SIAM J. OPTIM.
Vol. 31, No. 1, pp. 754–784

ACCELERATED ITERATIVE REGULARIZATION VIA DUAL DIAGONAL DESCENT*

LUCA CALATRONI[†], GUILLAUME GARRIGOS[‡], LORENZO ROSASCO[§], AND SILVIA VILLA[§]

Abstract. We propose and analyze an accelerated iterative dual diagonal descent algorithm for the solution of linear inverse problems with strongly convex regularization and general data-fit functions. We develop an inertial approach of which we analyze both convergence and stability properties. Using tools from inexact proximal calculus, we prove early stopping results with optimal convergence rates for additive data terms and further consider more general cases, such as the Kullback–Leibler divergence, for which different type of proximal point approximations hold.

Key words. iterative regularization, duality, acceleration, forward-backward splitting, diagonal methods, stability and convergence analysis

AMS subject classifications. 90C25, 49N45, 49N15, 68U10, 90C06

DOI. 10.1137/19M1308888

1. Introduction.

We are interested in solving the linear inverse problem:

$$\text{find} \quad \bar{x} \in \mathcal{X} \quad \text{s.t.} \quad A\bar{x} = \bar{y}, \tag{1.1}$$

where $A : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between two Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and $\bar{y} \in \mathcal{Y}$ is a given measurement of some unknown quantity $\bar{x} \in \mathcal{X}$ we want to recover. In general, the inverse problem (1.1) is ill-posed as its solution (if it exists) may lack some fundamental properties like uniqueness or stability. A standard modeling hypothesis in inverse problems [42, 27] is assuming that the desired $\bar{x}$ is well-approximated by $x^\dagger \in \mathcal{X}$ solving

$$\text{find } x^\dagger \in \operatorname{argmin} \left\{ R(x) \text{ s.t. } x \in \underset{x' \in \mathcal{X}}{\operatorname{argmin}}\, \ell(Ax'; \bar{y}) \right\}. \tag{$P_0(\bar{y})$}$$

Here, R is a regularization function enforcing a priori knowledge on the desired solution $\bar{x}$, while $\ell : \mathcal{Y}^2 \to \mathbb{R} \cup \{+\infty\}$ is a data-fit function. In practical situations, the data is subject to noise due to, e.g., possible transmission and/or acquisition problems. As a consequence, only an inexact version $\hat{y}$ of $\bar{y}$ is accessible. Replacing $\hat{y}$ in $(P_0(\bar{y}))$ no longer provides a suitable solution of problem (1.1), hence a regularization method is needed. Regularization methods can be seen as a way to explore the space of solutions $\mathcal{X}$ to find a good approximation of $x^\dagger$ in the presence of noise. More precisely, they have the following characteristics:

*Received by the editors December 27, 2019; accepted for publication (in revised form) October 29, 2020; published electronically March 1, 2021.
https://doi.org/10.1137/19M1308888

Funding: The first and second authors received financial support from the PEPS grant Efficient Iterative Regularization Methods for Inverse Problems and Machine Learning, awarded by the INSMI, CNRS. The first, third, and fourth authors were supported by the European Union's Horizon 2020 Research and Innovation Programme NoMADS, grant agreement 777826.

[†] Université Côte d'Azur, CNRS, Inria, I3S, 06903 Sophia-Antipolis, France (calatroni@i3s.unice.fr).

[‡] LPSM, Université de Paris, Sorbonne Université, CNRS, 75205 Paris, France (garrigos@lpsm.paris).

[§] MaLGa, DIBRIS, University of Genoa, 16126 Genoa GE, Italy (lorenzo.rosasco@unige.it, silvia.villa@unige.it).

1. Given any data $y \in \mathcal{Y}$, the method generates a *regularization path* $\{x_{\mathsf{p}}(y)\}_{\mathsf{p}\in\mathsf{P}}$ where $\mathsf{P} \subset \mathbb{R}$ is a set of *regularization parameters.*
2. Given the true data $\bar{y}$, there exists an accumulation point p_0 of P such that the regularization path converge to the ideal solution $x^\dagger$ of $(P_0(\bar{y}))$, i.e., $\lim_{\mathsf{p}\to\mathsf{p}_0} x_{\mathsf{p}}(\bar{y}) = x^\dagger$.
3. For any given noise level $\delta > 0$ and noisy data $\hat{y}$ such that $\|\bar{y} - \hat{y}\| \leqslant \delta$, there exists a regularization parameter $\mathsf{p}(\delta) \in \mathsf{P}$ such that

$$\|x_{\mathsf{p}(\delta)}(\hat{y}) - x^\dagger\| = O(\delta^\alpha) \quad \text{for some } \alpha > 0. \tag{1.2}$$

The quantity $O(\delta^\alpha)$ is often called the *convergence rate* of the considered regularization method, and the exponent α quantifies its efficiency: the larger α is, the closer the regularized solution $x_{\mathsf{p}(\delta)}(\hat{y})$ will be to the desired $x^\dagger$ and hence less affected by noise. Convergence and rates depend on the chosen regularization method and the properties of the considered problem. We briefly review in the following two well-known families of regularization methods.

Tikhonov regularization. This is the most classical regularization approach, which, for a given $\lambda > 0$, relies on the following family of penalized optimization problems:

$$(P_\lambda(\hat{y})) \qquad \text{find } \hat{x}_\lambda \in \underset{x\in\mathcal{X}}{\operatorname{argmin}} \left\{ p_\lambda(x) := R(x) + \frac{1}{\lambda}\ell(Ax; \hat{y}) \right\}.$$

Intuitively, the so-called regularization parameter λ balances the trust in the data $\hat{y}$ with the regularization enforced by R. In other words, it parametrizes a regularization path $\{x_\lambda(\hat{y})\}_{\lambda>0}$ along which we look for a good approximation of $x^\dagger$ (see Figure 1 for an illustrative example). In practice, this requires two steps. First, problem $(P_\lambda(\hat{y}))$ needs to be solved for various choices of λ by means of a suitable optimization algorithm (see, e.g., [38]). Second, all the computed solutions are compared using some validation criterion (e.g., discrepancy principles [42], SURE [56, 40], cross-validation [57], and many more) and an optimal parameter λ^* is computed along with the corresponding solution $\hat{x}_{\lambda^*}$.

There are a number of related regularization methods based on variational problems. For instance, one can replace $(P_\lambda(\hat{y}))$ with a constrained formulation, such as min $R(x)$ subject to $\ell(Ax; y) \leqslant \sigma$, for a given error level $\sigma \geqslant 0$, which can be solved by appropriate optimization methods; see, for instance, [28, 4]. Next, we discuss a class of regularization methods based on quite different ideas.

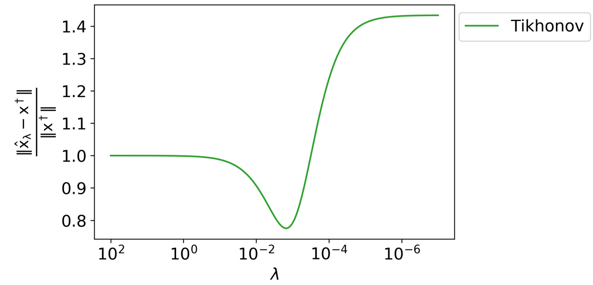

FIG. 1. *Tikhonov regularization path on a simple problem. After computing the solution $\hat{x}_\lambda := x_\lambda(\hat{y})$ of the problem $(P_\lambda(\hat{y}))$ for several values of λ, the best parameter λ is selected. In this example, $\lambda \simeq 10^{-3}$ minimizes $\|\hat{x}_\lambda - x^\dagger\|$.*

Iterative regularization. The choice of the optimal parameter λ in a Tikhonov regularization approach is in general very costly computationally. The family of so-called iterative regularization methods provides an accurate and more efficient alternative approach [42, 21, 45]. Iterative regularization methods are regularization methods for which the regularization path $\{x_k(\hat{y})\}_k$ is parametrized by the iterate index k of algorithms which can easily compute the iterates in terms of R, ℓ, and A. These algorithms are usually designed to iteratively solve $(P_0(\bar{y}))$ in a stable way with respect to errors on $\bar{y}$. Using these methods, it is therefore possible to find an approximation of $x^\dagger$ given noisy data $\hat{y}$ by "stopping" the algorithm when close to $x^\dagger$ [19, 32, 33, 29] (see Figure 2). In these methods, the number of iterations plays the role of a regularization parameter, controlling at the same time the accuracy of the solution and the computational cost. In practice, the selection of this regularization parameter is made using the similar validation criterion as the ones described for Tikhonov regularization.

Previous results. For quadratic data-fit terms ℓ and square-norm regularization R, both Tikhonov and iterative regularization approaches (such as the Landweber algorithm) have been shown to be *optimal*, in the sense that their reconstruction error in (1.2) has optimal rate $O(\delta^{\frac{1}{2}})$ [42]. Optimal results with possibly fewer iterations have also been obtained by considering accelerated approaches [42, 52]. For quadratic data-fit terms and general strongly convex regularizers, an iterative regularization procedure combined with a Morozov-type discrepancy principle was also shown to be optimal in [33], and accelerated approaches based on a dual accelerated gradient descent were shown to be optimal with fewer iterations in [49]. Iterative regularization methods also have been studied in the case of general convex regularizers in [32], where estimates in terms of the Bregman distance were proved (see also [33, 19] for Tikhonov-type approaches), but no explicit rates in the form (1.2) were shown. More general iterative algorithms defined in Banach spaces have been studied in [46, 47, 31] for linear and nonlinear inverse problems and in [29] for L^1 and total variation regularization. For data-fit terms different from the squared norm, the literature is more scarce. In the context of iterative regularization methods, we mention [26] for results in the framework of Bregman distances and [43], where the Dual Diagonal Descent (3D) algorithm is considered. Here, the authors provide convergence rates for general data-fit terms, but the latter is suboptimal in the quadratic case.

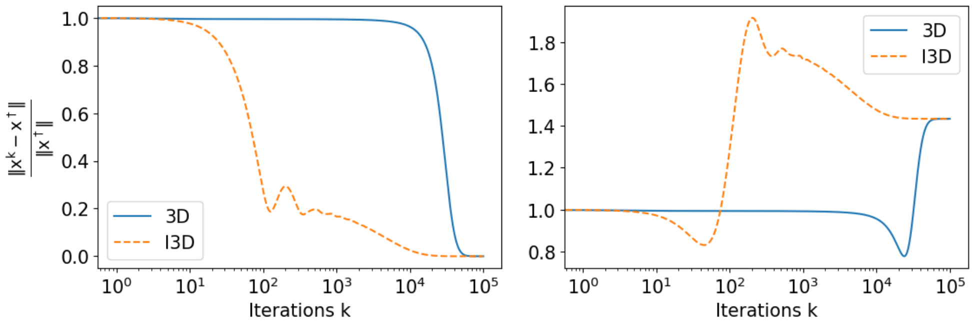

FIG. 2. *Illustration of two iterative regularization methods (Dual Diagonal Descent* (3D) [43] *and its Inertial variant* (I3D) *proposed in this work) on a simple problem. Left: given the true data $\bar{y}$, the iterates converge to the ideal solution $x^\dagger$. Right: given noisy data $\hat{y}$, the iterates $x_k(\hat{y}) =: \hat{x}_k$ approach $x^\dagger$ before tending away from it. Regularization holds by early stopping the algorithms at a suitably chosen iterate k^*. In this example, $k^* \simeq 2 \times 10^4$ (resp., 50) minimizes $\|\hat{x}_k - x^\dagger\|$ for* (3D) *(resp.,* (I3D) *).*

Contribution and organization of the paper. In this paper, we study a novel accelerated iterative regularization algorithm with strongly convex regularization and general data-fit terms. To the best of our knowledge, *accelerated* iterative regularization approaches have not been studied in this general setting. Our Inertial Dual Diagonal Descent algorithm, dubbed (I3D), extends the (3D) iterative algorithm studied in [43] by introducing an inertial term which yields acceleration.

Our main contribution is the analysis of convergence rates for this method. We show that these rates depend on how the noise interacts with the data-fit term considered. By introducing acceleration, we prove that the same or better convergence rates than those of (3D) can be achieved with much *fewer* iterations (see Figure 2 for an illustration). This extends similar observations previously made in the quadratic case in, e.g., [52, 49]. For the latter case, in particular, we obtain the optimal rate $O(\delta^{\frac{1}{2}})$. In addition, we show that this rate holds more generally for *every* additive data-fit, including, for instance, the ℓ^1 data term. From an optimization perspective, the rationale behind this fact is that inertial dynamics are able to exploit information in previous iterates to converge faster to an optimal solution. However, as pointed out in [41], inertial methods suffer from error accumulation that need to be controlled along the iterations and balanced with the improvement observed in the convergence speed, which makes their analysis in an inverse problem framework nontrivial.

The paper is organized as follows. In section 2 we introduce the notation and the main assumptions. In section 3 we introduce and analyze the inertial continuous dynamical system corresponding to (I3D) and, in particular, study its asymptotic behavior in Theorem 3.3. In section 4, we derive the algorithm (I3D) as a discretization in time of the continuous dynamics. We study its convergence properties in Theorem 4.6, showing fast convergence of the iterates to $x^\dagger$ in the noiseless case. In section 5 we study the stability properties of (I3D) in the presence of errors due to noise, proving a general abstract stability result in Theorem 5.5. We specialize this result in Theorems 5.6, 5.7, and 5.8 showing how convergence rates change depending on which type of error is assumed. Finally, in section 6, we provide explicit convergence rates for data-fit terms used in practice, including the Kullback–Leibler (KL) divergence.

2. Main assumptions and background on diagonal methods.

We begin fixing the notation. Let $\mathcal{H}$ be a Hilbert space with scalar product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|$. Given $y \in \mathcal{H}$ and $\varrho \in \mathbb{R}_+$, let $\mathbb{B}(y,\varrho)$ be the open ball of center y and radius ϱ. We denote by $\Gamma_0(\mathcal{H})$ the set of proper, convex, and lower semicontinuous functions from $\mathcal{H}$ to $]-\infty,+\infty]$. We say that $f \in \Gamma_0(\mathcal{H})$ is *σ-strongly convex* if $f-\sigma\|\cdot\|^2/2 \in \Gamma_0(\mathcal{H})$, with $\sigma \in]0,+\infty[$. We recall that the *subdifferential* of $f \in \Gamma_0(\mathcal{H})$ is the multivalued operator $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$ defined by

$$(2.1)\qquad (\forall x \in \mathcal{H}) \quad \partial f(x) := \{u \in \mathcal{H} : f(x') - f(x) - \langle u, x'-x\rangle \geqslant 0 \ \forall\ x' \in \mathcal{H}\}.$$

If f is Gateaux differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$; see, e.g., [22, Proposition 17.31(i)]. For all $x \in \mathcal{H}$ and $\tau > 0$, we also recall the definition of the proximity operator $\operatorname{prox}_{\tau f} : \mathcal{H} \to \mathcal{H}$ of $f \in \Gamma_0(\mathcal{H})$ with parameter τ, which is defined by

$$(2.2)\qquad \operatorname{prox}_{\tau f}(x) = (I + \partial f)^{-1}(x) = \underset{x' \in \mathcal{H}}{\operatorname{argmin}} \left\{ f(x') + \frac{1}{2\tau}\|x'-x\|^2 \right\}.$$

For a given $f \in \Gamma_0(\mathcal{H})$, we will then denote by $f^* : \mathcal{H} \to [-\infty,+\infty]$ the *Fenchel conjugate* of f, i.e., the function defined by

$$(\forall u \in \mathcal{H}) \quad f^*(u) = \sup_{x \in \mathcal{H}} \{\langle u, x\rangle - f(x)\}.$$

The Fenchel conjugate f^* of f belongs to $\Gamma_0(\mathcal{H})$ and is differentiable at any point with a σ^{-1}-Lipschitz continuous gradient when f is σ-strongly convex; see, e.g., [22, Theorem 18.15]. Furthermore, the following property holds (see [22, Theorem 16.23]):

$$(\forall (x,u) \in \mathcal{H}^2) \quad u \in \partial f(x) \Leftrightarrow x \in \partial f^*(u).$$

Given $\Omega \subset \mathcal{H}$ and $q \geqslant 1$, we say that f is *q-conditioned on* Ω if argmin $f \neq \emptyset$ and

$$(\exists \gamma > 0)(\forall x \in \Omega) \quad \frac{\gamma}{q}\text{dist }(x, \text{argmin } f)^q \leqslant f(x) - \inf f,$$

and say that f is *globally* q-conditioned when it is q-conditioned on $\mathcal{H}$. Further, we say that f is *locally* q-conditioned if, for any $\tilde{x} \in$ argmin f, it is q-conditioned on $\mathbb{B}(\tilde{x}, \varrho)$ for some $\varrho \in \mathbb{R}_+$. Finally, given two sequences $(a_k)_{k\geqslant 1}$ and $(b_k)_{k\geqslant 1}$ of real numbers, we will write $a_k = O(b_k)$ whenever there exists a positive constant $M > 0$ such that $a_k \leqslant M b_k$ for all $k \geqslant 1$. We will further use the more precise notation $a_k = \Theta(b_k)$ if both conditions $a_k = O(b_k)$ and $b_k = O(a_k)$ hold. Note also that we will use the notation $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the norm and the scalar product in all the Hilbert spaces considered.

2.1. Main assumptions.

We make the following assumptions on the data-fit ℓ and the regularizer R:

(L_1) For all $y \in \mathcal{Y}$, the function $\ell_y := \ell(\cdot, y) \in \Gamma_0(\mathcal{Y})$ and is coercive.
(L_2) For all $(y_1, y_2) \in \mathcal{Y}^2$, $\ell(y_1, y_2) \geqslant 0$ and $\ell(y_1, y_2) = 0 \iff y_1 = y_2$.
(L_3) For given "true" data $\bar{y} \in \mathcal{Y}$, $\ell_{\bar{y}}$ is locally q-conditioned for some $q \in [1, +\infty[$.
(R_1) R is σ-strongly convex with $\sigma \in]0, +\infty[$,
(R_2) $\partial R(x^\dagger) \cap \operatorname{Im} A^* \neq \emptyset$.

Observe that, in light of assumption (L_2), assumption (L_3) can be rewritten as

$$(\exists \varrho > 0)(\exists \gamma > 0)(\forall y \in \mathbb{B}(\bar{y}, \varrho)) \quad \frac{\gamma}{q}\|y - \bar{y}\|^q \leqslant \ell(y, \bar{y}).$$

These assumptions on ℓ and R cover a wide range of inverse problems, as discussed next.

DEFINITION 2.1. *A data-fit is said to be additive if there exists $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ such that*

$$(\forall (y_1, y_2) \in \mathcal{Y}^2) \quad \ell(y_1, y_2) = \mathcal{N}(y_1 - y_2).$$

Example 2.2 (data-fit functions). For $\mathcal{Y} = \mathbb{R}^d$, the additive data-fit functions defined by the functions $\mathcal{N}$ below trivially satisfy (L_1)–(L_2). In addition, $\ell_{\bar{y}}$ satisfies (L_3) if and only if $\mathcal{N}$ is locally q-conditioned for some $q \geqslant 1$. We report here some examples of locally and globally conditioned functions $\mathcal{N}$. Many of them are indeed globally conditioned.

- $\mathcal{N}(y) = \frac{1}{2}\|y\|^2$ is globally 2-conditioned, with $\gamma = 1$.
- $\mathcal{N}(y) = \frac{1}{q}\|y\|_q^q$, for $q \geqslant 1$, is globally q-conditioned, with $\gamma = d^r$, where $r = \min(\frac{1}{q} - \frac{1}{2}, 0)$. Note that this includes the case of the ℓ^1-norm.
- The weighted sum [44] $\mathcal{N}(y) = \alpha\|y\|_1 + \frac{1}{2}\|y\|_2^2$, for $\alpha > 0$, is globally 1-conditioned, with $\gamma = \alpha$.
- The Huber data-fit function [35] $\mathcal{N}(y) = \sum_{i=1}^d h_\nu(y^i)$, where $h_\nu \colon \mathbb{R} \to \mathbb{R}_+$ is the Huber smoothing function, defined for $\nu > 0$ by

$$(\forall t \in \mathbb{R}) \quad h_\nu(t) := \begin{cases} \frac{1}{2\nu}t^2 & \text{if } |t| \leqslant \nu, \\ |t| - \frac{\nu}{2} & \text{otherwise.} \end{cases} \tag{2.3}$$

For every $\varrho \in]0, +\infty[$, it is 2-conditioned on $\mathbb{B}(0, \varrho)$, with $\gamma = \min\{\frac{1}{\nu}, \frac{2\varrho - \nu}{\varrho^2}\}$.

- The exact penalization defined by $\mathcal{N}(y) = 0$ if $y = 0$ and $\mathcal{N}(y) = +\infty$ otherwise is globally 1-conditioned, with $\gamma = 1$.

We also mention here a nonadditive data-fit function used in several applications, which also satisfies assumption (L_3):

- The KL divergence, defined by

$$\ell(y_2, y_1) = \mathrm{KL}(y_1, y_2) := \sum_{i=1}^{d} \mathrm{kl}(y_1^i, y_2^i), \tag{2.4}$$

where

$$(\forall (t_1, t_2) \in \mathbb{R}^2) \quad \mathrm{kl}(t_1, t_2) := \begin{cases} t_1 \log \dfrac{t_1}{t_2} - t_1 + t_2 & \text{if } (t_1, t_2) \in \left]0, +\infty\right[^2, \\ +\infty & \text{otherwise.} \end{cases}$$

For every $\varrho \in \left]0, +\infty\right[$, $\ell_{\bar{y}}(\cdot) = \mathrm{KL}(\bar{y}, \cdot)$ is 2-conditioned on $\mathbb{B}(\bar{y}, \varrho)$, with $\gamma = \frac{2}{\varrho c^2} + \frac{2}{\varrho^2 c} \ln \frac{c}{\varrho + c}$, and $c = d\|\bar{y}\|_\infty$ (see Lemma A.2).

Example 2.3 (regularizers). A classical regularizer widely used in signal/image processing as a sparsifying prior is the ℓ^1-norm of the coefficients with respect to an orthonormal basis or, more generally, of a dictionary. Another popular choice in imaging is the total variation seminorm [53], due to its ability to preserve edges, together with its generalizations [30, 37]. For some specific tasks in computer vision and machine learning, there is also a need for structured sparsity. This can be enforced by means of group sparsity inducing norms [60, 18]. While not being strongly convex, these regularizers can be included in our framework by simply adding a quadratic term $\frac{\sigma}{2}\|\cdot\|^2$ where σ is small positive parameter, in the flavor of the elastic net regularization [62].

2.2. Iterative methods based on continuous and discrete dynamics.

It is useful to review some approaches designed for solving (1.1), the hierarchical problem $(P_0(\bar{y}))$, and the Tikhonov-regularized problem $(P_\lambda(\hat{y}))$. In particular, we focus on approaches based on duality and/or combined with diagonal dynamics.

Mirror descent approaches. A class of methods solving (1.1) consider the problem

$$\text{find } x^\dagger \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ R(x) + \delta_{\bar{y}}(Ax) \right\}, \tag{2.5}$$

where the constraint (1.1) is encoded by the indicator function $\delta_{\bar{y}}$. Using Fenchel–Rockafeller duality the corresponding dual problem reads

$$\text{find } u^\dagger \in \operatorname*{argmin}_{u \in \mathcal{Y}} \ \left\{ d_0(u) := R^*(-A^*u) + \langle \bar{y}, u \rangle \right\}. \tag{D_0}$$

Since R^* is smooth (see (ii) in Lemma A.1), a gradient method can be used to solve (D_0); see [25, 49]. This coincides, up to a change of variables, with mirror descent approaches [23] and linearized Bregman iterations [33, 19], where R plays the role of the mirror function. However, extending this approach for solving $(P_0(\bar{y}))$ is not clear.

Primal diagonal dynamics. A classical approach to solve hierarchical problems like $(P_0(\bar{y}))$ is based on the *diagonal principle*, which essentially states that when $\hat{y} = \bar{y}$ and $\lambda \to 0$, problem $(P_\lambda(\hat{y}))$ converges toward $(P_0(\bar{y}))$ in an appropriate sense

[5, Theorem 2.6]. In this view, diagonal approaches have been considered as nonautonomous dynamics solving $(P_\lambda(\hat{y}))$ with a parameter λ monotonically decreasing to zero. The simplest example of a continuous diagonal dynamic is the diagonal steepest descent differential inclusion defined for an initial $t_0 > 0$, which reads

$$(PD_\lambda) \qquad x(t_0) = x_0, \quad \lambda(t) \searrow 0, \quad \dot{x}(t) + \partial p_{\lambda(t)}(x(t)) \ni 0,$$

where $p_{\lambda(t)}(x(t))$ is defined in $(P_\lambda(\hat{y}))$. This dynamic is studied in [11, 13, 7], where convergence of $x(t)$ to $x^\dagger$ was guaranteed provided that $\lambda(t) \to 0$ *fast enough*, i.e., $\lambda \in L^{1/(q-1)}([t_0, +\infty))$, where $q \in [1, +\infty)$ is the exponent in (L_3); see [7, Corollary 3.3, Remark 4.4]. Discrete counterparts of (PD_λ) have also been studied [20, 12, 39]. They can be seen as a variant of the forward-backward algorithm applied to solve problem $(P_\lambda(\hat{y}))$, where the penalization parameter tends to zero along the iterations. A main drawback of this type of algorithms is that they are expensive for nonsmooth data-fit terms, since they require computing the proximal operator of the composition $\ell_{\bar{y}} \circ A$. A possible way to overcome this issue consists in applying Fenchel–Rockafellar duality to $(P_\lambda(\hat{y}))$, thus considering the dual problem (D_λ), where the linear operator appears only in composition with the smooth function R^*. Then, it is possible to apply an explicit gradient step to $R^* \circ (-A^*)$, while the nonsmooth data-fit term can be cheaply treated via its proximal operator.

Dual diagonal dynamics. The dual problem of $(P_\lambda(\hat{y}))$ is

$$(D_\lambda) \qquad \text{find } u_\lambda \in \operatorname*{argmin}_{u \in \mathcal{Y}} \left\{ d_\lambda(u) := R^*(-A^*u) + \frac{1}{\lambda}\ell^*(\lambda u; \hat{y}) \right\}.$$

Solutions of (D_λ) are related to those of $(P_\lambda(\hat{y}))$ via the formula $x_\lambda = \nabla R^*(-A^* u_\lambda)$, which holds thanks to the strong convexity of R. A natural question is whether the diagonal principle can be applied to the dual problem (D_λ) as well. The corresponding dual diagonal continuous dynamics read

$$(DD_\lambda) \qquad u(t_0) = u_0, \quad \lambda(t) \searrow 0, \quad \begin{cases} x(t) = \nabla R^*(-A^*u(t)), \\ \dot{u}(t) + \partial d_{\lambda(t)}(u(t)) \ni 0, \end{cases}$$

where, similarly as before, provided that $\lambda \in L^{1/(q-1)}([t_0, +\infty))$, the trajectory $x(t)$ is guaranteed to converge to $x^\dagger$. The discrete counterpart of (DD_λ) has been studied in [43] under the name of Dual Diagonal Descent algorithm, (3D) where its convergence and stability properties have been investigated. For $\hat{y} \in \mathcal{Y}$ such that $\|\hat{y} - \bar{y}\| \leqslant \delta$ and additive data-fit functions, the authors showed that stopping the algorithm at $k_\delta = \Theta(\delta^{-2/3})$ guarantees that the convergence rate (1.2) holds with $\alpha = 1/3$. However, this rate is not optimal for quadratic data terms [42]. In this paper, we propose a dual diagonal approach which, thanks to the use of acceleration, provides optimal convergence rates and an earlier stopping time.

3. Continuous inertial dual diagonal dynamic.

First-order inertial algorithms are popular in optimization due to their faster convergence on smooth and nonsmooth convex problems; see, e.g., [51, 24]. In several papers continuous inertial dynamics have been studied considering appropriate Lyapunov functions [58, 48, 3]. As already discussed, their regularization properties are also known for quadratic data-fit terms [52, 49]. We propose an inertial approach for general data-fit terms, considering a variant of the dynamic in (DD_λ). Namely, for a given $\alpha > 0$ and initial $t_0 > 0$, we consider

$$(IDD_\lambda)\;\; (u(t_0), \dot{u}(t_0)) = (u_0, \dot{u}_0), \quad \lambda(t) \searrow 0, \quad \begin{cases} x(t) = \nabla R^*(-A^* u(t)), \\ \ddot{u}(t) + \dfrac{\alpha}{t}\dot{u}(t) + \partial d_{\lambda(t)}(u(t)) \ni 0. \end{cases}$$

The asymptotic behavior of the trajectories of this inertial differential inclusion will be analyzed next, while its discrete counterpart will be studied in the rest of the paper.

Remark 3.1. The idea of coupling inertia with Tikhonov regularization is not new. In [9], an inertial variant of the primal dynamic (PD_λ) is proposed for $R = \|\cdot\|^2/2$. The corresponding inertial primal diagonal approach is

$$(IPD_\lambda)\;\; (x(t_0), \dot{x}(t_0)) = (x_0, \dot{x}_0), \quad \lambda(t) \searrow 0, \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \lambda(t)\partial p_{\lambda(t)}(x(t)) \ni 0.$$

Under a suitable decay assumption on $\lambda(\cdot)$ the authors guarantee fast convergence and regularization [9, section 6]. Compared to (IPD_λ), in our dual formulation (IDD_λ) we take advantage of a different scaling between the data-fit and the regularizer. Indeed, to derive (IDD_λ) the data-fit in the primal problem $(P_\lambda(\hat{y}))$ is multiplied by $\lambda(t)^{-1} \to +\infty$, while in (IPD_λ) the regularizer is multiplied by $\lambda(t) \to 0$. For first-order systems this difference is inessential, the two approaches being equivalent for an appropriate change of variables [13]. However, for second-order systems these two scalings describe different dynamics [14, section 4]. This difference can be understood looking at the limits (in the Γ-convergence sense) of the corresponding parametrized functions, which read

$$\text{(3.1)} \qquad \text{if } \lambda \searrow 0, \quad p_\lambda \to p_0 := R + \delta_{\operatorname{argmin} \ell_y \circ A} \quad \text{and} \quad \lambda p_\lambda \to \delta_{\operatorname{dom} R} + \ell_y \circ A.$$

3.1. Convergence of the continuous inertial dual diagonal dynamic.

In this section we study the convergence properties of the trajectories of (IDD_λ), assuming their existence to simplify the analysis. We remark that if d_λ is assumed to be differentiable with a Lipschitz continuous gradient, global existence and uniqueness results of a classical $C^2([t_0, +\infty), \mathbb{R}_+)$ solution to (IDD_λ) hold by the Cauchy–Lipschitz theorem. However, this assumption requires the data-fidelity function $\ell_{\bar{y}}$ to be strongly convex (see [22, Theorem 18.15]), which is in general not the case for most of the data-fit terms; see Example 2.2. We refer to [34, 3] for further details. In the following theorem, we show that the inertial term (IDD_λ) ensures that the dual function values $d_{\lambda(t)}(u(t))$ tend to $\inf d_0$ at a $O(t^{-2})$ rate as expected for inertial methods. Further, switching from the dual to the primal problem by means of the formula $x(t) = \nabla R^*(-A^* u(t))$, we prove the convergence of $x(\cdot)$ to $x^\dagger$. To prove these results, some assumptions on the decay of $\lambda(\cdot)$ are needed, as it is usual for dynamics such as (PD_λ) and (IPD_λ). We thus consider the following assumption:

- (Λ) $\lambda : [0, +\infty[\;\to\;]0, +\infty[$ is a nonincreasing differentiable function such that $\lim_{t\to\infty} \lambda(t) = 0$. If q defined in assumption (L_3) is strictly greater than 1, we assume that the quantity $\Lambda_c := \int_{t_0}^{+\infty} t\lambda^{\frac{1}{q-1}}(t)\, dt$ is finite.

Remark 3.2. A sufficient condition ensuring the validity of (Λ) is that $\lambda(\cdot) \in L^{\frac{1}{2(q-1)}}([t_0, +\infty))$; see Lemma A.4 in the appendix.

We are now ready to state the main convergence result for continuous dynamics. Note that Lemma A.1(iii) ensures that the set of solutions of problem (D_0) is nonempty. To prove fast convergence results of the dual function values, we follow the approach considered in [8, 58, 3] and define a suitable Lyapunov-type function.

THEOREM 3.3. *Let the assumptions* (L_1)–(L_3), (R_1)–(R_2), (Λ) *hold true. Let* $u^\dagger \in \operatorname{argmin} d_0$ *and assume that* $\lambda(t_0)\|u^\dagger\| \leqslant \gamma\varrho^{q-1}/q$. *Let* $\alpha \geqslant 3$ *and let the pair* $(x(\cdot), u(\cdot))$ *be a solution to* (IDD_λ) *in the following sense:*

- $u \in \mathcal{C}^1([t_0, +\infty[, \mathcal{Y})$, *and* $x = \nabla R^* \circ (-A^*) \circ u$,
- *for every* $T > t_0$, $\dot{u}$ *and* $d_{\lambda(\cdot)} \circ u$ *are absolutely continuous on* $[t_0, T]$,
- *for a.e.* $t \in [t_0, +\infty[$, $-\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t) \in \partial d_{\lambda(t)}(u(t))$.

Then, there exists an explicitly computable constant $C \in\,]0, +\infty[$ *such that*

$$\forall t > t_0 \quad d_{\lambda(t)}(u(t)) - \inf d_0 \leqslant \frac{C}{t^2} \qquad \textit{and} \quad \|x(t) - x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma} t}.$$

Proof. We define the following energy:

$$(\forall t \geqslant t_0) \quad \mathcal{E}(t) := t^2 \left(d_{\lambda(t)}(u(t)) - \inf d_0\right) + \frac{1}{2}\|(\alpha - 1)(u(t) - u^\dagger) + t\dot{u}(t)\|^2. \tag{3.2}$$

From now on, we will use the shorthand notation

$$R_A^* := R^* \circ (-A^*), \qquad \ell_{\bar{y}}^*(\cdot) := \ell(\cdot, \bar{y})^* \tag{3.3}$$

so that the composite dual function d_λ can be written as $d_\lambda(u) = R_A^*(u) + \lambda^{-1}\ell_{\bar{y}}^*(\lambda u)$ for every $u \in \mathcal{Y}$. Since $\partial d_{\lambda(t)}(u(t)) = \nabla R_A^*(u(t)) + \partial \ell_{\bar{y}}^*(\lambda(t)u(t))$ [22, Proposition 16.6 and Corollary 16.53], the notion of solution introduced entails the existence of some $\eta : [t_0, +\infty) \to \mathcal{Y}$ such that

$$\text{for a.e. } t > t_0, \quad \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \nabla R_A^*(u(t)) + \eta(t) = 0 \ \text{ and } \ \eta(t) \in \partial \ell_{\bar{y}}^*(\lambda(t)u(t)).$$

We divide the proof into two steps.

Step 1. *Fast convergence rates.* The function $\mathcal{E}$ is differentiable a.e. on $[t_0, +\infty[$ since it is absolutely continuous. We thus compute its derivative and obtain

$$\begin{aligned} \dot{\mathcal{E}}(t) = 2t\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + \frac{t^2\dot{\lambda}(t)}{\lambda^2(t)}\Big(\langle \eta(t), \lambda(t)u(t)\rangle - \ell_{\bar{y}}^*(\lambda(t)u(t))\Big) \\ + t^2\left\langle \dot{u}(t), \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \nabla R_A^*(u(t)) + \eta(t)\right\rangle + t(\alpha - 1)\left\langle u(t) - u^\dagger, \frac{\alpha}{t}\dot{u}(t) + \ddot{u}(t)\right\rangle. \end{aligned}$$

The second term in the expression above is nonpositive because λ is differentiable and decreasing and, moreover, by convexity of $\ell_{\bar{y}}(\cdot)$ together with Lemma A.1(ii), there holds

$$\ell_{\bar{y}}^*(\lambda(t)u(t)) - \langle \eta(t), \lambda(t)u(t)\rangle \leqslant \ell_{\bar{y}}^*(0) = 0.$$

Furthermore, the third term is equal to zero a.e. since $u(\cdot)$ is a solution of (IDD_λ) by assumption. We thus deduce that for a.e. $t > t_0$

$$\dot{\mathcal{E}}(t) \leqslant 2t\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + t(\alpha - 1)\left\langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\right\rangle. \tag{3.4}$$

Using that $-\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t) \in \partial d_{\lambda(t)}(u(t))$ and from the convexity of $d_{\lambda(t)}(\cdot)$ we have

$$\text{for a.e. } t > t_0 \quad \left\langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\right\rangle \leqslant d_{\lambda(t)}(u^\dagger) - d_{\lambda(t)}(u(t)). \tag{3.5}$$

We now add and subtract $\inf d_0 = d_0(u^\dagger)$ and define $r_{\lambda(t)}(u^\dagger) := d_{\lambda(t)}(u^\dagger) - \inf d_0$. We get

$$\left\langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\right\rangle \leqslant r_{\lambda(t)}(u^\dagger) + \inf d_0 - d_{\lambda(t)}(u(t)). \tag{3.6}$$

Applying this inequality to (3.4), since $\alpha \geqslant 3$ and $d_{\lambda(t)}(u(t)) - \inf d_0 \geqslant 0$ (see Proposition A.1.(iv)), we get

$$\dot{\mathcal{E}}(t) \leqslant t(3-\alpha)\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + t(\alpha-1)r_{\lambda(t)}(u^\dagger) \leqslant t(\alpha-1)r_{\lambda(t)}(u^\dagger). \tag{3.7}$$

To bound the right-hand side, we now apply Lemma A.1(vi) and deduce that, since $\lambda(t) \leqslant \lambda(t_0)$,

$$\dot{\mathcal{E}}(t) \leqslant c(\alpha-1)t\lambda(t)^{\frac{1}{q-1}},$$

where the constant c is defined as

$$c := \begin{cases} 0 & \text{if } q = 1, \\ (1-(1/q))\gamma^{-1/(q-1)}\|u^\dagger\|^{q/(q-1)} & \text{if } q > 1 \end{cases} \tag{3.8}$$

and is finite in both cases. Since the above inequality holds for a.e. $t > t_0$, assumption (Λ) yields that for a.e. $t > t_0$,

$$\mathcal{E}(t) = \mathcal{E}(t_0) + \int_{t_0}^{t} \dot{\mathcal{E}}(t) \leqslant \mathcal{E}(t_0) + c(\alpha-1)\Lambda_c.$$

By now defining $C := \mathcal{E}(t_0) + c(\alpha-1)\Lambda_c$, we derive

$$d_{\lambda(t)}(u(t)) - \inf d_0 \leqslant \frac{C}{t^2}. \tag{3.9}$$

Step 2. *Convergence rate for the primal iterates.* From (3.9), used in combination with Lemma A.1(v), we get

$$\begin{aligned}\frac{\sigma}{2}\|x(t)-x^\dagger\|^2 \leqslant d_0(u(t)) - \inf d_0 &= (d_0(u(t)) - d_{\lambda(t)}(u(t))) + (d_{\lambda(t)}(u(t)) - \inf d_0) \\ &\leqslant (d_0(u(t)) - d_{\lambda(t)}(u(t))) + \frac{C}{t^2}.\end{aligned}$$

The monotonicity property of Lemma A.1(iv) implies that the first term on the right-hand side above is nonpositive, whence we get

$$\|x(t)-x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma}t}. \tag{3.10}$$

□

4. Inertial Dual Diagonal Descent (I3D) algorithm.

In this section, we study the convergence properties of the discrete analogue of (IDD_λ), thus deriving an accelerated version of the (3D) algorithm studied in [43].

4.1. From the continuous dynamic to the discrete algorithm.

We follow here a standard approach for computing the time-discretization of continuous dynamical systems considered, e.g., in [1, 16, 58, 8]. Recalling the notation (3.3), we note that (IDD_λ) can be equivalently written as

$$\begin{cases} x(t) = \nabla R^*(-A^*u(t)), \\ \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \partial\ell_{\bar{y}}^*(\lambda(t)u(t)) + \nabla R_A^*(u(t)) \ni 0. \end{cases} \tag{4.1}$$

We discretize (4.1) *explicitly* with respect to the smooth component ∇R_A^* and *semi-implicitly* with respect to the nonsmooth term $\partial\ell_{\bar{y}}^*$. In other words, we discretize implicitly the trajectories, while leaving explicit the dependence on the discretized values λ_k. For $k \geqslant 0$, a fixed time step size $h > 0$ and for time-discretization points $t_k = kh$, we set $u_k := u(t_k)$, $\lambda_k := \lambda(t_k)$ and derive the finite difference scheme

$$\begin{cases} x_k = \nabla R^*(-A^* u_k), \\ \frac{1}{h^2}(u_{k+1} - 2u_k + u_{k-1}) + \frac{\alpha}{kh^2}(u_k - u_{k-1}) + \partial \ell^*_{\bar{y}}(\lambda_k u_{k+1}) + \nabla R^*_A(w_k) \ni 0, \end{cases}$$

where w_k is a linear combination of u_k and u_{k-1} which will be made clear in the following. After straightforward calculations, we rewrite the system above as

$$\begin{cases} x_k = \nabla R^*(-A^* u_k), \\ u_{k+1} + h^2 \partial \ell^*_{\bar{y}}(\lambda_k u_{k+1}) \ni u_k + \left(1 - \frac{\alpha}{k}\right)(u_k - u_{k-1}) - h^2 \nabla R^*_A(w_k). \end{cases} \tag{4.2}$$

By setting $\alpha_k = 1 - \alpha/k$, $\tau := h^2$, and $w_k := u_k + \alpha_k(u_k - u_{k-1})$, we get

$$\begin{cases} w_k \quad = u_k + \alpha_k(u_k - u_{k-1}), \\ u_{k+1} = \left(I + \frac{\tau}{\lambda_k} \partial \ell^*_{\bar{y}}(\lambda_k \cdot)\right)^{-1} (w_k - \tau \nabla R^*_A(w_k)), \\ x_{k+1} = \nabla R^*(-A^* u_{k+1}). \end{cases}$$

Note that the proximal operator of the map $\ell^*_{\bar{y}}(\lambda_k \cdot)$ with parameter τ/λ_k appears, in combination with an explicit gradient step for R^*_A. We can thus introduce the (I3D) algorithm:

(I3D)

For $u_0 = u_1 \in \mathcal{Y}$, compute for $k \geqslant 1$
$$\begin{cases} w_k \quad = u_k + \alpha_k(u_k - u_{k-1}), \\ u_{k+1} = \operatorname{prox}_{\frac{\tau}{\lambda_k}\ell^*_{\bar{y}}(\lambda_k \cdot)} (w_k - \tau \nabla R^*_A(w_k)), \\ x_{k+1} = \nabla R^*(-A^* u_{k+1}). \end{cases}$$

This algorithm depends on three parameters: the step size $\tau > 0$, the relaxation parameters $(\lambda_k)_k$, and the friction parameters $(\alpha_k)_k$. The step size will be chosen depending on the value of the Lipschitz constant of ∇R^*_A. For the choice of the relaxation parameters, we will consider a discrete analogue of the assumption (Λ) formulated in the continuous setting. For the friction parameters α_k, we will allow more general values than the ones above.

We gather the requirements on these parameters in the following assumptions:

(P_1) $\tau \in (0, \frac{\sigma^2}{\|A\|^2}]$, where $\sigma > 0$ is defined in assumption (R_1).

(P_2) α_k is nonnegative and for every $k \geqslant 1$ and $t_k := 1 + \sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j$ is finite, with $t_k = \Theta(k)$.

(P_3) (λ_k) is a strictly positive nonincreasing sequence such that $\lim_{k\to\infty} \lambda_k = 0$. Moreover, by defining

$$\Lambda := \begin{cases} \sum_{k \geqslant 1} t_{k+1} \lambda_k^{1/(q-1)} & \text{if } q > 1, \\ 0 & \text{if } q = 1, \end{cases} \tag{4.3}$$

we have that $\Lambda < +\infty$.

(P_4) For $u^\dagger \in \operatorname{argmin} d_0$, we have $\lambda_0 \|u^\dagger\| \leqslant \gamma \varrho^{q-1}/q$.

Remark 4.1 (on assumption (P_3)). As commented in Remark 3.2, one can check that a sufficient condition for (P_3) to hold is that $\lambda \in \ell^{\frac{1}{2(q-1)}}(\mathbb{N})$. In particular, if we consider a sequence verifying $\lambda_k = O\left(k^{-\theta}\right)$ for some $\theta > 0$, it is easy to verify that (P_3) holds as long as $\theta > 2(q-1)$. For $q = 1$ (for instance, if $\ell(y_1, y_2) = \|y_1 - y_2\|_1$), no summability condition is required. Roughly speaking, the assumption $\lambda \in \ell^{\frac{1}{2(q-1)}}(\mathbb{N})$ means in this case that $\lambda \in \ell^\infty(\mathbb{N})$, which is already implied by $\lim_{k\to\infty} \lambda_k = 0$.

Remark 4.2 (on assumption (P_4)). For many choices of data-fits, $\varrho = +\infty$ (see Example 2.2), in which case the assumption is automatically satisfied. Also, note that in assumption (P_3), we require λ_k to tend to zero. This means that $\lambda_K\|u^\dagger\| \leqslant \gamma\varrho^{q-1}/q$ for some $K \in \mathbb{N}$. In this case, up to a time rescaling $k \leftarrow k+K$, the required estimates always hold true.

Following [6], we require the sequence of friction parameters (α_k) to satisfy (P_2), a particular summability property guaranteeing a technical condition crucial in the following proofs. We summarize such a requirement and the resulting condition in the following lemma.

Lemma 4.3 (see [6, Lemma 2.1]). *Assume that (α_k) is nonnegative and satisfies*

$$\sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j < +\infty \quad \text{for every } k \geqslant 1. \tag{4.4}$$

Then, the sequence defined by

$$t_k := 1 + \sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j \tag{4.5}$$

is well-defined (P_2), and satisfies for every $k \geqslant 1$ the following properties:

$$1 + \alpha_k t_{k+1} = t_k, \qquad t_{k+1}^2 - t_k^2 \leqslant t_{k+1}. \tag{4.6}$$

Remark 4.4 (classical choices of α_k and t_k). Definitions (4.4) and (4.5) above accommodate standard choices of sequences (α_k) and (t_k). For example, in his seminal work Nesterov [50] considered

$$\alpha_k = \frac{t_k - 1}{t_{k+1}} \qquad \text{and} \qquad t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}, \qquad t_1 = 1, \tag{4.7}$$

which can be shown to verify the two conditions (4.4) and (4.5), as well as $k/2 \leqslant t_k \leqslant k$. For a given $\alpha > 1$, the two asymptotically equivalent choices

$$\alpha_k = 1 - \frac{\alpha}{k}, \quad t_{k+1} = \frac{k}{\alpha - 1}, \quad \text{and} \quad \alpha_k = \frac{k-1}{k+\alpha-1}, \quad t_{k+1} = \frac{k+\alpha-1}{\alpha-1}$$

have been recently considered in [36, 2, 10] and can be shown to satisfy (P_2). Note that for $\alpha = 3$ these sequences are asymptotically equivalent to the Nesterov sequences (4.7).

Remark 4.5 (splitting of the loss). In [43] the decomposition of the loss function $\ell_{\bar{y}} = \phi_{\bar{y}} \,\square\, \psi_{\bar{y}}$ was considered, where $\square$ is the infimal convolution and $\psi_{\bar{y}}$ is the possible strongly convex component of $\ell_{\bar{y}}$. In such case, the dual function $\ell_{\bar{y}}^*(\cdot)$ can be expressed as $\ell_{\bar{y}}^* = \psi_{\bar{y}}^* + \phi_{\bar{y}}^*$, where $\phi_{\bar{y}}^*$ is in general nonsmooth, while $\phi_{\bar{y}}^*$ has Lipschitz gradient and can therefore be incorporated with the smooth term R_A^* in the dual function d_λ. For several data discrepancies, however, $\psi_{\bar{y}} = \delta_{\{0\}}$ (see [43, section 4.3]). To simplify the presentation, we do not consider this decomposition in this work.

4.2. Fast convergence of the algorithm. We now prove the discrete analogue of Theorem 3.3 for (I3D). We follow the approach considered in [25, 15, 58, 8, 6].

THEOREM 4.6 (fast convergence). *Let the assumptions* (L_1)*–*(L_3)*,* (R_1)*–*(R_2)*,* (P_1)*–*(P_4) *hold true. Let* (x_k) *and* (u_k) *be the sequences generated by algorithm* (I3D). *Then, there exists* $C \in]0, +\infty[$ *such that*

$$d_{\lambda_k}(u_k) - \inf d_0 \leqslant \frac{C}{t_k^2} \qquad \text{and} \qquad \|x_k - x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma} t_k}. \tag{4.8}$$

Proof. Let $u^\dagger \in \operatorname{argmin} d_0$ be the minimizer of d_0 for which assumption (P_4) holds, and define, for every $k \geqslant 1$, the discrete Lyapunov energy function

$$\mathcal{E}(k) := t_k^2 \Big(d_{\lambda_k}(u_k) - \inf d_0 \Big) + \frac{1}{2\tau} \|z_k - u^\dagger\|^2, \tag{4.9}$$

where z_k is defined as

$$z_k := u_{k-1} + t_k(u_k - u_{k-1}). \tag{4.10}$$

Our goal is to get an estimate on the decay of $\mathcal{E}$ along time. In particular, we will show that for every $k \geqslant 1$

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leqslant t_{k+1} \Big(d_{\lambda_k}(u^\dagger) - \inf d_0 \Big), \tag{4.11}$$

which can be seen as a discrete analogue of (3.7), and from which the desired accelerated convergence rates will follow in a straightforward manner.

For simplicity, let us denote by ℓ_k the function defined by setting

$$(\forall u \in \mathcal{Y}) \quad \ell_k(u) := \lambda_k^{-1} \ell_{\bar{y}}^*(\lambda_k u). \tag{4.12}$$

To prove (4.11), we define for every $k \geqslant 1$ the operator $G_k : \mathcal{Y} \to \mathcal{Y}$ as

$$G_k(z) := \frac{1}{\tau} \Big(z - \operatorname{prox}_{\tau \ell_k}(z - \tau \nabla R_A^*(z)) \Big) \tag{4.13}$$

and notice that the proximal step of (I3D) can be written in terms of G_k as $u_{k+1} = w_k - \tau G_k(w_k)$. The descent lemma (see, e.g., [6, 36]) yields

$$d_{\lambda_k}(w - \tau G_k(w)) \leqslant d_{\lambda_k}(u) + \langle G_k(w), w - u \rangle - \frac{\tau}{2} \|G_k(w)\|^2 \quad \forall\, w, u \in \mathcal{Y}. \tag{4.14}$$

Evaluating (4.14) for $u = u_k$ and $w = w_k$, we get

$$d_{\lambda_k}(u_{k+1}) \leqslant d_{\lambda_k}(u_k) + \langle G_k(w_k), w_k - u_k \rangle - \frac{\tau}{2} \|G_k(w_k)\|^2. \tag{4.15}$$

Similarly, evaluating (4.14) for $u = u^\dagger$ and $w = w_k$, we derive

$$d_{\lambda_k}(u_{k+1}) \leqslant d_{\lambda_k}(u^\dagger) + \langle G_k(w_k), w_k - u^\dagger \rangle - \frac{\tau}{2} \|G_k(w_k)\|^2. \tag{4.16}$$

We now multiply (4.15) by $t_{k+1} - 1$ and we add it to (4.16), thus obtaining

$$\begin{aligned} t_{k+1} d_{\lambda_k}(u_{k+1}) &\leqslant (t_{k+1} - 1) d_{\lambda_k}(u_k) + d_{\lambda_k}(u^\dagger) \\ &\quad + \langle G_k(w_k), (t_{k+1} - 1)(w_k - u_k) + (w_k - u^\dagger) \rangle - \frac{\tau}{2} t_{k+1} \|G_k(w_k)\|^2. \end{aligned} \tag{4.17}$$

As an immediate consequence of Lemma 4.3, we observe that

$$\begin{aligned}(t_{k+1}-1)(w_k-u_k)+w_k &= u_k+t_{k+1}(w_k-u_k)\\ &= u_k+t_{k+1}\alpha_k(u_k-u_{k-1})\\ &= u_{k-1}+(1+t_{k+1}\alpha_k)(u_k-u_{k-1})\\ &= u_{k-1}+t_k(u_k-u_{k-1})=z_k.\end{aligned} \tag{4.18}$$

Thanks to (4.10), the fact that $z_k-\tau t_{k+1}G_k(w_k)=z_{k+1}$, and the previous equality, we can now reorder the terms in (4.17) and rewrite it as

$$\begin{aligned}t_{k+1}(d_{\lambda_k}(u_{k+1})-d_{\lambda_k}(u^\dagger)) \leqslant{}& (t_{k+1}-1)(d_{\lambda_k}(u_k)-d_{\lambda_k}(u^\dagger))\\ &+\frac{1}{2\tau t_{k+1}}\left(\|z_k-u^\dagger\|^2-\|z_{k+1}-u^\dagger\|^2\right).\end{aligned}$$

We now multiply everything by t_{k+1}, rearrange, and get

$$\begin{aligned}&t_{k+1}^2(d_{\lambda_k}(u_{k+1})-d_{\lambda_k}(u^\dagger))+\frac{1}{2\tau}\|z_{k+1}-u^\dagger\|^2\\ &\quad\leqslant (t_{k+1}^2-t_{k+1})(d_{\lambda_k}(u_k)-d_{\lambda_k}(u^\dagger))+\frac{1}{2\tau}\|z_k-u^\dagger\|^2,\end{aligned} \tag{4.19}$$

which can be equivalently rewritten as

$$\begin{aligned}&t_{k+1}^2\Big(d_{\lambda_k}(u_{k+1})-d_{\lambda_k}(u^\dagger)\Big)+\frac{1}{2\tau}\|z_{k+1}-u^\dagger\|^2\\ &\quad\leqslant t_k^2\Big(d_{\lambda_k}(u_k)-d_{\lambda_k}(u^\dagger)\Big)+(t_{k+1}^2-t_{k+1}-t_k^2)\Big(d_{\lambda_k}(u_k)-d_{\lambda_k}(u^\dagger)\Big)+\frac{1}{2\tau}\|z_k-u^\dagger\|^2.\end{aligned}$$

To get the desired terms, we first use on the left-hand side the monotonicity property of the function $d_{\lambda_k}(\cdot)$ as a function of k (see Lemma A.1(iv)) and then add and subtract in the parentheses the term $\inf d_0$, thus getting

$$\begin{aligned}&t_{k+1}^2\Big(d_{\lambda_{k+1}}(u_{k+1})-\inf d_0\Big)+\frac{1}{2\tau}\|z_{k+1}-u^\dagger\|^2\\ &\quad\leqslant t_k^2\Big(d_{\lambda_k}(u_k)-\inf d_0\Big)+(t_{k+1}^2-t_{k+1}-t_k^2)\Big(d_{\lambda_k}(u_k)-\inf d_0\Big)\\ &\qquad+t_{k+1}\Big(d_{\lambda_k}(u^\dagger)-\inf d_0\Big)+\frac{1}{2\tau}\|z_k-u^\dagger\|^2.\end{aligned} \tag{4.20}$$

After rearranging and recalling the definition of $\mathcal{E}$ in (4.9), we deduce

$$\mathcal{E}(k+1)+(t_k^2+t_{k+1}-t_{k+1}^2)\Big(d_{\lambda_k}(u_k)-\inf d_0\Big)\leqslant \mathcal{E}(k)+t_{k+1}\Big(d_{\lambda_k}(u^\dagger)-\inf d_0\Big).$$

Thanks to (4.6) and Lemma A.1(iv), we can now neglect the second term on the left-hand side of the above inequality, finally getting the desired inequality (4.11). Iterating this inequality recursively entails

$$\mathcal{E}(k)\leqslant \mathcal{E}(1)+\sum_{j=1}^{k-1}t_{j+1}\Big(d_{\lambda_j}(u^\dagger)-\inf d_0\Big). \tag{4.21}$$

To bound the sum appearing on the right-hand side, we need to analyze the residuals $r_j:=d_{\lambda_j}(u^\dagger)-\inf d_0$. Similarly as for the estimation obtained in the continuous case, we can use for this purpose the property in Lemma A.1(vi) and get that for some fixed constant $c > 0$ independent on j (defined analogously as in (3.8)), we have

$$r_j \leqslant c\lambda_j^{\frac{1}{q-1}} \qquad \text{for every } j \geqslant 1.$$

By assumption (P_3), with Λ as in (4.3), we thus conclude that

$$\sum_{j=1}^{k-1} t_{j+1} r_j \leqslant c \sum_{j=1}^{k-1} t_{j+1} \lambda_j^{\frac{1}{q-1}} \leqslant c\Lambda < +\infty.$$

This allows us to deduce from (4.21) the convergence rate on the dual values in (4.8) by simply taking $C := \mathcal{E}(1) + c\Lambda$. Finally, the convergence rate on the primal iterates in (4.8) follows from Lemma A.1(v). □

Remark 4.7 (Nesterov scheme as a special case). Let f be any differentiable function in $\Gamma_0(\mathcal{X})$ with Lipschitz-continuous gradient. Take $R = f^*$, $A = -I$, $\bar{y} = 0$, and $\ell(y_1, y_2) = \delta_0(y_2 - y_1)$, so that assumptions (L_1)–(L_3) and (R_1)–(R_2) are verified. In that case, $d_0 = f$, and (I3D) reads

$$u_0 = u_1 \in \mathcal{Y}, \quad \text{compute for } k \geqslant 1 \quad \begin{cases} w_k &= u_k + \alpha_k (u_k - u_{k-1}), \\ u_{k+1} &= w_k - \tau \nabla f(w_k), \\ x_{k+1} &= \nabla f(u_{k+1}), \end{cases}$$

which in the dual exactly performs Nesterov's method [51]. From our rates and Lemma A.1(iv), we deduce that $f(u_k) - \inf f = O(k^{-2})$. Furthermore, according to the Nemirovski and Yudin optimality result [51, Theorem 2.1.7], these rates are optimal over the class of Lipschitz smooth convex functions.

Remark 4.8 (different growth for t_k). In assumption (P_2) we require the sequence (t_k) to satisfy $t_k = \Theta(k)$, but this is actually not used in the proof of Theorem 4.6. What is crucial there is that $t_k < +\infty$, so that Lemma 4.3 can be used. Indeed, one might ask whether it is possible to require $t_k = \Theta(k^\beta)$, with $\beta > 1$ to improve the rates in (3.9). It is a simple exercise to verify that this is not possible, since (4.6) implies $t_k \leqslant t_1 k$, hence we must have $\beta \leqslant 1$ so that the best rates are actually achieved for $\beta = 1$.

5. Stability properties in the presence of errors.

We now study the iterative regularization properties of (I3D) in the presence of noisy data $\hat{y} \in \mathcal{Y}$. We thus consider

(5.1)
$$\text{For } \hat{u}_0 = \hat{u}_1 \in \mathcal{Y}, \text{ compute for } k \geqslant 1 \quad \begin{cases} \hat{w}_k &= \hat{u}_k + \alpha_k (\hat{u}_k - \hat{u}_{k-1}), \\ \hat{u}_{k+1} &= \operatorname{prox}_{\frac{\tau}{\lambda_k} \ell^*_{\hat{y}}(\lambda_k \cdot)} \left(\hat{w}_k - \tau \nabla R^*_A(\hat{w}_k) \right), \\ \hat{x}_{k+1} &= \nabla R^*(-A^* \hat{u}_{k+1}). \end{cases}$$

A first natural question one may ask is how much the dual and primal iterates $\hat{u}_k$ and $\hat{x}_k$ are affected by noise in terms of both convergence and stability. We discuss these issues showing that the noisy perturbation can be interpreted as an error in the calculation of the proximal step of the (I3D) algorithm. Before starting, we motivate the following with an example.

Example 5.1. Assume $\mathcal{Y} = \mathbb{R}$ and $\hat{y} = \bar{y} + \delta$ for some $\bar{y}$, $\delta > 0$. The (I3D) algorithm makes use of the datum only for the evaluation of the proximal operator $\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}$. One possible way to measure the impact of the noise consists then in finding an upper bound for $|\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda\cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}(w)|$ for $w \in \mathcal{Y}$ (see [43, Lemma 10]). Consider the following two illustrative cases:

- $\ell_y = \frac{1}{2}|\cdot - y|^2$. We have

$$\sup_{\bar{y}\in\mathcal{Y}} \sup_{w\in\mathcal{Y}} |\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda\cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}(w)| = \frac{\tau\delta}{1+\tau\lambda}.$$

- $\ell_y = \mathrm{kl}(y;\cdot)$. We have

$$\sup_{\bar{y}\in\mathcal{Y}} \sup_{w\in\mathcal{Y}} |\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda\cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}(w)| = \sqrt{\frac{\tau\delta}{\lambda}}.$$

In the former case, the error assumed in the evaluation of $\bar{y}$ has order δ. However, a different behavior is observed for the latter example. The square-root dependence on δ makes the estimate worse in a small noise regime, when $\delta \ll 1$. Further, notice that in a diagonal regime the sequence (λ_k) converges to zero (P_3), which makes the overall error grow fast along the iterations.

Example 5.1 shows that data-fit terms behave differently in the presence of noise. We thus need to provide an analysis flexible enough to take these differences into account and avoid suboptimal results via worst-case estimates. This is the purpose of the following discussion, where we will see that additive data terms (in the sense of Definition 2.1) behave essentially like $\frac{1}{2}|\cdot - y|^2$, while the KL data term belongs to a class of less stable losses.

5.1. ε-subdifferentials and inexact proximal calculus. In this section, we make precise the notion of noise perturbation we intend to use. To do so, we first recall standard definitions regarding the approximate subdifferential and proximal-type minimization problems.

DEFINITION 5.2 (ε-subdifferential [61]). *Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$, and $\varepsilon \geqslant 0$. The ε-subdifferential of f at $x \in \operatorname{dom} f$ is the set*

$$\partial_\varepsilon f(x) = \{u \in \mathcal{H} : f(x') \geqslant f(x) + \langle u, x' - x\rangle - \varepsilon \ \forall\ x' \in \mathcal{H}\}. \tag{5.2}$$

Such a notion generalizes that of the subdifferential recalled in (2.1). In particular, if $\varepsilon \geqslant 0$, then $\partial f(x) \subset \partial_\varepsilon f(x)$ for any $x \in \mathcal{H}$, and we have

$$0 \in \partial_\varepsilon f(x) \quad \Longleftrightarrow \quad x \in \operatorname{argmin}_\varepsilon f = \{x' \in \mathcal{H} : f(x') \leqslant \inf f + \varepsilon\}. \tag{5.3}$$

We recall now some useful characterizations of the proximal operator of $f \in \Gamma_0(\mathcal{H})$ with parameter $\eta > 0$,

$$p = \operatorname{prox}_{\eta f}(x) \Leftrightarrow \frac{x-p}{\eta} \in \partial f(p) \Leftrightarrow p = \underset{z}{\operatorname{argmin}} \left\{ f(z) + \frac{1}{2\eta}\|z - x\|^2 \right\}. \tag{5.4}$$

Next, we introduce notions of approximation of proximal points that can be seen as relaxed conditions of the characterizations in (5.4) (for details see [54, 17]).

DEFINITION 5.3 (approximation of proximal points). *Let $f \in \Gamma_0(\mathcal{H})$, $x \in \mathcal{H}$, $\eta > 0$, and $p := \operatorname{prox}_{\eta f}(x)$. We say that $\hat{p} \in \mathcal{H}$ is*

- *a type* 1 *approximation of* p *with precision* ε_1*, and we write* $\hat{p} \approx_1^{\varepsilon_1} p$*, if*

$$\exists e \in \mathcal{H},\ \exists(\varepsilon_1, \varepsilon_2, \varepsilon_3) \in [0, +\infty[^2,\ \|e\| \leqslant \varepsilon_3,\ \varepsilon_2^2 + \varepsilon_3^2 \leqslant \varepsilon_1^2,\ \frac{x + e - \hat{p}}{\eta} \in \partial_{\frac{\varepsilon_2^2}{2\eta}} f(\hat{p});$$

- *a type* 2 *approximation of* p *with precision* ε_2*, and we write* $\hat{p} \approx_2^{\varepsilon_2} p$*, if*

$$\exists \varepsilon_2 \in [0, +\infty[,\ \frac{x - \hat{p}}{\eta} \in \partial_{\frac{\varepsilon_2^2}{2\eta}} f(\hat{p}); \tag{5.5}$$

- *a type* 3 *approximation of* p *with precision* ε_3*, and we write* $\hat{p} \approx_3^{\varepsilon_3} p$*, if*

$$\exists e \in \mathcal{H}, \exists \varepsilon_3 \in [0, +\infty],\ \|e\| \leqslant \varepsilon_3,\ \ \frac{x + e - \hat{p}}{\eta} \in \partial f(\hat{p}). \tag{5.6}$$

Type 3 approximations simply describe the presence of an additive error in the argument of the proximal map, i.e., $\hat{p} = \operatorname{prox}_{\eta f}(x + e)$. We show in section 6.1 that this type of error arises naturally when additive data-fit functions are used. Type 2 approximations correspond to the presence of errors in the subdifferential operator. Type 1 approximations can be seen as a combination of type 2 and type 3 approximations, and the following lemma provides an easy characterization.

Lemma 5.4 (see [55, 54]). *Let* $f \in \Gamma_0(\mathcal{H})$*,* $x \in \mathcal{H}$*,* $\eta > 0$*. Then*

$$\hat{p} \approx_1^{\varepsilon_1} \operatorname{prox}_{\eta f}(x) \quad \Leftrightarrow \quad \hat{p} \in \operatorname{argmin}_{\varepsilon_1} \left\{ f(\cdot) + \frac{1}{2\eta} \| \cdot - x\|^2 \right\}. \tag{5.7}$$

We are now ready to study the stability properties of the (I3D) algorithm.

5.2. Stability estimates in the presence of errors. Using the notions introduced in the previous section, we can quantify the error due to the replacement of $\bar{y}$ by $\hat{y}$. In particular, recalling Definition 5.3, we assume that at each iteration the proximal step with $\hat{y}$ is an i-type approximation of the proximal step with $\bar{y}$, where $i \in \{1, 2, 3\}$:

$$(E_i) \quad (\forall k \geqslant 1)(\exists \varepsilon_{i,k} \geqslant 0) \ \text{ s.t.} \quad (\forall w \in \mathcal{Y}) \quad \operatorname{prox}_{\frac{\tau}{\lambda_k} \ell^*_{\hat{y}}(\lambda \cdot)}(w) \approx_i^{\varepsilon_{i,k}} \operatorname{prox}_{\frac{\tau}{\lambda_k} \ell^*_{\bar{y}}(\lambda \cdot)}(w).$$

In section 6 we show that this is indeed a natural assumption for standard data-fit terms.

We can now prove our second main result for (I3D) which provides error estimates under assumption (E_i) with $i = 1$. Stability results for type 2 and type 3 approximations are deduced as particular cases after noticing that for these choices the error terms with $\varepsilon_{3,k}$ and $\varepsilon_{2,k}$ vanish, respectively, for every k.

Theorem 5.5 (error estimates for type 1 errors). *Assume that* (L_1)*–*(L_3)*,* (R_1)*–*(R_2)*,* (P_1)*–*(P_4) *hold true. Let* $(\hat{x}_k)$*,* $(\hat{u}_k)$ *be the sequences generated by* (I3D) *with noisy datum* $\hat{y}$*, and suppose that* (E_i) *holds with* $i = 1$*. Then, the following stability estimate holds true:*

$$(\forall k \geqslant 1) \quad t_k^2 \frac{\sigma\tau}{2} \|\hat{x}_k - x^\dagger\|^2 \leqslant C + \sum_{j=1}^{k-1} t_{j+1}^2 \varepsilon_{2,j}^2 + \frac{5}{2} \left(\sum_{j=1}^{k-1} t_{j+1} \varepsilon_{3,j} \right)^2, \tag{5.8}$$

where the constant C is defined as $C := 2\tau t_1^2(d_1(\hat{u}_0) - \inf d_0) + \|\hat{u}_0 - u^\dagger\|^2 + C_q$ *with*

$$C_q := \begin{cases} 0 & \text{if } q = 1, \\ 2\tau\Lambda(1 - \frac{1}{q})\gamma^{-1/(q-1)}\|u^\dagger\|^{q/(q-1)} & \text{if } q > 1. \end{cases}$$

Proof. Following the proof of Theorem 4.6, we define the discrete energy function

$$\hat{\mathcal{E}}(k) := t_k^2\Big(d_{\lambda_k}(\hat{u}_k) - \inf d_0\Big) + \frac{1}{2\tau}\|\hat{z}_k - u^\dagger\|^2 \tag{5.9}$$

for $k \geqslant 1$, where $u^\dagger \in \operatorname{argmin} d_0$ (so that $\inf d_0 = d_0(u^\dagger)$) and $\hat{z}_k$ is defined as

$$\hat{z}_k := \hat{u}_{k-1} + t_k(\hat{u}_k - \hat{u}_{k-1}). \tag{5.10}$$

Since $\hat{u}_{k+1} \approx_1^{\varepsilon_{1,k}} \operatorname{prox}_{\tau\lambda_k^{-1}\ell_{\hat{y}}^*(\lambda_k\ \cdot)}(\hat{w}_k - \tau\nabla\mathcal{R}_A(\hat{w}_k))$, using Definition 5.3, we have

$$\xi_k := \frac{\hat{w}_k + e_k - \hat{u}_{k+1}}{\tau}, \qquad \xi_k - \nabla\mathcal{R}_A(\hat{w}_k) \in \partial_{\frac{\varepsilon_{2,k}^2}{2\tau}}\ell_{\hat{y}}^*(\lambda_k\hat{u}_{k+1}), \tag{5.11}$$

where $e_k \in \mathcal{H}$, $\varepsilon_{2,k}^2 + \varepsilon_{3,k}^2 \leqslant \varepsilon_{1,k}^2$, and $\|e_k\| \leqslant \varepsilon_{3,k}$. Without loss of generality, we can assume that $\varepsilon_{2,k}^2 + \varepsilon_{3,k}^2 = \varepsilon_{1,k}^2$. Thus, thanks to the descent lemma proved in [59, Lemma 4.1] and applied to $d_{\lambda_k} = \mathcal{R}_A + \lambda_k^{-1}\ell_{\hat{y}}^*(\lambda_k\ \cdot)$, we derive

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(u) + \langle \hat{u}_{k+1} - u, \xi_k\rangle + \frac{L}{2}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \frac{\varepsilon_{2,k}^2}{2\tau} \qquad \forall u \in \mathcal{Y}, \tag{5.12}$$

where $L = \|A\|^2/\sigma^2$. Using the fact that $\tau L \leqslant 1$ by (P_1), rearranging and neglecting nonpositive quantities, we obtain that for all $u \in \mathcal{Y}$

$$\begin{aligned} d_{\lambda_k}(\hat{u}_{k+1}) &\leqslant d_{\lambda_k}(u) - \frac{1}{\tau}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \left\langle \hat{u}_{k+1} - \hat{w}_k, \frac{e_k}{\tau}\right\rangle + \langle \hat{w}_k - u, \xi_k\rangle \\ &\quad + \frac{1}{2\tau}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \frac{\varepsilon_{2,k}^2}{2\tau} \\ &= d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\frac{\hat{u}_{k+1} - \hat{w}_k}{\tau}\|^2 + \tau\left\langle \frac{\hat{u}_{k+1} - \hat{w}_k}{\tau}, \frac{e_k}{\tau}\right\rangle + \frac{\varepsilon_{2,k}^2}{2\tau} \\ &= d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{1}{2\tau}\left(\|e_k\|^2 + \varepsilon_{2,k}^2\right) \\ &\leqslant d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau}, \end{aligned} \tag{5.13}$$

which can be seen as a noisy version of (4.14). We divide the rest of the proof into three steps. Since the former ones are analogous to the calculations done in the error-free case, we will skip some of the details for those.

Step 1. We show that for every $k \geqslant 1$, there holds

$$\hat{\mathcal{E}}(k+1) - \hat{\mathcal{E}}(k) \leqslant t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big) + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_k - u^\dagger\rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2. \tag{5.14}$$

To prove this, we write the descent inequality (5.13) first for $u = \hat{u}_k$,

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(\hat{u}_k) + \langle \hat{w}_k - \hat{u}_k, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau}, \tag{5.15}$$

and then for $u = u^\dagger$,

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(u^\dagger) + \langle \hat{w}_k - u^\dagger, \xi_k \rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau}. \tag{5.16}$$

We now multiply (5.15) by $t_{k+1} - 1$ and add it to (5.16), thus getting

$$\begin{aligned} t_{k+1} d_{\lambda_k}(\hat{u}_{k+1}) &\leqslant (t_{k+1} - 1) d_{\lambda_k}(\hat{u}_k) + d_{\lambda_k}(u^\dagger) \\ &\quad + \langle \xi_k, (t_{k+1} - 1)(\hat{w}_k - \hat{u}_k) + \hat{w}_k - u^\dagger \rangle - \frac{t_{k+1}\tau}{2}\|\xi_k\|^2 + \frac{t_{k+1}}{2\tau}\varepsilon_{1,k}^2. \end{aligned} \tag{5.17}$$

We apply the property $(t_{k+1} - 1)(\hat{w}_k - \hat{u}_k) + \hat{w}_k = \hat{z}_k$ (see (4.10)) and write (5.17) as

$$\begin{aligned} t_{k+1}(d_{\lambda_k}(\hat{u}_{k+1}) - d_{\lambda_k}(u^\dagger)) &\leqslant (t_{k+1} - 1)(d_{\lambda_k}(\hat{u}_k) - d_{\lambda_k}(u^\dagger)) \\ &+ \frac{1}{2\tau t_{k+1}}\left(\|\hat{z}_k - u^\dagger\|^2 - \|\hat{z}_k - u^\dagger - \tau t_{k+1}\xi_k\|^2\right) + \frac{t_{k+1}}{2\tau}\varepsilon_{1,k}^2. \end{aligned}$$

From the identity $-\tau t_{k+1}\xi_k = \hat{z}_{k+1} - \hat{z}_k - t_{k+1} e_k$, we deduce

$$\begin{aligned} & t_{k+1}(d_{\lambda_k}(\hat{u}_{k+1}) - d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_{k+1} - u^\dagger\|^2 \\ &\leqslant (t_{k+1} - 1)(d_{\lambda_k}(\hat{u}_k) - d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_k - u^\dagger\|^2 \\ &\quad + \frac{1}{\tau}\langle \hat{z}_{k+1} - u^\dagger, e_k \rangle + \frac{t_{k+1}}{2\tau}\left(\varepsilon_{1,k}^2 - \|e_k\|^2\right). \\ &= (t_{k+1} - 1)(d_{\lambda_k}(\hat{u}_k) - d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_k - u^\dagger\|^2 \\ &\quad + \frac{1}{\tau}\langle \hat{z}_{k+1} - u^\dagger, e_k \rangle + \frac{t_{k+1}}{2\tau}\varepsilon_{2,k}^2. \end{aligned}$$

We now multiply everything by t_{k+1}, rearrange, and get

$$\begin{aligned} & t_{k+1}^2\Big(d_{\lambda_k}(\hat{u}_{k+1}) - d_{\lambda_k}(u^\dagger)\Big) + \frac{1}{2\tau}\|\hat{z}_{k+1} - u^\dagger\|^2 \\ &\leqslant t_k^2\Big(d_{\lambda_k}(\hat{u}_k) - d_{\lambda_k}(u^\dagger)\Big) + (t_{k+1}^2 - t_{k+1} - t_k^2)\Big(d_{\lambda_k}(\hat{u}_k) - d_{\lambda_k}(u^\dagger)\Big) \\ &\quad + \frac{1}{2\tau}\|\hat{z}_k - u^\dagger\|^2 + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_{k+1} - u^\dagger \rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2. \end{aligned}$$

Using now that $d_{\lambda_k}(\hat{u}_k) \geqslant \inf d_0$ (see Lemma A.1(iv)), adding and subtracting in the parentheses the term $\inf d_0$, and after recalling the definition of $\mathcal{E}$ in (5.9), we get

$$\begin{aligned} & \hat{\mathcal{E}}(k+1) + (t_k^2 + t_{k+1} - t_{k+1}^2)\Big(d_{\lambda_k}(u_k) - \inf d_0\Big) \\ &\qquad \leqslant \hat{\mathcal{E}}(k) + t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big) + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_k - u^\dagger \rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2, \end{aligned} \tag{5.18}$$

whence we deduce condition (5.14) since $t_k^2 + t_{k+1} - t_{k+1}^2 \geqslant 0$ and $d_{\lambda_k}(u^\dagger) - \inf d_0 \geqslant 0$ (see (4.6)). Iterating recursively (5.14), the Cauchy–Schwarz inequality yields

$$\hat{\mathcal{E}}(k) \leqslant \hat{\mathcal{E}}(1) + \sum_{j=1}^{k-1} t_{j+1}\Big(d_{\lambda_j}(u^\dagger) - \inf d_0\Big) + \sum_{j=1}^{k-1}\frac{t_{j+1}}{\tau}\varepsilon_{3,j}\|\hat{z}_{j+1} - u^\dagger\| + \sum_{j=1}^{k-1}\frac{t_{j+1}^2}{2\tau}\varepsilon_{2,j}^2, \tag{5.19}$$

which is the starting point used in the following to deduce the desired stability estimate. We now study separately the sums appearing on the right-hand side of (5.19).

Step 2. For the first term in (5.19), following the proof of Theorem 4.6, we get

$$\sum_{j=1}^{k-1} t_{j+1}\Big(d_{\lambda_j}(u^\dagger) - \inf d_0\Big) \leqslant c\sum_{j=1}^{k-1} t_{j+1}\lambda_{\lambda_j}^{\frac{1}{q-1}} \leqslant c\Lambda < +\infty,$$

where c is defined in (3.8), and Λ is finite thanks to assumption (P_3).

Step 3. To bound the second sum in (5.19), we observe that by definition $\hat{\mathcal{E}}(k) \geqslant \frac{1}{2\tau}\|\hat{z}_k - u^\dagger\|^2$. Then, we set $C = 2\tau(\hat{\mathcal{E}}(1) + c\Lambda)$ and derive

$$\|\hat{z}_k - u^\dagger\|^2 \leqslant C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2 + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\|\hat{z}_{j+1} - u^\dagger\|. \tag{5.20}$$

We now recall Lemma A.5, which applied to $a_k = \|\hat{z}_k - u^\dagger\|$, $b_k = 2t_{k+1}\varepsilon_{3,k}$, $c_{k-1} = C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2$ implies

$$\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\|\hat{z}_{j+1} - u^\dagger\| \leqslant \left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right).$$

Combining altogether in (5.19), we thus deduce

$$\hat{\mathcal{E}}(k) \leqslant \frac{C}{2\tau} + \sum_{j=1}^{k-1}\frac{t_{j+1}^2}{2\tau}\varepsilon_{2,j}^2 + \frac{1}{\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right). \tag{5.21}$$

Young's inequality applied to the product appearing on the right-hand side of (5.21) yields

$$\begin{aligned}
&\frac{1}{\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\\
&\quad = \frac{1}{\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2}\right) + \frac{2}{\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2\\
&\quad \leqslant \frac{5}{2\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2 + \frac{C}{2\tau} + \frac{1}{2\tau}\sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{3,j}^2.
\end{aligned}$$

Hence, we thus obtain from (5.21)

$$\hat{\mathcal{E}}(k) \leqslant \frac{C}{\tau} + \frac{1}{\tau}\sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2 + \frac{5}{2\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2. \tag{5.22}$$

To conclude, we use Lemma A.1(v) and deduce

$$\hat{\mathcal{E}}(k) \geqslant t_k^2(d_{\lambda_k}(\hat{u}_k) - \inf d_0) \geqslant t_k^2(d_0(\hat{u}_k) - \inf d_0) \geqslant \frac{t_k^2\sigma}{2}\|\hat{x}_k - x^\dagger\|^2, \tag{5.23}$$

which combined with (5.22) provides the desired stability estimate (5.8). ∎

5.3. Early stopping. Starting from the stability estimate (5.22), in this section we provide early stopping results guaranteeing the iterative regularization properties of (I3D). These results quantify the reconstruction error $\|\hat{x}_{k(\delta)} - x^\dagger\|$ that can be achieved by *stopping* the algorithm on noisy data at a suitable *early* iteration k_δ. As expected, when errors are small we can recover a good reconstruction by stopping the algorithm later. On the other hand, when the errors are large, the algorithm needs to be stopped earlier to guarantee a good reconstruction. Note that these errors can be constant, or even increasing along iterations. We also show that the convergence rates we obtain depend on the *type* of error considered (see Definition 5.3). Adapting Theorem 5.5 to the three cases of assumption (E_i) for $i \in \{1, 2, 3\}$, we thus derive the following three theorems.

THEOREM 5.6 (early stopping for type 1 errors). *Assume that* (L_1)–(L_3), (R_1)–(R_2), (P_1)–(P_4) *hold true, and suppose that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$. *Let* $(\hat{x}_k)$ *be the sequence generated by* (I3D) *with noisy datum* $\hat{y}$, *and assume that* (E_i) *holds with* $i = 1$, $\varepsilon_{2,k} = O(\delta\lambda_k^{-r_2})$, $\varepsilon_{3,k} = O(\delta\lambda_k^{-r_3})$ *for some* $\delta > 0$ *and* $r_2, r_3 \geqslant 0$. *Set*

$$\alpha := \max\left\{\frac{2}{3+2r_2\theta}, \frac{1}{2+r_3\theta}\right\}. \tag{5.24}$$

Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-\alpha})$ *verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^{\alpha}\right) \quad \textit{for } \delta \searrow 0. \tag{5.25}$$

Proof. We apply the stability estimate (5.8) provided by Theorem 5.5. After substituting the expression for $\varepsilon_{2,k}$ and $\varepsilon_{3,k}$, we get

$$\begin{aligned} t_k^2\|\hat{x}_k - x^\dagger\|^2 &= O\left(1 + \sum_{j=1}^{k-1} t_{j+1}^2 \varepsilon_{2,j}^2 + \left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2\right) \\ &= O(1 + \delta^2 k^{3+2r_2\theta} + \delta^2 k^{4+2r_3\theta}). \end{aligned} \tag{5.26}$$

In correspondence with the stopping time $k(\delta)$, and using the fact that $t_{k(\delta)} = \Theta(k(\delta))$, we deduce from above

$$\begin{aligned} \|\hat{x}_{k(\delta)} - x^\dagger\|^2 &= O\left(\delta^{2\alpha} + \delta^{2-\alpha(1+2r_2\theta)} + \delta^{2-2\alpha(1+r_3\theta)}\right) \\ &= O\left(\delta^{\min\{2\alpha;2-\alpha(1+2r_2\theta),2-2\alpha(1+r_3\theta)\}}\right). \end{aligned} \tag{5.27}$$

Let us now define $\beta := \min\{\frac{1}{2} + r_2\theta; 1 + r_3\theta\}$. We easily see that

$$\min\{2-\alpha(1+2r_2\theta); 2-2\alpha(1+r_3\theta)\} = 2 - 2\alpha\beta, \tag{5.28}$$

so that $\min\{2\alpha, 2-\alpha(1+2r_2\theta), 2-2\alpha(1+r_3\theta)\} = \min\{2\alpha, 2-2\alpha\beta\}$, which is maximal for $\alpha = \frac{1}{1+\beta}$. ∎

The analogous results for errors of types 2 and 3 are straightforward.

THEOREM 5.7 (early stopping for type 2 errors). *Assume that the assumptions* (L_1)–(L_3), (R_1)–(R_2), (P_1)–(P_4) *hold true, and suppose that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$. *Let* $(\hat{x}_k)$ *be the sequence generated by* (I3D) *with noisy datum* $\hat{y}$, *and assume that* (E_i) *holds with* $i = 2$, $\varepsilon_{2,k} = O(\delta\lambda_k^{-r_2})$ *for some* $\delta > 0$ *and* $r_2 \geqslant 0$. *Then, any early stopping rule with* $k(\delta) = \Theta(\delta^{-\frac{2}{3+2\theta r}})$ *verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^{\frac{2}{3+2\theta r}}\right) \quad \textit{for } \delta \searrow 0. \tag{5.29}$$

Proof. For type 2 approximation (5.8) $\varepsilon_{3,k} \equiv 0$, and we get

$$\begin{aligned} t_k^2\|\hat{x}_k - x^\dagger\|^2 &= O\left(1 + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2\right) = O\left(1 + \sum_{j=1}^{k-1} \delta^2 j^{2+2r\theta}\right) \\ &= O(1+\delta^2 k^{3+2r\theta}). \end{aligned} \tag{5.30}$$

In correspondence with any stopping time $k(\delta) = \Theta(\delta^{-\alpha})$, we thus have

$$\|\hat{x}_{k(\delta)} - x^\dagger\|^2 = O\left(k(\delta)^{-2} + \delta^2 k(\delta)^{1+2r\theta}\right) = O\left(\delta^{2\alpha} + \delta^{2-\alpha(1+2r\theta)}\right). \tag{5.31}$$

The term on the right-hand side is minimized when $\alpha = \frac{2}{3+2\theta r}$. ◻

Theorem 5.8 (early stopping for type 3 errors). *Assume that the assumptions (L_1)–(L_3), (R_1)–(R_2), (P_1)–(P_4) hold true, and suppose that $\lambda_k = \Theta(k^{-\theta})$ with $\theta > 2(q-1)$. Let $(\hat{x}_k)$ be the sequence generated by* (I3D) *with noisy datum $\hat{y}$, and assume that (E_i) holds with $i = 3$ with $\varepsilon_{3,k} = O(\delta\lambda_k^{-r_3})$ for some $\delta > 0$ and $r_3 \geqslant 0$. Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-\frac{1}{2+\theta r}})$ verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^{\frac{1}{2+\theta r}}\right) \quad \text{for } \delta \searrow 0. \tag{5.32}$$

Proof. Assuming type 3 errors means that in the estimate (5.8) $\varepsilon_{2,k} \equiv 0$, so that

$$t_k^2\|\hat{x}_k - x^\dagger\|^2 = O(1) + O\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2 = O(1) + O\left(\sum_{j=1}^{k-1} \delta j^{1+r\theta}\right)^2 = O(1+\delta^2 k^{4+2r\theta}). \tag{5.33}$$

In correspondence with the stopping time $k(\delta) = \Theta(\delta^{-\alpha})$, we thus deduce

$$\|\hat{x}_{k(\delta)} - x^\dagger\|^2 = O\left(k(\delta)^{-2} + \delta^2 k(\delta)^{2+2r\theta}\right) = O\left(\delta^{2\alpha} + \delta^{2-2\alpha(1+r\theta)}\right). \tag{5.34}$$

The term on the right-hand side is minimal whenever $\alpha = \frac{1}{2+\theta r}$. ◻

6. Applications to specific data-fit terms.

We now apply the results from section 5.3 to some standard data-fit terms relevant in several applications. We introduce the following definition of noise perturbation.

Definition 6.1 (δ-perturbation). *For given $\bar{y},\ \hat{y} \in \mathcal{Y}$ and $\delta \in \mathbb{R}_{++}$, we say that $\hat{y}$ is a δ-perturbation of $\bar{y}$ according to ℓ if*

$$\ell_{\hat{y}}(\bar{y}) = \ell(\bar{y}, \hat{y}) \leqslant \delta^q,$$

where $q \in [1, +\infty)$ is the conditioning exponent appearing in (L_3).

We now show that a δ-perturbation $\hat{y}$ of $\bar{y}$ corresponds to considering a proximal mapping of $\ell_{\hat{y}}^*$ approximating the corresponding proximal mapping of $\ell_{\bar{y}}^*$ in the sense of Definition 5.3 with some precision $\varepsilon(\delta)$ depending on the noise level δ.

6.1. Additive data-fit terms.

For additive data-fit terms (see Example 2.2), a δ-perturbation corresponds to a type 3 approximation of the proximal mapping.

Proposition 6.2 (additive data-fit terms lead to type 3 errors). *Let $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ and assume that $\ell_{y_2}(y_1) = \mathcal{N}(y_2 - y_1)$ for every $(y_1, y_2) \in \mathcal{Y}^2$. For given $(\delta, \tau, \lambda) \in (0, +\infty)^3$, let $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$ be a δ-perturbation of $\bar{y}$ in the sense of Definition* 6.1. *Then*

$$(\forall z \in \mathcal{Y}) \quad \hat{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell_{\hat{y}}^*(\lambda\cdot)}(z) \ \approx_3^\varepsilon \ \bar{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell_{\bar{y}}^*(\lambda\cdot)}(z)$$

with precision $\varepsilon = \tau\delta(q/\gamma)^{1/q}$ *and where* $q \geqslant 1$ *and* $\gamma > 0$ *are the conditioning parameters appearing in assumption* (L_3)*.*

Proof. We need to find $e \in \mathcal{Y}$ and $\varepsilon \geqslant 0$ such that $\|e\| \leqslant \varepsilon$ and

$$\frac{z + e - \hat{p}}{\tau} \in \frac{1}{\lambda}\partial\ell^*_{\bar{y}}(\lambda\cdot)(\hat{p}). \tag{6.1}$$

Due to the special form of the data-fit we start noting that for any $u \in \mathcal{Y}$ we have

$$\ell^*_{\bar{y}}(u) = \mathcal{N}^*(u) + \langle \bar{y}, u\rangle,$$

and the same holds for $\ell^*_{\hat{y}}$. Then

$$\partial\ell^*_{\hat{y}}(\lambda\cdot)(\hat{p}) = \lambda\partial\ell^*_{\hat{y}}(\lambda\hat{p}) = \lambda\partial\Big(\mathcal{N}^* + \langle\hat{y},\cdot\rangle\Big)(\lambda\hat{p}) = \lambda\partial\mathcal{N}^*(\lambda\hat{p}) + \lambda\hat{y}. \tag{6.2}$$

By definition of $\hat{p}$ we have that $(z - \hat{p})/\tau \in (1/\lambda)\partial\ell^*_{\hat{y}}(\lambda\cdot)(\hat{p}) = \partial\mathcal{N}^*(\lambda\hat{p}) + \hat{y}$, which, by simple algebraic manipulations, entails the required condition (6.1), since

$$\frac{z - \hat{p}}{\tau} \in \partial\mathcal{N}^*(\lambda\hat{p}) + \bar{y} + (\hat{y} - \bar{y}) \Longleftrightarrow \frac{z - \hat{p} + \tau(\bar{y} - \hat{y})}{\tau} \in \partial\mathcal{N}^*(\lambda\hat{p}) + \bar{y} = \frac{1}{\lambda}\partial\ell^*_{\bar{y}}(\lambda\cdot)(\hat{p}).$$

By now setting $e = \tau(\bar{y} - \hat{y})$, we can find the required value of ε combining the q-conditioning of the function $\ell_{\bar{y}}$ on $\mathbb{B}(\bar{y}, \varrho)$ assumed in (L_3) with the δ-perturbation assumption:

$$\|e\| = \tau\|\bar{y} - \hat{y}\| \leqslant \tau\left(\frac{q}{\gamma}\ell(\hat{y}, \bar{y})\right)^{1/q} \leqslant \tau\left(\frac{q}{\gamma}\right)^{1/q}\delta =: \varepsilon,$$

where $\gamma > 0$ and $q \geqslant 1$ are the conditioning parameters. We can thus conclude that $\hat{p}$ is a ε-approximation of $\bar{p}$ with precision ε, as required. ◻

Thanks to Proposition 6.2, we can now derive the early stopping result for additive data-fit terms by applying Theorem 5.8 with the above choice of ε.

COROLLARY 6.3 (early stopping for additive data-fit terms). *Let* $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ *and set* $\ell_{y_2}(y_1) = \mathcal{N}(y_2 - y_1)$ *for every* $(y_1, y_2) \in \mathcal{Y}^2$*. Assume that the assumptions* (L_1)*–*(L_3)*,* (R_1)*–*(R_2)*,* (P_1)*–*(P_4) *hold and that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$*. Let* $(\hat{x}_k)$ *be the sequence generated by* (I3D) *with* $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$*, such that* $\hat{y}$ *is a* δ*-perturbation of* $\bar{y}$*. Then, any early stopping rule with* $k(\delta) = \Theta(\delta^{-1/2})$ *verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{1}{2}}) \quad \textit{for } \delta \searrow 0. \tag{6.3}$$

Remark 6.4 (optimality of the rates). The convergence rate in (6.3) is optimal for regularization methods with additive data-fit terms [42]. Among inertial algorithms, optimal convergence rates for different choices of regularizers but only quadratic data-fit terms have been proved in [52, 49]. For more general additive data-fits (e.g., the ℓ^1-norm; see Example 2.2), in [26] the authors prove a rate $O(\delta^{1/2})$ in terms of the Bregman distance, which is different from (6.3). To our knowledge, our result is the first one showing optimal convergence rates for iterative regularization methods when general data-fit terms are considered and improving the estimates obtained in [43] that showed a suboptimal rate $O(\delta^{1/3})$.

Remark 6.5 (different growth for t_k). As noted in Remark 4.8, if we replace $t_k = \Theta(t_k)$ by $t_k = \Theta(k^\beta)$, then $\beta \leqslant 1$, and $\beta = 1$ gives the fastest convergence rate for true datum $\bar{y}$. Corollary 6.3 implies that also for noisy data $\hat{y}$, any stopping rule with $k(\delta) = \Theta(\delta^{-1/(1+\beta)})$ verifies $\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{\beta}{\beta+1}})$ for $\delta \searrow 0$, where again the best rate is achieved for $\beta = 1$.

6.2. KL divergence. We consider the KL divergence as an example of a nonadditive data-fit term. KL divergence is often used in the presence of Poisson noise in the measurements. We show that for the KL divergence, δ-perturbations lead to type 2 approximations. We recall that the KL divergence is locally 2-conditioned (see Example 2.2).

PROPOSITION 6.6. *Assume that* $\ell_{y_2}(y_1) = \mathrm{KL}(y_2; y_1)$ *for every* $(y_1, y_2) \in \mathcal{Y}^2$*. For* $(\delta, \tau, \lambda) \in (0, +\infty)^3$*, let* $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$ *be a* δ*-perturbation of* $\bar{y}$*. Then*

$$(\forall z \in \mathcal{Y}) \quad \hat{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}(z) \ \approx^{\varepsilon}_{2} \ \bar{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda\cdot)}(z)$$

with $\varepsilon = \sqrt{2\tau}\delta/\lambda$*.*

Proof. It is enough to prove that for all $z \in \mathcal{Y}$

$$\frac{\lambda(z-\hat{p})}{\tau} \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\cdot)(\hat{p}) = \lambda\partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}) \quad \Longleftrightarrow \quad \frac{z-\hat{p}}{\tau} \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}). \tag{6.4}$$

We set $x = (z-\hat{p})/\tau \in \mathcal{Y}$ and consider the function $g : \mathcal{Y} \to \mathbb{R}^d \cup \{+\infty\}$ defined by

$$g(w) = \frac{\mathrm{KL}_{\bar{y}}}{\lambda}(w) \quad \forall\ w \in \mathcal{Y}. \tag{6.5}$$

By the standard property of convex conjugates we have that for any $u \in \mathcal{Y}$

$$g^*(u) = \left(\frac{\mathrm{KL}_{\bar{y}}}{\lambda}\right)^*(u) = \frac{1}{\lambda}\mathrm{KL}^*_{\bar{y}}(\lambda u). \tag{6.6}$$

We now claim that $x \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} g^*(\hat{p})$. To show that, we apply the Young–Fenchel inequality (A.14) of Lemma A.6 to g with $x^* = \hat{p}$. Our objective is thus to show that

$$g(x) + g^*(\hat{p}) \leqslant \langle x, \hat{p}\rangle + \frac{\lambda\varepsilon^2}{2\tau},$$

which, by definitions (6.5) and (6.6) and upon multiplication by λ, coincides with

$$\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}) \leqslant \langle x, \lambda\hat{p}\rangle + \frac{\lambda^2\varepsilon^2}{2\tau}. \tag{6.7}$$

Using the expression of KL and of its convex conjugate given by (A.5), we express the sum on the left-hand side of (6.7) as

$$\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}) = \sum_{i=1}^{d}\left(\bar{y}_i \log\frac{\bar{y}_i}{x_i} - \bar{y}_i + x_i - \bar{y}_i\log(1-\lambda\hat{p}_i)\right). \tag{6.8}$$

Furthermore, by definition of $\hat{p}$, we have that componentwise there holds

$$\frac{\lambda}{\tau}(z_i - \hat{p}_i) \in \lambda\partial\mathrm{kl}^*_{\hat{y}_i}(\lambda\hat{p}_i) \quad \Longleftrightarrow \quad x_i \in \partial\mathrm{kl}^*_{\hat{y}_i}(\lambda\hat{p}_i),$$

which, since $\mathrm{kl}^*_{\hat{y}_i}$ is differentiable (see formula (A.5)), entails that for every $i = 1, \ldots, d$ the element x_i can be written as $x_i = \hat{y}_i/1 - \lambda \hat{p}_i$. Substitute this expression in the formula (6.8) to derive

$$\begin{aligned}
\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda \hat{p}) &= \sum_{i=1}^{d} \underbrace{\bar{y}_i \log \bar{y}_i - \bar{y}_i \log \hat{y}_i - \bar{y}_i + \hat{y}_i}_{\mathrm{kl}(\bar{y}_i;\hat{y}_i)} \\
&\quad + \cancel{\bar{y}_i \log(1 - \lambda \hat{p}_i)} + \underbrace{\big(\hat{y}_i/(1 - \lambda \hat{p}_i)\big)}_{x_i} \lambda \hat{p}_i - \cancel{\bar{y}_i \log(1 - \lambda \hat{p}_i)} \\
&= \mathrm{KL}_{\bar{y}}(\hat{y}) + \langle x, \lambda \hat{p} \rangle \\
&\leqslant \delta^2 + \langle x, \lambda \hat{p} \rangle, \qquad (6.9)
\end{aligned}$$

where the last inequality follows from the perturbation assumption $\mathrm{KL}_{\bar{y}}(\hat{y}) \leqslant \delta^2$. We thus get (6.7) by choosing $\varepsilon = \sqrt{2\tau}\delta/\lambda$, which concludes the proof. □

From Proposition 6.6 and Theorem 5.7, we derive stopping rules for the KL divergence.

COROLLARY 6.7 (early stopping for KL divergence). *Let $\ell_{y_2}(y_1) = \mathrm{KL}(y_2; y_1)$ for every $(y_1, y_2) \in \mathcal{Y}^2$. Assume that the assumptions* (L_1)–(L_3), (R_1)–(R_2), (P_1)–(P_4) *hold true, and suppose that $\lambda_k = \Theta(k^{-\theta})$ with $\theta > 2$. Let $(\hat{x}_k)$ be the sequence generated by* (I3D) *given $\hat{y}$, such that $\hat{y}$ is a δ-perturbation of $\bar{y}$ in the sense of Definition* 6.1. *Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-\frac{2}{3+2\theta}})$ verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{2}{3+2\theta}}) \quad \text{for } \delta \searrow 0. \qquad (6.10)$$

Remark 6.8. It is hard to assess the quality of the rate in (6.10) since the notion of optimality in [42] only applies to additive noise. In the context of Bregman divergences, some analysis has been pursued in [26, section 4.2, estimate (4.3)]. The estimates obtained therein lead to a rate of order $\delta^{1/4}$ for suitable choices of the regularization parameter. In comparison, our estimate (6.10) is sharper and more explicit. Furthermore, as for additive data-fit terms, the use of inertia improves the rates in [43].

Remark 6.9 (the KL divergence does not lead to type 3 errors). The convergence rates for additive data-fit terms proved in Corollary 6.3 are better than the rate for the KL divergence, due to the fact that for the KL divergence we proved that δ-perturbations correspond to type 2 errors, instead of type 3 errors. Indeed, Lemma A.3 in the appendix shows that the error in the evaluation of proximal points for the KL divergence cannot be cast in a type 3 approximation.

7. Conclusions and outlook.

In this paper we proposed an inertial dual diagonal method to solve inverse problems for a wide class of data-fit and regularization terms, possibly corrupted by noise. On the one hand, we established convergence results for both continuous and discrete dynamics. On the other hand, we derived stability results and corresponding early stopping rules, characterizing the regularization properties of the proposed method. A number of open questions are left for future study. It would be interesting to consider a wider class of problems, for example, allowing for regularization terms that are convex but not strongly convex, and possibly nonconvex data fidelity terms. From an algorithmic point of view, it would be interesting to consider alternative approaches, such as stochastic methods. Finally, it would be interesting to investigate the numerical properties of the proposed method for practical problems.

Appendix A. Auxiliary results.

We gather in this appendix some relevant results used in this work.

A.1. Properties of the dual diagonal function.

We first consider $\mathcal{R}_A$, ℓ_y^* defined in (3.3) and on the diagonal dual function d_λ and its limit d_0 defined in (D_λ) and (D_0), respectively. For similar results see also [43].

LEMMA A.1. *Under the assumptions (L_1)–(L_3) and (R_1)–(R_2), we have that the following properties hold:*

(i) *$\mathcal{R}_A$ is differentiable and ∇R_A^* is Lipschitz continuous, with Lipschitz constant equal to $\sigma^{-1}\|A\|^2$.*

(ii) *For all $y \in \mathcal{Y}$, $\ell_y^*(0) = 0$ and $\partial \ell_y^*(0) = \{y\}$.*

(iii) *There holds $\operatorname{argmin} d_0 \neq \emptyset$.*

(iv) *For all $u \in \mathcal{Y}$, the function $\lambda \in [0, +\infty) \mapsto\ d_\lambda(u)$ is nondecreasing.*

(v) *For all $t > 0$, and $u \in \mathcal{Y}$, if $x := \nabla R^*(-A^*u)$, then $\frac{\sigma}{2}\|x - x^\dagger\|^2 \leqslant d_0(u) - \inf d_0$.*

(vi) *For all $u^\dagger \in \operatorname{argmin} d_0$, if $\lambda\|u^\dagger\| \leqslant \frac{\gamma}{q}\varrho^{q-1}$, then*

$$d_\lambda(u^\dagger) - \inf d_0 \leqslant \begin{cases} 0 & \textit{if } q = 1, \\ (1 - \frac{1}{q})\gamma^{-1/(q-1)}\|u^\dagger\|^{q/(q-1)}\lambda^{1/(q-1)} & \textit{if } q > 1. \end{cases} \tag{A.1}$$

Proof. (i) This follows from the strong convexity of R; see, e.g., [22, Theorem 18.15].

(ii) It is a simple consequence of the properties of the Fenchel transform as it can be found, e.g., in [22, Proposition 13.10(i) and Corollary 16.30].

(iii) and (v) These follow from [43, Lemma 5] by simply taking $f = R$ and $g = \delta_{\{\bar{y}\}}$, while property (iv) has been proved in [43, Proposition 2(i)].

(vi) It is enough to verify that $\ell_{\bar{y}}(\cdot)$ is q-well-conditioned in the sense of [43, Definition 1], while assumption (L_3) holds only locally. To check this, we introduce the function $\psi : \mathbb{R} \to \mathbb{R}$ defined for the $\varrho > 0$ appearing in (L_3) by

$$\psi t \mapsto \begin{cases} \frac{\gamma}{q}|t|^q & \text{if } |t| \leqslant \varrho, \\ \frac{\gamma}{q}\varrho^{q-1}|t| & \text{if } |t| > \varrho. \end{cases} \tag{A.2}$$

From (L_3), we easily deduce that $\ell_{\bar{y}}(y) \geqslant \psi(\|y - \bar{y}\|)$ for all $y \in \mathcal{Y}$ (see [61, Corollary 3.4.2]). Note that ψ is not convex for $q > 1$, so in this case we consider instead the function

$$m : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \begin{cases} \frac{\gamma}{q}|t|^q & \text{if } |t| \leqslant q^{1/(1-q)}\varrho, \\ \frac{\gamma}{q}\varrho^{q-1}|t| - \frac{\gamma}{q^{\frac{q}{q-1}}}\varrho^q(1 - \frac{1}{q}) & \text{if } |t| > q^{1/(1-q)}\varrho, \end{cases} \tag{A.3}$$

and define $m := \psi$ for $q = 1$. It is an easy exercise to verify that m is indeed a convex function on $\mathbb{R}$ and that $m(w) \leqslant \psi(w)$ for all $w \in \mathbb{R}$. Now, we can make use of [43, Lemma 2], which tells us that $d_\lambda(u) - \inf d_0 \leqslant \lambda^{-1}m^*(\|u\|\lambda)$. The desired result now follows from the computation of the Fenchel transform of m. If $q = 1$, we have that $m(t) = \gamma|t|$, so classic Fenchel calculus entails that m^* is $\delta_{[-\gamma,\gamma]}$, the indicator function of $[-\gamma,\gamma]$. If $q > 1$, easy computations show that m^* reads

$$\text{(A.4)} \qquad m^* : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \begin{cases} (1-\frac{1}{q})\gamma^{-1/(q-1)}|s|^{\frac{q}{q-1}} & \text{if } |s| \leqslant \frac{\gamma}{q}\varrho^{q-1}, \\ +\infty & \text{if } |s| > \frac{\gamma}{q}\varrho^{q-1}. \end{cases}$$

By now applying [43, Lemma 2] we conclude. ∎

A.2. Useful tools for KL computations. In this section, we report some computations and properties concerning the KL divergence defined in (2.4). For any $(u, y) \in (\mathbb{R}^d)^2$ we define $\mathrm{KL}(y, u)$ as in (2.4) for all $i = 1, \dots, d$. Consider now the functions KL and kl with respect to the first argument only, and define $\mathrm{KL}_y(u) := \mathrm{KL}(y; u)$ and, similarly, its ith component $\mathrm{kl}_{y_i}(u_i)$ for a fixed $y \in \mathbb{R}^d$. The componentwise expression for $\mathrm{KL}_y^*(w) = \sum_{i=1}^d \mathrm{kl}_{y_i}^*(w_i)$ can then be found simply by Fenchel calculus. It reads

$$\text{(A.5)} \qquad \mathrm{kl}_{y_i}^*(w_i) = \begin{cases} -y_i \log(1-w_i) & \text{if } 1-w_i > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Proximal maps. For every $i = 1, \dots, d$, straightforward calculations show that

$$\text{(A.6)} \qquad \mathrm{prox}_{\frac{\tau}{\lambda}\mathrm{kl}_{y_i}}(u_i) = \frac{1}{2}\left(u_i - \frac{\tau}{\lambda} + \sqrt{\left(u_i - \frac{\tau}{\lambda}\right)^2 + 4\frac{\tau}{\lambda}y_i}\right).$$

Furthermore, by applying Moreau's identity we have

$$\text{(A.7)} \qquad \mathrm{prox}_{\frac{\tau}{\lambda}\mathrm{kl}_{y_i}^*(\lambda\cdot)}(w_i) = \frac{1}{2\lambda}\left((1+\lambda w_i) - \sqrt{(1-\lambda w_i)^2 + 4\lambda\tau y_i}\right).$$

The following lemma implies the q-conditioning of the KL divergence.

Lemma A.2 (2-conditioning of the KL data-fit). *Let $\bar{y} \in \mathbb{R}^d$ and $\varrho \in]0, +\infty[$. Then,*

$$\text{(A.8)} \quad (\forall y \in \mathbb{B}(\bar{y}, \varrho)) \quad \mathrm{KL}(\bar{y}, y) \geqslant \left(\frac{1}{\varrho c^2} + \frac{1}{\varrho^2 c}\ln\frac{c}{\varrho + c}\right)\|y - \bar{y}\|^2, \ \textit{where } c = d\|\bar{y}\|_\infty.$$

Proof. Let $y \in \mathbb{B}(\bar{y}, \varrho)$. By [43, Lemma 10.2], we have that

$$\text{(A.9)} \qquad \mathrm{KL}(\bar{y}, y) \geqslant cm(\|y - \bar{y}\|), \ \text{where } m(t) = c^{-1}|t| - \ln\left(1 + c^{-1}|t|\right).$$

To get the desired result, we need to find a quadratic lower bound for m over $[-\varrho, \varrho]$. For simplicity, let us consider the change of variable $s = c^{-1}|t| \in [0, c^{-1}\varrho]$. Since the statement is trivially valid for $y = \bar{y}$, we can assume that $s > 0$ and write

$$s - \ln(1+s) = s^2\phi(s), \ \text{where } \phi(s) := \frac{s - \ln(1+s)}{s^2}.$$

To conclude, we only need to verify that ϕ is decreasing on $]0, +\infty[$. Indeed, this would imply that $m(t) \geqslant c^{-2}t^2\phi(c^{-1}\varrho)$, which together with (A.9) would complete the proof. To see that ϕ is decreasing, we compute explicitly its derivative on $]0, +\infty[$ and see that $\phi'(s) \leqslant 0$ if and only if $\psi(s) := s(s+2) - 2(1+s)\ln(1+s) \geqslant 0$. Combining this with the fact that $\psi(0) = 0$ and that $\psi'(s) = 2(s - \ln(1+s))$ is positive $]0, +\infty[$ we conclude the proof. ∎

The following result deals with the approximation of proximal points of the dual of the KL divergence, corresponding to noise-free and noisy data $\bar{y}$ and $\hat{y}$, respectively. As shown in Proposition 6.6, a type 2 approximation in the sense of Definition 5.3 holds. The following proposition provides a one-dimensional counterexample showing that a type 3 approximation—for which better convergence rates can be obtained—cannot hold.

PROPOSITION A.3. *Let* $w \in \mathbb{R}$ *and* $\bar{y}, \hat{y} \in]0, +\infty[$. *If* $\operatorname{prox}_{\mathrm{kl}^*_{\hat{y}}}(w) \approx^{\varepsilon}_{3} \operatorname{prox}_{\mathrm{kl}^*_{\bar{y}}}(w)$ *holds in the sense of Definition* 5.3 *for some* $\varepsilon > 0$, *then*

$$\varepsilon \geqslant \frac{2|\hat{y} - \bar{y}|}{(1-w) + \sqrt{(1-w)^2 + 4\hat{y}}}.$$

In particular, $\varepsilon \to +\infty$ *when* $w \to +\infty$.

Proof. Let $\varepsilon \geqslant 0$ such that the type 3 approximation property holds. By Definition 5.3, there exists $e \in \mathbb{R}$ such that $|e| \leqslant \varepsilon$ and $\operatorname{prox}_{\mathrm{kl}^*_{\hat{y}}}(w) = \operatorname{prox}_{\mathrm{kl}^*_{\bar{y}}}(w+e)$. Using the formula (A.7), we see that this is equivalent to

$$\frac{1}{2}\left[(1+w) - \sqrt{(1-w)^2 + 4\hat{y}}\right] = \frac{1}{2}\left[(1+w+e) - \sqrt{(1-w-e)^2 + 4\bar{y}}\right] \tag{A.10}$$

and we complete the proof by noting that the above equality is equivalent to

$$e\frac{1}{2}\left[(1-w) + \sqrt{(1-w)^2 + 4\hat{y}}\right] = \bar{y} - \hat{y}. \tag{A.11}$$

□

A.3. Miscellaneous. We here recall some technical lemmas which are used in several sections of the manuscript. The following lemma is useful to characterize the speed of decay of the diagonal term $\lambda(\cdot)$ in assumption (Λ); see also Remark 3.2.

LEMMA A.4. *Let* $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ *a decreasing function such that* $\int_{\mathbb{R}_+} |\lambda(t)|^{1/2}\, dt < +\infty$. *Then, the function* $t \mapsto t\lambda(t)$ *is integrable on* $\mathbb{R}_+$.

Proof. We first show that the function $t \mapsto t\sqrt{\lambda(t)}$ tends to zero as $t \to +\infty$. We have that for every $T > 0$,

$$\int_{T/2}^{+\infty} \sqrt{\lambda(t)}\, dt \geqslant \int_{T/2}^{T} \sqrt{\lambda(t)}\, dt \geqslant \frac{T}{2}\sqrt{\lambda(T)},$$

where the last inequality follows from the decreasing property of λ in the interval $[T/2, T]$. By taking limits, we get the required property:

$$\limsup_{T \to +\infty} \frac{T}{2}\sqrt{\lambda(T)} \leqslant \lim_{T \to +\infty} \int_{T/2}^{+\infty} \sqrt{\lambda(t)}\, dt = 0.$$

Now, from the observation

$$\lim_{t \to +\infty} \frac{t\lambda(t)}{\sqrt{\lambda(t)}} = \lim_{t \to +\infty} t\sqrt{\lambda(t)} = 0,$$

we deduce that there exists some $\overline{T} > 0$ such that $t\lambda(t) \leqslant \sqrt{\lambda(t)}$ for all $t \geqslant \overline{T}$. By thus taking $T > \overline{T}$, we have

$$\int_0^T t\lambda(t)\, dt = \int_0^{\overline{T}} t\lambda(t)\, dt + \int_{\overline{T}}^{T} t\lambda(t)\, dt \leqslant \int_0^{\overline{T}} t\lambda(t)\, dt + \int_{\overline{T}}^{T} \sqrt{\lambda(t)}\, dt,$$

which by taking the supremum over all $T > \bar{T}$ on both sides entails

$$\int_{\mathbb{R}_+} t\lambda(t)\ dt \leqslant \int_0^{\bar{T}} t\lambda(t)\ dt + \int_{\bar{T}}^{+\infty} \sqrt{\lambda(t)}\ dt < +\infty. \qquad \square$$

Next, we state and prove a variant of [8, Lemma 5.14] which we have used in the proof of Theorem 5.8 to get the final stability estimate (5.8).

Lemma A.5. *Let $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(c_k)_{k\in\mathbb{N}}$ be positive sequences, and assume that c_k is increasing. If*

$$(\forall k \in \mathbb{N}) \quad a_k^2 \leqslant c_k + \sum_{j=1}^{k-1} b_j a_{j+1}, \tag{A.12}$$

then $\max_{j=1,\ldots,k} a_j \leqslant \sqrt{c_k} + \sum_{j=1}^{k-1} b_j$ for every $k \in \mathbb{N}$.

Proof. Take $k \in \mathbb{N}$, and let $A_k := \max_{m=1,\ldots,k} a_m$. Then, for all $1 \leqslant m \leqslant k$,

$$a_m^2 \leqslant c_m + \sum_{j=1}^{m-1} b_j a_{j+1} \leqslant c_k + A_k \sum_{j=1}^{k-1} b_j, \tag{A.13}$$

because c_k is increasing and b_j is positive. Therefore $A_k^2 \leqslant c_k + A_k \sum_{j=1}^{k-1} b_j$. Define $S_k = \sum_{j=1}^{k-1} b_j$. By computing and bounding the solutions of the previous inequality we conclude that

$$A_k \leqslant \frac{S_k + \sqrt{S_k + 4c_k}}{2} \leqslant S_k + \sqrt{c_k}. \qquad \square$$

We recall a useful characterization of the elements in the ε-subdifferential of a function in $\Gamma_0(\mathcal{H})$. This property is used to prove Proposition 6.6; see also [61].

Lemma A.6 (see [61, Theorem 2.4.2]). *Let $\mathcal{H}$ be a Hilbert space, let $f \in \Gamma_0(\mathcal{H})$, let $(x, u) \in \mathcal{H}^2$, and let $\varepsilon > 0$. Then, the following statements are equivalent:*

(i) $u \in \partial_\varepsilon f(x)$.

(ii) *The following ε-Young–Fenchel inequality holds:*

$$f(x) + f^*(u) \leqslant \langle u, x\rangle + \varepsilon. \tag{A.14}$$

(iii) $x \in \partial_\varepsilon f^*(u)$.

REFERENCES

[1] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] V. Apidopoulos, J.-F. Aujol, and C. Dossal, *Convergence rate of inertial forward–backward algorithm beyond Nesterov's rule*, Math. Program., 180 (2020), pp. 137–156.

[3] V. Apidopoulos, J.-F. Aujol, and C. Dossal, *The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leqslant 3$*, SIAM J. Optim., 28 (2018), pp. 551–574.

[4] A. Aravkin, J. Burke, D. Drusvyatskiy, M. Friedlander, and S. Roy, *Level-set methods for convex optimization*, Math. Program., 174 (2019), pp. 359–390.

[5] H. Attouch, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.

[6] H. Attouch, A. Cabot, Z. Chbani, and H. Riahi, *Inertial forward–backward algorithms with perturbations: Application to Tikhonov regularization*, J. Optim. Theory Appl., 179 (2018), pp. 1–36.

[7] H. Attouch, A. Cabot, and M.-O. Czarnecki, *Asymptotic behavior of nonautonomous monotone and subgradient evolution equations*, Trans. Amer. Math. Soc., 370 (2018), pp. 755–790.

[8] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program., 168 (2018), pp. 123–175.

[9] H. Attouch, Z. Chbani, and H. Riahi, *Combining fast inertial dynamics for convex optimization with Tikhonov regularization*, J. Math. Anal. Appl., 457 (2018), pp. 1065–1094.

[10] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leqslant 3$, ESAIM Control Optim. Calc. Var., 25 (2019).

[11] H. Attouch and R. Cominetti, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.

[12] H. Attouch, M. Czarnecki, and J. Peypouquet, *Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities*, SIAM J. Optim., 21 (2011), pp. 1251–1274.

[13] H. Attouch and M.-O. Czarnecki, *Asymptotic behavior of coupled dynamical systems with multiscale aspects*, J. Differential Equations, 248 (2010), pp. 1315–1344.

[14] H. Attouch and M.-O. Czarnecki, *Asymptotic behavior of gradient-like dynamical systems involving inertia and multiscale aspects*, J. Differential Equations, 262 (2017), pp. 2745–2770.

[15] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM J. Optim., 26 (2016), pp. 1824–1834.

[16] H. Attouch, J. Peypouquet, and P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM J. Optim., 24 (2014), pp. 232–256.

[17] J. Aujol and C. Dossal, *Stability of over-relaxations for the forward-backward algorithm, application to FISTA*, SIAM J. Optim., 25 (2015), pp. 2408–2433.

[18] F. R. Bach, *Exploring large feature spaces with hierarchical multiple kernel learning*, in Advances in Neural Information Processing Systems, 2009, pp. 105–112.

[19] M. Bachmayr and M. Burger, *Iterative total variation schemes for nonlinear inverse problems*, Inverse Problems, 25 (2009).

[20] M. A. Bahraoui and B. Lemaire, *Convergence of diagonally stationary sequences in convex optimization*, Set-Valued Anal., 2 (1994), pp. 49–61.

[21] A. B. Bakushinsky and M. Y. Kokurin, *Iterative Methods for Approximate Solution of Inverse Problems*, Math. Appl. 577, Springer, New York, 2005.

[22] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Math., Springer, New York, 2017.

[23] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167 – 175.

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

[25] A. Beck and M. Teboulle, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett., 42 (2014), pp. 1–6.

[26] M. Benning and M. Burger, *Error estimates for general fidelities*, Electron. Trans. Numer. Anal., 38 (2011), pp. 44–68.

[27] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numer., 27 (2018), pp. 1–111.

[28] E. V. D. Berg and M. P. Friedlander, *Probing the Pareto frontier for basis pursuit solutions*, SIAM J. Sci. Comput., 31 (2009), pp. 890–912.

[29] R. I. Boţ and T. Hein, *Iterative regularization with a general penalty term—theory and application to L1 and TV regularization*, Inverse Problems, 28 (2012).

[30] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM J. Imaging Sci., 3 (2010), pp. 492–526.

[31] P. Brianzi, F. Di Benedetto, and C. Estatico, *Preconditioned iterative regularization in banach spaces*, Comput. Optim. Appl., 54 (2013), pp. 263–282.

[32] M. Burger and S. Osher, *A guide to the TV zoo*, in Level Set and PDE Based Reconstruction Methods in Imaging, Lecture Notes in Math. 2090, Springer, New York, 2013, pp. 1–70.

[33] M. Burger, E. Resmerita, and L. He, *Error estimation for Bregman iterations and inverse scale space methods in image restoration*, Computing, 81 (2007), pp. 109–135.

[34] A. Cabot and L. Paoli, *Asymptotics for some vibro-impact problems with a linear dissipation term*, J. Math. Pures Appl., 87 (2007), pp. 291–323.

[35] L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb, *Infimal convolution of data discrepancies for mixed noise removal*, SIAM J. Imaging Sci., 10 (2017), pp. 1196–1233.

[36] A. Chambolle and C. Dossal, *On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm,"* J. Optim. Theory Appl., 166 (2015), pp. 968–982.

[37] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Mathe., 76 (1997), pp. 167–188.

[38] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[39] M.-O. Czarnecki, N. Noun, and J. Peypouquet, *Splitting forward-backward penalty scheme for constrained variational problems*, J. Convex Anal., 23 (2016), pp. 531–565.

[40] C.-A. Deledalle, S. Vaiter, J. M. Fadili, and G. Peyré, *Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection*, SIAM J. Imaging Sci., 7 (2014), pp. 2448–2487.

[41] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Math. Program., 146 (2014), pp. 37–75.

[42] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Springer, New York, Media, 1996.

[43] G. Garrigos, L. Rosasco, and S. Villa, *Iterative regularization via dual diagonal descent*, J. Math. Imaging Vis., 60 (2018), pp. 189–215.

[44] M. Hintermüller and A. Langer, *Subspace correction methods for a class of nonsmooth and nonadditive convex variational problems with mixed l^1/l^2 data-fidelity in image processing*, SIAM J. Imaging Sci., 6 (2013), pp. 2134–2173.

[45] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, De Gruyter, Berlin, 2008.

[46] B. Kaltenbacher, F. Schöpfer, and T. Schuster, *Iterative methods for nonlinear ill-posed problems in banach spaces: Convergence and applications to parameter identification problems*, Inverse Problems, 25 (2009).

[47] B. Kaltenbacher and I. Tomba, *Convergence rates for an iteratively regularized Newton–Landweber iteration in banach space*, Inverse Problems, 29 (2013).

[48] W. Krichene, A. Bayen, and P. L. Bartlett, *Accelerated Mirror Descent in Continuous and Discrete Time*, in Advances in Neural Information Processing Systems, 2015, pp. 2827–2835.

[49] S. Matet, L. Rosasco, S. Villa, and B. L. Vu, *Don't Relax: Early Stopping for Convex Regularization*, preprint, https://arxiv.org/abs/1707.05422, 2017.

[50] Y. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Sov. Math. Dokl., 269 (1983), pp. 543–547.

[51] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Appl. Optim. 87, Springer, New York, 2004.

[52] A. Neubauer, *On Nesterov acceleration for landweber iteration of linear ill-posed problems*, J. Inverse Ill-Posed Problems, 25 (2016).

[53] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[54] S. Salzo and S. Villa, *Inexact and accelerated proximal point algorithms*, J. Convex Anal., 19 (2012), pp. 1167–1192.

[55] M. Schmidt, N. L. Roux, and F. R. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, 2011, pp. 1458–1466.

[56] C. M. Stein, *Estimation of the mean of a multivariate normal distribution*, Ann. Statist., 9 (1981), pp. 1135–1151.

[57] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York, 2008.

[58] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.

[59] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, *Accelerated and inexact forward-backward algorithms*, SIAM J. Optim., 23 (2013), pp. 1607–1633.

[60] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Stat. Method., 68 (2006), pp. 49–67.

[61] C. Zalinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

[62] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2005), pp. 301–320.