Generalized Bregman Distances and Convergence Rates for Non-convex Regularization Methods

Markus Grasmair

Computational Science Center
University of Vienna
Nordbergstr. 15
A–1090 Vienna, Austria

E-mail: `Markus.Grasmair@univie.ac.at`

Abstract. We generalize the notion of Bregman distance using concepts from abstract convexity in order to derive convergence rates for Tikhonov regularization with non-convex regularization terms. In particular, we study the non-convex regularization of linear operator equations on Hilbert spaces, showing that the conditions required for the application of the convergence rates results are strongly related to the standard range conditions from the convex case. Moreover, we consider the setting of sparse regularization, where we show that a rate of order $\delta^{1/p}$ holds, if the regularization term has a slightly faster growth at zero than $|t|^p$.

AMS classification scheme numbers: 65J20; 47A52, 52A01

Keywords: Tikhonov regularization, non-convex regularization, sparsity, convergence rates, Bregman distance, abstract convexity.

Submitted to: *Inverse Problems*

1. Introduction

Convergence rates in regularization theory provide quantitative, asymptotic estimates about the quality of the approximative solution of an ill-posed operator equation as the noise level decreases to zero. In the case of classical, quadratic Tikhonov regularization for the solution of a linear equation $Fx = y$, $F\colon X \to Y$ being some bounded, linear operator between the Hilbert spaces X and Y, which can be formulated as the minimization of the Tikhonov functional

$$\mathcal{T}(x,y) := \|Fx - y\|^2 + \alpha\|x\|^2 ,$$

a classical result reads as follows (see, for instance, [9]): Let $x^\dagger$ be the norm minimizing solution of the equation $Fx = y^\dagger$. If y^δ are noisy data satisfying $\|y^\dagger - y^\delta\| \le \delta$ for some $\delta > 0$ and $x_\alpha^\delta := \arg\min_x \mathcal{T}(x, y^\delta)$ is the regularized solution of the equation $Fx = y^\delta$, then we obtain, with a parameter choice $\alpha \sim \delta$, a convergence rate

$$\|x_\alpha^\delta - x^\dagger\| = O(\sqrt{\delta}) \qquad \text{as } \delta \to 0 ,$$

provided the solution $x^\dagger$ satisfies the *range condition* $x^\dagger \in \operatorname{Ran} F^*$. More generally, it is possible to derive similar convergence rates of power type, if the solution $x^\dagger$ satisfies a range condition of the form $x^\dagger \in \operatorname{Ran}(F^*F)^{\lambda/2}$ for some $0 < \lambda \le 2$. For instance, the condition $x^\dagger \in \operatorname{Ran}(F^*F)$ implies the convergence rate $\|x_\alpha^\delta - x^\dagger\| = O(\delta^{2/3})$ for a choice of the regularization parameter $\alpha \sim \delta^{2/3}$.

In the case of regularization on Banach spaces, but also for non-quadratic regularization on Hilbert spaces with general convex regularization functionals, the situation is more complex. The first problem is, how to formulate range conditions in the case of Banach spaces, where the adjoint of F is an operator mapping into the dual of X, which, in general, is not isomorphic to X. The second problem is that it is not obvious why convergence rates in the norm should hold at all. Indeed, many of the results to be derived arrive at rates only in considerably weaker distance measures.

In [4], it has been argued that the *Bregman distance* is the correct measure for determining the quality of the approximate solution x_α^δ in the case of convex regularization on Banach spaces. This distance is defined as the difference between the (convex) regularization functional and its linear approximation around $x^\dagger$. More precisely, if $\mathcal{R}$ is a convex and differentiable functional on a Banach space X and $x^\dagger$, $x \in X$, then the Bregman distance between $x^\dagger$ and x is defined as

$$D(x^\dagger; x) := \mathcal{R}(x) - \mathcal{R}(x^\dagger) - \mathcal{R}'(x^\dagger)(x - x^\dagger) .$$

Then, if the true solution $x^\dagger$ satisfies the range condition $\mathcal{R}'(x^\dagger) \in \operatorname{Ran} F^*$, one obtains the convergence rate

$$D(x^\dagger; x_\alpha^\delta) = O(\delta) \qquad \text{as } \delta \to 0 .$$

In addition, it has been shown in [17] that the improved rate $D(x^\dagger; x_\alpha^\delta) = O(\delta^{4/3})$ holds for a parameter choice $\alpha \sim \delta^{2/3}$, if the target space Y is a Hilbert space and the range condition $\mathcal{R}'(x^\dagger) \in \operatorname{Ran}(F^*F)$ is satisfied. In the particular case of quadratic regularization on Hilbert spaces, these results recover precisely the classical convergence rates, because, in this situation, $D(x^\dagger; x_\alpha^\delta) = \|x_\alpha^\delta - x^\dagger\|^2$ and $\mathcal{R}'(x^\dagger) = 2x^\dagger$.

While the results stated above only apply if the operator F is linear, there also exist analogous convergence rates for nonlinear operators. For their definition, it is, however, necessary to restrict the non-linearity of the operator. For quadratic regularization on Hilbert spaces, typically, one requires that F is Fréchet differentiable and its Fréchet derivative satisfies an additional continuity condition. Then, convergence rates like in the linear case can be derived under the range condition $x^\dagger \in \operatorname{Ran}(F'(x^\dagger)^* F'(x^\dagger))^{\lambda/2}$.

Although the smoothness assumptions on F appear natural, it has been shown in [14] that, in fact, they are not necessary. Observing that the range condition $\mathcal{R}'(x^\dagger) \in \operatorname{Ran}(F^*)$ can be interpreted as a separation condition for the functionals $x \mapsto \mathcal{R}(x) - \mathcal{R}(x^\dagger)$ and $x \mapsto \|Fx - y^\dagger\|^2$ and that the smoothness conditions in the non-linear case only serve for guaranteeing a similar separation locally near $x^\dagger$, the authors argue that convergence rates results for more general operators can be obtained, if one postulates such a separation directly instead of first linearizing the functional at $x^\dagger$ and then imposing restrictions on the behaviour of F' near $x^\dagger$. Thus, they arrive at conditions in the form of *variational inequalities*, which can be written as

$$\beta D(x^\dagger; x_\alpha^\delta) \leq (\mathcal{R}(x) - \mathcal{R}(x^\dagger)) + \gamma \|F(x) - y^\dagger\| \,. \tag{1}$$

Indeed, this inequality is a direct generalization of the range condition for the linear case: the exposition in [20, Section 3.2] shows that, for bounded linear functionals F, the inequality (1) holds if and only if $x^\dagger$ satisfies the range condition $\mathcal{R}'(x^\dagger) \in \operatorname{Ran} F^*$.

In this paper, we will show that the method of variational inequalities can also be generalized to work with non-convex regularization functionals $\mathcal{R}$. To that end, we define a generalized notion of the Bregman distance, as, in its original definition, it only makes sense for convex functions. This generalization uses an abstraction of the notion of convexity that relies an more general dualities than the one between a Banach space X and the Banach space X^* of all bounded linear functionals on X. Moreover, generalizing the results of [16], we consider more general similarity terms than the squared norm of the residual.

The main result of this paper is Theorem 3.3, which provides bounds on the (generalized) Bregman distance between the regularized solution x_α^δ of the perturbed equation and the true solution $x^\dagger$. In addition, Corollary 3.4 yields convergence rates for a parameter choice that depends on the behaviour of the similarity term near $F(x^\dagger)$. For the case of metric regularization, where the similarity term is some power of the distance on the target space Y, these rates reduce precisely to the ones derived in [14, 16, 20] (see Example 3.6).

In Sections 4 and 5 two exemplary applications of the results derived in Section 3 are presented. The first deals with regularization on Hilbert spaces using a non-convex regularization term and a power of the norm of the residual as similarity term. Assuming that $\mathcal{R}$ has a proximal subdifferential in the sense of Clarke et al. [6], that is, it can be approximated from below by a quadratic function, we show that a variational inequality with respect to the Bregman distance derived from these approximation holds, if the element $x^\dagger$ satisfies a condition of range type (see Proposition 4.1). In the convex case this condition reduces precisely to the standard condition $\partial\mathcal{R}(x^\dagger) \cap \operatorname{Ran} F^* \neq \emptyset$. As a concrete example, a functional suited for phase separation is treated.

Section 5 considers the setting of *sparse regularization*, which aims at enforcing sparsity of the regularized solutions with respect to some given basis (or frame) of the space X. This can be achieved by applying a sub-quadratic penalization on the coefficients with respect to this basis. For instance, in [7], the usage of the ℓ^q norm of the coefficients with $1 \leq q < 2$ has been suggested. Note that the particular case $q = 1$ is strongly related to the field of *compressed sensing* [5, 8]. It has been shown in [11] by using a variational inequality for the q-th power of the norm that in such a situation it is possible to obtain a rate $\|x_\alpha^\delta - x^\dagger\| = O(\delta^{1/q})$, if one assumes sparsity of $x^\dagger$ and a restricted injectivity of the (linear) operator F. These result have been generalized in [3] to arbitrary convex functionals of q-linear growth near zero. In addition, similar results were obtained in [12] for the residual method. In this paper we will derive convergence rates for non-convex regularization functionals of the same type. It turns out that the obtained rates are only slightly weaker than the ones that hold in the convex case.

2. Abstract Convexity

In this section, we introduce the notion of Bregman distance that will be used for the derivation of convergence rates. The definition is based on an abstract approach to convexity, which largely follows the exposition in [21, Chapter 8], though in a simplified setting.

Before we can define the generalized notions of convexity, it is necessary to introduce some notation concerning addition and subtraction on the extended real line $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$.

Definition 2.1 *We define the upper and lower addition and subtraction on* $\mathbb{R} \cup \{\pm\infty\}$ *as the extensions of the usual definitions satisfying*

$$
\begin{aligned}
+\infty \dot{+} (-\infty) &= -\infty \dot{+} \infty = +\infty\,,\\
+\infty \underset{\cdot}{+} (-\infty) &= -\infty \underset{\cdot}{+} \infty = -\infty\,,\\
+\infty \dot{-} (+\infty) &= -\infty \dot{-} (-\infty) = +\infty\,,\\
+\infty \underset{\cdot}{-} (+\infty) &= -\infty \underset{\cdot}{-} (-\infty) = -\infty\,.
\end{aligned}
$$

Definition 2.2 (Generalized Conjugation) *Let X be some set and W a family of functions $w\colon X \to \bar{\mathbb{R}}$. The (generalized)* Fenchel conjugate *with respect to W of a function $\mathcal{R}\colon X \to \bar{\mathbb{R}}$ is the function $\mathcal{R}^*\colon W \to \bar{\mathbb{R}}$ defined by*

$$
\mathcal{R}^*(w) := \sup_{x\in X}\Big[w(x) \underset{\cdot}{-} \mathcal{R}(x)\Big]\,.
$$

The double conjugate of $\mathcal{R}$ is the function $\mathcal{R}^{**}\colon X \to \bar{\mathbb{R}}$ given by

$$
\mathcal{R}^{**}(x) := \sup_{w\in W}\Big[w(x) \underset{\cdot}{-} \mathcal{R}^*(w)\Big] = \sup_{w\in W}\inf_{\tilde{x}\in X}\Big[w(x) \underset{\cdot}{+} (\mathcal{R}(\tilde{x}) \dot{-} w(\tilde{x}))\Big]\,.
$$

The function $\mathcal{R}$ is convex *with respect to W, if $\mathcal{R}^{**} = \mathcal{R}$.*

The function $\mathcal{R}$ is locally convex at $x \in X$ *with respect to W, if $\mathcal{R}^{**}(x) = \mathcal{R}(x)$.*

Remark 2.3 *With the above definition of convexity, a function $\mathcal{R}$ is locally convex at $x \in X$ with respect to W, if and only if for every $\varepsilon > 0$ there exists $w_\varepsilon \in W$ such that*

$$\mathcal{R}(\tilde{x}) \geq \mathcal{R}(x) \dotplus (w_\varepsilon(\tilde{x}) \underset{\cdot}{-} w_\varepsilon(x)) - \varepsilon \tag{2}$$

for all $\tilde{x} \in X$. Therefore, local convexity is in some sense a global property of $\mathcal{R}$, as it requires knowledge of the function $\mathcal{R}$ on the whole domain X, not only in a neighbourhood of x.

Lemma 2.4 *The function $\mathcal{R}$ is locally convex at $x \in X$ with respect to W, if and only if*

$$\mathcal{R}(x) = \sup\Big\{w(x)+c : w \in W,\ c \in \mathbb{R},\ w(\tilde{x})+c \leq \mathcal{R}(\tilde{x}) \textit{ for all } \tilde{x} \in X\Big\}\ . \tag{3}$$

Proof: See [21, Corollary 8.2, Remark 8.15(b)]. □

Remark 2.5 *In [21, Chapter 5], a slightly different definition of convexity with respect to W has been introduced. There, the mapping $\mathcal{R}\colon X \to \bar{\mathbb{R}}$ is said to be convex with respect to W, if*

$$\mathcal{R} = \sup\{w \in W : w \leq \mathcal{R}\}\ .$$

Lemma 2.4 implies that this coincides with the notion of convexity introduced in Definition 2.2, if the set W is closed with respect to addition of scalars, that is, if $w \in W$ implies that $w + c \in W$ for all $c \in \mathbb{R}$.

In the following, we always assume that W is a family of functions $w\colon X \to \bar{\mathbb{R}}$. In order to exclude trivialities we assume that W is non-empty.

Definition 2.6 (W-subdifferential) *Let $\mathcal{R}$ be locally convex at $x \in X$ with respect to W and assume that $\mathcal{R}(x) \in \mathbb{R}$. The W*-subdifferential *of $\mathcal{R}$ at $x \in X$, denoted by $\partial_W \mathcal{R}(x)$, is defined as the set of all $w \in W$ that satisfy $w(x) \in \mathbb{R}$ and*

$$\mathcal{R}(\tilde{x}) \geq \mathcal{R}(x) + (w(\tilde{x}) - w(x))$$

for all $\tilde{x} \in X$.

Remark 2.7 *The W-subdifferential is related to the generalized Fenchel conjugate in the usual manner. That is, we have $w \in \partial_W \mathcal{R}(x)$ if and only if $\mathcal{R}(x) = w(x) - \mathcal{R}^*(w)$.*

Definition 2.8 (W-Bregman Distance) *Let $\mathcal{R}$ be locally convex at $x \in X$ with respect to W and assume that $\partial_W \mathcal{R}(x) \neq \emptyset$. For $w \in \partial_W \mathcal{R}(x)$ and $\tilde{x} \in X$ we define the* W-Bregman distance between x and $\tilde{x}$ with respect to w *as*

$$D_W^w(x;\tilde{x}) := (\mathcal{R}(\tilde{x}) - \mathcal{R}(x)) \dot{-} (w(\tilde{x}) - w(x))\ .$$

The W-Bregman distance is non-negative and satisfies $D_W^w(x;x) = 0$.

Example 2.9 (Convexity) *Assume that X is a Banach space and $W = X^*$ is the space of bounded linear mappings on X. Then a function $\mathcal{R}$ is convex with respect to X^* if and only if it is lower semi-continuous and convex in the usual sense. Moreover, the W-Bregman distance coincides with the usual Bregman distance defined in standard convex analysis defined by*

$$D^{\xi}(x;\tilde{x}) = \mathcal{R}(\tilde{x}) - \mathcal{R}(x) - \langle \xi, \tilde{x} - x\rangle \tag{4}$$

with $\xi \in \partial\mathcal{R}(x)$. Therefore, the abstract notion of convexity used in this paper is indeed a generalization of the standard notion.

Example 2.10 (Generalized subdifferentiability) *Let X be a locally convex space and consider the space W of all negative semi-definite, continuous quadratic functions on X. That is, $w \in W$ if and only if there exist $c \in \mathbb{R}$, $\xi \in X^*$, and a positive semi-definite, symmetric, bounded quadratic form A on X such that*

$$w(x) = c + \langle \xi, x\rangle - A(x,x)$$

for all $x \in X$. Then (2) implies that a function $\mathcal{R}\colon X \to \bar{\mathbb{R}}$ is locally convex at $x \in X$ with respect to W, if and only if there exist for every $\varepsilon > 0$ some $\xi_\varepsilon \in X^*$ and a positive semi-definite, symmetric, bounded quadratic form A_ε on X such that

$$\begin{aligned}\mathcal{R}(\tilde{x}) + \varepsilon &\geq \mathcal{R}(x) + \langle \xi_\varepsilon, \tilde{x}\rangle - A_\varepsilon(\tilde{x},\tilde{x}) - \langle \xi_\varepsilon, x\rangle + A_\varepsilon(x,x)\\ &= \mathcal{R}(x) + \langle \xi_\varepsilon, \tilde{x} - x\rangle - 2A_\varepsilon(x, \tilde{x} - x) - A_\varepsilon(\tilde{x} - x, \tilde{x} - x)\ .\end{aligned}$$

Defining $\tilde{\xi}_\varepsilon \in X^*$ by $\langle \tilde{\xi}_\varepsilon, \hat{x}\rangle = \langle \xi_\varepsilon, \hat{x}\rangle - 2A_\varepsilon(x,\hat{x})$, it follows that $\mathcal{R}\colon X \to \bar{\mathbb{R}}$ is locally convex at $x \in X$ with respect to W, if and only if there exist $\tilde{\xi}_\varepsilon \in X^*$ and a positive semi-definite, symmetric, bounded quadratic form A_ε on X such that

$$\mathcal{R}(\tilde{x}) + \varepsilon \geq \mathcal{R}(x) + \langle \tilde{\xi}_\varepsilon, \tilde{x} - x\rangle - A_\varepsilon(\tilde{x} - x, \tilde{x} - x)$$

for all $\tilde{x} \in X$. Moreover, the W-subdifferential of $\mathcal{R}$ at x consists of all functions $w(\tilde{x}) = c + \langle \xi, \tilde{x} - x\rangle - A(\tilde{x} - x, \tilde{x} - x)$ that satisfy

$$\mathcal{R}(\tilde{x}) \geq \mathcal{R}(x) + \langle \xi, \tilde{x} - x\rangle - A(\tilde{x} - x, \tilde{x} - x)$$

for all $\tilde{x} \in X$. The W-Bregman distance between x and $\tilde{x}$ therefore reads as

$$D_W^w(x;\tilde{x}) = \mathcal{R}(\tilde{x}) - \mathcal{R}(x) - \langle \xi, \tilde{x} - x\rangle + A(\tilde{x} - x, \tilde{x} - x)\ . \tag{5}$$

Formally, the sole difference between the standard convex Bregman distance (4) and this generalized definition is the additional quadratic term in (5), which guarantees that the Bregman distance stays non-negative.

Example 2.11 (Generalized local subdifferentiability) *In the following we consider a localized variant of the generalized differentiability introduced in Example 2.10. To that end we assume again that X is a locally convex space and consider the set W_l of all locally negative semi-definite, continuous quadratic functions on X. That is, $w \in W_l$, if and only if there exist $x_0 \in X$, a neighbourhood U of x_0, $c \in \mathbb{R}$, $\xi \in X^*$, and a positive semi-definite, symmetric, bounded quadratic form A on X such that*

$$w(x) = c + \langle \xi, x\rangle - A(x,x) \qquad \text{for all } x \in U\ .$$

As in Example 2.10, it follows that $w \in \partial_{W_l}\mathcal{R}(x)$ *if and only if there exist a neighbourhood* U *of* x*,* $\xi \in X^*$*, and a positive semi-definite, symmetric, bounded quadratic form* A *such that*

$$\mathcal{R}(\tilde{x}) \geq \mathcal{R}(x) + \langle \xi, \tilde{x} - x \rangle - A(\tilde{x} - x, \tilde{x} - x) \qquad \text{for all } \tilde{x} \in U \,. \tag{6}$$

Assume now that X *is a Banach space. Then the inequality (6) is closely related to the notion of* proximal differentiability *of* $\mathcal{R}$ *(see [6]): Recall that the* proximal subdifferential $\partial_P\mathcal{R}(x)$ *of* $\mathcal{R}$ *at* x *is defined as the set of all* $\xi \in X^*$ *that satisfy, for some* $\sigma > 0$ *and* $\varepsilon > 0$*, the inequality*

$$\mathcal{R}(\tilde{x}) \geq \mathcal{R}(x) + \langle \xi, \tilde{x} - x \rangle - \sigma \|\tilde{x} - x\|^2 \tag{7}$$

for all $\tilde{x} \in X$ *with* $\|\tilde{x} - x\| < \varepsilon$*. The only difference between the inequalities (6) and (7) is that the quadratic form in the latter is simply a multiple of the squared norm on* X*, while in the first case any bounded quadratic form* A *can be used.*

In particular it follows that a function $\mathcal{R}$ *has a proximal subdifferential at* $x \in X$ *if and only if* $\mathcal{R}$ *is locally* W_l*-convex at* x *and its subdifferential with respect to* W_l *is non-empty. Moreover the proximal subdifferential consists of all the linear parts of all* W_l*-subgradients at* x *when written in the form (6).*

Remark 2.12 *In Examples 2.9–2.11, the local convexity of the topological vector space* X *is only needed in order to guarantee that the dual space* X^* *is rich enough as to permit the existence of non-trivial continuous linear functionals; else the only* W*-convex functionals would be constant and thus the applicability of the theory rather limited. In principle, however, the same definitions can also be used for arbitrary topological vector spaces. In particular, this applies to the interesting case of* ℓ^p *spaces with* $0 < p < 1$*, which are not locally convex, but whose topological dual space is* ℓ^∞*.*

3. Generalized Convergence Rates

We will now apply the definitions introduced in Section 2 for the derivation of (generalized) convergence rates for Tikhonov regularization. The goal is the stable solution of an operator equation

$$F(x) = y^\dagger \,, \tag{8}$$

where F is a mapping between the sets X and Y.

We assume that the right hand side in (8) is known only approximately, that is, instead of $y^\dagger$ we are only given noisy data $y^\delta \in Y$ close to $y^\dagger$. In addition, we assume that we can estimate the difference between y^δ and the unknown true data $y^\dagger$ in terms of some distance like functional $\mathcal{S}\colon Y \times Y \to \bar{\mathbb{R}}_{\geq 0}$ on the space Y. More precisely, we know that y^δ satisfies the inequality $\mathcal{S}(y^\dagger, y^\delta) \leq \delta$.

For the stable approximate solution of (8) we consider some regularization term $\mathcal{R}\colon X \to \bar{\mathbb{R}}_{\geq 0}$ and define the *Tikhonov functional* $\mathcal{T}_\alpha\colon X \times Y \to \bar{\mathbb{R}}$ as

$$\mathcal{T}_\alpha(x, y) := \mathcal{S}(F(x), y) + \alpha\mathcal{R}(x) \,.$$

Given some noise level $\delta > 0$ and noisy data $y^\delta \in Y$ we denote for every regularization parameter $\alpha > 0$ the approximate solution of $F(x) = y^\delta$ by

$$x_\alpha^\delta := \arg\min_{x \in X} \mathcal{T}_\alpha(x, y^\delta) \; .$$

In case we have no uniqueness, we denote by x_α^δ any minimizer of $\mathcal{T}_\alpha(\cdot, y^\delta)$.

Collecting the results of [14, 16, 20], we see that the minimization of $\mathcal{T}_\alpha$ is a well-defined regularization method (that is, it attains a solution that is stable with respect to data perturbations and converges to the true solution as the noise level decreases to zero), if the following conditions are satisfied for some topologies on X and Y:

$(A1)$ The distance measure $\mathcal{S}$ satisfies, for some $s \geq 1$, the quasi-triangle inequality

$$\mathcal{S}(y_0, y_1) \leq s(\mathcal{S}(y_0, y_2) + \mathcal{S}(y_2, y_1)) \tag{9}$$

for all y_0, y_1, $y_2 \in Y$.

$(A2)$ We have $\mathcal{S}(y_0, y_1) = 0$ if and only if $y_0 = y_1$.

$(A3)$ If $(y_k)_{k \in \mathbb{N}} \subset Y$ is a sequence satisfying $\mathcal{S}(y_k, y) \to 0$, then $y_k \to y$.

$(A4)$ For all $y \in Y$, the functional $(x, y) \mapsto \mathcal{S}(F(x), y)$ is sequentially lower semi-continuous. Here we define $\mathcal{S}(F(x), y) := +\infty$ if $x \notin \operatorname{Dom}(F)$.

$(A5)$ The functional $\mathcal{R}$ is sequentially lower semi-continuous.

$(A6)$ For all $\alpha > 0$, $y \in Y$, and $t \in \mathbb{R}$ the set $\{x \in X : \mathcal{T}_\alpha(x, y) \leq t\}$ is sequentially pre-compact.

Basically, assumptions $(A4)$–$(A6)$ guarantee the existence of a minimizer of $\mathcal{T}_\alpha$, while $(A1)$–$(A3)$ are required to obtain stability of the method, and convergence as the noise level decreases to zero. In fact, as far as the well-posedness of the regularization method is concerned, assumption $(A1)$ can be weakened: One only requires that for every $y \in Y$ there exist $\delta > 0$ such that $\mathcal{S}(y_0, y_1) < \infty$ whenever $\mathcal{S}(y_0, y) < \delta$ and $\mathcal{S}(y, y_1) < \infty$. The stronger assumption $(A1)$, however, will be required below for the derivation of convergence rates. More details on Tikhonov regularization in such a general setting can be found in [16].

For the derivation of convergence rates with respect to the generalized Bregman distance we will employ the method of variational inequalities, which has been introduced in [14] for proving convergence rates for Tikhonov regularization on Banach spaces, where the operator F is non-linear (and even possibly non-smooth).

In the following we always denote by $x^\dagger$ any $\mathcal{R}$-minimizing solution of the equation $F(x) = y^\dagger$, that is,

$$x^\dagger \in \arg\min\{\mathcal{R}(x) : x \in X, \; F(x) = y^\dagger\} \; .$$

Definition 3.1 *Let W be a family of extended real valued functions on X, and assume that $\mathcal{R}$ is W-convex at $x^\dagger$ and $\partial_W \mathcal{R}(x^\dagger) \neq \emptyset$. We say that the regularization method satisfies a* variational inequality *at $x^\dagger \in X$ with respect to W, if there exist $\beta > 0$, $\varepsilon > 0$,* *a neighbourhood U of $x^\dagger$, $w \in \partial_W \mathcal{R}(x^\dagger)$, and a concave, continuous, strictly increasing function $\Phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $\Phi(0) = 0$ such that*

$$\beta D_W^w(x^\dagger; x) \leq (\mathcal{R}(x) - \mathcal{R}(x^\dagger)) + \Phi(\mathcal{S}(F(x), F(x^\dagger))) \tag{10}$$

for all $x \in \operatorname{Dom} F \cap U$ satisfying $|\mathcal{R}(x) - \mathcal{R}(x^\dagger)| < \varepsilon$.

We note that a similar form of variational inequalities has been recently considered in [2], though only within the setting of standard convex analysis.

Remark 3.2 *Let X be a Banach space and assume that assumptions (A4)–(A6) hold with respect to the weak topology on X. Then, the regularization method satisfies a variational inequality at $x^\dagger$ with respect to W, if there exists a neighbourhood U of $x^\dagger$ with respect to the weak topology such that (10) holds for all $x \in \operatorname{Dom} F \cap U$ with $\mathcal{R}(x)$ sufficiently close to $\mathcal{R}(x^\dagger)$.*

Now recall that the functional $\mathcal{R}$ satisfies the Radon–Riesz *property, if the weak convergence of a sequence $(x_k)_{k\in\mathbb{N}}$ to some $x \in X$ together with the convergence $\mathcal{R}(x_k) \to \mathcal{R}(x) \in \mathbb{R}$ implies that $\|x_k - x\| \to 0$. If this is the case, then it is sufficient to verify (10) on a norm ball around $x^\dagger$ in order to prove the validity of a variational inequality. This can be of advantage, because the norm topology is often easier to deal with than the weak topology (see for instance the proof of Theorem 5.2 below).*

Theorem 3.3 *Assume that a variational inequality at $x^\dagger \in X$ with respect to W is satisfied, and let $\beta > 0$ and $\Phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be as in Definition 3.1. Let $\delta > 0$ and assume that $y^\delta \in Y$ satisfies $\mathcal{S}(y^\dagger, y^\delta) \leq \delta$ and $\mathcal{S}(y^\delta, y^\dagger) \leq \delta$. Moreover let $x_\alpha^\delta \in \arg\min_x \mathcal{T}_\alpha(x, y^\delta)$. Then, for δ small enough, the following hold:*

(i) If $\gamma := \lim_{t\to 0^+} \Phi(t)/t < +\infty$ and $\alpha \leq 1/(\gamma s)$, we have the estimate

$$\beta D_W^w(x^\dagger; x_\alpha^\delta) \leq \frac{\delta}{\alpha} + \Phi(s\delta)\,. \tag{11}$$

(ii) If $\lim_{t\to 0^+} \Phi(t)/t = +\infty$, let $\Psi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the conjugate of the convex mapping $t \mapsto \Phi^{-1}(st)$. Then we have, for α sufficiently small, the estimate

$$\beta D_W^w(x^\dagger; x_\alpha^\delta) \leq \frac{\delta}{\alpha} + \Phi(s\delta) + \frac{\Psi(\alpha)}{\alpha}\,. \tag{12}$$

Proof: Since x_α^δ is a minimizer of the Tikhonov functional $\mathcal{T}_\alpha(\cdot, y^\delta)$ and

$$\mathcal{S}(F(x^\dagger), y^\delta) = \mathcal{S}(y^\dagger, y^\delta) \leq \delta\,,$$

we have the inequality

$$\mathcal{S}(F(x_\alpha^\delta), y^\delta) + \alpha\mathcal{R}(x_\alpha^\delta) \leq \mathcal{S}(F(x^\dagger), y^\delta) + \alpha\mathcal{R}(x^\dagger) \leq \delta + \alpha\mathcal{R}(x^\dagger)\,.$$

Let now $\varepsilon > 0$ and $x^\dagger \in U \subset X$ be as in Definition 3.1. Because the regularization method is convergent, it follows that x_α^δ satisfies $x_\alpha^\delta \in U$ and $|\mathcal{R}(x_\alpha^\delta) - \mathcal{R}(x_\alpha^\delta)| < \varepsilon$ for δ small enough. Therefore (10), (9), and the sub-additivity of Φ imply that

$$\begin{aligned}
\delta &\geq \mathcal{S}(F(x_\alpha^\delta), y^\delta) + \alpha(\mathcal{R}(x_\alpha^\delta) - \mathcal{R}(x^\dagger)) \\
&\geq \mathcal{S}(F(x_\alpha^\delta), y^\delta) + \alpha\beta D_W^w(x^\dagger; x_\alpha^\delta) - \alpha\Phi(\mathcal{S}(F(x_\alpha^\delta), F(x^\dagger))) \\
&\geq \mathcal{S}(F(x_\alpha^\delta), y^\delta) + \alpha\beta D_W^w(x^\dagger; x_\alpha^\delta) - \alpha\Phi(s\delta + s\mathcal{S}(F(x_\alpha^\delta), y^\delta)) \\
&\geq \mathcal{S}(F(x_\alpha^\delta), y^\delta) + \alpha\beta D_W^w(x^\dagger; x_\alpha^\delta) - \alpha\Phi(s\delta) - \alpha\Phi(s\mathcal{S}(F(x_\alpha^\delta), y^\delta))\,.
\end{aligned} \tag{13}$$

Now assume that $\gamma = \lim_{t\to 0^+} \Phi(t)/t < +\infty$. Because Φ is concave and $\Phi(0) = 0$, it follows that $\Phi(t) \le \gamma t$ for all $t \ge 0$. Consequently, we obtain from (13) the inequality

$$\alpha\beta D_W^w(x^\dagger; x_\alpha^\delta) + (1 - \alpha s\gamma)\mathcal{S}(F(x_\alpha^\delta), y^\delta) \le \delta + \alpha\Phi(s\delta) .$$

For $\alpha \le 1/(s\gamma)$, the term $(1 - \alpha s\gamma)$ is non-negative, and thus (11) holds.

Now consider the case $\lim_{t\to 0^+} \Phi(t)/t = +\infty$. From Young's inequality (see [13, Thm. 13.2]) we obtain that

$$\alpha\Phi(s\mathcal{S}(F(x_\alpha^\delta), y^\delta)) \le \Psi(\alpha) + \Psi^*(\Phi(s\mathcal{S}(F(x_\alpha^\delta), y^\delta))) .$$

Now the definition of Ψ implies that $\Psi^*(\Phi(st)) = t$ for all $t \in \mathbb{R}$. Therefore,

$$\alpha\Phi(s\mathcal{S}(F(x_\alpha^\delta), y^\delta)) \le \Psi(\alpha) + \mathcal{S}(F(x_\alpha^\delta), y^\delta) .$$

Thus we obtain from (13) that

$$\beta D_W^w(x^\dagger; x_\alpha^\delta) \le \frac{\delta}{\alpha} + \Phi(s\delta) + \frac{\Psi(\alpha)}{\alpha} ,$$

which shows (12). □

Corollary 3.4 *Let the assumptions of Theorem 3.3 be satisfied.*

(i) If $\gamma := \lim_{t\to 0^+} \Phi(t)/t < +\infty$ *we have for a constant parameter choice* $\alpha \le 1/(\gamma s)$ *the convergence rate*

$$D_W^w(x^\dagger; x_\alpha^\delta) = O(\delta) .$$

(ii) If $\lim_{t\to 0^+} \Phi(t)/t = +\infty$*, then we have for a parameter choice* $\alpha \sim \delta/\Phi(s\delta)$ *the convergence rate*

$$D_W^w(x^\dagger; x_\alpha^\delta) = O(\Phi(s\delta)) .$$

Proof: In case $\gamma = \lim_{t\to 0^+} \Phi(t)/t < +\infty$, the assertion is an immediate consequence of Theorem 3.3 and the inequality $\Phi(s\delta) \le \gamma s\delta$.

Now assume that $\lim_{t\to 0^+} \Phi(t)/t = +\infty$ and $\alpha \sim \delta/\Phi(s\delta)$. Then Theorem 3.3 implies that

$$\begin{aligned}\beta D_W^w(x^\dagger; x_\alpha^\delta) &\le \frac{\delta}{\alpha} + \Phi(s\delta) + \frac{\Psi(\alpha)}{\alpha} \\ &\sim \frac{\delta\Phi(s\delta)}{\delta} + \Phi(s\delta) + \frac{\Psi(\delta/\Phi(s\delta))}{\delta}\Phi(s\delta) .\end{aligned}$$

Thus it remains to show that $\Psi(\delta/\Phi(s\delta))/\delta$ stays bounded for $\delta \to 0$.

Because Ψ is convex, its right derivative Ψ' exists everywhere and satisfies $t\Psi'(t) \ge \Psi(t) - \Psi(0) = \Psi(t)$ for every $t > 0$. Furthermore, because $\Psi(0) = 0$ and $\Psi(t) \ge 0$ for all t, the convexity of Ψ implies that Ψ is non-decreasing. Thus we obtain that

$$\frac{\Psi(\delta/\Phi(s\delta))}{\delta} \le \frac{\Psi(s\delta/\Phi(s\delta))}{s\delta} \le \frac{\Psi'(s\delta/\Phi(s\delta))}{\Phi(s\delta)} .$$

Setting $t := \Phi(s\delta)$, we obtain that

$$\limsup_{\delta\to 0^+} \frac{\Psi(\delta/\Phi(s\delta))}{\delta} \le \limsup_{\delta\to 0^+} \frac{\Psi'(s\delta/\Phi(s\delta))}{\Phi(s\delta)} = \limsup_{t\to 0^+} \frac{\Psi'(\Phi^{-1}(t)/t)}{t} .$$

Now the convexity of Φ^{-1} and the definition of Ψ imply that

$$\Psi'(\Phi^{-1}(t)/t) \leq \Psi'((\Phi^{-1})'(t)) = t/s \, .$$

This shows that

$$\limsup_{\delta\to 0^+} \frac{\Psi(\delta/\Phi(s\delta))}{\delta} \leq 1/s \, ,$$

which proves the assertion. □

Remark 3.5 *In the case* $\lim_{t\to 0^+} \Phi(t)/t < +\infty$ *it follows from (11) that the Bregman distance between the regularized solution and* $x^\dagger$ *is zero if no noise is present. If, in addition, the Bregman distance satisfies* $D_W^w(x^\dagger; x) > 0$ *for* $x \neq x^\dagger$*, then this implies that the exact solution can be recovered from exact data using a sufficiently small but non-zero regularization parameter. For this reason, regularization methods with this property have been called* exact penalization methods *in [4]. These methods share the somewhat surprising property that the best convergence rates are achieved if the regularization parameter does not tend to zero but rather stays bounded away from zero at some small enough value. This typically happens, if the distance measure* $\mathcal{S}$ *is non-smooth across the diagonal (see for instance Examples 3.6).*

In particular, the preceding results allow the recovery of the convergence rates results derived in [14, 16, 20]:

Example 3.6 (Metric Regularization) *Assume that* Y *is a metric space with metric* d *and that*

$$\mathcal{S}(y_1, y_2) := (d(y_1, y_2))^p$$

for some $p > 1$*. Then (9) holds with* $s = 2^{p-1}$*.*

With $\Phi(t) = \gamma t^{1/p}$*, the variational inequality (10) reads as*

$$\beta D_W^w(x^\dagger; x) \leq (\mathcal{R}(x^\dagger) - \mathcal{R}(x)) + \gamma d(F(x), F(x^\dagger)) \, , \tag{14}$$

which is the metric equivalent of the condition applied in [14]. Denote by p_* *the conjugate of* p *defined by* $1/p + 1/p_* = 1$*. Then*

$$\Psi(t) = \frac{\gamma^{p_*} t^{p_*}}{2^p p_* p^{p_*}} \, .$$

Thus (12) reads as

$$\beta D_W^w(x^\dagger; x_\alpha^\delta) \leq \frac{d(y^\dagger, y^\delta)^p}{\alpha} + \gamma 2^{1/p_*} d(y^\dagger, y^\delta) + \frac{\gamma^{p_*} \alpha^{p_*-1}}{2^p p_* p^{p_*}} \, .$$

Moreover, we obtain for a parameter choice $\alpha \sim d(y^\dagger, y^\delta)^{p-1}$ *a convergence rate*

$$D_W^w(x^\dagger; x_\alpha^\delta) \leq O(d(y^\dagger, y^\delta)) \, .$$

Similarly, if $p = 1$ *and (14) holds, then we have* $\Phi(t) = \gamma t$*, implying that we are in the case of exact penalization methods. In this situation (11) implies the estimate*

$$\beta D_W^w(x^\dagger; x_\alpha^\delta) \leq (\gamma + 1/\alpha)\, d(y^\dagger, y^\delta)$$

for $\alpha \leq 1/\gamma$*.*

4. Regularization on Hilbert Spaces

Let X and Y be Hilbert spaces and let $F\colon X \to Y$ be a bounded linear operator. We consider the space W_l defined in Example 2.11, which consists of all locally negative semi-definite, continuous quadratic functions on X. Then the regularization term $\mathcal{R}$ is locally convex at $x^\dagger$ with respect to W_l and the W_l-subgradient at $x^\dagger$ non-empty, if and only if $\mathcal{R}$ has a proximal subdifferential at $x^\dagger$ (see Example 2.11).

Moreover, we consider the similarity term

$$\mathcal{S}(y,z) := \Theta(\|y-z\|_Y)\,,$$

where $\Theta\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a convex and strictly increasing function satisfying $\Theta(0) = 0$.

Let $x^\dagger \in X$ be an $\mathcal{R}$-minimizing solution of the equation $Fx = y$. If $w \in \partial_{W_l}\mathcal{R}(x^\dagger)$, then there exist $\varepsilon > 0$, $\xi \in X$, and a positive semi-definite, symmetric, bounded quadratic form A on X such that

$$w(x) = \langle \xi, x - x^\dagger\rangle - A(x-x^\dagger, x-x^\dagger)$$

for all $x \in X$ with $\|x - x^\dagger\| < \varepsilon$. Moreover, there exists a bounded linear, self-adjoint mapping $L\colon X \to X$ such that

$$A(x,\hat{x}) = \langle Lx, \hat{x}\rangle_X$$

for all x, $\hat{x} \in X$. In particular, we have the inequality

$$\mathcal{R}(x) \geq \mathcal{R}(x^\dagger) + \langle \xi, x-x^\dagger\rangle - \langle L(x-x^\dagger), x-x^\dagger\rangle \tag{15}$$

for all $x \in X$ satisfying $\|x - x^\dagger\| < \varepsilon$.

Proposition 4.1 *Assume that $\mathcal{R}$ satisfies the Radon–Riesz property. Let $\xi \in X$ and $L\colon X \to X$ satisfy (15). Assume that there exists some $\mu > 0$ such that the mapping $\mu^2 F^*F - L$ is positive semi-definite and the range condition*

$$\xi \in \operatorname{Ran}(\sqrt{\mu^2 F^*F - L})$$

holds. Then the regularization method satisfies a variational inequality at $x^\dagger$ with respect to W_l with $\Phi(t) := \gamma\Theta^{-1}(t)$ for some $\gamma > 0$. In particular, we obtain for a parameter choice according to Corollary 3.4 a convergence rate $D^w_{W_l}(x^\dagger; x^\delta_\alpha) = O(\|y^\delta - y^\dagger\|)$.

Proof: By definition we have

$$D^w_{W_l}(x^\dagger; x) = \mathcal{R}(x) - \mathcal{R}(x^\dagger) - \langle \xi, x-x^\dagger\rangle + \langle L(x-x^\dagger), x-x^\dagger\rangle\,.$$

Because $\xi \in \operatorname{Ran}(\sqrt{\mu^2F^*F - L})$, there exists a constant $C > 0$ such that

$$\begin{aligned}\langle \xi, x-x^\dagger\rangle^2 &\leq C^2\|(\mu^2F^*F-L)^{1/2}(x-x^\dagger)\|^2\\ &= C^2\mu^2\|F(x-x^\dagger)\|^2 - C^2\langle L(x-x^\dagger), x-x^\dagger\rangle\\ &= C^2\mu^2\|F(x-x^\dagger)\|^2 - C^2A(x-x^\dagger, x-x^\dagger)\,.\end{aligned}$$

Thus, for $A(x-x^\dagger, x-x^\dagger) \leq 1/C^2$, the estimate

$$|\langle \xi, x-x^\dagger\rangle| \leq 2C\mu\|F(x-x^\dagger)\| - A(x-x^\dagger, x-x^\dagger)$$

holds. Collecting the above inequalities, we obtain

$$D^w_{W_l}(x^\dagger; x) \le \mathcal{R}(x) - \mathcal{R}(x^\dagger) + 2C\mu\|F(x - x^\dagger)\|\,,$$

which proves the assertion. □

Remark 4.2 *Let again (15) be satisfied by $\xi \in X$ and $L\colon X \to X$. Then the same inequality is also satisfied by $\lambda^2 L$ for every $\lambda^2 > 1$. Now assume in addition, that the range condition $\xi \in \operatorname{Ran}(\sqrt{\mu^2 F^*F - L})$ holds. Then, obviously, also the range condition $\xi \in \operatorname{Ran}(\sqrt{\lambda^2\mu^2 F^*F - \lambda^2 L})$ is satisfied, and thus a convergence rate $D^{w_\lambda}_W(x^\dagger; x^\delta_\alpha)$ with respect to the subgradient $w_\lambda(x) = \langle \xi, x - x^\dagger\rangle - \lambda^2\langle L(x - x^\dagger), (x - x^\dagger)\rangle$ holds. Now note that $D^{w_\lambda}_W$ satisfies the inequality*

$$D^{w_\lambda}_W(x^\dagger; x) \ge (\lambda^2 - 1)\langle L(x - x^\dagger), x - x^\dagger\rangle = (\lambda^2 - 1)A(x - x^\dagger, x - x^\dagger)\,.$$

Thus the range condition of Proposition 4.1 implies at the worst a rate

$$A(x^\delta_\alpha - x^\dagger, x^\delta_\alpha - x^\dagger) = O(\|y^\delta - y^\dagger\|)\,.$$

Remark 4.3 *Note that the assumption in Proposition 4.1 that $\mathcal{R}$ satisfies the Radon–Riesz property can be dropped, if instead of W_l one uses the space W of all negative semi-definite, bounded quadratic forms on X. That is, one requires (15) to hold* globally *on X instead of merely locally around $x^\dagger$.*

Example 4.4 *Assume that $\Omega \subset \mathbb{R}^n$ is a bounded Lipschitz domain and $F\colon L^2(\Omega) \to Y$ a bounded linear operator. We consider the regularization functional*

$$\mathcal{R}(x) = \int_\Omega f(x(s))\,ds + |Dx|(\Omega)$$

where

$$f(t) := (t^2 - 1)^2$$

and $|Dx|(\Omega)$ denotes the total variation of the function x on Ω (see [1]). That is, the regularization functional encourages x to take only the values $+1$ and -1, with these two phases separated by regular hypersurfaces. Functionals of this type are, for instance, used in the theory of phase transitions [18]. Also, an application of similar functionals to image classification and denoising has been proposed in [19].

We now consider the two components of $\mathcal{R}$ separately and define $\mathcal{R}_1(x) := \int_\Omega f(x(s))\,dx$ and $\mathcal{R}_2(x) := |Dx|(\Omega)$. The function $\mathcal{R}_2$ is convex in the classical sense and its sub-differential at $x^\dagger$ consists of all functions $\xi_2 \in L^2(\Omega)$ of the form

$$\xi_2(s) = -\operatorname{div}\Bigl(\frac{Dx^\dagger(s)}{|Dx^\dagger(s)|}\Bigr)\,.$$

Moreover the proximal subdifferential $\partial_P\mathcal{R}_1(x^\dagger)$ is the function

$$\xi_1(s) = f'(x^\dagger(x)) = 4x^\dagger(s)(x^\dagger(s)^2 - 1)\,.$$

Now let

$$\eta_0(s) := \max\{2(1 - x^\dagger(s)^2), 0\}$$

and define the positive semi-definite bounded quadratic form $A_0 \colon L^2(\Omega) \to \mathbb{R}$ as

$$A_0(x_1, x_2) := \int_\Omega \eta_0(s)\, x_1(s)\, x_2(s)\, ds\ .$$

Then $\partial_W \mathcal{R}_2(x^\dagger)$ consists of all mappings w_2 of the form

$$w_2(x) = \langle \xi_1, x - x^\dagger \rangle - A(x - x^\dagger, x - x^\dagger)$$

with $A \geq A_0$.

We now consider for simplicity only sub-differentials w of $\mathcal{R}$, where the quadratic part is of the form $A(x_1, x_2) = \int_\Omega \eta(s)\, x_1(s)\, x_2(s)\, ds$ with $\eta \geq \eta_0$. Then, the corresponding self-adjoint operator L is the diagonal operator $x \mapsto \eta x$. The range condition of Proposition 4.1 therefore reads as

$$4x^\dagger({x^\dagger}^2 - 1) - \operatorname{div}\Bigl(\frac{Dx^\dagger}{|Dx^\dagger|}\Bigr) \in \operatorname{Ran}(\sqrt{\mu^2 F^* F - \eta})\ . \tag{16}$$

Moreover, the Bregman distance defined by w equals

$$\begin{aligned} D_W^w(x^\dagger; x) &= \mathcal{R}(x) - \mathcal{R}(x^\dagger) - \langle \xi_1 + \xi_2, x - x^\dagger \rangle + A(x - x^\dagger, x - x^\dagger) \\ &= \int_\Omega f(x(s)) - f(x^\dagger(s)) - f'(x^\dagger(s))(x(s) - x^\dagger(s))\, ds \\ &\quad + \int_\Omega \eta(s)(x(s) - x^\dagger(s))^2\, ds \\ &\quad + |Dx|(\Omega) + \int_\Omega \operatorname{div}\Bigl(\frac{Dx^\dagger(s)}{|Dx^\dagger(s)|}\Bigr) x(s)\, ds\ . \end{aligned}$$

5. Sparse Regularization on Sequence Spaces

We now assume that the space X is the sequence space $\ell^2(\Lambda)$, where Λ is some countable index set. We consider, for a given function $\phi \colon \mathbb{R} \to \bar{\mathbb{R}}_{\geq 0}$ and weights ω_λ, $\lambda \in \Lambda$, the regularization functional

$$\mathcal{R}(x) = \sum_{\lambda \in \Lambda} \omega_\lambda \phi(x_\lambda)\ .$$

Here we denote by x_λ the coefficients of the sequence $x = (x_\lambda)_{\lambda \in \Lambda} \in \ell^2(\Lambda)$.

We assume that the following conditions hold:

($B1$) The mapping ϕ is lower semi-continuous and $\phi(0) = 0$.

($B2$) We have $\lim_{t \to \pm\infty} \phi(t) = +\infty$.

($B3$) There exists $C > 0$ such that

$$\phi(t) \geq \frac{Ct^2}{1 + t^2}$$

for all $t \in \mathbb{R}$.

($B4$) The weights satisfy $\inf_\lambda \omega_\lambda > 0$.

It has been shown in [10] that, under these assumptions, the functional $\mathcal{R}$ is weakly lower semi-continuous and coercive (that is, its level sets are bounded and therefore weakly pre-compact), and satisfies the Radon–Riesz property.

In addition, we assume that the operator F is a bounded linear operator from $\ell^2(\Lambda)$ to the Hilbert space Y and that the distance measure equals the squared norm on Y. That is, the Tikhonov functionals reads as

$$\mathcal{T}_\alpha(x,y) = \mathcal{S}(F(x),y) + \alpha\mathcal{R}(x) = \|Fx-y\|^2 + \alpha\sum_{\lambda\in\Lambda}\omega_\lambda\phi(x_\lambda)\ .$$

In case ϕ, and therefore $\mathcal{R}$, is convex, it is well known that the variational inequality (10) holds with $\Phi(t) = \sqrt{t}$, if and only if the $\mathcal{R}$-minimizing solution $x^\dagger$ of the equation $Fx^\dagger = y$ satisfies the *range condition*

$$\operatorname{Ran} F^* \cap \partial\mathcal{R}(x^\dagger) \neq \emptyset$$

(see [20, Prop. 3.38]). In this situation, therefore, one obtains, for every $\xi \in \operatorname{Ran} F^* \cap \partial\mathcal{R}(x^\dagger)$ and a parameter choice $\alpha \sim \|y-y^\delta\|$, a convergence rate $D^\xi(x^\dagger; x_\alpha^\delta) = O(\|y-y^\delta\|)$. Moreover, if $\phi(t) = |t|^q$ with $1 < q < 2$, one can derive a convergence rate $\|x_\alpha^\delta - x^\dagger\|^2 = O(\|y-y^\delta\|)$ (see [15]).

The results in [3, 11], however, show that better rates are possible, if the true solution is finitely supported and the operator F is injective on the support of $x^\dagger$. Here, the support of $x^\dagger$ is defined as

$$\operatorname{supp}(x^\dagger) := \{\lambda\in\Lambda : x_\lambda \neq 0\}\ .$$

Then, if ϕ is convex, the range condition $\operatorname{Ran} F^* \cap \partial\mathcal{R}(x^\dagger) \neq \emptyset$ implies a rate $\|x_\alpha^\delta - x^\dagger\|^q = O(\|y-y^\delta\|)$, provided a growth condition of the form

$$\phi(t) \geq C|t|^q$$

holds. The next result shows that a slightly weaker rate holds for non-convex functions.

Definition 5.1 *Let $1 \leq p \leq 2$. We define W_p as the set of all functions $w\colon \ell^2(\Lambda)\to\mathbb{R}$ for which there exist $\varepsilon > 0$, a sparse element $x\in\ell^2(\Lambda)$, $\xi\in\ell^2(\Lambda)$, and $c > 0$ such that*

$$w(\tilde{x}) = \langle\xi, \tilde{x}-x\rangle - c\sum_{\lambda\in\operatorname{supp}(x)}|\tilde{x}_\lambda - x_\lambda|^p$$

for all $\tilde{x}\in\ell^2(\Lambda)$ with $\|x-\tilde{x}\| < \varepsilon$.

Theorem 5.2 *Let $p > q > 0$, and assume that there exists $C > 0$ such that*

$$\phi(t) \geq \frac{C|t|^q}{1+|t|^q}$$

for all $t\in\mathbb{R}$. Assume moreover that $x^\dagger$ is the unique solution of the equation $Fx = y^\dagger$, that $x^\dagger$ is sparse and that the restriction of F to $\ell^2(\operatorname{supp}(x^\dagger))$ is injective. Then the following hold:

(i) If $p = 1$, then

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^1} = O(\|y^\delta - y^\dagger\|) \qquad \text{as } \delta\to 0\ .$$

(ii) If $p > 1$ and there exists $w \in \partial_{W_p}\mathcal{R}(x^\dagger)$ such that the linear part ξ of w satisfies $\xi \in \operatorname{Ran}(F^*)$ and $\xi_\lambda = 0$ for $\lambda \notin \operatorname{supp}(x^\dagger)$, then

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^p} = O(\|y^\delta - y^\dagger\|^{1/p}) \qquad \textit{as } \delta \to 0\ .$$

Proof: The assertion for $p = 1$ has already been shown in [10]. It is therefore sufficient to only consider the case $p > 1$. We will show that there exist $\beta > 0$ and $\gamma > 0$ such that

$$\beta D^w_{W_p}(x^\dagger; x) \le (\mathcal{R}(x) - \mathcal{R}(x^\dagger)) + \gamma\|F(x - x^\dagger)\|$$

for all x sufficiently close to $x^\dagger$. Because $\mathcal{R}$ satisfies the Radon–Riesz property, this allows us to apply Corollary 3.4.

Denote now for simplicity $\Omega := \operatorname{supp}(x^\dagger)$ and define $x_\Omega \in \ell^2(\Lambda)$ by $(x_\Omega)_\lambda := x_\lambda$ if $\lambda \in \Omega$ and $(x_\Omega)_\lambda = 0$ else. Let moreover $x_\Omega^\perp = x - x_\Omega$. By definition of W_p there exist $\varepsilon > 0$ and $c > 0$ such that

$$w(x) = \langle \xi, x - x^\dagger\rangle - c\|x_\Omega - x^\dagger\|_{\ell^p}^p$$

for all x satisfying $\|x - x^\dagger\| < \varepsilon$. Choosing $\beta = 1/2$, we therefore have to show that there exists $\gamma > 0$ such that

$$\frac{c}{2}\|x_\Omega - x^\dagger\|_{\ell^p}^p - \frac{1}{2}\langle \xi, x - x^\dagger\rangle \le \frac{1}{2}(\mathcal{R}(x) - \mathcal{R}(x^\dagger)) + \gamma\|F(x - x^\dagger)\| \tag{17}$$

for all x satisfying $\|x - x^\dagger\|_{\ell^p} < \varepsilon$.

Because the restriction of F to $\ell^2(\Omega)$ is injective and Ω is a finite set, there exists $\gamma_1 > 0$ such that

$$\gamma_1\|F(x_\Omega^\perp - x^\dagger)\|^p \ge \|x_\Omega^\perp - x^\dagger\|_{\ell^p}^p$$

for all $x \in \ell^2(\Lambda)$. Consequently, there exist $\gamma_2 > 0$ and $\gamma_3 > 0$ such that

$$\begin{aligned} c\|x_\Omega - x^\dagger\|_{\ell^p}^p &\le c\gamma_1\|F(x_\Omega - x^\dagger)\|^p \\ &\le 2^{p-1}c\gamma_1\|F(x - x^\dagger)\|^p + 2^{p-1}c\gamma_1\|Fx_\Omega^\perp\|^p \\ &\le \gamma_2\|F(x - x^\dagger)\| + \gamma_3\|x_\Omega^\perp\|_{\ell^p}^p \end{aligned} \tag{18}$$

for all $x \in \ell^2(\Lambda)$ satisfying $\|x - x^\dagger\|_{\ell^p} < \varepsilon$. Now note that, after possibly choosing a smaller $\varepsilon > 0$,

$$\begin{aligned} \mathcal{R}(x) - \mathcal{R}(x^\dagger) &= \mathcal{R}(x_\Omega^\perp) + \mathcal{R}(x_\Omega) - \mathcal{R}(x^\dagger) \\ &= \sum_{\lambda \notin \Omega} \omega_\lambda \phi(x_\lambda) + \mathcal{R}(x_\Omega) - \mathcal{R}(x^\dagger) \\ &\ge \frac{C \inf_\lambda \omega_\lambda}{2}\|x_\Omega^\perp\|_{\ell^q}^q + \langle \xi, x_\Omega - x^\dagger\rangle - c\|x_\Omega - x^\dagger\|_{\ell^p}^p\ . \end{aligned} \tag{19}$$

By assumption, $\xi_\lambda = 0$ for $\lambda \notin \Omega$, implying that

$$\langle \xi, x_\Omega - x^\dagger\rangle = \langle \xi, x - x^\dagger\rangle\ .$$

Because $p > q$, it follows that, again after possibly choosing a smaller $\varepsilon > 0$,

$$\frac{C \inf_\lambda \omega_\lambda}{2}\|x_\Omega^\perp\|_{\ell^q}^q \ge 2\gamma_3\|x_\Omega^\perp\|_{\ell^p}^p\ .$$

Therefore we obtain from (19) the inequality

$$\mathcal{R}(x) - \mathcal{R}(x^\dagger) \geq 2\gamma_3 \|x_\Omega^\perp\|_{\ell^p}^p + \langle \xi, x - x^\dagger\rangle - c\|x_\Omega - x^\dagger\|_{\ell^p}^p \,.$$

Together with (18), this implies that

$$\mathcal{R}(x) - \mathcal{R}(x^\dagger) \geq \gamma_3 \|x_\Omega^\perp\|_{\ell^p}^p + \langle \xi, x - x^\dagger\rangle - \gamma_2\|F(x - x^\dagger)\| \,. \tag{20}$$

The range condition $\xi \in \operatorname{Ran}(F^*)$ implies the existence of $\gamma_4 > 0$ such that

$$|\langle \xi, x - x^\dagger\rangle| \leq \gamma_4 \|F(x - x^\dagger)\| \,.$$

From (20) we obtain therefore the estimate

$$\gamma_3 \|x_\Omega^\perp\|_{\ell^p}^p \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) - (\gamma_2 + \gamma_4)\|F(x - x^\dagger)\| \,.$$

Therefore (18) implies that

$$\begin{aligned}\frac{c}{2}\|x_\Omega^\perp - x^\dagger\|_{\ell^p}^p - \frac{1}{2}\langle \xi, x - x^\dagger\rangle &\leq \frac{\gamma_2 + \gamma_4}{2}\|F(x - x^\dagger)\| + \frac{\gamma_3}{2}\|x_\Omega^\perp\|_{\ell^p}^p \\ &\leq (\gamma_2 + \gamma_4)\|F(x - x^\dagger)\| + \frac{1}{2}(\mathcal{R}(x) - \mathcal{R}(x^\dagger)) \,.\end{aligned}$$

Setting $\gamma := \gamma_2 + \gamma_4$, this proves (17). Corollary 3.4 therefore implies the rate

$$D_{W_p}^w(x^\dagger; x_\alpha^\delta) = O(\|y^\delta - y^\dagger\|) \,.$$

Now note that the same rate can be derived with w replaced by $\tilde{w}(x) := \langle \xi, x - x^\dagger\rangle - 2c\|x_\Omega - x^\dagger\|_{\ell^p}^p$. Moreover, we have the estimate

$$\begin{aligned}D_{W_p}^{\tilde{w}}(x^\dagger; x) &\geq c\|x_\Omega - x^\dagger\|_{\ell^p}^p + C\sum_{\lambda\notin\Omega} \omega_\lambda \frac{|x_\lambda|^q}{1 + |x_\lambda|^q} \\ &\geq c\|x_\Omega - x^\dagger\|_{\ell^p}^p + \frac{C\inf_\lambda \omega_\lambda}{2}\|x_\Omega^\perp - x^\dagger\|_{\ell^q}^q\end{aligned}$$

for x sufficiently close to $x^\dagger$. Because $p > q$, this proves the rate

$$\|x_\alpha^\delta - x^\dagger\|_{\ell^p}^p = O(\|y^\delta - y^\dagger\|) \,.$$

□

6. Conclusion

In this paper we have introduced a generalized notion of Bregman distances that allows the derivation of convergence rates for Tikhonov regularization with non-convex regularization terms. The proof of the rates is based on the method of variational inequalities introduced in [14], which can be extended without modifications to the abstract setting of convexity on which the generalized Bregman distances are based. We have demonstrated by means of two examples that the generalized theory can yield relevant results.

The first example concerns the regularized solution of linear operator equations on Hilbert spaces. Assuming that the regularization functional can be approximated from below by a quadratic functional, we have shown that the corresponding variational inequality can be derived from a range condition that reduces precisely to the standard range condition $\partial\mathcal{R}(x^\dagger) \in \operatorname{Ran} F^*$ from the convex theory, if the quadratic functional degenerates to an affine one and the regularization functional becomes locally convex at $x^\dagger$.

The second example treats sparse regularization with non-convex regularization functionals. In the convex case recent results have shown that a rate $\|x_\alpha^\delta - x^\dagger\| = O(\|y^\delta - y^\dagger\|^{1/q})$ holds, if the regularization functional satisfies a growth condition $\mathcal{R}(x) \geq C\|x\|_{\ell^q}^q$ near zero and the operator F is injective on the support of $x^\dagger$. The results of this paper show that the same conditions also imply convergence rates for non-convex regularization, albeit slightly weaker ones: only a rate of order $\|x_\alpha^\delta - x^\dagger\| = O(\|y^\delta - y^\dagger\|^{1/p})$ for any $p > q$ is shown to hold in the non-convex case.

Acknowledgments

This work has been supported by the Austrian Science Fund (FWF) within the national research network *Industrial Geometry*, project 9203-N12.

References

[1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford University Press, New York, 2000.

[2] R. Boţ and B. Hofmann. An extension of the variational inequality approach for obtaining convergence rates in regularization of nonlinear ill-posed problems. Technical report, Department of Mathematics, Chemnitz University of Technology, 2010.

[3] K. Bredies and D. Lorenz. Regularization with non-convex separable constraints. *Inverse Probl.*, 25(8):085011 (14pp), 2009.

[4] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Probl.*, 20(5):1411–1421, 2004.

[5] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[6] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*, volume 178 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[7] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[8] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[9] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[10] M. Grasmair. Non-convex sparse regularisation. *J. Math. Anal. Appl.*, 365(1):19–28, 2010.

[11] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with l^q penalty term. *Inverse Problems*, 24(5):055020, 13, 2008.

[12] M. Grasmair, M. Haltmeier, and O. Scherzer. The residual method for regularizing ill-posed problems. Reports of FSP S105 - ”Photoacoustic Imaging” 14, University of Innsbruck, Austria, 2009.

[13] E. Hewitt and K. Stromberg. *Real and Abstract Analysis*. Springer Verlag, New York, 1965.

[14] 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(3):987–1010, 2007.

[15] D. Lorenz. Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. *J. Inverse Ill-Posed Probl.*, 16(5):463–478, 2008.

[16] C. Pöschl. *Tikhonov Regularization with General Residual Term.* PhD thesis, University of Innsbruck, Austria, Innsbruck, October 2008.

[17] E. Resmerita. Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Probl.*, 21(4):1303–1314, 2005.

[18] P. Rosenau. Free-energy functionals at the high-gradient limit. *Phys. Rev. A*, 41(4):2227–2230, 1990.

[19] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia. A variational model for image classification and restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(5):460–472, 2000.

[20] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences.* Springer, New York, 2009.

[21] I. Singer. *Abstract convex analysis.* Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley & Sons Inc., New York, 1997. With a foreword by A. M. Rubinov, A Wiley-Interscience Publication.