{ exist } A \in \mathbb{R}^d \text{ and } b \in \mathbb{R} \text{ with } q(x)|_T = A + bx \text{ for all } x \in T\}$ . It seeks the minimizer  $(u_h^{\text{LS}}, \sigma_h^{\text{LS}}) \in U_h \times RT_0(\mathcal{T})$  that minimizes over all  $(u_h, \tau_h) \in U_h \times RT_0(\mathcal{T})$  the functional

$$\|\tau_h - \varepsilon \nabla u_h + \beta u_h\|_{L^2(\Omega)}^2 + \|\operatorname{div} \tau_h - cu_h + f\|_{L^2(\Omega)}^2.$$

Adaptive mesh refinements are driven by the local contributions

$$\|\sigma_h^{\text{LS}} - \varepsilon \nabla u_h^{\text{LS}} + \beta u_h^{\text{LS}}\|_{L^2(T)}^2 + \|\operatorname{div} \sigma_h^{\text{LS}} - cu_h^{\text{LS}} + f\|_{L^2(T)}^2 \quad \text{for all } T \in \mathcal{T}.$$

- (c) The third alternative is the minimal residual method introduced in [17, Example 2.2 (i)], which seeks the solution to the minimization problem

$$u_h^{\text{Min}} = \operatorname{argmin}_{u_h \in U_h} \sup_{v_h \in V_h \setminus \{0\}} \frac{b(u_h, v_h) - F(v_h)}{\|\nabla v_h\|_{L^2(\Omega)}}.$$

Adaptive mesh refinements are driven by the local contributions  $\|\eta_h\|_{L^2(T)}^2$  for all  $T \in \mathcal{T}$  of the Riesz representative  $\eta_h \in V_h$  with

$$\int_{\Omega} \nabla \eta_h \cdot \nabla v_h \, dx = b(u_h, v_h) - F(v_h) \quad \text{for all } v_h \in V_h.$$

The adaptive schemes use the Dörfler marking strategy with bulk parameter 0.5.

## 7.1 Experiment 1 (Viscosity Solution 1D)

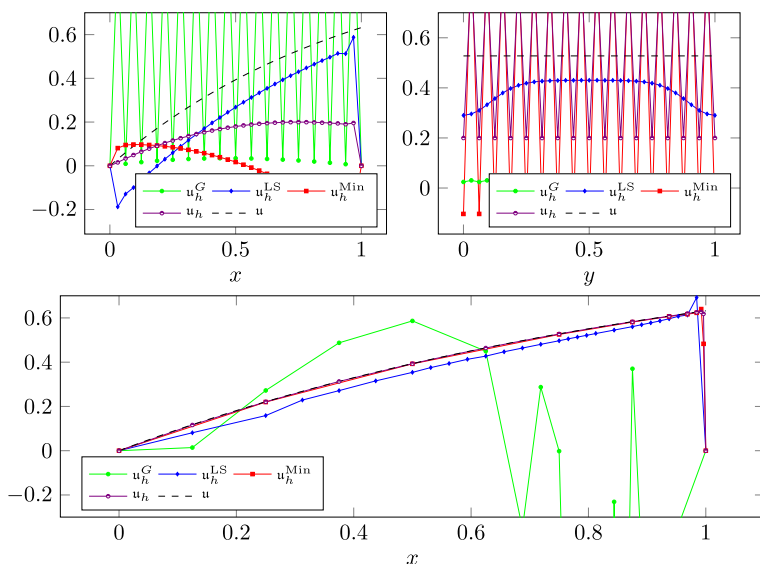
Our first experiment considers the one dimensional problem

$$u' + u = 1 \text{ in } \Omega := (0, 1) \quad \text{with} \quad u(0) = u(1) = 0. \quad (27)$$

In other words, we solve the problem in (25) with  $\varepsilon = 0$  and  $\beta = c = 1$ . This overdetermined ODE has no classical solution, but can be seen as the limiting case  $\varepsilon \rightarrow 0$  of the problem

$$-\varepsilon u'' + u' + u = 1 \text{ in } \Omega \quad \text{with} \quad u_\varepsilon(0) = u_\varepsilon(1) = 0.$$

These functions  $u_\varepsilon$  converge towards the viscosity solution  $u(x) = 1 - \exp(-x)$ , cf. [14, Chap. 1] and [9, Sec. 4.6]. Figure 1 displays the resulting approximations on a partition of the unit interval into  $2^5$  equidistant intervals. The solution  $u_h$  to (1) yields, apart from a tiny oscillation on the last intervals, a very accurate approximation of the viscosity solution. In contrast, the Galerkin FEM results in a highly oscillating function that does not resemble any



**Fig. 2** Approximations with about 1000 degrees of freedom of the viscosity solution  $u$  with (28) evaluated at  $(x, 1/2)$  (top left) and  $(3/4, y)$  (top right) with uniform mesh refinements and at  $(x, 1/2)$  (bottom) with adaptive mesh refinement

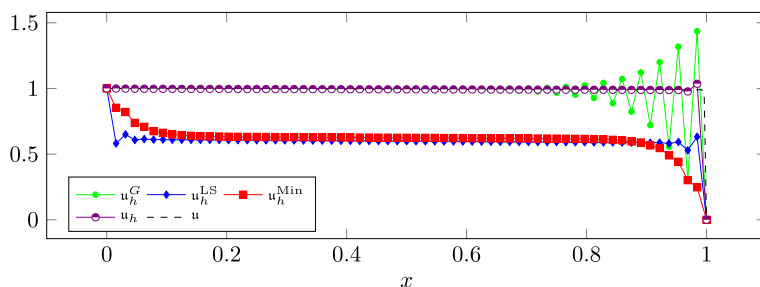
of the solutions characteristics at all. The solution to the minimization methods in (b) and (c) experiences some fast decay on the first interval. Thereafter, the approximation increase and experience a similar (but stronger) oscillation at the last two intervals. The accuracy of the solutions to (a)–(c) does not improve under uniform mesh refinement. Adaptive mesh refinements, driven by the local residuals of these methods, overcome this problem partially for the methods in (b)–(c). Apart from an oscillation near  $x = 1$ , the adaptive LSFEM shows some small oscillation near the origin  $x = 0$  and the adaptive minimal residual method (c) shows an oscillation near  $x = 1/2$ . This indicates severe difficulties of the methods in (a)–(c) for problems with small viscosity parameter  $\varepsilon \ll 1$ , that can be overcome by the minimal residual method (1) in this paper.

## 7.2 Experiment 2 (Viscosity Solution 2D)

In our second experiment we extend the first experiment to two dimensions: We seek the viscosity solution to

$$\frac{d}{dx}u + u = 1 \text{ in } \Omega := (0, 1)^2 \quad \text{with } u(0, \cdot) = u(1, \cdot) = 0. \quad (28)$$

In other words, we solve the problem in (25) with  $\varepsilon = 0$ ,  $\beta = (1, 0)^\top$ , and  $c = 1$  with modified boundary conditions. The viscosity solution reads  $u(x, y) = 1 - \exp(-x)$ . The resulting approximations are displayed in Fig. 2. All methods fail on uniform meshes: While the Galerkin FEM leads to strong oscillations along the  $x$ -axis, the minimal residual method in (c) and our suggested methods in (1) lead to strong oscillations along the  $y$ -axis. The LSFEM seems to be more robust, but does not provide a good approximation as well. Uniform mesh refinements do not seem to overcome these difficulties. However, adaptive mesh refinement



**Fig. 3** Approximations of the solution to (29) with  $\varepsilon = 10^{-3}$ , uniform mesh, and  $\dim U_h = 4225$  evaluated at  $(x, 1/2)$

overcomes this problem for the minimal residual methods. The adaptive LSFEM solution seems to converge to the exact solution but is still much worse than the adaptively computed solutions to the method in (c) and our approach in (1). Indeed, the solution to (c) shows only some tiny oscillation near  $x = 1$ , the solution to our scheme in (1) does not show any oscillation at all and provides a very accurate approximation, cf. Fig. 2. This shows that adaptivity might be a key in the convergence of our approximation.

### 7.3 Experiment 3 (Eriksson and Johnson)

Our third experiment has been introduced by Eriksson and Johnson in [8]. We seek the solution to (26) with  $\beta = (1, 0)^T$ , right-hand side  $f = 0$ , and initial data  $u(0, y) = \sin(\pi y)$  and  $u(x, y) = 0$  for  $x = 1$  or  $y \in \{0, 1\}$  with unit square domain  $\Omega = (0, 1)^2$ . In other words, we seek the solution to

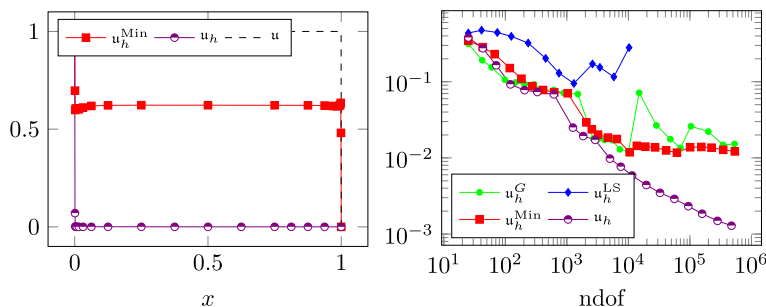
$$\begin{aligned} -\varepsilon \Delta u + \frac{d}{dx} u &= 0 & \text{in } \Omega, \\ u(x, y) &= 0 & \text{if } y \in 0, 1 \text{ or } x = 1, \\ u(x, y) &= \sin(\pi y) & \text{if } x = 0. \end{aligned} \quad (29)$$

Let  $s_1 := (1 + \sqrt{1 + 4\pi^2 \varepsilon^2})(2\varepsilon)^{-1}$  and  $s_2 := (1 - \sqrt{1 + 4\pi^2 \varepsilon^2})(2\varepsilon)^{-1}$ . The exact solution reads

$$u(x, y) = \frac{\exp(s_1(x-1)) - \exp(s_2(x-1))}{\exp(-s_1) - \exp(-s_2)} \sin(\pi y).$$

In our first computation we set  $\varepsilon := 10^{-3}$  and use uniformly refined meshes. In contrast to the previous calculations all methods converge as the mesh is uniformly refined. On coarse grids our minimization schemes leads to superior results as depicted in Fig. 3.

The situation changes drastically when we solve the problem with very small diffusion coefficient  $\varepsilon := 10^{-6}$ . For uniform mesh refinements the solutions to the minimal residual method in (1) and the minimal residual method in (c) show as in Experiment 2 strong oscillations along the  $y$ -axis. The direct solver in FEniCS (MUMPS) was not able to solve the resulting system for the Galerkin FEM solution with more than 80 degrees of freedom. The LSFEM solution seems to converge towards a function  $u \approx \gamma \sin(\pi y)$  with some constant  $\gamma$  slight larger than  $1/2$  as the mesh is uniformly refined. Unfortunately, adaptivity does not overcome this problem: The direct solver in FEniCS (MUMPS) was not able to compute a solution to the Galerkin FEM and the LSFEM with adaptive mesh refinements for meshes



**Fig. 4** The left-hand side shows approximations evaluated at  $(x, 1/2)$  with adaptive mesh refinements and fixed parameter  $\varepsilon = 10^{-6}$  with  $\dim U_h \approx 5000$  and the right-hand side shows the convergence history plot of the  $L^2(\Omega)$  error for  $\varepsilon = 10^{-6}$  with adapted diffusion parameter in (30)

with more than about 300 and 1000 degrees of freedom, respectively. The adaptive scheme for our method in (1) refines strongly near  $x = 0$  and the approximation seems to converge point-wise to zero, cf. Fig. 4. The adaptive minimal residual method in (c) refines strongly near  $x = 0$  and  $x = 1$  and the approximations look roughly like  $u \approx \gamma \sin(\pi y)$  with some constant  $\gamma$  slightly larger than  $1/2$ . All in all, non of the schemes converges towards the exact solution. We overcome this challenge by slowly adapting the diffusion parameter  $\varepsilon$  in our computations in the sense that we set

$$\varepsilon := \begin{cases} 10^{-2} & \text{if } \dim U_h \in [0, 1000), \\ 10^{-3} & \text{if } \dim U_h \in [1000, 5000), \\ 10^{-4} & \text{if } \dim U_h \in [5000, 10000), \\ 10^{-5} & \text{if } \dim U_h \in [10000, 50000), \\ 10^{-6} & \text{else.} \end{cases} \quad (30)$$

Fig. 4 shows the resulting convergence history plot of the error measured in the  $L^2(\Omega)$  norm. Initially adapting  $\varepsilon$  helps all methods, but as  $\dim U_h$  exceeds  $10^3$  (and so  $\varepsilon$  is set to  $10^{-3}$ ), the LSFEM starts to struggle. The same happens for the Galerkin and minimal residual method in (a) and (c) as  $\dim U_h$  exceed  $10^4$  (and so  $\varepsilon$  is set to  $10^{-5}$ ). In contrast, our method in (1) still converges as the number of degrees of freedom is increased. In order to save computational power, we did not use the adaptive scheme suggested in Sect. 6. Instead, we fixed the relaxation interval  $\zeta := [10^{-2}, 10^2]$  and computed only two Kačanov iterations on each mesh. An alternative calculation, using the adaptive strategy in Sect. 6 with the large weight  $w = 100$  causing about five Kačanov iterations on each mesh, led to similar convergence results.

## 8 Conclusion

We have introduced a novel numerical scheme that solves minimal residual methods in  $W^{-1,p'}(\Omega)$ . Additionally, we suggested an iterative scheme that converges towards the discrete solution of the resulting non-linear minimization problem. The scheme converges even for large exponents like  $p = 100$ . The resulting approximations are beneficial for solving challenging PDE's like convection-dominated diffusion problems compared to other schemes like the Galerkin FEM or minimal residual methods in Hilbert spaces. However, in



these challenging situations the convergence of our scheme seems to require some suitable mesh design. This can be done adaptively with some suitable designed initial mesh. We thus suggest a scheme where we increase the diffusion parameter depending on the degrees of freedom. This allowed for the approximation of convection-dominated diffusion problems with tiny diffusion parameters like  $\varepsilon = 10^{-6}$ .

**Funding** Open Access funding enabled and organized by Projekt DEAL. The work of the author was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1283/2 2021 – 317210226.

**Data Availability** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors have not disclosed any Conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Balci, A. Kh., Dening, L., Storn, J.: Relaxed Kačanov Scheme for the p-Laplacian with large exponent. *SIAM J. Numer. Anal.* **61**(6), 2775–2794 (2023). <https://doi.org/10.1137/22M1528550>
- Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Applied Mathematical Sciences, vol. 166, pp. xxii+660. Springer, New York (2009). <https://doi.org/10.1007/b13382>
- Carstensen, C., Demkowicz, L., Gopalakrishnan, J.: A posteriori error control for DPG methods. *SIAM J. Numer. Anal.* **52**(3), 1335–1353 (2014). <https://doi.org/10.1137/130924913>
- Carstensen, C., Demkowicz, L., Gopalakrishnan, J.: Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.* **72**(3), 494–522 (2016). <https://doi.org/10.1016/j.camwa.2016.05.004>
- Dening, L., Fornasier, M., Tomasi, R., Wank, M.: A relaxed Kačanov iteration for the p-Poisson problem. *Numer. Math.* **145**(1), 1–34 (2020). <https://doi.org/10.1007/s00211-020-01107-1>
- Demlow, A., Leykekhman, D., Schatz, A.H., Wahlbin, L.B.: Best approximation property in the  $W_\infty^1$  norm for finite element methods on graded meshes. *Math. Comput.* **81**(278), 743–764 (2012). <https://doi.org/10.1090/S0025-5718-2011-02546-9>
- Dening, L., Rolfes, J., Salgado, A.J.: Pointwise gradient estimate of the Ritz projection (2023). arXiv preprint [arXiv:2305.03575](https://arxiv.org/abs/2305.03575). <https://doi.org/10.48550/arXiv.2305.03575>
- Eriksson, K., Johnson, C.: Adaptive streamline diffusion finite element methods for stationary convection–diffusion problems. *Math. Comput.* **60**(201), 167–188 (1993). <https://doi.org/10.2307/2153160>
- Guermond, J.L.: A finite element technique for solving first-order PDEs in LP. *SIAM J. Numer. Anal.* **42**(2), 714–737 (2004). <https://doi.org/10.1137/S0036142902417054>
- Hanner, O.: On the uniform convexity of  $L_p$  and  $l_p$ . *Ark. Mat.* **3**, 239–244 (1956). <https://doi.org/10.1007/BF02589410>
- Houston, P., Muga, I., Roggendorf, S., van der Zee, K.G.: The convection–diffusion–reaction equation in non-Hilbert Sobolev spaces: a direct proof of the Inf-Sup condition and stability of Galerkin fs Method. *Comput. Methods Appl. Math.* **19**(3), 503–522 (2019). <https://doi.org/10.1515/cmam-2018-0198>
- Houston, P., Roggendorf, S., van der Zee, K.G.: Gibbs phenomena for  $L_q$ -best approximation in finite element spaces. *ESAIM Math. Model. Numer. Anal.* **56**(1), 177–211 (2022). <https://doi.org/10.1051/m2an/2021086>

13. Jerison, D., Kenig, C.E.: The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Funct. Anal.* **130**(1), 161–219 (1995). <https://doi.org/10.1006/jfan.1995.1067>
14. Katzourakis, N.: *An Introduction to Viscosity Solutions for Fully Nonlinear PDE with Applications to Calculus of Variations in  $L^\infty$* . Springer Briefs in Mathematics. Springer, Cham, pp. xii+123 (2015). <https://doi.org/10.1007/978-3-319-12829-0>
15. Li, J., Demkowicz, L.: An Lp-DPG method with application to 2D convection–diffusion problems. *Comput. Methods Appl. Math.* **22**(3), 649–662 (2022). <https://doi.org/10.1515/cmam-2021-0158>
16. Millar, F., Muga, I., Rojas, S., Van der Zee, K.G.: Projection in negative norms and the regularization of rough linear functionals. *Numer. Math.* **150**(4), 1087–1121 (2022). <https://doi.org/10.1007/s00211-022-01278-z>
17. Monsuur, H., Stevenson, R., Storn, J.: Minimal residual methods in negative or fractional Sobolev norms (2023). <https://doi.org/10.48550/2301.10484>
18. Muga, I., van der Zee, K.G.: Discretization of linear problems in Banach spaces: residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods. *SIAM J. Numer. Anal.* **58**(6), 3406–3426 (2020). <https://doi.org/10.1137/20M1324338>
19. Storn, J.: On a relation of discontinuous Petrov–Galerkin and least-squares finite element methods. *Comput. Math. Appl.* **79**(12), 3588–3611 (2020). <https://doi.org/10.1016/j.camwa.2020.02.018>
20. Verfürth, R.: *A Review of Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley, Hoboken (1996)
21. Zeidler, E.: *Applied Functional Analysis*. Applied Mathematical Sciences. Main Principles and Their Applications, vol. 109, pp. xvi+404. Springer, New York (1995)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.