SIAM J. NUMER. ANAL.
Vol. 58, No. 3, pp. 1773–1800

CONVERGENCE OF HEURISTIC PARAMETER CHOICE RULES FOR CONVEX TIKHONOV REGULARIZATION*

STEFAN KINDERMANN† AND KEMAL RAIK†

Abstract. We investigate the convergence theory of several known as well as new heuristic parameter choice rules for convex Tikhonov regularization. The success of such methods is dependent on whether certain restrictions on the noise are satisfied. In the linear theory, such conditions are well understood and hold for typically irregular noise. In this paper, we extend the convergence analysis of heuristic rules using noise restrictions to the convex setting and prove convergence of the aforementioned methods therewith. The convergence theory is exemplified for the case of an ill-posed problem with a diagonal forward operator in ℓ^q spaces. Numerical examples also provide further insight.

Key words. ill-posed problems, convex regularization, heuristic parameter choice rules

AMS subject classifications. 34A55, 49N45, 47A52, 65F22, 65J20, 46N10

DOI. 10.1137/19M1263066

1. Introduction.

Let X be a Banach and Y a Hilbert space. We consider the ill-posed problem

$$Ax = y,$$

where $A : X \to Y$ is a continuous linear operator and only noisy data $y^\delta = y + e$ are available and δ such that $\|e\| \leq \delta$ is defined to be the noise level. In other words, we assume that the solution does not depend continuously on the data. We therefore determine a regularized solution à la Tikhonov:

$$x_\alpha^\delta \in \operatorname*{argmin}_{x \in X} \mathcal{T}_\alpha^\delta(x), \qquad \text{where} \qquad \mathcal{T}_\alpha^\delta(x) := \frac{1}{2}\|Ax - y^\delta\|^2 + \alpha\mathcal{R}(x), \tag{1.1}$$

is the *Tikhonov functional*, and the regularization term given by the functional $\mathcal{R} : X \to \mathbb{R}\cup\{\infty\}$ is assumed to be convex, proper, coercive, and weak-$*$ lower semicontinuous and $\alpha \in (0, \alpha_{\max})$ is the so-called regularization parameter. The aforementioned properties ensure the existence of a minimizer for $\mathcal{T}_\alpha^\delta$ (cf. [34, Theorem 3.22]). In this way, we seek to approximate an $\mathcal{R}$*-minimizing solution*, $x^\dagger \in X$ (cf. [34, Definition 3.24]).

The choice of the regularization parameter is pivotal for any reasonable approximation of $x^\dagger$. There are several classes of rules which select this parameter (cf., e.g., [15]), the best-known being the a posteriori rules which choose a regularization parameter $\alpha_* := \alpha(\delta, y^\delta)$ in dependence of the noise level and the measured noisy data. A classic example is Morozov's discrepancy principle (cf. [29]). The drawback is that in practical situations, the noise level is usually unknown. In this case, one may opt to use another class of parameter choice rules, namely the so-called heuristic

*Received by the editors May 20, 2019; accepted for publication (in revised form) April 6, 2020; published electronically June 9, 2020.
https://doi.org/10.1137/19M1263066
Funding: The work of the authors was supported by Austrian Science Fund (FWF) grant P 30157-N31.
†Industrial Mathematics Institute, Johannes Kepler University Linz, Altenbergerstraße 69, A-4040 Linz, Austria (kindermann@indmath.uni-linz.ac.at, kemal.raik@indmath.uni-linz.ac.at).

rules, which select $\alpha_* := \alpha(y^\delta)$ in dependence of the measured data alone (i.e., without knowledge of the noise level). Their pitfall, however, comes in the form of the Bakushinskii veto (cf. [1]), which essentially asserts that a heuristic rule cannot yield a convergent regularization method in the worst case scenario. However, it was proven recently in [25, 22] for linear regularization methods that certain heuristic rules yield a convergent method if some *noise conditions* are postulated. It is important to note that the aforementioned noise conditions utilized the spectral theory for self-adjoint linear operators. A recent discussion and extension of the noise conditions within the linear theory may be found in [27].

The topic of this paper is the corresponding analysis for the convex case. Note that the tools from the linear theory are no longer applicable due to the absence of spectral theory. For the mentioned setting, some heuristic parameter choice rules were considered in the literature: B. Jin and Lorenz in [19] discussed the heuristic discrepancy rule and a version of the discrete quasi-optimality rule. The heuristic discrepancy rule was also considered for the augmented Lagrangian method and Bregman iteration for nonlinear operators in the work of Q. Jin (cf. [20, 21, 36]). A numerical study of certain heuristic rules was investigated in [24].

Crucial to the convergence theory of heuristic rules are restrictions on the noise. In the linear theory, they take the form of Muckenhoupt-type conditions. In the convex case, some rather abstract conditions were proposed in [19, 20]. However, the validity of these conditions remains unclear.

In this paper, we propose several heuristic rules and, as our main contribution, provide a convergence analysis by postulating so-called auto-regularization conditions, and, in the special case of ℓ^q regularization with a diagonal operator A, we show that so-called Muckenhoupt-type conditions suffice for the auto-regularization condition, which allows us also to verify their validity in typical cases. The main results are Theorem 3.4 in section 3 (with abstract conditions) and Theorems 4.2, 4.5, and 4.7 (specific convergence conditions for the diagonal case) in section 4. Furthermore, we provide a detailed numerical case study of these heuristic methods in section 5.

2. Preliminaries.

Note that in contrast with linear regularization theory, one cannot (in general) prove convergence of the regularized solution to $x^\dagger$ in the norm, but it is common to use the *Bregman distance* (cf. [6, 7]): for any $x_1, x_2 \in X$,

$$D_{\xi_2}(x_1, x_2) := \mathcal{R}(x_1) - \mathcal{R}(x_2) - \langle \xi_2, x_1 - x_2 \rangle_{X^* \times X}$$

defines the Bregman distance, with respect to $\mathcal{R}$, between x_1 and x_2 in the direction of the subgradient $\xi_2 \in \partial\mathcal{R}(x_2)$.

The Bregman distance is not a distance (a.k.a. a metric) as it does not satisfy the triangle inequality, nor is it in general symmetric. We do, however, have the following useful so-called three point identity (cf. [8, Lemma 3.1]):

$$D_{\xi_2}(x_1, x_2) = D_{\xi_3}(x_1, x_3) + D_{\xi_3}(x_3, x_2) + \langle \xi_3 - \xi_2, x_1 - x_3 \rangle. \tag{2.1}$$

One can also define the *symmetric Bregman distance* as

$$D^{\mathrm{sym}}_{\xi_1,\xi_2}(x_1, x_2) := \langle \xi_1 - \xi_2, x_1 - x_2 \rangle = D_{\xi_2}(x_1, x_2) + D_{\xi_1}(x_2, x_1).$$

In the following, we use a super/subscripted δ to indicate variables associated with noisy data and its absence indicates the corresponding variables for exact data. For instance,

$$x_\alpha^\delta \in \operatorname*{argmin}_{x \in X} \mathcal{T}_\alpha^\delta(x) \qquad \text{and} \qquad x_\alpha \in \operatorname*{argmin}_{x \in X} \mathcal{T}_\alpha(x),$$

where $\mathcal{T}_\alpha$ indicates the Tikhonov functional with exact data y replacing y^δ.

For the formulation of the heuristic rules, we require *Bregman iteration*, which can be seen as a nonlinear analogue of the well-known iterated Tikhonov regularization, (cf. [6, 30]). In particular, the *second Bregman iterate* $x_{\alpha,\delta}^{II}$ may be computed as a minimizer of a Tikhonov-type functional with the Bregman distance as regularization term or, equivalently, as a minimizer of the original Tikhonov functional with y^δ replaced by $y^\delta + y^\delta - Ax_\alpha^\delta$ (a.k.a. "adding back the noise" strategy) [35]:

$$\begin{aligned} &x_{\alpha,\delta}^{II} \in \operatorname*{argmin}_{x\in X} \frac{1}{2}\|Ax - y^\delta\|^2 + \alpha D_{\xi_\alpha^\delta}\left(x, x_\alpha^\delta\right), \\ \Leftrightarrow\ &x_{\alpha,\delta}^{II} \in \operatorname*{argmin}_{x\in X} \frac{1}{2}\left\|Ax - \left(y^\delta + y^\delta - Ax_\alpha^\delta\right)\right\|^2 + \alpha\mathcal{R}(x). \end{aligned} \tag{2.2}$$

We may define specific selections of the subgradients of $\mathcal{R}$ at x_α^δ and $x_{\alpha,\delta}^{II}$ as follows:

$$\begin{aligned} \xi_\alpha^\delta &:= -\frac{1}{\alpha}A^*\left(Ax_\alpha^\delta - y^\delta\right) \in \partial\mathcal{R}(x_\alpha^\delta), \\ \xi_{\alpha,\delta}^{II} &:= -\frac{1}{\alpha}A^*\left(Ax_{\alpha,\delta}^{II} - y^\delta - \left(y^\delta - Ax_\alpha^\delta\right)\right) \in \partial\mathcal{R}\left({x_\alpha^\delta}^{II}\right), \end{aligned}$$

and denote by ξ_α and ξ_α^{II} their respective noise-free variants.

It is useful to define the residuals of the first and second (i.e., Bregman) Tikhonov regularization as variables:

$$\begin{aligned} p_\alpha^\delta &:= y^\delta - Ax_\alpha^\delta, \qquad & p_\alpha &:= y - Ax_\alpha, \\ p_{\alpha,\delta}^{II} &:= y^\delta - Ax_{\alpha,\delta}^{II}, \qquad & p_\alpha^{II} &:= y - Ax_\alpha^{II}. \end{aligned}$$

Moreover, it will be convenient to write the residual variables by means of a proximal mapping operator. In the following proposition, $A^* : Y \to X^*$ denotes the adjoint operator and $\mathcal{R}^*$ denotes the *Fenchel conjugate* of $\mathcal{R}$ (cf. [5, 4]).

PROPOSITION 2.1. *The residuals* p_α^δ *and* $p_{\alpha,\delta}^{II}$ *may be expressed in terms of a proximal mapping operator,* $\operatorname{prox}_\mathcal{J} : Y \to Y$*,*

$$\operatorname{prox}_\mathcal{J} = (I + \partial\mathcal{J})^{-1}, \tag{2.3}$$

with the functional $\mathcal{J} = \alpha\mathcal{R}^* \circ \frac{1}{\alpha}A^*$ *in the following form:*

$$p_\alpha^\delta := \operatorname{prox}_\mathcal{J}(y^\delta), \qquad p_{\alpha,\delta}^{II} := \operatorname{prox}_\mathcal{J}\left(y + p_\alpha^\delta\right) - p_\alpha^\delta. \tag{2.4}$$

Proof. It follows from the optimality condition for the Tikhonov functional that

$$\partial\mathcal{R}(x_\alpha^\delta) \ni -A^*\left(Ax_\alpha^\delta - y^\delta\right)/\alpha \iff x_\alpha^\delta \in \partial\mathcal{R}^*\left(-A^*\left(Ax_\alpha^\delta - y^\delta\right)/\alpha\right), \tag{2.5}$$

since $x^* \in \partial\mathcal{R}(x) \iff x \in \partial\mathcal{R}^*(x^*)$ for all $x \in X$ and $x^* \in X^*$ (cf. [9]). Furthermore, (2.5) is equivalent to

$$Ax_\alpha^\delta - y^\delta \in A\partial\mathcal{R}^*\left(-A^*\left(Ax_\alpha^\delta - y^\delta\right)/\alpha\right) - y^\delta \iff (-p_\alpha^\delta) - A\partial\mathcal{R}^*\left(A^*p_\alpha^\delta/\alpha\right) \in -y^\delta.$$

We can rewrite the above as

$$(I + A\partial\mathcal{R}^*(A^*\cdot/\alpha))(p_\alpha^\delta) \in y^\delta \iff (I + \alpha\partial(\mathcal{R}^*(A^*\cdot/\alpha)))\left(p_\alpha^\delta\right) \in y^\delta,$$

due to the identity $A(\partial\mathcal{R}^*(A^*\cdot)) = \partial(\mathcal{R}^*(A^*\cdot))$, which holds true if $\mathcal{R}^*$ is finite and continuous at a point in the range of A^* (cf. [9, Prop. 5.7]), for instance, at $0 = A^*0$. By a result of Rockafellar [33, Theorems 4C, 7A], this follows if $\mathcal{R}$ has bounded sublevel sets, which is a consequence of the assumed coercivity. By definition of the proximal mapping, (2.4) follows for p_α^δ and analogously for $p_{\alpha,\delta}^{II}$. ∎

We will make use of the firm nonexpansivity of the proximal mapping operator:

$$\langle \operatorname{prox}_{\mathcal{J}}(y_1) - \operatorname{prox}_{\mathcal{J}}(y_2), y_1 - y_2 \rangle \geq \| \operatorname{prox}_{\mathcal{J}}(y_1) - \operatorname{prox}_{\mathcal{J}}(y_2)\|^2; \tag{2.6}$$

cf. [5, 4].

For notational purposes, we define shorthand for differences of noisy and noise-free variables:

$$\Delta y = y^\delta - y, \qquad \Delta p_\alpha = p_\alpha^\delta - p_\alpha, \qquad \Delta p_\alpha^{II} = p_{\alpha,\delta}^{II} - p_\alpha^{II}.$$

We state a useful estimate.

Lemma 2.2. *We have the following upper bound for the data propagation error:*

$$D_{\xi_\alpha}\left(x_\alpha^\delta, x_\alpha\right) \leq \frac{1}{\alpha}\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha\rangle \tag{2.7}$$

for all $\alpha \in (0, \alpha_{\max})$ *and* $y, y^\delta \in Y$.

Proof. We may estimate

$$D_{\xi_\alpha}\left(x_\alpha^\delta, x_\alpha\right) \leq D^{\mathrm{sym}}_{\xi_\alpha^\delta, \xi_\alpha}\left(x_\alpha^\delta, x_\alpha\right) = \frac{1}{\alpha}\left\langle \Delta p_\alpha, A\left(x_\alpha^\delta - x_\alpha\right)\right\rangle = \frac{1}{\alpha}\langle \Delta p_\alpha, \Delta y - \Delta p_\alpha\rangle,$$

which proves the desired result. ∎

2.1. Error estimates. Convergence results for convex regularization are well known. We state some standard results [7, 34]: we assume henceforth a regularity condition on $x^\dagger$, namely, that it fulfills the following *source condition*:

$$\partial\mathcal{R}(x^\dagger) \in \operatorname{range}(A^*) \iff \exists\, w : A^* w \in \partial\mathcal{R}(x^\dagger). \tag{2.8}$$

Subsequently, $\xi := A^* w$ is a subgradient of $\mathcal{R}$ at $x^\dagger$. We remark that much of the analysis (concerning the convergence results, not the rates) below is valid with (2.8) replaced by weaker conditions, e.g., in the form of variational source conditions [18, 10, 11]. We have the following error estimates from [19, Proposition 2.1, 2.2].

Proposition 2.3. *Let* $x^\dagger$ *satisfy the source condition* (2.8). *Then*

$$D_\xi(x_\alpha, x^\dagger) \leq \frac{\|w\|^2}{2}\alpha, \qquad \|Ax_\alpha - y\| \leq 2\|w\|\alpha,$$
$$D_{\xi_\alpha}(x_\alpha^\delta, x_\alpha) \leq \frac{\delta^2}{2\alpha}, \qquad \|A(x_\alpha^\delta - x_\alpha)\| \leq 2\delta,$$

and

$$D_\xi(x_\alpha^\delta, x^\dagger) \leq \frac{1}{2}\left(\frac{\delta}{\sqrt{\alpha}} + \sqrt{\alpha}\|w\|\right)^2, \qquad \|Ax_\alpha^\delta - y^\delta\| \leq \delta + 2\|w\|\alpha, \tag{2.9}$$

for all $\alpha \in (0, \alpha_{\max})$ *and* $y, y^\delta \in Y$.

Note also that

$$\|p_{\alpha,\delta}^{II}\| \leq \|p_\alpha^\delta\| \qquad \text{and} \qquad \|p_{\alpha,\delta}^{II}\|^2 \leq \langle p_{\alpha,\delta}^{II}, p_\alpha^\delta\rangle \tag{2.10}$$

for all $\alpha \in (0, \alpha_{\max})$ and $y^\delta \in Y$, where the former follows from the monotonicity of Bregman iteration (cf. [30]) and the latter follows from (2.4) and (2.6) with $y_1 = y^\delta + p_\alpha^\delta$ and $y_2 = y^\delta$. The inequalities (2.10), of course, hold analogously for the noise-free variables. From (2.8) and (2.1) with $x_1 = x_\alpha^\delta$, $x_2 = x^\dagger$, and $x_3 = x_\alpha$, we obtain the following estimate:

$$D_\xi\left(x_\alpha^\delta, x^\dagger\right) \le D_{\xi_\alpha}\left(x_\alpha^\delta, x_\alpha\right) + D_\xi\left(x_\alpha, x^\dagger\right) + 6\|w\|\delta; \tag{2.11}$$

cf. [19, p. 1214]. We will utilize this in the convergence proofs to come in the subsequent sections.

3. Heuristic parameter choice rules. The heuristic rules we consider select the parameter α in the Tikhonov functional (1.1)

$$\alpha = \alpha_* \in \underset{\alpha\in(0,\alpha_{\max})}{\operatorname{argmin}}\ \psi\left(\alpha, y^\delta\right) \tag{3.1}$$

as the global minimizer of a functional

$$\psi : (0, \alpha_{\max}) \times Y \to \mathbb{R},$$

where $\alpha_{\max}$ is a positive chosen constant. In the linear theory, this might be chosen as $\|A\|^2$. In the convex case, its choice is irrelevant for the convergence theory, although relevant for the numerical implementation (cf. section 5). In practice, one might choose it similarly as in the linear case. In case no α_* exists in the defined interval, we refer to [22, (2.13), p. 237], which details how one can extend the definition of ψ to overcome that issue. For simplicity, we persevere with (3.1) and simply assume existence.

The conceptual basis for (3.1) is that the surrogate functional ψ should act as an error estimator, i.e., $\psi \sim D_\xi(x_\alpha^\delta, x^\dagger)$, so that its minimizer should, in theory, be a good approximation of the optimal parameter choice. Note that the error estimating behavior is not guaranteed to hold in general (due to the Bakushinskii veto) unless restrictions on the noise are postulated. In the linear theory, they can have the abstract form of so-called auto-regularization conditions, i.e., for all $\alpha \in (0, \alpha_{\max})$, the inequality $\|x_\alpha^\delta - x_\alpha\| \le C\psi(\alpha, \Delta y)$ holds. However, such conditions are difficult to understand and verify as they exhibit a quite complicated dependence on the noise variable. Subsequently, in [25] these abstract conditions were replaced by more insightful and sufficient conditions in the form of Muckenhoupt-type inequalities (see (4.9) below). For the present convex case, the generalization of the aforementioned auto-regularization and Muckenhoupt conditions is not at all obvious, which thus serves as a significant motivation for this paper.

Any specific choice of ψ defines a heuristic parameter selection rule. We investigate four rules based on the following functionals:

$$\begin{aligned}
\psi_{\mathrm{HD}}\left(\alpha, y^\delta\right) &:= \frac{1}{\alpha}\|p_\alpha^\delta\|^2, && \text{the heuristic discrepancy rule (HD)},\\
\psi_{\mathrm{HR}}\left(\alpha, y^\delta\right) &:= \frac{1}{\alpha}\langle p_{\alpha,\delta}^{II}, p_\alpha^\delta\rangle, && \text{the Hanke–Raus rule (HR)},\\
\psi_{\mathrm{SQO}}\left(\alpha, y^\delta\right) &:= D^{\mathrm{sym}}_{\xi_{\alpha,\delta}^{II},\xi_\alpha^\delta}(x_{\alpha,\delta}^{II}, x_\alpha^\delta), && \text{the symmetric quasi-optimality rule (SQO)},\\
\psi_{\mathrm{RQO}}\left(\alpha, y^\delta\right) &:= D_{\xi_\alpha^\delta}(x_{\alpha,\delta}^{II}, x_\alpha^\delta), && \text{the right quasi-optimality rule (RQO)}.
\end{aligned}$$

Note that the heuristic discrepancy rule is sometimes also referred to as the Hanke–Raus rule (as the rules coincide for Landweber iteration). For clarity, it is preferable to name this method the heuristic analogue of the classical discrepancy rule. In particular, this rule is the only one for which some convergence analysis has been done (cf. [19, 21, 20, 36]). A discrete version of the quasi-optimality rules was also investigated in [19].

These rules are well established in the linear case. However, except for the HD rule, their extension to the convex setting is certainly not obvious. We consider Bregman iteration as the natural analogue of Tikhonov iteration, as stated previously, and we opt to define the latter three rules utilizing the second Bregman iterate. The HR rule was considered in [24], while the quasi-optimality rules considered here are entirely novel.

The principle behind the quasi-optimality rule is to minimize the difference of two successive approximations of the solution. In the linear case, the difference is measured with the norm, but in the convex setting, we use the Bregman distance. Therefore, possibilities for the quasi-optimality rule include choosing α_* as the minimizer of $D^{\mathrm{sym}}_{\xi^{II}_{\alpha,\delta},\xi^{\delta}_{\alpha}}(x^{II}_{\alpha,\delta}, x^{\delta}_{\alpha})$, $D_{\xi^{II}_{\alpha,\delta}}(x^{\delta}_{\alpha}, x^{II}_{\alpha,\delta})$, or $D_{\xi^{\delta}_{\alpha}}(x^{II}_{\alpha,\delta}, x^{\delta}_{\alpha})$.

Similarly as for the Hanke–Raus rule, the authors are not aware of any other analysis of the quasi-optimality rule defined as above. As mentioned, the discrete version was considered in [19]. There, the noise condition postulated was a generalization into the convex setting of the auto-regularization set of [12], and the numerical performance of the rule with $D_{\xi^{II}_{\alpha,\delta}}(x^{\delta}_{\alpha}, x^{II}_{\alpha,\delta})$ was tested in [24]. The performance of the latter rule in the aforementioned reference and also in our own numerical experiments proved to be quite poor and therefore we omit it. The rule with the symmetric Bregman distance performs reasonably on occasions, on the other hand. The remaining version is generally the best performing of all the quasi-optimality rules. More light will be shed on this, however, in the numerics section.

Similarly as for the heuristic discrepancy and Hanke–Raus rules, the symmetric quasi-optimality functional may also be expressed in terms of residuals.

PROPOSITION 3.1. *We have that*

$$\psi_{\mathrm{SQO}}\left(\alpha, y^{\delta}\right) = \frac{1}{\alpha}\left\langle p^{\delta}_{\alpha} - p^{II}_{\alpha,\delta}, p^{II}_{\alpha,\delta}\right\rangle$$

for all $\alpha \in (0, \alpha_{\max})$ *and* $y^{\delta} \in Y$.

Proof. We have

$$\begin{aligned} D^{\mathrm{sym}}_{\xi^{II}_{\alpha,\delta},\xi^{\delta}_{\alpha}}\left(x^{II}_{\alpha,\delta}, x^{\delta}_{\alpha}\right) &= \left\langle \xi^{II}_{\alpha,\delta} - \xi^{\delta}_{\alpha}, x^{II}_{\alpha,\delta} - x^{\delta}_{\alpha}\right\rangle \\ &= \frac{1}{\alpha}\left\langle A^*\left(Ax^{\delta}_{\alpha} - y^{\delta}\right) - A^*\left(Ax^{II}_{\alpha,\delta} - y^{\delta} + Ax^{\delta}_{\alpha} - y^{\delta}\right), x^{II}_{\alpha,\delta} - x^{\delta}_{\alpha}\right\rangle \\ &= \frac{1}{\alpha}\left\langle A\left(x^{\delta}_{\alpha} - x^{II}_{\alpha,\delta}\right), Ax^{II}_{\alpha,\delta} - y^{\delta}\right\rangle = \frac{1}{\alpha}\left\langle Ax^{\delta}_{\alpha} - y^{\delta} - \left(Ax^{II}_{\alpha,\delta} - y^{\delta}\right), Ax^{II}_{\alpha,\delta} - y^{\delta}\right\rangle \end{aligned}$$

for all $\alpha \in (0, \alpha_{\max})$ and $y^{\delta} \in Y$, which is what we wanted to show. ∎

PROPOSITION 3.2. *For all* $\alpha \in (0, \alpha_{\max})$ *and* $y^{\delta} \in Y$, *we have*

$$0 \le \psi_{\mathrm{RQO}}(\alpha, y^{\delta}) \le \psi_{\mathrm{SQO}}(\alpha, y^{\delta}) \le \psi_{HR}(\alpha, y^{\delta}) \le \psi_{\mathrm{HD}}(\alpha, y^{\delta}). \tag{3.2}$$

Moreover, if (2.8) *holds, then*

$$\psi_{\mathrm{HD}}(\alpha, y^{\delta}) \le \left(\frac{\delta}{\sqrt{\alpha}} + 2\|w\|\sqrt{\alpha}\right)^2. \tag{3.3}$$

In particular, with (2.8) *and* α_* *selected as* (3.1) *with* $\psi \in \{\psi_{\mathrm{HD}}, \psi_{\mathrm{HR}}, \psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}\}$, *we have that*

$$\lim_{\delta \to 0} \psi(\alpha_*, y^\delta) = 0. \tag{3.4}$$

Proof. Since

$$\psi_{\mathrm{SQO}}(\alpha, y^\delta) = \psi_{\mathrm{RQO}}(\alpha, y^\delta) + D_{\xi^{II}_{\alpha,\delta}}(x^\delta_\alpha, x^{II}_{\alpha,\delta}),$$

it follows immediately that $\psi_{\mathrm{SQO}} \geq \psi_{\mathrm{RQO}}$. Moreover,

$$\begin{aligned}\psi_{\mathrm{SQO}}\left(\alpha, y^\delta\right) &= \frac{1}{\alpha}\langle p^\delta_\alpha - p^{II}_{\alpha,\delta}, p^{II}_{\alpha,\delta}\rangle = \frac{1}{\alpha}\langle p^\delta_\alpha, p^{II}_{\alpha,\delta}\rangle - \frac{1}{\alpha}\|p^{II}_{\alpha,\delta}\|^2 \leq \frac{1}{\alpha}\langle p^\delta_\alpha, p^{II}_{\alpha,\delta}\rangle \\ = \psi_{\mathrm{HR}}\left(\alpha, y^\delta\right) &= \frac{1}{\alpha}\left\langle \mathrm{prox}_{\mathcal{J}^*}\left(y^\delta + p^\delta_\alpha\right) - \mathrm{prox}_{\mathcal{J}^*}(y^\delta), y^\delta + p^\delta_\alpha - y^\delta\right\rangle \\ &\leq \frac{1}{\alpha}\|\,\mathrm{prox}_{\mathcal{J}^*}\left(y^\delta + p^\delta_\alpha\right) - \mathrm{prox}_{\mathcal{J}^*}(y^\delta)\|\|p^\delta_\alpha\| \overset{(2.10)}{\leq} \frac{1}{\alpha}\|p^\delta_\alpha\|^2 = \psi_{\mathrm{HD}}\left(\alpha, y^\delta\right).\end{aligned}$$

The last estimate (3.3) follows from (2.9). As the respective α_* are the minimizers, we may estimate $\psi(\alpha_*, y^\delta)$ by the minimizer over α of (3.3), which is easily shown to be of the order of δ and thus tends to 0. ◻

3.1. Convergence. For heuristic rules and in the linear case, it is often standard to show convergence of the selected regularization parameter α_* as the noise level tends to zero. This is not necessarily true in the convex case. The next lemma deals with the (exceptional) case in which $\lim_{\delta \to 0} \alpha_* \neq 0$.

LEMMA 3.3. *Assume that $A : X \to Y$ is compact. Let α_* be the minimizer of $\psi \in \{\psi_{\mathrm{HD}}, \psi_{\mathrm{HR}}, \psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}\}$. In case of $\psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}$, we additionally assume that $\mathcal{R}$ is strictly convex. Suppose that $\lim_{\delta\to 0} \alpha_* = \bar{\alpha} > 0$. Then*

$$\lim_{\delta \to 0} D_\xi\left(x^\delta_{\alpha_*}, x^\dagger\right) = 0.$$

Proof. We show that any subsequence of $D_\xi(x^\delta_{\alpha_*}, x^\dagger)$ has a convergent subsequence with limit 0. In case of ψ_{HD}, the result follows from [19, Proof of Theorem 3.5] even without assumming compactness of A. From the same proof, it follows also that $x^\delta_{\alpha_*}$ is bounded, and similarly, we may show that $x^{II}_{\alpha_*,\delta}$ is bounded as $\delta \to 0$. Hence, there exist weakly (or weakly-*) convergent subsequences with

$$x^\delta_{\alpha_*} \rightharpoonup v, \qquad x^{II}_{\alpha_*,\delta} \rightharpoonup z.$$

In [19], it was shown that $v = x_{\bar{\alpha}}$, i.e., the limit of $x^\delta_{\alpha_*}$ is the regularized solution for exact data and regularization parameter $\bar{\alpha}$. Now using the compactness of A, we may find strongly convergent subsequences for the residuals as $\delta \to 0$:

$$p^\delta_{\alpha_*} \to y - Ax_{\bar{\alpha}} =: p_{\bar{\alpha}}, \qquad p^{II}_{\alpha_*,\delta} \to y - Az.$$

From lower semicontinuity, the minimization property of $x^{II}_{\alpha_*,\delta}$, and the strong convergence of $p^\delta_{\alpha_*}$, we obtain that for any $x \in X$

$$\begin{aligned}&\frac{1}{2}\|Az - y - (y - Ax_{\bar{\alpha}})\|^2 + \bar{\alpha}\mathcal{R}(z) \\ &\leq \liminf_{\delta \to 0} \frac{1}{2}\|Ax^{II}_{\alpha_*,\delta} - y^\delta - (y^\delta - Ax^\delta_{\alpha_*})\|^2 + \alpha_*\mathcal{R}(x^\delta_\alpha) \\ &\leq \liminf_{\delta \to 0} \frac{1}{2}\|Ax - y^\delta - (y^\delta - Ax^\delta_{\alpha_*})\|^2 + \alpha_*\mathcal{R}(x) \\ &= \frac{1}{2}\|Ax - y - (y - Ax_{\bar{\alpha}})\|^2 + \bar{\alpha}\mathcal{R}(x).\end{aligned}$$

Hence, z is the minimizer of the functional on the left-hand side, and by its uniqueness, it follows that $z = x_{\bar\alpha}^{II}$ and thus $p_{\alpha_*,\delta}^{II} \to y - Ax_{\bar\alpha}^{II} =: p_{\bar\alpha}^{II}$. From the optimality conditions, we furthermore obtain the strong convergence of $\xi_{\alpha_*}^{\delta} \to \xi_{\bar\alpha}$.

Consider ψ_{HR}: it follows from (2.6) that $\frac{1}{\alpha_*}\|p_{\alpha_*,\delta}^{II}\|^2 \leq \frac{1}{\alpha_*}\langle p_{\alpha_*}^{\delta}, p_{\alpha_*,\delta}^{II}\rangle = \psi_{\mathrm{HR}}(\alpha_*, y^\delta)$ and by (3.4), we find that $p_{\bar\alpha}^{II} = 0$. Since $p_{\bar\alpha}^{II} = \operatorname{prox}_{\mathcal{J}}(y + p_{\bar\alpha}) - p_{\bar\alpha}$, it follows that

$$\operatorname{prox}_{\mathcal{J}}(y + p_{\bar\alpha}) = p_{\bar\alpha} = \operatorname{prox}_{\mathcal{J}}(y).$$

As $\operatorname{prox}_{\mathcal{J}}$ is bijective, we obtain that $p_{\bar\alpha} = 0$. From this, the result can be concluded as in [19, Proof of Theorem 3.5].

In case of $\psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}$, we have, by (3.4), that $D_{\xi_{\alpha_*}^{\delta}}(x_{\alpha_*,\delta}^{II}, x_{\alpha_*}^{\delta}) \to 0$, and from the lower semicontinuity and the strong convergence of the subgradient ξ_α^δ, it follows that

$$D_{\xi_{\bar\alpha}}\left(x_{\bar\alpha}^{II}, x_{\bar\alpha}\right) = 0.$$

By the assumed strict convexity of $\mathcal{R}$, the Bregman distance is strictly positive for nonidentical arguments; thus $x_{\bar\alpha}^{II} = x_{\bar\alpha} \Rightarrow p_{\bar\alpha} = p_{\bar\alpha}^{II}$. Employing the proximal representation, we thus have that

$$\begin{aligned}&\operatorname{prox}_{\mathcal{J}}(y + p_{\bar\alpha}) - p_{\bar\alpha} = p_{\bar\alpha} \Leftrightarrow 2p_{\bar\alpha} + \partial\mathcal{J}(2p_{\bar\alpha}) = y + p_{\bar\alpha}\\ &\Leftrightarrow p_{\bar\alpha} + \partial\mathcal{J}(2p_{\bar\alpha}) = y = p_{\bar\alpha} + \partial\mathcal{J}(p_{\bar\alpha}) \Leftrightarrow \partial\mathcal{J}(2p_{\bar\alpha}) = \partial\mathcal{J}(p_{\bar\alpha}).\end{aligned}$$

The strict convexity implies $2p_{\bar\alpha} = p_{\bar\alpha}$, hence $p_{\bar\alpha} = 0$. The results then follow as before from those in [19]. ∎

In order to prove convergence, the most difficult part is to derive a condition that prevents α_* from decaying too rapidly, which involves some restriction on the noise. For instance, the condition

$$\|Q(y - y^\delta)\| \geq \varepsilon\|y - y^\delta\| \tag{3.5}$$

with Q the orthogonal projection onto $\operatorname{range}(A)^\perp$ and $\varepsilon > 0$ was introduced by Hanke and Raus in [16] in the linear case and also employed in [19] in the convex case. Also in [19], the slightly more general condition

$$\langle y^\delta - y, z\rangle \leq (1-\varepsilon)\|y^\delta - y\|\|z\| \qquad \forall z \in \overline{A(\operatorname{dom}\partial\mathcal{R})} \tag{3.6}$$

was postulated and used in the convergence proof. However, the validity of the above conditions is not well understood and it is therefore the main aim of this paper to replace these conditions with more practical ones.

In the next theorem, we impose such a noise restriction in the form of an *auto-regularization* condition. As mentioned previously, in the linear theory, these take on the form $\|x_\alpha^\delta - x_\alpha\| \leq C\psi(\alpha, \Delta y)$. The generalization of such inequalities to the nonlinear case is not obvious, though. In the following theorem, we state such new auto-regularization conditions which are at the heart of the convergence theory.

THEOREM 3.4. *Let $A : X \to Y$ be compact, the source condition* (2.8) *be satisfied, and α_* be the minimizer of $\psi \in \{\psi_{\mathrm{HD}}, \psi_{\mathrm{HR}}, \psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}\}$, and assume there exist constants $C > 0$ such that the respective auto-regularization condition*

$$\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha\rangle \leq C\|\Delta p_\alpha\|^2, \tag{ARC-HD}$$

$$\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha\rangle \leq C\langle \Delta p_\alpha^{II}, \Delta p_\alpha\rangle, \tag{ARC-HR}$$

$$\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha\rangle \leq C\left\langle \Delta p_\alpha - \Delta p_\alpha^{II}, \Delta p_\alpha^{II}\right\rangle, \tag{ARC-SQR}$$

$$\frac{1}{\alpha}\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha\rangle \leq CD_{\xi_\alpha^\delta}(x_{\alpha,\delta}^{II}, x_\alpha^\delta) + \mathcal{O}(\alpha) \tag{ARC-RQO}$$

holds for all $\alpha \in (0, \alpha_{\max})$ and $y^\delta \in Y$ for the heuristic discrepancy, Hanke–Raus, symmetric quasi-optimality, and right quasi-optimality rules. If $\psi \in \{\psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}\}$, assume in addition that $\mathcal{R}$ is strictly convex. Then it follows that the method converges, i.e.,

$$D_\xi(x_{\alpha_*}^\delta, x^\dagger) \to 0,$$

as $\delta \to 0$.

Proof. Take an arbitrary subsequence of $D_\xi(x_{\alpha_*}^\delta, x^\dagger)$. We show that it contains a convergent subsequence with limit 0, which proves the statement. Since α_* is bounded, we may consider a subsequence which in addition satisfies $\alpha_* \to \bar{\alpha}$. In case $\alpha_* \to \bar{\alpha} > 0$, the result $\lim_{\delta\to 0} D_\xi(x_{\alpha_*}^\delta, x^\dagger) = 0$ follows from Lemma 3.3 without even needing to invoke the auto-regularization conditions.

Thus, let us assume that $\alpha_* \to 0$. Then it follows from the estimate (2.11) and Proposition 2.3 that one only needs to prove convergence of the data propagation error $D_{\xi_\alpha}(x_\alpha^\delta, x_\alpha)$ which we can immediately estimate via (2.7) and the respective auto-regularization condition in Theorem 3.4.

For α_* minimizing the heuristic discrepancy functional, it follows from Proposition 2.3 and (3.4) that $\lim_{\delta\to} \psi_{\mathrm{HD}}(\alpha_*, y) = 0$ and $\lim_{\delta\to} \psi_{\mathrm{HD}}(\alpha_*, y^\delta) = 0$.

Thus, we may conclude that

$$\frac{\|\Delta p_{\alpha_*}\|^2}{\alpha_*} \leq \left(\frac{\|p_{\alpha_*}^\delta\|}{\sqrt{\alpha_*}} + \frac{\|p_{\alpha_*}\|}{\sqrt{\alpha_*}} \right)^2 = \left(\sqrt{\psi_{\mathrm{HD}}(\alpha_*, y^\delta)} + \sqrt{\psi_{\mathrm{HD}}(\alpha_*, y)} \right)^2 \to 0,$$

as $\delta \to 0$. Thus, using (2.7) and (ARC-HD) yields that $D_{\xi_{\alpha_*}}(x_{\alpha_*}^\delta, x_{\alpha_*}) \to 0$ as $\delta \to 0$. For the approximation error, it follows from Proposition 2.3 that $D_\xi(x_{\alpha_*}, x^\dagger) \leq C\alpha_* \to 0$. Hence, each term in (2.11) tends to 0 as $\delta \to 0$.

Let α_* be the minimizer of the Hanke–Raus functional. Then, as before, we estimate the Bregman distance as (2.11) and from (ARC-HR) deduce that

$$\begin{aligned} D_{\xi_{\alpha_*}}(x_{\alpha_*}^\delta, x_{\alpha_*}) &\leq \frac{C}{\alpha_*} \left\langle \Delta p_{\alpha_*}^{II}, \Delta p_{\alpha_*} \right\rangle = \frac{C}{\alpha_*} \langle p_{\alpha_*}^{II}, \Delta p_{\alpha_*} \rangle - \frac{C}{\alpha_*} \left\langle p_{\alpha_*,\delta}^{II}, \Delta p_{\alpha_*} \right\rangle \\ &= \frac{C}{\alpha_*} \left\langle p_{\alpha_*}^{II}, p_{\alpha_*} \right\rangle - \frac{C}{\alpha_*} \left\langle p_{\alpha_*}^{II}, p_{\alpha_*}^\delta \right\rangle + \frac{C}{\alpha_*} \left\langle p_{\alpha_*,\delta}^{II}, p_{\alpha_*}^\delta \right\rangle - \frac{C}{\alpha_*} \left\langle p_{\alpha_*,\delta}^{II}, p_{\alpha_*} \right\rangle \\ &\leq C\psi_{\mathrm{HR}}\left(\alpha, y^\delta\right) + C\psi_{\mathrm{HR}}(\alpha_*, y) - \frac{C}{\alpha_*} \left\langle p_{\alpha_*}^{II}, p_{\alpha_*}^\delta \right\rangle - \frac{C}{\alpha_*} \left\langle p_{\alpha_*,\delta}^{II}, p_{\alpha_*} \right\rangle . \end{aligned} \tag{3.7}$$

The last two terms can be estimated via the Cauchy–Schwarz inequality and (2.10), and are bounded from above by Proposition 2.3,

$$C\frac{1}{\alpha_*} \|p_{\alpha_*}^\delta\| \|p_{\alpha_*}\| \leq \frac{1}{\alpha_*} 2\omega\alpha_*(\delta + 2\|\omega\|\alpha_*).$$

Moreover, the last terms decay to zero as $\delta \to 0$.

For α_* minimizing ψ_{SQO}, note that from (2.11) and (ARC-SQR), it remains to estimate

$$\begin{aligned} D_{\xi_{\alpha_*}}(x_{\alpha_*}^\delta, x_{\alpha_*}) &\leq \frac{1}{\alpha_*} \langle p_{\alpha_*}^\delta - p_{\alpha_*,\delta}^{II}, p_{\alpha_*,\delta}^{II} \rangle + \frac{1}{\alpha_*} \langle p_{\alpha_*} - p_{\alpha_*}^{II}, p_{\alpha_*}^{II} \rangle \\ &\quad - \frac{1}{\alpha_*} \langle p_{\alpha_*} - p_{\alpha_*}^{II}, p_{\alpha_*,\delta}^{II} \rangle - \frac{1}{\alpha_*} \langle p_{\alpha_*}^\delta - p_{\alpha_*,\delta}^{II}, p_{\alpha_*}^{II} \rangle \\ &\leq \psi_{\mathrm{SQO}}\left(\alpha, y^\delta\right) + \psi_{\mathrm{SQO}}(\alpha_*, y) - \frac{1}{\alpha_*} \langle p_{\alpha_*} - p_{\alpha_*}^{II}, p_{\alpha_*,\delta}^{II} \rangle - \frac{1}{\alpha_*} \langle p_{\alpha_*}^\delta - p_{\alpha_*,\delta}^{II}, p_{\alpha_*}^{II} \rangle. \end{aligned}$$

As before, the result follows very similarly by estimating the first two terms from above (cf. Proposition 3.2) the "remainder" terms via the Cauchy–Schwarz inequality, the triangle inequality, and (2.10), all of which vanish as the noise decays.

We omit the proof for the right quasi-optimality rule as it is analogous to the above. ∎

Remark 3.5. Note that Theorem 3.4 holds true if the left-hand side of the auto-regularization conditions, i.e., $\langle \Delta y - \Delta p_\alpha, \Delta p_\alpha \rangle$ is replaced by $D_{\xi_\alpha}(x_\alpha^\delta, x_\alpha)$. Moreover, it is easy to see that for (ARC-HD) it is enough to prove

$$\langle \Delta p_\alpha, \Delta y \rangle \leq C \|\Delta p_\alpha\|^2, \tag{3.8}$$

for some positive constant C.

Remark 3.6. Observe that the heuristic discrepancy, the Hanke–Raus, and the symmetric quasi-optimality rules can all be expressed in terms of the residuals of the Bregman iteration $p_\alpha^\delta, p_{\alpha,\delta}^{II}$. It should be noted that in the linear case they can all be subsumed under the so-called family of R1-rules [32]. The similarity of patterns in the formulae for ψ may provide a hint that such a larger family of rules could be defined in the convex case as well.

The auto-regularization condition is an implicit condition on the noise. One may observe that it resembles the analogous condition of [22, eqn. (3.9)] in the linear case. To gain a better understanding and in particular show that they can be satisfied in practical situations, we will in section 4 derive sufficient and more transparent conditions in the form of Muckenhoupt-type inequalities.

3.2. Convergence rates. With the aid of the source condition, the auto-regularization condition, and an additional regularity condition, we can even derive rates of convergence. We start with the following proposition.

PROPOSITION 3.7. *Suppose that $\partial\mathcal{R}^*$ satisfies the following condition:*

$$x \to 0 \Rightarrow \xi \in \partial\mathcal{R}^*(x) \rightharpoonup 0. \tag{3.9}$$

Then, for all $D > 0$, there exists another constant $D_1 > 0$ such that for all $y^\delta \in Y$ with $\|y^\delta\| \geq D$, it holds that

$$\alpha \leq D_1 \psi_{\mathrm{HD}}(\alpha, y^\delta) \qquad \forall \alpha \in (0, \alpha_{\max}). \tag{3.10}$$

Proof. Suppose that the statement (3.10) is not true. Then there exists a constant $D > 0$ and a sequence of y^{δ_k} such that $\|y^{\delta_k}\| \geq D$ and a sequence of α_k with

$$\left\| \frac{p_{\alpha_k}^{\delta_k}}{\alpha_k} \right\| = \frac{\psi_{\mathrm{HD}}(\alpha_k, y^{\delta_k})}{\alpha_k} \to 0 \qquad \text{as } k \to \infty.$$

Define $z_k := \frac{p_{\alpha_k}^{\delta_k}}{\alpha_k}$. From its representation as a proximal mapping, we have that z_k satisfies

$$y^{\delta_k} = \alpha_k z_k + A\partial\mathcal{R}^*(A^* z_k).$$

Thus, from $z_k \to 0$, the boundedness of α_k, $A^* z_k \to 0$, and (3.9), we obtain

$$0 < D \leq \|y^{\delta_k}\|^2 = \alpha_k \langle z_k, y^{\delta_k} \rangle + \langle \partial\mathcal{R}^*(A^* z_k), A^* y^{\delta_k} \rangle_{X^*, X^{**}} \to 0,$$

which yields a contradiction; hence the statement is true. ∎

We now state the main convergence rates result.

PROPOSITION 3.8. *Let the source condition* (2.8) *hold, let* α_* *minimize* ψ_{HD}, *and suppose the auto-regularization condition* (ARC-HD) *is satisfied. Assume that* $\|y^\delta\| \geq C$ *and, in addition, that* (3.9) *holds. Then*

$$D_\xi(x^\delta_{\alpha_*}, x^\dagger) = \mathcal{O}(\delta)$$

for $\delta \to 0$.

Proof. Note that from Proposition 3.7 and since α_* is the global minimizer, for any α, we have that $\alpha_* \leq C\psi_{\mathrm{HD}}(\alpha_*, y^\delta) \leq C\psi_{\mathrm{HD}}(\alpha, y^\delta)$. Observe that from (2.11), it follows that

$$\begin{aligned}
D_\xi(x^\delta_{\alpha_*}, x^\dagger) &\leq D_{\xi_{\alpha_*}}(x^\delta_{\alpha_*}, x_{\alpha_*}) + \frac{\|w\|^2}{2}\alpha_* + 6\|w\|\delta \\
&\leq \left(\sqrt{\psi_{\mathrm{HD}}\left(\alpha, y^\delta\right)} + \sqrt{\psi_{\mathrm{HD}}(\alpha_*, y)}\right)^2 + C\delta + C\alpha_* \quad \text{via (ARC-HD) \& triangle ineq.} \\
&\leq \left(\frac{\delta}{\sqrt{\alpha}} + C\sqrt{\alpha} + C\sqrt{\alpha_*}\right)^2 + C\delta + C\alpha_* \quad \text{via Proposition 2.3} \\
&= \mathcal{O}\left(\left(\frac{\delta}{\sqrt{\alpha}} + \sqrt{\alpha}\right)^2 + \frac{\delta^2}{\alpha} + \alpha + \delta\right) \qquad \text{since } \alpha_* \leq C\psi_{\mathrm{HD}}\left(\alpha, y^\delta\right), \\
&= \mathcal{O}\left(\delta\right),
\end{aligned}$$

choosing $\alpha = \alpha(\delta) = \delta$. ∎

Note that if the source condition (2.8) holds, α_* is selected as the minimizer of $\psi \in \{\psi_{\mathrm{HR}}, \psi_{\mathrm{SQO}}, \psi_{\mathrm{RQO}}\}$ and the respective auto-regularization conditions ((ARC-HR), (ARC-SQR), or (ARC-RQO)) are satisfied, and additionally, if for some $\mu > 0$

$$\alpha_*^\mu \leq C\psi(\alpha_*, y^\delta), \tag{3.11}$$

then one may also prove, analogously, that

$$D_\xi(x^\delta_\alpha, x^\dagger) = \mathcal{O}(\delta^{\frac{1}{\mu}})$$

for $\delta \to 0$. In the linear theory, the inequality (3.11) can be shown to hold, and under certain additional regularity conditions on $x^\dagger$, optimal convergence rates are obtained. An analogous analysis with optimal rates for ψ other than ψ_{HD} is beyond the scope of this paper.

We note that the condition on $\partial\mathcal{R}^*$, (3.9), holds if $\mathcal{R}^*$ is continuously differentiable in 0 with $\nabla\mathcal{R}(0) = 0$. This is true, for instance, for $\mathcal{R}^*(x) = \|x\|^p_{\ell^p}$ with $1 < p < \infty$. On the other hand we observe that the conclusion in Proposition 3.7 is not satisfied for exact penalization methods [7] (such as ℓ^1-regularization) as then the residual p_α (and hence ψ_{HD}) could vanish for nonzero α, which does not concur with the estimate (3.10).

4. Diagonal operator. In the following analysis, we consider the case of the operator $A : X \to Y$ being diagonal between spaces of summable sequences; in particular, $X = \ell^q(\mathbb{N})$, with $1 < q < \infty$, and $Y = \ell^2(\mathbb{N})$, and the regularization functional is selected as the ℓ^q-norm to the qth power. The main objective in this setting is to state sufficient conditions for the auto-regularization conditions in the form of Muckenhoupt-type inequalities and to illustrate their validity for specific instances.

Let $\{e_n\}_{n\in\mathbb{N}}$ be the canonical (i.e., Cartesian) basis for X and also Y, and let $\{\lambda_n\}_{n\in\mathbb{N}}$ be a sequence of real (and for simplicity) positive scalars monotonically decaying to 0. Then we define a diagonal operator $A : \ell^q(\mathbb{N}) \to \ell^2(\mathbb{N})$,

$$Ae_n = \lambda_n e_n. \tag{4.1}$$

The regularization functional is chosen as the qth power of the ℓ^q-norm:

$$\mathcal{R} := \frac{1}{q}\|\cdot\|_{\ell^q}^q \qquad \text{with} \qquad \partial\mathcal{R}(x) = \{|x_n|^{q-1}\operatorname{sgn}(x_n)\}_{n\in\mathbb{N}}, \qquad q \in (1,\infty). \tag{4.2}$$

As we assume $q > 1$, the choice of sgn at 0 is not relevant and we may assume $\operatorname{sgn}(0) = 0$ throughout. In the present situation, the problem decouples and the components of the regularized solution can be computed independently of each other. Thus, for notational purposes, we opt to omit the sequence index n for the components of the regularized solutions and write

$$x_\alpha^\delta =: \{x_{\alpha,n}^\delta\}_n =: \{\chi_\alpha^\delta\}_n, \quad x_\alpha =: \{x_{\alpha,n}\}_n =: \{\chi_\alpha\}_n, \quad y^\delta =: \{y^\delta\}_n, \quad y =: \{y\}_n,$$

where $\chi_\alpha, \chi_\alpha^\delta, y^\delta, y \in \mathbb{R}$. As the problem decouples, χ_α and χ_α^δ can be computed by an optimization problem on $\mathbb{R}$, i.e., the optimality conditions read

$$\chi_\alpha^\delta + \gamma_n|\chi_\alpha^\delta|^{q-1}\operatorname{sgn}(x) = \frac{y^\delta}{\lambda_n}, \qquad \text{with } \gamma_n := \frac{\alpha}{\lambda_n^2},$$

and similar expressions hold for $\chi_{\alpha,\delta}^{II} := x_{\alpha_n,\delta_n}^{II}$. Because the term $|\chi_\alpha^\delta|^{q-1}$ is homogeneous of degree $q-1$, by an appropriate scaling, we can further simplify expressions, i.e., define the components of $p_\alpha^\delta, p_\alpha$ as

$$p_\alpha^\delta := y^\delta - \lambda_n\chi_\alpha^\delta, \qquad p_\alpha := y - \lambda_n\chi_\alpha,$$

and we use the expressions y, y^δ, Δy, Δp_α, Δp_α^{II} to denote the components of y^δ, y, Δy, Δp_α, Δp_α^{II}, respectively, where we again omit the sequence index n in the notation:

$$y^\delta = y_n^\delta, \quad y = y_n, \quad \Delta y := y_n - y_n^\delta, \qquad \Delta p_\alpha = p_{\alpha n} - p_{\alpha n}^\delta, \qquad \Delta p_\alpha^{II} := p_{\alpha_n}^{II} - p_{\alpha,\delta}^{II}.$$

Letting

$$h_q(x) := x + |x|^{q-1}\operatorname{sgn}(x), \qquad x \in \mathbb{R}, \qquad \eta_n := \gamma_n^{\frac{1}{2-q}} = \left(\frac{\alpha}{\lambda_n^2}\right)^{\frac{1}{2-q}},$$
$$\Phi_q(y) := h_q^{-1}(y),$$

we obtain via some simple calculations that

$$\chi_\alpha^\delta = \eta_n\Phi_q\left(\frac{y^\delta}{\eta_n\lambda_n}\right), \qquad \chi_{\alpha,\delta}^{II} = \eta_n\Phi_q\left(\frac{y^\delta}{\eta_n\lambda_n}, +\Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n}\right)\right), \tag{4.3}$$

$$p_\alpha^\delta = \lambda_n\eta_n\Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n}\right), \quad p_{\alpha,\delta}^{II} = \lambda_n\eta_n\left(\Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n} + \Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n}\right)\right) - \Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n}\right)\right), \tag{4.4}$$

where q^* is the conjugate index to q. Note that Φ_q corresponds to a proximal operator on $\mathbb{R}$. For $x > 0$ we have

$$x^{q-1} \leq h_q(x). \tag{4.5}$$

Moreover, Φ_q is monotonically increasing and it is not difficult to verify that for any $1 < q < \infty$,

$$x \mapsto \Phi_{q^*}(x^{\frac{1}{2-q}})^{q^*-2} \quad \text{is increasing for } x > 0. \tag{4.6}$$

We now state useful estimates for the function Φ_q.

LEMMA 4.1. *For $1 < q < \infty$, $q \neq 2$, there exist constants $\underline{D}_p, D_q$, and for any $\tau > 0$, a constant $D_{q,\tau}$, such that for all $x_1 > 0$ and $|x_2| \leq x_1$,*

$$\frac{1}{1 + \underline{D}_p \Phi_q(x_1)^{q-2}} \leq \frac{\Phi_q(x_1) - \Phi_q(x_2)}{x_1 - x_2} \leq \frac{1}{1 + \overline{D}_p \Phi_q(x_1)^{q-2}}, \tag{4.7}$$

$$\frac{1}{1 + \overline{D}_p \Phi_q(x_1)^{q-2}} \leq \begin{cases} \dfrac{1}{D_q} x_1^{\frac{2-q}{q-1}} & \text{if } 1 < q < 2, \\ \dfrac{1}{D_{q,\tau}} x_1^{\frac{2-q}{q-1}} & \text{if } q > 2 \text{ and } \forall x_1 > \tau. \end{cases} \tag{4.8}$$

Proof. For any $z_1 > 0, |z_2| \leq z_1$, we have

$$\frac{h_q(z_1) - h_q(z_2)}{z_1 - z_2} = 1 + \frac{z_1^{q-1} - |z_2|^{q-1}\operatorname{sgn}(z_2)}{z_1 - z_2} = 1 + z_1^{q-2} k\left(\frac{z_2}{z_1}\right) \quad \begin{cases} \leq 1 + z_1^{q-2}\underline{D}_p, \\ \geq 1 + z_1^{q-2}\overline{D}_p, \end{cases}$$

where

$$\overline{D}_p \leq k(z) := \frac{1 - |z|^{q-1}\operatorname{sgn}(z)}{1 - z} \leq \underline{D}_p \qquad \forall z \in [-1, 1].$$

Replacing z_i by $\Phi_q(x_i)$ yields the lower bound and the first upper bound in (4.7). In case $1 < q < 2$, we find that

$$\frac{1}{1 + \overline{D}_p \Phi_q(x_1)^{q-2}} = \frac{\Phi_q(x_1)^{2-q}}{\Phi_q(x_1)^{2-q} + \overline{D}_p} \leq \frac{x_1^{\frac{2-q}{q-1}}}{\overline{D}_p},$$

where we used that $\Phi_q(x_1)^{2-q} \geq 0$ in the denominator and the estimate $\Phi_q(x_1) \leq x_1^{\frac{1}{q-1}}$ that follows from (4.5). Now consider the case $q > 2$. Then

$$\frac{1}{1 + \overline{D}_p \Phi_q(x_1)^{q-2}} \leq \frac{C_\tau^{\frac{2-q}{q-1}}}{\overline{D}_p} x_1^{\frac{2-q}{q-1}} \qquad \forall x_1 \geq \tau,$$

where we used the estimate $h_q(x) \leq C_\tau x^{q-1}$ for $x \geq h_q(\tau)$. This yields the result. $\square$

4.1. Muckenhoupt conditions. In case the forward operator is diagonal, we may specialize the auto-regularization condition to Muckenhoupt-type inequalities [25, 22, 26] similar to the linear case. If we consider $A : X \to Y$ a compact operator and $\mathcal{R} = \|\cdot\|_{\ell^2}^2$, then the Muckenhoupt-type conditions take the following form, with some $t \in \{1, 2\}$: There exist constants C_1, C_2 such that for all Δy

$$\sum_{\{n : \frac{\sigma_n^2}{\alpha} \geq C_1\}} |\Delta y|^2 \frac{\alpha}{\sigma_n^2} \leq C_2 \sum_{\{n : \frac{\sigma_n^2}{\alpha} < C_1\}} |\Delta y|^2 \left(\frac{\sigma_n^2}{\alpha}\right)^{t-1} \qquad \forall \alpha \in (0, \alpha_{\max}), \tag{4.9}$$

where $\Delta y = \langle y^\delta - y, u_n\rangle$ with u_n the eigenvectors and σ_n^2 the eigenvalues of AA^*. The integer t is taken as $t = 1$ for the heuristic discrepancy and Hanke–Raus rules and $t = 2$ for the quasi-optimality rules. In this case, the linear auto-regularization conditions hold for the respective rules and one can prove convergence of the method. The Muckenhoupt inequalities hold in many situations; see, e.g., [25, 26].

In order to realize the insight (4.9) provides, one can observe that the right-hand side of the inequality (4.9) represents the high frequency components of the noise. Thus, in order for this upper bound to hold, one can interpret this as requiring that the noise be sufficiently irregular. In the case of the diagonal setting above, we have that $\sigma_n = \lambda_n$ and the definition of Δy agrees with that in (4.9).

For later reference, we define the following sequence of positive numbers:

$$\theta_{q,n} := \lambda_n^q \max\{|y|, |y^\delta|\}^{2-q}.$$

Then the following theorem provides a sufficient condition for the auto-regularization condition to hold.

4.1.1. The heuristic discrepancy rule.

THEOREM 4.2. *Let A be a diagonal operator* (4.1) *and $\mathcal{R}$ the regularization functional in* (4.2) *with $q \in (1,\infty)$. Suppose that for some constant C_1, there exists a constant C_2 such that for all y^δ and $\alpha \in (0, \alpha_{\max})$*

$$\sum_{\{n:\frac{\theta_{q,n}}{\alpha}\geq C_1\}} |\Delta y|^2 \frac{\alpha}{\theta_{q,n}} \leq C_2 \sum_{\{n:\frac{\theta_{q,n}}{\alpha}< C_1\}} |\Delta y|^2. \tag{4.10}$$

Then the auto-regularization condition (ARC-HD) *holds for the heuristic discrepancy principle.*

Proof. For A being a diagonal operator, the condition to prove, (3.8), may be rewritten as

$$\sum_{n\in\mathbb{N}} \Delta p_\alpha \Delta y \leq C \sum_{n\in\mathbb{N}} |\Delta p_\alpha|^2. \tag{4.11}$$

Let $I_{\mathrm{HD}} \subset \mathbb{N}$ be a set of indices where, for some fixed constants β_1, β_2, it holds that

$$n \in I_{\mathrm{HD}} \Rightarrow |\Delta y| \leq \beta_1 |\Delta p_\alpha|, \tag{4.12}$$

$$n \notin I_{\mathrm{HD}} \Rightarrow |\Delta p_\alpha| \leq \beta_2 |\Delta y| \frac{\alpha}{\theta_{q,n}}. \tag{4.13}$$

Then, we first show that for (4.11), it is sufficient that there exists a constant C_2 such that

$$\sum_{n\notin I_{\mathrm{HD}}} |\Delta y|^2 \frac{\alpha}{\theta_{q,n}} \leq C_2 \sum_{n\in I_{\mathrm{HD}}} |\Delta y|^2. \tag{4.14}$$

Indeed, splitting the sum in (4.11) into two parts and using (4.13), (4.12), and (4.14), and noting that Δp_α always has the same sign as Δy, we obtain

$$\begin{aligned}\sum_{n\in\mathbb{N}} \Delta p_\alpha \Delta y &= \sum_{n\notin I_{\mathrm{HD}}} \Delta p_\alpha \Delta y + \sum_{n\in I_{\mathrm{HD}}} \Delta p_\alpha \Delta y \leq \beta_2 \sum_{n\notin I_{\mathrm{HD}}} |\Delta y|^2 \frac{\alpha}{\theta_{q,n}} + \beta_1 \sum_{n\in I_{\mathrm{HD}}} |\Delta p_\alpha|^2 \\ &\leq \beta_2 C_2 \sum_{n\in I_{\mathrm{HD}}} |\Delta y|^2 + \beta_1 \sum_{n\in I_{\mathrm{HD}}} |\Delta p_\alpha|^2 \leq (\beta_2 C_2 \beta_1^2 + \beta_1) \sum_{n\in I_{\mathrm{HD}}} |\Delta p_\alpha|^2.\end{aligned}$$

The Lipschitz continuity of the proximal mapping $|\Delta p_\alpha| \leq |\Delta y|$ now implies (4.11). Note that (4.10) has the form of (4.14) with

$$I_{\mathrm{HD}} := \left\{ n : \frac{\theta_{q,n}}{\alpha} < C_1 \right\}.$$

Thus, it remains to verify that for this index set, there exist constants β_1, β_2 for which (4.13) and (4.12) hold.

We note that by monotonicity, the ratio $\frac{\Delta p_\alpha}{\Delta y}$ is always positive and invariant when y, y^δ are switched and when y, y^δ are replaced by $-y$, $-y^\delta$. Thus, without loss of generality, we may assume that $y > 0$ and $|y^\delta| \leq y$ such that $y = \max\{|y|, |y^\delta|\}$. Using this convention, noting that $\lambda_n \eta_n = \left(\frac{\alpha}{\lambda^q}\right)^{\frac{1}{2-q}}$, we have that

$$\left(\frac{y}{\lambda_n \eta_n}\right)^{2-q} = \frac{\max\{|y|, |y^\delta|\}^{2-q} \lambda_n^q}{\alpha} = \frac{\theta_{q,n}}{\alpha}. \tag{4.15}$$

Thus for $n \in I_{\mathrm{HD}}$ and by (4.6) we have that $\Phi_{q^*}(\frac{y}{\lambda_n \eta_n})^{q^*-2} \leq \Phi_{q^*}(C_1^{\frac{1}{2-q}})^{q^*-2}$. Using (4.4), Lemma 4.1 with $x_1 = \frac{y}{\lambda_n \eta_n} = \frac{\max\{|y|, |y^\delta|\}}{\lambda_n \eta_n}$ and $x_2 = \frac{y^\delta}{\lambda_n \eta_n}$, we find that for $n \in I_{\mathrm{HD}}$

$$\frac{\Delta p_\alpha}{\Delta y} \geq \frac{1}{1 + \underline{D}_p \Phi_{q^*}\left(\frac{y}{\lambda_n \eta_n}\right)^{q^*-2}} \geq \frac{1}{1 + \underline{D}_p \Phi_{q^*}\left(C_1^{\frac{1}{2-q}}\right)^{q^*-2}} > 0, \tag{4.16}$$

which verifies (4.12).

In view of (4.13), let $n \notin I_{\mathrm{HD}}$; then as $\frac{\theta_{q,n}}{\alpha} \geq C_1$, we conclude that

$$\frac{y}{\lambda_n \eta_n} \geq C_1^{\frac{1}{2-q}} \qquad \text{if } q \leq 2 \quad (\text{i.e., } q^* > 2). \tag{4.17}$$

Applying Lemma 4.1 with Φ_{q^*}, $x_1 = \frac{y}{\lambda_n \eta_n}$, we observe that the conditions on the right-hand side of (4.8) hold true by (4.17). Noting that the exponents in (4.7) satisfy $\frac{2-q^*}{q^*-1} = q - 2$, we obtain the upper bound

$$\frac{\Delta p_\alpha}{\Delta y} \leq \tilde{C}\left(\frac{y}{\lambda_n \eta_n}\right)^{\frac{2-q^*}{q^*-1}} = \tilde{C}\left(\frac{y}{\lambda_n \eta_n}\right)^{q-2} = \tilde{C}\frac{\alpha}{\theta_{q,n}}, \tag{4.18}$$

which verifies (4.13) and thus completes the proof. ∎

4.1.2. The Hanke–Raus rule. Contrary to the heuristic discrepancy case, we have to impose a restriction on the regularization functional exponent q in order to keep certain expressions positive.

LEMMA 4.3. *If $q \geq \frac{3}{2}$, then it follows that*

$$\Delta p_\alpha^{II} \Delta p_\alpha \geq 0$$

for all $\alpha \in (0, \alpha_{\max})$ and $y^\delta \in Y$.

Proof. Setting $z_1 = \frac{y^\delta}{\eta_n \lambda_n}$, $z_2 = \frac{y}{\eta_n \lambda_n}$, noting (4.4) and the identity $\Phi_{q^*}(z_1 + \Phi_{q^*}(z_1)) = \Phi_{q^*}\left(h_{q^*}(\Phi_{q^*}(z_1)) + \Phi_{q^*}(z_1)\right)$, it is enough to verify that the mapping

$$F : p \mapsto \Phi_{q^*}\left(h_{q^*}(p) + p\right) - p$$

is monotonically increasing. As this function is differentiable everywhere except at $p = 0$, it suffices to prove the inequality

$$0 \le F'(p) = \frac{2 + (q^* - 1)|p|^{q^*-2}}{1 + (q^* - 1)|\Phi_{q^*}(h_{q^*}(p) + p)|^{q^*-2}} - 1$$

for any $p \in \mathbb{R}$. Since F is antisymmetric and hence F' is symmetric, it is in fact sufficient to prove this inequality for $p > 0$. Setting $r = \Phi_{q^*}(h_{q^*}(p) + p) \ge p$, we thus have to show that

$$\frac{2 + (q^* - 1)|p|^{q^*-2}}{1 + (q^* - 1)|r|^{q^*-2}} \ge 1, \qquad \text{where } h_{q^*}(r) = h_{q^*}(p) + p. \tag{4.19}$$

Defining the number ζ implicitly by $h_{q^*}(\zeta p) = h_{q^*}(p) + p$ (i.e., $r = \zeta p$), it follows that $\zeta \in [1, 2]$ and that $p^{q^*-2} = \frac{2-\zeta}{\zeta^{q^*-1}-1}$. Plugging this formula and that for r into the inequality (4.19), we obtain that monotonicity holds if

$$(q^* - 1)\frac{(2 - \zeta)(\zeta^{q^*-2} - 1)}{\zeta^{q^*-1} - 1} \le 1 \qquad \forall \zeta \in [1, 2]. \tag{4.20}$$

Some detailed analysis reveals that this is satisfied for $q^* \le 3$, which is equivalent to $q \ge \frac{3}{2}$. ∎

The next lemma is needed to estimate a term in (ARC-HR).

Lemma 4.4. *Let $n \in \mathbb{N}$ be such that*

$$\frac{\theta_{q,n}}{\alpha} \le C_1 \tag{4.21}$$

with a constant C_1 that is sufficiently small. Then there is a constant β_1 depending on C_1 but independent of n with

$$\Delta p_\alpha \Delta y \le \beta_1 \Delta p_\alpha^{II} \Delta p_\alpha.$$

Proof. Define

$$x_1 = \frac{y}{\eta_n\lambda_n} + \Phi_{q^*}\left(\frac{y}{\eta_n\lambda_n}\right) = \frac{y + p_\alpha}{\eta_n\lambda_n}, \qquad x_2 = \frac{y^\delta}{\eta_n\lambda_n} + \Phi_{q^*}\left(\frac{y^\delta}{\eta_n\lambda_n}\right) = \frac{y^\delta + p_\alpha^\delta}{\eta_n\lambda_n}.$$

From the definition of $p_\alpha, p_\alpha^\delta$ it follows that $\operatorname{sgn}(x_1) = \operatorname{sgn}(y)$, $\operatorname{sgn}(x_2) = \operatorname{sgn}(y^\delta)$, and x_1, x_2 are increasing functions of $p_\alpha, p_\alpha^\delta$, respectively. Thus the ratio

$$R^{II} := \frac{\Delta p_\alpha^{II} + \Delta p_\alpha}{\Delta p_\alpha + \Delta y} = \frac{\Phi_{q^*}(x_1) - \Phi_{q^*}(x_2)}{x_1 - x_2}$$

is always positive and, moreover, invariant when x_1, x_2 are switched and, respectively, replaced by $-x_1, -x_2$. Thus, we may assume without loss of generality (otherwise we redefine the variables x_1, x_2) that $x_1 > 0$ and $|x_2| \le x_1$, which is equivalent to $y > 0$ and $|y^\delta| \le y$. Applying Lemma 4.1 yields then

$$R^{II} = \frac{\Delta p_\alpha^{II} + \Delta p_\alpha}{\Delta p_\alpha + \Delta y} = \frac{\Phi_{q^*}(x_1) - \Phi_{q^*}(x_2)}{x_1 - x_2} \ge \frac{1}{1 + \underline{D}_p \Phi_{q^*}(x_1)^{q^*-2}}.$$

It follows from (4.21) and $y \leq y + p_\alpha \leq 2y$ that

$$x_1^{2-q} = \left(\frac{y+p_\alpha}{\lambda\eta_n}\right)^{2-q} \leq \left(C_4\frac{y}{\lambda\eta_n}\right)^{2-q} = \left(C_4\frac{\theta_{q,n}}{\alpha}\right)^{2-q} = C_4^{2-q}C_1,$$

where $C_4 \in \{2,1\}$ depending on whether $q > 2$ or $q < 2$. In any case, we obtain with (4.6) that as before, $\Phi_{q^*}(\frac{y+p_\alpha}{\lambda_n\eta_n})^{q^*-2} \leq \Phi_{q^*}(C_4C_1^{\frac{1}{2-q}})^{q^*-2}$ and hence

$$R^{II} \geq \frac{1}{1+\underline{D}_p\Phi_{q^*}\left(C_4C_1^{\frac{1}{2-q}}\right)^{q^*-2}}.$$

Some standard calculus furthermore reveals that $\lim_{C_1\to 0}\Phi_{q^*}(C_4C_1^{\frac{1}{2-q}})^{q^*-2} = 0$. Thus we may choose C_1 sufficiently small such that

$$\underline{D}_p\Phi_{q^*}\left(C_4C_1^{\frac{1}{2-q}}\right)^{q^*-2} \leq \theta < 1, \tag{4.22}$$

as then $R^{II} \geq \frac{1}{1+\theta} > \frac{1}{2}$. From this inequality and using Lipschitz continuity of the residuals, $|\Delta p_\alpha| \leq |\Delta y|$, we find that

$$\begin{aligned}\Delta p_\alpha^{II}\Delta p_\alpha &= R^{II}\left(|\Delta p_\alpha|^2 + \Delta y\Delta p_\alpha\right) - |\Delta p_\alpha|^2 \geq \frac{1}{1+\theta}\Delta y\Delta p_\alpha - \left(1-\frac{1}{1+\theta}\right)|\Delta p_\alpha|^2\\ &\geq \left(\frac{2}{1+\theta}-1\right)\Delta y\Delta p_\alpha.\end{aligned}$$

This completes the proof. ∎

THEOREM 4.5. *Let A and $\mathcal{R}$ be as in Theorem* 4.2 *with $\frac{3}{2} \leq q < \infty$. Suppose that there is a sufficiently small constant C_1 and a constant C_2 such that for all y^δ,* (4.10) *holds. Then the auto-regularization condition* (ARC-HR) *holds for the Hanke–Raus rule.*

Proof. As in the HD case, we define a set of indices I_{HR} with the property

$$n \in I_{\mathrm{HR}} \Rightarrow \Delta p_\alpha\Delta y \leq \beta_1\Delta p_\alpha^{II}\Delta p_\alpha \quad \text{ and } \Delta y \leq \beta_2\Delta p_\alpha, \tag{4.23}$$

$$n \notin I_{\mathrm{HR}} \Rightarrow |\Delta p_\alpha| \leq \beta_3|\Delta y|\frac{\alpha}{\theta_{q,n}}. \tag{4.24}$$

Then, sufficient for (ARC-HR) is that a constant C_2 exists with

$$\sum_{n\notin I_{\mathrm{HR}}}|\Delta y|^2\frac{\alpha}{\theta_{q,n}} \leq C_2\sum_{n\in I_{\mathrm{HR}}}|\Delta y|^2. \tag{4.25}$$

This can be seen as follows:

$$\begin{aligned}&\sum_{n\in\mathbb{N}}\left(\Delta p_\alpha\Delta y - |\Delta p_\alpha|^2\right) \leq \sum_{n\in I_{\mathrm{HR}}}\Delta p_\alpha\Delta y + \sum_{n\notin I_{\mathrm{HR}}}\Delta p_\alpha\Delta y\\ &\leq \beta_1\sum_{n\in I_{\mathrm{HR}}}\Delta p_\alpha^{II}\Delta p_\alpha + \beta_3\sum_{n\notin I_{\mathrm{HR}}}|\Delta y|^2\frac{\alpha}{\theta_{q,n}}\\ &\leq \beta_1\sum_{n\in I_{\mathrm{HR}}}\Delta p_\alpha^{II}\Delta p_\alpha + \beta_3C_2\sum_{n\in I_{\mathrm{HR}}}|\Delta y|^2 \leq (\beta_1+\beta_3C_2\beta_2\beta_1)\sum_{n\in I_{\mathrm{HR}}}\Delta p_\alpha^{II}\Delta p_\alpha\\ &\leq (\beta_1+\beta_3C_2\beta_2\beta_1)\sum_{n\in\mathbb{N}}\Delta p_\alpha^{II}\Delta p_\alpha,\end{aligned}$$

where we used that $\Delta p_\alpha^{II} \Delta p_\alpha \geq 0$ in the last step. Hence (ARC-HR) follows from (4.25).

Note that (4.10) has the form (4.25) with

$$I_{\mathrm{HR}} := \left\{ n : \frac{\theta_{q,n}}{\alpha} \leq C_1 \right\}$$

and C_1 sufficiently small. We have already shown that for such indices, $\Delta y \leq \beta_2 \Delta p_\alpha$ holds, and for $n \notin I_{\mathrm{HR}}$, (4.24) holds. Moreover, from Lemma 4.4 it follows that on I_{HR} also $\Delta p_\alpha \Delta y \leq \beta_1 \Delta p_\alpha^{II} \Delta p_\alpha$ holds. Thus collecting these results yields that (4.10) implies the auto-regularization condition (ARC-HR). ∎

The smallness condition on C_1 is given by (4.22).

4.1.3. The symmetric quasi-optimality rule. Similarly as for the Hanke–Raus rule, we first have to verify the nonnegativity of certain expressions.

LEMMA 4.6. *If $q \geq \frac{3}{2}$, then*

$$(\Delta p_\alpha - \Delta p_\alpha^{II}) \Delta p_\alpha^{II} \geq 0$$

for all $\alpha \in (0, \alpha_{\max})$ and $y, y^\delta \in \mathbb{R}$.

Proof. Recall the mapping $F : p_\alpha \mapsto p_\alpha^{II}$ defined in the proof of Lemma 4.3. In order to prove the statement, it is enough to show that

$$((p_1 - F(p_1)) - (p_2 - F(p_2)))\,(F(p_1) - F(p_2)) \geq 0 \qquad \forall p_1, p_2.$$

It is not difficult to see that if F is monotone and Lipschitz continuous, then this inequality holds true. Thus, we have to prove that

$$0 \leq F'(p) \leq 1 \qquad \forall p.$$

As in the proof of Lemma 4.3, we may employ the variable $\zeta \in [1, 2]$, where also the inequality $0 \leq F'(p)$ was verified for $q \geq \frac{3}{2}$. The additional condition $F'(p) \leq 1$ leads to $\zeta^{q^*-2} \geq \frac{1}{2}$, which holds for any $q^* \geq 1$. This shows the result. ∎

THEOREM 4.7. *Let A and $\mathcal{R}$ be as in Theorem 4.2 with $\frac{3}{2} \leq q < \infty$. Suppose that there are constants C_1, C_2, C_3, with C_1 sufficiently small, such that for all y^δ and $\alpha \in (0, \alpha_{\max})$*

(4.26)
$$\sum_{\{n: \frac{\theta_{q,n}}{\alpha} \geq C_1\}} |\Delta y|^2 \frac{\alpha}{\theta_{q,n}} + \sum_{\{n: \frac{\theta_{q,n}}{\alpha} \leq C_1\} \cap I_2^c} |\Delta y|^2 \leq C_2 \sum_{\{n: \frac{\theta_{q,n}}{\alpha} < C_1\} \cap I_2} \left[\frac{\theta_{q,n}}{\alpha} \right]^{\frac{1}{q-1}} |\Delta y|^2,$$

where

$$I_2 = \left\{ n \in \mathbb{N} : |\Delta p_\alpha - \Delta y| \leq C_3 |\Delta p_\alpha - \Delta p_\alpha^{II}| \right\}.$$

Then the auto-regularization condition (ARC-SQR) *holds.*

Proof. We define an index set $\mathcal{I}_{\mathrm{SOR}}$ with the property that

$$n \in \mathcal{I}_{\mathrm{SOR}} \Rightarrow |\Delta p_\alpha - \Delta y| \leq \beta_1 |\Delta p_\alpha - \Delta p_\alpha^{II}| \text{ and } |\Delta p_\alpha| \leq \beta_2 |\Delta p_\alpha^{II}|. \tag{4.27}$$

Then, for (ARC-SQR) it is sufficient to prove that

$$\sum_{n \notin \mathcal{I}_{\mathrm{SOR}}} (\Delta y - \Delta p_\alpha)\Delta p_\alpha \leq C \sum_{n \in \mathcal{I}_{\mathrm{SOR}}} (\Delta p_\alpha - \Delta p_\alpha^{II})\Delta p_\alpha^{II}. \tag{4.28}$$

This can be seen as in the previous cases since the sum $\sum_{n\in\mathcal{I}_{\mathrm{SOR}}} (\Delta y - \Delta p_\alpha)\Delta p_\alpha$ can be bounded by $\sum_{n\in\mathcal{I}_{\mathrm{SOR}}} (\Delta p_\alpha - \Delta p_\alpha^{II})\Delta p_\alpha^{II}$ by the definition of $\mathcal{I}_{\mathrm{SOR}}$ and the sum $\sum_{n\notin\mathcal{I}_{\mathrm{SOR}}}$ on the right can be bounded from below by 0 according to Lemma 4.6.

Now take

$$\mathcal{I}_{\mathrm{SOR}} = \mathcal{I}_{\mathrm{HR}} \cap \mathcal{I}_2.$$

Similarly as for the Hanke–Raus rule, the inequality $|\Delta p_\alpha| \leq \beta_2 |\Delta p_\alpha^{II}|$ holds true on $\mathcal{I}_{\mathrm{HR}}$ with C_1 sufficiently small, thus $\mathcal{I}_{\mathrm{SOR}}$ satisfies the requirements (4.27). Thus, it remains to show that the stated condition (4.26) implies (4.28). The left-hand side in (4.28) can be bounded from above by

$$\begin{aligned}\sum_{n \notin \mathcal{I}_{\mathrm{SOR}}} (\Delta y - \Delta p_\alpha)\Delta p_\alpha &\leq \sum_{n \notin \mathcal{I}_{\mathrm{HR}}} (\Delta y - \Delta p_\alpha)\Delta p_\alpha + \sum_{n \mathcal{I}_{\mathrm{HR}} \cap I_2^c} (\Delta y - \Delta p_\alpha)\Delta p_\alpha \\ &\leq \sum_{n \notin \mathcal{I}_{\mathrm{HR}}} |\Delta y|^2 \frac{\alpha}{\theta_{q,n}} + \sum_{n \in \mathcal{I}_{\mathrm{HR}} \cap I_2^c} |\Delta y|^2,\end{aligned}$$

where we used the estimates (4.13) on the complement of $\mathcal{I}_{\mathrm{HR}}$ and the Lipschitz continuity of the proximal mapping for the second sum:

$$(\Delta y - \Delta p_\alpha)\Delta p_\alpha \leq |\Delta y||\Delta p_\alpha| \leq |\Delta y|^2.$$

Thus, the left-hand side of (4.26) serves as an upper bound for the left-hand side of (4.28).

The sum on the right-hand side of (4.28) can be bounded from below as follows: the summation index n is in $I_{\mathrm{SQO}} \subset I_{\mathrm{HR}}$, hence

$$\begin{aligned}&(\Delta p_\alpha - \Delta p_\alpha^{II})\Delta p_\alpha^{II} \geq |\Delta p_\alpha - \Delta y||\Delta p_\alpha| \geq \beta_1 (\Delta p_\alpha - \Delta y)\Delta y \geq \beta_1 |\Delta y|^2 \left(\frac{\Delta p_\alpha}{\Delta y} - 1\right) \\ &\geq \beta_1 |\Delta y|^2 \left(\frac{1}{1 + \underline{D}_p \Phi_{q^*}\left(\frac{y}{\lambda_n \eta_n}\right)^{q^*-2}} - 1 \right) = \beta_1 |\Delta y|^2 \frac{\underline{D}_p \Phi_{q^*}\left(\left(\frac{\theta_{q,n}}{\alpha}\right)^{\frac{1}{2-q}}\right)^{q^*-2}}{1 + \underline{D}_p \Phi_{q^*}\left(\left(\frac{\theta_{q,n}}{\alpha}\right)^{\frac{1}{2-q}}\right)^{q^*-2}} \\ &\geq \beta_1 |\Delta y|^2 \left(\frac{C}{1 + \underline{D}_p \Phi_{q^*}\left(C_1^{\frac{1}{2-q}}\right)^{q^*-2}} \right) \left(\frac{\theta_{q,n}}{\alpha}\right)^{\frac{1}{q-1}},\end{aligned}$$

where we used (4.15), (4.16), and a bound for $z > 0$ on I_{SQO} of the form

$$\Phi_{q^*}\left(z^{\frac{1}{2-q}}\right)^{q^*-2} \geq C' z^{\frac{1}{q-1}}$$

that can be obtained by similar means as above. Thus, the right-hand side of (4.26) is a lower bound for the right-hand side of (4.28). Together, (4.26) implies (4.28) and thus the desired auto-regularization condition. ∎

Remark 4.8. The condition in (4.26) has an additional sum over the index set $I_{HR} \cap I_2^c$ on the left-hand side. It might be possible to prove that this set is empty, e.g., if $I_{HR} \subset I_2$. Then the corresponding sum would vanish, and this happens in the linear case ($q = 2$). However, we postpone a more detailed analysis of this issue to the future.

We also point out that the Muckenhoupt-type conditions (4.10), and (4.26) (except for the additional sum) agree with the respective ones for the linear case $q = 2$ so that they appear, in fact, as natural extensions of the linear convergence theory.

4.2. Case study of noise restrictions. For the cases that the operator ill-posedness, the regularity of the exact solution, and the noise show some typical behavior, we investigate the restrictions that the Muckenhoupt-type condition (4.10) impose on the noise. In particular, we would like to point out that the restrictions are not at all unrealistic and they are satisfied in paradigmatic situations.

Consider a polynomially ill-posed problem, with a given decay of the exact soluton and a polynomial decay of the error:

$$\lambda_n = \frac{D_1}{n^\beta}, \quad |y| = \frac{D_2}{n^\nu}, \quad \Delta y = \delta s_n \frac{1}{n^\kappa}, \qquad \nu > \beta > 0, 0 < \kappa < \nu, \quad s_n \in \{-1, 1\}.$$

The restrictions $\kappa < \nu$, $\nu > \beta > 0$ are natural as the noise is usually less regular than the exact solution and the exact solution has higher decay rates than λ_n due to regularity. In the linear case, Muckenhoupt-type conditions lead to restrictions on the regularity of the noise, i.e., upper bounds for the decay rate κ. This is perfectly in line with their interpretation as conditions for sufficiently irregular noise.

In the following we write $\sim$ if the left and right expressions can be estimated by constants independent of n. (There might be a q-dependence, however).

The numbers $\theta_{q,n}$ that appear in (4.10) now read as

$$\begin{aligned}\theta_{q,n} &:= \max\{|y|, |y^\delta|\}^{2-q} \lambda_n^q = \max\{1, |y^\delta|/|y|\}^{q-2} |y|^{2-q} \lambda_n^q \\ &\sim \frac{1}{n^{\beta q + \nu(2-q)}} \max\left\{1, \left|1 + \frac{s_n \delta}{C_2} n^{\nu-\kappa}\right|\right\}^{2-q}.\end{aligned}$$

We additionally impose the restriction that for sufficiently large n, $\theta_{q,n} \to 0$ monotonically. If $2 - q > 0$, this is trivially satisfied, while for $2 - q < 0$, we require that

$$\beta q + \kappa(2 - q) > 0 \qquad \text{if } 2 - q < 0. \tag{4.29}$$

Under these assumptions, for any α sufficiently small, we find an n^* such that $\theta_{q,n} = C_1 \alpha$ and $\theta_{q,n} \le C_1 \alpha$ for $n \ge n^*$. Expressing α in terms of $\theta_{q,n}$ yields a sufficient condition for (4.10), as

$$\theta_{q,n^*} \sum_{n=1}^{n^*} \frac{|\Delta y|^2}{\theta_{q,n}} \le C \sum_{n=n^*+1}^{\infty} |\Delta y|^2 \sim \frac{1}{{n^*}^{2\kappa-1}}. \tag{4.30}$$

By the straightforward estimate $\max\{1, |1 + \frac{s_n \delta}{C_2} n^{\nu-\kappa}|\} \sim 1 + \delta n^{\nu-\kappa}$, the inequality (4.30) reduces to

$$\frac{(1 + \delta {n^*}^{\nu-\kappa})^{2-q}}{{n^*}^{\beta q + \nu(2-q) - 2\kappa}} \sum_{n=1}^{n^*} \frac{n^{\beta q + \nu(2-q) - 2\kappa}}{(1 + \delta n^{\nu-\kappa})^{2-q}} \le C n^*. \tag{4.31}$$

For any $x \ge 0$ and $0 \le z \le 1$, it holds that

$$1 \le \frac{1 + x}{1 + zx} \le \frac{1}{z}.$$

We use this inequality with $z = \frac{n}{n^*}$ and $x = \delta n^*$. Then, we obtain the sufficient conditions

$$\begin{cases} \dfrac{1}{{n^*}^{\beta q+\nu(2-q)-2\kappa-(2-q)(\nu-\kappa)}} \displaystyle\sum_{n=1}^{n^*} n^{\beta q+\nu(2-q)-2\kappa-(2-q)(\nu-\kappa)} \leq Cn^*, & 2-q>0, \\ \dfrac{1}{{n^*}^{\beta q+\nu(2-q)-2\kappa}} \displaystyle\sum_{n=1}^{n^*} n^{\beta q+\nu(2-q)-2\kappa} \leq Cn^*, & 2-q<0. \end{cases}$$

These inequalities are satisfied if the exponent for n is strictly larger than -1. This finally leads to the restrictions

$$\begin{cases} \kappa \leq \beta + \dfrac{1}{q}, & q<2, \\ \kappa \leq \dfrac{q}{2}\beta + \dfrac{2-q}{q}\nu + \dfrac{1}{2}, & q>2. \end{cases}$$

Note that for $q > 2$, we additionally require (4.29).

We hope to have the reader convinced that the imposed conditions on the noise are not too restrictive and, in particular, the set of noise that satisfies them is nonempty. These conditions provide a hint for which cases the methods may work or fail.

In case $\frac{3}{2} \leq q \leq 2$, both the heuristic discrepancy and the Hanke–Raus are reasonable rules. The conditions on the noise are less restrictive the smaller q is.

In case $1 < q < \frac{3}{2}$, our convergence analysis only applies to the heuristic discrepancy rule, as the nonnegativity condition of the Hanke–Raus rule is not guaranteed in this case. It could be said that the heuristic discrepancy rule is the more robust one then.

In case $q > 2$, we observe that the restriction on the noise depends on the regularity of the exact solution. For highly regular exact solutions ($\nu \gg 1$) the noise condition might fail to be satisfied as q becomes very large. This happens for both the heuristic discrepancy and the Hanke–Raus rules.

We did not include the quasi-optimality condition in this analysis as it still requires further analysis. However, the conditions for it are usually even more restrictive than for the Hanke–Raus rules and we expect similar problems for the case $q > 2$.

5. Numerical experiments. In this section, we illustrate and verify the theoretical findings of the preceding sections. It should be stressed that the preceding convergence analysis provides an important piece in understanding the behavior, but there exist further factors that influence the actual quality of the results using heuristic rules (e.g., for optimal-order results, a regularity condition on $x^\dagger$ is often required and the values of the constants in the estimates are important). It should be noted as well that our results only proved sufficient conditions for convergence, which does not necessarily mean that the methods fail when the conditions are violated. (For instance, by including δ-dependent estimates, one may find weaker convergence conditions that hold for certain ranges of the noise level). Still, a preliminary understanding can be gained from the numerical experiments in this section.

In all experiments, we consider the discretized space $\mathbb{R}^n$ and for the regularization functional, $\mathcal{R} = \frac{1}{q}\|\cdot\|_{\ell^q}^q$ with q to be specified and $\mathcal{R} = TV$ defined below. Due to the discretization, we opt to choose the parameter $\alpha_* \in [\alpha_{\min}, \alpha_{\max}]$. We also choose to select $\alpha_{\max} = \|A\|^2$ (apart from for total variation (TV) regularization below), and for the more tricky issue of the lower bound, we set $\alpha_{\min} = \sigma_{\min}$, the smallest singular value of A^*A (again excluding the TV regularization case, for which we fix $\alpha_{\min} = 10^{-9}$). Other methodologies for selecting $\alpha_{\min}$ were suggested in [23, 14].

A difficult issue in the parameter selection procedure is how global minima appearing at the boundary, i.e., $\alpha_* = \alpha_{\min}$ or $\alpha_* = \alpha_{\max}$, should be treated. For linear regularization methods, usually only the lower bound at $\alpha_{\min}$ is an issue, which can be handled with the above mentioned techniques. However, for the present convex regularization case, we observed several instances (in particular for ℓ^1-penalties), where the global minimum was at the right boundary at $\alpha_{\max}$ leading to a suboptimal choice of α_*, which is quite an unusual phenomenon (in the linear case). We treated this issue through "brute force" by explicitly excluding boundary minima and selecting α_* from the set of interior minima, even when the boundary values of ψ would have smaller values. Only when the set of interior minima is empty do we consider boundary minima again.

Additionally, we always rescale the forward operator and exact solution so that $\|A\| = \|x^\dagger\| = 1$. In each experiment, we compute 10 different realizations of the noise in which the noise level is logarithmically increasing. We also compute the error (the measure for which will differ for the various regularization schemes) induced by each parameter choice rule, as well as the optimal parameter choice, which will be computed as the minimizer of the respective error functional itself. Moreover, we include plots of the total error at each noise level for all considered scenarios (left-hand-side plots), and we also provide a supplementary plot of the parameter choice functionals (right-hand-side plots) for an example noise level which is specified in the error plot via a dotted vertical line. Note that in case we plot the ψ-functionals for $\delta = 10\%$, the aforementioned line will not be visible due to it lying on the boundary of the error graph.

For our operators, we use the tomography operator (`tomo`) from Hansen's *regularization tools* (cf. [17]) with $n = 625$ and $\mathtt{f} = 1$. We also define a diagonal matrix $A \in \mathbb{R}^{n\times n}$, initially with eigenvalues $\lambda_i = C\frac{1}{i^\beta}$ and subsequently also with $\lambda_i = C\exp(-i\beta)$ which simulate mildly and severely ill-posed problems, respectively. In either case, we take an exact solution $x^\dagger = C \cdot s_i \frac{1}{i^\nu}$ and data perturbed by noise $e_i = C\mathcal{N}(1,0)\frac{1}{i^\kappa}$, where $s_i \in \{-1, 1\}$ are random, and set the parameters as $n = 20$, $\beta = 4$, $\nu = 2$, and $\kappa = 1$.

5.1. ℓ^1 regularization. A particularly interesting application of convex variational Tikhonov regularization is the case in which $q = 1$, since it is sparsity enforcing. In fact, it is the most sparsity enforcing regularization method while still remaining a convex minimization problem. Significant work in the area of sparse regularization includes [13, 28, 31]. While it does not fit with the Muckenhoupt-type conditions we derived earlier, it is nevertheless an interesting regularization scheme for the practitioner who would be eager to see the performance of the studied rules. Note that in this case, we minimize the Tikhonov and Bregman functionals using FISTA (cf. [5]). The corresponding proximal mapping operator is the soft thresholding operator. In this experiment, we use the tomography operator defined above.

The solution $x^\dagger$ in our experiment is chosen to be sparse in a custom-built manner. For each parameter choice rule, we compute the error as

$$\mathrm{Err}_{\ell^1}(\alpha_*) = \|x^\delta_{\alpha_*} - x^\dagger\|_{\ell^1}.$$

We may observe in Figure 1 that, for smaller noise levels, the heuristic discrepancy rule appears to be the best performing, while for larger noise levels, the right quasi-optimality rule performs best. The two worst performers are the symmetric QO and HR rules, where the latter bucks this trend for small noise.

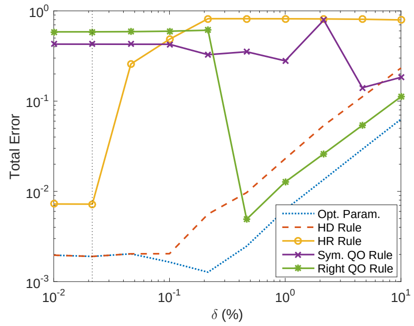

FIG. 1. *Error plots* $\mathrm{Err}_{\ell^1}(\alpha_*)$ *for different rules and optimal choice:* ℓ^1 *regularization,* `tomo` *operator.*

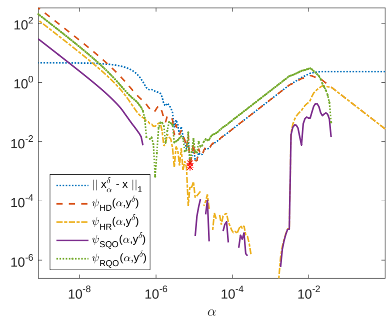

FIG. 2. *Plot of* ψ*-functionals:* ℓ^1 *regularization,* $\delta = 0.02\%$.

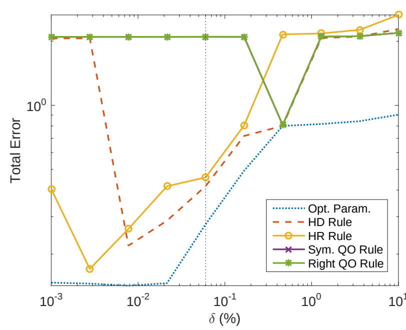

FIG. 3. *Error plots* $\mathrm{Err}_{\ell^q}(\alpha_*)$ *for different rules and optimal choice:* $\ell^{\frac{3}{2}}$ *regularization, mildly ill-posed.*

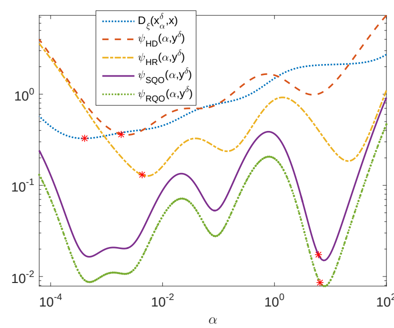

FIG. 4. *Plot of* ψ*-functionals:* $\ell^{\frac{3}{2}}$ *regularization,* $\delta = 0.06\%$.

In Figure 2, we display a plot of the surrogate functionals ψ for a specific example. In particular, the Hanke–Raus, symmetric, and right quasi-optimality functionals exhibit some erroneous negative values which hamper their performances. Indeed, in this example, negative values were frequently observed for the symmetric QO rule, less so for the HR rule, and less still for the right QO rule. This issue will be discussed later on.

5.2. $\ell^{\frac{3}{2}}$ regularization. An interesting case for the purposes of illustrating our theory is when $q = \frac{3}{2}$. Additionally, as with the previous regularization, we have an analytic formula for the proximal mapping operator corresponding to the regularization functional. In this scenario, we use the diagonal operator defined above with the given parameters and we compute the error with the Bregman distance, namely,

$$\mathrm{Err}_{\ell^q}(\alpha_*) = D_\xi(x_{\alpha_*}^\delta, x^\dagger), \qquad q \in (1, \infty).$$

We first consider the mildly ill-posed case and observe in Figure 3 that the Hanke–Raus rule is the best performing one in case that the noise level is relatively small, although for midrange noise levels, the heuristic discrepancy rule performs slightly better, and for larger noise levels still, the quasi-optimality rules match the heuristic discrepancy rule. Note that the quasi-optimality rules appear indistinguishable in this plot and we remark too that the plots of their respective functionals were very similar (see Figure 4).

The relatively poor performance of the quasi-optimality rules may be explained by the observation in Figure 4 that the selected minimizers of the quasi-optimality functionals are suboptimal. If the other *local* (instead of global) minima were selected (e.g., those left of $\alpha = 10^{-2}$), then the results would be much improved. This is a common phenomenon in many of our experiments involving the diagonal operator with $q = \frac{3}{2}$. We observe that the HD and HR functionals oscillate as well, although in Figure 4, at least, the "correct" minimizers were chosen.

In the severely ill-posed setting, we notice in Figure 5 that the heuristic rules display a mixed performance. Observe in Figure 6 that the global minimum of the quasi-optimality functionals appears to be the "wrong" choice as in the previous experiment. Indeed, the optimal parameter choice almost coincides with a local minimum of the quasi-optimality functionals. We also see in this example plot that the HD and HR rules grossly overestimate the parameter. A similar pattern was observed for other considered noise levels.

5.3. ℓ^3 regularization. Based on the Muckenhoupt-type conditions in the preceding sections, we postulated that for $q > 2$, the parameter choice rules we consider are likely to face mishaps. Consequently, we have elected to run a numerical experiment with $q = 3$ in order to illustrate what happens in practice. As in the previous experiment, we consider the diagonal operator and compute the error induced by the parameter choice rules as before.

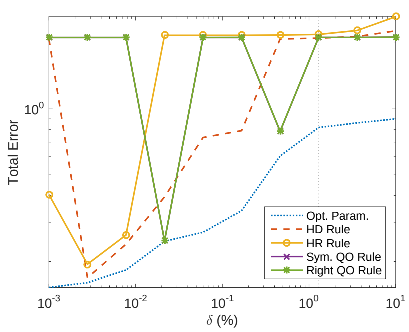

FIG. 5. *Error plots* $\mathrm{Err}_{\ell^q}(\alpha_*)$ *for different rules and optimal choice:* $\ell^{\frac{3}{2}}$, *severely ill-posed.*

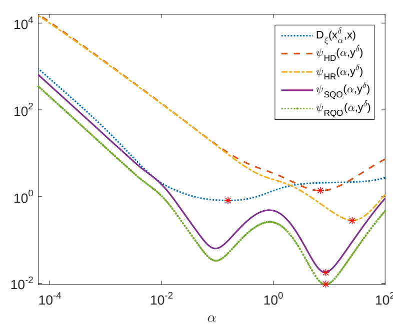

FIG. 6. *Plot of* ψ*-functionals:* $\ell^{\frac{3}{2}}$ *severely ill-posed,* $\delta = 1.3\%$.

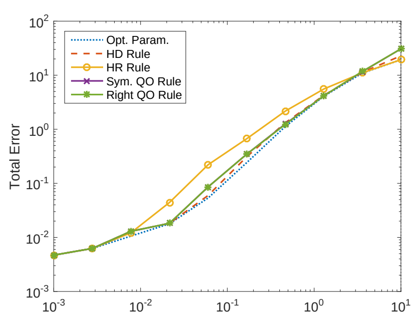

FIG. 7. *Error plots* $\mathrm{Err}_{\ell^q}(\alpha_*)$ *for different rules and optimal choice:* ℓ^3 *regularization, mildly ill-posed.*

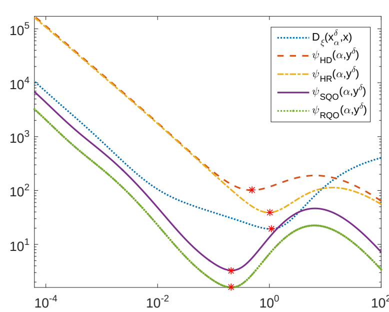

FIG. 8. *Plot of* ψ*-functionals:* ℓ^3 *regularization,* $\delta = 10\%$.

First considering the mildly ill-posed case, we see in Figure 7 that all the rules appear to perform very well. In Figure 8, we examine an example case in which the heuristic discrepancy would have selected $\alpha_* = \alpha_{\max}$, being its global minimum, were it not for our method, which effectively disqualifies it.

For the severely ill-posed setting, we observe in Figures 9 and 10 that the rules perform very well in general, with the Hanke–Raus rule being the worst and the quasi-optimality rules being the best of the bunch.

5.4. TV regularization. Selecting

$$\mathcal{R}(x) := \sup_{\substack{\phi \in C_0^\infty(\Omega;\mathbb{R}^n) \\ \|\phi\|_\infty \leq 1}} \int_\Omega x(t) \operatorname{div} \phi(t) \, \mathrm{d}t,$$

with div denoting the divergence and $\Omega \subset \mathbb{R}^n$ an open subset, yields TV regularization. For the numerical treatment and for functions on the real line, this is often discretized as $\mathcal{R} = \sum \|\Delta x\|_{\ell^1}$ with a (e.g., forward) difference operator Δ. For our numerical implementation, we used the FISTA algorithm with the proximal mapping operator for the TV functional being computed using a fast Newton-type method, courtesy of the code provided by [2, 3].

Note that in this case, we choose $\alpha_{\max}$ such that $\|x^\delta_{\alpha_{\max}}\| \leq C$ for a reasonable constant. Moreover, for each parameter choice rule, we compute the error via the so-called strict metric,

$$\mathrm{Err}_{\mathrm{TV}}(\alpha_*) = |\mathcal{R}(x^\delta_{\alpha_*}) - \mathcal{R}(x^\dagger)| + \|x^\delta_{\alpha_*} - x^\dagger\|_{\ell^1},$$

which was suggested in [24]. In this instance, we consider the tomography operator.

One may observe in Figure 11 that the right quasi-optimality rule appears to overall be the best performing one, although for smaller noise levels, the heuristic discrepancy rule seems to be preferable. The Hanke–Raus and symmetric quasi-optimality rules are generally the worst performing with a tendency to exhibit negative values; see, e.g., Figure 12.

In Figure 12, we display a plot of the ψ-functionals for one particular example in which we can see a demonstration of the symmetric quasi-optimality functional's proneness to exhibiting negative values due to numerical errors. In this particular plot, we see that even when the Hanke–Raus functional is nonnegative, it overestimates the parameter. Recall that the same phenomenon of negativity was observed for ℓ^1 regularization (cf. Figure 2), although on this occasion, the right QO rule is better behaved.

The existence of negative values is a numerical artifact as all ψ-functionals are, by definition, nonnegative. The symmetric QO functional appears to be the most susceptible of them. The negative values are only occurrent for ℓ^1 and TV penalties. Hence, we conjecture that this could be related to the lack of strict convexity of the penalty functional $\mathcal{R}$. The negative values are likely caused by numerical cancellation and, perhaps more importantly, approximation errors in the numerical computation of the minimizer of the Tikohnov functional. Indeed, when we calculate the minimizer more accurately, then negative values are less occurrent.

5.5. Summary. To summarize the numerical experiments presented above, we remark that the rules worked reasonably well, even in instances contrary to the expectations set by the theory. We observed that while none of the studied parameter choice rules were completely immune to mishaps, the heuristic discrepancy rule could

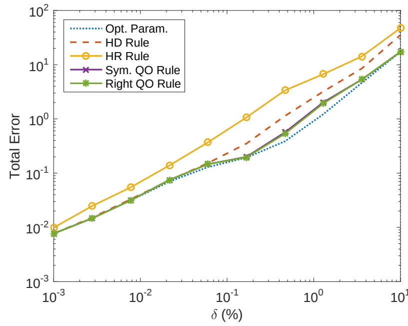

FIG. 9. *Error plots* $\mathrm{Err}_{\ell^q}(\alpha_*)$ *for different rules and optimal choice:* ℓ^3*, severely ill-posed.*

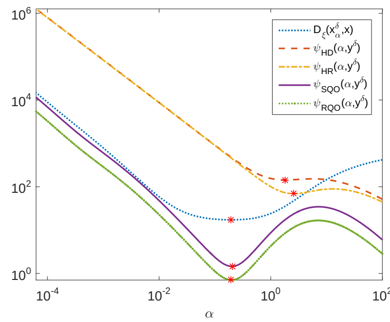

FIG. 10. *Plot of* ψ*-functionals:* ℓ^3 *severely ill-posed,* $\delta = 10\%$.

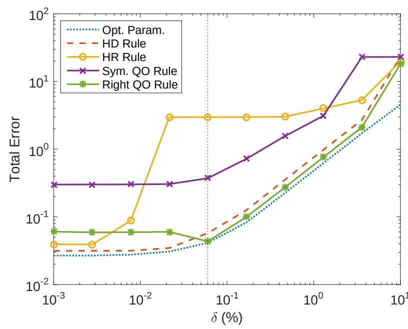

FIG. 11. *Error plots* $\mathrm{Err}_{TV}(\alpha_*)$ *for different rules and optimal choice: TV regularization,* `tomo` *operator.*

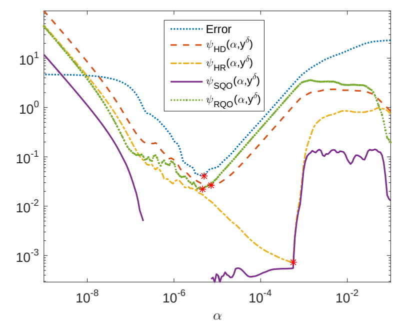

FIG. 12. *Plot of* ψ*-functionals: TV regularization,* $\delta = 0.06\%$.

perhaps be said to be the most robust overall with the right quasi-optimality rule often presenting itself as the better choice for medium to large noise levels. The symmetric quasi-optimality rule, on the other hand, appears to be less reliable (cf. sections 5.1 and 5.4) and prone to numerical errors especially for nonstrict convex penalties. For such cases, one should use other methods instead. The Hanke–Raus rule is not a stellar performer either. An important issue that seems to influence the actual performance is that the surrogate functionals ψ sometimes do not estimate the error precisely enough, for instance, when a "wrong" local minimum is selected. In the linear theory this issue can be related to the lack of certain regularity conditions for $x^\dagger$ being satisfied [22]; however, for the present convex case no such analysis is available yet, which, also in light of the above experiments, makes it difficult to offer a particular recommendation for a rule.

6. Conclusion. In conclusion, we introduced four heuristic parameter choice rules for convex Tikhonov regularization and presented a detailed analysis of the conditions we postulated for when the aforementioned rules are convergent regularization methods. This involved the more general auto-regularization conditions, as well as the reduction to the more specific Muckenhoupt-type conditions in case the forward operator is diagonal and the vector spaces are ℓ^q. Indeed, the analysis for the heuristic discrepancy and Hanke–Raus rules was more in-depth and further investigation of the conditions presented for the symmetric quasi-optimality rule presents room for further research.

We furthermore provided a numerical study to demonstrate the performance of the rules examined in this paper and also illustrate our theoretical findings. Indeed, all the rules in question performed at least reasonably well overall and allow one to conclude that heuristic rules present themselves as viable and in fact, on occasion, attractive options for convex Tikhonov regularization.

Acknowledgment. We would like to thank the referees for their valuable input, which contributed in particular to the clarity and quality of the paper.

REFERENCES

[1] A. Bakushinskiy, *Remarks on choosing a regularization parameter using quasi-optimality and ratio criterion*, USSR Comput. Math. Math. Phys., 24 (1985), pp. 181–182.

[2] A. Barbero and S. Sra, *Fast Newton-type methods for total variation regularization*, in Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 313–320.

[3] A. Barbero and S. Sra, *Modular Proximal Optimization for Multidimensional Total-Variation Regularization*, preprint, https://arxiv.org/abs/1411.0589, 2014.

[4] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Math., Springer, New York, 2017, https://books.google.at/books?id=PrM-DgAAQBAJ.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[6] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[7] M. Burger and S. Osher, *Convergence rates of convex variational regularization*, Inverse Problems, 20 (2004), pp. 1411–1421.

[8] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[9] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North-Holland, Amsterdam, 1976, https://books.google.at/books?id=gj8alAEACAAJ.

[10] J. Flemming, *Variational Source Conditions, Quadratic Inverse Problems, Sparsity Promoting Regularization: New Results in Modern Theory of Inverse Problems and an Application in Laser Optics*, Frontiers in Math., Birkhäuser/Springer, Cham, 2018, https://doi.org/10.1007/978-3-319-95264-2.

[11] J. Flemming and B. Hofmann, *A new approach to source conditions in regularization with general residual term*, Numer. Funct. Anal. Optim., 31 (2010), pp. 254–284, https://doi.org/10.1080/01630561003765721.

[12] V. Glasko and Y. Kriskin, *On the quasi-optimality principle for ill-posed problems in Hilbert space*, USSR Comput. Math. Math. Phys., 24 (1984), pp. 1–7.

[13] M. Grasmair, M. Haltmeier, and O. Scherzer, *Sparse regularization with l^q penalty term*, Inverse Problems, 24 (2008), 055020, http://stacks.iop.org/0266-5611/24/i=5/a=055020.

[14] U. Hämarik, U. Kangro, S. Kindermann, and K. Raik, *Semi-heuristic parameter choice rules for Tikhonov regularisation with operator perturbations*, J. Inverse Ill-posed Problems, 27 (2019), pp. 117–131.

[15] U. Hämarik, R. Palm, and T. Raus, *Comparison of parameter choices in regularization algorithms in case of different information about noise level*, Calcolo, 48 (2011), pp. 47–59, https://doi.org/10.1007/s10092-010-0027-4.

[16] M. Hanke and T. Raus, *A general heuristic for choosing the regularization parameter in ill-posed problems*, SIAM J. Sci. Comput., 17 (1996), pp. 956–972, https://doi.org/10.1137/0917062.

[17] P. C. Hansen, *Regularization tools version* 4.0 *for MATLAB* 7.3, Numer. Algorithms, 46 (2007), pp. 189–194.

[18] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer, *A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators*, Inverse Problems, 23 (2007), pp. 987–1010, https://doi.org/10.1088/0266-5611/23/3/009.

[19] B. Jin and D. A. Lorenz, *Heuristic parameter-choice rules for convex variational regularization based on error estimates*, SIAM J. Numer. Anal., 48 (2010), pp. 1208–1229.

[20] Q. Jin, *Hanke-Raus heuristic rule for variational regularization in banach spaces*, Inverse Problems, 32 (2016), 085008, https://doi.org/10.1088/0266-5611/32/8/085008.

[21] Q. Jin, *On a heuristic stopping rule for the regularization of inverse problems by the augmented lagrangian method*, Numer. Math., 136 (2017), pp. 973–992, https://doi.org/10.1007/s00211-016-0860-8.

[22] S. Kindermann, *Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems*, Electron. Trans. Numer. Anal., 38 (2011), pp. 233–257.

[23] S. Kindermann, *Discretization independent convergence rates for noise level-free parameter choice rules for the regularization of ill-conditioned problems*, Electron. Trans. Numer. Anal, 40 (2013), pp. 58–81.

[24] S. Kindermann, L. D. Mutimbu, and E. Resmerita, *A numerical study of heuristic parameter choice rules for total variation regularization*, J. Inverse Ill-Posed Problems, 22 (2014), pp. 63–94.

[25] S. Kindermann and A. Neubauer, *On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization*, Inverse Problems Imaging, 2 (2008), pp. 291–299.

[26] S. Kindermann, S. Pereverzyev, Jr., and A. Pilipenko, *The quasi-optimality criterion in the linear functional strategy*, Inverse Problems, 34 (2018), 075001.

[27] S. Kindermann and K. Raik, *Heuristic parameter choice rules for tikhonov regularization with weakly bounded noise*, Numer. Funct. Anal. Optim., 40 (2019), pp. 1373–1392.

[28] D. A. Lorenz, P. Maass, and P. Q. Muoi, *Gradient descent for Tikhonov functionals with sparsity constraints: Theory and numerical comparison of step size rules*, Electron. Trans. Numer. Anal., 39 (2012), pp. 437–463.

[29] V. Morozov, *On the solution of functional equations by the method of regularization*, Dokl. Akad. Nauk SSSR, 167 (1966), pp. 510–512.

[30] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Model. Simul., 4 (2005), pp. 460–489.

[31] R. Ramlau and E. Resmerita, *Convergence rates for regularization with sparsity constraints*, Electron. Trans. Numer. Anal, 37 (2010), pp. 87–104.

[32] T. Raus, *About regularization parameter choice in case of approximately given error bounds of data*, Acta Comment. Univ. Tartu, 937 (1992), pp. 77–89.

[33] R. T. Rockafellar, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63, https://doi.org/10.2307/1994612.

[34] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging*, Appl. Math. Sci., Springer, New York, 2008, https://books.google.at/books?id=MnzI4Rffr8sC.

[35] W. Yin, *Analysis and generalizations of the linearized bregman method*, SIAM J. Imaging Sci., 3 (2010), pp. 856–877, https://doi.org/10.1137/090760350.

[36] Z. Zhang and Q. Jin, *Heuristic rule for non-stationary iterated tikhonov regularization in banach spaces*, Inverse Problems, 34 (2018), 115002, https://doi.org/10.1088/1361-6420/aad918.