Convergence rates for Morozov's Discrepancy Principle using Variational Inequalities

Stephan W Anzengruber* Ronny Ramlau*†

Abstract

We derive convergence rates for Tikhonov-type regularization with convex penalty terms, where the regularization parameter is chosen according to Morozov's discrepancy principle and variational inequalities are used to generalize classical source and nonlinearity conditions. Rates are obtained first with respect to the Bregman distance and a Taylor-type distance and those results are combined to derive rates in norm and the penalty term topology.

For the special case of the sparsity promoting weighted ℓ_p-norms as penalty terms and for a searched-for solution, which is known to be sparse, the above results give convergence rates of up to linear order.

Keywords. Inverse problems, Morozov's discrepancy principle, Variational inequalities, Convergence rates, Regularization, Sparsity.

AMS subject classification. 47J06; 65J20; 49N45.

1 Introduction

Many problems arising in phyiscal applications can be modeled mathematically as an operator equation

$$F(x) = y, \tag{1}$$

where one is interested in finding a quantity x from observed data y. Examples include, but are by no means limited to, medical and astronomical imaging, inverse scattering and mathematical finance. Frequently the data y will be corrupted by noise, for instance, if they were obtained through a measurement process which is subject to inaccuracy. We will indicate the noisy version of the data by y^δ. If the operator under consideration is ill-posed, even small data errors may lead to large errors in the reconstruction.

If the data y, y^δ belong to a normed space Y, as a first step towards making problem (1) mathematically more tangible, one can consider instead the minimization of the least-squares functional

$$J(x) = \left\|F(x) - y^\delta\right\|_Y^2, \tag{2}$$

and denote the noise level by δ, i.e., $\left\|y - y^\delta\right\| \leq \delta$. This formulation allows for the definition of a generalized solution even if the noisy data do not belong to the range of the operator,

*Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, 4040 Linz, Austria, email: stephan.anzengruber@ricam.oeaw.ac.at

†Institute for Industrial Mathematics, Johannes Kepler University, Altenbergerstraße 69, 4040 Linz, Austria

$y^\delta \notin \mathrm{rg}(F)$, but problem (2) remains ill-posed. One approach to overcome the aforementioned difficulties of ill-posed problems is to replace (2) by a family $\{J_\alpha\}_{\alpha>0}$ of neighboring well-posed (or at least stable) problems, which incorporate additional a-priori knowledge of properties of the searched-for solution $x^\dagger$ through a regularizing functional $\Psi(x)$. For the purpose of this paper we will only be concerned with convex Ψ. The approximate solutions are taken to be the minimizers – denoted by x_α^δ – of the resulting variational functional

$$J_\alpha(x) = \left\|F(x) - y^\delta\right\|_Y^2 + \alpha\Psi(x). \tag{3}$$

The choice $\Psi(x) = \|x\|_X^2$, for x belonging to some Hilbert space X, constitutes the classical Tikhonov regularization. We refer the reader to [5] for further details in this respect.

Other choices which have received considerable attention in recent years, are total variation and sparsity promoting weighted ℓ_p-norms with respect to a given basis or frame $\{\phi_\lambda\}_{\lambda\in\Lambda} \subset X$,

$$\Psi_{w,p}(x) = \sum_{\lambda\in\Lambda} w_\lambda \left|\langle \phi_\lambda, x\rangle\right|^p, \qquad 0 < w_0 \le w_\lambda, 1 \le p \le 2. \tag{4}$$

Sparse representations of solutions are of strong interest, for example, in signal compression and astronomical imaging, where objects of interest like images are sparse, but their standard reconstructions are not. Enforcing sparsity adds knowledge on the solution and therefore improves the reconstruction.

The choice of the regularization parameter α in (3) turns out to be of crucial importance for the quality of the resulting reconstructions. Many strategies have been proposed in the literature and they typically lead to a somewhat different behaviour in terms of convergence and, especially, rates of convergence in the chosen topology of the regularized solutions $x_\alpha^\delta \to x^\dagger$ as $\delta \to 0$, where $x^\dagger$ denotes the searched-for solution to problem (1). For *a-priori* parameter choice rules, where $\alpha = \alpha(\delta)$ depends on the noise level δ only, convergence rates have been shown with respect to the Bregman distance in [4, 18, 19]. For the functionals $\Psi_{w,p}$ in (4) these results could even be used to obtain convergence rates in norm (cf. [7, 13, 17]).

When using Morozov's Discrepancy Principle (henceforth, MDP), which belongs to the class of *a-posteriori* parameter choice rules, i.e., the regularization parameter depends not only on the noise level, but also on the noisy data y^δ, we choose $\alpha = \alpha(\delta, y^\delta)$ such that for some minimizer x_α^δ of (3)

$$\tau_1\delta \le \left\|F\left(x^\delta_{\alpha(\delta,y^\delta)}\right) - y^\delta\right\| \le \tau_2\delta, \qquad 1 \le \tau_1 \le \tau_2$$

holds. This way of choosing the regularization parameter has been studied in great detail in its present formulation [1, 3, 15] as well as in several related variations [5, 8, 12, 14, 21].

It has been shown in [5] that for linear operator equations and certain classes of regularization methods defined via spectral decomposition in Hilbert spaces the discrepancy principle gives order optimal convergence rates. These results cover the classical Tikhonov regularization mentioned above, but not variational regularization methods with general convex penalty terms as in (3), and in particular not the functionals $\Psi_{p,w}$ in (4) for $p < 2$.

For the special case of denoising, where the operator under consideration is the identity in $L^2(\mathbb{R}^d)$, with L^1 or ℓ_1-penalty term, optimal order convergence rate results were obtained in [12]. It was also shown that the resulting regularization method does not saturate, in which case the discrepancy principle yields the same convergence rates as a-priori parameter choice rules.

For Tikhonov-type regularization of linear operator equations with general convex penalty terms as in (3) and regularization parameter chosen according to MDP, Bonesky [3] showed convergence rate results with respect to the Bregman distance in reflexive Banach spaces and his results were generalized to non-linear operators in [1] adding an additional condition on the structure of the non-linearity in F.

Finally, the residual method was studied in the report [8]. It is closely related to the discrepancy priniple and in r-convex Banach spaces, $r \geq 2$, convergence rates in norm were derived when using the penalty term $\Psi(x) = \|x\|_X^r / r$. Moreover, for linear operators additional convergence rates were provided under the assumption that the unknown solution is sparse.

In the present paper we study non-linear operators in reflexive Banach spaces and obtain convergence rates in norm of up to linear order with respect to the data error using source and non-linearity conditions formulated through variational inequalities in combination with Morozov's discrepancy principle. We show that if the searched-for solution $x^\dagger$ is sparse, then a linear convergence rate can be obtained when penalizing with $\Psi_{w,1}$ as defined in (4).

The paper is structured as follows. In Section 2 we specify our setting. Then we derive the main results about convergence rates both with respect to the Bregman distance and in norm in Section 3. Sparse recovery in Hilbert spaces is studied in Section 4 as an example, which will be found to be a special case of the framework described in Section 3, and we show that convergence rates of up to linear order can be observed for non-linear operators using MDP in combination with variational inequalities. Finally, Section 5 provides a discussion which illustrates the link between the variational inequalities used in our analysis and classical source and non-linearity conditions as well as previously considered formulations.

2 Preliminaries

Throughout this paper we assume the operator $F : \mathrm{dom}(F) \subset X \to Y$, with $0 \in \mathrm{dom}(F)$, to be weakly continuous between reflexive Banach spaces X and Y with dual spaces X^* and Y^*, respectively. However, Example 2.2 which we study in more detail in Section 4 is formulated in Hilbert spaces. Moreover, we assume that the penalty term $\Psi(x)$ fulfills the following

Condition 2.1. Let $\Psi : \mathrm{dom}(\Psi) \subset X \to \mathbb{R}^+$, with $0 \in \mathrm{dom}(\Psi)$, be a convex functional such that

(i) $\Psi(x) = 0$ if and only if $x = 0$,

(ii) Ψ is weakly sequentially lower semicontinous w.r.t. the norm in X,

(iii) Ψ is weakly coercive, i.e. $\|x_n\| \to \infty \implies \Psi(x_n) \to \infty$.

We will repeatedly encounter the following example of penalty terms throughout this paper.

Example 2.2. Let X be a Hilbert space and let $\Phi = \{\phi_\lambda\}_{\lambda\in\Lambda} \subset X$ be a *frame* for X, which means that there exist constants $A, B > 0$ such that

$$A \|x\|^2 \leq \sum_{\lambda\in\Lambda} |\langle \phi_\lambda, x\rangle|^2 \leq B \|x\|^2$$

holds for all $x \in X$. Then, for a fixed sequence $w = \{w_\lambda\}_{\lambda\in\Lambda}$ with $0 < w_0 \leq w_\lambda$ for all $\lambda \in \Lambda$ and for $1 \leq p \leq 2$, we define

$$\Psi_{p,w}(x) = \sum_{\lambda\in\Lambda} w_\lambda |\langle \phi_\lambda, x\rangle|^p , \tag{5}$$

with

$$\mathrm{dom}(\Psi_{p,w}) = \{x \in X : \Psi_{p,w}(x) < \infty\}.$$

These penalty terms have been shown to be weakly sequentially lower semicontinuous in [20, Section 3.3]. They also fulfill the remaining assumptions in Condition 2.1 as an immediate consequence of the following Lemma.

Lemma 2.3. *If $\Psi_{p,w}(x)$ is as in Example 2.2 with given $\Phi = \{\phi_\lambda\}_{\lambda\in\Lambda}$, $w = \{w_\lambda\}_{\lambda\in\Lambda}, w_0$ and p, then*

$$\|x\|^p \leq \frac{1}{w_0\sqrt{A^p}}\Psi_{p,w}(x) \tag{6}$$

holds for all $x \in X$, where A denotes the lower frame bound of Φ.

Proof. From the definition of the frame Φ, the continuous embedding of ℓ_p into ℓ_2 for $p \leq 2$ and $w_0 \leq w_\lambda$ we obtain

$$\begin{aligned} \|x\| &\leq \frac{1}{\sqrt{A}}\left(\sum_{\lambda\in\Lambda} |\langle\phi_\lambda, x\rangle|^2\right)^{1/2} \leq \frac{1}{\sqrt{A}}\left(\sum_{\lambda\in\Lambda} |\langle\phi_\lambda, x\rangle|^p\right)^{1/p} \\ &\leq \frac{1}{w_0^{1/p}\sqrt{A}}\left(\sum_{\lambda\in\Lambda} w_\lambda\, |\langle\phi_\lambda, x\rangle|^p\right)^{1/p}, \end{aligned}$$

which gives the assertion when taken to the p-th power. □

At this point we would like to fix some notational conventions.

Definition 2.4. We denote the set of all Ψ-minimizing solutions of $F(x) = y$ by $\mathcal{L}$, i.e.

$$\mathcal{L} = \{x^\dagger \in X|\ F(x^\dagger) = y \text{ and } \Psi(x^\dagger) \leq \Psi(x)\ \forall x \text{ s.t. } F(x) = y\}. \tag{7}$$

Throughout this work we assume $\mathcal{L} \neq \emptyset$ and write $\psi^\dagger$ for the common value of the penalty functional Ψ evaluated at any $x^\dagger \in \mathcal{L}$, i.e.,

$$\psi^\dagger = \Psi(x^\dagger). \tag{8}$$

Our regularization method consists in minimizing Tikhonov-type variational functionals $J_\alpha(x)$ defined as

$$J_\alpha(x) = \begin{cases} \left\|F(x) - y^\delta\right\|^q + \alpha\Psi(x) & \text{if} \quad x \in \mathcal{D} \\ +\infty & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{D} := \mathrm{dom}(F) \cap \mathrm{dom}(\Psi)$ and $q > 0$ is fixed. Hence, the regularized solutions are chosen to be minimizers of these functionals,

$$x_\alpha^\delta \in \mathcal{M}_\alpha = \underset{x\in X}{\arg\min}\{J_\alpha(x)\}. \tag{10}$$

In general, the minimizers of (9) will not be unique.

Now, let us come to the parameter choice rule of interest to us.

Definition 2.5. When using Morozov's Discrepancy Principle (MDP) we choose the regularization parameter $\alpha = \alpha(\delta, y^\delta)$ such that for some $x_\alpha^\delta \in \mathcal{M}_\alpha$

$$\tau_1\delta \leq \left\|F(x_\alpha^\delta) - y^\delta\right\| \leq \tau_2\delta \tag{11}$$

holds with fixed constants $1 \leq \tau_1 \leq \tau_2$.

It has been shown in [1] that the following condition is sufficient for the existence of $\alpha = \alpha(\delta, y^\delta)$ and $x_\alpha^\delta \in \mathcal{M}_\alpha$ fulfilling (11).

Condition 2.6. Let $\tau_1 > 1$ and assume that y^δ satisfies

$$\left\|y - y^\delta\right\| \leq \delta < \tau_2\delta < \left\|F(0) - y^\delta\right\|, \tag{12}$$

and that there is no $\alpha > 0$ with minimizers $x_1, x_2 \in \mathcal{M}_\alpha$ such that

$$\left\|F(x_1) - y^\delta\right\| < \tau_1\delta \leq \tau_2\delta < \left\|F(x_2) - y^\delta\right\|.$$

For the purpose of this paper we will assume, henceforth, that α, x_α^δ as in (11) can indeed be found, which is certainly the case if Condition 2.6 holds true.

Remark 2.7. An immediate consequence of (11) which we will need repeatedly, is that

$$\left\|F\left(x_\alpha^\delta\right) - F\left(x^\dagger\right)\right\| \leq \left\|F\left(x_\alpha^\delta\right) - y^\delta\right\| + \left\|y^\delta - y\right\| \leq (\tau_2 + 1)\delta. \tag{13}$$

■

The next Lemma can be found in [1], we give a proof here for the convenience of the reader.

Lemma 2.8. *If α is chosen according to MDP, then*

$$\Psi(x_\alpha^\delta) \leq \psi^\dagger$$

holds for all $x^\dagger \in \mathcal{L}$ and $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfying (11).

Proof. Using (11) and the minimizing property of $x_\alpha^\delta \in \mathcal{M}_\alpha$ we see that

$$\tau_1^q\delta^q + \alpha\Psi(x_\alpha^\delta) \leq \left\|F\left(x_\alpha^\delta\right) - y^\delta\right\|^q + \alpha\Psi(x_\alpha^\delta) \leq \delta^q + \alpha\psi^\dagger.$$

For $\tau_1 \geq 1$ we thus get

$$0 \leq (\tau_1^q - 1)\frac{\delta^q}{\alpha} \leq \psi^\dagger - \Psi(x_\alpha^\delta),$$

which completes the proof. □

The above assumptions are sufficient to obtain weak convergence of the regularized solutions x_α^δ to the set $\mathcal{L}$ of Ψ-minimizing solutions. For the discrepancy principle in Definition 2.5 this was proven in [1], where it was also shown that for penalty terms fulfilling the following generalized Kadec-property the convergence even takes place in the norm topology.

Condition 2.9. (Kadec property) Let $\{x_n\} \subset X$ be such that $x_n \rightharpoonup \bar{x} \in X$ and $\Psi(x_n) \to \Psi(\bar{x}) < \infty$, then x_n converges strongly to $\bar{x}$, i.e., $\|x_n - \bar{x}\| \to 0$.

Theorem 2.10. *Let $\delta_n \to 0$ and the data y^{δ_n} be such that $\left\|y - y^{\delta_n}\right\| \leq \delta_n$. If $\alpha_n = \alpha(\delta_n, y^{\delta_n})$ is chosen according to MDP and $x_n \in \mathcal{M}_{\alpha_n}$ satisfies* (11)*, then the sequence $\{x_n\}$ converges weakly to $\mathcal{L}$. If, moreover, Condition 2.9 holds, then $\{x_n\}$ converges to $\mathcal{L}$ strongly.*

For the proof we refer the reader to [1, Corollary 4.5 and Remark 4.6].

3 Convergence rates

In order to formulate the variational inequalities as well as to measure and estimate convergence rates, we will make use of the Bregman distance.

Definition 3.1. Let $\partial\Psi(x) \subset X^*$ denote the subdifferential of Ψ at $x \in X$. The generalized Bregman distance with respect to Ψ of two elements $x, z \in X$ is defined as

$$D_\Psi(x,z) = \{D_\Psi^\xi(x,z) : \xi \in \partial\Psi(z) \neq \emptyset\},$$

where

$$D_\Psi^\xi(x,z) = \Psi(x) - \Psi(z) - \langle \xi, x - z\rangle \tag{14}$$

denotes the Bregman distance with respect to Ψ and $\xi \in \partial\Psi(z)$. We remark that from here on $\langle\cdot,\cdot\rangle$ denotes the dual pairing in X^*, X or Y^*, Y and not the inner product on a Hilbert space (unless noted otherwise).

Throughout the remainder of this paper we will assume that the operator $F : X \to Y$ is Fréchet differentiable at arbitrary but fixed $x^\dagger \in \mathcal{L}$. We start by introducing the following notational conventions.

Definition 3.2. For $x \in X, x^\dagger \in \mathcal{L}$, we denote the norm of the second order Taylor remainder by

$$\mathcal{T}(x,x^\dagger) = \left\|F\big(x\big) - F\big(x^\dagger\big) - F'(x^\dagger)(x - x^\dagger)\right\|, \tag{15}$$

and call

$$D_\mathcal{T}(x,x^\dagger) = \left\|F'(x^\dagger)(x - x^\dagger)\right\| \tag{16}$$

the *Taylor distance* of x and $x^\dagger$.

In the recent work [2] Boţ and Hofmann formulated the conjecture, that convergence rate results cannot be proven if the operator F fails to satisfy a structural condition of the form

$$D_\mathcal{T}(x,x^\dagger) \leq C\sigma(\left\|F\big(x\big) - F\big(x^\dagger\big)\right\|), \tag{17}$$

where σ is a continuous, strictly increasing function through the origin and $C > 0$. We now introduce variational inequalities which are generalizations of the standard source and nonlinearity conditions discussed in Example 5.1 below, and which ultimately also fall into the framework of (17). Similar inequalities were also used in the recent works [2, 6, 7, 11].

Condition 3.3. (Variational inequalities) Define

$$V_\mathcal{L}(\rho) = \left\{x \in \mathcal{D} \;\middle|\; \Psi(x) \leq \psi^\dagger \text{ and } \left\|F\big(x\big) - y\right\| \leq \rho\right\} \tag{18}$$

and assume that for $x^\dagger \in \mathcal{L}$ there exist $\xi \in \partial\Psi(x^\dagger)$, $0 < \kappa \leq 1$, $\rho > 0$ and $\beta_i, \gamma_i \geq 0$, $i = 1,2,3$, such that

$$-\langle \xi, x - x^\dagger\rangle \leq \beta_1 D_\Psi^\xi(x,x^\dagger) + \beta_2 D_\mathcal{T}(x,x^\dagger) + \beta_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa \tag{19}$$

$$\mathcal{T}(x,x^\dagger) \leq \gamma_1 D_\Psi^\xi(x,x^\dagger) + \gamma_2 D_\mathcal{T}(x,x^\dagger) + \gamma_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa, \tag{20}$$

holds for all $x \in V_\mathcal{L}(\rho)$ and the constants β_i, γ_i fulfill

$$\beta_1 < 1, \qquad \gamma_2 < 1 \quad \text{and} \quad \frac{\beta_2\gamma_1}{(1-\beta_1)(1-\gamma_2)} < 1. \tag{21}$$

Examples of applications where variational inequalities as in Condition 3.3 are fulfilled are phase retrieval problems and inverse option pricing, which were studied in [9].

We now show that the regularized solutions $x_\alpha^\delta \in \mathcal{M}_{\alpha,y^\delta}$ belong to the sets $V_\mathcal{L}(\rho)$ for any $\rho > 0$ once δ is small enough.

Lemma 3.4. *If $0 < \delta \le \rho/(\tau_2+1)$, $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP and $x_\alpha^\delta \in \mathcal{M}_\alpha$ fulfills* (11)*, then $x_\alpha^\delta \in V_\mathcal{L}(\rho)$.*

Proof. According to Lemma 2.8 and Remark 2.7 for α and $x_\alpha^\delta \in \mathcal{M}_\alpha$ fulfilling (11), we know $\Psi(x_\alpha^\delta) \le \psi^\dagger$ and $\|F(x_\alpha^\delta) - y\| \le (\tau_2+1)\delta \le \rho$. □

Here and below we denote by $B_\varepsilon(x^\dagger)$ the open ball with radius ε centered at $x^\dagger$. Under certain conditions it suffices for the variational inequalities (19) and (20) to only hold locally.

Lemma 3.5. *Let the penalty term $\Psi(x)$ fulfill Condition 2.9 and the set $\mathcal{L}$ of Ψ-minimizing solutions consists of $x^\dagger$ only. Then to every $\varepsilon > 0$ there exists $\delta^* > 0$ such that for all $0 < \delta \le \delta^*$ and y^δ with $\|y - y^\delta\| \le \delta$ it holds that*

$$x_\alpha^\delta \in B_\varepsilon(x^\dagger) \cap V_\mathcal{L}(\rho).$$

Here, $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP and $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfies (11)*.*

Proof. Theorem 2.10 asserts that $x_\alpha^\delta \to x^\dagger$ under the assumptions of the Lemma and since $x_\alpha^\delta \in V_\mathcal{L}(\rho)$ according to Lemma 3.4 the existence of such δ^* follows. □

When using parameter choice rules other than the discrepancy principle, e.g. a-priori rules, then Lemma 2.8 may no longer hold true and the only information concerning the relation between the values of the penalty term at the regularized solutions x_α^δ and the Ψ-minimizing solutions $x^\dagger$ we might have at hand is

$$\lim_{\delta \to 0} \Psi(x_\alpha^\delta) = \Psi(x^\dagger).$$

This is, however, not sufficient to ensure that $x_\alpha^\delta \in V_\mathcal{L}(\rho)$ as defined in (18). Therefore, when working with other parameter choice rules, it might be necessary to consider larger sets

$$S_\alpha(\sigma) = \left\{ x \in \mathcal{D} \;\middle|\; \|F(x) - y\|^q + \alpha\Psi(x) \le \sigma \right\}, \tag{22}$$

with $\sigma > \alpha\psi^\dagger$ (see, for example, [2]), or

$$\tilde{V}_\mathcal{L}(\rho, \eta) = \left\{ x \in \mathcal{D} \;\middle|\; \Psi(x) \le \psi^\dagger + \eta \text{ and } \|F(x) - F(x^\dagger)\| \le \rho \right\} \tag{23}$$

for some $\rho, \eta > 0$ ([7]). We will see that these sets contain $V_\mathcal{L}(\rho)$ in Lemma 5.3. But even though this distinction is seemingly small, most of the terms in Condition 3.3 become redundant when working with MDP:

Lemma 3.6. *Condition 3.3 is equivalent to: Assume that for $x^\dagger \in \mathcal{L}$ there exist $\xi \in \partial\Psi(x^\dagger)$, $0 < \kappa \le 1$, $\rho > 0$ and $\tilde{\beta}_3, \tilde{\gamma}_3 \ge 0$ such that*

$$-\langle \xi, x - x^\dagger \rangle \le \tilde{\beta}_3 \|F(x) - F(x^\dagger)\|^\kappa \tag{24}$$

$$\mathcal{T}(x, x^\dagger) \le \tilde{\gamma}_3 \|F(x) - F(x^\dagger)\|^\kappa \tag{25}$$

holds for all $x \in V_\mathcal{L}(\rho)$.

Proof. Note first, that if (24) and (25) hold for $\tilde{\beta}_3, \tilde{\gamma}_3 \geq 0$, then Condition 3.3 is clearly fulfilled with $\beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 0$, $\beta_3 = \tilde{\beta}_3$ and $\gamma_3 = \tilde{\gamma}_3$.

To show the other implication, we write $\beta_i' = \beta_i/(1-\beta_1)$ and $\gamma_i' = \gamma_i/(1-\gamma_2)$ for $i = 1, 2, 3$. For $x \in V_{\mathcal{L}}(\rho)$ we have using $\Psi(x) \leq \psi^\dagger$

$$D_\Psi^\xi \left(x, x^\dagger\right) = \Psi(x) - \Psi(x^\dagger) - \langle \xi,\ x - x^\dagger \rangle \leq -\langle \xi,\ x - x^\dagger \rangle$$

and with $0 < \kappa \leq 1$ also

$$\left\| F\big(x\big) - F\big(x^\dagger\big) \right\| \leq \max(1, \rho) \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa .$$

Thus, if (19) holds, then from $\beta_1 < 1$ and $D_{\mathcal{T}}\left(x, x^\dagger\right) \leq \mathcal{T}(x, x^\dagger) + \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|$ we obtain for $x \in V_{\mathcal{L}}(\rho)$

$$\begin{aligned}
-\langle \xi,\ x - x^\dagger \rangle &\leq \beta_2' D_{\mathcal{T}}\left(x, x^\dagger\right) + \beta_3' \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa \\
&\leq \beta_2' \left(\mathcal{T}(x, x^\dagger) + \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|\right) + \beta_3' \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa . \\
&\leq \beta_2' \mathcal{T}(x, x^\dagger) + \left(\beta_2' \max(1, \rho) + \beta_3'\right) \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa .
\end{aligned}$$

Consequently, if (20) holds and $\gamma_2 < 1$, one gets

$$\begin{aligned}
\mathcal{T}(x, x^\dagger) &\leq -\gamma_1' \langle \xi,\ x - x^\dagger \rangle + \gamma_2' \left\| F\big(x\big) - F\big(x^\dagger\big) \right\| + \gamma_3' \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa \\
&\leq \gamma_1' \beta_2' \mathcal{T}(x, x^\dagger) + \left((\gamma_1' \beta_2' + \gamma_2') \max(1, \rho) + \gamma_1' \beta_3' + \gamma_3'\right) \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa
\end{aligned}$$

and due to (21) we find that (24) and (25) hold with

$$\begin{aligned}
\tilde{\gamma}_3 &= \frac{(\gamma_1' \beta_2' + \gamma_2') \max(1, \rho) + \gamma_1' \beta_3' + \gamma_3'}{1 - \gamma_1' \beta_2'} \\
\tilde{\beta}_3 &= \beta_2' \tilde{\gamma}_3 + \beta_2' \max(1, \rho) + \beta_3'.
\end{aligned}$$

□

Lemma 3.7. *If Condition 3.3 holds, there exist $\bar{\beta}_3, \bar{\gamma}_3 \geq 0$, such that for all $x \in V_{\mathcal{L}}(\rho)$ it holds that*

$$D_\Psi^\xi \left(x, x^\dagger\right) \leq \bar{\beta}_3 \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa \tag{26}$$

$$D_{\mathcal{T}} \left(x, x^\dagger\right) \leq \bar{\gamma}_3 \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa . \tag{27}$$

Proof. We choose $\tilde{\beta}_3, \tilde{\gamma}_3$ as in Lemma 3.6 and obtain that for all $x \in V_{\mathcal{L}}(\rho)$

$$D_\Psi^\xi \left(x, x^\dagger\right) \leq -\langle \xi, x - x^\dagger \rangle \leq \tilde{\beta}_3 \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa \tag{28}$$

As in the proof of Lemma 3.6 we use that for $x \in V_{\mathcal{L}}(\rho)$ and $0 < \kappa \leq 1$

$$\left\| F\big(x\big) - F\big(x^\dagger\big) \right\| \leq \max(1, \rho) \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa ,$$

whence in combination with (25) if follows that

$$\begin{aligned}
D_{\mathcal{T}}\left(x, x^\dagger\right) &\leq \mathcal{T}(x, x^\dagger) + \left\| F\big(x\big) - F\big(x^\dagger\big) \right\| \\
&\leq \left(\tilde{\gamma}_3 + \max(1, \rho)\right) \left\| F\big(x\big) - F\big(x^\dagger\big) \right\|^\kappa .
\end{aligned}$$

Consequently, setting $\bar{\beta}_3 = \tilde{\beta}_3$ and $\bar{\gamma}_3 = \tilde{\gamma}_3 + \max(1, \rho)$ finishes the proof. $\square$

In [1] it has been proven that for the parameter choice rule MDP, the source condition from Example 5.1 (i) and a nonlinearity condition as in (46) yield a convergence rate of order $\mathcal{O}(\delta)$ in the Bregman distance. We will now show that similar results still hold under the more general Condition 3.3 with respect to the Bregman distance and also in the Taylor distance $D_{\mathcal{T}}(x, x^\dagger)$.

Theorem 3.8. *Let Condition 3.3 hold for $x^\dagger \in \mathcal{L}$, $\xi \in \partial\Psi(x^\dagger)$. If $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP then for $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfying* (11)*, it holds that*

$$D_\Psi^\xi(x_\alpha^\delta, x^\dagger) = \mathcal{O}(\delta^\kappa) \quad \text{as} \quad \delta \to 0, \tag{29}$$

$$D_{\mathcal{T}}(x_\alpha^\delta, x^\dagger) = \mathcal{O}(\delta^\kappa) \quad \text{as} \quad \delta \to 0. \tag{30}$$

Proof. According to Lemma 3.4 we know that $x_\alpha^\delta \in V_{\mathcal{L}}(\rho)$ whenever δ is small enough. Thus, if Condition 3.3 holds, we apply Lemma 3.7 to obtain $\bar{\beta}_3$ and $\bar{\gamma}_3$ such that

$$D_\Psi^\xi\left(x_\alpha^\delta, x^\dagger\right) \le \bar{\beta}_3 \left\|F\left(x_\alpha^\delta\right) - F\left(x^\dagger\right)\right\|^\kappa \le \bar{\beta}_3 \tau_2^\kappa \delta^\kappa = \mathcal{O}(\delta^\kappa)$$

as $\delta \to 0$, where the last estimate stems from the definition of MDP in (11). Similarly,

$$D_{\mathcal{T}}\left(x_\alpha^\delta, x^\dagger\right) \le \bar{\gamma}_3 \left\|F\left(x_\alpha^\delta\right) - F\left(x^\dagger\right)\right\|^\kappa \le \bar{\gamma}_3 \tau_2^\kappa \delta^\kappa = \mathcal{O}(\delta^\kappa).$$

$\square$

To prove convergence rates with respect to the topology induced by the penalty term we introduce another variational inequality.

Condition 3.9. Let $x^\dagger \in \mathcal{L}$, $\xi \in \partial\Psi(x^\dagger)$ and assume there exist $\mu_i \ge 0, r, \rho > 0$, and $0 < \kappa \le 1$ such that for all $x \in V_{\mathcal{L}}(\rho)$ it holds

$$\Psi(x - x^\dagger)^r \le \mu_1 D_\Psi^\xi(x, x^\dagger) + \mu_2 D_{\mathcal{T}}(x, x^\dagger) + \mu_3 \left\|F(x) - F(x^\dagger)\right\|^\kappa. \tag{31}$$

Remark 3.10. If we succeed in proving that the regularized solutions x_α^δ converge strongly to $x^\dagger$ without the use of Condition 3.9, then it suffices to assume (31) for $x \in B_\varepsilon(x^\dagger) \cap V_{\mathcal{L}}(\rho)$ for arbitrarily small $\varepsilon > 0$. As shown in Lemma 3.5 this is possible if the penalty term Ψ fulfills Condition 2.9 and $\mathcal{L} = \{x^\dagger\}$. Additionally, by virtue of Theorem 3.8, convergence is ensured if Condition 3.3 holds and the set

$$\left\{z \in X \mid D_\Psi^\xi(z, x^\dagger) = D_{\mathcal{T}}(z, x^\dagger) = 0\right\}$$

consists of $x^\dagger$ only. This would be the case, for example, if Ψ is strictly convex or if $F'(x^\dagger)$ is injective. $\blacksquare$

In Section 4 below, we will see that in the context of sparse recovery Condition 3.9 is satisfied. When using MDP as the parameter choice rule the additional variational inequalitiy (31) immediately yields convergence rates.

Theorem 3.11. *If Conditions 3.3 and 3.9 hold for $x^\dagger \in \mathcal{L}$ and $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP, then*

$$\Psi(x_\alpha^\delta - x^\dagger) = \mathcal{O}(\delta^{\kappa/r}) \quad \text{as} \quad \delta \to 0 \tag{32}$$

holds for any $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfying (11)*.*

Proof. The assumptions of Theorem 3.8 hold for $x = x_\alpha^\delta$ and from (31), (29), (30), and (11) we get

$$\Psi(x_\alpha^\delta - x^\dagger)^r \leq \mu_1 D_\Psi^\xi(x_\alpha^\delta, x^\dagger) + \mu_2 D_\mathcal{T}(x_\alpha^\delta, x^\dagger) + \mu_3 \left\|F\big(x_\alpha^\delta\big) - F\big(x^\dagger\big)\right\|^\kappa = \mathcal{O}(\delta^\kappa)$$

as $\delta \to 0$, which is the desired convergence rate. □

Remark 3.12. It is worthwhile noting, that if one has

$$\left\|x - x^\dagger\right\|_X^r \leq \mu_1 D_\Psi^\xi(x, x^\dagger) + \mu_2 D_\mathcal{T}(x, x^\dagger) + \mu_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa . \tag{33}$$

instead of (31), then in complete analogy to Theorem 3.11 one obtains a convergence rate in norm, namely

$$\left\|x_\alpha^\delta - x^\dagger\right\|_X = \mathcal{O}(\delta^{\kappa/r}) \quad \text{as} \quad \delta \to 0. \tag{34}$$

This would be the case, for example, if $\Psi(x)$ is q-coercive for $2 \leq q < \infty$, i.e., for some $c_q, \rho > 0$,

$$\left\|x - x^\dagger\right\|_X^q \leq c_q D_\Psi(x, x^\dagger) \tag{35}$$

holds for all $x \in X$. It is well known, that the sparsity constraints $\Psi_{p,w}(x)$ defined in (5) fulfill (35) with $q = 2$, and for the optimal case $\kappa = 1$ we would obtain the classical rate

$$\left\|x_\alpha^\delta - x^\dagger\right\|_X = \mathcal{O}(\delta^{1/2}) \quad \text{as} \quad \delta \to 0. \tag{36}$$

But – as we will see in Section 4 – even (31) with $r = 1$ holds true in this setting and the resulting convergence result with respect to $\Psi_{p,w}$ is stronger than convergence in norm, which is why we prefer to work with formulation (31). Nevertheless, for different choices of the penalty term Ψ it may be more suitable to use (33) instead. ■

We now summarize our findings for the special case of linear operators F, where (20) becomes a tautology and

$$D_\mathcal{T}(x, x^\dagger) = \left\|F\big(x\big) - F\big(x^\dagger\big)\right\| .$$

Without loss of generality we may thus assume $\beta_2 = \gamma_i = \mu_2 = 0$ for $i = 1, 2, 3$.

Corollary 3.13. *Assume that F is a linear operator and that for $x^\dagger \in \mathcal{L}$ there exist $\xi \in \partial\Psi(x^\dagger)$, $0 < \kappa \leq 1$ and $\rho > 0$ such that*

$$-\langle \xi, x - x^\dagger \rangle \leq \beta_1 D_\Psi^\xi(x, x^\dagger) + \beta_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa$$

holds for all $x \in V_\mathcal{L}(\rho)$ with $0 \leq \beta_1 < 1$ and $\beta_3 \geq 0$. Then, if $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP and $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfies (11), we have

$$D_\Psi^\xi(x_\alpha^\delta, x^\dagger) = \mathcal{O}(\delta^\kappa) \quad \text{and} \quad D_\mathcal{T}(x_\alpha^\delta, x^\dagger) = \mathcal{O}(\delta^\kappa),$$

as $\delta \to 0$. If, furthermore, there exist $\mu_1, \mu_3 \geq 0$ and $r > 0$ such that either

$$\Psi(x - x^\dagger)^r \leq \mu_1 D_\Psi^\xi(x, x^\dagger) + \mu_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa ,$$

or

$$\left\|x - x^\dagger\right\|^r \leq \mu_1 D_\Psi^\xi(x, x^\dagger) + \mu_3 \left\|F\big(x\big) - F\big(x^\dagger\big)\right\|^\kappa ,$$

holds for all $x \in V_\mathcal{L}(\rho)$, then

$$\Psi(x_\alpha^\delta - x^\dagger) = \mathcal{O}(\delta^{\kappa/r}) \quad \text{or} \quad \left\|x_\alpha^\delta - x^\dagger\right\| = \mathcal{O}(\delta^{\kappa/r}),$$

as $\delta \to 0$, respectively.

Proof. Since Conditions 3.3 and 3.9 are satisfied, we obtain the respective results from Theorems 3.8 and 3.11 together with Remark 3.12. □

Remark 3.14. If the linear operator F is injective or if Ψ is strictly convex, then according to Remark 3.10 it would suffice for (31) to hold only locally, i.e., for $x \in B_\varepsilon(x^\dagger) \cap V_{\mathcal{L}}(\rho)$. Moreover, the set $\mathcal{L}$ of Ψ-minimizing solutions is known to be single-valued under these assumptions and therefore, if Ψ satisfies Condition 2.9, then Lemma 3.5 asserts that also (19) and (20) are only required to hold locally. ■

4 Sparse recovery

As a prominent case study, we will show now that the convergence rate from (36) for the sparsity promoting penalty terms $\Psi_{p,w}$ defined in (5) can be improved significantly when including the a-priori information that the $\Psi_{p,w}$-minimizing solution $x^\dagger$ is also sparse. To this end we show that a variational inequality as in Condition 3.9 holds true for this method.

Condition 4.1. Let X be a Hilbert space and $\{\phi_\lambda\}_{\lambda\in\Lambda}$ be a fixed frame for X (cf. Example 2.2). For elements $x \in X$ we use the shorthand notation

$$x_\lambda = \langle \phi_\lambda, x \rangle.$$

Throughout this section we assume that $x^\dagger \in \mathcal{L}$ is sparse w.r.t. $\{\phi_\lambda\}_{\lambda\in\Lambda}$, i.e., that only finitely many $x^\dagger_\lambda$ are nonzero. For any $\xi \in \partial\Psi(x^\dagger) \subset X^* = X$ the sequence $\{\xi_\lambda\}_{\lambda\in\Lambda}$ belongs to $\ell_2(\Lambda)$ and thus the set

$$J = \{\lambda \in \Lambda \mid x^\dagger_\lambda \neq 0 \vee |\xi_\lambda| \geq w_0\} \tag{37}$$

is also finite. We denote the subspace spanned by elements with indices in J by

$$U = \operatorname{span}\{\phi_\lambda \mid \lambda \in J\},$$

and the projections of X onto U and $U^\perp$ by π and $\pi^\perp$, respectively.

Moreover, we assume that $F'(x^\dagger)|_U$ is injective, i.e., for all $x, z \in X$ from $F'(x^\dagger)(x-z) = \pi^\perp(x-z) = 0$ it follows that $x = z$. Note that $F'(x^\dagger)|_U$ is clearly injective, if $F'(x^\dagger)$ satisfies the so-called *FBI property* (see, e.g., [13] and the references therein for further information).

A variational inequality of type (31) indeed holds in the sparse recovery case which will allow us to derive convergence rates whenever Condion 3.3 is satisfied. This is shown in Theorem 4.4 and Corollary 4.5 below. The proof is based on techniques from [7] and uses the following technical Lemma.

Lemma 4.2. *If condition 4.1 is satisfied, then there exists $c > 0$ such that for any $x \in X, \xi \in \partial\Psi_{p,w}(x^\dagger)$ and $\lambda \notin J$*

$$w_\lambda\, |x_\lambda|^p \leq c\, (w_\lambda\, |x_\lambda|^p - \xi_\lambda x_\lambda) \tag{38}$$

holds.

Proof. We first consider the case $p > 1$, then the unique element in the subgradient $\xi \in \partial\Psi_{p,w}(x^\dagger)$,

$$\xi = \sum_{\lambda\in\Lambda} p\, w_\lambda \operatorname{sign}(\langle\phi_\lambda, x^\dagger\rangle)\, \left|\langle\phi_\lambda, x^\dagger\rangle\right|^{p-1} \phi_\lambda,$$

satisfies $\xi_\lambda = 0$ whenever $x^\dagger_\lambda = 0$, which in turn holds for all $\lambda \notin J$, so that (38) clearly holds for all $c \geq 1$.

On the other hand, for $p = 1$ we define

$$m = \max_{\lambda \notin J} |\xi_\lambda| < w_0.$$

Here the maximum is attained since $\xi \in X^* = X$ and thus the sequence $\{\xi_\lambda\}_{\lambda\in\Lambda}$ belongs to $\ell_2(\Lambda)$. Since $0 \leq |\xi_\lambda| \leq m < w_0 \leq w_\lambda$ the choice $c = 2w_0/(w_0 - m)$ yields

$$\begin{aligned} \frac{w_\lambda}{c} |x_\lambda| &= w_\lambda |x_\lambda| - \frac{w_\lambda}{w_0} \frac{w_0 + m}{2} |x_\lambda| \\ &\leq w_\lambda |x_\lambda| - m |x_\lambda| \leq w_\lambda |x_\lambda| - \xi_\lambda x_\lambda, \end{aligned}$$

and (38) follows. □

We now show that a variational inequality (31) holds in the sparse recovery case.

Lemma 4.3. *If condition 4.1 is satisfied, then there exist $\mu_1, \mu_2 > 0$ such that for all $x \in X$*

$$\Psi_{p,w}(x - x^\dagger) \leq \mu_1 D^\xi_{\Psi_{p,w}}(x, x^\dagger) + \mu_2 D_\mathcal{T}(x, x^\dagger)^p, \qquad 1 \leq p \leq 2. \tag{39}$$

holds.

Proof. In order to estimate the difference between x and $x^\dagger$ with respect to the penalty term, we use the splitting

$$\Psi_{p,w}(x - x^\dagger) = \sum_{\lambda\in J} w_\lambda \left| x_\lambda - x^\dagger_\lambda \right|^p + \sum_{\lambda\notin J} w_\lambda \left| x_\lambda - x^\dagger_\lambda \right|^p, \tag{40}$$

and write

$$c_w = \sup_{\lambda\in J}\{w_\lambda\},$$

which is a finite number because the set J, defined in (37), is finite. Using the equivalence of norms on finite dimensional spaces, we find a constant c_p such that

$$\begin{aligned} \sum_{\lambda\in J} w_\lambda \left| x_\lambda - x^\dagger_\lambda \right|^p &\leq c_w \left\| \{x_\lambda - x^\dagger_\lambda\}_{\lambda\in J} \right\|^p_{\ell_p(J)} \\ &\leq c_w c_p \left\| \{x_\lambda - x^\dagger_\lambda\}_{\lambda\in J} \right\|^p_{\ell_2(J)} \\ &= c_w c_p \left\| \pi(x - x^\dagger) \right\|^p \end{aligned}$$

Due to the injectivity of $F'(x^\dagger)$ on U, the boundedness of $F'(x^\dagger)$ and the inequality $(a+b)^p \leq 2(a^p + b^p)$ for $a, b \geq 0$ and $p \geq 1$, we get the following estimate.

$$\begin{aligned} \left\| \pi(x - x^\dagger) \right\|^p &\leq c' \left\| F'(x^\dagger)\pi(x - x^\dagger) \right\|^p \\ &\leq 2c' \left(\left\| F'(x^\dagger)(x - x^\dagger) \right\|^p + \left\| F'(x^\dagger) \right\|^p \left\| \pi^\perp x \right\|^p \right). \end{aligned}$$

From the well known inequality $\|.\|_{\ell_2} \leq \|.\|_{\ell_p}$ for $1 \leq p \leq 2$, Lemma 4.2 and $x^\dagger_\lambda = 0$ for all

$\lambda \notin J$, it follows that

$$\begin{aligned}\left\|\pi^{\perp}x\right\|^{p} &= \Big(\sum_{\lambda\notin J} |x_\lambda|^2\Big)^{p/2} \le \sum_{\lambda\notin J} \frac{w_\lambda}{w_0}\ |x_\lambda|^p \\ &= \frac{c}{w_0}\sum_{\lambda\notin J} w_\lambda\ |x_\lambda|^p - w_\lambda\ \left|x_\lambda^\dagger\right|^p - \xi_\lambda(x_\lambda - x_\lambda^\dagger) \\ &\le \frac{c}{w_0} D^{\xi}_{\Psi_{p,w}}(x,x^\dagger),\end{aligned}$$

where the last inequality holds because all remaining summands for $\lambda \in J$ are Bregman distances $D^{\xi_\lambda}_{w_\lambda\,|.|}(x_\lambda, x_\lambda^\dagger)$, where $\xi_\lambda \in \partial(w_\lambda\ |.|)(x_\lambda^\dagger)$, and hence nonnegative.

To obtain the remaining estimates for terms corresponding to $\lambda \notin J$ in (40), we again use Lemma 4.2 and $x_\lambda^\dagger = 0$ for all $\lambda \notin J$.

$$\begin{aligned}\sum_{\lambda\notin J} w_\lambda\ \left|x_\lambda - x_\lambda^\dagger\right|^p &\le c\sum_{\lambda\notin J} w_\lambda\ |x_\lambda|^p - w_\lambda\ \left|x_\lambda^\dagger\right|^p - \xi_\lambda(x_\lambda - x_\lambda^\dagger) \\ &\le c\ D^{\xi}_{\Psi_{p,w}}(x,x^\dagger).\end{aligned}$$

Finally, collecting the above inequalities we find that

$$\begin{aligned}\Psi_{p,w}(x-x^\dagger) &= \sum_{\lambda\notin J} w_\lambda\ \left|x_\lambda - x_\lambda^\dagger\right|^p + \sum_{\lambda\in J} w_\lambda\ \left|x_\lambda - x_\lambda^\dagger\right|^p \\ &\le \mu_1\ D^{\xi}_{\Psi_{p,w}}(x,x^\dagger) + \mu_2\ \ \left\|F'(x^\dagger)(x-x^\dagger)\right\|^p\end{aligned}$$

holds for all $x \in X$, where

$$\mu_1 = 2c'c_w c_p\ \left\|F'(x^\dagger)\right\|^p \frac{c}{w_0} + c \qquad \text{and} \qquad \mu_2 = 2c'.$$

□

Theorem 4.4. *If Condition 4.1 is satisfied, then for* $\varepsilon < \left\|F'(x^\dagger)\right\|^{-1}$ *and* $x \in B_\varepsilon(x^\dagger)$ *it holds that*

$$\Psi_{p,w}(x-x^\dagger) \le \mu_1 D^{\xi}_{\Psi_{p,w}}(x,x^\dagger) + \mu_2 D_{\mathcal{T}}(x,x^\dagger), \qquad 1 \le p \le 2, \tag{41}$$

with μ_1, μ_2 *as in Lemma 4.3. Moreover, if additionally Condition 3.3 holds for arbitrary* $\rho > 0$*, then there exist* $\mu_1, \mu_2 > 0$ *such that* (41) *holds for all* $x \in V_{\mathcal{L}}(\rho)$ *as defined in* (18).

Proof. The assumptions of Lemma 4.3 are fulfilled and using (39) and that $D_{\mathcal{T}}(x,x^\dagger) \le 1$ whenever $x \in B_\varepsilon(x^\dagger)$, we find that

$$\begin{aligned}\Psi_{p,w}(x-x^\dagger) &\le \mu_1\ D^{\xi}_{\Psi_{p,w}}(x,x^\dagger) + \mu_2\ D_{\mathcal{T}}(x,x^\dagger)^p \\ &\le \mu_1 D^{\xi}_{\Psi_{p,w}}(x,x^\dagger) + \mu_2\ D_{\mathcal{T}}(x,x^\dagger).\end{aligned}$$

holds for all $x \in B_\varepsilon(x^\dagger)$. If, on the other hand, Condition 3.3 holds and $x \in V_{\mathcal{L}}(\rho)$, then according to Lemma 3.7

$$D_{\mathcal{T}}(x,x^\dagger) \le \bar{\gamma}_3\ \left\|F(x) - F(x^\dagger)\right\|^\kappa \le \bar{\gamma}_3\rho^\kappa =: C$$

Using (39) we thus obtain

$$\begin{aligned}\Psi_{p,w}(x-x^\dagger) &\le \mu_1\ D^{\xi}_{\Psi_{p,w}}(x,x^\dagger)+\mu_2\, D_{\mathcal{T}}(x,x^\dagger)^p\\ &\le \mu_1 D^{\xi}_{\Psi_{p,w}}(x,x^\dagger)+\mu_2 \max(1,C^p)\, D_{\mathcal{T}}(x,x^\dagger).\end{aligned}$$

□

As a corollary we obtain the convergence rate result.

Corollary 4.5. *If $x^\dagger \in \mathcal{L}$ satisfies Condition 3.3 and 4.1 and $\alpha = \alpha(\delta, y^\delta)$ is chosen according to MDP, then for $x_\alpha^\delta \in \mathcal{M}_\alpha$ satisfying* (11) *we obtain a convergence rate*

$$\Psi_{p,w}(x_\alpha^\delta - x^\dagger) = \mathcal{O}(\delta^\kappa) \quad \textit{as} \quad \delta \to 0. \tag{42}$$

Proof. According to Theorem 4.4 we find that Condition 3.9 is satisfied with $r = 1$ and thus Theorem 3.11 is applicable and provides the result. □

If we take X to be the sequence space ℓ_2 with the canonical basis, then the penalty terms $\Psi_{p,w}$ are powers of the weighted ℓ_p-norms, namely

$$\Psi_{p,w}(x) = \|x\|_{p,w}^p$$

and therefore (42) corresponds to a convergence rate

$$\left\|x_\alpha^\delta - x^\dagger\right\|_{p,w} = \mathcal{O}(\delta^{\kappa/p})$$

and if $\kappa = 1$ we obtain linear convergence speed for ℓ_1-regularization (compare [7, 8]).

5 Relation of variational inequalities to different types of source and nonlinearity conditions

Variational formulations of source and nonlinearity conditions have been used earlier in order to obtain convergence rate results. In this Section, we would like to draw a connection between inequalities (19) and (20) in Condition 3.3 and classical source and nonlinearity conditions. Variational inequalities can be seen as a generalization of the latter as the following examples illustrate.

Example 5.1. (i) If $\xi \in \partial\Psi(x^\dagger)$ fulfills the classical source condition

$$\xi = F'(x^\dagger)^* w, \tag{43}$$

with $w \in Y^*$, then it follows that

$$-\langle \xi, x - x^\dagger\rangle \le \left|\langle w, F'(x^\dagger)(x-x^\dagger)\rangle\right| \le \|w\|_{Y*}\, D_{\mathcal{T}}(x,x^\dagger), \tag{44}$$

and thus (19) holds with $\beta_2 = \|w\|_{Y*}$, and $\beta_1 = \beta_3 = 0$. Note, that the presence of the term $D_{\mathcal{T}}(x,x^\dagger)$ in (19) allows us to express this classical source condition through only the first variational inequality. Omitting this term and using an alternative formulation

$$-\langle \xi, x-x^\dagger\rangle \le \beta_1 D_\Psi^\xi(x,x^\dagger) + \beta_3\left\|F(x)-F(x^\dagger)\right\|^\kappa, \tag{45}$$

which has been considered, e.g., in [2, 20], one always needs to also employ some sort of structural nonlinearity condition to include this standard case in the setting. Nevertheless, (45) combined with a nonlinearity condition (such as (20)) is equivalent to Condition 3.3 in the sense of Lemma 3.6.

(ii) One of the first structural assumptions regarding nonlinearity (see, e.g., [5, 16]), was that F be Fréchet differentiable between Hilbert spaces X, Y and that the derivative be Lipschitz continuous, i.e., for some $\rho > 0$ and all $x, z \in X$ it holds

$$\|F'(x) - F'(z)\| \leq c\,\|x - z\|\,.$$

Under this assumption one can show that for classical Tikhonov regularization, where $\Psi(x) = \|x\|^2$, the following estimate holds

$$\mathcal{T}(x, x^\dagger) \leq \frac{c}{2}\,\|x - x^\dagger\|^2 = \frac{c}{2} D_\Psi^\xi(x, x^\dagger),$$

which is (20) with $\gamma_1 = c/2$ and $\gamma_2 = \gamma_3 = 0$. In [19] the variational inequality

$$\mathcal{T}(x, x^\dagger) \leq c\, D_\Psi^\xi(x, x^\dagger). \tag{46}$$

has been used as the nonlinearity condition for regularization with more general, convex penalty terms.

(iii) In [10] an operator F is defined to be *nonlinear of degree* (n_1, n_2, n_3) locally near $x^\dagger$, with $n_1, n_2 \in [0, 1], n_3 \in [0, 2]$, if for some $c, \rho > 0$ and all $x \in V_\mathcal{L}(\rho)$:

$$\mathcal{T}(x, x^\dagger) \leq c D_\mathcal{T}(x, x^\dagger)^{n_1}\,\|F(x) - F(x^\dagger)\|^{n_2}\,\|x - x^\dagger\|^{n_3}\,.$$

Taking into account the problem under consideration in [10], where X,Y are Hilbert spaces and $\Psi = \|.\|^2$, this definition may be generalized within our framework to

$$\mathcal{T}(x, x^\dagger) \leq c D_\mathcal{T}(x, x^\dagger)^{n_1}\,\|F(x) - F(x^\dagger)\|^{n_2}\, D_\Psi^\xi(x, x^\dagger)^{n_3}.$$

Applying Young's inequality twice to that last inequality we find that – whenever $n_1 + n_3 < 1$ – there exist constants γ_i such that (20) holds for x sufficiently close to $x^\dagger$ with

$$\kappa = \min\left(1, \frac{n_2}{1 - n_1 - n_3}\right).$$

Note that if one considers examples (i) and (ii) together for $\beta_1 = \gamma_2 = 0$ (cf. Lemma 3.6), then (21) becomes the well-known smallness condition

$$\frac{c}{2}\,\|w\|_{Y*} < 1.$$

Remark 5.2. Regarding the third variational inequality (31) in Condition 3.9 (or more precisely (33)), it is related to Assumption 1 in [7], where the following formulation was considered: For $x^\dagger \in \mathcal{L}$ assume that there exist $\rho, \eta, r, c_1, c_2 > 0$ such that

$$\Psi(x) - \Psi(x^\dagger) \geq c_1\,\|x - x^\dagger\|^r - c_2\,\|F(x) - F(x^\dagger)\|\,, \tag{47}$$

holds for all $x \in \tilde{V}_\mathcal{L}(\rho, \eta)$ as defined in (23).

If α is chosen according to MDP we restrict our attention to the sets $V_\mathcal{L}(\rho)$ in (18) (or even subsets thereof, cf. Lemma 3.5), where (47) is a stronger assumption than (31) because

$$\Psi(x) - \Psi(x^\dagger) \leq 0 \leq D_\Psi^\xi(x, x^\dagger) \tag{48}$$

holds for $x \in V_\mathcal{L}(\rho)$.

In [7] also the variational inequalities in Condition 3.3 are formulated with $D_\Psi^\xi(x,x^\dagger)$ replaced by $\Psi(x)-\Psi(x^\dagger)$, which imply (19) and (20) with $\beta_1=0$ and $\gamma_1=0$, respectively, due to (48). Therefore Theorems 3.8 and 3.11 remain applicable and we obtain the same convergence rates.

Furthermore, in the special case $\gamma_1=\mu_1=0$ assumption (19) can be dropped entirely as it is only needed to estimate the Bregman distance which no longer appears in our estimates.

If, on the other hand, a stronger nonlinearity condition

$$\mathcal{T}(x,x^\dagger)\le\gamma_1\big(\Psi(x)-\Psi(x^\dagger)\big)+\gamma_2 D_{\mathcal{T}}(x,x^\dagger)+\gamma_3\left\|F(x)-F(x^\dagger)\right\| \tag{49}$$

is satisfied locally for $x\in B_\varepsilon(x^\dagger)\subset\mathcal{D}$ (assuming $\mathcal{D}$ contains such a ball) and the penalty term Ψ under consideration is differentiable, then (19) necessarily holds true. As argued in [7] this can be seen by fixing $z\neq 0$ and applying (49) to $z_t=x^\dagger+tz$ (which belongs to $B_\varepsilon(x^\dagger)$ for $t>0$ small enough) and dividing by t, which yields

$$\begin{aligned}&\frac{1}{t}\left\|F(x^\dagger+tz)-F(x^\dagger)-F'(x^\dagger)(tz)\right\|\\ \le\gamma_1\frac{\Psi(x^\dagger+tz)-\Psi(x^\dagger)}{t}&+\gamma_2\left\|F'(x^\dagger)z\right\|+\gamma_3\frac{1}{t}\left\|F(x^\dagger+tz)-F(x^\dagger)\right\|.\end{aligned}$$

Taking the limit $t\to 0^+$ and choosing $z=x-x^\dagger$ for $x\in X\backslash\{x^\dagger\}$ (note, that if $x=x^\dagger$, then (19) is satisfied trivially) we obtain

$$0\le\gamma_1\langle\Psi'(x^\dagger),x-x^\dagger\rangle+(\gamma_2+\gamma_3)\left\|F'(x^\dagger)(x-x^\dagger)\right\|,$$

which is a special case of (19). This is to say that assumption (49) is strong enough locally to ensure that for differentiable penalty terms a variational source condition (19) holds globally as well. ■

Let us now establish a relation between the sets $V_{\mathcal{L}}(\rho)$, $\tilde V_{\mathcal{L}}(\rho,\eta)$ and $S_\alpha(\sigma)$ as defined in (18), (23) and (22), respectively, for different values of ρ,α and σ.

Lemma 5.3. *If $\sigma>\alpha\psi^\dagger$, then there exists $\rho>0$ such that*

$$V_{\mathcal{L}}(\rho)\subset S_\alpha(\sigma).$$

Moreover, $V_{\mathcal{L}}(\rho)\subset\tilde V_{\mathcal{L}}(\rho,\eta)$ holds for all $\rho,\eta>0$.

Proof. If $\sigma>\alpha\psi^\dagger$ and $x\in V_{\mathcal{L}}(\rho)$ for $0<\rho\le(\sigma-\alpha\psi^\dagger)^{1/q}$, then

$$\left\|F(x)-y\right\|^q+\alpha\Psi(x)\le\rho^q+\alpha\psi^\dagger\le\sigma,$$

so that $x\in S_\alpha(\sigma)$. The inclusion $V_{\mathcal{L}}(\rho)\subset\tilde V_{\mathcal{L}}(\rho,\eta)$ is an immediate consequence of the definitions of these sets. □

However, the sets $V_{\mathcal{L}}(\rho)$ do in general not contain any sublevelset $S_\alpha(\sigma)$ for a combination of parameters α and σ such that $\sigma>\alpha\psi^\dagger$. This can be seen from the following counterexample. It is in this sense that the sets $V_{\mathcal{L}}(\rho)$ are smaller than the sets $S_\alpha(\sigma)$.

Lemma 5.4. *Let F be a linear operator and let the penalty term $\Psi(x)$ be p-homogeneous (with $p>0$), i.e.,*

$$\Psi(tx)=t^p\Psi(x),\qquad\forall t\in\mathbb{R}_0^+,x\in X.$$

If $\alpha,\psi^\dagger>0$ and $\sigma>\alpha\psi^\dagger$, then there exists $\bar c>1$ such that for all $x^\dagger\in\mathcal{L}$ it holds that $\bar c x^\dagger\in S_\alpha(\rho)$, but there is no $\rho\ge 0$ such that $\bar c x^\dagger\in V_{\mathcal{L}}(\rho)$.

Proof. According to our assumption $\sigma > \alpha\psi^\dagger > 0$, it is

$$c = \frac{\sigma}{\alpha\psi^\dagger} > 1$$

and we may choose $d > 1$ such that $d^p \in (1, c)$. Consequently, in case $y \neq 0$

$$\varepsilon = \frac{(\sigma - d^p\alpha\psi^\dagger)^{1/q}}{\|y\|} > \frac{(\sigma - c\alpha\psi^\dagger)^{1/q}}{\|y\|} = 0,$$

so that $\bar{c} = \min(d, 1+\varepsilon) > 1$. Due to our choices it follows for arbitrary $x^\dagger \in \mathcal{L}$ that

$$\begin{aligned}\left\|F\left(\bar{c}x^\dagger\right) - y\right\|^q + \alpha\Psi(\bar{c}x^\dagger) = {} & \left\|\bar{c}F\left(x^\dagger\right) - y\right\|^q + \bar{c}^p\alpha\Psi(x^\dagger) \\ & = (\bar{c}-1)^q \ \|y\|^q + \bar{c}^p\alpha\psi^\dagger \\ & \leq \varepsilon^q \ \|y\|^q + d^p\alpha\psi^\dagger \leq \sigma,\end{aligned}$$

and $\bar{c}x^\dagger \in S_\alpha(\sigma)$. Note, that in case $y = 0$ we may simple choose $\bar{c} = d$ to obtain

$$\left\|F\left(\bar{c}x^\dagger\right) - y\right\|^q + \alpha\Psi(\bar{c}x^\dagger) = \bar{c}^p\alpha\psi^\dagger \leq \sigma.$$

However, in both of the above cases $\bar{c} > 1$ and thus

$$\Psi(\bar{c}x^\dagger) = \bar{c}^p\psi^\dagger > \psi^\dagger$$

which shows that $\bar{c}x^\dagger \notin V_{\mathcal{L}}(\rho)$ for any $\rho > 0$. □

Conclusion

We have studied a Tikhonov-type regularization method for ill-posed, possibly non-linear operator equations with general convex penalty terms, where the regularization parameter is chosen according to Morozov's discrepancy principle.

Under the assumption that the searched-for solution $x^\dagger$ satisfies a generalized source condition and the operator under consideration satisfies a generalized nonlinearity condition, which were formulated as variational inequalities, we found that the difference between the regularized solution obtained through our method and $x^\dagger$ when measured in the Bregman distance or a Taylor-type distance, goes to zero at a rate of δ^κ as $\delta \to 0$, where the parameter $\kappa \in (0, 1]$ allows for a relaxation of the classical source and non-linearity conditions, which are related to the case $\kappa = 1$.

Using another variational inequality (compare to (31)), which links the aforementioned Bregman- and Taylor distances to the Banach space norm or the penalty term topology, we could use the rates established for these distances to obtain a convergence rate $\mathcal{O}(\delta^{\kappa/r})$ in norm. Here the parameter r stems from the third variational inequality and even though such a constant r may be found from properties of the underlying (Banach) space and the penalty functional alone, it may be improved by additional knowledge about the true solution $x^\dagger$.

This behaviour could be observed when analyzing the situation of a solution which is known to be sparse in a Hilbert space setting, where the penalty term was chosen to be $\Psi_{p,w}$ with $1 \leq p \leq 2$ as defined in (5). A rate with $r = 2$ can always be achieved for these choices, but using the sparsity assumption the third variational inequality could be shown to hold even for $r = p$, which yields convergence rates of up to linear order, $\mathcal{O}(\delta)$, in the limiting case $\kappa = p = 1$.

Acknowledgment

The authors acknowledge support from the Austrian Science Fund (FWF) within the Austrian Academy of Sciences, project P19496-N18 and thank the anonymous referees for their valuable remarks which led to substantial improvements in the paper.

References

[1] Anzengruber S W and Ramlau R 2010 Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators *Inverse Problems* 26(2) 025001 doi:10.1088/0266-5611/26/2/025001

[2] Boţ R I and Hofmann B 2010 An extension of the variational inequality approach for obtaining convergence rates in regularization of nonlinear ill-posed problems *Journal of Integral Equations and Applications* 22(3)

[3] Bonesky T 2009 Morozov's discrepancy principle and Tikhonov-type functionals *Inverse Problems* 25(1) 015015 doi:10.1088/0266-5611/25/1/015015

[4] Burger M and Osher S 2004 Convergence rates of convex variational regularization *Inverse Problems* 20(5) 1411–21 doi:10.1088/0266-5611/20/5/005

[5] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* vol. 375 of *Mathematics and its Application* (Dordrecht: Kluwer Academic Publishers)

[6] Flemming J and Hofmann B 2010 A new approach to source conditions in regularization with general residual term *Numer. Funct. Anal. Optimiz.* 31 254–284

[7] Grasmair M, Haltmeier M and Scherzer O 2008 Sparse regularization with ℓ^q penalty term *Inverse Problems* 24(5) 1–13

[8] Grasmair M, Haltmeier M and Scherzer O 2009 The Residual Method for Regularizing Ill-posed Problems *Photoacoustic Imaging in Medicine and Biology Report* 14

[9] Hofmann B, Kaltenbacher B, Poeschl C and Scherzer O 2007 A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators *Inverse Problems* 23(3) 987–1010

[10] Hofmann B and Scherzer O 1994 Factors influencing the ill-posedness of nonlinear problems *Inverse Problems* 10(6) 1277–97 doi:10.1088/0266-5611/10/6/007

[11] Hofmann B and Yamamoto M 2010 On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems *Applicable Analysis* 89(11) 1705–1727 doi: 10.1080/00036810903208148

[12] Justen L and Ramlau R 2009 A general framework for soft-shrinkage with applications to blind deconvolution and wavelet denoising *Appl. and Comp. Harmonic Anal.* 26(1) 43–63

[13] Lorenz D A 2008 Convergence rates and source conditions for Tikhonov regularization with sparsity constraints *Journal of Inverse and Ill-Posed Problems* 16 463–478

[14] Morozov V A 1984 *Methods for solving incorrectly posed problems* (New York: Springer-Verlag)

[15] Ramlau R 2002 Morozov's discrepancy principle for Tikhonov regularization of non-linear operators *Journal for Numer. Funct. Anal. and Opt.* 23 147–172

[16] Ramlau R 2008 Regularization properties of Tikhonov regularization with sparsity constraints *Electron. Trans. Numer. Anal.* 30 54–74

[17] Ramlau R and Resmerita E 2010 Convergence rates for regularization with sparsity constraints *Electronic Transactions on Numerical Analysis* 37 87–104

[18] Resmerita E 2005 Regularization of ill-posed problems in Banach spaces: convergence rates *Inverse Problems* 21 1303–1314

[19] Resmerita E and Scherzer O 2006 Error estimates for non-quadratic regularization and the relation to enhancement *Inverse Problems* 22 801–814

[20] Scherzer O, Grasmair M, Grossauer H, Haltmeier M and Lenzen F 2009 *Variational Methods in Imaging* (New York: Springer-Verlag)

[21] Tikhonov A N, Leonov A S and Yagola A G 1998 *Nonlinear Ill-posed Problems* (London: Chapman & Hall)