ON THE EXISTENCE OF OPTIMAL MULTI-VALUED DECODERS AND THEIR ACCURACY BOUNDS FOR UNDERSAMPLED INVERSE PROBLEMS

NINA M. GOTTSCHLING, PAOLO CAMPODONICO, VEGARD ANTUN, AND ANDERS C. HANSEN

ABSTRACT. Undersampled inverse problems occur everywhere in the sciences including medical imaging, radar, astronomy etc., yielding underdetermined linear or non-linear reconstruction problems. There are now a myriad of techniques to design decoders that can tackle such problems, ranging from optimization based approaches, such as compressed sensing, to deep learning (DL), and variants in between the two techniques. The variety of methods begs for a unifying approach to determine the existence of optimal decoders and fundamental accuracy bounds, in order to facilitate a theoretical and empirical understanding of the performance of existing and future methods. Such a theory must allow for both single-valued and multi-valued decoders, as underdetermined inverse problems typically have multiple solutions. Indeed, multi-valued decoders arise due to non-uniqueness of minimizers in optimisation problems, such as in compressed sensing, and for DL based decoders in generative adversarial models, such as diffusion models and ensemble models. In this work we provide a framework for assessing the lowest possible reconstruction accuracy in terms of worst- and average-case errors. The universal bounds bounds only depend on the measurement model F, the model class $\mathcal{M}_1 \subseteq \mathcal{X}$ and the noise model $\mathcal{E}$. For linear F these bounds depend on its kernel, and in the non-linear case the concept of kernel is generalized for undersampled settings. Additionally, we provide multi-valued variational solutions that obtain the lowest possible reconstruction error.

1. INTRODUCTION

Finite dimensional inverse problems are ubiquitous in the computational sciences as they naturally appear in a plethora of applications. An incomplete list of examples includes most types of computational imaging [2, 14, 25], matrix completion [28], parametric PDEs/system identification [1], phase retrieval [30, 46, 86], quantized sampling [24] etc. Most often, we can express these problems using the following general model:

$$\text{Given noisy measurements } y = F(x, e) \text{ of } x \in \mathcal{M}_1 \subset \mathbb{C}^N \text{ and } e \in \mathcal{E} \subset \mathbb{C}^k, \text{ recover } x. \tag{1}$$

Here x represents the signal of interest, while e represents the noise. The sets $\mathcal{M}_1$ and $\mathcal{E}$ describe the signals of interest and the potential noise, respectively. The function $F\colon \mathcal{X} \times \mathcal{E} \to \mathbb{C}^m$ models the forward process, and is deliberately kept general to encompass many known models. Examples include

$$y = G(x) + e, \qquad \text{(additive noise)} \tag{2}$$

$$y = G(x) \odot e, \qquad \text{(multiplicative noise)} \tag{3}$$

$$y = G(x) \odot e_1 + e_2, \qquad \text{(mixture of multiplicative and additive noise)} \tag{4}$$

where $G\colon \mathbb{C}^N \to \mathbb{C}^m$ is a linear or non-linear forward map, and the dimension of the set $\mathcal{E}$ depends on the considered model in such a way that the point-wise multiplication $\odot$ is defined.

In most instances of interest, the problem of recovering x given y is ill-posed or undersampled, unless further assumptions are made. That is, the forward model F is either poorly conditioned or dimensionality reducing. A standard assumption for making the problem (1) tractable is to assume that the noiseless forward model is injective when restricted to the model class $\mathcal{M}_1$. This assumption is a cornerstone for models with additive noise, as it allows for accurate reconstruction up to the noise level in both the linear [26, 51, 93] and non-linear [65] setting. Traditionally, the set $\mathcal{M}_1$ has been given a precise mathematical description and reconstruction methods have been designed for the given choice of $\mathcal{M}_1$. Examples of sets $\mathcal{M}_1$ include sparse vectors [29], union of subspaces [22], manifolds [13], sparsity in a given basis [3] or frame [83], matrices with low rank [31, 62], etc.

More recently, data-driven approaches [11, 61] have emerged as an alternative to many of these standard methods, often promising superior performance [101]. Methods based on data typically do not specify the solution set $\mathcal{M}_1$, but aim to learn the reconstruction mapping $\Psi\colon \mathbb{C}^m \to \mathcal{M}_1$ given a finite number of training data $\mathcal{T} \subset \mathcal{M}_1$ and access to the forward model F. This approach has the advantage that the learned set $\mathcal{M}_1$ might provide a better approximation of the underlying data stemming from a real-world process, rather than a potentially simplistic abstract mathematical modelling of the set $\mathcal{M}_1$. However, the challenge with these methods is that the learned mapping Ψ tends to produce accurate reconstructions of all elements in $\mathcal{T}$, regardless of whether the noiseless problem is injective on $\mathcal{M}_1$ or not. This has made these methods susceptible to both hallucinations [77] and instabilities [9].

The purpose of this work is to develop a mathematical framework which provides a notion of optimality when F is not necessarily injective on $\mathcal{M}_1$, and to provide variational expression for the optimal mappings in this setting. Our main contributions are the following:

(I) *We provide a framework which allows for the classification of the lowest achievable reconstruction accuracy for* (1). Note that many such frameworks exist in the literature. For example, Gelfand widths in the noiseless linear setting [81], best k-term approximation [35], and to obtain the worst-case bounds set-valued decoders on Banach spaces have been considered in [10] and then extended to metric spaces in [73], and when noise is included one has the notion of optimal learning [21] and generalized instance optimality [26]. See also [75] for a survey of early works on optimal recovery. In many of these works, the underlying spaces are infinite or finite dimensional Banach spaces, and the reconstruction error is measured by the given norm. A contribution of this work is to consider the general setting of metric spaces, to allow for set-valued decoders which naturally arise as solutions of undersampled inverse problems, and to extend previous studies on worst-case scenarios to the average error scenario.

(II) *We derive explicit formulas for set-valued reconstruction mappings which achieve the lowest possible reconstruction error, both in a worst-case sense and in a probabilistic sense.* Finding and analyzing reconstruction mappings which achieve optimal or near-optimal recovery – given some notion of optimality – is an essential question in inverse problems, and many mappings exists [21, 26, 51, 82, 93]. In this work, we provide explicit formulas for possibly set-valued reconstruction mappings that achieve the lowest possible reconstruction error. Moreover, a contribution of this work is to prove that the optimal mappings in the average error scenario are measurable, compact valued and admits a measurable single-valued selector by utilising the Measurable Maximum Theorem, [5, Thm. 18.19] and Hausdorff distance. As such, this work provides a first and necessary step, before any numerical procedure can provide statistical approximations to these mappings.

(III) *We provide lower and upper bounds on the lowest achievable reconstruction accuracy described in* (I). The bounds solely depend on the forward operator, signal and noise class of the inverse problem, and not on the method or decoder used to solve (1). Hence, the performance of any – possibly set-valued – method can be evaluated with respect to the problem's intrinsic optimal accuracy. However, there is a trade-off for obtaining lower bounds compared to other approximation error bounds, such as in Compressed Sensing (CS) that only provides upper bounds. We refer to [51] for a comprehensive treatment in CS and the work of [26] as an extension thereof. The key point in order to obtain lower bounds on reconstruction accuracy is to consider worst-case and average errors instead of point-wise approximation errors. Such worst-case error bounds can be related to the Chebycheff radius, as in [21], and analogous worst- and average- case error bounds are established in [82] in the case of normed spaces and single-valued decoders .

Remark 1.1. Extending previous work from normed spaces to metric spaces is of particular relevance in imaging applications. In fact, the ℓ^2 or ℓ^1 norm are not the only candidates considered for image quality

assessment. A prominent example is the structural similarity index (SSI), firstly proposed to assess image quality in [97], does not constitute a norm but can rather be related to a metric [27].

Remark 1.2 (**Multi-valued decoders, computability and randomised algorithms**). Many of the decoders provided in theory will not be computable, as the phenomenon of generalised hardness of approximation [8, 16,39,47,54] – within the Solvability Complexity Index (SCI) hierarchy [18,19,37,38,56] – happen in many inverse problems. This includes a wide range of neural network decoders, as well as compressed sensing decoders. Typically, they can only be computed to a certain accuracy $\epsilon_0 > 0$, referred to as the approximation threshold, and it is the size of the approximation threshold that determines if the decoder can be used in practice. However, as demonstrated in [16], when the decoder is multi-valued, randomised algorithms may actually help mitigating the non-computability. This is only possible for multi-valued decoders, as single-valued decoders that are non-computable will have no help from randomised algorithms [16,42].

2. Preliminaries

2.1. **The forward map and the model class.** In the following we introduce the main assumptions and setup. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be non-empt sets. Furthermore, consider subsets $\mathcal{M}_1 \subset \mathcal{X}$ and $\mathcal{E} \subset \mathcal{Z}$, referred to as the *model class* and the *noise class*, respectively. As outlined in the introduction, we consider inverse problems with a given forward map $F\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, and we seek to solve the problem:

$$\text{Given noisy measurements } y = F(x,e) \text{ of } x \in \mathcal{M}_1 \text{ and } e \in \mathcal{E}, \text{ recover } x, \tag{5}$$

where e represents noise and x is the signal of interest. The goal of solving the inverse problem is to produce an approximation $\hat{x}$ of the true solution x. In order to quantify the error of the approximation $\hat{x} \in \mathcal{X}$ to the sample signal $x \in \mathcal{M}_1$, it is common to assume [35,51,81] that $\mathcal{X}$ is a normed space, or even a Banach or Hilbert space. In the current work, we extend the scope to the more general setting of metric spaces. The use of metrics – rather than norms – allows for a more general theory, in which normed vector spaces are a special case. Thus, we equip the sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ with metrics $d_{\mathcal{X}}$, $d_{\mathcal{Y}}$ and $d_{\mathcal{Z}}$, respectively, and on each set we consider the topology induced by the respective metric.

To avoid certain pathological cases for metric spaces, we make assumptions on the choice of metrics involved. Our first assumption is the following.

Assumption 1.

(i) We assume that the metrics $d_{\mathcal{X}}$, $d_{\mathcal{Y}}$, $d_{\mathcal{Z}}$ are chosen so that the topologies induced by the metrics are second countable (they admit a countable base).

(ii) We assume that the metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Z}}$ satisfy the Heine-Borel property, i.e., that all closed and bounded sets are compact.

Here, the first assumption ensures that the topologies induced by the metrics are separable, which is a convenient assumption for both theoretical and practical purposes. In particular, this assumption excludes metrics such as the discrete metric over an uncountable set, as one cannot find a countable base for an uncountable set using this metric.

For the second assumption, we recall that in metric spaces, a compact subset is automatically closed and bounded; however, the converse does not necessarily hold, unless the metrics satisfy the Heine-Borel property. Note that this (ii) is not implied by (i), as for example the bounded metric $d_{\mathcal{X}}(x,y) = \min\{\|x - y\|_{\ell^2}, 1\}$ on $\mathcal{X} = \mathbb{C}^N$ satisfies assumption (i) (since $d_{\mathcal{X}}$ induces the same topology as the Euclidean topology, which is second countable), but not assumption (ii) (since $\mathbb{C}^N$ is closed and bounded but not compact with respect to $d_{\mathcal{X}}$). Finally, note that the Heine-Borel property implies in particular that $\mathcal{X}$ is a complete separable metric space, usually referred to as a *Polish space*.

2.2. **Multivalued reconstruction maps.** Now that the preliminaries have been clarified, we return to the problem of solving the inverse problem in (5). Let

$$\mathcal{M}_2^{\mathcal{E}} = F(\mathcal{M}_1 \times \mathcal{E}) = \{y \in \mathcal{Y} : \exists (x, e) \in \mathcal{M}_1 \times \mathcal{E} \text{ s.t. } y = F(x, e)\}$$

denote the image of $F\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ on the class $\mathcal{M}_1 \times \mathcal{E}$. A reconstruction mapping for (5) is a mapping that takes a measurement $y = F(x, e)$ and returns one or many approximations $\hat{x}$ to the acquired signal x. That is, a *reconstruction mapping* is a multi-valued function $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows 2^{\mathcal{X}}$, where $2^{\mathcal{X}}$ denotes the power set of the set $\mathcal{X}$. Note that it is necessary to consider set-valued maps, as model-based methods often write the reconstruction mapping as an optimization problem. These problems might not have unique solutions, see e.g., [39, SI, Sec. 3.B] for a few standard examples. Finally, we mention that given a multivalued mapping $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows 2^{\mathcal{X}}$, we call a function $f\colon \mathcal{M}_2^{\mathcal{E}} \to \mathcal{X}$ a *selector* for φ if it satisfies $f(y) \in \varphi(y)$ for each $y \in \mathcal{M}_2^{\mathcal{E}}$.

Distances between sets in $\mathcal{X}$ will be measured via the Hasudorff distance. Given subsets $A, B \subseteq \mathcal{X}$, their *Hasudorff distance* is

$$d_{\mathcal{X}}^H(A, B) := \max\Big\{ \sup_{a\in A} \inf_{b\in B} d_{\mathcal{X}}(a, b), \sup_{b\in B} \inf_{a\in A} d_{\mathcal{X}}(a, b)\Big\}.$$

The Hausdorff distance satisfies all the properties of a metric only when restricted to the subsets of $\mathcal{X}$ that are bounded (which ensures $d_{\mathcal{X}}^H$ is finite) and closed (which ensures that $d_{\mathcal{X}}^H(A, B) = 0$ only if $A = B$). With a slight abuse of notation, we will denote $d_{\mathcal{X}}^H(\{a\}, B)$ by $d_{\mathcal{X}}^H(a, B)$. The Hausdorff distance between a point and a set should not be confused with the usual distance between a point and a set, $\operatorname{dist}_{\mathcal{X}}(a, B) := \inf_{b\in B} d_{\mathcal{X}}(a, b)$. Indeed, note that $\operatorname{dist}_{\mathcal{X}}(a, B) = \inf_{b\in B} d_{\mathcal{X}}(a, b) \leqslant \sup_{b\in B} d_{\mathcal{X}}(a, b) = d_{\mathcal{X}}^H(a, B)$.

We require an extra assumption that relates the forward map F and the model class $\mathcal{M}_1$. In particular, we require that for every y, the set of possible true solutions x's is bounded. Throughout the text, we denote by $\pi_1\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}, (x, e) \mapsto x$ the projection on the first component.

Assumption 2. We assume that for every $y \in \mathcal{M}_2^{\mathcal{E}}$ the *feasible set*

$$F_y := \pi_1(F^{-1}(y)) \cap \mathcal{M}_1 = \{x \in \mathcal{M}_1 : \exists e \in \mathcal{E} \text{ s.t. } F(x, e) = y\} \tag{6}$$

is bounded.

The feasible set F_y consists of all the candidate solutions x's that are consistent with the measurement y, for some realisation of the noise e.

The previous condition is satisfied, for example, if the model class $\mathcal{M}_1$ is compact, as assumed in e.g., [21].

Finally, we clarify our notation for objects in a metric space (X, d). We denote by $B_d(x, r)$ the *closed* ball of center $x \in X$ and radius $r \geqslant 0$. If $A \subseteq X$, then the closed ball around A of radius $r \geqslant 0$ is

$$B(A, r) = \{x \in X : \operatorname{dist}_d(x, A) \leqslant r\} = \bigcup_{x\in A} B(x, r)$$

where we recall that $\operatorname{dist}_d(x, A) = \inf_{a\in A} d(x, a)$. Finally, the diameter of $A \subseteq X$ is $\operatorname{diam}_d(A) := \sup\{d(a, a') : a, a' \in A\}$.

3. Main results

In this section we describe the considered framework. The section is divided into two subsections that both introduce the notion of an *optimal mapping* and the *kernel size*, but for different measures of accuracy. In particular, Section 3.1 presents an analysis based on the *worst-case* error, whereas Section 3.2 considers a probabilistic model class, where an *average* error is analyzed. Theorems 3.4 and 3.9 contain the main results, in Section 3.1 and 3.2, respectively. These theorems bound the so-called *optimality constant* of a optimal mapping in terms of the *kernel size* of the problem, and provide an explicit expression for an optimal reconstruction mapping. Finally, in Section 4 we elaborate on how these quantities relate to common frameworks in the literature. The proofs of the main results are referred to Section 7.

3.1. **Worstcase optimality bounds and a optimal map with worst-case noise.** What is the best possible reconstruction error one can achieve for a given model class $\mathcal{M}_1$? This is an old question [75], that has been asked many times, and in different settings [21, 26, 35, 44, 50, 65, 93, 94]. In the definition below, we investigate this question for measurements contaminated by worst-case noise. This definition has appeared in [44,50] for the linear model with additive noise in a normed vector space, and is here generalized to metric spaces.

Definition 3.1 (Optimality constant with worst-case noise)**.** The *optimality constant with worst-case noise* of the inverse problem (1) is

$$c_{\mathrm{opt}}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) = \inf_{\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}} \sup_{x\in\mathcal{M}_1} \sup_{e\in\mathcal{E}} d_{\mathcal{X}}^H(x, \varphi(F(x,e))).$$

A mapping $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ that attains such an infimum is called an *optimal map with worst-case noise*.

Given the optimality constant with worst-case noise of an inverse problem (1), the question arises on how to upper and lower bound such optimality constant. As shown in Theorem 3.4, the upper and lower bounds can be obtained by referring to a constant intrinsic to the problem, its *kernel size with worst-case noise*, as defined below.

Definition 3.2 (Kernel size with worst-case noise)**.** The *kernel size with worst-case noise* of the problem (1) is

$$\mathrm{kersize}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) = \sup_{\substack{(x,e),(x',e')\in\mathcal{M}_1\times\mathcal{E} \text{ s.t.} \\ F(x,e)=F(x',e')}} d_{\mathcal{X}}(x, x'). \tag{7}$$

Remark 3.3. The kernel size with worst-case noise has been considered under different names in previous work. This includes, but is not limited to, the supremum taken over the measurements y of diameter of the Chebycheff balls of the feasible sets F_y, similar to [21], or the diameter of information in [82]. However, as this work focuses on characterizing accuracy bounds for undersampled inverse problems, we opt for referring to the above defined constant as kernel size. In Example 1 in Section 4, more details on this relation are presented.

The kernel size with worst-case noise gives the maximum distance between any two points $x, x' \in \mathcal{M}_1$ with identical measurement $y = F(x,e) = F(x',e')$ for some noise vectors $e, e' \in \mathcal{E}$. It is worth observing that the optimality constant and the kernel size with worst-case noise also can be expressed as follows.

$$c_{\mathrm{opt}}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) = \inf_{\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}} \sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \sup_{x\in F_y} d_{\mathcal{X}}^H(x, \varphi(y))$$

$$\mathrm{kersize}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) = \sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \mathrm{diam}_{d_{\mathcal{X}}}(F_y).$$

This should clarify the connection with the feasible sets F_y. In our first theorem, we provide an upper and lower bound on the optimality constant in terms of the kernel size with worst-case noise. Moreover, we provide a variational expression for an optimal map with worst-case noise.

Theorem 3.4 (Worst case optimality bounds)**.** *Under Assumptions 1 and 2, the following holds.*

(i) We have that

$$\mathrm{kersize}^{w}(F, \mathcal{M}_1, \mathcal{E})/2 \leqslant c_{\mathrm{opt}}^{w}(F, \mathcal{M}_1, \mathcal{E}) \leqslant \mathrm{kersize}^{w}(F, \mathcal{M}_1, \mathcal{E}). \tag{8}$$

(ii) The map

$$\Psi(y) = \operatorname*{argmin}_{z\in\mathcal{X}} \sup_{(x,e)\in F_y} d_{\mathcal{X}}(x, z) = \operatorname*{argmin}_{z\in\mathcal{X}} d_{\mathcal{X}}^H(z, F_y), \tag{9}$$

has non-empty, compact values and it is an optimal map with worst-case noise.

The previous theorem illustrates a fundamental limit for the inverse problem (5). Indeed, for (5) one would hope to find a solution whose error is as close to zero as possible. However, the lower bound in (8) shows that there is a fundamental constant intrinsic to the problem – the kernel size with worst-case noise – such that no worst-case reconstruction error can be made smaller than this constant for all possible choices of $x \in \mathcal{M}_1$.

Remark 3.5. Note that analogous bounds as in (8) under slightly different assumptions have been obtained in a variety of previous works, including but not limited to [10,21,73,82]. However, to the best of our knowledge the characterisation of the multi-valued optimal map (9) as compact-valued has not been considered under the same assumptions.

3.2. **Average optimality bounds and optimal map with average error.** The previous theorem provides fundamental limits when performance is assessed by considering worst-case reconstruction error for a given inverse problem of the form (5). Such bounds naturally lead to very pessimistic estimates of performance. This motivates the need for model that considers the average error of a reconstruction mapping, given a probabilistic model of the data. Below, we will adapt the worst-case setup considered in the previous section to a probabilistic model. Note, that in the machine learning literature, this notion of average error is often referred to as the *risk* of a given reconstruction mapping [85, Ch. 3] [49, Ch. 1].

We start by introducing the notation. For a topological space (X, τ), the *Borel σ-algebra* of X, denoted by $\mathcal{B}(X)$, is the σ-algebra generated by the family τ of open sets. Elements of $\mathcal{B}(X)$ are called *Borel sets*. If X is a metric space, then $\mathcal{B}(X)$ is the smallest family of sets containing all the open sets that are closed under countable intersections and countable disjoint unions. See, e.g., [5, Cor. 4.16] for further reference. A *Borel measure* on X is a measure defined on the Borel σ-algebra $\mathcal{B}(X)$. Borel measurable functions are defined analogously.

Assumption 3. We consider the measure spaces $(\mathcal{X} \times \mathcal{Z}, \mathcal{B}(\mathcal{X} \times \mathcal{Z}))$, $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ and assume that $F\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is Borel measurable. We equip $\mathcal{X} \times \mathcal{Z}$ with a finite Borel measure μ supported on $\mathcal{M}_1 \times \mathcal{E}$, and we equip $\mathcal{Y}$ with the pushforward measure $F_*\mu$ given by $(F_*\mu)(E) = \mu(F^{-1}(E))$ for every $E \in \mathcal{B}(\mathcal{Y})$.

Next, we need to define conditional probabilities on $\mathcal{M}_1 \times \mathcal{E}$ for the considered measure μ for different values of $y \in \mathcal{M}_2^{\mathcal{E}}$. To compute such conditional probabilities, we consider a *disintegration* of the measure μ. Intuitively, a disintegration of the measure μ given the mapping F, is a family of probability measures $\{\mu^y\}_{y \in \mathcal{M}_2^{\mathcal{E}}}$ such that for each $y \in \mathcal{M}_2^{\mathcal{E}}$, μ^y is concentrated on the set $F^{-1}(y)$, and which allows to reconstruct the original measure μ by integration. Note that these set $\{F^{-1}(y)\}_{y \in \mathcal{M}_2^{\mathcal{E}}}$ might have measure zero with respect to μ. The concept of disintegration is presented the following definition, which is taken from [34] and adapted to our Assumption 3. See also [23, Ch. 10] or [64, Ch. 1] for more on disintegration. Moreover, note that in [82] a similar tool – that of a *regular conditional probability density* – is used in order to prove analogous bounds when restricted to normed spaces and single-valued decoders.

Definition 3.6. A *disintegration* of the measure μ along the measurable function F is a family $\{\mu^y\}_{y \in \mathcal{M}_2^{\mathcal{E}}}$ of probability measures on $\mathcal{M}_1 \times \mathcal{E}$ such that

(i) for $F_*\mu$-almost every $y \in \mathcal{M}_2^{\mathcal{E}}$, μ^y is a probability measure concentrated on $F^{-1}(y)$, i.e., $\mu^y(\mathcal{M}_1 \times \mathcal{E} \backslash F^{-1}(y)) = 0$ for $F_*\mu$-almost all $y \in \mathcal{M}_2^{\mathcal{E}}$.

and such that, for each non-negative Borel measurable function f on $\mathcal{M}_1 \times \mathcal{E}$,

(ii) the function

$$y \mapsto \int_{\mathcal{M}_1 \times \mathcal{E}} f(x, e)\, d\mu^y(x, e), \quad \text{with} \quad y \in \mathcal{M}_2^{\mathcal{E}},$$

is Borel measurable,

(iii) and

$$\int_{\mathcal{M}_1\times\mathcal{E}} f(x,e)\,d\mu(x,e) = \int_{\mathcal{M}_2^{\mathcal{E}}} \left(\int_{F^{-1}(y)} f(x,e)\,d\mu^y(x,e) \right) d(F_*\mu)(y).$$

Note that while we assume that μ is a finite measure (and not necessarily a probability measure), the definition above require that μ^y's are probability measures. In section 7.1 we show that a disintegration of μ always exists and that it is essentially unique in our setting given Assumption 4, stated below.

Assumption 4. We assume that $\mathcal{M}_1$ is compact and we assume that $\mathcal{E}$ is a Borel set.

Note that Assumption 4 implies Assumption 2, which will therefore be omitted from now on. Next, let

$$r_\varphi : \mathcal{M}_1 \times \mathcal{E} \to [0,+\infty], \quad r_\varphi(x,e) = d_{\mathcal{X}}^H(x,\varphi(F(x,e))$$

denote the *residual map* of a given reconstruction mapping $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ and let

$$\mathcal{C} \coloneqq \{\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X} : r_\varphi \text{ is Borel measurable}\}. \tag{10}$$

denote the set of all reconstruction mappings with a Borel measurable residual map. For $p \in [1,\infty]$ and a given a reconstruction map $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$, we denote its pth order error by

$$\mathrm{Err}^{\mathrm{a}}(\varphi,p) = \begin{cases} \left(\int_{\mathcal{M}_1\times\mathcal{E}} d_{\mathcal{X}}^H\Big(x,\varphi(F(x,e))\Big)^p \, d\mu(x,e) \right)^{\frac{1}{p}} & \text{for } 1 \leqslant p < \infty \\ \operatorname{essup}_{(x,e)\in\mathcal{M}_1\times\mathcal{E}} \; d_{\mathcal{X}}^H(x,\varphi(F(x,e))) & \text{for } p = \infty \end{cases} \tag{11}$$

where the essential supremum for $p=\infty$ is taken with respect to the measure μ. Concretely, for $p \in [1,\infty]$ this quantity is the $L^p(\mathcal{M}_1\times\mathcal{E},\mu)$ norm of the residual function r_φ.

This definition of reconstruction error generalizes the usual L^p-norms to include a set-valued mapping and a metric $d_{\mathcal{X}}$ that is not necessarily induced by a norm in the single-valued case. However, the proposed setup includes many standard expectations as special cases. Indeed, let φ be a single-valued map and let $d_{\mathcal{X}}$ be given by an ℓ^q norm. Then, if we choose $p=q=1$, we recover the expected absolute error, whereas for $p=q=2$ we find the expected root squared error. Moreover, for $p=\infty$, the choice of metric $d_{\mathcal{X}}$ corresponds to the choice of the loss function in a machine learning setting [85, Ch. 3] [49, Ch. 1]. One can also, view the case where $p=\infty$, as a weak form of worst-case error where all irregularities on sets of measure zero are ignored.

With the error term in (11) defined, we can now establish the notion of optimality constant with average error and kernel constant with average error.

Definition 3.7 (Optimality constant with average error)**.** For $p \in [1,\infty]$ the *optimality constant with average error* of order p for the inverse problem (5) is

$$c_{\mathrm{opt}}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p) = \inf_{\varphi\in\mathcal{C}} \mathrm{Err}^{\mathrm{a}}(\varphi,p). \tag{12}$$

A map $\varphi \in \mathcal{C}$ attaining the infimum in (12) for a given p is called an *optimal map with average error* of order p.

Definition 3.8 (Average kernel size)**.** The average kernel size of the inverse problem (1) for $p \in [1,\infty)$ is given by

$$\mathrm{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p) = \left(\int_{\mathcal{M}_2^{\mathcal{E}}} \int_{F_y} \int_{F_y} d_{\mathcal{X}}(x,x')^p \, d\mu^y(x,e)\, d\mu^y(x',e')\, d(F_*\mu)(y) \right)^{\frac{1}{p}},$$

and for $p=\infty$ it is

$$\mathrm{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},\infty) = \operatorname*{essup}_{y\in\mathcal{M}_2^{\mathcal{E}}} \operatorname*{essup}_{\substack{(x,e)\in F_y \\ (x',e')\in F_y}} d_{\mathcal{X}}(x,x'), \tag{13}$$

where the left and right essential suprema are taken with respect to $F_*\mu$ on $\mathcal{M}_2^{\mathcal{E}}$ and μ^y on F_y, respectively.

Intuitively, the average kernel size measures the average distance between the x-components of $(x, e), (x', e') \in \mathcal{M}_1 \times \mathcal{E}$ that have the same measurement $F(x, e) = F(x', e') = y$ in the case of $p \in [1, \infty)$, and the maximum of such distances (up to negligible sets) when $p = \infty$.

Our main theorem in this section is stated below. It bounds the average optimality constant in terms of the average kernel size. It also ensures the existence of an optimal map, provides a variational expression for one such map and establishes some regularity properties of this mapping.

Theorem 3.9. *Under Assumptions 1, 3, 4, the following holds for every $p \in [1, \infty]$,*

(i) We have that

$$\operatorname{kersize}^a(F, \mathcal{M}_1, \mathcal{E}, p)/2 \leqslant c^a_{\mathrm{opt}}(F, \mathcal{M}_1, \mathcal{E}, p) \leqslant \operatorname{kersize}^a(F, \mathcal{M}_1, \mathcal{E}, p).$$

(ii) Consider $p \in [1, \infty)$, and let $\Psi \colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ be given by

$$\Psi(y) = \operatorname*{argmin}_{z \in \mathcal{X}} \int_{F_y} d_{\mathcal{X}}(x, z)^p \, d\mu^y(x, e). \tag{14}$$

Then we have the following: Ψ has compact values, it is measurable and it admits a measurable selector. Moreover, Ψ is an optimal map with average error of order p.

(iii) Consider $p = \infty$, and let $\Psi \colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ be given by

$$\Psi(y) = \operatorname*{argmin}_{z \in \mathcal{X}} \operatorname*{essup}_{(x,e) \in F_y} d_{\mathcal{X}}(x, z). \tag{15}$$

Then we have the following: Ψ has compact values, it is measurable and it admits a measurable selector. Moreover, Ψ is an optimal map with average error of order p.

Remark 3.10 (Measurability of multi-valued mappings). There exist several notions of measurability for multi-valued mappings. The precise definition of the term used in the theorem above can be found in Definition 7.6.

A key finding in Theorem 3.9 is that the possibly multi-valued Ψ admits a measurable selector, i.e. a *single-valued* function $\psi : \mathcal{M}_2^{\mathcal{E}} \to \mathcal{X}$ which is Borel measurable and such that $\psi(y) \in \Psi(y)$ for every y. Thus, Theorem 3.9 not only provides fundamental limits for the reconstruction error of an inverse problem, but also gives a (variational) expression for a mapping that is optimal. Moreover, such a reconstruction mapping satisfies some regularity properties, which one usually hopes for when solving an inverse problem. One can argue that measurability is still a weak condition, but in order to obtain stronger conditions (such as continuity of the reconstruction), one would need to impose stronger assumptions on the problem.

4. Examples and Applications

Example 1: Linear inverse problems with the robust null-space property. The robust null space property was first introduced as a necessary condition for uniform recovery of sparse vectors from linear measurements. See [51] for a historical overview. Later it has been extended to a wide range of model classes $\mathcal{M}_1$, including low-rank [62] and sparsity in levels [15] models. The most general version of the property appeared in [26] which considers general sets $\mathcal{M}_1$. We now recall a specialized version of the property from [26], and we wi ll see how this relates to the notion of worst-case kernel size.

Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be vector spaces, with $\mathcal{Y} = \mathcal{Z}$, and let $A \colon \mathcal{X} \to \mathcal{Y}$ be a linear mapping that is onto $\mathcal{Y}$. We consider the additive linear model $F(x, e) = Ax + e$. Let $|||\cdot|||_1$ and $|||\cdot|||_2$ be norms on $\mathcal{X}$, and let $|||\cdot|||_3$ be a norm on $\mathcal{Y}$ (slightly weaker conditions are used in [26]). Furthermore, let $\mathcal{M}_1 - \mathcal{M}_1$ denote the set

$\{x - x' : x, x' \in \mathcal{M}_1\}$, and let $\mathrm{dist}_2(h, \mathcal{M}_1 - \mathcal{M}_1) = \inf_{z \in \mathcal{M}_1 - \mathcal{M}_1} |||z - h|||_2$. Then the linear mapping A is said to satisfy the *robust null-space property* with constants $D_1, D_2 > 0$, if

$$|||h|||_1 \leqslant D_1 \mathrm{dist}_2(h, \mathcal{M}_1 - \mathcal{M}_1) + D_2 |||A(h)|||_3 \quad \text{for all } h \in \mathcal{X}. \tag{16}$$

We note that this condition implies that A is injective on $\mathcal{M}_1$. Indeed, suppose that $x, x' \in \mathcal{M}_1$ are such that $Ax = Ax'$ then applying (16) with $h \coloneqq x - x'$ leads to $|||x - x'|||_1 \leqslant D_1 \mathrm{dist}_2(x - x', \mathcal{M}_1 - \mathcal{M}_1) = 0$, which implies that $h = x - x' = 0$. Hence A is injective on $\mathcal{M}_1$.

Next, assume for simplicity that Assumption 4 holds, i.e., that $\mathcal{M}_1$ is compact. For linear inverse problems, this assumption is often used on the noise level, whereas the set $\mathcal{M}_1$ is often an unbounded set (sparse vector, matrices with low rank etc.). However, when modelling any practical application it is not unreasonable to assume that the set of interest is closed and bounded.

Now let $\mathrm{diam}_{d_3}(\mathcal{E}) = \eta$. We want to show that worst-case kernel size with metrics induced by $|||\cdot|||_1$ and $|||\cdot|||_3$ and under the assumptions above is bounded by

$$\mathrm{kersize}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) \leqslant D_2 \eta. \tag{17}$$

To this end let $x, x' \in \mathcal{M}_1$ and $e, e' \in \mathcal{E}$ be such that $Ax + e = Ax' + e'$. From (16) we then have that

$$\big|\big|\big|x - x'\big|\big|\big|_1 \leqslant D_1 \mathrm{dist}_{d_1}(x - x', \mathcal{M}_1 - \mathcal{M}_1) + D_2 \big|\big|\big|A(x - x')\big|\big|\big|_3 = D_2 \big|\big|\big|e - e'\big|\big|\big|_3 \leqslant D_2 \eta.$$

Taking supremum over the quantities above yields the bound in (17). Note in particular, that if we take $\eta = 0$, i.e., the noiseless model $\mathcal{E} = \{0\}$, we have that $\mathrm{kersize}^{\mathrm{w}}(F, \mathcal{M}_1, \mathcal{E}) = 0$.

Example 2: Optimal learning, Chebyschev centers and Gelfand widths. Recently the notion of optimal learning was introduced in [21]. Herein, one considers the setting in which the set $\mathcal{X}$ is a Banach space with norm $\|\cdot\|_{\mathcal{X}}$ and $\mathcal{M}_1$ is a compact subset of $\mathcal{X}$. To sample an element $x \in \mathcal{M}_1$, one uses m linear functionals $\lambda_1, \ldots, \lambda_m \in \mathcal{X}^*$ from the dual space of $\mathcal{X}$. Now, for a given $x \in \mathcal{M}_1$, let $y = (\lambda_1(x), \ldots, \lambda_m(x)) \in \mathbb{C}^m$ and let

$$K_y = \{x \in \mathcal{M}_1 : \lambda_i(x) = y_i, i \in \{1, \ldots, m\}\}$$

denote the set of solutions which are data-consistent. The above measurement model is noiseless, so define for convenience $\mathcal{M}_2 = \{y = (\lambda_1(x), \ldots, \lambda_m(x)) : x \in \mathcal{M}_1\}$ as the set of noiseless measurements.

Now, let $B_{\mathcal{X}}(z, r)$ denote the closed ball in $\mathcal{X}$ with center z and radius r. For a compact set $\mathcal{S} \subset \mathcal{X}$, the Chebyshev radius of $\mathcal{S}$ is given by

$$R_{\mathcal{X}}(\mathcal{S}) \coloneqq \inf\{r > 0 : \mathcal{S} \subset B_{\mathcal{X}}(z, r) \text{ for some } z \in \mathcal{X}\}.$$

In [21] the quantity $R_{\mathcal{X}}(K_y)$ is defined as the *optimal recovery rate* for a given $x \in \mathcal{M}_1$, with $y = (\lambda_1(x), \ldots, \lambda_m(y))$. This quantity is zero if K_y is a singleton, however, here we follow [21] and focus on the cases for which $R_{\mathcal{X}}(K_y) > 0$.

While the framework developed in this paper does not extend to general Banach spaces, it does cover the special cases of $\mathbb{C}^N$ with metric induced by a norm $\|\cdot\|_{\mathcal{X}}$. Indeed, by letting $A \in \mathbb{C}^{m \times N}$ be the matrix given by the linear functionals λ_i, taking $\mathcal{E} = \{0\}$, and using the linear additive model $F(x, e) = Ax + e$, it is straightforward to see that $R_X(K_y) = \frac{1}{2}\mathrm{diam}_X(F_y)$, i.e., if we take the supremum over all $y \in \mathcal{M}_2$ we recover the worst case kernel size. To further analyze the noisy setting, we refer to [21]

The framework proposed in this work is closely related to the concept of Chebychev centers. In fact, the optimal maps Ψ defined in (9) and (14) correspond to the problem of finding the Chebychev center or the p-center (see [79]) of the set F_y of candidate solutions. The setting proposed in this paper guarantees that such centres exist, so that the optimal maps are well-defined; however, Chebychev centers may not exist in general, and other sufficient and necessary conditions for their existence can be found in the literature (see [7], [6], [98], [84]).

In this work, we have kept the forward model F fixed. However, in the special case of a linear forward model, it is reasonable to also consider the question of how well we can perform given a fixed budget of m

linear functionals. Fortunately, this question is well studied in the literature [81], and is given by the Gelfand width of $\mathcal{M}_1$,

$$w^m(\mathcal{M}_1) \coloneqq \inf_{A\in\mathbb{C}^{m\times N}} \sup_{x\in\mathcal{M}_1} R_{\mathcal{X}}(K_{Ax}).$$

We note that this presents an lower bound for $c^{\mathrm{w}}_{\mathrm{opt}}(A, \mathcal{M}_1, \{0\})$ in the noiseless setting. We do not consider the question of what the optimal forward model would be any further in this work, but remark that it is an interesting question worth investigating.

Example 3: Bayesian Inverse Problems. Bayesian inverse problems [11,90] take a probabilisitc approach. Instead of using a measurement y to output a *single* candidate solution x, the goal of a Bayesian approach is to use a measurement y to output a *distribution* on the possible solutions x's.

Typically, a Bayesian inverse problem setting assumes that the measurements are realisations of a random variable $Y \sim \mu_{Y|X=x}$ distributed according to a measure conditional on the value of the true solution x; moreover, the unknown x itself is assumed to be the realisation of a random variable $X \sim \mu_X$ distributed according to some prior distribution μ_X on $\mathcal{X}$. The goal of Bayesian inverse problem is to use an observed measurement y to update the prior μ_X distribution into a posterior distribution $\mu_{X|Y=y}$. Tools from probability and statistics can then be implemented to analyse the posterior distribution, such as computing its mean, its mode(s), variance, quantiles, and other quantities of interest.

A typical setting for Bayesian inverse problems involves a model of the form $Y = \mathcal{G}(X) + E$, where the unknown X is a random variable, and $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ is the 'ideal' forward model which is corrupted by some random noise $E \sim \mu_E$ in $\mathcal{Y}$ (often assumed independent from X). For a fixed a realisation x of X, the randomness on Y is induced by the noise E: for example, when μ_E admits a probability density function f_E defined on $\mathcal{Y}$, then the conditional density of $Y|X = x$ can be espressed as $f_{Y|X=x} = f_E(\cdot - \mathcal{G}(x))$.

Our proposed framework encompasses the general case of Bayesian inverse problems. In fact, we interpret the measure μ on $\mathcal{X} \times \mathcal{Z}$ as being the joint law of a random variable (X, E) that models both the unkown x and the noise e, so that $\mu = \mu_{(X,E)}$. We consider the random variable $Y \coloneqq F(X, E)$, whose law is precisely $\mu_Y = F_*\mu$. While the Bayesian prior and posterior distributions are measures on $\mathcal{X}$, in our framework both μ and the disintegrations $\{\mu^y\}_{y\in\mathcal{Y}}$ are measures on the product $\mathcal{X} \times \mathcal{Z}$; it turns out that it is sufficient to take the pushforward along the projection $\pi_1 : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ to recover the familiar distributions of the Bayesian approach. Moreover, it also turns out that the optimal map in (14) coincides with the Bayesian estimator of the *posterior mean* in a typical setting. These claims are summarised in the following result.

Proposition 4.1. *In our setting as described above, it holds that:*

(1) the Bayesian prior distribution of X is $\mu_X \coloneqq \pi_{1*}\,\mu$;

(2) the Bayesian posteriors $\{\mu_{X|Y=y}\}_{y\in\mathcal{Y}}$ can be obtained via the disintegration as

$$\mu_{X|Y=y} = \pi_{1*}\,\mu^y.$$

(3) In the case where $(\mathcal{X}, d) = (\mathbb{R}^N, d_{\|\cdot\|_2})$ is an Euclidean space and $p = 2$, the posterior mean is an optimal map according to (12)*, as defined in* (14)*. Explicitly, for every $y \in \mathcal{Y}$ it holds that*

$$\Psi(y) = \mathbb{E}_{\mu_{X|Y=y}}[X].$$

Example 4: Worst- vs. average-case: a simple comparison. Consider the problem of recovering a point $x = (x_1, x_2) \in \mathbb{R}^2$ from its first coordinate x_1. Equip $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \mathbb{R}$ with the Euclidean metric, and assume a noiseless model $\mathcal{Z} = \mathcal{E} = \{0\}$. The forward map is explicitly $F((x_1, x_2), e) = x_1$. Consider a model class of two points aligned vertically and at distance 1 from each other, such as $\mathcal{M}_1 = \{(0, 0), (0, 1)\} \subseteq \mathbb{R}^2$. The space of all possible measurements is is $\mathcal{M}_2^{\mathcal{E}} = \{0\} \subseteq \mathbb{R}$.

(1) In the worst-case setting, we have

$$\mathrm{kersize}^w(F, \mathcal{M}_1, \mathcal{E}) = 1, \quad c_{opt}(F, \mathcal{M}_1, \mathcal{E}) = \frac{1}{2}$$

and the optimal map is given by the midpoint $\Psi(0) = (0, \frac{1}{2})$.

(2) Now equip the space $\mathcal{M}_1 \times \mathcal{E} \cong \mathcal{M}_1$ with the probability measure

$$\mu = \mu_\alpha := \alpha\delta_{(0,0)} + (1-\alpha)\delta_{(0,1)}$$

depending on the parameter $\alpha \in [0,1]$. The parameter α acts as a weight that associates increasingly importance to one of the two points, in this case $(0,0)$. In this case, we have for $p \in [0, +\infty)$

$$\text{kersize}^{\text{a}}(F, \mathcal{M}_1, \mathcal{E}, p)^p = 2\alpha(1-\alpha)$$

and for $p = \infty$ it is

$$\text{kersize}^{\text{a}}(F, \mathcal{M}_1, \mathcal{E}, \infty) = \begin{cases} 0 & \text{if } \alpha \in \{0,1\} \\ 1 & \text{if } \alpha \in (0,1). \end{cases} \tag{18}$$

For example, for $p = 2$ the optimal map is given by

$$\Psi(0) = (0, 1-\alpha).$$

So that the optimality constant for $p = 2$ is

$$c^a_{opt}(F, \mathcal{M}_1, \mathcal{E})^2 = \alpha(1-\alpha).$$

This can be interpreted in the following way: if both points in $\mathcal{M}_1$ are assumed to have the same importance, it is most appropriate to consider the *worst-case* optimality constant. Then, the optimum reconstruction is the mid point. However, if one point is more important than the other, then the *average* optimality constant is more pertinent. In this case, the optimum is not the mid-point, but a point closer to the more important point. The average case presented here is the simplest toy example for what happens in Deep Learning (DL): a learning procedure, that optimizes the Euclidian distance as in the example, tends to stay closer to the most important points.

5. Relation to previous work

As an ongoing topic of research in many areas of mathematics, undersampled inverse problems have been studied in many different areas. As a non-extensive list of examples, they have been studied from a theoretical standpoint [43] and using a statistical perspective [80], or by using iterative deep neural networks [4], regularization [40] and in more applied fields such as radio tomography of the ionosphere [52]. Different moise models have been studied, and while an assumption of additive noise appears ubiquitously [20,59,63], often multiplicative noise models can be of interest in applications [12,60,87,100]. To the best of the authors' knowledge fundamental accuracy bounds for undersampled inverse problems with multiplicative noise have not been produced, and the framework presented in the current work also encompasses this case as a special case. Moreover, despite this extensive amount of research, there is little to be found on fundamental accuracy bounds of approximate multi-valued solutions to undersampled inverse problems.

Accuracy bounds in approximation theory: Fundamental accuracy bounds, the Gelfand Widths, have been established by A. Pinkus [81], in the noiseless linear setting. A framework for obtaining the best k-term approximation and corresponding accuracy bounds is established in the work of A. Cohen, W. Dahmen and R. DeVore [35]. In [36] A. Cohen, R. DeVore, G. Petrova and P. Wojtaszczyk propose a framework to measure the optimal performance for nonlinear methods of approximation. The notion of optimal learning in a noisy setting and upper worst-case accuracy bounds based on the Chebycheff radius are presented by P. Binev et al. in [21]. B. Adcock et al. [3] propose a framework for accuracy bounds extending the classical assumption of Compressed Sensing (CS). Generalized instance optimality and the corresponding accuracy bounds have been presented by A. Bourrier et al. [26]. Related to the CS framework, the Restricted Isometry Property (RIP) is generalized by Y. Traonmilin and R. Gribonval in [93] and the null space property is

generalized by H. Tran and C. Webster in [92]. Related to our work, analogous worst- and average- case error bounds for normed spaces and single-valued decoders are established in [82].

Optimal (multi-valued) decoders for undersampled inverse problems: The study of optimal and near-optimal recovery in inverse problem has always been a central question, and many single-valued near-optimal mappings have been proposed and analyzed [21,26,51,82,93]. A survey of early works on optimal recovery can be found in [75]. The case of multi-valued optimal decoders has been considered for obtaining the worst case bounds on Banach spaces in [10] and extended to metric spaces in [73]. As presented in Example 3, the Bayesian approach to inverse problems aims at recovering a distribution valued decoder [11,90]. There exist some a posteriori accuracy bounds in the case of normed spaces [68,69,96].

(Multi-valued) Decoders arising in Deep Learning: In the recent work by I. Daubechies, R. DeVore, S. Foucart et. al., [41], the ability of deep neural networks to nonlinearly approximate functions, and thus also decoders, is investigated. A range of results on how the resulting decoder may be constituted is presented by M. Unser in [95]. From an application-based perspective, a detailed overview of deep learning in inverse problems and stability of robustness for deep learning is given by M. McCann et al. in [74] and the review [11, 72]. However, compared to standard methods, data-driven approaches using deep learning for solving inverse problems (1) have reported superior accuracy in different applications [17,89,101]. This can potentially lead to instabilities, which is also highlighted by V. Antun et al. in [9] and N. M. Gottschling et al. in [55]. In fact, there is a variety of research that has established that artificial intelligence techniques based on deep learning are unstable, firstly in image classification [45,66,76,78,91], and later in applications ranging from audio and speech recognition [32, 33, 99] to natural language processing [70] and automatic diagnosis in medicine [48]. Instabilities, such as false positives, false negatives, and especially AI hallucinations, have been an issue in the fastMRI challenge, [77] and also in microscopy [17,58]. To the best of our knowledge there do not exist fundamental performance and accuracy limits for data-driven approaches using deep learning for solving the inverse problems. Moreover, many DL based approaches implicity include multi-valued functions, where examples include, but are not limited to, deep ensembles [67] or model-based probabilistic conditional diffusion models as in [71]. In fact, any probabilistic DL model used to an inverse problem that uses sampling – see the work of J. Gawlikowski [53] for an introduction – can be considered to be a multi-valued decoder.

6. Conclusion

The proposed theoretical framework for undersampled inverse problems can be seen as a generalisation of a variety previous frameworks to multi-valued decoders. Additionally, commonly used assumptions, such as the convexity of the set $\mathcal{M}_1$ and a linear A and additive noise, as well as the condition $(\mathcal{M}_1 - \mathcal{M}_1) \cap \mathcal{N}(A) = \{0\}$, which is for example implied by the RIP, are extended or entirely obliterated. Under general assumptions, our work provides relevant accuracy bounds for possible multi-valued decoders and a variational expression for optimal decoders. Due to the generality of the assumptions, these bounds provide a mean to bridge the gap between theory and practice, as the lower bounds can be used for assessing accuracy and performance of a wide range of models, including ensemble models, diffusion models and sampling-based approaches that have multi-valued solutions.

7. Proofs

7.1. Existence and uniqueness of a disintegration of the measure μ given Assumptions 1, 3 and 4.

In this section we prove the following proposition, which ensures that our setting guarantees the existence of a disintegration of the measure μ.

Proposition 7.1. *Under Assumptions 1, 3 and 4, there exists a disintegration $\{\mu^y\}_{y\in\mathcal{M}_2^\varepsilon}$ of the measure μ along F. Moreover, such disintegration is essentially unique: if $\{\tilde{\mu}^y\}_{y\in\mathcal{M}_2^\varepsilon}$ is another family satisfying $(i)-(iii)$ in Definition 3.6, then $\tilde{\mu}^y = \mu^y$ for almost every y.*

The proof of Proposition 7.1 is a special case of what is found in [34]. Before jumping to the proof, in this section we start by recalling some well known definitions for measures, and state the relevant theorems from [34]. Let $(X, \mathcal{B}(X))$ be the Borel measurable space associated with the topological space X. A measure μ on $(X, \mathcal{B}(X))$ is said to be a *Radon measure* (sometimes also called a regular measure [5]) if

(i) $\mu(K) < +\infty$ for each compact $K \in \mathcal{B}(X)$,
(ii) $\mu(B) = \inf\{\mu(V) : V \in \mathcal{B}(X), V \text{ open, and } B \subset V\}$, for every $B \in \mathcal{B}(X)$ (μ is outer regular), and
(iii) $\mu(B) = \sup\{\mu(K) : K \in \mathcal{B}(X), K \text{ compact, and } K \subset B\}$, for every $B \in \mathcal{B}(X)$ (μ is tight).

Moreover, a measure μ is said to *dominate* a measure ν if $\mu(B) = 0$ implies $\nu(B) = 0$ for every $B \in \mathcal{B}(X)$. Suppose that X can be covered by at most countably many Borel measurable sets $\{B_i\}_{i\in I}$, $I \subset \mathbb{N}$, and that $\mu(B_i) < \infty$ for each $i \in I$. Then, μ is said to be a σ-finite measure. In particular, every finite measure is σ-finite. We also have the following results for finite measures, that are used in the proof of Proposition 7.1.

Theorem 7.2 ([5, Thm. 12.7])**.** *A finite measure on a Polish space (i.e. a complete separable metric space) is Radon.*

In Definition 3.6 one may replace the measure $F_*\mu$ with another measure ρ, in which case $\{\mu^y\}_y$ would be called a (F, ρ) disintegration of μ. We refer the reader to [34, Def. 1] for the detailed statement.

Theorem 7.3 ([34, Thm. 1])**.** *Let μ be a σ-finite Radon measure on a metric space X and let F be a measurable map from $(X, \mathcal{B}(X))$ to the measurable space (Y, Σ). Let ρ be a σ-finite measure on Σ that dominates the pushforward measure $F_*\mu$. If Σ is countably generated and contains all the singleton sets $\{y\}$, then μ has a (F, ρ)-disintegration. The μ^y measures are uniquely determined up to an almost sure equivalence: if $\{\mu^y_*\}$ is another (F, ρ)-disintegration then $\rho(\{y \in \mathcal{Y} : \mu^y_* = \mu^y\}) = 0$.*

Theorem 7.4 ([34, Thm. 2 (iii)])**.** *Let μ have a (F, ρ)-disintegration $\{\mu^y\}$, with μ and ρ each σ-finite. Then, the measures $\{\mu^y\}$ are probabilities for ρ-almost all $y \in \mathcal{Y}$ if and only if $\rho = F_*\mu$.*

Proof of Proposition 7.1. Let $d := \max\{d_{\mathcal{X}}, d_{\mathcal{Z}}\}$ be one of the standard metrics on the product $\mathcal{X} \times \mathcal{Z}$. The spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Z}, d_{\mathcal{Z}})$ are separable by Assumption 1, and complete by Assumption 1(ii), thus the product space $(\mathcal{X} \times \mathcal{Z}, d)$ is complete and separable, [88] (pg. 26, Invariance properties, Table 1). Thus, $(\mathcal{X} \times \mathcal{Z}, d)$ is a Polish space. Since μ is a finite measure on the Polish space $\mathcal{X} \times \mathcal{Z}$, Theorem 7.2 guarantees that μ is Radon. Moreover, since μ is finite, it is in particular σ-finite. Taking $\rho = F_*\mu$, clearly ρ dominates $F_*\mu$ trivially, as they are the same measure. By Assumption 1, $(\mathcal{M}_2^{\mathcal{E}}, d_{\mathcal{Y}})$ is second countable, hence its topology is countably generated, and so is the corresponding Borel σ-algebra $\mathcal{B} = \mathcal{B}(\mathcal{M}_2^{\mathcal{E}})$. Furthermore, since singletons $\{y\} \subseteq \mathcal{M}_2^{\mathcal{E}}$ are closed, the Borel σ-algebra contains all singletons. Lastly, the mapping F is measurable by Assumption 3. Thus, all the conditions of Theorem 7.3 are satisfied. Hence, there exists a disintegration of the measure μ along F. Such a disintegration is essentially unique in the sense of Theorem 7.3. Finally, since μ and $\rho = F_*\mu$ are both finite and in particular σ-finite, Theorem 7.4 guarantees that $\rho = F_*\mu$ implies that the measures μ^y are probability measures for $F_*\mu$-almost every $y \in \mathcal{M}_2^{\mathcal{E}}$. □

7.2. **Proof of Theorem 3.4.**

Proof of Theorem 3.4. We begin with the proof of (ii). First, let us introduce some notation: for fixed $y \in \mathcal{M}_2^{\mathcal{E}}$, define the function

$$f_y \colon \mathcal{X} \to [0, +\infty], \quad f_y(z) = \sup_{(x,e)\in F_y} d_{\mathcal{X}}(x, z) = d^H_{\mathcal{X}}(z, F_y)$$

and notice that f_y appears in the definition of Ψ in (9). That is, we have the relation,

$$\Psi(y) = \operatorname*{argmin}_{z\in\mathcal{X}} f_y(z). \tag{19}$$

Furthermore, for each $y \in \mathcal{M}_2^{\mathcal{E}}$ let

$$r_y = \sup_{\substack{(x,e)\in F_y \\ (x',e')\in F_y}} d_{\mathcal{X}}(x,x') = \operatorname{diam}_{d_{\mathcal{X}}}(F_y).$$

Next, we make a claim which, once it is established, we use to prove that Ψ has non-empty, compact values. This ensures that Ψ is well-defined.

Claim. For every $y \in \mathcal{M}_2^{\mathcal{E}}$, we have that,

(I) f_y is continuous,

(II) for any $x \in F_y$, we have that

$$\operatorname*{argmin}_{z\in\mathcal{X}} f_y(z) = \operatorname*{argmin}_{z\in B(x,r_y)} f_y(z),$$

(III) $\Psi(y)$ is closed.

We proceed to prove the claim and start by considering (I). We consider a general setting, and let (X,d) be a metric space, $A \subseteq X$ be a bounded subset and $g(x) \coloneqq d^H(x,A) = \sup_{a\in A} d(x,a)$. We claim that $|g(x)-g(y)| \leqslant d(x,y)$ for all $x,y \in X$. To see this, consider $a,x,y \in X$ and note that

$$d(x,a) \leqslant d(x,y) + d(y,a) \quad \Longrightarrow \quad \sup_{a\in A} d(x,a) \leqslant d(x,y) + \sup_{a\in A} d(y,a).$$

Switching the roles of x and y, leads to the desired inequality. It follows that g is continuous. Moreover, F_y is bounded by Assumption 2. Thus, letting $(X,d) = (\mathcal{X}, d_{\mathcal{X}})$, $A = F_y$, $g = f_y$ above, proves (I).

To prove (II), we show that no point outside of the closed ball $B_{d_{\mathcal{X}}}(x, r_y)$ can be a minimiser of f_y. First, note that $r_y = \operatorname{diam}_{d_{\mathcal{X}}}(\pi_1(F_y)) < \infty$ as $\pi_1(F_y)$ is bounded. Moreover, by definition, we have that

$$r_y = \sup_{(x,e)\in F_y} \sup_{(x',e')\in F_y} d_{\mathcal{X}}(x,x') = \sup_{(x,e)\in F_y} f_y(x),$$

which implies that $r_y \geqslant f_y(x)$ for all $x \in F_y$. Now, pick an $\hat{x} \in F_y$. If $z \in \mathcal{X}\backslash B(\hat{x}, r_y)$ is a point outside of the ball, then $d_{\mathcal{X}}(z,\hat{x}) > r_y$, and

$$f_y(z) = \sup_{(x'e')\in F_y} d_{\mathcal{X}}(x',z) \geqslant d_{\mathcal{X}}(\hat{x},z) > r_y \geqslant f_y(\hat{x}).$$

It follows that any minimizer of f_y must lie in the ball $B(\hat{x}, r_y)$. This proves our claim in (II).

Part (III) of the claim follows directly from the continuity of f_y. Indeed, let $\alpha_y \coloneqq \min f_y \in [0, +\infty)$. Then, since $\{\alpha_y\} \subset [0,\infty)$ is closed and f_y is continuous, the preimage $f_y^{-1}(\{\alpha_y\}) = \Psi(y)$ is closed. This proves (III), and concludes our proof of the claim.

Next, we prove the following properties of Ψ. These properties finalize the proof of statement (ii) of the theorem.

(a) Ψ *has non-empty and compact values*;

(b) Ψ *is an optimal map*;

We start with the proof of (a). Let $y \in \mathcal{M}_2^{\mathcal{E}}$ and consider $x \inf F_y$. From (II) and (19), we see that

$$\Psi(y) = \operatorname*{argmin}_{z\in\mathcal{X}} f_y(z) = \operatorname*{argmin}_{z\in B(x,r_y)} f_y(z). \tag{20}$$

By Assumption 1(ii), $d_{\mathcal{X}}$ satisfies the Heine-Borel property, which implies that the closed and bounded ball $\mathcal{B}(x,r_y)$ is compact. Moreover, from (I) we know that the objective function f_y is continuous. Thus, it follows from the Extreme Value Theorem that f_y attains it minimum in (20). This implies that Ψ has non-empty values. To see that Ψ has compact values, observe that $\Psi(y)$ is closed by (III) and that $\Psi(y) \subseteq B(x,r_y)$, where $B(x,r_y)$ is compact. Since a closed subset of a compact set is compact, we conclude that $\Psi(y)$ is compact.

Next, we prove (b). We will show that the minimum worst-case reconstruction error of Ψ equals the optimality constant as given in Definition 3.1. To this end, let $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ and fix $y \in \mathcal{M}_2^{\mathcal{E}}$. By construction we have that

$$\sup_{(x,e)\in F_y} d_{\mathcal{X}}^H(\Psi(y),x) \leqslant \sup_{(x,e)\in F_y} d_{\mathcal{X}}(x,z), \quad \text{for all } z \in \varphi(y).$$

It follows that

$$\sup_{(x,e)\in F_y} d_{\mathcal{X}}^H(\Psi(y),x) \leqslant \sup_{(x,e)\in F_y} \sup_{z\in\varphi(y)} d_{\mathcal{X}}(x,z) = \sup_{(x,e)\in F_y} d_{\mathcal{X}}^H(x,\varphi(y)).$$

By taking the supremum with respect to $y \in \mathcal{M}_2^{\mathcal{E}}$ on both sides, we obtain

$$\sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \sup_{(x,e)\in F_y} d_1^H(\Psi(y)),x) \leqslant \sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \sup_{(x,e)\in F_y} d_1^H(\varphi(y),x).$$

As $F\colon \mathcal{M}_1 \times \mathcal{E} \to \mathcal{M}_2^{\mathcal{E}}$ is surjective, the previous inequality can be rewritten as

$$\sup_{(x,e)\in\mathcal{M}_1\times\mathcal{E}} d_1^H(\Psi(F(x,e))),x) \leqslant \sup_{(x,e)\in\mathcal{M}_1\times\mathcal{E}} d_1^H(\varphi(F(x,e))),x).$$

Now, since $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ was arbitrary, the above inequality holds for any $\varphi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$. Taking the infimum over all mappings on the right hand, side we obtain the optimality constant:

$$\sup_{(x,e)\in\mathcal{M}_1\times\mathcal{E}} d_1^H(\Psi(F(x,e)),x) \leqslant c_{\mathrm{opt}}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}).$$

The opposite inequality is trivial, since $\Psi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ is one of the reconstruction mappings over which the infimum is taken. Therefore,

$$\sup_{(x,e)\in\mathcal{M}_1\times\mathcal{E}} d_1^H(\Psi(F(x,e)),x) = c_{\mathrm{opt}}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}),$$

and we conclude that Ψ is an optimal map. This concludes the proof of statement (ii) in the theorem.

We proceed with the proof of (i) and start with the lower bound in (8). Let $\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ and $y \in \mathcal{M}_2^{\mathcal{E}}$. Then

$$\mathrm{diam}_{d_{\mathcal{X}}}(F_y) = \sup_{x,x'\in F_y} d_{\mathcal{X}}(x,x') \leqslant \sup_{x\in F_y} 2d_{\mathcal{X}}^H(x,\varphi(y)).$$

Now, taking the supremum over all $y \in \mathcal{M}_2^{\mathcal{E}}$ gives the inequality

$$\mathrm{kersize}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}) \leqslant \sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \sup_{x\in F_y} 2d_{\mathcal{X}}^H(x,\varphi(y)). \tag{21}$$

Finally, since φ was arbitrary, taking the infimum over all $\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ in (21) gives

$$\mathrm{kersize}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}) \leqslant 2c_{\mathrm{opt}}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}).$$

which proves the lower bound in (8).

For proving the upper bound in (8), we will make use of the mapping Ψ in (9), which we know from statement (ii) is an optimal map. Let $y \in \mathcal{M}_2^{\mathcal{E}}$ and $x' \in F_y$. Then by construction of Ψ, we have that

$$\sup_{x\in F_y} d_{\mathcal{X}}^H(\Psi(y),x) \leqslant \sup_{x\in F_y} d_{\mathcal{X}}(x,x') \leqslant \mathrm{diam}(F_y).$$

Taking the supremum over all $y \in \mathcal{M}_2^{\mathcal{E}}$ on both sides above, gives

$$\sup_{y\in\mathcal{M}_2^{\mathcal{E}}} \sup_{x\in F_y} d_{\mathcal{X}}^H(\Psi(y),x) = c_{\mathrm{opt}}(F,\mathcal{M}_1,\mathcal{E}) \leqslant \mathrm{kersize}^{\mathrm{w}}(F,\mathcal{M}_1,\mathcal{E}),$$

where we used the fact that Ψ is an optimal map in the first equality. This establishes statement (i).

□

7.3. **Preliminaries from measure theory.** In order to prove Theorem 3.9 we need to establish some results from measure theory.

7.3.1. *Carathéodory functions.* Let $(S, \Sigma_S), (X, \Sigma_X)$ and (Y, Σ_Y) be measurable spaces. For functions $f\colon S \times X \to Y$, whose domain is a product of two measurable spaces, several notions of measurability exist. The function f could be *jointly measurable*, i.e., measurable with respect to the product σ-algebra $\Sigma_S \otimes \Sigma_X$. For fixed $s \in S$ or $x \in X$, the functions $f^s = f(s, \cdot)\colon X \to Y$ and $f^x = f(\cdot, x)\colon S \to Y$, could be measurable with respect to Σ_X and Σ_S, respectively. A function f which is measurable in one variable for each fixed $s \in S$ and for each fixed $x \in X$ is said to be *separately measurable*. In general, joint measurability implies separate measurability, but the converse is not true [5, p. 152].

A class of functions which are jointly measurable in many important cases are Carathéodory functions.

Definition 7.5 ([5, Def. 4.50])**.** Let (S, Σ) be a measurable space, and let X and Y be topological spaces. A function $f\colon S \times X \to Y$ is a *Carathéodory function* if:

(i) for each $x \in X$, the function $f^x = f(\cdot, x)\colon S \to Y$ is $(\Sigma, \mathcal{B}(Y))$-measurable, and
(ii) for each $s \in S$, the function $f^s = f(s, \cdot)\colon X \to Y$ is continuous.

In particular, if X is a separable metric space and Y is a metric space, then every Carathéodory function $f\colon S \times X \to Y$ is jointly measurable [5, Lem. 4.51]. It is also straightforward to see that if f is continuous in both arguments separately, then f is a Carathéodory function.

7.3.2. *Measurability of multi-valued functions.* For a single-valued function $f\colon Y \to X$ the inverse image of a set $A \subset X$ is $f^{-1}(A) \coloneqq \{y \in Y\colon f(y) \in A\}$. For a multi-valued function $\varphi\colon Y \rightrightarrows X$ there are different ways to generalize the concept of an inverse image: Given a set $A \subset X$, we say that the *upper inverse* of A is $\varphi^{\mathrm{u}}(A) \coloneqq \{y \in Y : \varphi(y) \subset A\}$ and we say that the *lower inverse* of A is

$$\varphi^{\ell}(A) \coloneqq \{y \in Y : \varphi(y) \cap A \neq \varnothing\}.$$

Naturally, the notion of continuity of multi-valued mappings between topological spaces depends on the definition of inverse. Interested readers are referred to [5, Ch. 17]. In this work we shall be most concerned with the lower inverse of a multi-valued mapping, since this notion of inverse allows us to pick a measurable selector. Next, we define two different notions of measurability for multi-valued mappings based on the concept of lower inverse.

Definition 7.6 ([5, Def. 18.1])**.** Let (Y, Σ) be a measurable space and X a topological space. We say that a multi-valued mapping $\varphi : Y \rightrightarrows X$ is:

- *weakly measurable*, if $\varphi^{\ell}(V) \in \Sigma$ for each open subset V of X,
- *measurable*, if $\varphi^{\ell}(A) \in \Sigma$ for each closed subset A of X.

Note that if the topology of X above is induced by a metric, then every multi-valued measurable mapping is also weakly measurable [5, Lem. 18.2]. Our main tool for proving Theorem 3.9 is the following result.

Theorem 7.7 (Measurable Maximum Theorem, [5, Thm. 18.19])**.** *Let X be a separable metrisable space and (S, Σ) a measurable space. Let $\varphi\colon S \rightrightarrows X$ be weakly measurable with non-empty compact values, and suppose $f\colon S \times X \to \mathbb{R}$ is a Carathéodory function. Define the value function $m\colon S \to \mathbb{R}$ by*

$$m(s) = \max_{x \in \varphi(s)} f(s, x),$$

and the correspondence $\Phi\colon S \rightrightarrows X$ of maximisers by

$$\Phi(s) = \{x \in \varphi(s) \ : \ f(s, x) = m(s)\} = \operatorname*{argmax}_{x \in \varphi(s)} f(s, x).$$

Then m is measurable, Φ has non-empty and compact values, and Φ is measurable and admits a measurable selector.

The Measurable Maximum Theorem will be applied in two ways. The first can be found in Proposition 7.8, which gives sufficient conditions for a map $\phi\colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ to be in the set $\mathcal{C}$ given by (10). The second use of the theorem above is found in the proof of part (ii) and (iii) of Theorem 3.9.

Proposition 7.8. *If $\phi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ is measurable and has non-empty compact values, then $\phi \in \mathcal{C}$.*

Proof. By definition of $\mathcal{C}$, we need to prove that the function

$$r_\phi : \mathcal{M}_1 \times \mathcal{E} \to \mathbb{R}, \quad r_\phi(x,e) = d^H(x, \phi(F(x,e))) = \sup_{z \in \phi(F(x,e))} d_{\mathcal{X}}(x,z)$$

is measurable. Denote $S := \mathcal{M}_1 \times \mathcal{E}$, $X := \mathcal{X}$, and

$$\begin{aligned} &\varphi := \phi \circ F : \mathcal{M}_1 \times \mathcal{E} \rightrightarrows \mathcal{X}, && \varphi(x,e) = \phi(F(x,e)), \\ &f := d_{\mathcal{X}} \circ (\pi_1, \mathrm{id}_{\mathcal{X}}) : (\mathcal{M}_1 \times \mathcal{E}) \times \mathcal{X} \to \mathbb{R}, && f((x,e),z) = d_{\mathcal{X}}(x,z). \end{aligned}$$

The function φ is measurable, since it is the composition of the measurable functions F and ϕ. Now, since $\mathcal{X}$ is a metric space, φ is also weakly measurable [5, Lem. 18.2]. Moreover, φ has non-empty compact values because ϕ has non-empty compact values by assumption. Finally, the function f is continuous in both arguments and hence Carathéodory. Thus, the assumptions in Theorem 7.7 are satisfied, and its application implies that the value function $m : \mathcal{M}_1 \times \mathcal{E} \to \mathbb{R}$

$$m(x,e) = \max_{z \in \phi(F(x,e))} d_{\mathcal{X}}(x,z)$$

is measurable. It is immediate to notice that $m = r_\phi$, so we conclude that r_ϕ is measurable. Hence we have proven that $\phi \in \mathcal{C}$. □

7.3.3. *Disintegration of measures.* Next we prove a useful proposition that will be applied in the proof of Theorem 3.9. For completeness, we start by recalling that every time we write an expression of the form

$$\operatorname*{essup}_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} f(x,e), \qquad \operatorname*{essup}_{(x,e) \in F_y} f(x,e), \qquad \operatorname*{essup}_{y \in \mathcal{M}_2^{\mathcal{E}}} f(y).$$

It is implicit that we are taking the essential supremum with respect to μ on $\mathcal{M}_1 \times \mathcal{E}$, with respect to μ^y on F_y, and with respect to $F_*\mu$ on $\mathcal{M}_2^{\mathcal{E}}$.

We also recall that $\pi_1\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ refers to the projection on the first component $\pi_1(x,e) = x$.

Proposition 7.9. *Given Assumptions 1, 3 and 4, we have the following:*

(i) For $A \in \mathcal{B}(\mathcal{M}_1 \times \mathcal{E})$, $\mu(A) = 0$ if and only if $\mu^y(A) = 0$ for almost every $y \in \mathcal{M}_2^{\mathcal{E}}$.

(ii) Let $f : \mathcal{M}_1 \times \mathcal{E} \to [0, +\infty)$ be Borel measurable. Then the function

$$m\colon \mathcal{M}_2^{\mathcal{E}} \to \mathbb{R}, \quad m(y) = \operatorname*{essup}_{(x,e) \in F_y} f(x,e)$$

is Borel measurable.

(iii) Let $f : \mathcal{M}_1 \times \mathcal{E} \to [0, +\infty)$ be Borel measurable. Then

$$\operatorname*{essup}_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} f(x,e) = \operatorname*{essup}_{y \in \mathcal{M}_2^{\mathcal{E}}} \operatorname*{essup}_{(x,e) \in F_y} f(x,e).$$

(iv) Let $g : \mathcal{M}_1 \to [0, +\infty)$ be Borel measurable. Then

$$\int_{F_y} g(\pi_1(x,e)) d\mu^y(x,e) = \int_{F_y} g(x) d\big((\pi_1)_* \mu^y\big)(x),$$

and

$$\operatorname*{essup}_{(x,e) \in F_y} g(\pi_1(x,e)) = \operatorname*{essup}_{x \in F_y} g(x).$$

Proof. We start with (i). Let $\mathbb{1}_A$ denote the indicator function on $A \in \mathcal{B}(\mathcal{M}_1 \times \mathcal{E})$. Suppose that $\mu(A) = 0$, then by (iii) in Definition 3.6, we have that

$$0 = \mu(A) = \int_{\mathcal{M}_1 \times \mathcal{E}} \mathbb{1}_A \, d\mu = \int_{\mathcal{M}_2^{\mathcal{E}}} \left(\int_{F_y} \mathbb{1}_A \, d\mu^y \right) d(F_* \mu)(y).$$

The integral of a positive function with respect to a positive measure is zero if and only if the integrand function is zero almost-everywhere, so the previous equality is equivalent to

$$\int_{F_y} \mathbb{1}_A \, d\mu^y = \mu^y(A) = 0 \quad \text{for almost every } y \in \mathcal{M}_2^{\mathcal{E}}.$$

The reverse argument can be made with the same steps.

Let us now prove (ii). Since the Borel σ-algebra on $\mathbb{R}$ is generated by sets of form $(a, +\infty)$, we proceed to show that $m^{-1}((a, +\infty))$ is Borel measurable for any $a \in \mathbb{R}$. We have

$$\begin{aligned} m^{-1}((a, +\infty)) &= \{y \in \mathcal{M}_2^{\mathcal{E}} : m(y) > a\} \\ &= \{y \in \mathcal{M}_2^{\mathcal{E}} : \mu^y(\{(x, e) \in \mathcal{M}_1 \times \mathcal{E} : f(x, e) > a\}) > 0\} \\ &= \{y \in \mathcal{M}_2^{\mathcal{E}} : \int_{\mathcal{M}_1 \times \mathcal{E}} \mathbb{1}_{f^{-1}((a,\infty))} \, d\mu^y > 0\} \\ &= \{y \in \mathcal{M}_2^{\mathcal{E}} : h_a(y) > 0\} = h_a^{-1}((0, +\infty)), \end{aligned}$$

where $h_a(y) = \int_{\mathcal{M}_1 \times \mathcal{E}} \mathbb{1}_{f^{-1}((a,\infty))} \, d\mu^y$. In particular, since f is Borel measurable, $f^{-1}((a, \infty))$ is a measurable set and hence $\mathbb{1}_{f^{-1}((a,\infty))}$ is a measurable function on $\mathcal{M}_1 \times \mathcal{E}$. Now, from (ii) in Definition 3.6 we know that the function $y \mapsto h_a(y) = \int_{\mathcal{M}_1 \times \mathcal{E}} \mathbb{1}_{g^{-1}((a,\infty))} \, d\mu^y$ is measurable and so $h_a^{-1}((0, +\infty)) = m^{-1}((a, +\infty))$ is a measurable subset of Y. As $a \in \mathbb{R}$ was arbitrary, this proves that m is Borel-measurable. This concludes (ii).

Next we consider (iii), and let m be as in (ii). We start by showing that

$$\operatorname*{essup}_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} f(x, e) \geqslant \operatorname*{essup}_{y \in \mathcal{M}_2^{\mathcal{E}}} m(y). \tag{22}$$

Let $K = \operatorname{essup}_{\mathcal{M}_1 \times \mathcal{E}} f$. By definition we have $\mu(\{(x, e) \in \mathcal{M}_1 \times \mathcal{E} : f(x, e) > K\}) = 0$. Hence, by (i) and the fact that μ^y is concentrated on F_y, we have

$$0 = \mu^y(\{(x, e) \in \mathcal{M}_1 \times \mathcal{E} : f(x, e) > K\}) = \mu^y(\{(x, e) \in F_y : f(x, e) > K\})$$

for almost every $y \in \mathcal{M}_2^{\mathcal{E}}$. So $m(y) = \operatorname{essup}_{(x,e) \in F_y} f(x, e) \leqslant K = \operatorname{essup}_{\mathcal{M}_1 \times \mathcal{E}} f$ for almost every $y \in \mathcal{M}_2^{\mathcal{E}}$. This proves (22).

Next, we consider the reverse inequality. By definition of $\operatorname{essup}_{\mathcal{M}_2^{\mathcal{E}}} m$, we have

$$m(y) = \operatorname*{essup}_{(x,e) \in F_y} f(x, e) \leqslant \operatorname*{essup}_{\mathcal{M}_2^{\mathcal{E}}} m$$

for almost every $y \in \mathcal{M}_2^{\mathcal{E}}$. Hence,

$$\mu^y(\{(x, e) \in F_y : f(x, e) > \operatorname*{essup}_{\mathcal{M}_2^{\mathcal{E}}} m\}) = \mu^y(\{(x, e) \in \mathcal{M}_1 \times \mathcal{E} : f(x, e) > \operatorname*{essup}_{\mathcal{M}_2^{\mathcal{E}}} m\}) = 0$$

for almost every $y \in \mathcal{M}_2^{\mathcal{E}}$. By (i), this is equivalent to $\mu\{(x, e) \in \mathcal{M}_1 \times \mathcal{E} : f(x, e) > \operatorname{essup}_{\mathcal{M}_2^{\mathcal{E}}} m\} = 0$, which implies $\operatorname{essup}_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} f(x, e) \leqslant \operatorname{essup}_{y \in \mathcal{M}_2^{\mathcal{E}}} m(y)$, as desired. This concludes the proof of (iii).

Finally, let us prove (iv). Equality between integrals

$$\int_{F_y} (g \circ \pi_1) \, d\mu^y = \int_{F_y} g d(\pi_1)_* \mu^y$$

comes directly from the definition of pushforward measure, and can be found in detail [5, Theorem 13.46]. To see the equality between essential suprema, take $M \in [0, +\infty)$, then note that

$$((\pi_1)_* \mu^y)\big(g^{-1}(M, +\infty)\big) = \mu^y(\pi_1^{-1}(g^{-1}(M, +\infty))) = \mu^y((g \circ \pi_1)^{-1}(M, +\infty)).$$

Taking $M = \operatorname{essup}_{F_y}(g\circ\pi_1)$, then the definition of essential supremum gives $\mu^y((g\circ\pi_1)^{-1}(M,+\infty)) = 0$, which by the previous chain of equalities implies that $0 = ((\pi_1)_*\mu^y)(\{x \in F_y : g(x) > \operatorname{essup}_{F_y}(g\circ\pi_1)\})$, so that $g(x) \leqslant \operatorname{essup}_{F_y}(g\circ\pi_1)$ for $(\pi_1)_*\mu^y$-almost every $x \in F_y$. This gives $\operatorname{essup}_{x\in F_y} g(x) \leqslant \operatorname{essup}_{F_y}(g\circ\pi_1)$, By taking instead $M = \operatorname{essup}_{x\in F_y} g(x)$, the same reasoning proves the reverse inequality between essential suprema. Hence equality is proven, concluding the proof of (iv). □

7.4. **Proof of Theorem 3.9.** Having established the above preliminares, we now proceed to proving Theorem 3.9. In order to do so, we will split the Theorem up into two results, Proposition 7.10 and Proposition 7.11 and prove these. These results combined provide the proof of Theorem 3.9.

Proposition 7.10 (Lower bound of part (i), Theorem 3.9)**.** *Given the assumptions in Theorem 3.9, then the following holds for every $p \in [1,\infty]$:*

$$\operatorname{kersize}^a(F,\mathcal{M}_1,\mathcal{E},p) \leqslant 2c^a_{\mathrm{opt}}(F,\mathcal{M}_1,\mathcal{E},p).$$

Proof of Proposition 7.10. Let $\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ be an arbitrary reconstruction mapping. Fix $y \in \mathcal{M}_2^{\mathcal{E}}$ and consider $(x,e),(x',e') \in \mathcal{M}_1\times\mathcal{E}$ such that $F(x,e) = F(x',e') = y$. Then, by the triangle inequality, we deduce

$$d_{\mathcal{X}}(x,x') \leqslant d^H_{\mathcal{X}}(x,\varphi(F(x,e))) + d^H_{\mathcal{X}}(x',\varphi(F(x',e'))). \tag{23}$$

Let us now distinguish the two cases where $p=\infty$ and $p\in[1,\infty)$. The structure of the proof will be very similar in the two cases, with the main difference that the case $p=\infty$ involves essential suprema, while the case $p\in[1,+\infty)$ involves integrals. We we will fully prove the case $p\in[1,\infty)$ and provide a sketch of the proof of the case $p=\infty$, as it is virtually identical. In both cases, however, equation (23) will play a crucial role. To ensure that the essential suprema and integrals are well-defined we assume that $\varphi\in\mathcal{C}$.

Case $p\in[1,\infty)$; integrating (23) twice with respect to μ^y and using that $(a+b)^p \leqslant 2^p(a^p+b^p)$ for $a,b\geqslant 0$, we obtain

$$\begin{aligned}
&\int_{F_y}\int_{F_y} d_{\mathcal{X}}(x,x')^p\, d\mu^y(x,e)\, d\mu^y(x',e')\\
\leqslant&\int_{F_y}\int_{F_y}\left(d^H_{\mathcal{X}}(x,\varphi(F(x,e))) + d^H_{\mathcal{X}}(x',\varphi(F(x',e')))\right)^p d\mu^y(x,e)\, d\mu^y(x',e')\\
=&2^p\int_{F_y} d^H_{\mathcal{X}}(x,\varphi(F(x,e)))^p\, d\mu^y(x,e)
\end{aligned}$$

where in the last step we also used the fact that μ^y is a probability measure. Now, integrating the above inequality with respect to $F_*\mu$ on $\mathcal{M}_2^{\mathcal{E}}$ and raising to the power $\frac{1}{p}$ gives that

$$\begin{aligned}
\operatorname{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p) &= \left(\int_{\mathcal{M}_2^{\mathcal{E}}}\int_{F_y}\int_{F_y} d_{\mathcal{X}}(x,x')^p\, d\mu^y(x,e)\, d\mu^y(x',e')\, d(F_*\mu)(y)\right)^{\frac{1}{p}} \leqslant\\
&\leqslant \left(2^p\int_{\mathcal{M}_2^{\mathcal{E}}}\int_{F_y} d^H_{\mathcal{X}}(x,\varphi(F(x,e)))^p\ d\mu^y(x,e)\, d(F_*\mu)(y)\right)^{\frac{1}{p}}\\
&= 2\left(\int_{\mathcal{M}_1\times\mathcal{E}} d^H_{\mathcal{X}}(x,\varphi(F(x,e)))^p\, d\mu(x,e)\right)^{\frac{1}{p}}.
\end{aligned}$$

Since $\varphi\in\mathcal{C}$ was arbitrary, by taking the infimum over $\varphi\in\mathcal{C}$ we obtain:

$$\begin{aligned}
\operatorname{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p) &\leqslant 2\inf_{\varphi\in\mathcal{C}}\left(\int_{\mathcal{M}_1\times\mathcal{E}} d^H_{\mathcal{X}}(x,\varphi(F(x,e)))^p\, d\mu(x,e)\right)^{\frac{1}{p}}\\
&= 2c^{\mathrm{a}}_{\mathrm{opt}}(F,\mathcal{M}_1,\mathcal{E},p).
\end{aligned}$$

The proposition is therefore also proven in the case $p\in[1,+\infty)$.

Case $p = \infty$; the proof follows the basic structure of the case $p \in [1, \infty)$. In (23) instead of integrating we take the essential supremum respect to the measure μ^y. Then in the next step, instead of integrating, we take essential supremum in y with respect to the measure $F_*\mu$ on $\mathcal{M}_2^{\mathcal{E}}$ and apply Proposition 7.9. Finally, as $\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ was arbitrary, taking the infimum over $\varphi \in \mathcal{C}$ concludes the proof.

□

Proposition 7.11 (Part (ii) and (ii) and upper bound in (i) of Theorem 3.9)**.** *The following holds for every $p \in [1, \infty]$:*

$$c^a_{\mathrm{opt}}(F, \mathcal{M}_1, \mathcal{E}, p) \leqslant \mathrm{kersize}^a(F, \mathcal{M}_1, \mathcal{E}, p),$$

and the map $\Psi \colon \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ given by,

$$\Psi(y) = \operatorname*{argmin}_{z \in \mathcal{X}} \operatorname*{essup}_{(x,e) \in F_y} d_{\mathcal{X}}(x, z) \qquad (p = \infty) \tag{24}$$

$$\Psi(y) = \operatorname*{argmin}_{z \in \mathcal{X}} \int_{F_y} d_{\mathcal{X}}(x, z)^p \, d\mu^y(x, e) \qquad (p \in [1, \infty)) \tag{25}$$

is an optimal map with average error of order p. Moreover, Ψ has non-empty compact values, is measurable and it admits a measurable selector.

Proof of Proposition 7.11. We distinguish the cases $p = \infty$ and $p \in [1, \infty)$. In both cases, the structure of the proof consists in proving the following steps:

(a) Ψ*, defined either by* (24) *or* (25)*, has non-empty values;*
(b) Ψ *is measurable, has compact values and admits a measurable selector;*
(c) $\Psi \in \mathcal{C} = \{\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X} : (x, e) \mapsto d^H_{\mathcal{X}}(x, \varphi(F(x, e)) \text{ is measurable}\}$;
(d) Ψ *is an optimal map with average error of order p*;
(e) *Upper bound:* $c^a_{\mathrm{opt}}(F, \mathcal{M}_1, \mathcal{E}, p) \leqslant \mathrm{kersize}^a(F, \mathcal{M}_1, \mathcal{E}, p)$ *with $p = \infty$ or $p \in [1, +\infty)$.*

First we consider the case $p \in [1, \infty)$. Again, let us first introduce some notation, similar to above. Fix $y \in \mathcal{M}_2^{\mathcal{E}}$. Define $f_y : \mathcal{X} \to [0, \infty)$,

$$f_y(z) = \int_{F_y} d_{\mathcal{X}}(x, z)^p \, d\mu^y(x, e)$$

for $z \in \mathcal{X}$. Define also

$$r_y = \operatorname*{essup}_{\substack{(x,e) \in F_y \\ (x',e') \in F_y}} d_{\mathcal{X}}(x, x'),$$

$$E_{x,e} = \{(x', e') \in F_y : d_{\mathcal{X}}(x, x') > r_y\},$$

$$G_y = \{(x, e) \in F_y : \mu^y(E_{x,e}) = 0\}.$$

Claim. We claim that the following holds: for every $y \in \mathcal{M}_2^{\mathcal{E}}$

(I) f_y is continuous,
(II) $\operatorname*{argmin}_{\mathcal{X}} f_y = \operatorname*{argmin}_{B(x, 2r_y)} f_y$ for μ^y-almost every $(x, e) \in F_y$.

We proceed to prove the claim and start by considering (I). We consider a general setting, and let (X, d) be a metric space equipped with a probability measure ν concentrated on a bounded subset $A \subseteq X$, and define

$$g(x) \coloneqq \Big(\int_A d(x, a)^p \, d\nu(a) \Big)^{\frac{1}{p}} = \|d(x, \cdot)\|_{L^p(X,\nu)}.$$

We claim that $|g(x) - g(z)| \leqslant d(x, z)$ for all $x, y \in X$. To see this, consider $a, x, z \in X$ and note that by the triangle inequality $d(x, a) \leqslant d(x, z) + d(z, a)$. Taking $L^p(X, \nu)$-norms in the variable a and applying Minkowski's inequality gives

$$g(x) = \|d(x, \cdot)\|_p \leqslant \|d(x, z) + d(z, \cdot)\|_p \leqslant \|d(x, z)\|_p + \|d(z, \cdot)\|_p = d(x, z) + g(z),$$

where in the last passage we used that ν is a probability measure. Switching the roles of x and z, leads to the desired inequality. It follows that g is continuous.

By assumption 4, since $\mathcal{M}_1$ is compact and $F_y \subseteq \mathcal{M}_1$, then F_y is bounded. Thus, letting $(X, d) = (\mathcal{X}, d_{\mathcal{X}})$, $\nu = (\pi_1)_* \mu^y$, $A = F_y$ above, and recalling Proposition 7.9, (iv), then $g^p = f_y$, which proves (I).

To prove (II) we will show that for μ^y-almost every $(x, e) \in F^{-1}y$ we have

$$\Psi(y) = \operatorname*{argmin}_{z \in \mathcal{X}} \int_{F_y} d_{\mathcal{X}}(x', z)^p \, d\mu^y(x', e') = \operatorname*{argmin}_{z \in B_{d_{\mathcal{X}}}(x, 2r_y)} f_y(z). \tag{26}$$

Fix $(x, e) \in G_y$. If $z \in \mathcal{X} \backslash B_{d_{\mathcal{X}}}(x, 2r_y)$, then for μ^y-almost every $(x', e') \in F_y$,

$$d_{\mathcal{X}}(z, x') \geqslant d_{\mathcal{X}}(z, x) - d_{\mathcal{X}}(x, x') > 2r_y - r_y = r_y.$$

Thus,

$$\begin{aligned} f_y(z) &= \int_{F_y} d_{\mathcal{X}}(z, x')^p \, d\mu^y(x', e') \\ &> \int_{F_y} r_y^p \, d\mu^y(x', e') = r_y^p. \end{aligned}$$

The previous inequality holds for any $z \in \mathcal{X} \backslash B_{d_{\mathcal{X}}}(x, 2r_y)$. On the other hand,

$$\begin{aligned} f_y(x) &= \int_{F_y} d_{\mathcal{X}}(x, x')^p \, d\mu^y(x', e') \\ &\leqslant \int_{F_y} r_y^p \, d\mu^y(x', e') = r_y^p. \end{aligned}$$

Therefore $f_y(z) > f_y(x)$ whenever $z \notin B(x, 2r_y)$, which implies that points z outside of such ball cannot be minimisers, hence proving (26). This concludes part (II) of the claim.

Armed with the claim, we can prove the required (a)-(e) properties of Ψ. First, let us prove (a), namely that Ψ in (25) has non-empty values. By the Heine-Borel property of the metric $d_{\mathcal{X}}$, the set $B_{d_{\mathcal{X}}}(x, 2r_y) \subset \mathcal{X}$ is compact, since it is a closed ball with respect to the metric $d_{\mathcal{X}}$. Hence, the function f_y is continuous by (I) and its minimisers are by (II) restricted to the compact set $B_{d_{\mathcal{X}}}(x, 2r_y)$. Therefore, the minimum in (25) is attained by the Extreme Value Theorem. This shows that the argmin is non-empty, and hence that Ψ has non-empty values on $\mathcal{M}_2^{\mathcal{E}}$.

We now proceed to prove (b), namely that Ψ is measurable and has compact values. We will apply the Maximum Measurable Theorem, 7.7, with $S = \mathcal{M}_2^{\mathcal{E}}$, $X = \mathcal{X}$,

$$\begin{aligned} &\varphi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}, \qquad \varphi(y) = B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1)), \\ &f : \mathcal{M}_2^{\mathcal{E}} \times \mathcal{X} \to \mathbb{R}, \; f(y, z) = f_y(z) = \int_{F_y} d_{\mathcal{X}}(x, z)^p \, d\mu^y(x, e). \end{aligned}$$

In order to apply the theorem, we need to verify that the assumptions are satisfied. We start by proving that φ is weakly-measurable with non-empty compact values. Firstly, it is clear that, since φ is constant, then φ is weakly measurable and has non-empty values. Moreover, the only value φ takes is $B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1))$, which we now prove to be compact. Note that, since $\mathcal{M}_1$ is compact by Assumption 4, then $\mathcal{M}_1$ is bounded and hence $B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1)) = \{x \in \mathcal{X} : \text{dist}_{d_{\mathcal{X}}}(x, \mathcal{M}_1) \leqslant 2\text{diam}(\mathcal{M}_1)\}$ is bounded too. Moreover, since the function $d(\cdot, \mathcal{M}_1)$ is continuous, the set $B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1)) = d(\cdot, \mathcal{M}_1)^{-1}([0, 2\,\text{diam}(\mathcal{M}_1)])$ is closed. Hence, we have proven that $B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1))$ is closed and bounded, and by the Heine Borel property of $d_{\mathcal{X}}$ granted by Assumption 4 it follows that $B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\text{diam}(\mathcal{M}_1))$ is compact. This proves that φ is weakly-measurable with non-empty compact values. Secondly, to prove that f is Carathéodory, we need to show that $f(y, \cdot) = f_y$ is continuous for every fixed $y \in \mathcal{M}_2^{\mathcal{E}}$ and that $f(\cdot, z)$ is measurable for every fixed $z \in \mathcal{M}_1$. On the one hand, for every fixed $y \in \mathcal{M}_2^{\mathcal{E}}$, the function f_y is continuous on $\mathcal{X}$ as proven in claim (I). On the other hand, for every fixed z, the function $f(\cdot, z) : y \mapsto \int_{F_y} d_{\mathcal{X}}(x, z)^p \, d\mu^y(x, e)$ is

Borel measurable due to the definition of disintegration of measure 3.6. Then, by Theorem 7.7, the possibly multi-valued function $\Phi : \mathcal{M}_2^{\mathcal{E}} \rightrightarrows \mathcal{X}$ given by

$$\Phi(y) = \operatorname*{argmin}_{z \in B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\mathrm{diam}(\mathcal{M}_1))} \int_{F^{-1}y} d_{\mathcal{X}}(x,z)^p \, d\mu^y(x,e)$$

is measurable and has non-empty, compact values. Moreover, by combining (26) and the fact that for $(x,e) \in G_y$, $B_{d_{\mathcal{X}}}(x, 2r_y) \subseteq B_{d_{\mathcal{X}}}(\mathcal{M}_1, 2\mathrm{diam}(\mathcal{M}_1))$, we deduce that $\Phi = \Psi$. Hence, Ψ is measurable and has non-empty, compact values.

We now prove (c), namely that $\Psi \in \mathcal{C}$. This follows directly from Proposition 7.8.

We now proceed to prove (d), namely that Ψ is an optimal map. Let $\varphi \in \mathcal{C}$. By the minimising definition of Ψ,

$$\int_{F_y} d_{\mathcal{X}}^H(\Psi(y), x)^p \, d\mu^y(x,e) \leqslant \int_{F_y} d_{\mathcal{X}}(z,x)^p \, d\mu^y(x,e)$$

for every $z \in \varphi(y)$. In particular, taking the supremum with respect to $z \in \varphi(y)$, which coincides with considering the Hausdorff distance, and using Fatou's Lemma yields

$$\begin{aligned}\int_{F_y} d_{\mathcal{X}}^H(\Psi(y), x)^p \, d\mu^y(x,e) &\leqslant \sup_{z \in \varphi(y)} \int_{F_y} d_{\mathcal{X}}(z,x)^p \, d\mu^y(x,e) \\ &\leqslant \int_{F_y} \sup_{z \in \varphi(y)} d_{\mathcal{X}}(z,x)^p \, d\mu^y(x,e) \\ &\leqslant \int_{F_y} d_{\mathcal{X}}^H(\varphi(y), x)^p \, d\mu^y(x,e).\end{aligned}$$

By integrating with respect to $y \in \mathcal{M}_2^{\mathcal{E}}$, we obtain

$$\int_{y \in \mathcal{M}_2^{\mathcal{E}}} \int_{F_y} d_{\mathcal{X}}^H(\Psi(y), x)^p \, d\mu^y(x,e) \, d(F_*\mu)(y) \leqslant \int_{y \in \mathcal{M}_2^{\mathcal{E}}} \int_{F_y} d_{\mathcal{X}}^H(\varphi(y), x)^p \, d\mu^y(x,e) \, d(F_*\mu)(y).$$

Thanks to the disintegration of measure, this can be rewritten as

$$\int_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} d_{\mathcal{X}}^H(\Psi(F(x,e)), x) \, d\mu(x,e) \leqslant \int_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} d_{\mathcal{X}}^H(\varphi(F(x,e)), x) \, d\mu(x,e).$$

Now, as $\varphi \in \mathcal{C}$ was arbitray and by raising both sides to the power $\frac{1}{p}$, we obtain

$$\left(\int_{(x,e) \in \mathcal{M}_1 \times \mathcal{E}} d_{\mathcal{X}}^H(\Psi(F(x,e)), x)^p \, d\mu(x,e) \right)^{\frac{1}{p}} \leqslant c_{\mathrm{opt}}^{\mathrm{a}}(F, \mathcal{M}_1, \mathcal{E}, p).$$

The opposite inequality holds trivially, as $\Psi \in \mathcal{C}$. Therefore, Ψ is an optimal map.

Finally, we proceed to prove (e), namely the upper bound $c_{\mathrm{opt}}^{\mathrm{a}}(F, \mathcal{M}_1, \mathcal{E}, p) \leqslant \mathrm{kersize}^{\mathrm{a}}(F, \mathcal{M}_1, \mathcal{E}, p)$. By the minimisation property of Φ, hence also of Ψ, for every $(x', e') \in F_y$:

$$\int_{F^{-1}y} d_{\mathcal{X}}^H(x, \Psi(y))^p \, d\mu^y(x,e) \leqslant \int_{F^{-1}y} d_{\mathcal{X}}(x,x')^p \, d\mu^y(x,e). \tag{27}$$

Integrating (27) with respect to μ^y yields,

$$\int_{F_y} d_{\mathcal{X}}^H(x, \Psi(y))^p \, d\mu^y(x,e) \leqslant \int_{F_y} \int_{F_y} d_{\mathcal{X}}(x,x')^p \, d\mu^y(x,e) \, d\mu^y(x',e').$$

where we used that μ^y is a probability measure. Integrating both sides over $y \in \mathcal{M}_2^{\mathcal{E}}$ with respect to $F_*\mu$ we obtain

$$\begin{aligned}&\int_{\mathcal{M}_2^{\mathcal{E}}} \int_{F_y} d_{\mathcal{X}}^H(x, \Psi(y))^p \, d\mu^y(x,e) \, d(F_*\mu)(y) \\ &\leqslant \int_{\mathcal{M}_2^{\mathcal{E}}} \int_{F_y} \int_{F_y} d_{\mathcal{X}}(x,x')^p \, d\mu^y(x,e) \, d\mu^y(x',e') \, d(F_*\mu)(y).\end{aligned}$$

Using the definition of disintegration of the measure μ on the left hand side of the above inequality yields

$$\int_{\mathcal{M}_1\times\mathcal{E}} d_{\mathcal{X}}^H(x,\Psi(F(x,e)))^p \, d\mu(x,e) \leqslant \operatorname{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p)^p.$$

Then, raising both sides to the $\frac{1}{p}$-th power, gives

$$\left(\int_{\mathcal{M}_1\times\mathcal{E}} d_{\mathcal{X}}^H(x,\Psi(F(x,e)))^p \, d\mu(x,e)\right)^{\frac{1}{p}} \leqslant \operatorname{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p).$$

And finally, since $\Psi \in \mathcal{C}$, we conclude

$$c_{\mathrm{opt}}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p) \leqslant \operatorname{kersize}^{\mathrm{a}}(F,\mathcal{M}_1,\mathcal{E},p).$$

This concludes the case $p \in [1,+\infty)$. Hence the proof is complete.

Now we consider the case $p = \infty$. The proof only requires minor modifications to the case $p \in [1,\infty)$, that we will describe in the following. The objective function for $y \in \mathcal{M}_2^{\mathcal{E}}$ and $z \in \mathcal{X}$ is

$$f_y(z) = \operatorname*{essup}_{(x,e)\in F_y} d_{\mathcal{X}}(x,z).$$

where the essential supremum is taken with respect to μ^y. Then $\Psi(y) = \operatorname{argmin}_{\mathcal{X}} f_y$. Define, as before,

$$r_y = \operatorname*{essup}_{\substack{(x,e)\in F_y \\ (x',e')\in F_y}} d_{\mathcal{X}}(x,x').$$

Claim. For every $y \in \mathcal{M}_2^{\mathcal{E}}$:

(I) f_y is continuous;

(II) $\operatorname*{argmin}_{\mathcal{X}} f_y = \operatorname*{argmin}_{B(x,2r_y)} f_y$ for μ^y-almost every $(x,e) \in F_y$.

The claims are proven analogously to the case $p \in [1,\infty)$, except that to prove (I), Minkowski's inequality is replaced by the following. We consider a general setting, and let (X,d) be a metric space, $A \subseteq X$ be a bounded subset, ν be a probability measure on X concentrated on A and $g(x) := \operatorname{essup}_{a\in A} d(x,a)$. We claim that $|g(x) - g(z)| \leqslant d(x,z)$ for all $x,z \in X$. To see this, consider $a,x,z \in X$ and notice that

$$d(x,a) \leqslant d(x,z) + d(z,a) \quad \Longrightarrow \quad \operatorname*{essup}_{a\in A} d(x,a) \leqslant d(x,z) + \operatorname*{essup}_{a\in A} d(z,a).$$

Switching the roles of x and z, leads to the desired inequality. It follows that g is continuous. Claim (II) follows the same line of reasoning as in the case $p \in [1,\infty)$ by replacing the integration with the essential supremum with respect to μ^y.

Armed with the claim, we can show the properties given by the list (a)-(e) stated at the beginning of the proof, similiarly to the case $p \in [1,\infty)$. For the sake of brevity, we provide a sketch of the proof.

Part (a) follows the same line of reasoning as in the case $p \in [1,\infty)$. In part (b) we again apply the Maximum Measurable Theorem 7.7 with $f(y,z) = f_y(z) := \operatorname{essup}_{(x,e)\in F_y} d_{\mathcal{X}}(x,z)$ and apply Proposition 7.9 to show that $f(\cdot,z) : y \mapsto \operatorname{essup}_{(x,e)\in F_y} d_{\mathcal{X}}(x,z)$ is Borel measurable. Part (c) is again a a direct consequence of (b) and Proposition 7.8. Parts (d) and (e) follows the same line of reasoning as in the case $p \in [1,\infty)$ by replacing the integration with the essential supremum. This concludes the proof of the proposition in the case $p = \infty$.

□

7.5. **Proof of Proposition 4.1.**

Proof. (1) Starting from the joint random variable $(X,E) \sim \mu_{(X,E)} = \mu$ consider the marginal $X = \pi_1 \circ (X,E)$ with marginal distribution $\mu_X = \pi_{1*}\mu_{(X,E)} = \pi_{1*}\mu = \mathbb{P}[X \in \cdot]$. This is precisely the Bayesian prior distribution of X.

(2) By Proposition 7.1 applied to $\pi_1 \colon (\mathcal{X} \times \mathcal{Z}, \mu_{(X,E)}) \to (\mathcal{X}, \mu_X)$, there exists a disintegration $\{\mu^x_{(X,E)} =: \mu_{(x,E)}\}_{x\in\mathcal{X}}$ on $\mathcal{X}\times\mathcal{Z}$, each of which concentrated on $\{x\}\times\mathcal{Z}$. It satisfies $\mu_{(X,E)}(A) = \int_{\mathcal{X}} \mu_{(x,E)}(A) d\mu_X(x)$.

This disintegration can be pushed forward via F: recall that $\mu_Y = F_*\,\mu_{(X,E)}$, and define $\mu_{Y|X=x} \coloneqq F_*\,\mu_{(x,E)}$ for every $x \in \mathcal{X}$. Then it is immediate to notice that $\{\mu_{Y|X=x}\}_{x\in\mathcal{X}}$ is a regular conditional distribution of Y given X. In fact, for every $A \in \mathcal{B}(\mathcal{Y})$ the defining property of conditional distributions is verified:

$$\mu_Y(A) = F_*\,\mu_{(X,E)}(A) = \mu_{(X,E)}(F^{-1}(A)) = \int_{\mathcal{X}} \mu_{(x,E)}(F^{-1}(A)) d\mu_X(x) =$$
$$= \int_{\mathcal{X}} \Big(F_*\,\mu_{(x,E)}(A)\Big) d\mu_X(x) = \int_{\mathcal{X}} \mu_{Y|X=x}(A) d\mu_X(x).$$

(3) In Bayesian terms, the Bayes posterior is a family of distributions $\{\mu_{X|Y=y}\}_{y\in\mathcal{Y}}$ that is uniquely defined (up to μ_Y-almost equivalence) by satisfying the condition

$$\mu_X(A) = \int_{\mathcal{Y}} \mu_{X|Y=y}(A) d\mu_Y(y).$$

Let us now prove that the family $\{\pi_{1*}\mu^y\}_{y\in\mathcal{Y}}$ satisfies the previous condition.

For $A \in \mathcal{B}(\mathcal{X})$, we have

$$\mu_X(A) = (\pi_{1*}\mu)(A) = \mu(\pi^{-1}(A)) = \int_{\mathcal{X}\times\mathcal{Z}} 1_{\pi^{-1}(A)} d\mu = \int_{\mathcal{Y}} \Big(\int_{\mathcal{X}\times\mathcal{Z}} 1_{\pi^{-1}(A)} d\mu^y \Big) d\mu_Y(y) = \tag{28}$$
$$= \int_{\mathcal{Y}} \pi_{1*}\mu^y(A) d\mu_Y(y) = \int_{\mathcal{Y}} \mu^y(\pi^{-1}(A)) d\mu_Y(y). \tag{29}$$

Therefore, the defining condition for the posterior is verified. By uniqueness of the posterior, we conclude that $\pi_{1*}\mu^y = \mu_{X|Y=y}$ for every $y \in \mathcal{Y}$.

(4) Note that the following is a special case of section 3.3 in [11] using the work of [57]. Moreover, the optimal map obtained in the following proof is not compact-valued, as we do not assume that $\mathcal{X}$ is compact. For fixed $y \in \mathcal{Y}$, by Proposition 7.1 applied to $\pi_1 \colon (\mathcal{X} \times \mathcal{Z}, \mu^y) \to (\mathcal{X}, \pi_{1*}\,\mu^y)$, there exists a disintegration $\{\mu^{y,x}_{(X,E)} =: \mu_{(x,E)|Y=y}\}_{x\in\mathcal{X}}$ on $\mathcal{X}\times\mathcal{Z}$, each of which concentrated on $\{x\}\times\mathcal{Z}$. It satisfies for every $A \in \mathcal{B}(\mathcal{X}\times\mathcal{Z})$ that $\mu^y(A) = \int_{\mathcal{X}} \mu^y_{(x,E)|Y=y}(A) d(\pi_{1*}\,\mu^y)(x)$. More generally, if $f : \mathcal{X}\times\mathcal{Z} \to [0, +\infty]$ is a positive measurable function, it holds that

$$\int_{\mathcal{X}\times\mathcal{Z}} f(x,e)\, d\mu^y(x,e) = \int_{\mathcal{X}} \Big(\int_{\mathcal{X}\times\mathcal{Z}} f(x,e)\, d\mu_{(x,E)|Y=y}(x,e) \Big) \dot{(}\pi_{1*}\mu^y)(x).$$

Notice in particular that, if f only depends on x, the previous condition simplifies to

$$\int_{\mathcal{X}\times\mathcal{Z}} f(x)\, d\mu^y(x) = \int_{\mathcal{X}} f(x) \Big(\int_{\mathcal{X}\times\mathcal{Z}} d\mu_{(x,E)|Y=y}(x,e) \Big) \dot{(}\pi_{1*}\mu^y)(x) \tag{30}$$
$$= \int_{\mathcal{X}} f(x) \dot{(}\pi_{1*}\mu^y)(x) \tag{31}$$

where we used the fact that $\mu_{(x,E)|Y=y}$ is a probability measure.

Therefore, the optimal map can be rewritten in terms of the posterior distribution in the following way: for every $y \in \mathcal{Y}$, by (30) it holds that

$$\int_{F_y} d_{\mathcal{X}}(x,z)^p\, d\mu^y(x,e) = \int_{\mathcal{X}} d_{\mathcal{X}}(x,z)^p\, d(\pi_{*1}\,\mu^y)(x)$$

since the noise e was not involved in the integrand function. Therefore, the optimal map can be rewritten in terms of the posterior distribution in the following way:

$$\Psi(y) \coloneqq \operatorname*{argmin}_{z\in\mathcal{X}} \int_{F_y} d_{\mathcal{X}}(x,z)^p\, d\mu^y(x,e) = \operatorname*{argmin}_{z\in\mathcal{X}} \int_{\mathcal{X}} d_{\mathcal{X}}(x,z)^p\, d(\pi_{*1}\,\mu^y)(x).$$

In the special case where $\mathcal{X} = \mathbb{R}^N$ is an Euclidean space equipped with the Euclidean metric $d_{\mathcal{X}} = d_{\|\cdot\|_2}$ induced by the 2-norm $\|\cdot\|_2$, and considering the exponent $p = 2$, the optimal map reduces to

$$\Psi(y) = \operatorname*{argmin}_{z\in\mathbb{R}^N} \int_{\mathbb{R}^N} \|x - z\|_2^2 \, d\mu_{X|Y=y}(x)$$

which is the Minimum Mean Squared Error (MMSE) for the posterior distribution $\mu_{X|Y=y}$. It is a well-known result that, for any distribution for which enough moments are finite (mean and variance), the point that minimises the mean squared error is the expected value. In fact, by differentiating the function $g\colon \mathbb{R}^N \to [0, +\infty)$

$$g(z) \coloneqq \int_{\mathbb{R}^N} \|z - x\|_2^2 \, d\mu_{X|Y=y}(x) = \sum_{i=1}^{N} \int_{\mathbb{R}^N} (z_i - x_i)^2 d\mu_{X|Y=y}(x)$$

we obtain for any $i \in \{1, \dots, N\}$ that

$$\begin{aligned}\Big(\frac{\partial}{\partial z_i} g\Big)(z) &= \int_{\mathbb{R}^N} 2(z_i - x_i) \, d\mu_{X|Y=y}(x) = \\ &= 2z_i - 2\int_{\mathbb{R}^N} x_i \, d\mu_{X|Y=y}(x) \\ &= 2\big(z_i - \mathbb{E}_{\mu_{X|Y=y}}[x_i]\big).\end{aligned}$$

Therefore, imposing $\frac{\partial}{\partial z_i} g = 0$ for every i, we obtain that the minimiser of the optimal map is given by

$$\Phi(y) = z = (z_1, \dots, z_N) = \Big(\mathbb{E}_{\mu_{X|Y=y}}[x_1], \dots, \mathbb{E}_{\mu_{X|Y=y}}[x_N]\Big) = \mathbb{E}_{\mu_{X|Y=y}}[X].$$

□

REFERENCES

[1] B. Adcock, S. Brugiapaglia, and C. G. Webster. *Sparse polynomial approximation of high-dimensional functions*, volume 25. SIAM, 2022.

[2] B. Adcock and A. C. Hansen. *Compressive Imaging: Structure, Sampling, Learning*. Cambridge University Press, Cambridge, UK, 2021.

[3] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5, page e4. Cambridge University Press, 2017.

[4] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[5] C. D. Aliprantis and K. C. Border. *Infinite dimensional analysis: A Hitchhiker's Guide*. Springer, 2006.

[6] D. Amir and J. Mach. Chebyshev centers in normed spaces. *Journal of approximation theory*, 40(4):364–374, 1984.

[7] D. Amir, J. Mach, and K. Saatkamp. Existence of Chebyshev centers, best n-nets and best compact approximants. *Transactions of the American Mathematical Society*, 271(2):513–524, 1982.

[8] V. Antun, M. J. Colbrook, and A. C. Hansen. Proving existence is not enough: Mathematical paradoxes unravel the limits of neural networks in artificial intelligence. *SIAM News*, 55(04):1–4, May 2022.

[9] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci. USA*, 117(48):30088–30095, 2020.

[10] V. Arestov. Optimal recovery of operators and related problems. *Collection of Papers from the All-Union School on the Theory of Functions*, pages 3–20, 1986.

[11] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numer.*, 28:1–174, 2019.

[12] G. Aubert and J.-F. Aujol. A variational approach to removing multiplicative noise. *SIAM journal on applied mathematics*, 68(4):925–946, 2008.

[13] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[14] H. H. Barrett and K. J. Myers. *Foundations of image science*. John Wiley & Sons, 2013.

[15] A. Bastounis and A. C. Hansen. On the absence of uniform recovery in many real-world applications of compressed sensing and the restricted isometry property and nullspace property in levels. *SIAM Journal on Imaging Sciences*, 10(1):335–371, 2017.

[16] A. Bastounis, A. C. Hansen, and V. Vlačić. The extended Smale's 9th problem – On computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs and learning. *arXiv:2110.15734*, 2021.

[17] C. Belthangady and L. A. Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature methods*, 16(12):1215–1225, 2019.

[18] J. Ben-Artzi, M. J. Colbrook, A. C. Hansen, O. Nevanlinna, and M. Seidel. Computing spectra – On the solvability complexity index hierarchy and towers of algorithms. *arXiv:1508.03280v5*, 2020.

[19] J. Ben-Artzi, M. Marletta, and F. Rösler. Computing scattering resonances. *Journal of the European Mathematical Society*, 2023.

[20] M. Bertero, P. Boccacci, and C. De Mol. *Introduction to inverse problems in imaging*. CRC press, 2021.

[21] P. Binev, A. Bonito, R. DeVore, and G. Petrova. Optimal Learning. *arXiv preprint arXiv:2203.15994*, 2022.

[22] T. Blumensath. Sampling and reconstructing signals from a union of linear subspaces. *IEEE Transactions on Information Theory*, 57(7):4660–4671, 2011.

[23] V. I. Bogachev. *Measure theory*, volume 1. Springer, 2007.

[24] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21, 2008.

[25] C. A. Bouman. *Foundations of Computational Imaging: A Model-Based Approach*. SIAM, Philadelphia, PA, 2022.

[26] A. Bourrier, M. E. Davies, T. Peleg, P. Pérez, and R. Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 60(12):7928–7946, 2014.

[27] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011.

[28] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[29] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[30] E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[31] E. J. Candès and Y. Plan. Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[32] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden Voice Commands. In *25th USENIX Security Symp.*, pages 513–530, 2016.

[33] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Worksh.*, pages 1–7, 2018.

[34] J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

[35] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[36] A. Cohen, R. Devore, G. Petrova, and P. Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

[37] M. Colbrook. On the computation of geometric features of spectra of linear operators on hilbert spaces. *Foundations of Computational Mathematics*, 2022 (online).

[38] M. Colbrook and A. C. Hansen. The foundations of spectral computations via the solvability complexity index hierarchy. *Journal of the European Mathematical Society*, 2022 (online).

[39] M. J. Colbrook, V. Antun, and A. C. Hansen. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. *Proc. Natl. Acad. Sci. USA*, 119(12):e2107151119, 2022.

[40] R. Cont and P. Tankov. Retrieving lévy processes from option prices: Regularization of an ill-posed inverse problem. *SIAM Journal on Control and Optimization*, 45(1):1–25, 2006.

[41] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear Approximation and (Deep) ReLU Neural Networks. *Constructive Approximation*, 55(1):127–172, 2022.

[42] K. de Leeuw, E. F. Moore, C. E. Shannon, and N. Shapiro. Computability by probabilistic machines. In *Automata studies*, volume no. 34 of *Ann. of Math. Stud.*, pages 183–212. Princeton Univ. Press, Princeton, NJ, 1956.

[43] H. W. Engl and C. W. Groetsch. *Inverse and ill-posed problems*, volume 4. Elsevier, 2014.

[44] M. Ettehad and S. Foucart. Instances of computational optimal recovery: dealing with observation errors. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1438–1456, 2021.

[45] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[46] A. Fannjiang and T. Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29:125–228, 2020.

[47] C. Fefferman, A. C. Hansen, and S. Jitomirskaya, editors. *Computational mathematics in computer assisted proofs*, American Institute of Mathematics Workshops. American Institute of Mathematics, 2022. Available online at `https://aimath.org/pastworkshops/compproofsvrep.pdf`.

[48] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[49] S. Foucart. *Mathematical Pictures at a Data Science Exhibition*. Cambridge University Press, 2022.

[50] S. Foucart, C. Liao, S. Shahrampour, and Y. Wang. Learning from non-random data in Hilbert spaces: an optimal recovery perspective. *Sampling Theory, Signal Processing, and Data Analysis*, 20(1):5, 2022.

[51] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Birkhauser, 2013.

[52] R. Garcia and F. Crespon. Radio tomography of the ionosphere: Analysis of an underdetermined, ill-posed inverse problem, and regional application. *Radio Science*, 43(02):1–13, 2008.

[53] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77, 2023.

[54] L. E. Gazdag and A. C. Hansen. Generalised hardness of approximation and the SCI hierarchy – On determining the boundaries of training algorithms in AI. *arXiv:2209.06715*, 2022.

[55] N. M. Gottschling, V. Antun, A. C. Hansen, and B. Adcock. The troublesome kernel – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems. *arXiv preprint arXiv:2001.01258*, 2023.

[56] A. C. Hansen. On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators. *Journal of the American Mathematical Society*, 24(1):81–124, 2011.

[57] T. Helin and M. Burger. Maximum a posteriori probability estimates in infinite-dimensional bayesian inverse problems. *Inverse Problems*, 31(8):085009, 2015.

[58] D. P. Hoffman, I. Slavitt, and C. A. Fitzpatrick. The promise and peril of deep learning in microscopy. *Nature Methods*, 18(2):131–132, 2021.

[59] A. Hofinger and H. K. Pikkarainen. Convergence rates for linear inverse problems in the presence of an additive normal noise. *Stochastic Analysis and Applications*, 27(2):240–257, 2009.

[60] Y.-M. Huang, M. K. Ng, and Y.-W. Wen. A new total variation method for multiplicative noise removal. *SIAM Journal on imaging sciences*, 2(1):20–40, 2009.

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

[62] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege. Stable low-rank matrix recovery via null space properties. *Information and Inference: A Journal of the IMA*, 5(4):405–441, 2016.

[63] J. P. Kaipio, V. Kolehmainen, M. Vauhkonen, and E. Somersalo. Inverse problems with structural prior information. *Inverse problems*, 15(3):713, 1999.

[64] O. Kallenberg. *Foundations of modern probability*. Springer, 2021. Third edition.

[65] N. Keriven and R. Gribonval. Instance optimal decoding and the restricted isometry property. In *Journal of Physics: Conference Series*, page 012002. IOP Publishing, 2018.

[66] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Int. Conf. on Learning Representations*, 2017.

[67] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[68] A. S. Leonov. A posteriori accuracy estimations of solutions to ill-posed inverse problems and extra-optimal regularizing algorithms for their solution. *Numerical Analysis and Applications*, 5(1):68–83, 2012.

[69] A. S. Leonov. Locally extra-optimal regularizing algorithms and a posteriori estimates of the accuracy for ill-posed problems with discontinuous solutions. *Computational Mathematics and Mathematical Physics*, 56(1):1–13, 2016.

[70] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi. Deep text classification can be fooled. In *The 27th Int. Joint Conf. on Artificial Intelligence*, 2017.

[71] J. Liu, R. Anirudh, J. J. Thiagarajan, S. He, K. A. Mohan, U. S. Kamilov, and H. Kim. DOLCE: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10508, 2023.

[72] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

[73] G. G. Magaril-Il'yaev and K. Y. Osipenko. Optimal recovery of functionals based on inaccurate data. *Mat. Zametki*, 50(6):85–93, 1991.

[74] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[75] C. A. Micchelli and T. J. Rivlin. A survey of optimal recovery. *Optimal estimation in approximation theory*, pages 1–54, 1977.

[76] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[77] M. J. Muckley, B. Riemenschneider, A. Radmanesh, S. Kim, G. Jeong, J. Ko, Y. Jun, H. Shin, D. Hwang, M. Mostapha, et al. State-of-the-art Machine Learning MRI Reconstruction in 2020: Results of the Second fastMRI Challenge. *arXiv preprint arXiv:2012.06318*, 2020.

[78] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[79] E. Novak and K. Ritter. A stochastic analog to Chebyshev centers and optimal average case algorithms. *Journal of Complexity*, 5(1):60–79, 1989.

[80] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518, 1986.

[81] A. Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

[82] L. Plaskota. *Noisy information and computational complexity*, volume 95. Cambridge University Press, 1996.

[83] C. Poon. Structure dependent sampling in compressed sensing: theoretical guarantees for tight frames. *Applied and Computational Harmonic Analysis*, 42(3):402–451, 2017.

[84] T. Rao. Chebyshev centres and centrable sets. *Proceedings of the American Mathematical Society*, 130(9):2593–2598, 2002.

[85] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[86] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

[87] J. Shi and S. Osher. A nonlinear inverse scale space method for a convex multiplicative noise model. *SIAM Journal on imaging sciences*, 1(3):294–321, 2008.

[88] L. A. Steen, J. A. Seebach, and L. A. Steen. *Counterexamples in topology*, volume 1. Springer, 1978.

[89] R. Strack. Imaging: AI transforms image reconstruction. *Nature Methods*, 15(5):309, 2018.

[90] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[91] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Int. Conf. on Learning Representations*, 2014.

[92] H. Tran and C. Webster. A class of null space conditions for sparse recovery via nonconvex, non-separable minimizations. *Results in Applied Mathematics*, 3:100011, 2019.

[93] Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all. *Applied and Computational Harmonic Analysis*, 45(1):170–205, 2018.

[94] Y. Traonmilin, R. Gribonval, and S. Vaiter. A theory of optimal convex regularization for low-dimensional recovery. *arXiv preprint arXiv:2112.03540*, 2021.

[95] M. Unser. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics*, pages 1–20, 2020.

[96] Y. Wang, A. S. Leonov, D. V. Lukyanenko, and A. G. Yagola. General Tikhonov regularization with applications in geoscience. *CSIAM Trans. Appl. Math*, 1:53–85, 2020.

[97] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[98] J. Ward. Chebyshev centers in spaces of continuous functions. *Pacific journal of mathematics*, 52(1):283–287, 1974.

[99] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. Dolphinattack: Inaudible voice commands. In *Proc. of the 2017 ACM SIGSAC Conf. on Computer and Commun. Security*, pages 103–117, 2017.

[100] X.-L. Zhao, F. Wang, and M. K. Ng. A new convex optimization model for multiplicative noise and blur removal. *SIAM Journal on Imaging Sciences*, 7(1):456–475, 2014.

[101] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 03 2018.

DEPARTMENT OF APPLIED MATHEMATICS AND THEORETICAL PHYSICS, UNIVERSITY OF CAMBRIDGE
Email address: nmg43@cam.ac.uk

DEPARTMENT OF APPLIED MATHEMATICS AND THEORETICAL PHYSICS, UNIVERSITY OF CAMBRIDGE
Email address: pc628@cam.ac.uk

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF OSLO
Email address: vegarant@math.uio.no

DEPARTMENT OF APPLIED MATHEMATICS AND THEORETICAL PHYSICS, UNIVERSITY OF CAMBRIDGE
Email address: ach70@cam.ac.uk