On Optimal Regularization Parameters via Bilevel Learning

Matthias J. Ehrhardt, Silvia Gazzola, and Sebastian J. Scott *

Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, UK

Abstract

Variational regularization is commonly used to solve linear inverse problems, and involves augmenting a data fidelity by a regularizer. The regularizer is used to promote a priori information, and is weighted by a regularization parameter. Selection of an appropriate regularization parameter is critical, with various choices leading to very different reconstructions. Existing strategies such as the discrepancy principle and L-curve can be used to determine a suitable parameter value, but in recent years a supervised machine learning approach called bilevel learning has been employed. Bilevel learning is a powerful framework to determine optimal parameters, and involves solving a nested optimisation problem. While previous strategies enjoy various theoretical results, the well-posedness of bilevel learning in this setting is still a developing field. One necessary property is positivity of the determined regularization parameter. In this work, we provide a new condition that better characterises positivity of optimal regularization parameters than the existing theory. Numerical results verify and explore this new condition for both small and large dimensional problems.

Keywords: Inverse problems; Machine learning; Variational regularization; Bilevel learning; Imaging; Regularization parameter

1 Introduction

Inverse problems are a class of mathematical problems where one is tasked to determine the input to a system given the output of the system, along with some knowledge about the properties of said system. Such problems arise in many important science and engineering applications such as biomedical, astronomical, and seismic imaging [5, 8, 24, 32].

We consider the class of linear inverse problems wherein we are interested in retrieving the ground truth input $x^\star \in \mathbb{R}^n$ given a matrix $A \in \mathbb{R}^{m\times n}$, and corrupted measurement $y \in \mathbb{R}^m$ satisfying

$$y \approx Ax^\star. \tag{1}$$

The challenge with inverse problems such as (1) is that almost all interesting applications are ill-posed in the sense of Hadamard [30], in that at least one of the following conditions regarding solutions is violated: existence; uniqueness; continuity with respect to the observed measurement

*Corresponding author: ss2767@bath.ac.uk

y. A classical approach to remedy the ill-posedness of (1) is via variational regularization [8, 15], wherein one solves a minimisation problem such as

$$\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|Ax-y\|^2+\alpha\mathcal{R}(x)\right\}. \tag{2}$$

In (2) we consider the popular squared Euclidean distance data fidelity, which can be statistically motivated by considering the negative log likelihood of an additive Gaussian noise corruption [5]. The regularizer $\mathcal{R}:\mathbb{R}^n\to\mathbb{R}$ is used to encourage a priori information of the solution $x^\star$ in reconstructions. In recent years a popular non-smooth choice has been total variation (TV) [43] which encourages sharp edges in reconstructions. While naturally smooth regularizers are explicitly used [40], one may be interested in a non-smooth regularizer [5, 8], such as TV, but, be it for computational or theoretical reasons [44], require a smooth approximation instead which can be achieved for example by the Huber norm [34]. The balance between the data fidelity and regularizer in (2) is controlled by the regularization parameter $\alpha\geq 0$. It is crucial to determine a suitable value of α, as a poor choice could lead to a noise dominated or oversmoothed reconstruction [31].

There are a variety of existing techniques to determine an appropriate parameter value, such as the discrepancy principle [31], generalised cross validation [29], or L-curve [31]. In particular, there is no one-method-fits-all and rather each method works under different assumptions to varying degrees of success [4, 6, 9, 24, 28, 31].

An alternative is machine learning, wherein an optimal parameter is found by minimising some appropriate loss function. This can achieved via bilevel learning - a popular data-driven approach to determine hyperparameters [5, 17, 21, 38] which sits in the wider class of bilevel optimisation [16, 45]. In this work, we put emphasis on the following class of bilevel learning problems:

$$\hat{\alpha}\in\arg\min_{\alpha\in[0,\infty]}\left\{\mathcal{J}(\alpha):=\frac{1}{2}\mathbb{E}\left[\|x^\alpha(y)-x^\star\|^2\right]\right\} \tag{3a}$$

$$\text{subject to } x^\alpha(y)=\arg\min_{x\in\mathbb{R}^n}\left\{\Phi_\alpha(x):=\frac{1}{2}\|Ax-y\|^2+\alpha\mathcal{R}(x)\right\}, \tag{3b}$$

where we assume A and $\mathcal{R}$ are such that the solution to (3b) is unique - more details given later. Minimisation problem (3a) is referred to as the upper level problem, and (3b) the lower level problem.

Although the bilevel learning problem (3) is phrased to optimise over a scalar α, the general framework extends to the multi-parameter setting. For example: to find the weights of a sum of different regularizers [20]; the sampling of the forward operator for MRI [44]; the weights in the field of experts model [14]; the parameters of an input-convex neural network acting as the regularizer [2, 39]. In this work we are interested in bilevel learning as a regularization parameter choice rule, so remain in the scalar setting.

We consider minimising the expected Mean Squared Error (MSE) in (3a) which is the most popular choice of loss function in bilevel learning [17], though other choices have been explored [21, 25]. By minimising the expected MSE, the determined parameter is expected to perform well on average. The expectation in (3a) is simply the total expectation and, unless specified otherwise in which case it will be denoted by a subscript, can be taken with respect to, say, some underlying distribution of the ground truth, noisy measurement, or the noise itself. We do not require any properties of these distributions other than the expectations being well defined.

While it is of theoretical importance to study the expected MSE upper level cost function, in practice we instead have a finite number of training data $(x_k^\star, y_k)$, $k = 1, \dots, K$. In this scenario the upper level cost function is the empirical risk

$$\frac{1}{2K}\sum_{k=1}^{K}\|x^\alpha(y_k) - x_k^\star\|^2,$$

and the bilevel learning problem would be solved to determine a single regularization parameter $\hat{\alpha}$ that performs well across the entire training dataset. Then, given some unseen measurement data which is similar to the training data, we can expect $\hat{\alpha}$ to be a reasonable choice of parameter and the variational problem (3b) can be solved. From now on, we suppress the dependence of the reconstruction on the observed measurement, that is, we denote $x^\alpha(y)$ simply as x^α.

A critical theoretical issue of (3) is the well-posedness of the learning and the characterisation of solutions, which can be used to inform the design of numerical methods. In recent years, literature has been developed to address these issues [19, 20, 33, 38]. One example is [33], where optimal parameters are assumed to live in an interval $[\underline{\alpha}, \bar{\alpha}]$ for $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$ chosen a priori. In this setting it is possible to prove stability of the lower level problem and existence of solutions to the upper level problem under certain assumptions. However, in imposing solutions lie in a bounded interval determined a priori, it is possible that for a given training dataset the determined parameter is suboptimal.

Removing this restriction and instead working on a domain more naturally associated with the parameter, namely,

$$[0, \infty] := \{\alpha \in \mathbb{R} : \alpha \geq 0\} \cup \{\infty\}, \tag{4}$$

is therefore natural and the focus of this paper. A consequence of this setting is that qualitative changes in reconstructions may occur for those associated with parameters at the boundary. Additionally, for an optimal parameter to reside at the boundary of $[0, \infty]$ can be an indication that the chosen regularizer is not well suited for the problem setting. Determining natural conditions which guarantee optimal parameters reside in the interior is therefore crucial to exclude these degenerate cases. Various works have contributed towards conditions that yield optimal parameters in the interior [19, 20, 38] and, while primarily considering the case $A = I$, have considered regularizers such as generalized Tikhonov [38], TV-like and their huberised counterparts [20], and more recently a broad class of lower semicontinuous regularizers [19].

Since the optimised parameter in (3) is the regularization parameter of a variational model, positivity of the determined solution is also necessary for regarding bilevel learning as a well-posed parameter choice rule [24]. Parameter choice rules enjoy a rich amount of existing theory [4, 6, 9, 24, 28, 31] so determining conditions for when bilevel learning is a valid parameter choice rule will allow access to a vast amount of well studied results.

In this work we will also consider the sets $(0, \infty]$ and $[0, \infty)$, defined in a similar fashion to (4). Furthermore, we will refer to case $\hat{\alpha} \in (0, \infty]$ as $\hat{\alpha}$ being strictly positive. In particular, we focus on conditions that guarantee $\hat{\alpha}$ is strictly positive, that is, non-zero and possibly infinite. This is in alignment with the setting of existing works [20]. We remark that we do not demand uniqueness of the solution to the upper level problem (3a). Indeed, uniqueness is not guaranteed in general.

1.1 Our contribution

In this work we provide a new sufficient condition to deduce positivity of the solution to the bilevel learning problem (3) that, not only is satisfied whenever the condition commonly used in existing theory [19, 20, 38] is satisfied, but is applicable to inverse problems with a general forward operator, rather than just the denoising setting. We provide an example illustrating that our condition can completely characterise positivity for certain applications. Furthermore, we show that our condition will always be satisfied in a very realistic denoising setting.

The class of regularizers that we consider is very general, with a full statement of the assumptions given in Section 2. While not a focus of this work, we briefly describe how the bilevel learning problem (3) can be solved in practice, as that will in part motivate the class of smooth regularizers that we consider.

1.2 Solving the bilevel learning problem

We first remark that the bilevel learning problem (3) optimises over a scalar parameter α, the regularization parameter. Because of the low dimensional parameter space, one can easily explore how α affects the upper level cost $\mathcal{J}$ and, for example, consider a finite number of values and select the one that achieves the lowest upper level cost or perform an interval search [1]. Indeed, this will be the approach taken in this report for the numerics. In the general multi-parameter case, a brute force grid search is not computationally feasible and other strategies must be developed, of which we discuss a few here.

Most approaches assume that the lower level cost function Φ_α is sufficiently smooth [26, 44] - with operations involving the Hessian of Φ_α being utilised. This limits the choice of lower level cost function and thus regularizer that can be considered. However, should the original cost function be non-smooth, smooth approximations are possible [34] and so the demand for a smooth Φ_α is not unreasonable.

One common approach is to rephrase the bilevel learning problem as a single level problem [44]. A main challenge towards this strategy is that, for a general lower level cost function Φ_α, the minimiser x^α is not differentiable with respect to the parameters that the upper level is optimising over, in our case the regularization parameter α. However, provided that the lower level cost function is sufficiently smooth, the gradient to the upper level cost function $\mathcal{J}$ can be derived using the implicit function theorem [37]. An issue with the approach of [44] is that exact solutions of the lower level problem are required - but are often computed numerically in practice. While results are still promising in spite of this, methods that acknowledge this inexactness have also been developed and studied [23, 22, 46].

1.3 Structure of the paper

The paper is organised as follows. In Section 2 we motivate and state the main results regarding positivity of solutions to the bilevel learning problem. Some useful properties of the lower level problem are covered in Section 3, and we prove the main results in Section 4. Finally, numerous numerical experiments are performed in Section 5 to validate the derived theory.

2 Main result

The choice of regularizer is problem specific, as what constitutes as a suitable reconstruction varies between applications. In general, $\mathcal{R}$ should attain a large evaluation for an x that exhibits undesirable properties. For the denoising application, we are trying to improve upon the noisy measurement y and in particular $\mathcal{R}$ should deem y less desirable than the ground truth $x^\star$. Thus, it is natural to assume that

$$\mathcal{R}(x^\star) < \mathcal{R}(y). \tag{5}$$

Table 1: Examples of Bregman distances for different regularizers, see [10] for details. Here sgn *denotes the sign function.*

Regularizer $\mathcal{R}$	Bregman distance $D_{\mathcal{R}}(x,z)$	Regularization name
$\frac{1}{2}\|x\|^2$	$\frac{1}{2}\|x-z\|^2$	Tikhonov
$\frac{1}{2}\|Kx\|^2$	$\frac{1}{2}\|K(x-z))\|^2$	Generalised Tikhonov
$\|x\|_1$	$\sum_{i=1}^{n}(\text{sgn}(x_i)-\text{sgn}(z_i))x_i$	Lasso

Indeed, condition (5) has been considered in recent works [19, 20] to deduce positivity of solutions to the bilevel learning problem.

While we do not demand it here, typically the regularizer involves a norm and in particular is an even function. Consequently, for a fixed $x^\star$, condition (5) is inherently circular around the origin and as such may underestimate the region of y for which solutions to (3) are strictly positive. While our condition will encompass applications with a very general forward operator, to give an intuition of how it compares to (5), in the denoising setting our condition will read as requiring the linearisation of $\mathcal{R}$ around y evaluated at $x^\star$ to be smaller than $\mathcal{R}(y)$ to conclude positivity of $\hat{\alpha}$. To represent this linearisation, we will find it useful to work with Bregman distances, which are defined as follows.

Definition 1 (Bregman distance)**.** *For a differentiable convex function $\psi : \mathbb{R}^n \to \mathbb{R}$ with gradient $\nabla\psi$, the Bregman distance is defined as*

$$D_\psi(x,\tilde{x}) := \psi(x) - \psi(\tilde{x}) - \langle \nabla\psi(\tilde{x}), x-\tilde{x}\rangle.$$

Bregman distances can be considered a generalisation of the squared Euclidean norm, and have nice properties such as convexity in the first argument and non-negativity [10]. We remark that while some definitions demand ψ be strictly convex [12], we do not require that here as convexity provides all the properties we need for the scope of this paper. Table 1 gives examples of Bregman distances for various functions. From the definition of the Bregman distance, it is clear that, when viewed as a function of x, it represents the distance between $\psi(x)$ and the linearisation of ψ around $\tilde{x}$ evaluated at x. To this end, we introduce the following definition.

Definition 2 (Linearisation around a point)**.** *For a differentiable convex function $\psi : \mathbb{R}^n \to \mathbb{R}$ with gradient $\nabla\psi$, the linearisation of ψ around $\tilde{x}$ evaluated at x is denoted*

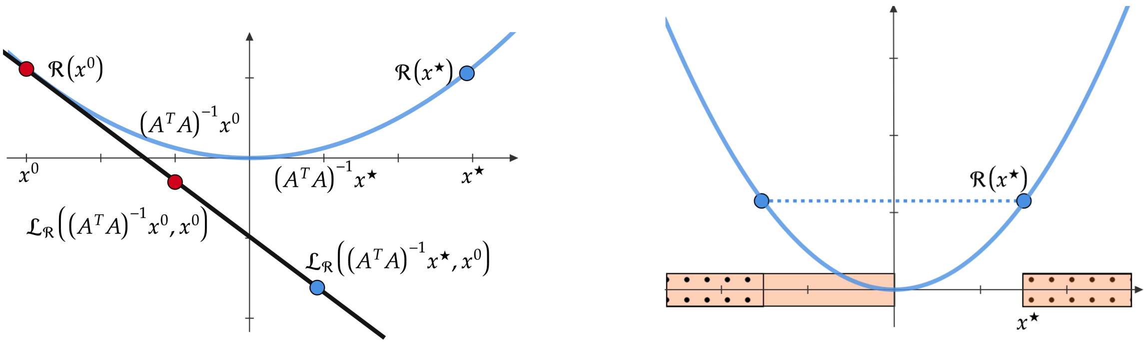
$$\mathcal{L}_\psi(x,\tilde{x}) := \psi(x) - D_\psi(x,\tilde{x}).$$

Rather than requiring (5), we merely require

$$\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0) < \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0) \tag{6}$$

to deduce positivity of $\hat{\alpha}$. If we are in the denoising setting and condition (5) is satisfied, we immediately get that (6) is also satisfied by the non-negativity of the Bregman distance and noticing that $\mathcal{L}_{\mathcal{R}}(x,x) = \mathcal{R}(x)$. Figure 1a illustrates the linearisation for a simple choice of regularizer. Figure 1b compares, for a fixed $x^\star$, how the regions of x^0 for which condition (5) and condition (6) are satisfied differ.

We now provide a motivating example illustrating that condition (6) can completely characterise positivity of optimal parameters.



(a) (b)

Figure 1: The case $n = 1$, $\mathcal{R}(x) = \frac{1}{2}\|x\|^2$, and $A = c \geq 1$. Pane (a): Illustration of the linearisation of the regularizer around x^0. Pane (b): Regions of x^0 for which (5) and (6) are satisfied, indicated by the dotted and shaded regions on the horizontal axis respectively.

Example 1. *Consider the denoising setting with Tikhonov regularization, $\mathcal{R} = \frac{1}{2}\|\cdot\|^2$, and a single data sample. In this setting the lower level solution is given analytically as $x^\alpha = y/(1+\alpha)$. Further, assume that $\|y\| \neq 0$ and $\langle y, x^\star \rangle \neq 0$. Now, solutions will occur either at boundary points of $[0, \infty]$, or in the interior. Evaluation of the upper level at the left boundary point yields*

$$\mathcal{J}(0) = \frac{1}{2}\|y - x^\star\|^2.$$

Notice that optimal solutions $\bar{\alpha}$ in the interior will satisfy $0 = \mathcal{J}'(\bar{\alpha})$, from which one can show that

$$\bar{\alpha} = \frac{\|y\|^2}{\langle y, x^\star \rangle} - 1. \tag{7}$$

It follows from (7) that $\bar{\alpha}$ is strictly positive if and only if

$$\frac{1}{2}\|x^\star\|^2 - \frac{1}{2}\|y - x^\star\|^2 < \frac{1}{2}\|y\|^2,$$

that is, condition (6) is satisfied. Moreover, one can show that the associated upper level cost is

$$\mathcal{J}(\bar{\alpha}) = \frac{1}{2}\|x^\star\|^2 - \frac{1}{2}\frac{\langle y, x^\star \rangle^2}{\|y\|^2}.$$

We claim that whenever condition (6) is satisfied, we also have that $\mathcal{J}(\bar{\alpha}) < \mathcal{J}(0)$. Indeed, using that in this problem setting (6) is equivalent to $\bar{\alpha} > 0$, it follows that

$$\begin{aligned} & 0 < \frac{1}{2}\left(\frac{\langle y, x^\star \rangle}{\|y\|} - \|y\|\right)^2 \\ \iff & \frac{1}{2}\|x^\star\|^2 - \frac{1}{2}\frac{\langle y, x^\star \rangle^2}{\|y\|^2} < \frac{1}{2}\|y - x^\star\|^2 \end{aligned}$$

and so the claim is true. To summarise, for Tikhonov denoising, condition (6) is satisfied if and only if the solution to the bilevel learning problem is strictly positive.

We now state the precise assumptions that we make on the choice of regularizer.

Assumption 1. *We assume the regularizer $\mathcal{R} : \mathbb{R}^n \to [0, \infty)$ is convex and differentiable with continuous gradient $\nabla\mathcal{R}$.*

Remark 1. *We assume the regularizer is non-negative and claim this can be done without loss of generality provided $\mathcal{R}$ is bounded below. Indeed, should $\mathcal{R}$ take negative values then, since $\mathcal{R}$ is assumed bounded below by, say, $C < 0$, one can instead consider a new regularizer $\tilde{\mathcal{R}} := \mathcal{R} - C$, which is non-negative by construction and otherwise inherits the properties of $\mathcal{R}$. In particular, the solution to the associated lower level problem is unchanged.*

We have two main results regarding positivity of $\hat{\alpha}$. The first regards a pointwise condition while the second considers an expectation but requires additional assumptions.

Theorem 1 (Positivity of the pointwise bilevel learning problem solution)**.** *Fix $x^\star \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0) < \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0).$$

Then solutions $\hat{\alpha}$ to the bilevel learning problem (3) *are strictly positive.*

Remark 2. *For the specific case of $A = I$ and $\mathcal{R}(x) = \frac{1}{2}\|Kx\|^2$ where $K \in \mathbb{R}^{p\times n}$, Theorem 1 is proven in Proposition 3.1 of [38] where an equivalent condition is assumed. Said choice of regularizer admits a closed form solution to the lower level problem, which can be inserted into the upper level and further analysed. The result of Theorem 1 covers a wide class of regularizers, and we only demand uniqueness of solutions to the lower level problem, rather than an analytic form.*

Theorem 2 (Positivity of the bilevel learning problem solution)**.** *Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0)\right] < \mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0)\right],$$

$$\mathbb{E}\left[\mathcal{R}((A^TA)^{-1}x^\star)\right] < \infty,$$

and for any $\alpha > 0$

$$\nabla\mathcal{R}((A^TA)^{-1}x^\alpha) = \nabla\mathcal{R}(x^\alpha).$$

Then solutions $\hat{\alpha}$ to the bilevel learning problem (3) *are strictly positive.*

Remark 3. *The assumption that for any $\alpha > 0$, $\nabla\mathcal{R}((A^TA)^{-1}x^\alpha) = \nabla\mathcal{R}(x^\alpha)$ is most restrictive to the A that are applicable to Theorem 2. While this condition is not required in Theorem 1, we note that it is clearly satisfied should $A = I$.*

We prove both results in Section 4. Before this, we first cover in Section 3 some fundamental results and properties of the lower level problem.

3 Preliminaries

We start by stating relevant definitions and properties of the lower level cost function. In particular, we require properties that are sufficient for existence and uniqueness of solutions to the lower level problem, as well as continuity of reconstructions with respect to the regularization parameter. The following definitions are taken from [7] and [13].

Definition 3 (Minimiser)**.** *We say that $\hat{x}$ is a global minimiser of $\psi : \mathbb{R}^n \to \mathbb{R}$ if $\psi(\hat{x}) \leq \psi(x)$ for all $x \in \mathbb{R}^n$.*

Definition 4 (Bounded below)**.** *A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is bounded below if there exists $C \geq 0$ such that*

$$\psi(x) \geq -C \quad \text{for all } x \in \mathbb{R}^n.$$

Definition 5 (Coercive)**.** *A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is coercive if*

$$\psi(x) \to \infty \qquad \text{as } \|x\| \to \infty.$$

Definition 6 (Convex function)**.** *A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for all $x, z \in \mathbb{R}^n$,*

$$D_\psi(x, z) \geq 0. \tag{8}$$

Moreover, ψ is said to be strictly convex if for $x \neq z$, inequality (8) *is strict.*

Remark 4. *The classical definition of (strict) convexity [41] is different to the one we use, however it can be easily recovered by the definition of the Bregman distance (e.g. see Propoisition 3.10 in [41]).*

We can now state a classical result regarding existence and uniqueness of minimisers to the lower level problem.

Lemma 1. *Let $\psi : \mathbb{R}^n \to \mathbb{R}$ be bounded below, coercive, continuous, and strictly convex. Then ψ has a unique global minimiser.*

Proof. Existence follows by the direct method in the calculus of variations [18]. Uniqueness follows by strict convexity. □

In particular, by the assumptions on A and $\mathcal{R}$, the lower level problem (3b) admits a unique minimizer, justifying our choice of notation.

Proposition 1 (Properties of the lower level cost function)**.** *Fix $\alpha \geq 0$. Let A be injective and $\mathcal{R}$ satisfy Assumption 1. Then the lower level cost function*

$$\Phi_\alpha(x) = \frac{1}{2}\|Ax - y\|^2 + \alpha\mathcal{R}(x)$$

is bounded below, coercive, continuous, and strictly convex. Moreover, it admits a unique global minimizer.

Proof. The properties of Φ_α follow from standard results in convex analysis. Existence and uniqueness of a minimizer then follows from Lemma 1.

□

We require continuity of the reconstruction map $\alpha \mapsto x^\alpha$ at $\alpha = 0$. This is proven in [20], which we include here for completeness.

Lemma 2 (Convergence of reconstructions at the boundary (Lemma 8 in [20]))**.** *Suppose $\{\alpha_k\} \subset (0, \infty)$ satisfies $\lim_{k\to\infty} \alpha_k = 0$. Then $\lim_{k\to\infty} x^{\alpha_k} = x^0$.*

Proof. Pick arbitrary $\delta > 0$. By the choice of datafit $\mathcal{F}(x) = \frac{1}{2}\|Ax - y\|^2$, and injectivity of A, x^0 is the unique minimiser of $\mathcal{F}$. By the minimality of $x^{\alpha_k} = \arg\min_x \Phi_{\alpha^k}(x)$ and non-negativity of $\mathcal{R}$, we have

$$\begin{aligned}\mathcal{F}(x^{\alpha_k}) \leq \Phi_{\alpha_k}(x^{\alpha_k}) &= \mathcal{F}(x^{\alpha_k}) + \alpha_k\mathcal{R}(x^{\alpha_k}) \\ &\leq \mathcal{F}(x^0) + \alpha_k\mathcal{R}(x^0).\end{aligned}$$

Since $\mathcal{R}(x^0)$ is fixed and $\alpha_k \to 0$, we may choose k large enough such that

$$\mathcal{F}(x^{\alpha_k}) \le \mathcal{F}(x^0) + \delta.$$

Letting $\delta \to 0_+$, we see that $\lim_{k\to\infty} \mathcal{F}(x^{\alpha_k}) \le \mathcal{F}(x^0)$. By the minimality of x^0, necessarily $\lim_{k\to\infty} \mathcal{F}(x^{\alpha_k}) = \mathcal{F}(x^0)$. By continuity of the datafit and uniqueness of its minimiser, it follows that $\lim_{k\to\infty} x^{\alpha_k} = x^0$ and we are done. □

4 Proof of main result

We aim to show that, under very natural conditions, $\alpha = 0$ is not a minimum of the upper level cost function $\mathcal{J}$. There are two main problems towards this: firstly, the reconstruction map $\alpha \mapsto x^\alpha$ is in general non-differentiable [35] and consequently, without stronger assumptions on the choice of regularizer $\mathcal{R}$ [44], the upper level cost function $\mathcal{J}$ is non-differentiable; secondly, the value we are interested in is on the domain boundary of $\mathcal{J}$. To this end, we work with a generalisation of the derivative known as Dini derivatives [3, 27, 36]. More precisely, we consider the upper right Dini derivative.

Definition 7 (Upper right Dini derivative). *Let $\mathcal{J}$ be any real valued function defined on $[0,\infty]$ and let $\tilde{\alpha} \in [0,\infty)$. We define the upper right Dini derivative of $\mathcal{J}$ evaluated at $\tilde{\alpha}$ as*

$$\mathcal{J}'_+(\tilde{\alpha}) := \limsup_{\alpha\to\tilde{\alpha}_+} \frac{\mathcal{J}(\alpha) - \mathcal{J}(\tilde{\alpha})}{\alpha - \tilde{\alpha}},$$

where $\alpha \to \tilde{\alpha}_+$ denotes the right-hand limit.

We remark that the we allow infinite limits in the above definition and so the upper right Dini derivative is always well defined. Dini derivatives follow some standard calculus rules and generalisations of the mean value theorem and Rolle's theorem can be stated [3, 27, 36].

Definition 8 (Local minimum). *Let $\mathcal{J}$ be any real valued function on $[0,\infty]$. We say that $\alpha^\star \in [0,\infty)$ is a local minimum of $\mathcal{J}$ if there exists $\delta \in (0,\infty]$ such that*

$$\mathcal{J}(\alpha^\star) \le \mathcal{J}(\alpha) \quad \text{for all } \alpha \in B_\delta(\alpha^\star) \cap [0,\infty],$$

where $B_\delta(\alpha^\star) := \{\alpha \in \mathbb{R} : \|\alpha - \alpha^\star\| \le \delta\}$.

We now use Dini derivatives to determine behaviour of $\mathcal{J}$ at the domain boundary.

Lemma 3. *Let $\mathcal{J} : [0,\infty] \to [0,\infty)$ be any function. If the upper right Dini derivative at 0 satisfies $\mathcal{J}'_+(0) < 0$, then 0 is not a local minimum of $\mathcal{J}$.*

Proof. Assume $\mathcal{J}'_+(0) < 0$. We then have existence of $\delta > 0$ such that

$$\frac{\mathcal{J}(\alpha) - \mathcal{J}(0)}{\alpha} < 0$$

for all $\alpha \in (0,\delta)$. It immediately follows that $\mathcal{J}(\alpha) < \mathcal{J}(0)$ in $(0,\delta)$ and so $\mathcal{J}$ is locally strictly decreasing at the domain boundary. In particular, 0 is not a local minimum of $\mathcal{J}$. □

Remark 5. *The condition in Lemma 3 is sufficient but not necessary for the solution of the bilevel learning problem* (3) *to be strictly positive. Indeed, we will see in Section 5 that the upper level cost function $\mathcal{J}(\alpha)$ is non-convex and in particular* 0 *may actually be a local minimum, and yet the global minimum of* (3a) *is achieved at a strictly positive parameter value (possibly infinity).*

The proof of the main results involves find an upper bound of $\mathcal{J}(\alpha) - \mathcal{J}(0)$ of the form $\alpha h(\alpha)$ and showing that $\limsup_{\alpha\to 0_+} h(\alpha) < 0$. We now justify that such a manipulation will also show that $\mathcal{J}'_+(0) < 0$.

Proposition 2. *Let $f, g : [0,\infty] \to \mathbb{R}$ be functions such that $f(\alpha) \leq g(\alpha)$ for all $\alpha \in [0,\infty]$. Then*

$$\limsup_{\alpha\to 0_+} \frac{f(\alpha)}{\alpha} \leq \limsup_{\alpha\to 0_+} \frac{g(\alpha)}{\alpha}.$$

Proof. Let $\{\alpha_n\} \subset (0,\infty)$ be a sequence such that $\lim_{n\to\infty} \alpha_n = 0$. We need to show that

$$\limsup_{n\to\infty} \frac{f(\alpha_n)}{\alpha_n} \leq \limsup_{n\to\infty} \frac{g(\alpha_n)}{\alpha_n}$$

or more precisely $\lim_{n\to\infty} u_n \leq \lim_{n\to\infty} v_n$ where

$$u_n := \sup\left\{\frac{f(\alpha_n)}{\alpha_n}, \frac{f(\alpha_{n+1})}{\alpha_{n+1}}, \cdots\right\}, \qquad v_n := \sup\left\{\frac{g(\alpha_n)}{\alpha_n}, \frac{g(\alpha_{n+1})}{\alpha_{n+1}}, \cdots\right\}.$$

Indeed, since each $\alpha_k > 0$ we have that $f(\alpha_k)/\alpha_k \leq g(\alpha_k)/\alpha_k$ and so for all n, $u_n \leq v_n$. It follows that $\lim_{n\to\infty} u_n \leq \lim_{n\to\infty} v_n$ and we are done. □

We aim to show that the upper right Dini derivative of $\mathcal{J}$ at 0 can be expressed in terms of the linearisation of the regularizer, $\mathcal{L}_{\mathcal{R}}$. We now state some useful results that will go towards showing this.

Proposition 3. *Let A be injective and $\mathcal{R}$ satisfy Assumption 1. Then*

(i) $\alpha\nabla\mathcal{R}(x^\alpha) = A^T A x^0 - A^T A x^\alpha$

(ii) For any $x, \tilde{x} \in \mathbb{R}^n$, $\mathcal{L}_{\mathcal{R}}(x,\tilde{x}) = \mathcal{R}(\tilde{x}) + \langle\nabla\mathcal{R}(\tilde{x}), x - \tilde{x}\rangle$

(iii) For any $x \in \mathbb{R}^n$, $\mathcal{L}_{\mathcal{R}}(x,x) = \mathcal{R}(x)$

(iv) If $u, v : [0,\infty] \to \mathbb{R}^n$ are such that $\lim_{\alpha\to 0_+} u(\alpha) = u(0)$ and $\lim_{\alpha\to 0_+} v(\alpha) = v(0)$, then

$$\lim_{\alpha\to 0_+} \mathcal{L}_{\mathcal{R}}(u(\alpha), v(\alpha)) = \mathcal{L}_{\mathcal{R}}(u(0), v(0)).$$

Proof. For (i), by the choice of the data fidelity and differentiability of $\mathcal{R}$, we have that

$$0 = \nabla\Phi_\alpha(x^\alpha) = A^T A x^\alpha - A^T y + \alpha\nabla\mathcal{R}(x^\alpha). \tag{9}$$

Since A is injective, x^0 is the unique least squares solution and in particular satisfies the normal equations

$$A^T A x^0 - A^T y = 0. \tag{10}$$

Combining (9) and (10) yields (i). (ii) is immediate by the definition of both $\mathcal{L}_{\mathcal{R}}$ and the Bregman distance $D_{\mathcal{R}}$. (iii) follows from (ii). By the continuity of both $\mathcal{R}$ and $\nabla\mathcal{R}$ and assumption on u and v, (iv) also follows from (ii). □

While the result of Theorem 2 involves the expectation, we will find it useful to work with the quantity that we are taking the expectation of, and to this end define

$$\tilde{\mathcal{J}}(\alpha) := \frac{1}{2}\|x^\alpha - x^\star\|^2.$$

We now state the form of $\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0)$ that will be utilised in the main proofs.

Proposition 4. *Let A be injective $\mathcal{R}$ satisfy Assumption 1. Then*

$$\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0) = \alpha\left(\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)\right) - \frac{1}{2}\left\|x^0 - x^\alpha\right\|^2. \tag{11}$$

Proof. By the definition of $\tilde{\mathcal{J}}$ we have

$$\tilde{\mathcal{J}}(\alpha) = \frac{1}{2}\|x^\alpha - x^\star\|^2 = \frac{1}{2}\|x^\alpha\|^2 - \langle x^\alpha, x^\star\rangle + \frac{1}{2}\|x^\star\|^2$$

and in particular

$$\begin{aligned}
\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0) &= \frac{1}{2}\|x^\alpha - x^\star\|^2 - \frac{1}{2}\left\|x^0 - x^\star\right\|^2 \\
&= \frac{1}{2}\|x^\alpha\|^2 - \frac{1}{2}\left\|x^0\right\|^2 + \langle x^0 - x^\alpha, x^\star\rangle \\
&= \frac{1}{2}\|x^\alpha\|^2 - \frac{1}{2}\left\|x^0\right\|^2 + \langle x^0 - x^\alpha, x^\star - x^\alpha\rangle + \langle x^0, x^\alpha\rangle - \langle x^\alpha, x^\alpha\rangle \\
&= \langle x^0 - x^\alpha, x^\star - x^\alpha\rangle - \frac{1}{2}\left\|x^0\right\|^2 + \langle x^0, x^\alpha\rangle - \frac{1}{2}\|x^\alpha\|^2 \\
&= \langle x^0 - x^\alpha, x^\star - x^\alpha\rangle - \frac{1}{2}\left\|x^0 - x^\alpha\right\|^2.
\end{aligned}$$

It remains to show that

$$\langle x^0 - x^\alpha, x^\star - x^\alpha\rangle = \alpha\left(\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)\right). \tag{12}$$

In order to phrase the inner product in (12) in terms of $\mathcal{L}_{\mathcal{R}}$, we must first introduce Bregman distances and consequently gradients of $\mathcal{R}$. Recall that by Proposition 3 (i), $\alpha\nabla\mathcal{R}(x^\alpha)$ involves A^TA, which we can freely introduce since it is invertible by the injectivity of A. Indeed,

$$\langle x^0 - x^\alpha, x^\star - x^\alpha\rangle = \langle (A^TA)^{-1}(A^TA)(x^0 - x^\alpha), x^\star - x^\alpha\rangle$$

by the symmetry of A^TA

$$\begin{aligned}
&= \langle A^TA(x^0 - x^\alpha), (A^TA)^{-1}(x^\star - x^\alpha)\rangle \\
&= \langle A^TA(x^0 - x^\alpha), (A^TA)^{-1}x^\star - x^\alpha - (A^TA)^{-1}x^\alpha + x^\alpha\rangle \\
&= \langle A^TA(x^0 - x^\alpha), (A^TA)^{-1}x^\star - x^\alpha\rangle - \langle A^TA(x^0 - x^\alpha), (A^TA)^{-1}x^\alpha - x^\alpha\rangle
\end{aligned}$$

by Proposition 3 (i)

$$= \alpha\left(\langle\nabla\mathcal{R}(x^\alpha), (A^TA)^{-1}x^\star - x^\alpha\rangle - \langle\nabla\mathcal{R}(x^\alpha), (A^TA)^{-1}x^\alpha - x^\alpha\rangle\right)$$

by Proposition 3 (ii)

$$= \alpha \left(\mathcal{L}_{\mathcal{R}}((A^T A)^{-1} x^\star, x^\alpha) - \mathcal{R}(x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^T A)^{-1} x^\alpha, x^\alpha) + \mathcal{R}(x^\alpha) \right)$$
$$= \alpha \left(\mathcal{L}_{\mathcal{R}}((A^T A)^{-1} x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^T A)^{-1} x^\alpha, x^\alpha) \right)$$

and we are done. □

In calculating the upper right Dini derivative of $\mathcal{J}$ at 0, we see from Proposition 4 that the quantity

$$\liminf_{\alpha \to 0_+} \frac{1}{\alpha} \left\| x^0 - x^\alpha \right\|^2 \tag{13}$$

will be encountered. We now show that (13) vanishes.

Proposition 5. *Let A be injective and $\mathcal{R}$ satisfy Assumption 1. Then*

$$\lim_{\alpha \to 0_+} \frac{1}{\alpha} \left\| x^0 - x^\alpha \right\|^2 = 0.$$

Proof. By the non-negativity of $\left\| x^0 - x^\alpha \right\|^2 / \alpha$, we immediately have

$$0 \le \liminf_{\alpha \to 0_+} \frac{1}{\alpha} \left\| x^0 - x^\alpha \right\|^2 \le \limsup_{\alpha \to 0_+} \frac{1}{\alpha} \left\| x^0 - x^\alpha \right\|^2$$

and so the result will follow if we can show that

$$\limsup_{\alpha \to 0_+} \frac{1}{\alpha} \left\| x^0 - x^\alpha \right\|^2 \le 0. \tag{14}$$

Since A is injective, we immediately have that

$$\left\| x^0 - x^\alpha \right\|^2 \le \frac{1}{\sigma_{\min}^2} \left\| A(x^0 - x^\alpha) \right\|^2,$$

where $\sigma_{\min} > 0$ is the smallest singular value of A. Thus, with Proposition 2 in mind, it suffices to show that

$$\limsup_{\alpha \to 0_+} \frac{1}{\alpha} \left\| A(x^0 - x^\alpha) \right\|^2 \le 0. \tag{15}$$

Indeed, by Proposition 3 (i) and convexity of $\mathcal{R}$,

$$\begin{aligned} \alpha(\mathcal{R}(x) - \mathcal{R}(x^\alpha)) &\ge \langle A^T A x^0 - A^T A x^\alpha, x - x^\alpha \rangle \\ &= \langle A^T A x^0 - A^T A x + A^T A x - A^T A x^\alpha, x - x^\alpha \rangle \\ &= \langle A^T A x^0 - A^T A x, x - x^\alpha \rangle + \| A(x - x^\alpha) \|^2. \end{aligned}$$

It follows that

$$\alpha(\mathcal{R}(x^0) - \mathcal{R}(x^\alpha)) \ge \left\| A(x^0 - x^\alpha) \right\|^2.$$

By Proposition 2

$$\limsup_{\alpha \to 0_+} \frac{1}{\alpha} \left\| A(x^0 - x^\alpha) \right\|^2 \le \limsup_{\alpha \to 0_+} \left(\mathcal{R}(x^0) - \mathcal{R}(x^\alpha) \right) = 0,$$

where the equality follows by the convexity (and hence continuity) of $\mathcal{R}$ and Lemma 2. In particular we have shown (15) and by Proposition 2 we have (14) and we are done. □

Remark 6. *In the proof of Proposition 5 we implicitly assume that $\mathcal{R}(x^0) < \infty$ to deduce the final equality. Indeed, the indeterminate form $\infty - \infty$ would otherwise be encountered. Regarding extending the result of this work to a more general class of regularizers, such as indicator functions, either this property must be assumed or an alternative strategy found.*

We now prove that under the pointwise condition, the upper right Dini derivative of $\tilde{\mathcal{J}}$ at 0 is strictly negative; this is a crucial result for the proof of both Theorem 1 and Theorem 2.

Lemma 4. *Fix $x^\star \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0) < \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0).$$

Then $\tilde{\mathcal{J}}'_+(0) < 0$.

Proof. By Proposition 4,

$$\begin{aligned}\tilde{\mathcal{J}}'_+(0) &= \limsup_{\alpha\to 0_+} \frac{\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0)}{\alpha} \\ &= \limsup_{\alpha\to 0_+} \left(\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha) - \frac{1}{2\alpha}\left\| x^0 - x^\alpha \right\|^2 \right). \end{aligned} \tag{16}$$

The last term in (16) will vanish in the limit by Proposition 5. With the choice of $u(\alpha) = (A^TA)^{-1}x^\star$ and $v(\alpha) = (A^TA)^{-1}x^\alpha$, it follows from Lemma 2 and Proposition 3 (iv) that

$$\tilde{\mathcal{J}}'_+(0) = \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0),$$

which is strictly negative by assumption. □

We are now ready prove the first main result.

Theorem 1 (Positivity of the pointwise bilevel learning problem solution)**.** *Fix $x^\star \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0) < \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0).$$

Then solutions $\hat{\alpha}$ to the bilevel learning problem (3) *are strictly positive.*

Proof. Since $x^\star$ and y are fixed, we have that $\mathcal{J} = \tilde{\mathcal{J}}$. By Lemma 4, $\mathcal{J}'_+(0) < 0$. It follows from Lemma 3 that 0 is not a local minimiser of $\mathcal{J}$ and in particular cannot be a global minimiser. Thus the global minimum is achieved at some $\hat{\alpha} \in (0, \infty]$, that is, $\hat{\alpha}$ is strictly positive. □

We now prove an analogous result of Lemma 4 for when we are taking expectations.

Lemma 5. *Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0)\right] < \mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0)\right], \tag{17}$$

$$\mathbb{E}\left[\mathcal{R}((A^TA)^{-1}x^\star)\right] < \infty, \tag{18}$$

and for any $\alpha > 0$

$$\nabla\mathcal{R}((A^TA)^{-1}x^\alpha) = \nabla\mathcal{R}(x^\alpha). \tag{19}$$

Then $\mathcal{J}'_+(0) < 0$.

Proof. The main challenge is to justify swapping the expectation and lim sup in

$$\mathcal{J}'_+(0) = \limsup_{\alpha \to 0_+} \mathbb{E}\left[\frac{\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0)}{\alpha}\right].$$

This can be justified (up to inequality) by the Reverse Fatou Lemma (see Corollary 5.3.2 in [42]). We now show that the conditions of the Reverse Fatou Lemma are satisfied, which in this setting requires showing that $(\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0))/\alpha \leq Z$ for some random variable Z independent of α and with finite expectation.

By Proposition 4 we have

$$\begin{aligned}\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0) &= \alpha\left(\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)\right) - \frac{1}{2}\left\|x^0 - x^\alpha\right\|^2 \\ &\leq \alpha\left(\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)\right)\end{aligned}$$

by the definition of $\mathcal{L}_{\mathcal{R}}$

$$= \alpha\left(\mathcal{R}((A^TA)^{-1}x^\star) - D_{\mathcal{R}}((A^TA)^{-1}x^\star, x^\alpha) - \mathcal{R}((A^TA)^{-1}x^\alpha) + D_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)\right).$$

Since $\mathcal{R}$, and $D_{\mathcal{R}}$ are non-negative it follows that

$$\frac{\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0)}{\alpha} \leq \mathcal{R}((A^TA)^{-1}x^\star) + D_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha).$$

By the definition of the Bregman distance (see also Proposition 2.4 in [10]), assumption (19) is equivalent to $D_{\mathcal{R}}((A^TA)^{-1}x^\alpha, x^\alpha)$ vanishing. Thus we have found an upper bound of $(\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0))/\alpha$ independent of α which by assumption (18) has finite expectation. It follows from the Reverse Fatou Lemma that

$$\mathcal{J}'_+(0) \leq \mathbb{E}\left[\limsup_{\alpha \to 0_+} \frac{\tilde{\mathcal{J}}(\alpha) - \tilde{\mathcal{J}}(0)}{\alpha}\right] = \mathbb{E}\left[\tilde{\mathcal{J}}'_+(0)\right]. \tag{20}$$

By assumption (17) and Lemma 4, (20) is strictly negative and we are done. □

Theorem 2 (Positivity of the bilevel learning problem solution)**.** *Let A be injective and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^\star, x^0)\right] < \mathbb{E}\left[\mathcal{L}_{\mathcal{R}}((A^TA)^{-1}x^0, x^0)\right], \tag{21}$$

$$\mathbb{E}\left[\mathcal{R}((A^TA)^{-1}x^\star)\right] < \infty$$

and

$$\nabla\mathcal{R}((A^TA)^{-1}x^\alpha) = \nabla\mathcal{R}(x^\alpha)$$

for all $\alpha > 0$. Then solutions $\hat{\alpha}$ to the bilevel learning problem (3) *are strictly positive.*

Proof. By the assumptions we have from Lemma 5 that $\mathcal{J}'_+(0) < 0$. It follows from Lemma 3 that 0 is not a local minimiser of $\mathcal{J}$ and in particular cannot be a global minimiser. Thus the global minimum is achieved at some $\hat{\alpha} \in (0, \infty]$, that is, $\hat{\alpha}$ is strictly positive. □

We now show that in the denoising setting where the measurement has been corrupted by additive noise of mean zero, if the regularizer is strictly convex then we are guaranteed $\hat{\alpha} > 0$.

Corollary 1. *If $A = I$, $x^\star$ is fixed and $y = x^\star + \epsilon$ where $\mathbb{E}_\epsilon[\epsilon] = 0$, then condition* (21) *reads*

$$0 < \mathbb{E}_\epsilon \left[D_R(y, x^\star) + D_{\mathcal{R}}(x^\star, y) \right], \tag{22}$$

which we recognise to be the symmetric Bregman distance (e.g. see [11]). Moreover, if $\mathcal{R}$ is strictly convex, then $\hat{\alpha} > 0$.

Proof. By definition of the Bregman distance, condition (21) is also given by

$$\begin{aligned} 0 &< \mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y), y - x^\star \rangle \right] \\ &= \mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(x^\star) + \nabla\mathcal{R}(x^\star), y - x^\star \rangle \right] \\ &= \mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(x^\star), y - x^\star \rangle + \langle \nabla\mathcal{R}(x^\star), y - x^\star \rangle \right]. \end{aligned} \tag{23}$$

In the denoising setting we have that $y - x^\star = \epsilon$. Since $x^\star$ is fixed and $\mathbb{E}_\epsilon[\epsilon] = 0$, it follows that (23) is given by

$$\mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(x^\star), y - x^\star \rangle \right] + \langle \nabla\mathcal{R}(x^\star), \mathbb{E}_\epsilon[\epsilon] \rangle = \mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(x^\star), y - x^\star \rangle \right].$$

By definition of the Bregman distance, we have

$$\mathbb{E}_\epsilon \left[\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(x^\star), y - x^\star \rangle \right] = \mathbb{E} \left[D_R(y, x^\star) + D_{\mathcal{R}}(x^\star, y) \right]$$

and we have shown (22). Since ϵ is a continuous random variable we have that $x^\star \neq y$ almost surely. From the definition of strict convexity it immediately follows that $D_{\mathcal{R}}(y, x^\star) > 0$ and so condition (22) is always satisfied. Since $A = I$ and $x^\star$ is fixed, the other conditions of Theorem 2 are always satisfied and so $\hat{\alpha} > 0$ by Theorem 2. □

4.1 Extension to a forward operator in the upper level

The bilevel learning problem (3) and thus results of Theorem 1 and Theorem 2 are specific to the MSE upper level cost. We now prove an analogous result for an alternative upper level cost function, namely, the predictive risk,

$$\mathcal{J}(\alpha) = \mathbb{E} \left[\frac{1}{2} \| Ax^\star - Ax^\alpha \|^2 \right],$$

where we will require that A is invertible.

For the predictive risk upper level cost, the associated bilevel learning problem is

$$\hat{\alpha} \in \arg\min_{\alpha \in [0,\infty]} \mathbb{E} \left[\frac{1}{2} \| Ax^\alpha - Ax^\star \|^2 \right], \tag{24a}$$

$$\text{subject to } x^\alpha = \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \| Ax - y \|^2 + \alpha \mathcal{R}(x) \right\}. \tag{24b}$$

Using Theorem 2 and the invertability of A, positivity of the solution $\hat{\alpha}$ of the bilevel learning problem (24) can be established.

Theorem 3. *Let A be invertible and $\mathcal{R}$ satisfy Assumption 1 and be such that*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{R}}(x^\star, x^0)\right] < \mathbb{E}\left[\mathcal{R}(x^0)\right], \tag{25}$$

and

$$\mathbb{E}\left[\mathcal{R}(x^\star)\right] < \infty. \tag{26}$$

Then solutions $\hat{\alpha}$ to the bilevel learning problem (24) *are strictly positive.*

Proof. Using the invertability of A, we intend to rephrase the bilevel learning problem (24) as a denoising problem and apply Theorem 2.

We first remark that $x^0 = A^{-1}y$. Since A is assumed invertible, the solution x^α to (24b) satisfies $Ax^\alpha = z^\alpha$, where

$$z^\alpha = \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\|z - y\|^2 + \alpha \mathcal{R}(A^{-1}z) \right\}. \tag{27}$$

By the invertability of A and assumption on $\mathcal{R}$ it follows that $\tilde{\mathcal{R}} := \mathcal{R} \circ A^{-1}$ also satisfies Assumption 1. Furthermore, notice that

$$\mathcal{L}_{\mathcal{R}}(x^\star, x^0) = \mathcal{L}_{\tilde{\mathcal{R}}}(Ax^\star, Ax^0) =: \mathcal{L}_{\tilde{\mathcal{R}}}(z^\star, y)$$

where we have defined $z^\star := Ax^\star$.

Thus, assumption (25) reads

$$\mathbb{E}\left[\mathcal{L}_{\tilde{\mathcal{R}}}(z^\star, y)\right] < \mathbb{E}\left[\tilde{\mathcal{R}}(y)\right]. \tag{28}$$

and also assumption (26) reads

$$\mathbb{E}\left[\tilde{\mathcal{R}}(z^\star)\right] < \infty. \tag{29}$$

Using the new notation, the upper level problem (24a) reads

$$\hat{\alpha} = \arg\min_{\alpha \in [0,\infty]} \mathbb{E}\left[\frac{1}{2}\|z^\alpha - z^\star\|^2\right]. \tag{30}$$

In particular, we have phrased the bilevel learning problem (24) as a denoising bilevel problem (30) and (27) with regularizer $\tilde{\mathcal{R}}$. By the properties of $\tilde{\mathcal{R}}$ and inequalities (28) and (29), it follows from Theorem 2 that $\hat{\alpha}$ is strictly positive. □

5 Numerical Experiments

We now explore the presented theory with some numerical examples. Although in practice the problem is high-dimensional, for a geometric and visual interpretation of the theory, we consider in Section 5.1 the small dimensional case of $n = 2$. Relevant large scale problems are provided in Section 5.2.

5.1 Low dimensional problems

We explore how well the results of Theorem 1 and Theorem 2 characterise positivity. In the following, we consider the area $\Omega := [-1.6, 1.6] \times [-1.6, 1.6] \subset \mathbb{R}^2$, discretised into a 100×100 grid. Since our results involve x^0 and $x^\star$, we interpret Ω as the reconstruction space, rather than the measurement space.

Table 2: Key for the boundary colours used in numerical examples. The shorthand labels used in the following legends are also indicated.

Associated result	Condition satisfied outside the boundary	Colour	Label
[19, 20]	$\mathcal{R}(x^\star) < \mathcal{R}(y)$	Red	Old
Theorem 1	$\mathcal{L}_\mathcal{R}((A^TA)^{-1}x^\star, x^0) < \mathcal{L}_\mathcal{R}((A^TA)^{-1}x^0, x^0)$	Blue	New
–	Numerical solution to (3) satisfying $\hat{\alpha} > 0$	Black	Numerical

We fix the ground truth $x^\star = [1, 0.5]^T$, which will be indicated by a yellow star in the upcoming plots. Considering each point in the grid Ω as a candidate x^0, we compute the boundary for when the condition of Theorem 1 becomes satisfied. If we are in the case $A = I$, we may also compute the boundary for when (5) becomes satisfied. Since the lower level problem requires a measurement y, in order to have a well defined mapping between the x^0 and y, we restrict ourselves to invertible A in this section. We can then numerically compute the solution to the bilevel learning problem with data $(x^\star, Ax^0)$ by considering parameters

$$[\alpha_1 = 0,\ \alpha_2 = 10^{-12}, \cdots,\ \alpha_{99} = 10^3,\ \alpha_{100} = 10^7],$$

where $[\log_{10}(\alpha_2),\ \cdots, \log_{10}(\alpha_{99})]$ is a linear discretisation of 98 points between -12 and 3, and select the parameter that achieves the smallest upper level cost. With this, we compute the boundary for which the numerical bilevel solution becomes strictly positive. A summary of the boundaries and their represented colours and legend names is provided in Table 2.

We consider two forward operators corresponding to denoising and deconvolution. Namely,

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0.7274 & 0.2726 \\ 0.2726 & 0.7274 \end{bmatrix},$$

where A_2 represents a Gaussian blur with standard deviation 0.8.

For each forward operator, we consider four different regularizers and see how the boundaries, detailed in the above and summarised in Table 2, change. In particular, we consider both generalised Tikhonov and generalised Huber norm, given by

$$\mathcal{R}(x) = \frac{1}{2}\|Kx\|^2 \quad \text{and} \quad \mathcal{R}(x) = \sum_{i=1}^{p} \mathrm{hub}_\gamma([Kx]_i)$$

respectively, where $K \in \mathbb{R}^{p\times n}$, not necessarily full rank, and

$$\mathrm{hub}_\gamma(s) = \begin{cases} |s| - \frac{\gamma}{2} & \text{if } |s| \geq \gamma \\ \frac{1}{2\gamma}(s)^2 & \text{if } |s| < \gamma. \end{cases}$$

Regarding the choice of K, we will consider both $K = I \in \mathbb{R}^{2\times 2}$, which will yield standard Tikhonov and Huber norm respectively, and also $K = [1\ -1] \in \mathbb{R}^{1\times 2}$ which can be interpreted as the discretisation of the first order finite difference operator for $n = 2$ [32]. For this latter choice of K, we refer to the regularizer as ℓ_2^2-grad and an $n = 2$ analogue of Huber TV respectively.

We and interested in how well Theorem 1 characterises positivity of the solution to the bilevel learning problem (3). Using the approach outlined above, we can determine the area of the region where the numerical solution to (3) is 0, and also the area where the condition of Theorem 1 is violated. Should Theorem 1 perfectly characterise positivity, we would expect both these areas to be the same. We compute the ratio between these areas for the different A and $\mathcal{R}$ mentioned above, and display the results in Table 3. In the denoising setting, we also compute the area where condition (5) is violated, to see how the new condition compares. In Figure 2a we see that, for Tikhonov denoising, Theorem 1 perfectly characterises positivity, as we would expect following Example 1. From Table 3, Theorem 1 characterises the positivity of (3) well for the considered problems, with many area ratios being around 1. Furthermore, we see in Figure 2 and Figure 3 that some instances where the ratio is close to but not exactly 1 is down to numerical error. For the denoising setting, we see in Figure 2 that condition (5) overestimates the region where $\hat{\alpha} = 0$ by a factor of 2 to 4. Compared to condition (5), Theorem 1 yields a better characterisation of positivity, particularly for points far away from $x^\star$ - as demonstrated in Figure 2a and Figure 2c.

Table 3: Ratio between the area where 0 *is the optimal parameter and area in the reconstruction space where the (old or new) theory condition is violated. Values close to* 1 *mean the condition is close to fully characterising positivity of* (3)*. Since condition* (5) *is only valid for $A = I$, we cannot compare for the $A \neq I$ case. As we only consider points in Ω, if the area where a condition is violated extends beyond Ω, we indicate the case with an asterisk beside the provided number. All numbers are given to 3 decimal points.*

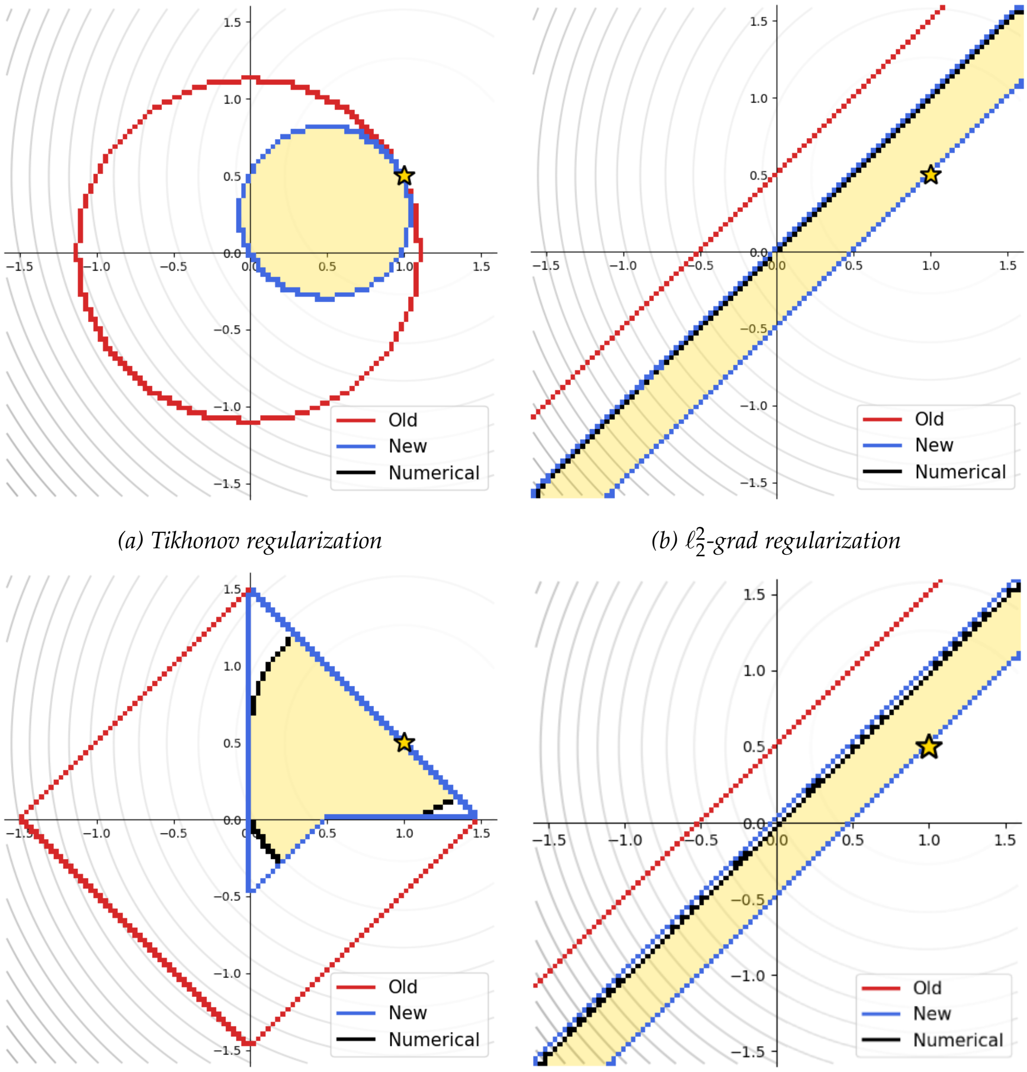
Problem	Condition violated	Regularizer			
		Tikhonov	ℓ_2^2-grad	Huber	Huber TV
Denoising	New	1	1.069*	1.171	1.129*
Denoising	Old	3.979	2.071*	4.214	2.182*
Deconvolution	New	1.028	1.020*	1.143	1.015*
Deconvolution	Old	—	—	—	—

We now demonstrate the result of Theorem 3, where the upper level cost is the predictive risk and A is invertible. In particular, we consider the same setup as above with forward operator A_2 and Tikhonov regularization. We plot the region for which $\hat{\alpha} = 0$ and boundary for when the condition of Theorem 3 holds in Figure 4. Since the upper level cost is different to the one considered in the above numerics, the contour plot of the upper level cost looks very different. We see that the condition of Theorem 3 characterises the positivity of $\hat{\alpha}$ well in this setting.

We now demonstrate the result of Corollary 1, where we are guaranteed positivity of $\hat{\alpha}$ provided that $A = I$, $\mathcal{R}$ is strictly convex, and the noise has zero mean. We fix ground truth $x^\star = [1, 0]^T$ and generate 1000 noisy realisations by perturbing $x^\star$ with Gaussian noise of mean $[0, 0]^T$, standard deviation $[0.1, 0.1]^T$. A plot of the ground truth and noisy realisations is shown in Figure 5a. To ensure the regularizer is strictly convex and differentiable, we consider

$$\mathcal{R}(x) = \frac{\beta}{2}\|x\|^2 + \sum_{i=1}^{n} \mathrm{hub}_\gamma(x_i),$$

for $\beta = \gamma = 0.01$. For regularization parameters in the linear discretisation of the interval $[0, 0.1]$ into 50 points, we plot the associated upper level cost in Figure 5b. We see that the optimal parameter is achieved at a strictly positive value. We now show that if the assumption on the noise is violated, we are not guaranteed positivity. For the same $x^\star$ we generate 1000 noisy realisations by perturbing $x^\star$ with Gaussian noise of mean $[-0.1, 0]^T$ and standard deviation $[0.1, 0.1]^T$. A plot of the ground truth and noisy realisations is shown in Figure 5c, and the associated upper level cost in Figure 5d. We see that 0 is the optimal parameter in this case.



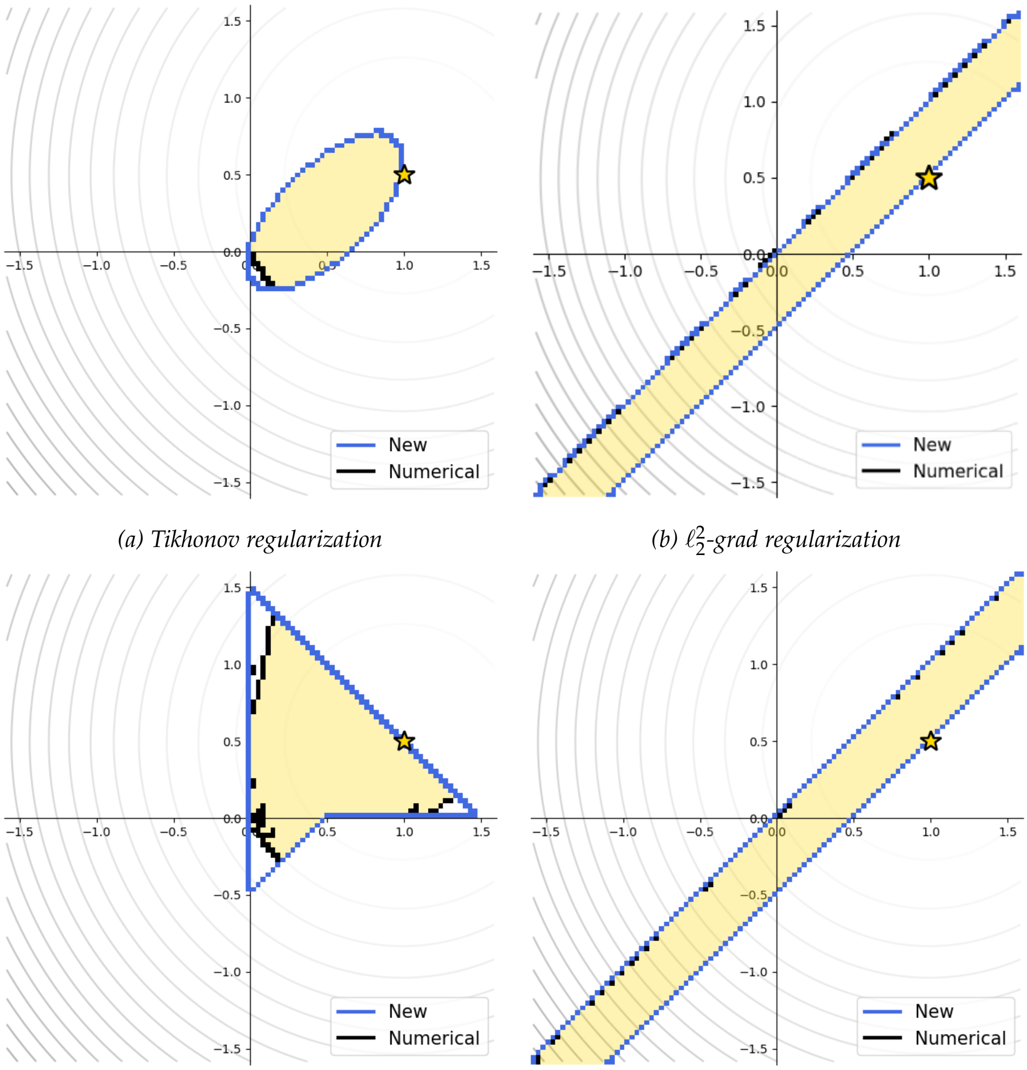
(a) Tikhonov regularization

(b) ℓ_2^2-grad regularization

(c) Huber regularization

(d) Huber TV regularization

Figure 2: Plots of the reconstruction space Ω for the denoising (A_1 forward operator) setting and various choices of regularizer, with the condition boundaries as detailed in Table 2. The ground truth $x^\star = [1, 0.5]$ is represented by a yellow star, and level sets of the upper level cost function are visible. The region where $\hat{\alpha} = 0$ is shaded yellow.



(a) Tikhonov regularization *(b) ℓ_2^2-grad regularization*

(c) Huber norm regularization *(d) Huber TV regularization*

Figure 3: Plots of the reconstruction space Ω for the deconvolution (A_2 forward operator) setting and various choices of regularizer, with the condition boundaries as detailed in Table 2. The ground truth $x^\star = [1, 0.5]^T$ is represented by a yellow star, and level sets of the upper level cost function are visible. The region where $\hat{\alpha} = 0$ is shaded yellow.

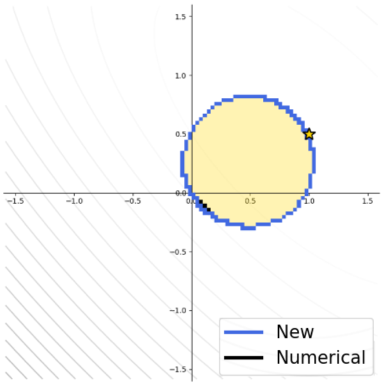


Figure 4: Plots of the reconstruction space Ω for the deconvolution (A_2 forward operator) setting, Tikhonov regularization, and predictive risk upper level cost. The ground truth $x^\star = [1, 0.5]^T$ is represented by a yellow star, and level sets of the upper level cost function are visible. The region where $\hat{\alpha} = 0$ is shaded yellow.

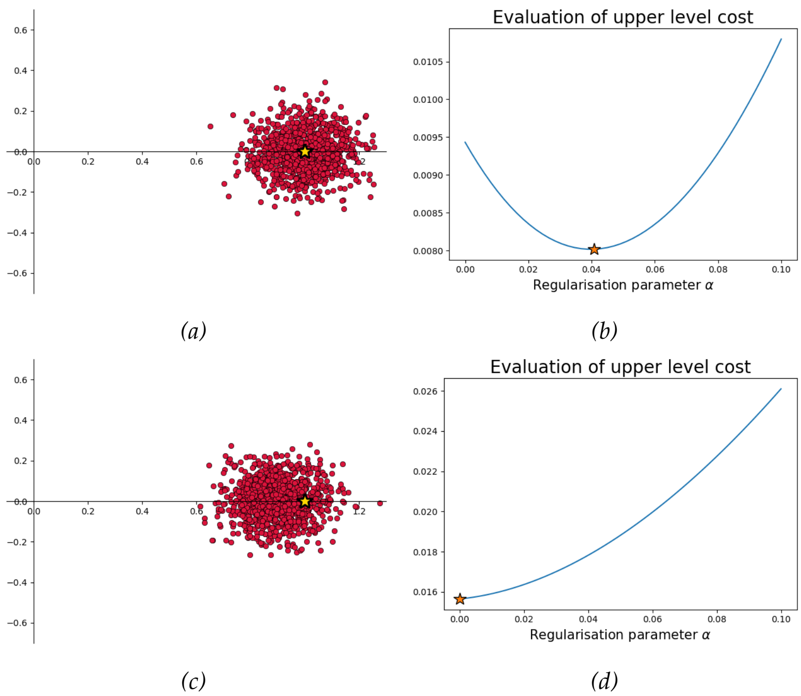
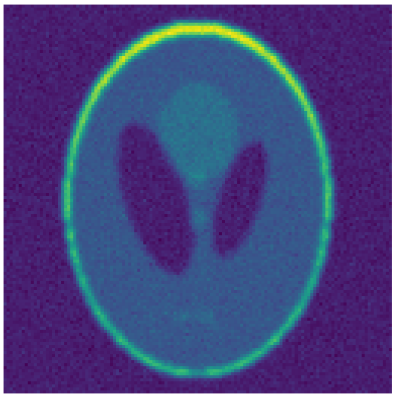
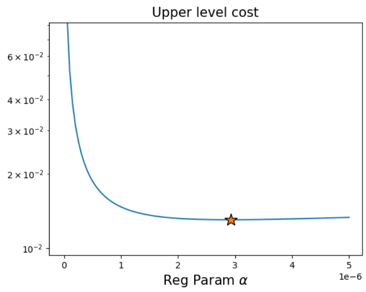


Figure 5: Pane 5a: ground truth $x^\star = [1, 0]^T$ indicated by a yellow star, and 1000 noisy realisations indicated by red dots where the corruption was additive Gaussian noise of mean $[-0.1, 0]^T$ standard deviation $[0.1, 0.1]^T$. Pane 5b: MSE upper level cost corresponding to the data in Pane 5a. The optimal regularization parameter is indicated by an orange star. Pane 5c: Similar plot as Pane 5a, but the noise has mean $[-0.1, 0]^T$ instead. Pane 5d: MSE upper level cost corresponding to the data in Pane 5c. The optimal regularization parameter is indicated by an orange star.

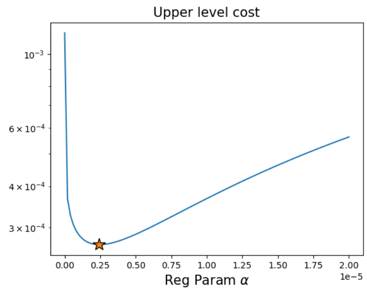
(a) Ground truth $x^\star$



(b) Corrupted measurement y



(c) MSE upper level cost



(d) Predictive risk upper level cost

Figure 6: Ground truth image, the 126×126 pixel image Shepp-Logan phantom, in Pane 6a, and the observed blurry and and noisy measurement in Pane 6b. Plot of the MSE and predictive risk upper level cost in Pane 6c and Pane 6d respectively. The optimal regularization parameter is indicated by an orange star.

5.2 Large scale problems

We now consider the well known Shepp-Logan phantom of size the 128×128 pixels, and displayed in Figure 6a. The observed measurement, displayed in Figure 6b, has been affected by a Gaussian blur with standard deviation 0.05 and then corrupted by Gaussian noise of mean zero and standard deviation $0.1\|Ax^\star\|$.

We consider $\mathcal{R}(x) = \frac{1}{2}\|\nabla x\|^2$, where ∇ calculates both the horizontal and vertical gradient of x and returns the vectorised concatenation of both results. The conditions of both Theorem 1 and Theorem 3 hold in this problem setup are so we are guaranteed that optimal regularization parameters are strictly positive. Indeed, plots of the MSE and predictive risk upper level cost are displayed in Figure 6c and Figure 6d respectively, and we see that the optimal regularization parameters lie away from 0 for both upper level cost functions.

6 Conclusion

In this work we determined a new sufficient condition involving the linearisation of the regularizer to deduce positivity of solutions to the bilevel learning problem (3), applicable to settings where the forward operator is injective and regularizer differentiable. In particular, a pointwise condition is presented, along with a result in expectation. While primarily focused on the MSE upper level cost, an extension to the predictive risk cost function was made. Furthermore, we showed that in a very realistic denoising setting our condition will always be satisfied and we are guaranteed positivity of optimal regularization parameters.

We have shown both analytically and empirically that the presented results characterise positivity well, and are an improvement on the condition used in existing theory.

Acknowledgements

MJE acknowledges support from EPSRC (EP/S026045/1, EP/T026693/1, EP/V026259/1) and the Leverhulme Trust (ECF-2019-478). The work of SG was partially supported by EPSRC under grant EP/T001593/1. SJS is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/S022945/1.

References

[1] Babak Maboudi Afkham, Julianne Chung, and Matthias Chung. Learning Regularization Parameters of Inverse Problems via Deep Neural Networks. *Inverse Problems*, 37(10):105017, 2021.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. Input Convex Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, page 146–155. JMLR, 2017.

[3] Qamrul Hasan Ansari, Lalitha C. S., and Monika Mehta. *General Convexity, Nonsmooth Variational Inequalities, and Nonsmooth Optimization*. CRC Press, Boca Raton, 1st edition, 2014.

[4] Stephan W. Anzengruber and Ronny Ramlau. Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26(2):025001, 2010.

[5] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta numerica*, 28:1–174, 2019.

[6] A.B. Bakushinskii. Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *USSR Computational Mathematics and Mathematical Physics*, 24(4):181–182, 1984.

[7] Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2017.

[8] Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta numerica*, 27:1–111, 2018.

[9] Thomas Bonesky. Morozov's discrepancy principle and Tikhonov-type functionals. *Inverse Problems*, 25(1):015015, 2009.

[10] Martin Burger. Bregman distances in inverse problems and partial differential equation. arXiv preprint arXiv:1505.05191, 2015.

[11] Martin Burger, E. Resmerita, and L. He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2/3):109–136, November 2007.

[12] Yair Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34:321–353, 1981.

[13] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta numerica*, 25:161–319, 2016.

[14] Yunjin Chen, Thomas Pock, René Ranftl, and Horst Bischof. Revisiting Loss-Specific Training of Filter-Based MRFs for Image Restoration. *Lecture notes in computer science*, 8142:271–281, 2014.

[15] Julianne Chung, Sarah Knepper, and James G. Nagy. Large-Scale Inverse Problems in Imaging. In *Handbook of Mathematical Methods in Imaging*, pages 43–86. Springer New York, New York, NY, 2011.

[16] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

[17] Caroline Crockett and Jeffrey A. Fessler. *Bilevel Methods for Image Reconstruction*, volume 15 of *Foundations and Trends® in Signal Processing*. 2022.

[18] Bernard Dacorogna. *Direct Methods in the Calculus of Variations*, volume 78 of *Applied Mathematical Sciences*. Springer New York, New York, NY, 2nd edition, 2008.

[19] Elisa Davoli, Rita Ferreira, Carolin Kreisbeck, and Hidde Schönberger. Structural changes in nonlocal denoising models arising through bi-level parameter learning. *Applied mathematics & optimization*, 88(1):9, 2023.

[20] Juan Carlos De los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. The structure of optimal parameters for image restoration problems. *Journal of mathematical analysis and applications*, 434(1):464–500, 2016.

[21] Juan Carlos De los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of mathematical imaging and vision*, 57(1):1–25, 2017.

[22] Matthias J. Ehrhardt and Lindon Roberts. Inexact Derivative-Free Optimization for Bilevel Learning. *Journal of mathematical imaging and vision*, 63(5):580–600, 2021.

[23] Matthias J. Ehrhardt and Lindon Roberts. Analyzing Inexact Hypergradients for Bilevel Learning. arXiv preprint arXiv:2301.04764, 2023.

[24] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer, 1996.

[25] Jonas Geiping and Michael Moeller. Parametric Majorization for Data-Driven Energy Minimization Methods. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[26] Saeed Ghadimi and Mengdi Wang. Approximation Methods for Bilevel Programming. arXiv preprint arXiv:1802.02246, 2018.

[27] Giorgio Giorgi and Sándor Komlósi. *Dini derivatives in optimization - Part I*, volume 15 of *Decisions in Economics and Finance*. Springer, 1992.

[28] Mark S. Gockenbach and Elaheh Gorgin. On the convergence of a heuristic parameter choice rule for Tikhonov regularization. *SIAM Journal on Scientific Computing*, 40(4):A2694–A2719, 2018.

[29] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, 1979.

[30] Jacques S. Hadamard. Sur les problemes aux derive espartielles et leur signification physique. In *Princeton University Bulletin*, pages 49–52, 1902.

[31] Per Christian Hansen. *Discrete inverse problems : insight and algorithms*. Fundamentals of algorithms. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 2010.

[32] Per Christian Hansen, James G. Nagy, and Dianne P. O'Leary. *Deblurring images : matrices, spectra, and filtering*. Fundamentals of algorithms. Society for Industrial and Applied Mathematics, 2006.

[33] Gernot Holler, Karl Kunisch, and Richard C Barnard. A Bilevel Approach for Parameter Learning in Inverse Problems. *Inverse Problems*, 34(11):115012, 2018.

[34] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[35] Bangti Jin and Peter Maass. Sparsity regularization for parameter identification problems. *Inverse Problems*, 28(12):123001–70, 2012.

[36] R. Kannan and Carole King Kreuger. *Advanced Analysis on the Real Line*. Universitext. Springer, 1996.

[37] Steven G Krantz and Harold R Parks. The Implicit Function Theorem. Springer, New York, 2012.

[38] Karl Kunisch and Thomas Pock. A Bilevel Optimization Approach for Parameter Learning in Variational Models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

[39] Subhadip Mukherjee, Sören Dittmer, Zakhar Shumaylov, Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Learned convex regularizers for inverse problems. arXiv preprint arXiv:2008.02839, 2020.

[40] Johan Nuyts, Dirk Bequé, Patrick Dupont, and L. Mortelmans. A concave prior penalizing relative differences for maximum-a-posteriori reconstruction in emission tomography. *IEEE Transactions on Nuclear Science*, 49(1):56–60, 2002.

[41] Juan Peypouquet. *Convex Optimization in Normed Spaces : Theory, Methods and Examples*, volume 0 of *SpringerBriefs in Optimization*. Springer, 1st edition, 2015.

[42] Sidney I. Resnick. *A Probability Path*. Modern Birkhäuser Classics. Birkhäuser, 1st edition, 2014.

[43] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[44] Ferdia Sherry, Martin Benning, Juan Carlos De los Reyes, Martin J. Graves, Georg Maierhofer, Guy Williams, Carola-Bibiane Schönlieb, and Matthias J. Ehrhardt. Learning the sampling pattern for mri. *IEEE Transactions on Medical Imaging*, 39(12):4310–4321, 2020.

[45] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A Review on Bilevel Optimization: From Classical to Evolutionary Approaches and Applications. *IEEE transactions on evolutionary computation*, 22(2):276–295, 2018.

[46] Nicolas Zucchet and João Sacramento. Beyond backpropagation: bilevel optimization through implicit differentiation and equilibrium propagation. arXiv preprint arXiv:2205.03076, 2022.