# Weakly Convex Regularisers for Inverse Problems: Convergence of Critical Points and Primal-Dual Optimisation

Zakhar Shumaylov [1]   Jeremy Budd [2]   Subhadip Mukherjee [3]   Carola-Bibiane Schönlieb [1]

## Abstract

Variational regularisation is the primary method for solving inverse problems, and recently there has been considerable work leveraging deeply learned regularisation for enhanced performance. However, few results exist addressing the convergence of such regularisation, particularly within the context of critical points as opposed to global minimisers. In this paper, we present a generalised formulation of convergent regularisation in terms of critical points, and show that this is achieved by a class of weakly convex regularisers. We prove convergence of the primal-dual hybrid gradient method for the associated variational problem, and, given a Kurdyka–Łojasiewicz condition, an $\mathcal{O}(\log k/k)$ ergodic convergence rate. Finally, applying this theory to learned regularisation, we prove universal approximation for input weakly convex neural networks (IWCNN), and show empirically that IWCNNs can lead to improved performance of learned adversarial regularisers for computed tomography (CT) reconstruction.

## 1. Introduction

In an inverse problem, one seeks to recover an unknown parameter (e.g., an image) that has undergone some, typically lossy and noisy, transformation (e.g., during measurement). Such problems arise frequently in science, e.g. in medical imaging such as in MRI, CT, and PET, and also beyond science, e.g. in art restoration (Natterer, 2001; Scherzer et al., 2009; Calatroni et al., 2018).

Mathematically, one seeks to estimate an unknown parame-

[1]University of Cambridge, UK [2]California Institute of Technology, USA [3]Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, India. Correspondence to: Zakhar Shumaylov <zs334@cam.ac.uk>.
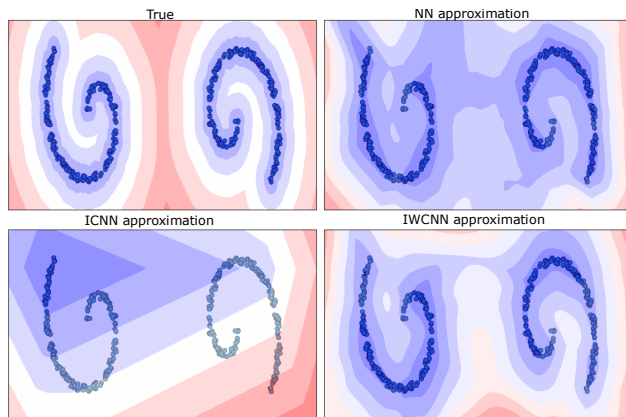
Figure 1: Contour plots comparing the distance function to a data manifold (top-left) with learned regularisers for denoising this data, via convex (bottom-left), standard (top-right), and weakly convex (bottom-right) adversarial regularisation. The convex regulariser lacks an interpretation as a distance function, whilst the weakly convex regulariser, introduced in this work, retains this feature and shows improved generalisation.

ter $x^* \in \mathcal{X}$ from a transformed and noisy measurement

$$y^\delta = \mathcal{A} x^* + \eta \in \mathcal{Y}, \tag{1.1}$$

see (Engl et al., 1996). Here $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces; $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is the *forward model*, assumed to be linear and bounded, e.g. representing imaging physics; and $\eta \in \mathcal{Y}$, with $\|\eta\|_\mathcal{Y} \leqslant \delta$, describes measurement noise. However, (1.1) is usually *ill-posed*, i.e. $x \in \mathcal{X}$ with $\mathcal{A} x = y^\delta$ may not exist, be unique, or be continuous in $y^\delta$.

An influential technique to overcome this ill-posedness has been the variational approach pioneered by (Tikhonov, 1963) and (Phillips, 1962):

$$x^\alpha(y^\delta) \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ \alpha \mathcal{R}(x) + \mathcal{D}\left(\mathcal{A} x, y^\delta\right) =: \mathcal{J}_{\alpha, y^\delta}(x) \right\}. \tag{1.2}$$

Here: $\mathcal{R} : \mathcal{X} \to [0, \infty)$ is the *regulariser*, which aims to incorporate prior knowledge about $x^*$ and penalise $x$ if it is not 'realistic'; $\mathcal{D} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is the *data-fidelity term*, which quantifies the distance between $y^\delta$ and $\mathcal{A} x$ and comes from the distribution of the noise $\eta$; and $\alpha > 0$ balances the relative importance of the two terms. A regularisation

method is *well-defined* if $x^\alpha(y^\delta)$ is stable with respect to data perturbations and converges, as $\delta \downarrow 0$, to a solution of the noiseless operator equation $\mathcal{A}\,x = y^0 := \mathcal{A}\,x^*$.

A good choice of regulariser $\mathcal{R}$ is important for achieving accurate results. Traditional regularisation was *knowledge-driven*: the regulariser was a hand-crafted functional designed to encourage the reconstruction $x$ to have structures known to be realistic. Over the past decades, a whole zoo of such regularisers have been proposed, see e.g. (Rudin et al., 1992; Bredies et al., 2010; Gilboa & Osher, 2009; Mallat, 2009; Kutyniok & Labate, 2012; Lassas et al., 2009; Davoli et al., 2023) and see (Benning & Burger, 2018) for a detailed overview. However, there is a limit to the complexity of patterns that can be described by a hand-crafted functional.

Seeking to overcome this limitation, there has been considerable recent interest in *data-driven* approaches to inverse problems, see (Arridge et al., 2019) for a detailed overview. A key data-driven technique is to learn the regulariser from data. Such methods include: dictionary learning (see (Chen & Needell, 2016) for an overview), plug-and-play priors (Venkatakrishnan et al., 2013; Kamilov et al., 2023), regularisation by denoising (Romano et al., 2017), deep image priors (Lempitsky et al., 2018; Dittmer et al., 2019), network Tikhonov (NETT) (Li et al., 2020), total deep variation (Kobler et al., 2020), generative regularisation (see (Dimakis, 2022) for an overview), and adversarial regularisation (Lunz et al., 2018). These methods typically outperform knowledge-driven regularisation in practice.

One of the major advantages of the learned regularisation paradigm compared to *end-to-end* learning methods for solving inverse problems, e.g. (Jin et al., 2017; Zhu et al., 2018; Adler & Öktem, 2018), is that it can operate in the *weakly supervised* setting, i.e. when one has data of ground truths (and measurements) but not of ground-truth-measurement pairs. This setting is often more realistic, especially in medical imaging where paired data can be hard to acquire. Furthermore, because they do not required paired training data, learned regularisation methods can adapt to other forward models without requiring retraining (see e.g. (Sun et al., 2020; Lunz, 2021)).

### 1.1. Motivation and contributions

A key challenge in the data-driven paradigm is proving guarantees about the behaviour of the learned models, both for the inverse problem and for optimisation methods for solving it. In (Mukherjee et al., 2023) the role of convexity, via the input convex neural network architecture (Amos et al., 2017), was emphasised as a method for achieving such guarantees, e.g. in the case of the adversarial convex regulariser (ACR) (Mukherjee et al., 2021). But convexity is quite a restrictive condition, and the ACR sacrifices the adversarial regulariser's interpretability as a distance function to the

data manifold (see Figure 1), contributing to worse performance. It has been observed in the literature that non-convex regularisers often have better performance (Mohimani et al., 2009; Roth & Black, 2009; Pieper & Petrosyan, 2022). Of particular interest is recent work on 'optimal' regularisation (Alberti et al., 2021; Leong et al., 2023), as such optimal regularisers are typically non-convex. This non-convexity returns us to a setting in which provable guarantees are challenging to achieve. However, the optimal regulariser of (Alberti et al., 2021) is often weakly convex, and weak convexity offers a more structured setting for reasoning about the non-convex case, see e.g. (Pinilla et al., 2022).

Optimisation methods in the context of weak convexity can be split into either subgradient or proximal based methods; see (Drusvyatskiy & Davis, 2020) for an introductory review. Subgradient methods admit strong iteration complexity guarantees (Davis et al., 2018; Davis & Drusvyatskiy, 2018; Goujon et al., 2024). However, the convergence settings for these methods are somewhat narrow. Proximal methods admit significantly improved complexity bounds, as well as objective convergence rates (Hurault et al., 2022). These rates can be further improved through prox-linearisation (Drusvyatskiy & Paquette, 2019; Davis & Drusvyatskiy, 2022) or adaptive steps (Böhm & Wright, 2021).

Of particular interest for optimising (1.2) is the proximal primal-dual method, first analysed for convex functions in (Chambolle & Pock, 2011). The first guarantees for primal-dual in the context of weak convexity were shown in (Möllenhoff et al., 2015), which assumed the (1.2) to be convex, with a unique minimiser. In the non-convex case, however, rather stringent assumptions had to be made. This was overcome in the Arrow–Hurwicz (Arrow et al., 1958) special case in (Sun et al., 2018), via a Kurdyka–Łojasiewicz assumption. In (Guo et al., 2023) a preconditioned primal-dual method was proposed for smooth primal functions, and convergence achieved under operator surjectivity. But for general weakly convex functions, no such convergence was shown. We are therefore motivated to fill this gap.

Inspired by the optimisation literature, there has been recent interest in learning weakly convex regularisers (Hurault et al., 2022; Goujon et al., 2024; Shumaylov et al., 2023). In (Shumaylov et al., 2023), an input weakly convex neural network (IWCNN) architecture was introduced, and incorporated into the adversarial regularisation framework to learn a convex-nonconvex regulariser (see (Lanza et al., 2022)) with provable guarantees. However, existing convergent regularisation guarantees focus on the convergence **of** global minimisers. In practice, optimisation schemes for (1.2) do not converge **to** global minimisers, but only **to** critical points, and as such one should consider convergence **of** critical points.

There exist only a few works focused on stability and con-

vergence of regularisation in the sense of critical points. Under injectivity of $\mathcal{A}$, (Durand & Nikolova, 2006) prove stability guarantees in the finite dimensional setting. In (Obmann & Haltmeier, 2022) stability and convergence was shown for so-called $\phi$-critical points, defined in terms of $\phi$-subgradients. But these $\phi$-critical points are *not* true critical points of $\mathcal{J}_{\alpha,y^\delta}$. They can be interpreted as bounded points, with bound quality depending critically on choice of $\phi$, motivating a return to the usual notions of subgradients. For exact definitions and further discussion, the interested reader is referred to (Obmann & Haltmeier, 2022; 2024). The setting of critical points of $\mathcal{J}_{\alpha,y^\delta}$ is a special case of (Obmann & Haltmeier, 2023), which showed stability and convergence assuming weak-to-weak continuity and appropriate coercivity of the gradient of the regulariser. But in the context of learned regularisation, differentiability has empirically been observed to worsen performance.

In this study, we show that imposing weak convexity on the regulariser achieves the best of both worlds. It yields guarantees for both inverse problems and optimisation, whilst also leading to robust numerical performance. We illustrate this from three sides. On the inverse problem front we:

  i. Prove guarantees about existence, in Theorem 3.2, and stability, in Theorem 3.3, of solutions to (1.2).

 ii. Formulate convergent regularisation in terms of critical points in a generalised way in Definition 2.4, and prove that it is achieved by a class of weakly convex regularisers in Theorem 3.4.

On the optimisation front we:

iii. Prove convergence guarantees of primal-dual methods for solving (1.2) under weak convexity in Theorem 4.2.

 iv. Prove an $\mathcal{O}(\log k/k)$ ergodic convergence rate for the primal dual scheme under KŁ in Theorem 4.6.

On the learned regularisation front we:

  v. Prove a universal approximation theorem for IWCNNs in Theorem 5.2.

 vi. Define an adversarial weak convex regulariser (AWCR) in Definition 5.5 and corroborate the theoretical results via numerical experiments with AWCRs for CT reconstruction, shown in Table 1 and Figure 2.

## 2. Groundwork

Throughout this paper, $\mathcal{X}$ and $\mathcal{Y}$ denote Banach spaces.

**Definition 2.1** ($\rho$-convexity)**.** Let $f : \mathcal{X} \to \mathbb{R}$. Then $f$ is said to be $\rho$-convex if there exists some $\rho \in \mathbb{R}$ such that

$f - \frac{1}{2}\rho\| \cdot \|_{\mathcal{X}}^2$ is convex. Then $f$ is said to be $\rho$-strongly convex if $\rho > 0$, convex if $\rho = 0$, and $(-\rho)$-weakly convex if $\rho < 0$. Hence, strongly convex entails convex, which entails weakly convex.

**Definition 2.2** (Subdifferential; see e.g. (Kruger, 2003))**.** Let $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$. Then the *(Frechét) subdifferential* of $f$ at $x$ is the empty set if $f(x) = \infty$, and if $f(x) < \infty$ is

$$\partial f(x) := \{\psi \in \mathcal{X}^* :$$
$$f(x') \geqslant f(x) + \psi(x' - x) + o(\|x' - x\|_{\mathcal{X}}) \text{ as } x' \to x\},$$

where $\mathcal{X}^*$ denotes the dual space of $\mathcal{X}$, i.e. the space of all continuous linear functionals on $\mathcal{X}$.

**Note 2.3.** If $x$ is a local minimiser of $f$, then $0 \in \partial f(x)$. If $f$ is (Frechét) differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$.

In what follows we are going to be interested in critical points of the objective function $\mathcal{J}_{\alpha,y^\delta}$. For this we define the set of critical points of $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ as:

$$\operatorname*{crit}_{x \in \mathcal{X}} f := \{x \in \mathcal{X} : 0 \in \partial f(x)\}.$$

For further details on convex analysis, see Appendix A.

### 2.1. Convergent regularisation

The idea behind convergent regularisation is that we do not want to over-regularise or regularise in the wrong way. We want to guarantee that as we tune the noise level $\delta$ down, we can also tune down the regularisation parameter $\alpha$ at an appropriate rate such that all limit points of the resulting reconstructions $x^\alpha(y^\delta)$ are regular and data-consistent. To make this precise, we make the following definitions.

**Definition 2.4** ($\mathcal{R}$-minimising and $\mathcal{R}$-criticising solutions)**.** Let $y^0 := \mathcal{A}\,x^*$ be the clean measurement. If

$$x^\dagger \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathcal{R}(x) \text{ subject to } \mathcal{A}\,x = y^0,$$

then following (Pöschl, 2008) we call $x^\dagger$ an $\mathcal{R}$-*minimising solution*, and considering critical points if

$$x^\dagger \in \operatorname*{crit}_{x \in \mathcal{X}} \mathcal{R}(x) + \iota_{\{0\}}(\mathcal{A}\,x - y^0)$$

then we call $x^\dagger$ an $\mathcal{R}$-*criticising solution*, where the indicator function

$$\iota_{\{0\}}(z) := \begin{cases} 0, & \text{if } z = 0, \\ \infty, & \text{otherwise,} \end{cases}$$

imposes the condition $\mathcal{A}\,x = y^0$. In particular, if $x^\dagger$ is $\mathcal{R}$-minimising, then it is also $\mathcal{R}$-criticising.

Traditionally in inverse problems, convergence of regularisation has been studied in the sense of conditions under which,

as $\delta \to 0$, there exist $\alpha(\delta) \to 0$ such that the limit points of the *global minimisers* of $\mathcal{J}_{\alpha(\delta),y^\delta}$ are $\mathcal{R}$-minimising solutions. I.e., convergence **of** the minimisers. But in practice optimisation methods rarely converge **to** global minimisers when $\mathcal{J}_{\alpha,y^\delta}$ is non-convex, converging instead **to** other critical points. Thus, in this work we will instead ask whether the limit points **of** the *critical points* of $\mathcal{J}_{\alpha(\delta),y^\delta}$ are $\mathcal{R}$-criticising solutions.

Well-defined regularisation via global minimisers is typically shown by assuming the following pre-compactness condition (Grasmair, 2010; Pöschl, 2008): For all $\alpha > 0$, $y \in \mathcal{Y}$, and $t \in \mathbb{R}$, the set $\{x \in \mathcal{X} : \mathcal{J}_{\alpha,y}(x) \leqslant t\}$ is sequentially pre-compact. In particular, this is satisfied by coercive $\mathcal{R}$, though coercivity may not be necessary, see (Lorenz & Worliczek, 2013). But this is insufficient for well-defined regularisation via critical points, as we now illustrate.

**Example 2.5.** Inspired by the example in (Obmann & Haltmeier, 2022), we consider $\mathcal{R}(x) = |x| + \cos(x)$, $\mathcal{A} = 0$, $\alpha > 0$, and a sequence of measurements $y_k = 0$. In this case the sequence $x_k = 2\pi k + \frac{\pi}{2}$ is such that each $x_k$ is a critical point of $\mathcal{J}_{\alpha,y_k}(x) := \alpha\mathcal{R}(x) + \mathcal{D}(\mathcal{A}x, y_k)$, but $(x_k)$ has no convergent subsequence. Yet $|x| + \cos(x)$ is coercive, and can be checked to be weakly convex. Therefore, unlike in the global minimisers setting, regulariser coercivity does not guarantee stability or convergence.

## 2.2. Set-up and bounding the critical points

We are interested in Tikhonov functionals $\mathcal{J}_{\alpha,y^\delta} : \mathcal{X} \to [0, \infty)$. Instead of working with global minimisers of $\mathcal{J}_{\alpha,y^\delta}$, we consider regularised solutions $x_\alpha^\delta \in \text{crit } \mathcal{J}_{\alpha,y^\delta}$ as critical points of $\mathcal{J}_{\alpha,y^\delta}$ We wish to construct non-convex regularisers which still achieve convergent regularisation in terms of critical points. By the discussion above, we will need to make assumptions on $\mathcal{R}$ which bound the set of critical points.

**Theorem 2.6** (Bounded critical points). *Let $\mathcal{R} = R_{wc} + R_{sc}$ where $R_{wc} : \mathcal{X} \to [0, \infty)$ is $\gamma$-weak convex and $R_{sc}$ is $\mu$-strongly convex. For any $\hat{x}$ a critical point of $\mathcal{R}$, we have that for all $z \in \mathcal{X}$: If $\gamma < 2\mu$,*

$$\|\hat{x} - z\|_{\mathcal{X}} \leqslant \frac{1}{(\mu - \frac{\gamma}{2})}\|\partial R_{sc}(z)\| + \sqrt{\frac{1}{(\mu - \frac{\gamma}{2})}R_{wc}(z)},$$

*or if $R_{wc}$ is $L_R$-Lipschitz continuous,*

$$\|\hat{x} - z\|_{\mathcal{X}} \leqslant \frac{1}{\mu}\left(L_R + \|\partial R_{sc}(z)\|\right).$$

*Here $\|\partial R_{sc}(z)\| := \sup\{\|\psi\|_{\mathcal{X}^*} : \psi \in \partial R_{sc}(z)\}$.*

*Proof.* Proven in Appendix B. □

**Note 2.7.** Note that $\mathcal{R}$ is $(\mu - \gamma)$-convex. Thus, $\mathcal{R}$ can posses multiple minimisers when $\mu \leqslant \gamma < 2\mu$. This convexity bound is optimal, up to equality, and we provide an

intuitive explanation for it, constructing an example with infinitely many unbounded minimisers in Appendix C.

## 3. Inverse problem guarantees

In this section, we show that weakly convex regularisation is well-defined, under the following assumptions, which are standard in the literature (see (Obmann & Haltmeier, 2022; Grasmair, 2010; Pöschl, 2008)), except for Assumption 3.1(3b), which is imposed to avoid a counterexample (see Appendix D.3) that arises in the infinite-dimensional setting.

**Assumption 3.1.** (1) $\mathcal{X}$ is a reflexive Banach space.

(2) $\mathcal{R}$ is weakly sequentially l.s.c.

(3) $\mathcal{R} = R_{wc} + R_{sc}$, where $R_{wc} : \mathcal{X} \to [0, \infty)$ is $\gamma$-weak convex, $R_{sc}$ is $\mu$-strongly convex, and either

   (a) $\gamma \leqslant \mu$ (i.e., $\mathcal{R}$ is convex), or
   (b) $\mu < \gamma < 2\mu$, and $R_{sc} - \frac{1}{2}\mu\|\cdot\|_{\mathcal{X}}^2$ is weakly sequentially l.s.c.

   This, in particular, means that $\mathcal{R}$ is coercive, and therefore so is $\mathcal{J}_{\alpha,y^\delta}$, since $\mathcal{D}$ is non-negative.

(4) $\mathcal{D}$ is weakly sequentially l.s.c., convex in its first argument, continuous in its second argument, and $\mathcal{D}(y_1, y_2) = 0$ if and only if $y_1 = y_2$.

(5) There exist $C > 0$ and $p \geqslant 1$ s.t. for all $y_1, y_2, y_3 \in \mathcal{Y}$, $\mathcal{D}(y_1, y_2) \leqslant C\left(\mathcal{D}(y_1, y_3) + \|y_2 - y_3\|_{\mathcal{Y}}^p\right)$.

### 3.1. Existence and stability of solutions

**Theorem 3.2** (Existence). *Under Assumption 3.1 solutions exist, i.e. for all $\alpha > 0$ and $y^\delta \in \mathcal{Y}$, crit $\mathcal{J}_{\alpha,y^\delta}$ is non-empty.*

*Proof.* Existence of minimisers of $\mathcal{J}_{\alpha,y^\delta}$ follows from the coercivity and the continuity assumptions on $\mathcal{J}_{\alpha,y^\delta}$. □

**Theorem 3.3** (Stability). *Let $y_k, y^\delta \in \mathcal{Y}$, $\alpha > 0$, and $y_k \to y^\delta$ (in norm), and assume that $x_k \in \mathcal{X}$ is such that $x_k \in$ crit $\mathcal{J}_{\alpha,y_k}$. Then under Assumption 3.1 the sequence $(x_k)$ has a weakly convergent subsequence and the weak limit $x_+$ of any such subsequence is a critical point of $\mathcal{J}_{\alpha,y^\delta}$.*

*Proof.* Proven in Appendix D.1. □

In words: the reconstruction from a measurement at a given noise level (i.e., a critical point of the variational energy) is (weakly) continuous with respect to perturbations of the measurement (up to subsequences).

## 3.2. Convergent regularisation

**Theorem 3.4.** *Let Assumption 3.1 hold. Let $y^0 \in \mathcal{Y}$ and assume that $y^0 = \mathcal{A} u$ for some $u \in \mathcal{X}$. Let $\delta_k \downarrow 0$, and choose $\alpha = \alpha(\delta)$ such that for $\alpha_k = \alpha(\delta_k)$ we have $\lim_k \alpha_k = \lim_k \delta_k^p / \alpha_k = 0$, where $p$ is the same $p$ from Assumption 3.1(5). Let $y^{\delta_k} \in \mathcal{Y}$ be any sequence of measurements with $\|y^{\delta_k} - y^0\|_{\mathcal{Y}} \leqslant \delta_k$, and let $x_k \in \mathcal{X}$ satisfy $x_k \in \text{crit } \mathcal{J}_{\alpha_k, y^{\delta_k}}$. Then the sequence $(x_k)$ has a weakly convergent subsequence and the weak limit $x_+$ of any such subsequence is an $\mathcal{R}$-criticising solution. If there is a unique $u \in \mathcal{X}$ such that $\mathcal{A} u = y^0$, then $x_k \rightharpoonup u$.*

*Proof.* Proven in Appendix D.2. □

In words: in the limit as the noise level vanishes, there is a regularisation parameter selection strategy under which reconstructions from measurements at each noise level converge to a solution of the noiseless operator equation, up to a subsequence unless such a solution is unique.

**Note 3.5.** Theorems 3.3 and 3.4 do not hold in the general setting where $\mathcal{R}$ is weakly convex without Assumption 3.1(2,3), even assuming that $\mathcal{R}$ is globally Lipschitz; for a counterexample see Appendix D.3.

**Note 3.6.** If $\mathcal{X}$ is a *separable* reflexive Banach space, then the subsequences in Theorems 3.3 and 3.4 can be explicitly constructed, as the Banach–Alaoglu theorem has a constructive proof in this special case. We omit the details.

**Note 3.7.** If we make the further assumption that $\mathcal{X}$ is a finite-dimensional Hilbert space then Theorems 3.3 and 3.4 hold with strong convergence of the iterates, with Assumption 3.1(3) replaced with that $R_{wc}$ is globally Lipschitz or that $\gamma < 2\mu$. In the remainder of this paper, we will assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional real Hilbert spaces.

## 4. Optimisation guarantees

In this section, we analyse the primal-dual algorithm in the weakly convex setting. All results are in the setting of Assumption 3.1 with the modification described in Note 3.7.

### 4.1. Primal-dual optimisation

The idea of primal-dual optimisation is to reformulate (1.2) as a minimax problem. First, we rewrite (1.2) as

$$\min_{x \in \mathcal{X}} \mathcal{R}(x) + F(\mathcal{A} x), \qquad (4.1)$$

where $F(y) := \mathcal{D}(y, y^\delta)$. By Assumption 3.1(4), $F$ is convex and l.s.c., so by Theorem A.2 we can rewrite (4.1) as the minimax problem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} L(x, y) := \mathcal{R}(x) + \langle \mathcal{A} x, y \rangle_{\mathcal{Y}} - F^*(y).$$

This can then be solved via the modified primal-dual hybrid gradient method (PDHGM) due to (Chambolle & Pock,

2011). The method consists of the following updates for step sizes $\tau, \sigma > 0$:

$$x^{k+1} := \underset{x \in \mathcal{X}}{\arg\min} \left\{ \mathcal{R}(x) + \langle y^k, \mathcal{A} x \rangle_{\mathcal{Y}} + \frac{1}{2\tau} \left\| x - x^k \right\|_{\mathcal{X}}^2 \right\},$$

$$x_\vartheta^{k+1} := x^{k+1} + \vartheta \left( x^{k+1} - x^k \right), \qquad (4.2)$$

$$y^{k+1} := \underset{y \in \mathcal{Y}}{\arg\min} \left\{ F^*(y) - \langle y, \mathcal{A} x_\vartheta^{k+1} \rangle_{\mathcal{Y}} + \frac{1}{2\sigma} \left\| y - y^k \right\|_{\mathcal{Y}}^2 \right\}.$$

Note that for non-convex $\mathcal{R}$, the first update may not be unique for all choices of $\tau$. However for $\rho$-weakly convex $\mathcal{R}$ it is unique for $\tau < 1/\rho$, as shown in Theorem A.4. This method, in the case of non-convexity and KŁ (see Appendix E.3) was analysed for $\vartheta = 0$ in (Sun et al., 2018). Here, we are interested in $\vartheta = 1$ similar to (Chambolle & Pock, 2011), due to connections to the proximal point method and the alternating direction method of multipliers (Lu & Yang, 2023). We will use weak convexity of $\mathcal{R}$ and $\mu$-strong convexity of $F^*$, satisfied, e.g., when $F$ is convex and $1/\mu$-smooth.

### 4.2. Convergence of PDHGM

First we rewrite our problem in a nice form in analogy with (Lu & Yang, 2023). Let $z := (x, y)$ and define

$$T(z) := \begin{pmatrix} \partial_x L(x, y) \\ -\partial_y L(x, y) \end{pmatrix} = \begin{pmatrix} \partial \mathcal{R}(x) + \mathcal{A}^* y \\ \partial F^*(y) - \mathcal{A} x \end{pmatrix},$$

where $\mathcal{A}^* : \mathcal{Y} \to \mathcal{X}$ is the adjoint of $\mathcal{A}$, i.e. for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\langle \mathcal{A} x, y \rangle_{\mathcal{Y}} = \langle x, \mathcal{A}^* y \rangle_{\mathcal{X}}$. Setting $z^k := (x^k, y^k)$, the update rule of PDHGM can be written as

$$M \left( z^k - z^{k+1} \right) \in T \left( z^{k+1} \right), \text{ for } M := \begin{pmatrix} \frac{1}{\tau} I & -\mathcal{A}^* \\ -\vartheta \mathcal{A} & \frac{1}{\sigma} I \end{pmatrix}.$$

For $\vartheta = 1$, $M$ is a self-adjoint positive semi-definite operator if $\tau \sigma \| \mathcal{A} \|^2 < 1$. Henceforth, we only consider $\vartheta = 1$. We define the following Lyapunov function, whose critical points coincide with those of $L$:

$$\mathcal{L}(z, z') := L(z) + \frac{1}{2} \|z' - z\|_M^2, \qquad (4.3)$$

where

$$\|z\|_M^2 := \langle z, M z \rangle_{\mathcal{X} \times \mathcal{Y}} = \frac{1}{\tau} \|x\|_{\mathcal{X}}^2 - 2 \langle \mathcal{A} x, y \rangle_{\mathcal{Y}} + \frac{1}{\sigma} \|y\|_{\mathcal{Y}}^2.$$

For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we denote $\|y\|_M := \| (0, y) \|_M = \frac{1}{\sqrt{\sigma}} \|y\|_{\mathcal{Y}}$, and similarly $\|x\|_M := \| (x, 0) \|_M = \frac{1}{\sqrt{\tau}} \|x\|_{\mathcal{X}}$.

**Theorem 4.1** (Strict descent). *For $(x^k, y^k)$ satisfying the PDHGM updates (4.2) with $\vartheta = 1$, $\mathcal{R}$ $\rho$-weak convex, $F^*$ $\mu$-strong convex, and step sizes satisfying $\tau \sigma \| \mathcal{A} \|^2 < 1$, $\tau \rho < 1$, and $\mu \sigma > 3$, the following descent holds for the Lyapunov function (4.3):*

$$\mathcal{L}(z^k, z^{k-1}) - \mathcal{L}(z^{k+1}, z^k) \geqslant$$

$$\frac{1}{2}(\mu \sigma - 3) \left\| y^k - y^{k+1} \right\|_M^2 + \frac{1}{2}(1 - \rho \tau) \left\| x^k - x^{k+1} \right\|_M^2$$

*Proof.* Proven in Appendix E.1. □

Using this, we can prove the following under iterate bound-edness, which is a standard assumption in analysis of non-convex optimisation (Sun et al., 2018; Bolte et al., 2014).

**Theorem 4.2.** *Assume that* $\inf_{x \in \mathcal{X}} L(x, y) > -\infty$ *for all* $y \in \mathcal{Y}$*, the* $z^k = (x^k, y^k)$ *are bounded,* $\mathcal{R}$ *is* $\rho$*-weak convex,* $F^*$ *is* $\mu$*-strong convex,* $\tau\sigma\|\mathcal{A}\|^2 < 1$*,* $\tau\rho < 1$*, and* $\mu\sigma > 3$*. Let* $\nu := \min\{\mu\sigma - 3, 1 - \rho\tau\}$*, then*

$$\min_{k \leqslant K} \operatorname{dist}\left(0, \partial L(z^k)\right) \leqslant \frac{2}{\sqrt{\nu K}}\sqrt{\mathcal{L}(z^1, z^0) - \mathcal{L}(z^{K+1}, z^K)}.$$

*Furthermore, the iterates are square-summable:*

$$\sum_{k=1}^{\infty}\left\|x^{k+1} - x^k\right\|_{\mathcal{X}}^2 < \infty, \quad \sum_{k=1}^{\infty}\left\|y^{k+1} - y^k\right\|_{\mathcal{Y}}^2 < \infty.$$

*Letting* $\mathcal{C}$ *denote the set of cluster points of* $z^k$*, we have that* $\mathcal{C}$ *is a nonempty compact set and*

$$\lim_{k \to \infty} \operatorname{dist}\left(\left(x^k, y^k\right), \mathcal{C}\right) = 0,$$

*with* $\mathcal{C} \subseteq \operatorname{crit} L$ *and* $L$ *is finite and constant on* $\mathcal{C}$*.*

*Proof.* Proven in Appendix E.2. □

**Note 4.3.** It is important to note that parameter bounds $\tau\sigma\|\mathcal{A}\|^2 < 1$, $\tau\rho < 1$, and $\mu\sigma > 3$ are likely not optimal. It is, however, always possible to choose step sizes to satisfy these, taking $\tau < \min\left\{\frac{1}{\rho}, \frac{\mu}{3\|\mathcal{A}\|^2}\right\}$ and $\sigma \in \left(\frac{3}{\mu}, \frac{1}{\tau\|\mathcal{A}\|^2}\right)$.

### 4.3. Ergodic convergence rate

We now prove ergodic convergence rates, bounding the primal-dual gap for the PDHGM. To derive these, we will need bounds on iterate distance from a general point. These are not entailed by weak convexity, so we make the further assumption that the Lyapunov function $\mathcal{L}$ has the Kurdyka–Łojasiewicz property (see Appendix E.3 for details). By the following note, this assumption is not too restrictive.

**Note 4.4.** By Theorem E.3, if $\mathcal{R}$ is a deep neural network with continuous, piecewise analytic activations with finitely many pieces (e.g., ReLU), then $\mathcal{R}$ is subanalytic. If $F(y) := \frac{1}{2\alpha}\|y - y^\delta\|_{\mathcal{Y}}^2$ then all the remaining terms in $\mathcal{L}$ are analytic. It follows that $\mathcal{L}$ is subanalytic (by e.g. Lemma 7.4 of (Budd et al., 2023)) and therefore is KŁ with KŁ exponent in $[0, 1)$ on all of its domain (Bolte et al., 2007).

The following results are standard in the literature, but will prove useful for proving ergodic convergence rates.

**Theorem 4.5.** *Suppose that* $\mathcal{L}$ *is KŁ and the* $z^k$ *are bounded, then* $z^k$ *converges to a critical point* $\hat{z}$ *of* $L$ *and*

$$\sum_{k=1}^{\infty}\left\|z^{k+1} - z^k\right\|_M < \infty.$$

*Furthermore, if* $\mathcal{L}$ *has KŁ exponent* $\theta \in [0, 1)$ *at* $\hat{z}$*, we have that there exist constants* $\nu_1, \nu_2 > 0$ *and* $\tau \in (0, 1)$*, s.t.:*
*If* $\theta = 0$*, the sequence* $z^k$ *converges in finite steps.*
*If* $\theta \in \left(0, \frac{1}{2}\right]$*,* $\left\|z^k - \hat{z}\right\|_M \leqslant \nu_1\tau^k$*.*
*If* $\theta \in \left(\frac{1}{2}, 1\right)$*,* $\left\|z^k - \hat{z}\right\|_M \leqslant \nu_2 k^{-\frac{1-\theta}{2\theta-1}}$*.*

*Proof.* See (Guo et al., 2023; Sun et al., 2018). □

This lets us derive the corresponding ergodic rate.

**Theorem 4.6.** *Assume that the bounded sequence* $z^k = (x^k, y^k)$ *are PDHGM iterates from (4.2) with* $\vartheta = 1$*,* $\mathcal{R}$ $\rho$*-weak convex and* $F^*$ $\mu$*-strong convex, satisfying* $\tau\sigma\|\mathcal{A}\|^2 < 1$*,* $\tau\rho < 1$*, and* $\mu\sigma > 3$*. Assume that* $\mathcal{L}$ *is KŁ, with KŁ exponent* $\theta \in [0, 1)$ *at* $\hat{z}$*, denoting the limit* $z^k \to \hat{z} = (\hat{x}, \hat{y})$*, and let* $\bar{x}^k$ *and* $\bar{y}^k$ *denote the means of* $(x^i)_{i=1}^k$ *and of* $(y^i)_{i=1}^k$*, respectively. Then, for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$*:*

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right)$$
$$\leqslant \frac{1}{2}\left(\rho\|x - \hat{x}\|_{\mathcal{X}}^2 - \mu\|y - \hat{y}\|_{\mathcal{Y}}^2\right) + \mathcal{O}\left(\frac{\log k}{k}\right).$$

*Proof.* Proven in Appendix E.4. □

**Note 4.7.** In the case of $\rho = \mu = 0$ we recover the usual convex ergodic rates (Lu & Yang, 2023). Unlike the convex case, here there is a non-$k$ dependent term, which does not tend to zero. This arises due to non-convexity, and hence non-uniqueness of critical points of the original problem.

**Note 4.8.** We have here only shown iterate convergence in the case of $\mathcal{R}$ $\rho$-weak convex and $F^*$ $\mu$-strong convex, but the proof of this result naturally extends to the case of $F^*$ being weak convex, as long as convergence can be shown.

## 5. Learning a weakly convex regulariser

Input weakly convex neural networks (IWCNNs) were introduced in (Shumaylov et al., 2023) with the following idea. As discussed in Section 1.1, non-convex regularisers have been observed to result in better performance than convex regularisers, and 'optimal' regularisers are typically non-convex yet often weakly convex. We therefore want a neural network which is weakly convex but not convex. A smooth neural network achieves this, but these are usually undesirable (Krizhevsky et al., 2017). But a neural network with all ReLU activations is weakly convex if and only if it is convex (Shumaylov et al., 2023). To get the best of both worlds, the IWCNN architecture uses the following result: if $f = g_c \circ g_{sm}$ for $g_c : \mathbb{R}^n \to \mathbb{R}$ convex and $L$-Lipschitz and $g_{sm} : \mathbb{R}^k \to \mathbb{R}^n$ $C^1$ with $\beta$-Lipschitz gradient, then $f$ is $L\beta$-weakly convex (Davis et al., 2018). (Shumaylov et al., 2023) utilises an input convex neural network (ICNN) (Amos et al., 2017) to define an IWCNN.

**Definition 5.1** (IWCNN; Definition 4.1 in (Shumaylov et al., 2023)). An IWCNN $f_\theta^{\text{IWCNN}}$ is defined by

$$f_\theta^{\text{IWCNN}} := g_{\theta_1}^{\text{ICNN}} \circ g_{\theta_2}^{\text{sm}},$$

where $g_{\theta_1}^{\text{ICNN}}$ is an input convex neural network and $g_{\theta_2}^{\text{sm}}$ is a neural network with Lipschitz smooth activations.

We now prove that IWCNNs can universally approximate continuous functions. We will use the fact (see (Sun & Yu, 2019) Theorem 6) that the weakly convex functions (on an open or closed convex domain) are dense in the continuous functions (on that domain) with respect to $\|\cdot\|_\infty$. Therefore, as long as IWCNNs can approximate any weakly convex function, they can approximate any continuous function.

**Theorem 5.2** (Universal Approximation of IWCNN). *Consider an IWCNN as in Definition 5.1, with intermediate dimension $m \in \mathbb{N}$. For any $m$, IWCNNs can uniformly approximate any continuous function on a compact domain.*

*Proof.* Proof provided in Appendix F. $\square$

### 5.1. Adversarial regularisation

Adversarial learning operates within the weakly supervised setting described in Section 1. We have datasets $(x_i) \in \mathcal{X}$ and $(y_j) \in \mathcal{Y}$ each i.i.d. sampled from the distributions of ground truth images $\mathbb{P}_r$ and measurements $\mathbb{P}_Y$, respectively. To get these two distributions into the same space, we push-forward $\mathbb{P}_Y$ from the measurement space $\mathcal{Y}$ to the image space $\mathcal{X}$ using some pseudo-inverse $\mathcal{A}^\dagger$ of the forward model, giving $\mathbb{P}_n := (\mathcal{A}^\dagger)_\# \mathbb{P}_Y$, a distribution of images with reconstruction artefacts.

The key idea is that regularisation is really a type of classification problem. We want $\mathcal{R}$ to be small on the ground truth images (i.e., $\mathbb{P}_r$) and large on artificial images (i.e., $\mathbb{P}_n$). Therefore, (Lunz et al., 2018) learned a neural network regulariser $\mathcal{R}_\theta$ by minimising the following loss functional:

$$\mathbb{E}_{X \sim \mathbb{P}_r}[\mathcal{R}_\theta(X)] - \mathbb{E}_{X \sim \mathbb{P}_n}[\mathcal{R}_\theta(X)] \\ + \lambda \cdot \mathbb{E}\left[(\|\partial_x \mathcal{R}_\theta(X)\| - 1)_+^2\right]. \quad (5.1)$$

The final term pushes $\mathcal{R}_\theta$ to be 1-Lipschitz, inspired by the Wasserstein GAN (WGAN) loss (Arjovsky et al., 2017), and the expectation is taken over all lines connecting samples in $\mathbb{P}_n$ and $\mathbb{P}_r$. This $\mathcal{R}_\theta$ has a key interpretation, given two assumptions. First, assume that the measure $\mathbb{P}_r$ is supported on the weakly compact set $\mathcal{M}$. This captures the intuition that real image data lies in a lower-dimensional non-linear subspace of the original space. Let $P_\mathcal{M}$ be the projection function onto $\mathcal{M}$, assumed to be defined $\mathbb{P}_n$-a.e. Then second, assume that the measures $\mathbb{P}_r$ and $\mathbb{P}_n$ satisfy $(P_\mathcal{M})_\#(\mathbb{P}_n) = \mathbb{P}_r$. This essentially means that the reconstruction artefacts are small enough to allow recovery of the real distribution by simply projecting the noisy distribution onto $\mathcal{M}$. Of the two assumptions, the latter is stronger.

**Theorem 5.3** (Optimal AR (Lunz et al., 2018)). *Under these assumptions, the distance function to $\mathcal{M}$, $d_\mathcal{M}(x) := \min_{z \in \mathcal{M}} \|x - z\|_\mathcal{X}$ is a maximiser of*

$$\sup_{f \in \text{1-Lip}} \mathbb{E}_{X \sim \mathbb{P}_n}[f(X)] - \mathbb{E}_{X \sim \mathbb{P}_r}[f(X)].$$

**Note 5.4** (Non-uniqueness). The functional in Theorem 5.3 does not necessarily have a unique maximiser. For a more in-depth discussion of when potentials and transports can be unique, see (Staudt et al., 2022; Milne et al., 2022).

In (Mukherjee et al., 2021) this approach was modified by parameterising $\mathcal{R}_\theta(x) = \mathcal{R}_\theta^{\text{ICNN}}(x) + \frac{\mu_0}{2}\|x\|_\mathcal{X}^2$, where $\mathcal{R}_\theta^{\text{ICNN}}$ is an ICNN, and then minimising (5.1) to learn an adversarial convex regulariser (ACR). In order to retain the interpretation of Theorem 5.3, we would have to assume that $d_\mathcal{M}$ is convex (so that it can be approximated by an ICNN), which is true if and only if $\mathcal{M}$ is a convex set. However, real data rarely lives on a convex set. This breakdown can lead to performance issues, as illustrated on Figure 1, where an ICNN completely fails to approximate the distance function.

### 5.2. Adversarial weakly convex and convex-nonconvex regularisation

In (Shumaylov et al., 2023), an adversarial convex-nonconvex regulariser (ACNCR) was defined by $\mathcal{R}_\theta(x) := \mathcal{R}_{\theta_1}^c(x) + \mathcal{R}_{\theta_2}^{wc}(\mathcal{A}\,x)$, for $\mathcal{R}_{\theta_1}^c$ parameterised the same as the ACR, and $\mathcal{R}_{\theta_2}^{wc}$ parameterised with an IWCNN. Then $\mathcal{R}_{\theta_1}^c$ and $\mathcal{R}_{\theta_2}^{wc}$ are trained in a decoupled way, with $\mathcal{R}_{\theta_1}^c$ minimising (5.1) and $\mathcal{R}_{\theta_2}^{wc}$ minimising a similar loss but with $\mathbb{P}_n$ replaced by $\mathbb{P}_Y$ and $\mathbb{P}_r$ replaced by $\mathbb{P}_{Y_r} := (\mathcal{A})_\# \mathbb{P}_r$, the push-forward of the ground truth distribution under the forward model. This was an important step towards greater interpretability, since, by Theorem 5.3, the optimal $\mathcal{R}_{\theta_2}^{wc}(y) \approx d_{\mathcal{A}[\mathcal{M}]}(y)$, but it is still unclear how to interpret $\mathcal{R}_{\theta_1}^c$. In this work, we define a more directly weak convex parameterisation, following the ACR.

**Definition 5.5** (Adversarial weak convex regulariser). The *adversarial weak convex regulariser* (AWCR) is parameterised by $\mathcal{R}_\theta(x) = \mathcal{R}_\theta^{\text{IWCNN}}(x) + \frac{\mu_0}{2}\|x\|_\mathcal{X}^2$, where $\mathcal{R}_\theta^{\text{IWCNN}}$ is an IWCNN. It is learned by minimising (5.1). This satisfies Assumption 3.1 with the modification described in Note 3.7, as an IWCNN is Lipschitz by construction, and so $\mu$ can be chosen to be arbitrarily small.

**Note 5.6.** Theorem 5.2 allows this AWCR to overcome the ACR's issue of ICNNs not being able to approximate a distance function to non-convex sets. An IWCNN is able to approximate the distance function to any compact set, since any such distance function is 1-Lipschitz continuous.

Ground-truth     FBP: 21.303 dB, 0.195     TV: 31.690 dB, 0.889     U-Net: 36.712 dB, 0.920     LPD: 36.810 dB, 0.912

AR: 36.694 dB, 0.907     ACR: 35.708 dB, 0.897     ACNCR: 36.533 dB, 0.921     AWCR: 37.603 dB, 0.918     AWCR-PD: 37.941 dB, 0.924
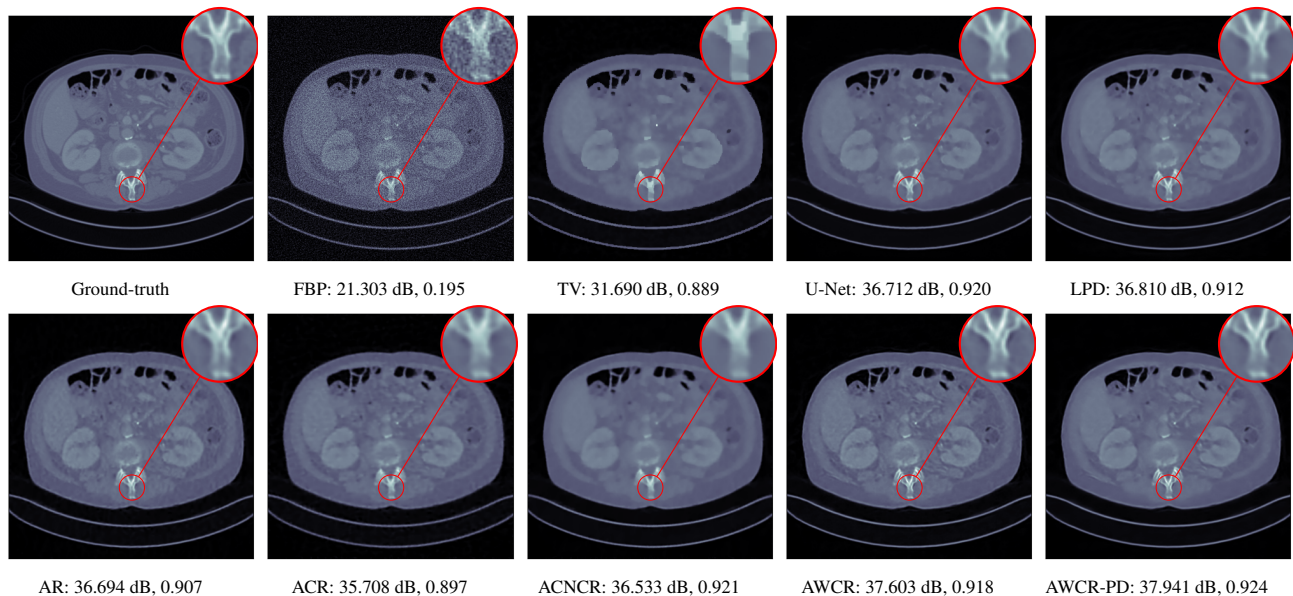
Figure 2: Reconstructed images obtained using different methods, along with the associated PSNR and SSIM, for sparse view CT. In this case the AWCR and AWCR-PD achieve the highest PSNR and SSIM. Furthermore, both AWCR methods retain the fine-structure in the reconstruction, unlike the ACNCR and ACR, the only other methods which possess convergence guarantees.

## 6. Numerical results

### 6.1. Distance function approximations

To illustrate the importance of proper neural network structure, we first discuss a toy example. We consider data lying on a product manifold $\mathcal{M}$ of two separate spirals embedded into $\mathbb{R}^2$ in the form of a double spiral, as shown in Figure 1. We then train an AR, an ACR (i.e., an ICNN), and an AWCR (i.e., an IWCNN) on the denoising problem for this data, and compare the learned regularisers with the true $d_{\mathcal{M}}$. As is clear from Figure 1, the ICNN entirely fails to approximate this non-convex $d_{\mathcal{M}}$. This limitation is overcome by the IWCNN, which despite the imposed constraints can approximate the true function well, and even shows improved extrapolation. Despite the simplicity of this example, the overall hierarchy of methods persists in the experiments explored in Section 6.2 on higher dimensional real data.

### 6.2. Computed Tomography (CT)

For evaluation of the proposed methodology, we consider two applications: CT reconstruction with (i) sparse-view and (ii) limited-angle projection. Supervised methods require access to large high-quality datasets for training, but outside of curated datasets, obtaining large amounts of high-quality paired data is unrealistic. For this reason weakly supervised methods, requiring access only to unpaired data, are of significant interest for this problem.

We consider two versions of the AWCR method, one using the subgradient method to solve (1.2) as in (Lunz et al.,

2018; Mukherjee et al., 2021; 2024), and one using PDHGM, denoted AWCR-PD. These are compared with:

- two standard knowledge-driven techniques: filtered back-projection (FBP) and total variation (TV) regularisation, which act as the baseline;

- two supervised data-driven methods: the learned primal-dual (LPD) method (Adler & Öktem, 2018) and U-Net-based post-processing of FBP (Jin et al., 2017), considered to be state of the art methods in end-to-end learned reconstruction from the two main paradigms: algorithm unrolling and learned post-processing;

- three weakly supervised methods: the AR, ACR, and the ACNCR. Adversarial regularisation methods were chosen as the currently best performing, to the authors' knowledge, weakly supervised methods for CT.

These comparisons illustrate the trade-offs in levels of constraints and supervision versus stability and performance. For details of the experimental set-up, see Appendix G.1. We measure the performance in terms of the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) (Wang et al., 2004). We report average test dataset results in Table 1, with further visual examples in Figure 2.

**Sparse view CT** As in (Lunz et al., 2018) performance of AR during reconstruction begins to deteriorate if the network is over-trained, so early stopping must be employed in training. For the ACR, ACNCR, and both AWCR methods

Table 1: Average test PSNR and SSIM in CT experiments.

| Methods | Limited | | Sparse | |
|---|---|---|---|---|
| | PSNR (dB) | SSIM | PSNR (dB) | SSIM |
| *Knowledge-driven* | | | | |
| FBP | 17.1949 | 0.1852 | 21.0157 | 0.1877 |
| TV | **25.6778** | **0.7934** | **31.7619** | **0.8883** |
| *Supervised* | | | | |
| LPD | 28.9480 | **0.8394** | **37.4868** | 0.9217 |
| U-Net | **29.1103** | 0.8067 | 37.1075 | **0.9265** |
| *Weakly Supervised* | | | | |
| AR | 23.6475 | 0.6257 | 36.4079 | 0.9101 |
| ACR | 26.4459 | **0.8184** | 34.5844 | 0.8765 |
| ACNCR | 26.5420 | 0.8161 | 35.6476 | 0.9094 |
| AWCR | **26.7233** | 0.8105 | 36.5323 | 0.9046 |
| AWCR-PD | 24.3254 | 0.7650 | **36.9993** | **0.9108** |

this does not occur due to reduced expressivity, yet both AWCR methods surpass the performance of AR. Furthermore, the AWCR methods perform qualitatively better than the AR, ACR, and ACNCR (especially the latter two) at reconstructing fine details, which in the medical setting can be more important than PSNR/SSIM accuracy. Indeed, the AWCR-PD method approaches the PSNR accuracy of the strongly supervised U-Net post-processing method.

**Limited view CT** In this setting a specific angular region contains no measurement, turning this into a severely ill-posed inverse problem, and a good image prior is crucial for reconstruction. As shown in (Mukherjee et al., 2021), the AR begins introducing artifacts during reconstruction, which is overcome for both the ACNCR and ACR due to the imposed convexity. The AWCR, on the other hand, is able to remain non-convex without experiencing deterioration, and slightly outperforms both the ACR and ACNCR in PSNR, but not SSIM. However, the AWCR-PD method performs worse in this setting, though still outperforming AR. This occurs due to the forward and the adjoint operator being severely ill-posed. For a visual comparison of the AWCR and AWCR-PD methods in this setting, see Appendix G.2.

## 7. Conclusions

In this work, we have shown that the framework of weak convexity is fertile ground for proving theoretical guarantees. We proved that weakly convex regularisation can yield existence, stability, and convergence in the sense of critical points. Furthermore, we filled a gap in the literature by proving convergence (with an ergodic rate) for the PDHGM optimisation method (with $\vartheta = 1$) in the weakly convex setting. All of these results hold when $\mathcal{R}$ is a weakly convex neural network. We showed that the IWCNN architecture can universally approximate continuous functions, allowing

the AWCR to retain interpretability as a distance function (empirically illustrated in Figure 1) whilst also benefiting from all of the provable guarantees from weak convexity. Finally, for the problem of CT reconstruction we showed that the AWCR can compete with or outperform the state-of-the-art in adversarial regularisation, and approach the lower bound performance of state-of-the-art supervised approaches for CT, despite being merely weakly supervised.

Although we have focused on adversarial regularisation in this work, we emphasise that other learned regularisation methods can be put into this weakly convex setting. For example, in (Hurault et al., 2023) it was shown that plug-and-play proximal gradient descent approaches (using sufficiently smooth gradient step denoisers) correspond to criticising an energy with a weakly convex regulariser. The total deep variation regulariser (Kobler et al., 2020)

$$\mathcal{R}(x) := \sum_{i=1}^{n} \sum_{j=1}^{\ell} \mathcal{N}_{ij}(Kx)w_j$$

is weakly convex so long as $w_j \geqslant 0$ for all $j$ and the $\mathcal{N}_{ij}$ are weakly convex for all $i, j$, which is achieved if the U-Net $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}^{n\ell}$ has the form of an IWCNN, using a uniformly convex U-Net as in (Bianchi et al., 2023). The NETT $\ell^q$ regulariser (Li et al., 2020) has the form:

$$\mathcal{R}(x) := \sum_{\lambda \in \Lambda} \|\Phi_{\lambda,\theta}(x)\|_q^q,$$

where $q \geqslant 1$ and $\Lambda$ is a finite set. For $q = 2$ and $|\Lambda| = 1$, this is also the form of the learned noise reduction regulariser used in (Liu et al., 2022). A sufficient condition for this $\mathcal{R}$ to be weakly convex is, for all $\lambda$, $\Phi_{\lambda,\theta}$ is at least one of: bounded and $C^1$ with Lipschitz gradient (by (Davis et al., 2018)), or weakly convex (e.g., an IWCNN) with $(\Phi_{\lambda,\theta}(x))_i \in [0, B_i]$ for all $x$ and $i$ (by Proposition A.1).

A number of open questions remain. Expressivity questions about IWCNNs beyond universal approximation, e.g. efficiency of representation, remain unanswered. It remains open whether the PDHGM algorithm can be shown to be convergent for a generic weak-weak splitting; or whether the boundedness assumption can be dropped in certain cases. Alternatively, given the structure of the regulariser, it may be beneficial to turn to prox-linear schemes instead, analysis of which is currently missing in the primal-dual setting. Also, the problem of extracting proximal operators directly from learned networks remains open. Finally, there is the broader question of comparing the effectiveness of different approaches for learned regularisation, especially (as noted by an anonymous reviewer) quantifying how this effectiveness depends on the ill-posedness of the inverse problem.

## Impact Statement

This paper presents work whose goal is to advance the fields of Machine Learning, Optimisation and Inverse Problems. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgements

## References

Adler, J. and Öktem, O. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018. (Cited on pages 2 and 8.)

Adler, J., Kohr, H., and Öktem, O. Operator discretization library (ODL). *Software available from https://github.com/odlgroup/odl*, 2017. (Cited on page 27.)

Alberti, G. S., De Vito, E., Lassas, M., Ratti, L., and Santacesaria, M. Learning the optimal Tikhonov regularizer for inverse problems. *Advances in Neural Information Processing Systems*, 34:25205–25216, 2021. (Cited on page 2.)

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005. (Cited on page 15.)

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017. (Cited on pages 2 and 6.)

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN, 2017. (Cited on page 7.)

Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059. (Cited on page 2.)

Arrow, K. J., Hurwicz, L., and Chenery, H. B. *Studies in linear and non-linear programming*. Stanford Mathematical Studies in the Social Sciences. Stanford University Press, Stanford, California, 1958. (Cited on page 2.)

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz Inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/40801236. (Cited on page 23.)

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974997. (Cited on page 15.)

Benning, M. and Burger, M. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018. doi: 10.1017/S0962492918000016. (Cited on page 2.)

Bianchi, D., Lai, G., and Li, W. Uniformly convex neural networks and non-stationary iterated network tikhonov (inett) method. *Inverse Problems*, 39 (5):055002, mar 2023. doi: 10.1088/1361-6420/acc2b6. URL https://dx.doi.org/10.1088/1361-6420/acc2b6. (Cited on page 9.)

Böhm, A. and Wright, S. J. Variable smoothing for weakly convex composite functions. *Journal of optimization theory and applications*, 188:628–649, 2021. (Cited on page 2.)

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. (Cited on pages 6 and 23.)

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014. (Cited on pages 6 and 22.)

Borwein, J. and Lewis, A. *Convex Analysis and Nonlinear Optimization*. CMS Books in Mathematics. Springer New York, NY, 2nd edition, 2006. doi: 10.1007/978-0-387-31256-9. (Cited on page 15.)

Bredies, K., Kunisch, K., and Pock, T. Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. doi: 10.1137/090769521. (Cited on page 2.)

Budd, J. M., van Gennip, Y., Latz, J., Parisotto, S., and Schönlieb, C.-B. Joint reconstruction-segmentation on graphs. *SIAM Journal on Imaging Sciences*, 16 (2):911–947, June 2023. ISSN 1936-4954. doi: 10. 1137/22m151546x. URL http://dx.doi.org/10. 1137/22M151546X. (Cited on page 6.)

Calatroni, L., d'Autume, M., Hocking, R., Panayotova, S., Parisotto, S., Ricciardi, P., and Schönlieb, C.-B. Unveiling the invisible: mathematical methods for restoring and interpreting illuminated manuscripts. *Heritage Science*, 6(1), Sep 2018. ISSN 2050-7445. doi: 10.1186/s40494-018-0216-z. URL http://dx.doi. org/10.1186/s40494-018-0216-z. (Cited on page 1.)

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40 (1):120–145, May 2011. ISSN 1573-7683. doi: 10. 1007/s10851-010-0251-1. URL https://doi.org/ 10.1007/s10851-010-0251-1. (Cited on pages 2 and 5.)

Chen, G. and Needell, D. Compressed sensing and dictionary learning. *Proc. Sympos. Appl. Math.*, 73:201–241, 2016. (Cited on page 2.)

Chen, Y., Shi, Y., and Zhang, B. Optimal control via neural networks: A convex approach. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum? id=H1MW72AcK7. (Cited on page 27.)

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018. (Cited on page 2.)

Davis, D. and Drusvyatskiy, D. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22(2):561–606, 2022. (Cited on pages 2 and 16.)

Davis, D., Drusvyatskiy, D., MacPhee, K. J., and Paquette, C. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, Dec 2018. ISSN 1573-2878. doi: 10.1007/s10957-018-1372-8. URL https://doi. org/10.1007/s10957-018-1372-8. (Cited on pages 2, 6, and 9.)

Davoli, E., Fonseca, I., and Liu, P. Adaptive image processing: First order PDE constraint regularizers and a bilevel training scheme. *Journal of Nonlinear Science*, 33(3):41, Mar 2023. ISSN 1432-1467. doi: 10.1007/s00332-023-09902-4. URL https://doi.

org/10.1007/s00332-023-09902-4. (Cited on page 2.)

Dimakis, A. G. *Deep Generative Models and Inverse Problems*, pp. 400–421. Cambridge University Press, 2022. doi: 10.1017/9781009025096.010. (Cited on page 2.)

Dittmer, S., Kluth, T., Maass, P., and Baguer, D. O. Regularization by architecture: A deep prior approach for inverse problems. *Journal of Mathematical Imaging and Vision*, pp. 1–15, 2019. (Cited on page 2.)

Drusvyatskiy, D. and Davis, D. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, 28:1–10, 2020. (Cited on page 2.)

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019. (Cited on page 2.)

Durand, S. and Nikolova, M. Stability of the minimizers of least squares with a non-convex regularization. Part I: Local behavior. *Applied Mathematics and Optimization*, 53:185–208, 2006. (Cited on page 3.)

Engl, H. W., Hanke, M., and Neubauer, A. *Regularization of inverse problems*, volume 375 of *Mathematics and Its Applications*. Springer Dordrecht, 1996. (Cited on page 1.)

Gilboa, G. and Osher, S. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2009. doi: 10.1137/070698592. URL https://doi.org/10.1137/070698592. (Cited on page 2.)

Goujon, A., Neumayer, S., and Unser, M. Learning weakly convex regularizers for convergent image-reconstruction algorithms. *SIAM Journal on Imaging Sciences*, 17(1): 91–115, 2024. doi: 10.1137/23M1565243. URL https: //doi.org/10.1137/23M1565243. (Cited on page 2.)

Grasmair, M. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse problems*, 26(11):115014, 2010. (Cited on page 4.)

Guo, J., Wang, X., and Xiao, X. Preconditioned primal-dual gradient methods for nonconvex composite and finite-sum optimization. *arXiv preprint arXiv:2309.13416*, 2023. (Cited on pages 2, 6, and 22.)

Hanin, B. Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations. *Mathematics*, 7(10):992, October 2019. ISSN 2227-7390. doi: 10.3390/math7100992. URL http://dx.doi. org/10.3390/math7100992. (Cited on page 27.)

Hurault, S., Leclaire, A., and Papadakis, N. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9483–9505. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hurault22a.html. (Cited on page 2.)

Hurault, S., Chambolle, A., Leclaire, A., and Papadakis, N. A relaxed proximal gradient descent algorithm for convergent plug-and-play with proximal denoiser. In Calatroni, L., Donatelli, M., Morigi, S., Prato, M., and Santacesaria, M. (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 379–392, Cham, 2023. Springer International Publishing. ISBN 978-3-031-31975-4. (Cited on page 9.)

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26 (9):4509–4522, 2017. (Cited on pages 2 and 8.)

Kamilov, U. S., Bouman, C. A., Buzzard, G. T., and Wohlberg, B. Plug-and-Play Methods for Integrating Physical and Learned Models in Computational Imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023. doi: 10.1109/MSP.2022.3199595. (Cited on page 2.)

Kidger, P. and Lyons, T. Universal Approximation with Deep Narrow Networks. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/kidger20a.html. (Cited on page 27.)

Kobler, E., Effland, A., Kunisch, K., and Pock, T. Total deep variation for linear inverse problems. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7546–7555, 2020. doi: 10.1109/CVPR42600.2020.00757. (Cited on pages 2 and 9.)

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. (Cited on page 6.)

Kruger, A. Y. On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, Jul 2003. ISSN 1573-8795. doi: 10.1023/A: 1023673105317. URL https://doi.org/10.1023/A:1023673105317. (Cited on page 3.)

Kurdyka, K. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pp. 769–783, 1998. (Cited on page 23.)

Kutyniok, G. and Labate, D. *Multiscale analysis for multivariate data*. Springer, 2012. (Cited on page 2.)

Lanza, A., Morigi, S., Selesnick, I. W., and Sgallari, F. Convex non-convex variational models. In *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pp. 1–57. Springer, 2022. (Cited on page 2.)

Lassas, M., Saksman, E., and Siltanen, S. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009. ISSN 1930-8337. doi: 10.3934/ipi.2009.3.87. URL https://www.aimsciences.org/article/id/2c169f6d-50f3-4e01-9fc0-24c836905147. (Cited on page 2.)

Lempitsky, V., Vedaldi, A., and Ulyanov, D. Deep image prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018. doi: 10.1109/CVPR.2018.00984. (Cited on page 2.)

Leong, O., O'Reilly, E., Soh, Y. S., and Chandrasekaran, V. Optimal regularization for a data source. *arXiv preprint arXiv:2212.13597*, 2023. (Cited on page 2.)

Li, H., Schwab, J., Antholzer, S., and Haltmeier, M. NETT: Solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020. (Cited on pages 2 and 9.)

Liu, R., Sun, Y., Zhu, J., Tian, L., and Kamilov, U. S. Recovery of continuous 3d refractive index maps from discrete intensity-only measurements using neural fields. *Nature Machine Intelligence*, 4(9):781–791, Sep 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00530-3. URL https://doi.org/10.1038/s42256-022-00530-3. (Cited on page 9.)

Lorenz, D. and Worliczek, N. Necessary conditions for variational regularization schemes. *Inverse Problems*, 29 (7):075016, 2013. (Cited on page 4.)

Lu, H. and Yang, J. On a Unified and Simplified Proof for the Ergodic Convergence Rates of PPM, PDHG and ADMM. *arXiv preprint arXiv:2305.02165*, 2023. (Cited on pages 5 and 6.)

Lunz, S. *Machine Learning in Inverse Problems - Learning Regularisation Functionals and Operator Corrections*. PhD thesis, Apollo - University of Cambridge Repository, 2021. URL https://www.repository.cam.ac.uk/handle/1810/336791. (Cited on page 2.)

Lunz, S., Öktem, O., and Schönlieb, C.-B. Adversarial regularizers in inverse problems. *Advances in neural information processing systems*, 31, 2018. (Cited on pages 2, 7, 8, and 28.)

Mallat, S. *A wavelet tour of signal processing*. Academic Press, Boston, third edition, 2009. ISBN 978-0-12-374370-1. doi: https://doi.org/10.1016/B978-0-12-374370-1.00021-5. URL https://www.sciencedirect.com/science/article/pii/B9780123743701000215. (Cited on page 2.)

McCollough, C. Tfg-207a-04: Overview of the low dose CT grand challenge. *Medical Physics*, 43(6):3759–3760, 2014. (Cited on page 27.)

Milne, T., Étienne Bilocq, and Nachman, A. A new method for determining Wasserstein 1 optimal transport maps from Kantorovich potentials, with deep learning applications. *arXiv preprint arXiv:2211.00820*, 2022. (Cited on page 7.)

Mohimani, H., Babaie-Zadeh, M., and Jutten, C. A fast approach for overcomplete sparse decomposition based on smoothed $\ell^0$ norm. *IEEE Transactions on Signal Processing*, 57(1):289–301, 2009. doi: 10.1109/TSP.2008.2007606. (Cited on page 2.)

Möllenhoff, T., Strekalovskiy, E., Moeller, M., and Cremers, D. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015. (Cited on page 2.)

Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Öktem, O., and Schönlieb, C.-B. Learned convex regularizers for inverse problems. *arXiv preprint arXiv:2008.02839*, 2021. (Cited on pages 2, 7, 8, 9, and 28.)

Mukherjee, S., Hauptmann, A., Öktem, O., Pereyra, M., and Schönlieb, C.-B. Learned reconstruction methods with convergence guarantees: a survey of concepts and applications. *IEEE Signal Processing Magazine*, 40(1):164–182, 2023. (Cited on page 2.)

Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Öktem, O., and Schönlieb, C.-B. Data-driven convex regularizers for inverse problems. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13386–13390. IEEE, 2024. (Cited on page 8.)

Natterer, F. *The mathematics of computerized tomography*. Society for Industrial and Applied Mathematics, USA, 2001. ISBN 0898714931. (Cited on page 1.)

Obmann, D. and Haltmeier, M. Convergence analysis of critical point regularization with non-convex regularizers. *Inverse Problems*, 2022. (Cited on pages 3 and 4.)

Obmann, D. and Haltmeier, M. Convergence analysis of equilibrium methods for inverse problems. *arXiv preprint arXiv:2306.01421*, 2023. (Cited on page 3.)

Obmann, D. and Haltmeier, M. Convergence rates for critical point regularization. *Applied Mathematics and Computation*, 471:128605, 2024. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2024.128605. URL https://www.sciencedirect.com/science/article/pii/S0096300324000778. (Cited on page 3.)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. *NIPS 2017 Workshop Autodiff*, 2017. (Cited on page 28.)

Phillips, D. L. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1):84–97, 1962. (Cited on page 1.)

Pieper, K. and Petrosyan, A. Nonconvex regularization for sparse neural networks. *Applied and Computational Harmonic Analysis*, 61:25–56, 2022. ISSN 1063-5203. doi: 10.1016/j.acha.2022.05.003. URL https://www.sciencedirect.com/science/article/pii/S1063520322000434. (Cited on page 2.)

Pinilla, S., Mu, T., Bourne, N., and Thiyagalingam, J. Improved imaging by invex regularizers with global optima guarantees. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VpHFHz57fT. (Cited on page 2.)

Pöschl, C. *Tikhonov regularization with general residual term*. PhD thesis, Leopold Franzens Universität Innsbruck, 2008. (Cited on pages 3 and 4.)

Romano, Y., Elad, M., and Milanfar, P. The Little Engine That Could: Regularization by Denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. doi: 10.1137/16M1102884. URL https://doi.org/10.1137/16M1102884. (Cited on page 2.)

Roth, S. and Black, M. J. Fields of experts. *International Journal of Computer Vision*, 82:205–229, 2009. (Cited on page 2.)

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: 10.1016/0167-2789(92)90242-F. (Cited on page 2.)

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2009. (Cited on page 1.)

Shumaylov, Z., Budd, J., Mukherjee, S., and Schönlieb, C.-B. Provably convergent data-driven convex-nonconvex regularization. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023. URL https://openreview.net/forum?id=yavtWi6ew9. (Cited on pages 2, 6, and 7.)

Staudt, T., Hundrieser, S., and Munk, A. On the uniqueness of Kantorovich potentials. *arXiv preprint arXiv:2201.08316*, 2022. (Cited on page 7.)

Strömberg, T. On regularization in Banach spaces. *Arkiv för Matematik*, 34(2):383 – 406, 1996. doi: 10.1007/BF02559552. URL https://doi.org/10.1007/BF02559552. (Cited on page 26.)

Sun, S. and Yu, Y. Least squares estimation of weakly convex functions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2271–2280. PMLR, 2019. (Cited on pages 7 and 27.)

Sun, T., Barrio, R., Cheng, L., and Jiang, H. Precompact convergence of the nonconvex primal–dual hybrid gradient algorithm. *Journal of Computational and Applied Mathematics*, 330:15–27, 2018. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2017.07.037. URL https://www.sciencedirect.com/science/article/pii/S0377042717303886. (Cited on pages 2, 5, and 6.)

Sun, Y., Liu, J., and Kamilov, U. S. Block coordinate regularization by denoising. *IEEE Transactions on Computational Imaging*, 6:908–921, 2020. doi: 10.1109/TCI.2020.2996385. (Cited on page 2.)

Tikhonov, A. N. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963. (Cited on page 1.)

Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048. (Cited on page 2.)

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. (Cited on page 8.)

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, Mar 2018. ISSN 1476-4687. doi: 10.1038/nature25988. URL https://doi.org/10.1038/nature25988. (Cited on page 2.)

## A. Weakly convex analysis review

For a locally Lipschitz $f : \mathcal{X} \to \mathbb{R}$, the following are equivalent to $\rho$-convexity (Ambrosio et al., 2005):

1. For all $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$,

$$f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leqslant \lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right) - \frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2.$$

2. For any $x, \hat{x} \in \mathcal{X}$ and $\psi \in \partial f(x)$, we have the following inequality: $f(\hat{x}) \geqslant f(x) + \psi(\hat{x} - x) + \frac{\rho}{2}\|\hat{x} - x\|_{\mathcal{X}}^2$.

**Proposition A.1.** *Let $f : \mathcal{X} \to [0, B]$ be $\rho$-weakly convex and $q \geqslant 1$. Then $f(x)^q$ is $qB^{q-1}\rho$-weakly convex.*

*Proof.* Note that by Taylor's theorem, for all $y, h \geqslant 0$, $(y + h)^q = y^q + q\xi^{q-1}h$ for some $\xi \in [y, y + h]$. We have that for all $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, $0 \leqslant \lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right) \leqslant B$ and

$$0 \leqslant f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leqslant \min\left\{\lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right) + \frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2, B\right\}.$$

Hence, since $y \mapsto y^q$ is monotonic on $[0, \infty)$,

$$
\begin{aligned}
f(\lambda x_1 &+ (1 - \lambda)x_2)^q \\
&\leqslant \left(\min\left\{\lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right) + \frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2, B\right\}\right)^q \\
&\leqslant \left(\lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right) + \min\left\{\frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2, B - \lambda f\left(x_1\right) - (1 - \lambda)f\left(x_2\right)\right\}\right)^q \\
&= (\lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right))^q + q\xi^{q-1}\min\left\{\frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2, B - \lambda f\left(x_1\right) - (1 - \lambda)f\left(x_2\right)\right\} \\
&\leqslant \lambda f(x_1)^q + (1 - \lambda)f(x_2)^q + qB^{q-1}\frac{\rho\lambda(1 - \lambda)}{2}\|x_1 - x_2\|_{\mathcal{X}}^2,
\end{aligned}
$$

since $\xi \in [\lambda f\left(x_1\right) + (1 - \lambda)f\left(x_2\right), B]$ and $y \mapsto y^q$ is convex. $\qquad\square$

For an extended real-valued function $f : \mathcal{X} \to (-\infty, +\infty]$, let $\operatorname{dom} f := \{x \in \mathcal{X} : f(x) < +\infty\}$ be its domain and

$$f^*(\psi) := \sup_{x \in \mathcal{X}}\{\psi(x) - f(x)\}, \qquad \psi \in \mathcal{X}^*,$$

be its conjugate function, which is always closed, convex, and lower semi-continuous (l.s.c.), see Theorem 4.3 in (Beck, 2017).

**Theorem A.2** (Fenchel–Moreau theorem; Theorem 4.2.1 in (Borwein & Lewis, 2006)). *Let $\mathcal{X}$ be a locally convex Hausdorff space, let $f : \mathcal{X} \to (-\infty, +\infty]$ be convex and l.s.c., and let $\phi : \mathcal{X} \to \mathcal{X}^{**}$ be the canonical embedding which sends $x$ to the pointwise evaluation map $\delta_x : \psi \mapsto \psi(x)$. Then $f = f^{**} \circ \phi$, i.e., for all $x \in \mathcal{X}$*

$$f(x) = \max_{\psi \in \mathcal{X}^*}\{\psi(x) - f^*(\psi)\}.$$

*If $\mathcal{X}$ is a Hilbert space, then by the Riesz representation theorem it follows that*

$$f(x) = \max_{x' \in \mathcal{X}}\{\langle x, x'\rangle_{\mathcal{X}} - f^*(x')\}.$$

**Definition A.3** (Moreau envelope and Proximal operator). For any $f : \mathcal{X} \to (-\infty, +\infty]$ and $\nu > 0$, the *Moreau envelope* and the *proximal mapping* are defined for all $x \in \mathcal{X}$ by

$$f_\nu(x) := \inf_{z \in \mathcal{X}}\left\{f(z) + \frac{1}{2\nu}\|z - x\|_{\mathcal{X}}^2\right\},$$

$$\operatorname{prox}_{\nu f}(x) := \operatorname*{argmin}_{z \in \mathcal{X}}\left\{f(z) + \frac{1}{2\nu}\|z - x\|_{\mathcal{X}}^2\right\}.$$

The Moreau envelope of the function has some rather nice properties with relation to the original function:

**Theorem A.4** (Moreau envelope and the proximal point map; Lemma 2.5 in (Davis & Drusvyatskiy, 2022))**.** *Consider a $\rho$-weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and fix a parameter $\mu < \rho^{-1}$. Then, the following are true:*

- *The envelope $f_\mu$ is $C^1$-smooth with its gradient given by*

$$\nabla f_\mu(x) = \mu^{-1} \left( x - \text{prox}_{\mu f}(x) \right).$$

- *The envelope $f_\mu(\cdot)$ is $\mu^{-1}$-smooth and $\frac{\rho}{1-\mu\rho}$-weakly convex meaning:*

$$-\frac{\rho}{2(1-\mu\rho)} \left\| x' - x \right\|^2 \leqslant f_\mu \left( x' \right) - f_\mu(x) - \langle \nabla f_\mu(x), x' - x \rangle \leqslant \frac{1}{2\mu} \left\| x' - x \right\|^2,$$

*for all $x, x' \in \mathbb{R}^d$.*

- *The proximal map $\text{prox}_{\mu f}(\cdot)$ is $\frac{1}{1-\mu\rho}$-Lipschitz continuous and the gradient map $\nabla f_\mu$ is Lipschitz continuous with constant $\max \left\{ \mu^{-1}, \frac{\rho}{1-\mu\rho} \right\}$.*

## B. Proof of Theorem 2.6

First, we introduce some useful notation. For $\psi \in \mathcal{X}^*$ and $x \in \mathcal{X}$, we define the pairing $\langle \psi, x \rangle_{\mathcal{X}^* \times \mathcal{X}} := \psi(x)$. For $B \subseteq \mathcal{X}^*$ and $x \in \mathcal{X}$, we define $\langle B, x \rangle_{\mathcal{X}^* \times \mathcal{X}} := \{\psi(x) : \psi \in B\}$. For sets $A, B$, we write $A \leqslant B$ if for all $a \in A$ and $b \in B$, $a \leqslant b$. We start with a lemma which will also prove useful in other proofs.

**Lemma B.1.** *Let $f : \mathcal{X} \to (-\infty, +\infty]$ be $\mu$-strongly convex. Then for all $x, z \in \mathcal{X}$*

$$\langle \partial f(x) - \partial f(z), x - z \rangle_{\mathcal{X}^* \times \mathcal{X}} \geqslant \mu \|x - z\|_{\mathcal{X}}^2,$$

*i.e., for all $\psi \in \partial f(x)$ and $\xi \in \partial f(z)$,*

$$\langle \psi - \xi, x - z \rangle_{\mathcal{X}^* \times \mathcal{X}} \geqslant \mu \|x - z\|_{\mathcal{X}}^2.$$

*Proof.* Because $f$ is $\mu$-strongly convex, it follows that for all $x, x' \in \mathcal{X}$ and $\psi \in \partial f(x)$,

$$f(x') \geqslant f(x) + \langle \psi, x' - x \rangle_{\mathcal{X}^* \times \mathcal{X}} + \frac{1}{2}\mu \|x' - x\|_{\mathcal{X}}^2,$$

and therefore

$$\langle \psi, x - z \rangle_{\mathcal{X}^* \times \mathcal{X}} \geqslant f(x) - f(z) + \frac{1}{2}\mu \|x - z\|_{\mathcal{X}}^2.$$

By the same argument,

$$\langle \xi, z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \geqslant f(z) - f(x) + \frac{1}{2}\mu \|x - z\|_{\mathcal{X}}^2,$$

and the result follows. □

**Lemma B.2.** *Let $f : \mathcal{X} \to (-\infty, +\infty]$ be weakly convex and $L$-Lipschitz. Then for all $x, z \in \mathcal{X}$ and $\psi \in \partial f(x)$, $\langle \psi, z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \leqslant L \|x - z\|_{\mathcal{X}}$.*

*Proof.* Let $x'_t := tz + (1-t)x$. Then as $t \to 0$, $x'_t \to x$ and so

$$f(x'_t) \geqslant f(x) + \langle \psi, x'_t - x \rangle_{\mathcal{X}^* \times \mathcal{X}} + o(\|x'_t - x\|_{\mathcal{X}})$$

and therefore, since $x'_t - x = t(z - x)$,

$$t\langle \psi, z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} + o(t) \leqslant f(x'_t) - f(x) \leqslant L\|x'_t - x\|_{\mathcal{X}} = tL\|z - x\|_{\mathcal{X}}.$$

Dividing both sides by $t$ and taking $t \to 0$ completes the proof. □

*Proof of Theorem 2.6.* For any $x, z \in \mathcal{X}$, by Lemma B.1,

$$\langle \partial R_{sc}(x) - \partial R_{sc}(z), x - z \rangle_{\mathcal{X}} \geqslant \mu \|x - z\|_{\mathcal{X}}^2.$$

Therefore, in the Lipschitz case, for $\mathcal{R} = R_{wc} + R_{sc}$, we have:

$$\langle (\partial R_{wc}(x) + \partial R_{sc}(x)), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \subseteq -\langle \partial R_{sc}(x) - \partial R_{sc}(z), x - z \rangle_{\mathcal{X}^* \times \mathcal{X}} + \langle \partial R_{wc}(x) + \partial R_{sc}(z), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}}$$
$$\leqslant -\mu \|x - z\|_{\mathcal{X}}^2 + (L_R + \|\partial R_{sc}(z)\|) \|x - z\|_{\mathcal{X}},$$

where the first half of the inequality follows from the above, and the second half since, for all $\psi \in \partial R_{wc}(x)$ and $\xi \in \partial R_{sc}(z)$

$$\langle \psi, z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \leqslant L_R \|x - z\|_{\mathcal{X}} \quad \text{by Lemma B.2, and}$$
$$\langle \xi, z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \leqslant \|\xi\|_{\mathcal{X}^*} \|x - z\|_{\mathcal{X}} \leqslant \|\partial R_{sc}(z)\| \|x - z\|_{\mathcal{X}}.$$

Then, for any $\hat{x}$ a stationary point we have that $0 \in \partial R_{wc}(\hat{x}) + \partial R_{sc}(\hat{x})$ and so

$$0 \leqslant -\mu \|x - z\|_{\mathcal{X}}^2 + (L_R + \|\partial R_{sc}(z)\|) \|x - z\|_{\mathcal{X}}$$

which gives

$$\|\hat{x} - z\|_{\mathcal{X}} \leqslant \frac{1}{\mu} (L_R + \|\partial R_{sc}(z)\|).$$

In the bounded weak convex case, using weak convexity instead:

$$R_{wc}(z) \geqslant R_{wc}(x) + \langle \partial R_{wc}(x), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} - \frac{\gamma}{2} \|z - x\|_{\mathcal{X}}^2.$$

Then we get the following:

$$\langle (\partial R_{wc}(x) + \partial R_{sc}(x)), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} \subseteq -\langle \partial R_{sc}(x) - \partial R_{sc}(z), x - z \rangle_{\mathcal{X}^* \times \mathcal{X}} + \langle \partial R_{wc}(x) + \partial R_{sc}(z), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}}$$
$$\leqslant -\mu \|x - z\|_{\mathcal{X}}^2 + \frac{\gamma}{2} \|x - z\|_{\mathcal{X}}^2 + (\|\partial R_{sc}(z)\|) \|x - z\|_{\mathcal{X}} + R_{wc}(z).$$

Taking $\hat{x}$ to be any stationary point and completing the square, this implies that

$$\left( \|\hat{x} - z\|_{\mathcal{X}} - \frac{1}{2(\mu - \frac{\gamma}{2})} (\|\partial R_{sc}(z)\|) \right)^2 \leqslant \frac{1}{4(\mu - \frac{\gamma}{2})^2} (\|\partial R_{sc}(z)\|)^2 + \frac{1}{(\mu - \frac{\gamma}{2})} R_{wc}(z),$$

or rearranging and using the triangle inequality:

$$\|\hat{x} - z\|_{\mathcal{X}} \leqslant \frac{1}{(\mu - \frac{\gamma}{2})} \|\partial R_{sc}(z)\| + \sqrt{\frac{1}{(\mu - \frac{\gamma}{2})} R_{wc}(z)}.$$

$\square$

## C. Unbounded critical points

Example construction of how to get infinitely many minimisers for the weak convex case.

**Note C.1.** It may not be immediately clear why the condition is $\gamma < 2\mu$ and not simply $\gamma < \mu$, thus requiring strong convexity of the overall functions. The idea for why arises from the lower boundedness of the weak convex function. As an example, one can consider the 1D case. The overall value of the function can not decrease arbitrarily, thus the derivative cannot be strictly negative. And thus heuristically, as the gradient of a strong convex function simply increases, the weak convex one has to decrease and then increase, overall resulting in the factor of 2, since it has to go up and then back down. Based on this heuristic understanding we can answer the question of whether it is necessary to have $\gamma < 2\mu$ to have bounded stationary points, or if it can be made larger. As it turns out, we can construct functions with $\gamma > 2\mu$ with unbounded stationary points. Assuming we are working in 1D on a positive real line, and letting $R_{sc} = \frac{1}{2}x^2$, assume we start with a stationary point at $x = x_0 = 1$, with value $R_{wc} = 0$, i.e. derivative intersects $y = -x$. Now, we require that the function is

bounded from below by 0 and thus gradient has to now become positive in such a way to increase and decrease quickly enough to once again cross $y = -x$. The largest area is found by taking the gradient to be $-\gamma x$. With this, we can find that the next intersection, and thus stationary point would be at $x_1 = \frac{\gamma}{\gamma-2} x_0$. And continuing in such an iterative way, we can construct a positive weakly convex function that has unboundedly many stationary points. For concreteness, we can write the following exact form for the gradient of the function, if we let $m = \frac{\gamma}{\gamma-2}$:

$$R'_{wc} = (m + \gamma) \, m^n - \gamma x, \quad \text{for } x \in [m^n, m^{n+1}), n \in \mathbb{Z}$$

Figure 3 illustrates what the resulting functions looks like, and in particular, we can see that for any $\gamma > 2$, we have infinitely many stationary points.
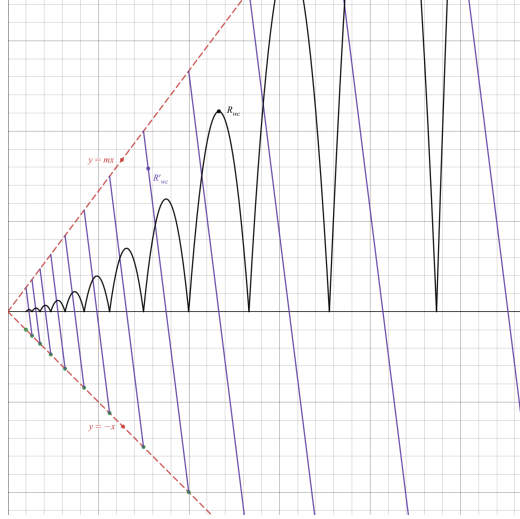


Figure 3: Example of a positive weakly convex function with $\gamma > 2\mu$ and unbounded stationary points.

# D. Proofs and a counterexample for Section 3

### D.1. Proof of Theorem 3.3

*Proof.* One of the main limitations in achieving provable stability and convergence guarantees lies in the ability of proving boundedness of iterates $x_k$. However, thanks to the form of the regulariser, this is achieved almost immediately. By Lemma B.1, for all $u \in \mathcal{X}$,

$$\langle \partial R_{sc}(x_k) - \partial R_{sc}(u), x_k - u \rangle_{\mathcal{X}^* \times \mathcal{X}} \geqslant \mu \|x_k - u\|_{\mathcal{X}}^2$$

and for $\mathcal{D}$ convex: for all $x, z \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\mathcal{D}(\mathcal{A} z, y) \geqslant \mathcal{D}(\mathcal{A} z, y) - \mathcal{D}(\mathcal{A} x, y) \geqslant \langle \partial_x \mathcal{D}(\mathcal{A} x, y), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}}$$

For any critical point $x_k$ and $z \in \mathcal{X}$, we have:

$$
\begin{aligned}
0 &\in \langle \partial_x \mathcal{D}(\mathcal{A} x_k, y_k) + \alpha \left( \partial \mathcal{R}_{wc}(x_k) + \partial \mathcal{R}_{sc}(x_k) \right), z - x_k \rangle_{\mathcal{X}^* \times \mathcal{X}} \\
&\subseteq \langle \partial_x \mathcal{D}(\mathcal{A} x_k, y_k), z - x \rangle_{\mathcal{X}^* \times \mathcal{X}} - \alpha \langle \partial \mathcal{R}_{sc}(x_k) - \partial \mathcal{R}_{sc}(z), x_k - z \rangle_{\mathcal{X}^* \times \mathcal{X}} + \alpha \langle \partial \mathcal{R}_{wc}(x_k) + \partial \mathcal{R}_{sc}(z), z - x_k \rangle_{\mathcal{X}^* \times \mathcal{X}} \\
&\leqslant \mathcal{D}(\mathcal{A} z, y_k) - \alpha \langle \partial \mathcal{R}_{sc}(x_k) - \partial \mathcal{R}_{sc}(z), x_k - z \rangle_{\times \mathcal{X}} + \alpha \langle \partial \mathcal{R}_{wc}(x_k) + \partial \mathcal{R}_{sc}(z), z - x_k \rangle_{\mathcal{X}}.
\end{aligned}
$$

Therefore, since $\gamma < 2\mu$, by Theorem 2.6 and by rearranging and completing the square, we get:

$$\left( \|x_k - z\|_{\mathcal{X}} - \frac{1}{2(\mu - \frac{\gamma}{2})} \left( \|\partial \mathcal{R}_{sc}(z)\| \right) \right)^2 \leqslant \frac{1}{\alpha(\mu - \frac{\gamma}{2})} \mathcal{D}(\mathcal{A} z, y_k) + \frac{1}{4(\mu - \frac{\gamma}{2})^2} \left( \|\partial \mathcal{R}_{sc}(z)\| \right)^2 + \frac{1}{(\mu - \frac{\gamma}{2})} \mathcal{R}_{wc}(z), \tag{D.1}$$

which together with convergence of $y_k$ and continuity of $\mathcal{D}$ implies that the $(x_k)$ are bounded, and hence have a weak convergent subsequence since $\mathcal{X}$ is reflexive, via the Banach–Alaoglu theorem. All that remains is to show that all cluster points $x_+$ must be critical points. Passing to a subsequence, let $x_k \rightharpoonup x_+$. Then in the case of $\gamma \leqslant \mu$:

$$\mathcal{D}(\mathcal{A}\,x_+, y^\delta) + \alpha\mathcal{R}(x_+) \leqslant \liminf_k \mathcal{D}(\mathcal{A}\,x_k, y_k) + \alpha\mathcal{R}(x_k) \quad \text{since } \mathcal{R} \text{ and } \mathcal{D} \text{ are weakly sequentially l.s.c.}$$

$$\leqslant \liminf_k \mathcal{D}(\mathcal{A}\,u, y_k) + \alpha\mathcal{R}(u) + \frac{\alpha}{2}(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2 \quad \forall u \in \mathcal{X}, \text{ by weak convexity}$$

$$\leqslant \mathcal{D}(\mathcal{A}\,u, y^\delta) + \alpha\mathcal{R}(u) + \frac{\alpha}{2}(\gamma - \mu)\|x_+ - u\|_{\mathcal{X}}^2 \quad \forall u \in \mathcal{X},,$$

where the last line follows since $x_k \rightharpoonup x_+$ and $\|\cdot\|_{\mathcal{X}}$ is weakly sequentially l.s.c., and therefore $0 \in \partial\mathcal{J}_{\alpha, y^\delta}(x_+)$ by definition of weak convexity.

In the case of $\mu < \gamma < 2\mu$, $R_{sc} - \frac{1}{2}\mu\|\cdot\|_{\mathcal{X}}^2$ weakly sequentially l.s.c.,

$$\mathcal{D}(\mathcal{A}\,x_+, y^\delta) + \alpha\mathcal{R}(x_+) - \frac{\alpha}{2}(\gamma - \mu)\|x_+ - u\|_{\mathcal{X}}^2$$

$$= \mathcal{D}(\mathcal{A}\,x_+, y^\delta) + \alpha\mathcal{R}(x_+) - \frac{\alpha}{2}(\gamma - \mu)\|x_+\|_{\mathcal{X}}^2 + \alpha(\gamma - \mu)\langle x_+, u\rangle_{\mathcal{X}} - \frac{\alpha}{2}(\gamma - \mu)\|u\|_{\mathcal{X}}^2$$

$$= \mathcal{D}(\mathcal{A}\,x_+, y^\delta) + \alpha\left(\mathcal{R}(x_+) - \frac{\mu}{2}\|x_+\|_{\mathcal{X}}^2\right) + \frac{\alpha}{2}(2\mu - \gamma)\|x_+\|_{\mathcal{X}}^2 + \alpha(\gamma - \mu)\langle x_+, u\rangle_{\mathcal{X}} - \frac{\alpha}{2}(\gamma - \mu)\|u\|_{\mathcal{X}}^2$$

$$\leqslant \liminf_k \mathcal{D}(\mathcal{A}\,x_k, y_k) + \alpha\left(\mathcal{R}(x_k) - \frac{\mu}{2}\|x_k\|_{\mathcal{X}}^2\right) + \frac{\alpha}{2}(2\mu - \gamma)\|x_k\|_{\mathcal{X}}^2 + \alpha(\gamma - \mu)\langle x_k, u\rangle_{\mathcal{X}} - \frac{\alpha}{2}(\gamma - \mu)\|u\|_{\mathcal{X}}^2$$

$$\text{since } \mathcal{R} - \frac{\mu}{2}\|\cdot\|_{\mathcal{X}}^2, \|\cdot\|_{\mathcal{X}}, \text{ and } \mathcal{D} \text{ are weakly sequentially l.s.c., and inner product is weakly continuous}$$

$$= \liminf_k \mathcal{D}(\mathcal{A}\,x_k, y_k) + \alpha\mathcal{R}(x_k) - \frac{\alpha}{2}(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2$$

$$\leqslant \liminf_k \mathcal{D}(\mathcal{A}\,u, y_k) + \alpha\mathcal{R}(u) \quad \forall u \in \mathcal{X}, \text{ by weak convexity and the criticality of } x_k$$

$$= \mathcal{D}(\mathcal{A}\,u, y^\delta) + \alpha\mathcal{R}(u) \quad \forall u \in \mathcal{X}, \text{ since } y_k \to y^\delta \text{ and } \mathcal{D} \text{ is continuous in its second argument.}$$

$\square$

## D.2. Proof of Theorem 3.4

*Proof.* Similar to the theorem above, we first want to show boundedness of iterates $x_k$. To achieve this, we use the fact that $y^0 = \mathcal{A}\,u$, and therefore by Assumption 3.1(5)

$$\mathcal{D}(\mathcal{A}\,u, y^{\delta_k}) = \mathcal{D}(y^0, y^{\delta_k}) \leqslant C\left(\mathcal{D}(y^0, y^0) + \|y^{\delta_k} - y^0\|^p\right) = C\delta_k^p.$$

Hence, by (D.1):

$$\left(\|x_k - u\|_{\mathcal{X}} - \frac{1}{2(\mu - \frac{\gamma}{2})}\left(\|\partial\mathcal{R}_{sc}(u)\|\right)\right)^2 \leqslant \frac{C}{\alpha_k(\mu - \frac{\gamma}{2})}\delta_k^p + \frac{1}{4(\mu - \frac{\gamma}{2})^2}\left(\|\partial\mathcal{R}_{sc}(u)\|\right)^2 + \frac{1}{(\mu - \frac{\gamma}{2})}\mathcal{R}_{wc}(u),$$

Therefore, as in the previous theorem, $(x_k)$ weakly has convergent subsequences (due to reflexivity and the Banach–Alaoglu theorem) since $\delta_k^p/\alpha_k$ is assumed to converge. Passing to such a subsequence, let $x_k \rightharpoonup x_+$. We wish to show that $x_+$ is $\mathcal{R}$-criticising. We first show that $\mathcal{A}\,x_+ = y^0$:

$$0 \leqslant \mathcal{D}(\mathcal{A}\,x_+, y^0) \leqslant \liminf_k \mathcal{D}(\mathcal{A}\,x_k, y^{\delta_k}) \quad \text{as } \mathcal{D} \text{ is weakly sequentially l.s.c.}$$

$$\leqslant \liminf_k \mathcal{D}(\mathcal{A}\,x_k, y^{\delta_k}) + \alpha_k\mathcal{R}(x_k) \quad \text{as } \alpha_k\mathcal{R}(x_k) \geqslant 0$$

$$\leqslant \liminf_k \mathcal{D}(\mathcal{A}\,u, y^{\delta_k}) + \alpha_k\mathcal{R}(u) + \frac{1}{2}\alpha_k(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2 \quad \text{since } x_k \in \text{crit } \mathcal{J}_{\alpha_k, y^{\delta_k}}, \text{ where } \mathcal{A}\,u = y^0$$

$$\leqslant \liminf_k C\delta_k^p + \alpha_k\mathcal{R}(u) + \frac{1}{2}\alpha_k(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2 \quad \text{by Assumption 3.1(5)}$$

$$= 0,$$

hence $\mathcal{A}\,x_+ = y^0$. It remains to prove that, for all $u \in \mathcal{X}$ such that $\mathcal{A}\,u = y^0$,

$$\mathcal{R}(u) \geqslant \mathcal{R}(x_+) + \frac{1}{2}(\mu - \gamma)\|u - x_+\|_{\mathcal{X}}^2.$$

By the criticality of $x_k$, the non-negativity of $\mathcal{D}$, and Assumption 3.1(5):

$$\alpha_k \mathcal{R}(x_k) \leqslant \mathcal{D}(\mathcal{A}\,x_k, y^{\delta_k}) + \alpha_k \mathcal{R}(x_k) \leqslant C\delta_k^p + \alpha_k \mathcal{R}(u) + \frac{1}{2}\alpha_k(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2,$$

and so

$$\mathcal{R}(x_k) \leqslant C\frac{\delta_k^p}{\alpha_k} + \mathcal{R}(u) + \frac{1}{2}(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2. \tag{D.2}$$

Therefore, if $\gamma \leqslant \mu$,

$$\mathcal{R}(x_+) \leqslant \liminf_k \mathcal{R}(x_k) \quad \text{as } \mathcal{R} \text{ is weakly sequentially l.s.c.}$$

$$\leqslant \liminf_k C\frac{\delta_k^p}{\alpha_k} + \mathcal{R}(u) + \frac{1}{2}(\gamma - \mu)\|x_k - u\|_{\mathcal{X}}^2 \quad \text{by (D.2)}$$

$$\leqslant \mathcal{R}(u) + \frac{1}{2}(\gamma - \mu)\|x_+ - u\|_{\mathcal{X}}^2, \quad \text{since } \|\cdot\|_{\mathcal{X}} \text{ is weakly sequentially l.s.c.,}$$

as desired. In the case $\mu < \gamma < 2\mu$, $R_{sc} - \frac{1}{2}\mu\|\cdot\|_{\mathcal{X}}^2$ is weakly sequentially l.s.c., the proof runs:

$$\mathcal{R}(x_+) - \frac{1}{2}\mu\|x_+ - u\|_{\mathcal{X}}^2 = \mathcal{R}(x_+) - \frac{1}{2}\mu\|x_+\| + \mu\langle x_+, u\rangle - \frac{1}{2}\mu\|u\|_{\mathcal{X}}^2$$

$$\leqslant \liminf_k \mathcal{R}(x_k) - \frac{1}{2}\mu\|x_k\|_{\mathcal{X}}^2 + \mu\langle x_k, u\rangle - \frac{1}{2}\mu\|u\|_{\mathcal{X}}^2$$

$$= \liminf_k \mathcal{R}(x_k) - \frac{1}{2}\mu\|x_k - u\|_{\mathcal{X}}^2$$

$$\leqslant \liminf_k C\frac{\delta_k^p}{\alpha_k} + \mathcal{R}(u) + \frac{1}{2}(\gamma - 2\mu)\|x_k - u\|_{\mathcal{X}}^2$$

$$\leqslant \mathcal{R}(u) + \frac{1}{2}(\gamma - 2\mu)\|x_+ - u\|_{\mathcal{X}}^2.$$

Finally, whenever the solution to $\mathcal{A}\,u = y^0$ is unique, then by the above every subsequence of $x_k$ has a subsubsequence converging weakly to that $u$. It follows that $x_k$ converges weakly to $u$, as if $x_k$ did not converge weakly to $u$, then there would be a neighbourhood $\mathcal{N}$ of $u$ (in the weak topology) and a subsequence $x'_\ell$ of $x_k$ such that $x'_\ell \notin \mathcal{N}$ for all $\ell$. This subsequence would have no subsubsequence converging weakly to $u$, contradicting the above. $\qquad\square$

### D.3. Counterexample to Theorems 3.3 and 3.4 in the general weakly convex setting

**Example D.1.** Let:

- $\mathcal{X} := \ell^2(\mathbb{N})$, i.e. $\mathcal{X} := \{(a_n)_{n\in\mathbb{N}} : \sum_{n=1}^{\infty} a_n^2 < \infty\}$ with $\langle (a_n), (b_n)\rangle_{\mathcal{X}} := \sum_{n=1}^{\infty} a_n b_n$,

- $\mathcal{Y} = \mathbb{R}$,

- $\mathcal{A} = 0$,

- $\mathcal{D}(y, z) := (y - z)^2$ (which can be checked to satisfy Assumption 3.1(4-5)) and so $\mathcal{D}(\mathcal{A}\,x, y) = y^2$,

- $y_k = y^\delta = y^0 = 0$, and

- $\mathcal{R}(x) := \begin{cases} f(x), & \|x\|_{\mathcal{X}} \leqslant \frac{1}{2}, \\ \left(\|x\|_{\mathcal{X}}^2 - 1\right)^2, & \|x\|_{\mathcal{X}} > \frac{1}{2}, \end{cases}$ where $f$ is any smooth weakly convex function for which $0$ is not a critical
  point and for which every derivative (including the $0^{\text{th}}$) of $f$ at any $x$ with $\|x\|_{\mathcal{X}} = \frac{1}{2}$ agrees with that of $\left(\|x\|_{\mathcal{X}}^2 - 1\right)^2$.

It follows that $\mathcal{R}$ is weakly convex because there exists $\rho \geqslant 0$ such that $f + \rho\|x\|_{\mathcal{X}}^2$ is convex, and $\left(\|x\|_{\mathcal{X}}^2 - 1\right)^2 + 2\|x\|_{\mathcal{X}}^2 = \|x\|_{\mathcal{X}}^4 + 1$ is convex since it is the composition of $t \mapsto t^4 + 1$ (which is a convex and monotone on $[0, \infty)$) with $\|\cdot\|_{\mathcal{X}}$ which is convex and has range $[0, \infty)$. Hence $\mathcal{R}(x) + \max\{\rho, 2\}\|x\|_{\mathcal{X}}$ is convex.

Then for all $\alpha > 0$ and $y \in \mathcal{Y}$, $\partial \mathcal{J}_{\alpha, y}(x) = \alpha \partial \mathcal{R}(x)$. Hence, for all $\alpha > 0$, $x_k \in \mathrm{crit}_{x \in \mathcal{X}} \mathcal{J}_{\alpha, y_k}$ if and only if $x_k \in \mathrm{crit}_{x \in \mathcal{X}} \mathcal{R}$. But by construction $\{\|x\|_{\mathcal{X}} = 1\} \subseteq \mathrm{crit}_{x \in \mathcal{X}} \mathcal{R}$ and $0 \notin \mathrm{crit}_{x \in \mathcal{X}} \mathcal{R}$. Therefore, we can choose $x_k = e_k$, i.e. the sequence which is 1 when $n = k$ and 0 otherwise. Then it is a well-known fact that $e_k \rightharpoonup 0$, which by construction is not a critical point of $\mathcal{J}_{\alpha, y^\delta}$. Hence Theorem 3.3 does not hold.

For Theorem 3.4, every $x \in \mathcal{X}$ has $\mathcal{A} x = y^0$. It follows that $x_+$ is an $\mathcal{R}$-criticising solution if and only if $x_+$ is a critical point of $\mathcal{R}$. We can once again take $x_k = e_k \in \mathrm{crit}_{x \in \mathcal{X}} \mathcal{J}_{\alpha_k, y_k}$, with weak limit 0, which is not a critical point of $\mathcal{R}$. Hence, Example D.1 is also a counterexample to Theorem 3.4 in the general setting of $\mathcal{R}$ weak convex.

**Note D.2.** The problematic critical points here can be chosen to indeed be global minimisers of $\mathcal{R}$, so this issue is not caused by the fact that we are in the setting of critical points.

**Note D.3.** We could modify $\mathcal{R}$ to also be globally Lipschitz by setting $\mathcal{R}(x) = g(x)$ for $\|x\|_{\mathcal{X}} > 2$ where $g$ is any smooth, globally Lipschitz, weakly convex function whose derivatives all agree with those of $\left(\|x\|_{\mathcal{X}}^2 - 1\right)^2$ at $\|x\|_{\mathcal{X}} = 2$.

**Note D.4.** This counterexample does not work in the finite-dimensional setting, as in that case the weak and norm topologies coincide, and hence $\{\|x\|_{\mathcal{X}} = 1\}$ contains all its limit points, and so does not have 0 as a limit point.

# E. Proofs and definitions for Section 4

## E.1. Proof of Theorem 4.1

*Proof.* Written this way, we have by weak convexity-concavity of $L(x, y)$, assuming that $F^*$ is $\mu$-strongly convex, that

$$
\begin{aligned}
L\left(x^{k+1}, y\right) - L\left(x, y^{k+1}\right) &= L\left(x^{k+1}, y\right) - L\left(x^{k+1}, y^{k+1}\right) + L\left(x^{k+1}, y^{k+1}\right) - L\left(x, y^{k+1}\right) \\
&\leqslant \left\langle T z^{k+1}, z^{k+1} - z \right\rangle_{\mathcal{X} \times \mathcal{Y}} + \frac{\rho}{2}\|x^{k+1} - x\|_{\mathcal{X}}^2 - \frac{\mu}{2}\|y^{k+1} - y\|_{\mathcal{Y}}^2 \\
&= \left\langle z^k - z^{k+1}, z^{k+1} - z \right\rangle_M + \frac{\rho}{2}\|x^{k+1} - x\|_{\mathcal{X}}^2 - \frac{\mu}{2}\|y^{k+1} - y\|_{\mathcal{Y}}^2 \\
&= \frac{1}{2}\left\|z^k - z\right\|_M^2 - \frac{1}{2}\left\|z^{k+1} - z\right\|_M^2 - \frac{1}{2}\left\|z^k - z^{k+1}\right\|_M^2 + \frac{\rho}{2}\|x^{k+1} - x\|_{\mathcal{X}}^2 - \frac{\mu}{2}\|y^{k+1} - y\|_{\mathcal{Y}}^2.
\end{aligned}
$$

$$\text{(E.1)}$$

But then adding together the inequality above for $k$ and $k + 1$ evaluated at $z = (x^k, y^{k+1})$

$$
\begin{aligned}
L&\left(x^{k+1}, y^{k+1}\right) - L\left(x^k, y^k\right) \\
&= L\left(x^{k+1}, y^{k+1}\right) - L\left(x^k, y^{k+1}\right) + L\left(x^k, y^{k+1}\right) - L\left(x^k, y^k\right) \\
&\leqslant \frac{1}{2}\left\|y^k - y^{k+1}\right\|_M^2 - \frac{1}{2}\left\|x^{k+1} - x^k\right\|_M^2 - \frac{1}{2}\left\|z^k - z^{k+1}\right\|_M^2 + \\
&\quad - \frac{1}{2}\left\|z^{k-1} - z^k\right\|_M^2 + \frac{\rho}{2}\|x^{k+1} - x^k\|_{\mathcal{X}}^2 - \frac{\mu}{2}\|y^{k+1} - y^k\|_{\mathcal{Y}}^2 + \frac{1}{2}\left\|\begin{pmatrix} x^{k-1} - x^k \\ y^{k-1} - y^{k+1} \end{pmatrix}\right\|_M^2 \\
&= \left\|y^k - y^{k+1}\right\|_M^2 - \frac{1}{2}\left\|x^k - x^{k+1}\right\|_M^2 - \frac{1}{2}\left\|z^k - z^{k+1}\right\|_M^2 + \frac{\rho}{2}\|x^k - x^{k+1}\|_{\mathcal{X}}^2 \\
&\quad - \frac{\mu}{2}\|y^{k+1} - y^k\|_{\mathcal{Y}}^2 + \left\langle \begin{pmatrix} x^{k-1} - x^k \\ y^{k-1} - y^k \end{pmatrix}, \begin{pmatrix} 0 \\ y^k - y^{k+1} \end{pmatrix} \right\rangle_M \\
&\leqslant \frac{3}{2}\left\|y^k - y^{k+1}\right\|_M^2 - \frac{1}{2}\left\|x^{k+1} - x^k\right\|_M^2 - \frac{1}{2}\left\|z^k - z^{k+1}\right\|_M^2 + \\
&\quad + \frac{1}{2}\left\|z^{k-1} - z^k\right\|_M^2 + \frac{\rho}{2}\|x^{k+1} - x^k\|_{\mathcal{X}}^2 - \frac{\mu}{2}\|y^{k+1} - y^k\|_{\mathcal{Y}}^2.
\end{aligned}
$$

We can write this illustrating descent:

$$
\begin{aligned}
L\left(x^k, y^k\right) + \frac{1}{2}\left\|z^{k-1} - z^k\right\|_M^2 \geqslant\; & L\left(x^{k+1}, y^{k+1}\right) + \frac{1}{2}\left\|z^k - z^{k+1}\right\|_M^2 \\
& + \left[\frac{1}{2}(\mu\sigma - 3)\left\|y^k - y^{k+1}\right\|_M^2 + \frac{1}{2}(1 - \rho\tau)\left\|x^{k+1} - x^k\right\|_M^2\right].
\end{aligned}
$$

This becomes descent for the following restrictions on the parameters:

$$\begin{cases} \tau\sigma < \frac{1}{\|\mathcal{A}\|^2}, \\ \tau\rho < 1, \\ \mu\sigma > 3. \end{cases}$$

$\square$

### E.2. Proof of Theorem 4.2

*Proof.* In analogy with (Bolte et al., 2014), we shall show require two results - one on descent, as in Theorem 4.1, and one on subgradient boundedness, which arises from the following:

$$\begin{aligned} \left\| \partial\mathcal{L}(z^{k+1}, z^k) \right\| &= \left\| \begin{pmatrix} \begin{pmatrix} \partial_x L(z^{k+1}) \\ \partial_y L(z^{k+1}) \end{pmatrix} + M[z^{k+1} - z^k] \\ -M[z^{k+1} - z^k] \end{pmatrix} \right\| \\ &\leqslant \left\| \begin{pmatrix} \begin{pmatrix} \partial_x L(z^{k+1}) \\ \partial_y L(z^{k+1}) \end{pmatrix} \\ 0 \end{pmatrix} \right\| + \left\| \begin{pmatrix} M[z^{k+1} - z^k] \\ -M[z^{k+1} - z^k] \end{pmatrix} \right\| \\ &= 3\| M(z^{k+1} - z^k) \|. \end{aligned}$$

Overall, the proof is similar to that of (Guo et al., 2023) and (Bolte et al., 2014), and only the first part will be shown here, for compactness. Now, from Theorem 4.1:

$$\sum_{k=1}^{K} \left[ \frac{1}{2}(\mu\sigma - 3) \left\| y^k - y^{k+1} \right\|_M^2 + \frac{1}{2}(1 - \rho\tau) \left\| x^{k+1} - x^k \right\|_M^2 \right] \leqslant \mathcal{L}(z^1, z^0) - \mathcal{L}(z^{K+1}, z^K).$$

But by assumptions above $\inf_k L(x_k, y_k) > -\infty$ and by boundedness $\mathcal{L}^k > -\infty$, we have $\mathcal{L}^k$ is non-increasing and thus converges to $\mathcal{L}^*$. Thus, taking $K \to \infty$:

$$\sum_{k=1}^{\infty} \left[ \frac{1}{2}(\mu\sigma - 3) \left\| y^k - y^{k+1} \right\|_M^2 + \frac{1}{2}(1 - \rho\tau) \left\| x^{k+1} - x^k \right\|_M^2 \right] \leqslant \mathcal{L}^0 - \mathcal{L}^*.$$

This, in particular, implies that the series is square-summable and furthermore that

$$\lim_{k\to\infty} \left\| x^{k+1} - x^k \right\|_{\mathcal{X}} = 0 \text{ and } \lim_{k\to\infty} \left\| y^{k+1} - y^k \right\|_{\mathcal{Y}} = 0.$$

Recalling that $\text{dist}(0, \partial L(z_k)) = \text{dist}(0, T(z_k)) = \left\| z^{k+1} - z^k \right\|_M \leqslant \left\| y^{k+1} - y^k \right\|_M + \left\| x^{k+1} - x^k \right\|_M \to 0$ as $k \to \infty$, and for $\nu = \min\{\mu\sigma - 3, 1 - \rho\tau\}$:

$$\frac{\nu}{4} \left\| z^{k+1} - z^k \right\|_M^2 \leqslant \left[ \frac{1}{2}(\mu\sigma - 3) \left\| y^k - y^{k+1} \right\|_M^2 + \frac{1}{2}(1 - \rho\tau) \left\| x^{k+1} - x^k \right\|_M^2 \right].$$

But also from above we find that

$$\min_k \left[ \frac{1}{2}(\mu\sigma - 3) \left\| y^k - y^{k+1} \right\|_M^2 + \frac{1}{2}(1 - \rho\tau) \left\| x^{k+1} - x^k \right\|_M^2 \right] \leqslant \frac{1}{K} \left( \mathcal{L}(z^1, z^0) - \mathcal{L}(z^{K+1}, z^K) \right).$$

Which we can combine to find:

$$\min_k \text{dist}\left(0, \partial L(z^k)\right) \leqslant \frac{2}{(\nu K)^{1/2}} \sqrt{\mathcal{L}(z^1, z^0) - \mathcal{L}(z^{K+1}, z^K)}.$$

$\square$

### E.3. The Kurdyka–Łojasiewicz inequality

In (Kurdyka, 1998) Kurdyka provided a generalisation of the Łojasiewicz inequality, with extensions to the nonsmooth setting in (Bolte et al., 2007).

**Definition E.1.** Let $\eta \in (0, +\infty]$. Denote with $\Phi_\eta$ the set of all concave and continuous functions $\varphi : [0, \eta) \to [0, +\infty)$ which satisfy: $\varphi(0) = 0$; $\varphi$ is $\mathcal{C}^1$ on $(0, \eta)$ and continuous at 0; and for all $s \in (0, \eta)$, $\varphi'(s) > 0$.

**Definition E.2.** Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semicontinuous (l.s.c.). Then $\Psi$ is said to have the *Kurdyka–Łojasiewicz (KŁ) property* at a point $\hat{v} \in \mathrm{dom}\, \partial\Psi := \{v \in \mathbb{R}^d : \partial\Psi(v) \neq \varnothing\}$, if there exists $\eta \in (0, +\infty]$, a neighborhood $V$ of $\hat{v}$, and a function $\varphi \in \Phi_\eta$ such that for any

$$v \in V \cap \{v \in \mathcal{X} : \Psi(\hat{v}) < \Psi(v) < \Psi(\hat{v}) + \eta\}$$

the Kurdyka–Łojasiewicz inequality holds

$$\varphi'(\Psi(v) - \Psi(\hat{v})) \cdot \mathrm{dist}(\mathbf{0}, \partial\Psi(v)) \geqslant 1.$$

If $\varphi(s) = cs^{1-\theta}$ for $c > 0$ and $\theta \in [0, 1)$, then $\theta$ is called the KŁ exponent of $\Psi$ at $\hat{v}$. If $\Psi$ has the KŁ property at each point of $\mathrm{dom}\, \partial\Psi$, then $\Psi$ is called a KŁ function (or just, KŁ).

Examples of KŁ functions include semialgebraic, subanalytic, uniformly convex functions (see (Attouch et al., 2010) and the references therein) and all typical neural networks.

**Theorem E.3.** *Let $\Psi : \mathbb{R}^d \to \mathbb{R}$ be a deep neural network with every activation function a continuous piecewise analytic function with finitely many pieces (e.g.,* ReLU*,* sigmoid*). Then $\Psi$ is a KŁ function and for all $\hat{v} \in \mathrm{dom}\, \partial\Psi$, there exists $\theta \in [0, 1)$ such that $\Psi$ has KŁ exponent $\theta$ at $\hat{v}$.*

*Proof.* The network $\Psi$ is a finite composition of continuous piecewise analytic functions with finitely many pieces, and hence is a continuous piecewise analytic function with finitely many pieces. It follows that $\Psi$ is subanalytic and so the result follows by Theorem 3.1 of (Bolte et al., 2007). $\square$

### E.4. Proof of Theorem 4.6

Before providing the proof, we are first going to provide two lemmas:

**Lemma E.4.** *For $f : \mathcal{X} \to \mathbb{R}$ $\rho$-convex, $(x^i) \in \mathcal{X}$, $\bar{x}^i := \frac{1}{i} \sum_{j=1}^i x^j$, and all $N \in \mathbb{N}$,*

$$f(\bar{x}^N) \leqslant \bar{f} - \frac{\rho}{2N} \sum_{i=1}^N \frac{i-1}{i} \|x^i - \bar{x}^{i-1}\|_\mathcal{X}^2,$$

*and furthermore emphasising the weak convex case, for any $x \in \mathcal{X}$:*

$$f(\bar{x}^N) \leqslant \bar{f} + \max(0, -\rho) \cdot \left[ \frac{1}{N} \sum_{i=1}^N \|x^i - x\|_\mathcal{X}^2 + \frac{1}{N} \sum_{i=1}^N \frac{1}{i} \sum_{j=1}^{i-1} \|x^j - x\|_\mathcal{X}^2 \right].$$

*Proof.* This arises from the following fact for $\rho$-convex functions:

$$
\begin{aligned}
f\left( \frac{1}{N}(x_1 + \cdots + x_N) \right) &\leqslant \frac{1}{N} f(x_N) - \frac{\rho}{2} \frac{1}{N}\left(1 - \frac{1}{N}\right) \|x_n - \bar{x}_{n-1}\|_\mathcal{X}^2 \\
&\quad + \frac{N-1}{N} f\left( \frac{1}{N-1}(x_1 + \cdots + x_{N-1}) \right) \\
&\leqslant \cdots \leqslant \bar{f} - \frac{\rho}{2N} \sum_{i=1}^k \frac{i-1}{i} \|x_i - \bar{x}_{i-1}\|_\mathcal{X}^2.
\end{aligned}
\tag{E.2}
$$

To arrive at the second result we wish to expand the last term in terms of iterate lengths:

$$\sum_{i=1}^{N} \frac{i-1}{i} \|x_i - \bar{x}_{i-1}\|_{\mathcal{X}}^2 \leqslant 2 \sum_{i=1}^{N} \frac{i-1}{i} \left( \|x_i - x\|_{\mathcal{X}}^2 + \|x - \bar{x}_{i-1}\|_{\mathcal{X}}^2 \right)$$

$$\leqslant 2 \sum_{i=1}^{N} \frac{i-1}{i} \left( \|x_i - x\|_{\mathcal{X}}^2 + \left\| \frac{1}{i-1} \sum_{j=1}^{i-1} (x_j - x) \right\|_{\mathcal{X}}^2 \right)$$

By Jensen's inequality: $\leqslant 2 \sum_{i=1}^{N} \frac{i-1}{i} \left( \|x_i - x\|_{\mathcal{X}}^2 + \frac{1}{i-1} \sum_{j=1}^{i-1} \|x_j - x\|_{\mathcal{X}}^2 \right).$

Plugging this into (E.2) yields the desired result. $\qquad\square$

And now by application of Lemma E.4 to $L$:

**Lemma E.5.** *Let* $\bar{z}^k := \left( \bar{x}^k, \bar{y}^k \right) := \frac{1}{k} \sum_{i=1}^{k} z^i$ *be the average iterate. Assuming that* $L(x,y)$ *is* $\rho$-*convex in the first argument and* $\mu$-*concave in the second, for any* $k \geqslant 1$ *and* $z$:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) \leqslant \frac{\|z - z^0\|_M^2}{2k} - \frac{1}{2k} \sum_{i=1}^{k} \|z^{i-1} - z^i\|_M^2$$

$$- \frac{\rho}{2k} \sum_{i=1}^{k} \|x^i - x\|_{\mathcal{X}}^2 - \frac{\rho}{2k} \sum_{i=1}^{k} \frac{i-1}{i} \|x^i - \bar{x}^{i-1}\|_{\mathcal{X}}^2$$

$$- \frac{\mu}{2k} \sum_{i=1}^{k} \|y^i - y\|_{\mathcal{Y}}^2 - \frac{\mu}{2k} \sum_{i=1}^{k} \frac{i-1}{i} \|y^i - \bar{y}^{i-1}\|_{\mathcal{Y}}^2.$$

*Further, setting* $D = \text{diag}\left(\max\left(0, -\rho\right), \max\left(0, -\mu\right)\right)$:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) \leqslant \frac{\|z - z^0\|_M^2}{2k} + \frac{3}{2k} \sum_{i=1}^{k} \|z^i - z\|_D^2 + \frac{1}{k} \sum_{i=1}^{k} \frac{1}{i} \sum_{j=1}^{i-1} \|z^j - z\|_D^2.$$

*Proof.* We can expand the original as:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) = \left( L\left(\bar{x}^k, y\right) - L\left(x, y\right) \right) + \left( L\left(x, y\right) - L\left(x, \bar{y}^k\right) \right)$$

Noting that the first bracket is convex in $\bar{x}^k$ and second is convex in $\bar{y}^k$, we can use Lemma E.4 to bound:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) = \left(L\left(\bar{x}^k, y\right) - L\left(x, y\right)\right) + \left(L\left(x, y\right) - L\left(x, \bar{y}^k\right)\right)$$

$$\leqslant \frac{1}{k}\sum_{i=1}^{k} L\left(x^i, y\right) - L\left(x, y^i\right)$$

$$- \frac{\rho}{2k}\sum_{i=1}^{k}\|x^i - x\|_{\mathcal{X}}^2 - \frac{\rho}{2k}\sum_{i=1}^{k}\frac{i-1}{i}\|x^i - \bar{x}^{i-1}\|_{\mathcal{X}}^2$$

$$- \frac{\mu}{2k}\sum_{i=1}^{k}\|y^i - y\|_{\mathcal{Y}}^2 - \frac{\mu}{2k}\sum_{i=1}^{k}\frac{i-1}{i}\|y^i - \bar{y}^{i-1}\|_{\mathcal{Y}}^2$$

$$\text{by (E.1)} \leqslant \frac{\left\|z - z^0\right\|_M^2}{2k} - \frac{1}{2k}\sum_{i=1}^{k}\|z^{i-1} - z^i\|_M^2$$

$$- \frac{\rho}{2k}\sum_{i=1}^{k}\|x^i - x\|_{\mathcal{X}}^2 - \frac{\rho}{2k}\sum_{i=1}^{k}\frac{i-1}{i}\|x^i - \bar{x}^{i-1}\|_{\mathcal{X}}^2$$

$$- \frac{\mu}{2k}\sum_{i=1}^{k}\|y^i - y\|_{\mathcal{Y}}^2 - \frac{\mu}{2k}\sum_{i=1}^{k}\frac{i-1}{i}\|y^i - \bar{y}^{i-1}\|_{\mathcal{Y}}^2$$

Using the second part of Lemma E.4, we can arrive at the second result:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) \leqslant \frac{\left\|z - z^0\right\|_M^2}{2k} - \frac{\rho}{2k}\sum_{i=1}^{k}\|x^i - x\|_{\mathcal{X}}^2 - \frac{\mu}{2k}\sum_{i=1}^{k}\|y^i - y\|_{\mathcal{Y}}^2$$

$$+ \max\left(0, -\rho\right) \cdot \left[\frac{1}{k}\sum_{i=1}^{k}\|x^i - x\|_{\mathcal{X}}^2 + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|x^j - x\|_{\mathcal{X}}^2\right]$$

$$+ \max\left(0, -\mu\right) \cdot \left[\frac{1}{k}\sum_{i=1}^{k}\|y^i - y\|_{\mathcal{Y}}^2 + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|y^j - y\|_{\mathcal{Y}}^2\right]$$

Yielding the desired result. $\qquad\square$

Now, with the lemmas above we can prove Theorem 4.6.

*Proof of Theorem 4.6.* Using $\mu$-strong concavity in the second argument and $\rho$-weak convexity in the first, and using that $\hat{z}$ is a stationary point of $L$, we can bound:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) \leqslant -\frac{\mu}{2}\|y - \hat{y}\|_{\mathcal{Y}}^2 + \frac{\rho}{2}\|x - \hat{x}\|_{\mathcal{X}}^2 + L\left(\bar{x}^k, \hat{y}\right) - L\left(\hat{x}, \bar{y}^k\right).$$

Now, from the second part of Lemma E.5 we have

$$L\left(\bar{x}^k, \hat{y}\right) - L\left(\hat{x}, \bar{y}^k\right) \leqslant \frac{\left\|\hat{z} - z^0\right\|_M^2}{2k} + \frac{3}{2k}\sum_{i=1}^{k}\|z^i - \hat{z}\|_D^2 + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|z^j - \hat{z}\|_D^2,$$

while from Theorem 4.5 (by redefining the constants appropriately), then we have constants $\nu > 0, 0 < \tau < 1$, such that

$$\left\|z^k - \hat{z}\right\|_D \leqslant \nu\tau^k,$$

which in turn implies the following two bounds:

$$\frac{1}{k}\sum_{i=1}^{k}\|z^i - \hat{z}\|_D^2 \leqslant \frac{1}{k}\nu^2\sum_{i=1}^{k}\tau^{2k} = \frac{1}{k}\nu^2\tau^2\frac{1 - \tau^{2k}}{1 - \tau^2} = \mathcal{O}\left(\frac{1}{k}\right)$$

and

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|z^j - \hat{z}\|_D^2 \leqslant \frac{1}{k}\nu^2\tau^2\sum_{i=1}^{k}\frac{1}{i}\cdot\frac{1-\tau^{2(i-1)}}{1-\tau^2}$$

$$\leqslant \frac{1}{k}\frac{\nu^2\tau^2}{1-\tau^2}\sum_{i=1}^{k}\frac{1}{i} \leqslant \frac{\nu^2\tau^2}{1-\tau^2}\frac{\log k + 1}{k} = \mathcal{O}\left(\frac{\log k}{k}\right),$$

by a standard inequality bounding harmonic series.

If instead $\theta \in \left(\frac{1}{2}, 1\right)$, then there exist a constant $\mu > 0$,

$$\|z^k - \hat{z}\|_D \leqslant \mu k^{-\frac{1-\theta}{2\theta-1}}.$$

But this implies the following two bounds:

$$\frac{1}{k}\sum_{i=1}^{k}\|z^i - \hat{z}\|_D^2 \leqslant \frac{1}{k}\mu^2\sum_{i=1}^{k}i^{-2\frac{1-\theta}{2\theta-1}} \leqslant \frac{1}{k}\mu^2\zeta\left(2\frac{1-\theta}{2\theta-1}\right) = \mathcal{O}\left(\frac{1}{k}\right)$$

and

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|z^j - \hat{z}\|_D^2 \leqslant \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\mu^2\zeta\left(2\frac{1-\theta}{2\theta-1}\right)$$

$$\leqslant \frac{1}{k}\mu^2\zeta\left(2\frac{1-\theta}{2\theta-1}\right)(\log k + 1) = \mathcal{O}\left(\frac{\log k}{k}\right),$$

where $\zeta$ denotes the Riemann zeta function.

Now, if $\theta = 0$, denoting $\sum_{i=1}^{\infty}\|z^i - \hat{z}\|_D^2$ as $d$, which is guarranteed to be finite, since we converge in a finite number of steps by Theorem 4.5, we have:

$$\frac{1}{k}\sum_{i=1}^{k}\|z^i - \hat{z}\|_D^2 \leqslant \frac{d}{k} = \mathcal{O}\left(\frac{1}{k}\right)$$

and

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i-1}\|z^j - \hat{z}\|_D^2 \leqslant \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}d \leqslant d\frac{\log k + 1}{k} = \mathcal{O}\left(\frac{\log k}{k}\right).$$

Therefore, in all three cases we achieve the following bound:

$$L\left(\bar{x}^k, y\right) - L\left(x, \bar{y}^k\right) \leqslant \frac{1}{2}\|z - \hat{z}\|_D^2 + \mathcal{O}\left(\frac{\log k}{k}\right).$$

$\square$

## F. Proof of Theorem 5.2

Before we can move onto proof of Theorem 5.2, we are going to need to establish a number of results relating to both universal approximation properties of standard neural networks, as well as approximation properties of Moreau envelopes.

**Lemma F.1** (Lemma 3 in (Strömberg, 1996)). *Let $f$ be a proper extended-real-valued l.s.c. bounded below function on $\mathcal{X}$.*

*For any $t > 0$ it holds that*

  *(i) $f_t$ is a real-valued minorant of $f$;*

  *(ii) $\inf f_t = \inf f$;*

*(iii)* $\arg\min f_t = \arg\min f$.

*Furthermore, when $t \downarrow 0$:*

*(iv)* $f_t \to f$ *pointwise;*

*(v)* $f_t \to f$ *with respect to the epi-distance topology provided $\|\cdot\|_{\mathcal{X}}$ is uniformly rotund;*

*(vi)* $f_t \to f$ *uniformly on bounded sets when $f$ is uniformly continuous on bounded sets and $\|\cdot\|_{\mathcal{X}}$ is uniformly rotund.*

Note, that e.g. the Euclidean norm is uniformly rotund. In what follows we use point (vi) in order to approximate the desired function with a weak convex smooth approximant.

**Theorem F.2** (Representation power of ICNN; Theorem 1 in (Chen et al., 2019))**.** *For any Lipschitz convex function over a compact domain, there exists a neural network with nonnegative weights and ReLU activation functions (i.e., an ICNN) that approximates it uniformly within $\varepsilon$, i.e. in the $\|\cdot\|_{\infty}$ norm.*

A further summary can be found in (Hanin, 2019). The last piece for the proof of universal approximation of IWCNNs is a universal approximation results for a neural network with smooth activations:

**Theorem F.3** (Theorem 3.2 in (Kidger & Lyons, 2020))**.** *Let $\rho : \mathbb{R} \to \mathbb{R}$ be any nonaffine continuous function which is continuously differentiable at at least one point, with nonzero derivative at that point. Let $K \subseteq \mathbb{R}^n$ be compact. Then $\mathcal{NN}^{\rho}_{n,m,n+m+2}$ is dense in $C(K;\mathbb{R}^m)$ with respect to the uniform norm.*

Here $\mathcal{NN}^{\rho}_{n,m,n+m+2}$ represents the class of functions $\mathbb{R}^n \to \mathbb{R}^m$ described by feedforward neural networks with $n$ neurons in the input layer, $m$ neurons in the output layer, and an arbitrary number of hidden layers, each with $n+m+2$ neurons and activation function $\rho$ as in (Kidger & Lyons, 2020).

With these results, we can now prove Theorem 5.2.

*Proof of Theorem 5.2.* Fix $\varepsilon > 0$. Consider some target proper extended-real-valued l.s.c. function $f \in \mathcal{C}_0(X)$, with $X$ compact. Then, by (Sun & Yu, 2019) Theorem 6, there exists $f_{wc} \in \mathcal{WC}(X) \cap \mathcal{C}_0(X)$ with $\|f - f_{wc}\|_{\infty} < \frac{\varepsilon}{2}$. In fact, we know exactly how to find such approximation using the Moreau envelope: $f_{wc} = f_t$ for some $t > 0$ by Lemma F.1, which furthermore is Lipschitz smooth by Theorem A.4. As a result, we can write $f_{wc} = \mathrm{Id} \circ f_t$. Now, more generically for $m \neq 1$, we can e.g. consider just the first element, i.e. $g^{\mathrm{conv}} : x \to x_1$ and $g^{\mathrm{sm}} : x \to (e_t f, 0, \cdots, 0)^{\top}$, with $f_{wc} = g^{\mathrm{conv}} \circ g^{\mathrm{sm}}$.

By Theorem F.3 there exists a neural network such that $\|g^{\mathrm{sm}}_{\theta_2} - g^{\mathrm{sm}}\|_{\infty} < \frac{\varepsilon}{2}$. Since $g^{\mathrm{sm}}_{\theta_2}$ is continuous and $X$ is compact, $g^{\mathrm{sm}}_{\theta_2}(X)$ is compact. Therefore, by Theorem F.2, as $g^{\mathrm{conv}}$ is Lipschitz convex, there exists an ICNN such that $\|g^{\mathrm{ICNN}}_{\theta_1} - g^{\mathrm{conv}}\|_{\infty} < \frac{\varepsilon}{2}$ on $\mathrm{Im}\, g^{\mathrm{sm}}_{\theta_2}$[1]. Therefore,

$$
\begin{aligned}
\|g^{\mathrm{ICNN}}_{\theta_1} \circ g^{\mathrm{sm}}_{\theta_2} - g^{\mathrm{conv}} \circ g^{\mathrm{sm}}\|_{\infty} &\leqslant \|g^{\mathrm{ICNN}}_{\theta_1} \circ g^{\mathrm{sm}}_{\theta_2} - g^{\mathrm{conv}} \circ g^{\mathrm{sm}}_{\theta_2}\|_{\infty} + \|g^{\mathrm{conv}} \circ g^{\mathrm{sm}}_{\theta_2} - g^{\mathrm{conv}} \circ g^{\mathrm{sm}}\|_{\infty} \\
&\leqslant \|g^{\mathrm{ICNN}}_{\theta_1} - g^{\mathrm{conv}}\|_{\infty} + \|g^{\mathrm{sm}}_{\theta_2} - g^{\mathrm{sm}}\|_{\infty} \\
&< \varepsilon,
\end{aligned}
$$

by 1-Lipschitzness of $g^{\mathrm{conv}}$. Therefore, IWCNNs are dense in the space of continuous functions. $\qquad\square$

# G. Experimental set-up and additional data visualisations for Section 6

### G.1. Experimental set-up

We use human abdominal CT scans for 10 patients provided by Mayo Clinic for the low-dose CT grand challenge (McCollough, 2014). The training dataset for CT experiments consists of a total of 2250 2D slices, each of dimension $512 \times 512$, corresponding to 9 patients. The remaining 128 slices corresponding to one patient are used to evaluate the reconstruction performance.

Projection data is simulated in ODL (Adler et al., 2017) with a GPU-accelerated *astra* back-end, using a parallel-beam acquisition geometry with 350 angles and 700 rays/angle, using additive Gaussian noise with $\sigma = 3.2$. The pseudoinverse

---

[1] Technically in this we are approximating the identity map, which can be exact with `ReLU`/`Leaky ReLU` activations.

reconstruction is taken to be images obtained using FBP. For limited angle experiments, data is simulated with a missing angular wedge of $60°$. The native `odl` power method is used to approximate the norm of the operator.

The TV method was computed by employing the ADMM-based solver in ODL.

All of the methods are implemented in *PyTorch* (Paszke et al., 2017). The LPD method is trained on pairs of target images and projected data, while the U-Net post-processor is trained on pairs of true images and the corresponding reconstructed FBPs. AR, ACR, ACNCR, and AWCR, in contrast, require ground-truth and FBP images drawn from their marginal distributions (and hence not necessarily paired). The hyperparameter $\mu_0$ from Definition 5.5 is chosen to be the same as in (Mukherjee et al., 2021): $\mu_0 = \log\left(1 + \exp(-9.0)\right)$, and during training diverges from this value minimally, having almost no effect on the convexity constant of the regulariser. To aid with training, $\lambda$ from (5.1) is first chosen to be small (0.1) and once the network is trained is increased to a larger value (10) as in (Mukherjee et al., 2021; Lunz et al., 2018)), and the network is fine-tuned. The *RMSprop* optimizer, following the recommendation in (Lunz et al., 2018; Mukherjee et al., 2021) with a learning rate of $5 \times 10^{-5}$ is used for training for a total of 50 epochs. For fine-tuning, the learning rate is reduced to $10^{-6}$. The AWCR architecture differed for the two experiments in the following way:

**Sparse view CT** The ICNN component of the AWCR is constructed with 5 convolutional layers, using `LeakyReLU` activations and $5 \times 5$ kernels with 16 channels. The smooth component of the AWCR is constructed using 6 convolutional layers, using `nn.SiLU` activations and $5 \times 5$ kernels with a doubling number of channels from 16 and stride of 2, similar to that of (Lunz et al., 2018), with the last layer containing 128 channels

**Limited view CT** The ICNN component of the AWCR is constructed with 5 convolutional layers, using `LeakyReLU` activations and $5 \times 5$ kernels with 16 channels. Guided by the observations of (Mukherjee et al., 2021), that reducing the total number of layers and the number of feature channels helps avoid overfitting in the limited angle setting, the smooth component of the AWCR is constructed using a single convolutional layers, using a `SiLU` activation, $7 \times 7$ with 32 channels. To remain close to the ACR formulation, a residual structure is further employed.

The reconstruction in all adversarial regularisation cases is performed by solving the variational problem via gradient-descent for 1000 iterations with a step size of $10^{-6}$. For the AWCR-PD, the primal-dual algorithm is used for solving the variational problem. The proximal operator of the network is approximated by performing gradient descent with backtracking on the objective in (4.2). The step sizes are chosen to be equal to $0.1/\|\mathcal{A}\|$ in the sparse case, and in the limited angle are chosen to be equal to $10/\|\mathcal{A}\|$ and $0.1/\|\mathcal{A}\|$ to ensure that iterates do not diverge and that steps are large enough to provide good results. Akin to AR, reconstruction performance of AWCR can sometimes deteriorate if early stopping is not applied. For a fair comparison, we report the highest PSNR achieved by all methods during reconstruction. The regularisation parameter $\alpha$ is chosen according to (Lunz et al., 2018) and is *not* tuned.

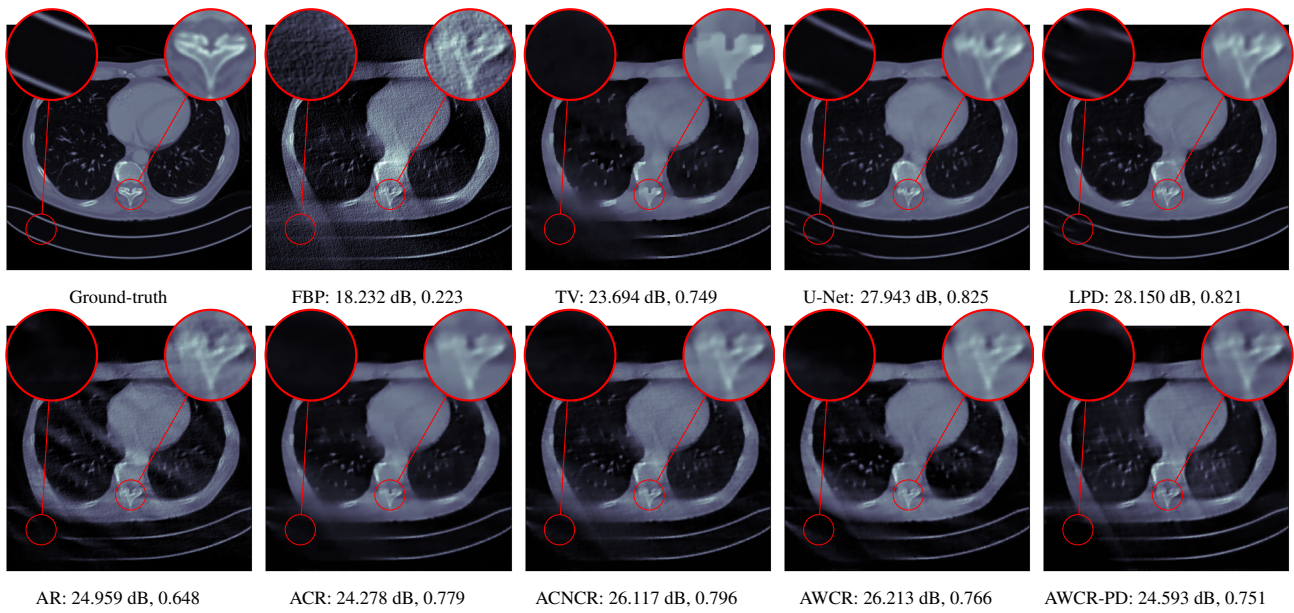### G.2. Additional data visualisations

Figure 4: Reconstructed images obtained using different methods, along with the associated PSNR and SSIM, for limited view CT.