Nr. 92
17. Feb. 2023

universität innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

Convergence rates for critical point regularization

D. Obmann, M. Haltmeier

AppliedMathematics

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

arXiv:2302.08830v1 [math.OC] 17 Feb 2023

Convergence rates for critical point regularization

Daniel Obmann

Department of Mathematics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck, Austria
E-mail: daniel.obmann@uibk.ac.at

Markus Haltmeier

Department of Mathematics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck, Austria
E-mail: markus.haltmeier@uibk.ac.at

February 17, 2023

Abstract

Tikhonov regularization involves minimizing the combination of a data discrepancy term and a regularizing term, and is the standard approach for solving inverse problems. The use of non-convex regularizers, such as those defined by trained neural networks, has been shown to be effective in many cases. However, finding global minimizers in non-convex situations can be challenging, making existing theory inapplicable. A recent development in regularization theory relaxes this requirement by providing convergence based on critical points instead of strict minimizers. This paper investigates convergence rates for the regularization with critical points using Bregman distances. Furthermore, we show that when implementing near-minimization through an iterative algorithm, a finite number of iterations is sufficient without affecting convergence rates.

Keywords: Inverse problems, regularization, critical points, convergence rates, variational methods

1 Introduction

Many practical applications such as in medical imaging or remote sensing, can be represented by an equation of the form $\mathbf{A}x + z = y^\delta$, where $\mathbf{A}\colon \mathbb{X} \to \mathbb{Y}$ describes the system, z is some noise with $\|z\| \leq \delta$, and x is the signal of interest to be recovered. However, such

problems are often ill-posed, making direct inversion impossible or unstable. To construct stable solutions, variational regularization methods are often used which minimize the combination of a data-fidelity term and a regularization term. Such methods are well-established and provide stable recovery under reasonable assumptions [5,8,14,15,17]. In addition, convergence rates can be derived, giving quantitative estimates of how close the regularized solution is to the true solution [1, 4, 7, 11, 16].

However, the theory of variational regularization assumes knowledge of exact minimizers, which can be difficult to obtain in the case of non-convex regularizers. In contrast, the critical point regularization introduced in [13] discards the requirements of having access to global minimizers. Instead, it uses critical points relative to a certain tolerance function $\phi\colon \mathbb{X} \to [0, \infty)$. It has been shown that this relaxed approach leads to a stable and convergent regularization method.

In this paper, we build upon [13] and derive convergence rates for critical point regularization using the absolute symmetric Bregman distance. With the additional assumption of having access to near-minimizers, we also establish convergence rates in the Bregman distance. Furthermore, we demonstrate that having access to near-minimizers is often a reasonable assumption, for instance, when an iterative minimization algorithm is available. This gives a new perspective on variational and, particularly, convex regularization methods and shows that inexactness in the minimization process for the construction of critical points does not reduce the convergence rate.

Outline: The rest of the paper is organized as follows. Section 2 gives background and an overview of the most relevant results of [13]. Section 3 presents convergence rates in the absolute symmetric Bregman-distance for exact and inexact critical points. In Section 4 we consider critical points that are close to global minimizers of the Tikhonov functional and derive additional convergence rates. Section 5 presents a simple numerical example showing that inexact minimization may lead to significantly slower rates. The paper finishes with a short conclusion presented in Section 6.

2 Background

Throughout this paper, $\mathbb{X}$ and $\mathbb{Y}$ denote Hilbert spaces and $\mathbf{A}\colon \mathbb{X} \to \mathbb{Y}$ is a linear and bounded operator. We consider the Tikhonov functional

$$\mathcal{H}_{\alpha,y^\delta} = \frac{1}{2}\|\mathbf{A}(\cdot) - y^\delta\|^2 + \alpha\mathcal{R}\,, \tag{2.1}$$

where $\|\mathbf{A}x - y\|^2/2$ is the data-fidelity term and $\mathcal{R}$ a regularization term.

2.1 Notation

The following concepts are introduced in [13].

Definiton 2.1 (Relative sub-differentiability)**.** Let $\mathcal{R}\colon \mathbb{X} \to \mathbb{R}$ and $\phi\colon \mathbb{X} \to [0, \infty)$. Then $\xi \in \mathbb{X}$ is called ϕ-relative subgradient of $\mathcal{R}$ at $x_0 \in \mathbb{X}$ if

$$\forall x \in \mathbb{X}\colon \quad \mathcal{R}(x_0) + \langle \xi, x - x_0 \rangle \leq \mathcal{R}(x) + \phi(x)\,. \tag{2.2}$$

The set of all ϕ-relative subgradients at x_0 is denoted by $\partial_\phi \mathcal{R}(x_0)$ and called ϕ-relative sub-differential of $\mathcal{R}$. Functional $\mathcal{R}$ is called ϕ-relative subdifferentiable or relative sub-differentiable with bound ϕ if $\partial_\phi \mathcal{R}(x) \neq \emptyset$ for all $x \in \mathbb{X}$.

Definiton 2.2 (Relative critical points)**.** Let $\mathcal{R}\colon \mathbb{X} \to \mathbb{R}$ and $\phi\colon \mathbb{X} \to [0, \infty)$. We call $x_0 \in \mathbb{X}$ a ϕ-critical point of $\mathcal{R}$ or relative critical bound with bound ϕ if $0 \in \partial_\phi \mathcal{F}(x_0)$.

From the definition it follows that $x_0 \in \mathbb{X}$ is a ϕ-critical point if and only if $\mathcal{R}(x_0) \leq \mathcal{R}(x) + \phi(x)$ for all $x \in \mathbb{X}$.

Definiton 2.3 (Gradient selection)**.** Let $\mathcal{R}\colon \mathbb{X} \to \mathbb{R}$ be a relatively subdifferentiable functional. Then any function $G\colon \mathbb{X} \to \mathbb{X}$ with $G(x) \in \partial_\phi \mathcal{R}(x)$ for all $x \in \mathbb{X}$ is called gradient selection for $\mathcal{R}$.

Important examples of gradient selections include $G(x) \in \partial_0 \mathcal{R}(x)$ if $\mathcal{R}$ is convex and subdifferentiable and $G(x) = \mathcal{R}'(x)$ if $\mathcal{R}$ is differentiable and ϕ is such that $\mathcal{R}'(x) \in \partial_\phi \mathcal{R}(x)$.

Definiton 2.4 (Bregman-distance)**.** The Bregman-distance with gradient selection G of a relatively subdifferentiable $\mathcal{R}$ is defined by

$$\begin{aligned} D_G\colon \mathbb{X} \times \mathbb{X} &\to \mathbb{R}\colon \\ (x, x_0) &\mapsto \mathcal{R}(x) - \mathcal{R}(x_0) - \langle G(x_0), x - x_0 \rangle\,. \end{aligned}$$

If the Bregman distance is used with fixed x_0 and $\xi = G(x_0)$ we write $D_\xi(x, x_0) = D_G(x, x_0)$ as it only depends on the gradient selection at x_0. This is different for the symmetric Bregman distance defined next, that depends on the gradient selection at both input elements.

Definiton 2.5 (Symmetric Bregman-distance)**.** The symmetric Bregman-distance with gradient selection G of a relatively subdifferentiable $\mathcal{R}$ is defined by

$$\begin{aligned} D_G^{\mathrm{sym}}\colon \mathbb{X} \times \mathbb{X} &\to \mathbb{R}\colon \\ (x, x_0) &\mapsto \langle G(x) - G(x_0), x - x_0 \rangle\,. \end{aligned}$$

The symmetric Bregman-distance is actually symmetric and satisfies $D_G^{\mathrm{sym}}(x, x_0) = D_G(x, x_0) + D_G(x_0, x)$. If $\mathcal{R}$ is convex, then D_G is non-negative with $D_G(x, x_0) \leq D_G^{\mathrm{sym}}(x, x_0)$ which in particular shows that the symmetric Bregman-distance in this case is an upper bound for the Bregman-distance. In the non-convex case, both Bregman distances may be negative and we thus derive convergence rates for the absolute values of it.

Throughout we write $\alpha(\delta) \asymp \delta$ if $C_1\delta \leq \alpha(\delta) \leq C_2\delta$ as $\delta \to 0$ for constants $C_1, C_2 > 0$.

2.2 Critical point regularization

The rates analysis use the following assumptions on the regularization functional $\mathcal{R}$.

Condition A (Critical point regularization)**.**

(A1) $\mathcal{R}$ is weakly lower semicontinuous

(A2) $\mathcal{R}$ is ϕ-relatively subdifferentiable

(A3) $\forall \alpha \, \forall y^\delta \colon \frac{1}{2}\|\mathbf{A}(\cdot) - y^\delta\|^2 + \alpha\mathcal{R}$ is coercive.

Critical point regularization then consists in finding $(\alpha\phi)$-critical points of the Tikhonov functional $\mathcal{H}_{\alpha,y^\delta}$. In particular, any x_α^δ is such a regularized solution provided that

$$0 \in \mathbf{A}^*(\mathbf{A}x_\alpha^\delta - y^\delta) + \alpha\partial_\phi\mathcal{R}(x_\alpha^\delta)\,.$$

The analysis of [13] implies the following.

Theorem 2.6 (Critical point regularization)**.** *Let $y \in \mathrm{ran}(\mathbf{A})$, $y^\delta \in \mathbb{Y}$, $\alpha > 0$ and let Condition A be satisfied. Then the following hold*

(1) Existence: $\mathcal{H}_{\alpha,y^\delta}$ has a ϕ-critical point.

(2) Stability: Let $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ converge to y^δ and let $(x_k)_k \in \mathbb{X}^{\mathbb{N}}$ with $0 \in \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha\partial_\phi\mathcal{R}(x_k)$.

- *$(x_k)_k$ has a weakly convergent subsequence*
- *Weak cluster points of $(x_k)_k$ are $(\alpha\phi)$-critical points of $\mathcal{H}_{\alpha,y^\delta}$.*

(3) Convergence: Let $(y_k)_k \in Y^{\mathbb{N}}$ satisfy $\|y - y_k\| \leq \delta_k$, let $\delta_k, \alpha_k, \delta_k^2/\alpha_k \to 0$ and let $(x_k)_k \in \mathbb{X}^{\mathbb{N}}$ with $0 \in \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k\partial_\phi\mathcal{R}(x_k)$.

- *$(x_k)_k$ has a weakly convergent subsequence and any weak cluster point $x^\ddagger$ of $(x_k)_k$ is a solution of $\mathbf{A}x = y$ with $\mathcal{R}(x^\ddagger) \leq \inf\{\mathcal{R}(x) + \phi(x) \mid \mathbf{A}x = y\}$.*
- *If the solution $x^\ddagger$ of $\mathbf{A}x = y$ is unique, then $(x_k)_k \rightharpoonup x^\ddagger$ as $k \to \infty$.*
- *Any cluster point ξ of $\xi_k \in \partial_\phi\mathcal{R}(x_k)$ satisfies $\xi \in \ker(\mathbf{A}) \cap \partial_\phi\mathcal{R}(x^\ddagger)$.*

2.3 Example

Before we begin our discussion of convergence rates we start with a simple example which shows that without any further assumptions on the $\alpha\phi$-critical points, convergence in the value of $\mathcal{R}$ and in the Bregman-distance does not hold.

Example 2.7 (Non-convergence in $\mathcal{R}$)**.** Define the operator $\mathbf{A}\colon \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ by $\mathbf{A}x = (x_1, 0, x_3/3, \dots)$ and the regularizer $\mathcal{R} = \|\cdot\|_2^2/2$. Let $y \in \operatorname{ran}(\mathbf{A}\mathbf{A}^*)$, $\|y_k - y\| \leq \delta_k$ where $\delta_k \to 0$ and $\operatorname{supp}(y_k) \subseteq 2\mathbb{N} - 1$ and $x_k = \arg\min \mathcal{H}_k$ with $\mathcal{H}_k \coloneqq \mathcal{H}_{\alpha_k, y_k}$, $\alpha_k \asymp \delta_k$. According to standard Tikhonov regularization, $(x_k)_k \to x^\ddagger = (y_1, 0, 3y_3, \dots)$ at the convergence rate $\|x_k - x^\ddagger\|^2 = \mathcal{O}(\delta_k)$. We consider two cases.

1. For given $\varepsilon > 0$ denote $z_k = x_k + \varepsilon e_{2k}$ where $e_{2k} \in \ell^2(\mathbb{N})$ has entries 0 except for the $(2k)$-th entry where the value is 1. Then $\mathcal{H}_k(x_{\varepsilon,k}) \leq \mathcal{H}_k(x) + \alpha_k\varepsilon^2/2$ and thus $x_{\varepsilon,k}$ is an $(\alpha_k\varepsilon^2/2)$-critical point of $\mathcal{H}_k$. Further, $z_k \rightharpoonup x^\ddagger$ and $\mathcal{R}(z_k) \to \mathcal{R}(x^\ddagger) + \varepsilon^2/2$ as $k \to \infty$.
2. For $\varepsilon_k \to 0$ consider $z_k = x_k + \varepsilon_k e_{2k}$. Then $z_k \rightharpoonup x^\ddagger$ and $|\mathcal{R}(z_k) - \mathcal{R}(x^\ddagger)| \leq \varepsilon_k^2/2$. In particular, $\mathcal{R}(z_k) \to \mathcal{R}(x^\ddagger)$ at a rate determined by ε_k.

These considerations show that without further assumptions on the critical points, the Bregman distance might not converge to zero and even if it converges, the convergence can be arbitrarily slow. To address this issue, we instead either work with the symmetric Bregman-distance (see Section 3) or use a near-minimization concept (see Section 4).

3 Rates in the symmetric Bregman-distance

Throughout this section, let $\mathcal{R}\colon \mathbb{X} \to [0, \infty]$ be a possibly non-convex regularization functional that satisfies Condition A with tolerance function ϕ. Further let G be a gradient selection for $\mathcal{R}$ and let the $\alpha\phi$-critical points x_α^δ of $\mathcal{H}_{\alpha,y^\delta}$ be chosen such that $\mathbf{A}^*(\mathbf{A}x_\alpha^\delta - y) + \alpha G(x_\alpha^\delta) = 0$.

3.1 Error estimates

Convergence rates will be derived under the following assumption on $x^\ddagger$.

Condition B (Convergence rates)**.**

(B1) $G(x^\ddagger) \in \operatorname{ran}(\mathbf{A}^*)$

(B2) $\exists c \forall z \in \mathcal{M}(x^\ddagger)\colon \langle G(z), x^\ddagger - z\rangle \leq c\|\mathbf{A}(z - x^\ddagger)\|$.
Here and below $\mathcal{M}(x^\ddagger) = \{x \in \mathbb{X}\colon G(x) \in \operatorname{ran}(\mathbf{A}^*) \wedge |\mathcal{R}(x) - \mathcal{R}(x^\ddagger)| < \phi(x^\ddagger)\}$.

We have the following result.

Theorem 3.1 (Convergence rates)**.** *Let Condition B hold, let $y \in \operatorname{ran}(\mathbf{A})$, $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ satisfy $\|y_k - y\| \leq \delta_k$ where $\delta_k \to 0$ and let $\alpha_k \asymp \delta_k$. Let x_k satisfy $\mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k G(x_k) = 0$ and let $(x_k)_k$ weakly converge to $x^\ddagger$. Then the following hold*

(1) $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$

(2) $|D_G^{\mathrm{sym}}(x_k, x^\ddagger)| = \mathcal{O}(\delta_k)$.

Proof. By definition of the symmetric Bregman distance we have

$$|D_G^{\mathrm{sym}}(x_k, x^\ddagger)| = \langle G(x_k) - G(x^\ddagger), x_k - x^\ddagger\rangle + \eta_k \langle G(x^\ddagger) - G(x_k), x_k - x^\ddagger\rangle$$

for $\eta_k \in \{0, 2\}$ depending on the sign of $\langle G(x_k) - G(x^\ddagger), x_k - x^\ddagger\rangle$. By construction of the critical points and following the proof in [13] we have $x_k \in \mathcal{M}(x^\ddagger)$ for k sufficiently large. By Condition B,

- $\langle -G(x^\ddagger), x_k - x^\ddagger\rangle \leq C_1 \left(\delta_k + \|\mathbf{A}x_k - y_k\|\right)$
- $\langle G(x^\ddagger) - G(x_k), x_k - x^\ddagger\rangle \leq C_2 \left(\delta_k + \|\mathbf{A}x_k - y_k\|\right)$

for some constants $C_1, C_2 > 0$. Using the definition of x_k, the convexity of the data-fit term, equality $\mathbf{A}x^\ddagger = y$ and the estimate $\|y - y_k\| \leq \delta_k$ we get

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k \langle G(x_k), x_k - x^\ddagger\rangle \\
&= \frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), x^\ddagger - x_k\rangle \\
&\leq \frac{1}{2}\|\mathbf{A}x^\ddagger - y_k\|^2 \leq \delta_k^2/2\,.
\end{aligned}$$

The above estimates together with Young's product-inequality gives

$$\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k |D_G^{\mathrm{sym}}(x_k, x^\ddagger)| \leq \frac{1}{2}\delta_k^2 + C_3\alpha_k\delta_k + C_4\alpha_k^2 + \frac{1}{4}\|\mathbf{A}x_k - y_k\|^2\,,$$

for some constants $C_3, C_4 > 0$. With the parameter choice $\alpha_k \asymp \delta_k$ we obtain (1), (2). □

Corollary 3.2 (Convex case)**.** *If $\mathcal{R}$ is convex and $\phi = 0$, then (B1) implies (1), (2) of Theorem 3.1.*

Proof. Since $\mathcal{R}$ is convex and $G(x)$ is a regular subgradient, $\langle G(z) - G(x^\ddagger), z - x^\ddagger\rangle \geq 0$ for any $z \in \mathbb{X}$. By (B1) we get $\langle G(z), x^\ddagger - z\rangle \leq \langle G(x^\ddagger), x^\ddagger - z\rangle \leq \|w\|\|\mathbf{A}x^\ddagger - \mathbf{A}z\|$. This implies (B2) and thus (1), (2) of Theorem 3.1. □

Corollary 3.2 with $\mathcal{R}(x) = \|\mathbf{K}x\|_2^2$ for a bounded linear $\mathbf{K}$ recovers the known rate $\|\mathbf{K}(x_k - x^\ddagger)\|^2 = \mathcal{O}(\delta_k)$. Next we discuss regularizers which locally behave convexly around the exact solution.

Remark 3.3 (Locally convex case)**.** Let $y \in \mathrm{ran}(\mathbf{A})$ and let $x^\ddagger$ be a solution to $\mathbf{A}x = y$ which satisfies (B1). Assume further that $\mathcal{R}$ is locally convex at $x^\ddagger$ and that $\mathcal{R}(x^\ddagger) + \langle G(x^\ddagger), x - x^\ddagger\rangle \leq \mathcal{R}(x)$ for some $r > 0$ and all $x \in B_r(x^\ddagger)$.

(1) If $x \in B_r(x^\ddagger)$ satsfies $\mathbf{A}x = y$, then (B1) and $x - x^\ddagger \in \ker(\mathbf{A})$ imply

$$\mathcal{R}(x^\ddagger) = \mathcal{R}(x^\ddagger) + \langle G(x^\ddagger), x - x^\ddagger\rangle \leq \mathcal{R}(x)\,.$$

Thus $x^\ddagger$ is a local minimizer of $\mathcal{R}$ on the set of all solutions of $\mathbf{A}x = y$.

(2) Assume further that $x^\ddagger$ is the weak limit of a sequence $(x_k)_k$ which was constructed according to Theorem 3.1. Then $(x_k)_k$ converges to a locally $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ with rates given by Theorem 3.1. This recovers a local version of the well known convergence rates for $\mathcal{R}$-minimizing solutions [16].

(3) Finally, let $\mathcal{R}$ be locally strongly convex around $x^\ddagger$ and $\langle G(z) - G(x^\ddagger), z - x^\ddagger\rangle \geq \mu\|z - x^\ddagger\|^2$ for all $z \in B_r(x^\ddagger)$ and some $r, \mu > 0$. By Theorem 3.1 we get $\|x_k - x^\ddagger\| = \mathcal{O}(\sqrt{\delta_k})$, if $x_k \in B_r(x^\ddagger)$.

Corollary 3.4 (Finite dimensional case)**.** *Let $\mathbb{X}$ be a finite dimensional, $\mathcal{R}$ coercive and G bounded on bounded sets. Then (B2) is satisfied and the convergence rates of Theorem 3.1 hold under (B1).*

Proof. Since $\mathcal{R}$ is coercive the set $\mathcal{M}(x^\ddagger)$ is bounded. From $G = \mathbf{A}^*(\mathbf{A}^*)^\ddagger G$ where $(\mathbf{A}^*)^\ddagger$ denotes the pseudo-inverse of $\mathbf{A}^*$ we get

$$|\langle G(z), x^\ddagger - z\rangle| = |\langle \mathbf{A}^*(\mathbf{A}^*)^\ddagger G(z), x^\ddagger - z\rangle| \leq \|\mathbf{A}x^\ddagger - \mathbf{A}z\| \sup_{z\in\mathcal{M}} \|(\mathbf{A}^*)^\ddagger G(z)\|\,.$$

Since $(\mathbf{A}^*)^\ddagger$ is continuous and G is bounded on bounded sets, $\sup_{z\in\mathcal{M}}\|(\mathbf{A}^*)^\ddagger G(z)\| < \infty$ which shows (B2). ∎

Corollary 3.4 shows that (B2) is always satisfied in the case where $\mathbb{X}$ is finite dimensional. Moreover, this corollary can readily be extended to infinite dimensional $\mathbb{X}$ if $\mathbf{A}$ has closed range.

Remark 3.5 (Smooth regularizer)**.** Consider assumption (B2) for a smooth regularizer with $G(x) = \mathcal{R}'(x)$. Assume that $\ker(\mathbf{A})$ is non-empty and choose $z \in \ker(\mathbf{A})$. For $t > 0$ and $z_\pm = x^\ddagger \pm tz$ we have $\langle G(z_\pm), x^\ddagger - z_1\rangle = -t\langle \mathcal{R}'(x^\ddagger \pm tz), \pm z\rangle \leq 0$. Adding these inequalities and dividing by $-t^2$ we find that

$$\frac{1}{t}\langle \mathcal{R}'(x^\ddagger + tz) - \mathcal{R}'(x^\ddagger - tz), z\rangle \geq 0.$$

Taking $t \to 0$ yields $D^2\mathcal{R}(x^\ddagger)(x_0, x_0) \geq 0$ which shows that $\mathcal{R}$ satisfies the necessary second order convexity around $x^\ddagger$ in any direction $z \in \ker(\mathbf{A})$. Conversely, if $z \in \mathbb{X}$ is such that $\langle \mathcal{R}'(z), x^\ddagger - z\rangle > 0$, then $z - x^\ddagger$ is a first order descent direction of $\mathcal{R}$ and walking away from $x^\ddagger$ in the direction $z - x^\ddagger$ has to result in an appropriate increase in data-error.

3.2 Converse result

Next we show that the source condition is not only sufficient, but essentially also necessary for the convergence rates to hold.

Proposition 3.6 (Necessity of source condition)**.** Let $y, (y_k)_k, (x_k)_k$, $(\delta_k)_k$, $(\alpha_k)_k$, $x^\ddagger$ be as in Theorem 3.1 and assume that G is weakly continuous. Then $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$ implies the range condition (B1) and the convergence rate $D_G^{\text{sym}}(x_k, x^\ddagger) = \mathcal{O}(\delta_k)$.

Proof. From $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$ it follows that $w_k = (\mathbf{A}x_k - y_k)/\alpha_k$ is bounded. Hence it has subsequence, again denoted by $(w_k)_k$, which weakly converges to w. Because $\mathbf{A}$ is continuous, $G(x_k) = -\mathbf{A}^* w_k \rightharpoonup -\mathbf{A}^* w$. Due to the weak continuity of G and the uniqueness of the limit it follows that $G(x^\ddagger) = -\mathbf{A}^* w$. Moreover,

$$|D_G^{\text{sym}}(x_k, x^\ddagger)| = |\langle \mathbf{A}^*(w_k - w), x_k - x^\ddagger\rangle| \leq C\left(\delta_k + \|\mathbf{A}x_k - y_k\|\right)$$

implies the desired rate. □

Proposition 3.6 suggests the Morozovs discrepancy principle (see [2, 3] and references therein) for selecting the regularization parameter, since in this case the condition $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$ is satisfied by definition. While this is an interesting line of future research such an analysis is beyond the scope of this paper.

3.3 Inexact critical points

In the following we consider inexact critical points where $\|\mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k G(x_k)\|$ is sufficiently small. This case has also been analyzed in [13] where it has been shown to yield to a convergent regularization. The following theorem provides rates in this case.

Theorem 3.7 (Inexact rates)**.** *Let Condition B hold, $y \in \text{ran}(\mathbf{A})$, $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$, $\|y_k - y\| \leq \delta_k \to 0$ and $\alpha_k \asymp \delta_k$. Assume that $(x_k)_k$ is such that for $z_k = \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k G(x_k)$ we have $\|z_k\| \leq \alpha_k \eta_k$ for some $\eta_k \to 0$ and $\langle z_k, x_k\rangle \leq 0$. Denote by $x^\ddagger$ the weak limit of $(x_k)_k$. Then the following hold*

(1) $\|\mathbf{A}x_k - y_k\| = \mathcal{O}\Big(\sqrt{\delta_k^2 + \delta_k \eta_k}\Big)$

(2) $|D_G^{\mathrm{sym}}(x_k, x^\ddagger)| = \mathcal{O}(\delta_k + \eta_k)$.

Proof. The proof is similar to the one of Theorem 3.1 and we only show main changes. By construction

$$\begin{aligned}
&\alpha_k \langle G(x_k), x_k - x^\ddagger \rangle \\
&\quad = \langle z_k, x_k - x^\ddagger \rangle + \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), x^\ddagger - x_k \rangle \\
&\quad \leq \|z_k\| \|x^\ddagger\| + \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), x^\ddagger - x_k \rangle \\
&\quad \leq C\delta_k \eta_k + \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), x^\ddagger - x_k \rangle \,.
\end{aligned}$$

Hence,

$$\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k \langle G(x_k), x_k - x^\ddagger \rangle \leq \frac{\delta_k^2}{2} + C\delta_k \eta_k \,.$$

Following the proof of Theorem 3.1 yields (1), (2). □

Theorem 3.7 provides convergence rates dependent on η_k as a measure for the exactness of the critical points. To recover the rates of Theorem 3.1 the choice $\eta_k = \delta_k$ is appropriate. While Theorem 3.7 requires $\eta_k \to 0$ the same proof can be given for a bounded sequence $(\eta_k)_k$. In this case the absolute symmetric Bregman-distance might not converge and the convergence in the discrepancy is only $\mathcal{O}\big(\sqrt{\delta_k}\big)$.

Remark 3.8 (Iterative minimization)**.** Suppose that the critical points are approximated using an iterative descent algorithm and write $z_k = \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k G(x_k)$ where x_k is the approximate critical point. In this context the conditions of Theorem 3.7 on z_k are quite natural. First, $\langle z_k, x_k \rangle < 0$ is a descent condition for a descent direction. Second, the condition $\|z_k\| \leq \alpha_k \eta_k$ is simply a common stopping criterion for the iteration dictated by the noise-level and the desired accuracy.

Note that the necessity of the source condition for inexact critical points can be derived as in Proposition 3.6. In Section 5 we will provide an example showing that the inexact choice can indeed result in a significantly slower convergence rate.

4 Rates for near minimizers

In the previous section we have derived convergence rates for exact and inexact critical points which can be very different from global minimizers. In this section we derive convergence rates for near-minimizers in the absolute Bregman-distance extending results of [10].

Throughout this section let $y \in \mathrm{ran}(\mathbf{A})$ and $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ be a sequence of noisy data with $\|y_k - y\| \leq \delta_k$. Further assume $\delta_k \to 0$ and $\alpha_k \asymp \delta_k$. Let x_k be an $(\alpha_k \phi)$-critical point of

the Tikhonov functional $\mathcal{H}_k = \frac{1}{2}\|\mathbf{A}(\cdot) - y_k\|^2 + \alpha_k \mathcal{R}$ where the regularizer $\mathcal{R}\colon \mathbb{X} \to [0, \infty)$ satisfies Condition A. Further, let $x^\ddagger$ be the weak limit $(x_k)_{k\in\mathbb{N}}$.

4.1 Error estimates

We call $\Delta_k(x) = \mathcal{H}_k(x_k) - \mathcal{H}_k(x)$ the Tikhonov gap between x_k and $x \in \mathbb{X}$. Convergence rates will then be derived under the following assumption.

Condition C (Rates using Tikhonov gap)**.**

(C1) $\exists \xi \in \partial_\phi \mathcal{R}(x^\ddagger) \cap \operatorname{ran}(\mathbf{A}^*)$.

(C2) $\exists c \forall z \in \mathcal{B}(x^\ddagger)\colon \mathcal{R}(x^\ddagger) - \mathcal{R}(z) \le c\|\mathbf{A}z - \mathbf{A}x^\ddagger\|$. Here $\mathcal{B}(x^\ddagger) \coloneqq \{x \in \mathbb{X}\colon |\mathcal{R}(x^\ddagger) - \mathcal{R}(x)| \le \varepsilon\}$ where $\varepsilon > \max\{0, \sup_k \Delta_k/\alpha_k\}$.

We have the following result.

Theorem 4.1 (Rates using Tikhonov gap)**.** *Let $y, y_k, \delta_k, \alpha_k$ and $x^\ddagger$ be as introduced above and let Assumption C be satisfied. Then*

(1) $\|\mathbf{A}x_k - y_k\| = \mathcal{O}\big(\sqrt{\delta_k^2 + \Delta_k}\big)$

(2) $|D_\xi(x_k, x^\ddagger)| = \mathcal{O}(\delta_k + \Delta_k/\delta_k)$.

Proof. By definition of D_ξ,

$$|D_\xi(x_k, x^\ddagger)| = \mathcal{R}(x_k) - \mathcal{R}(x^\ddagger) - \langle r, x_k - x^\ddagger\rangle + \eta_k \left(\mathcal{R}(x^\ddagger) - \mathcal{R}(x_k) + \langle r, x_k - x^\ddagger\rangle\right),$$

with $\eta_k \in \{0, 2\}$. By (C1), (C2)

$$\mathcal{R}(x^\ddagger) - \mathcal{R}(x_k) + \langle r, x_k - x^\ddagger\rangle \le C_1\|\mathbf{A}x_k - \mathbf{A}x^\ddagger\| \le C_1\left(\|\mathbf{A}x_k - y_k\| + \delta_k\right),$$

By definition of the Tikhonov gap Δ_k,

$$\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k\mathcal{R}(x_k) - \alpha_k\mathcal{R}(x^\ddagger) \le \frac{1}{2}\|\mathbf{A}x^\ddagger - y_k\|^2 + \Delta_k \le \delta_k^2/2 + \Delta_k.$$

Using once again (C1) it follows that

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k|D_\xi(x_k, x^\ddagger)| \\
&\le \Delta_k + \delta_k^2/2 + \tilde{C}\alpha_k\left(\|\mathbf{A}x_k - y_k\| + \delta_k\right) \\
&\le \Delta_k + \delta_k^2/2 + C_2(\alpha_k\delta_k + \alpha_k^2) + \|\mathbf{A}x_k - y_k\|^2/4\,,
\end{aligned}$$

where the last inequality follows again by Young's product-inequality. With $\alpha_k \asymp \delta_k$ we get (1), (2). □

Clearly, if $\Delta_k \leq \delta_k^2$ we obtain the classical convergence rates. Note, however, that Δ_k depends on $x^\ddagger$ and hence controlling Δ_k is challenging. We next discuss a special case where such an assumption is achievable.

Assume that we can construct x_k as a near-minimizer,

$$\mathcal{H}_k(x_k) \leq \inf_z \mathcal{H}_k(z) + \alpha_k \eta_k \,. \tag{4.1}$$

Since $(\eta_k)_k$ is bounded, [13] shows that $(x_k)_k$ has a weakly convergent subsequence. After restriction to such a convergent subsequence and denoting its limit by $x^\ddagger$ we get $\varepsilon_k(x^\ddagger) \leq \alpha_k \eta_k$. With $\alpha_k \asymp \delta_k$, according to Theorem 4.1 we get the rates $\|\mathbf{A}x_k - y_k\|^2 = \mathcal{O}(\delta_k^2 + \delta_k \eta_k)$ and $|D_\xi(x_k, x^\ddagger)| = \mathcal{O}(\delta_k + \eta_k)$.

4.2 Iterative minimization

In the following we assume that near minimization (4.1) of $\mathcal{H}_k$ is realized with some iterative algorithm.

Corollary 4.2 (Iterative minimization)**.** *Let $\mathcal{A}_k$ be an iterative algorithm for minimizing $\mathcal{H}_k$ such that for the n-th iterate $x_{k,n}$ we have $\mathcal{H}_k(x_{k,n}) \leq \inf \mathcal{H}_k + f_k(n)$ with $f_k(n) \to 0$ as $n \to \infty$. Let $x_{k,n(k)}$ satisfy* (4.1)*. Then*

(1) $\|\mathbf{A}x_{k,n(k)} - y_k\| = \mathcal{O}\Big(\sqrt{\delta_k^2 + \delta_k \eta_k}\Big)$

(2) $D_G(x_{k,n(k)}, x^\ddagger) = \mathcal{O}(\delta_k + \eta_k)$.

If $\eta_k \to 0$ then $(x_{k,n(k)})_k$ converges to an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ in the Bregman-distance.

Proof. The rates follow from Theorem 4.1 and because $x_{k,n(k)}$ is an $\alpha_k \eta_k$-minimizer. That $x^\ddagger$ is an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ is shown similar to [16]. □

An important feature of Corollary 4.2 is that the points $x_{k,n(k)}$ can be obtained in a finite number of steps of the algorithm $\mathcal{A}_k$ and still result in convergence rates of order $\mathcal{O}(\delta_k)$. Opposed to this, classical theory needs access to global minimizers which usually requires an infinite number of steps.

Remark 4.3 (Convex regularizers)**.** The assumption of having access to an algorithm $\mathcal{A}_k$ is applicable if $\mathcal{R}$ is convex. In this case $\mathcal{H}_k$ is convex and algorithms such as subgradient descent, heavy ball methods or accelerated gradient methods guarantee

convergence [6, 12]. For such algorithms the number of iterations in dependence of δ_k can be stated. Assume $\eta_k = \delta_k$ and that the algorithm $\mathcal{A}_k$ is convergnt with a rate of $f(n) = Cn^{-\beta}$ for some $\beta > 0$. According to Corollary 4.2, we need to perform on the order of $(\alpha_k \delta_k)^{-1/\beta}$ to guarantee convergence to an $\mathcal{R}$-minimizing solution. With the choice $\alpha_k \asymp \delta_k$ this yields that we need to perform on the order of $\delta_k^{-2/\beta}$ iterations. For example, if $\beta = 1$, e.g. in the case of the heavy ball method [6], we need on the order of δ_k^{-2} number of iterations which is comparable to the number of iterations for the Landweber iteration [5].

Theorem 4.1 suggests that slower rates might occur if $\mathcal{H}_k$ is not sufficiently small. Indeed, it is easy to construct examples where this is the case and convergence in the Bregman-distance is considerably slower than the rate $\mathcal{O}(\delta)$; compare Example 2.7. As a consequence, this means that depending on the algorithm $\mathcal{A}_k$, its initialization and the choice of hyper-parameters that in some cases the required number of iterations necessary is strict.

Remark 4.4 (Comparison of inexactness results)**.** Finally, we briefly compare the results of Theorem 3.7 and Theorem 4.1. Both Theorems deal with the case of inexactness in the construction of the regularized solution, but they rely on different measures for inexactness. As such these Theorems might be more suitable in different situations. Assume that in both cases regularized solutions are constructed using an iterative algorithm. The main advantage of Theorem 3.7 is that the condition $\|z_k\| \leq \alpha_k \eta_k$ for some user-defined $\eta_k > 0$ is easily checkable in an online manner during the the algorithm itself and no prior information other than the tolerance η_k is necessary. However, it may not be known a-priori how long this might take and the number of iterations could significantly increase depending on η_k. Opposed to this, Corollary 4.2 gives an estimate of the number of iterations necessary.

5 Numerical example

We perform a simple test to numerically check the convergence rates derived in the previous sections in the context of iterative minimization.

5.1 Setting

As inverse problem we adapt the "Depth Profiling and Depth Resolution" as presented in [9, Section 7.8]. We take $\mathbb{X} = \mathbb{Y} = L^2(0, \pi/2)$ and the linear operator $\mathbf{A}\colon \mathbb{X} \to \mathbb{Y}$ defined by

$$(\mathbf{A}x)(s) = \int_0^{\arcsin(\cos(s))} \exp(-\sin(\tau)) \cos(\tau) x(\tau) \mathrm{d}\tau,$$

δ	δ^β	$x_0 = 1$	$x_0 = 0$
10^{-2}	$6 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$1 \cdot 10^{-2}$
10^{-4}	$4 \cdot 10^{-1}$	$6 \cdot 10^{-2}$	$2 \cdot 10^{-4}$
10^{-6}	$3 \cdot 10^{-1}$	$2 \cdot 10^{-2}$	$3 \cdot 10^{-6}$

Table 5.1: Error $\|x_k - x^\ddagger\|^2$ in for different noise levels using gradient descent for two different initial values.

for $x \in \mathbb{X}$ and $s \in (0, \pi/2)$.

We consider the quadratic regularizer $\mathcal{R} = 1/2\|\cdot\|^2$ in which case according to Corollary 3.2, Condition B holds true whenever the source condition (B1) is satisfied. We construct near minimizers using gradient descent. As theoretical framework for the convergence rates we use Theorem 3.7 for which conditions on gradients can be checked during iteration; compare Remark 4.4. The source condition is satisfied, whenever $x^\ddagger \in \operatorname{ran}(\mathbf{A}^*)$.

5.2 Implementation details

For the presented results we choose the true signal $x^\ddagger = \mathbf{A}^* w$ with $w(s) = \cos(10s) + \sin(5s^2)$ which satisfies the source condition by definition.

We simulate noisy data y^δ by adding white noise for different noise levels $\delta \in \{10^{-k} \colon k = 2, \dots, 7\}$ to $y = \mathbf{A}x^\ddagger$. We choose the regularization parameter $\alpha_k = \delta_k$ and consider the Tikhonov functional $\mathcal{H}_k = \|\mathbf{A}(\cdot) - y_k\|^2/2 + \alpha_k\|\cdot\|^2/2$. We choose the tolerance level $\eta_k = \alpha_k^\beta$ with $\beta = 0.1$. Hence we stop the gradient descent iteration once we have $\|\nabla \mathcal{H}_k(x_k)\| \leq \delta_k^{1+\beta}$. It should be noted, that since we stop gradient descent before convergence, the resulting x_k depends on the initial value x_0 and hence we test for different choices of x_0 namely the constant 0 and constant 1 functions. The code for the numerical simulations is publicly available at `https://git.uibk.ac.at/c7021101/cpr-rates`.

5.3 Results

Numerical results are shown in Table 5.1 where the difference $\|x_k - x^\ddagger\|^2$ is given in dependence of δ for the two different initial values x_0. One notices that the convergence rate obtained with $x_0 = 0$ is way better than the one given in Theorem 3.7 and is closer to δ than δ^β. The estimated rate is around $\beta_{\text{est}} = 0.99$. On the other hand, the convergence rate for the initial value $x_0 = 1$ is closer to δ^β. The estimated rate in this case is $\beta_{\text{est}} = 0.22$. This shows that depending on the input-parameters of the algorithm the convergence rate obtained can significantly differ. This also indicates that without further assumptions better rates than the one in Theorem 3.7 can be expected.

6 Conclusion

In this paper, we have presented convergence rates in the absolute symmetric Bregman distance for the regularization of critical points under a classical source condition and an assumption on the nonconvexity of the regularizer $\mathcal{R}$. This result has been generalized to inexact critical points, where the inexactness is measured in the magnitude of the gradient of the Tikhonov functional. Making the additional assumption that almost-minimizers can be achieved, we derived convergence rates in the absolute Bregman distance. A direct consequence is that, in contrast to the classical theory, access to global minimizers is not necessary for regularization, while known rates of $\mathcal{O}(\delta)$ are preserved in the absolute Bregman distance. We have also shown that near-minimizers on the order of $\delta^{-2/\beta}$ iterations can be achieved using an iterative algorithm with rate $n^{-\beta}$.

We finally presented numerical simulations showing that non-exactness of the critical points can indeed lead to different convergence rates depending on the input parameters of the algorithm. Corollary 3.6 suggests Morozov's discrepancy principle for choosing the regularization parameters, and establishing conditions for when this leads to a convergent regularization method is an interesting line of future research. Other directions of future work could focus more on the practical aspect of minimization and to derive conditions under which rates can be improved under an inexactness assumption.

References

[1] V. Albani, P. Elbau, M. V. de Hoop, and O. Scherzer. Optimal convergence rates results for linear inverse problems in hilbert spaces. *Numerical functional analysis and optimization*, 37(5):521–540, 2016.

[2] S. Anzengruber and R. Ramlau. Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26(2):025001, 2009.

[3] T. Bonesky. Morozov's discrepancy principle and Tikhonov-type functionals. *Inverse Problems*, 25(1):015015, 2008.

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

[5] 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.

[6] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[7] M. Grasmair. Generalized bregman distances and convergence rates for non-convex regularization methods. *Inverse Problems*, 26(11):115014, 2010.

[8] C. Groetsch. The theory of Tikhonov regularization for fredholm equations. *104p, Boston Pitman Publication*, 1984.

[9] P. C. Hansen. *Discrete inverse problems: insight and algorithms*. SIAM, 2010.

[10] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: Solving inverse problems with deep neural networks. *Inverse Probl.*, 36(6):065005, 2020.

[11] D. Lorenz. Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. *Journal of Inverse & Ill-Posed Problems*, 16(5), 2009.

[12] Y. Nesterov. A method of solving a convex programming problem with convergence rate o(1/ k^2). *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[13] D. Obmann and M. Haltmeier. Convergence analysis of critical point regularization with non-convex regularizers, 2022. arXiv preprint.

[14] R. Ramlau and G. Teschke. A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. *Numerische Mathematik*, 104:177–203, 2006.

[15] A. Rieder. A wavelet multilevel method for ill-posed problems stabilized by Tikhonov regularization. *Numerische Mathematik*, 75:501–522, 1997.

[16] 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.

[17] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady akademii nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.