Optimal Prediction of Multivalued Functions from Point Samples

Simon Foucart*— Texas A&M University

Abstract

Predicting the value of a function f at a new point given its values at old points is an ubiquitous scientific endeavor, somewhat less developed when f produces multiple values that depend on one another, e.g. when it outputs likelihoods or concentrations. Considering the points as fixed (not random) entities and focusing on the worst-case, this article uncovers a prediction procedure that is optimal relatively to some model-set information about f. When the model sets are convex, this procedure turns out to be an affine map constructed by solving a convex optimization program. The theoretical result is specified in the two practical frameworks of (reproducing kernel) Hilbert spaces and of spaces of continuous functions.

Key words and phrases: Optimal recovery, Minimax problems, Prediction, Reproducing kernel Hilbert spaces, Moment-SoS hierarchy, Dominated extension theorem, Support function.

AMS classification: 41A65, 46N40, 65D15, 90C47.

1 Introduction

In this article, one considers the general problem of predicting the value $f(x^{(0)})$ of an unknown function f at a 'new' point $x^{(0)}$, given M pieces of data

$$y_m = f(x^{(m)}), \qquad i \in [1:M],$$

consisting of evaluations of f at 'old' points $x^{(1)}, \ldots, x^{(M)}$. This undertaking is a core problem of Statistical Learning Theory (see e.g. [12] for an initiation). There, the $x^{(m)}$'s are considered independent realizations of a random variable and the performance of a prediction process is assessed by a notion of risk, which reflects a focus on the average case. The risk decomposes as the sum of an estimation error, often carefully studied, and an approximation error, typically postulated to be small—thus implicitly assuming that the unknown function f is well approximated by elements of a chosen hypothesis class. Another view, favoring a focus on the worst case, is provided by the theory of Optimal Recovery (see e.g. [10] for an initiation). There, the $x^{(m)}$'s are considered

*S. F. is partially supported by a grant from the Office of Naval Research (N00014-20-1-2787). Part of this work was carried out during a stay at the Isaac Newton Institute, supported by a grant from the Heilbronn Institute.

fixed deterministic entities and one has some *a priori* knowledge about f in the form $f \in \mathcal{K}$—for instance, that f is approximated by elements of a hypothesis class up to some given accuracy. The prediction process, encapsulated by a map Δ taking the y_m's as input and returning an estimation of $f(x^{(0)})$ as an output, has its performance quantified by the (global[1]) worst-case error

$$\mathrm{gwce}(\Delta) := \sup_{f \in \mathcal{K}} \big\| f(x^{(0)}) - \Delta(f(x^{(1)}), \dots, f(x^{(M)})) \big\|. \tag{1}$$

A prime objective of Optimal Recovery is to find a map Δ that minimizes $\mathrm{gwce}(\Delta)$, or nearly does. A standard result in the field asserts that such an optimal Δ can be chosen as a linear map as soon as $\mathcal{K}$ is convex and symmetric about the origin and the evaluation map $f \mapsto f(x^{(0)})$ is a linear functional, i.e., the function f outputs a single value.

However, in many Data Science settings, the functions of interest instead output several values. As an important example, the functions $\mathbf{f}$ associated with a neural network are not only multivariate (the inputs are in $\mathbb{R}^D$) but also multivalued (the outputs are in $\mathbb{R}^N$). When neural networks are used for classification, the nth component f_n of $\mathbf{f}$ often represents the likelihood of the nth class, so $f_1, \dots, f_N$ are furthermore nonnegative and sum up to one. This situation is common in practical problems, e.g. $\mathbf{f}(x) = [f_1(x); \dots; f_N(x)]$ can represent chemical concentrations evolving with time x. The illustrative theorems stated below, dealing with such as case, instantiate the more general results proved later, namely Corollaries 11 and 12.

The first theorem takes place in the setting of a reproducing kernel Hilbert space H with kernel K, say. The dependence relation $f_1 + \cdots + f_N = 1$ is imposed, but not the nonnegativity of each component $f_n \in H$. These components are assumed to be well approximated by hypothesis classes which are linear subspaces of H with finite dimensions. Note that the dependence relation presumes that the constant functions belong to H, which excludes the Gaussian kernel, see [11]. Note also that the norm on $\mathbb{R}^N$ used in (1) to quantify the worst-case error is chosen to be the ℓ_∞-norm.

Theorem 1. In a reproducing kernel Hilbert space H containing the constant functions, consider multivalued functions $\mathbf{f} \in H^N$ satisfying $\mathrm{dist}_H(f_1, V_1) \le \varepsilon_1, \dots, \mathrm{dist}_H(f_N, V_N) \le \varepsilon_N$ for some finite-dimensional linear subspaces $V_1, \dots, V_N$ of H and some parameters $\varepsilon_1, \dots, \varepsilon_N \ge 0$, as well as $f_1 + \cdots + f_N = 1$. The prediction of $\mathbf{f}(x^{(0)}) \in \mathbb{R}^N$ from the point values $y_1 = \mathbf{f}(x^{(1)}), \dots, y_M = \mathbf{f}(x^{(M)}) \in \mathbb{R}^N$ is accomplished with minimal ℓ_∞-worst-case error by an affine recovery map $\Delta^{\mathrm{aff}} : y \in \mathbb{R}^{M \times N} \mapsto \sum_{m=1}^M \sum_{n=1}^N y_{m,n} \widehat{c}^{(m,n)} + \widehat{d} \in \mathbb{R}^N$. Its construction involves, for each $j \in [1:N]$, the solutions $\widehat{c}_j = \Big[\widehat{c}_j^{(m,n)}\Big]_{\substack{m \in [1:M] \\ n \in [1:N]}} \in \mathbb{R}^{M \times N}$ and $\widehat{d}_j \in \mathbb{R}$ of the convex optimization program

$$\underset{\substack{u^+, u^- \in H \\ c \in \mathbb{R}^{M \times N},\, d, e \in \mathbb{R}}}{\text{minimize}} \quad e \quad \text{s.to} \quad \pm \langle u^\pm, 1 \rangle + \sum_{n=1}^N \varepsilon_n \Big\| \delta_{j,n} K(\cdot, x^{(0)}) - \sum_{m=1}^M c_{m,n} K(\cdot, x^{(m)}) - u^\pm \Big\|_H \le e \pm d$$

$$\text{and} \quad \delta_{j,n} v(x^{(0)}) - \sum_{m=1}^M c_{m,n} v(x^{(m)}) - \langle u^\pm, v \rangle = 0 \quad \text{for all } v \in V_n,\ n \in [1:N].$$

[1] To distinguished from the local worst-case error, also known as Chebyshev radius.

This result provides a genuinely practical way of constructing an optimal recovery map, since the optimization program—a second-order come program—can be solved computationally, albeit not in closed form. This is true even in the (predominant) case where H is infinite dimensional: in the spirit of a representer theorem, the minimizers can be searched for in a finite-dimensional subspace of H, see Subsection 4.1 for details.

The second theorem takes place in a space $C(\mathcal{X})$ of continuous functions on a compact set $\mathcal{X}$. Still assuming that the components are well approximated by finite-dimensional subspaces and sum up to one, one also incorporates here the fact that they are nonnegative. In the statement below, the set $B(\mathcal{X})$, resp. $B_+(\mathcal{X})$, represents the set of signed, resp. nonnegative, Borel measures on $\mathcal{X}$.

Theorem 2. In a space $C(\mathcal{X})$ of continuous functions on a compact set $\mathcal{X}$, consider multivalued functions $\mathbf{f} \in C(\mathcal{X})^N$ satisfying $\mathrm{dist}_{C(\mathcal{X})}(f_1, V_1) \le \varepsilon_1, \dots, \mathrm{dist}_{C(\mathcal{X})}(f_N, V_N) \le \varepsilon_N$ for some finite-dimensional linear subspaces $V_1, \dots, V_N$ of $C(\mathcal{X})$ and some parameters $\varepsilon_1, \dots, \varepsilon_N \ge 0$, as well as $f_1 + \cdots + f_N = 1$ and $f_1 \ge 0, \dots, f_N \ge 0$. The prediction of $\mathbf{f}(x^{(0)}) \in \mathbb{R}^N$ from the point values $y_1 = \mathbf{f}(x^{(1)}), \dots, y_M = \mathbf{f}(x^{(M)}) \in \mathbb{R}^N$ is accomplished with minimal ℓ_∞-worst-case error by an affine recovery map $\Delta^{\mathrm{aff}} : y \in \mathbb{R}^{M\times N} \mapsto \sum_{m=1}^M \sum_{n=1}^N y_{m,n} \widehat{c}^{(m,n)} + \widehat{d} \in \mathbb{R}^N$. Its construction involves, for each $j \in [1:N]$, the solutions $\widehat{c}_j = \left[\widehat{c}_j^{(m,n)}\right]_{\substack{m\in[1:M]\\ n\in[1:N]}} \in \mathbb{R}^{M\times N}$ and $\widehat{d}_j \in \mathbb{R}$ of the convex optimization program

$$
\begin{aligned}
\underset{\substack{\mu^+,\mu^-\in B(\mathcal{X})\\ \boldsymbol{\nu}^+,\boldsymbol{\nu}^-\in B(\mathcal{X})^N\\ c\in\mathbb{R}^{M\times N},\, d,e\in\mathbb{R}}}{\text{minimize}} \quad e \qquad & \text{s.to } \pm \int_{\mathcal{X}} d\mu^{\pm} + \sum_{n=1}^N \varepsilon_n \int_{\mathcal{X}} d|\nu_n^{\pm}| \le e \pm d,\\
& \text{and } \int_{\mathcal{X}} v\, d\nu_n^{\pm} = 0 \quad \text{for all } v \in V_n,\ n \in [1:N],\\
& \text{and } \nu_n^{\pm} \mp \Big(\delta_{j,n}\delta_{x^{(0)}} - \sum_{m=1}^M c_{m,n}\delta_{x^{(m)}} - \mu^{\pm}\Big) \in B_+(\mathcal{X}), \quad n \in [1:N].
\end{aligned}
$$

The genuine practicality of the construction is now a more subtle question, because the optimization variables include measures, which are infinite dimensional objects. But optimizing over measures can be performed by solving a hierarchy of semidefinite programs, in the sense that solutions to the infinite-dimensional program are limits of solutions to finite-dimensional semidefinite programs of increasing sizes and, in favorable situations, the convergence occurs in a finite number of steps. The readers are referred to [9] and the references therein for more details on this so-called Moment-SOS hierarchy. The current situation has an extra advantage, namely that one is not really interested in the minimizing measures but rather in the minimizing coefficients $c \in \mathbb{R}^{M\times N}$ and $d \in \mathbb{R}$. This point is further discussed in Subsection 4.2.

Here is an outline of the article's content leading to Theorems 1 and 2 above. In Section 2, after a succinct rundown on Optimal Recovery, it is proved that affine maps provide optimal procedures for the prediction, and more generally for the estimation of linear functionals, of multiobjects.

This statement covers the case of independent components relative to any ℓ_p-worst-case error and the case of dependent components relative to the ℓ_∞-worst-case error. Section 3 concentrates on the latter case and transforms the existence result into a constructive, yet rather abstract, result (Theorem 8) involving the support functions of convex model sets. The main ingredient is a functional analytic result stemming from the dominated extension theorem and exploited through two of its consequences. The constructive result is also particularized for two specific model sets. Next, in Section 4, the results are applied in the more practical settings of Hilbert spaces and of spaces of continuous functions, for which the predecessors of Theorems 1 and 2 are derived. Computability is also discussed in this section. Section 5 gives an outlook on further aspects to be explored around the recovery of multivalued functions. The article finishes with an appendix, of interest in its own right, where some Optimal Recovery facts are combined and revisited.

Notation. Throughout this article, plain letters, such as f, g, h, are used to denote single-valued functions or more generally elements of a vector space F. Boldface letters, such as $\mathbf{f}, \mathbf{g}, \mathbf{h}$, are used to denote multivalued functions, or more generally elements in F^N, e.g. $\mathbf{f} = [f_1; \ldots; f_N]$ with $f_1, \ldots, f_N \in F$. Indices are written in lowercase, while their maximal value is written in uppercase, so e.g. $n \in [1 : N]$ means that the index n ranges from 1 to N. Linear functionals are designated by Greek letters, such as $\lambda, \mu, \nu \in F^*$, where F^* stands for the dual space of F. Elements of $(F^*)^N$ are again written in boldface, as in $\boldsymbol{\mu} = [\mu_1; \ldots; \mu_N]$. Particular linear functionals are the evaluations at points x, which are symbolized by δ_x, so that $\delta_x(f) = f(x)$ when f is a function. There should not be any confusion with the Kronecker symbol $\delta_{i,j}$. The Greek letter ρ is used for the Minkowski functional $\rho_{\mathcal{S}}$ associated with a set $\mathcal{S} \subseteq F$. As for the support function of $\mathcal{S}$, is is written as $\|\eta\|_{\mathcal{S}}$ when evaluated at some $\eta \in F^*$. The dual cone of $\mathcal{S}$, which is a subset of F^*, is denoted by $\mathcal{S}^{\mathrm{dual}}$. Specific spaces F coming into play are Hilbert spaces, designated by H, and spaces of continuous functions on a compact set $\mathcal{X}$, designated by $C(\mathcal{X})$. The set of signed, resp. nonnegative, Borel measures on $\mathcal{X}$ is denoted as $B(\mathcal{X})$, resp. $B_+(\mathcal{X})$.

2 Preliminary Considerations

In this section, after recalling some classical results from Optimal Recovery, one quickly unravels a solution to the problem of optimally predicting multivalued functions that present no dependence between their components. Then one establishes the optimality of affine maps when such a dependence is integrated as prior knowledge in the set $\mathcal{K}$, required to be convex, and when the output space is endowed with the ℓ_∞-norm.

2.1 Brief overview of Optimal Recovery

As a start, it is appropriate to recall some ingredients that are already available in the theory of Optimal Recovery. In its abstract form, one does not restrict f to be a function—it generally denotes an element from a vector space F, so multivalued functions are covered. The *a priori* information reads $f \in \mathcal{K}$ for some subset $\mathcal{K}$ of F. This so-called model set encapsulates e.g. scientific knowledge about f. As for the *a posteriori* information, it is provided by data $y_1 = \lambda_1(f), \ldots, y_M = \lambda_M(f)$ obtained from linear maps[2] $\lambda_m : F \to Y_m$ applied to f—they do not need to come from point evaluations. One condenses this information as $y = \Lambda f$, where the linear map $\Lambda : F \to Y_1 \times \cdots \times Y_M$ is called observation map. The goal is then to estimate $\Gamma(f)$, where the quantity of interest $\Gamma : F \to Z$ is a linear map. In the prediction problem, one takes Γ to be the point evaluation $\delta_{x^{(0)}} : f \in F \mapsto f(x^{(0)}) \in \mathbb{R}^N$, but again no restriction on Γ other than linearity is imposed. Thus, the choice $\Gamma = \mathrm{Id}_F$ corresponds to the full recovery problem. The estimation process boils down to a map (cognizant of $\mathcal{K}$ and Λ) $\Delta : Y_1 \times \cdots \times Y_M \to Z$, whose performance is assessed, as in (1), by the worst-case error

$$\mathrm{gwce}(\Delta) := \sup_{f \in \mathcal{K}} \big\| \Gamma(f) - \Delta(\Lambda f) \big\|_Z.$$

The practical construction of recovery maps Δ^{opt} that are optimal (meaning $\mathrm{gwce}(\Delta^{\mathrm{opt}}) \le \mathrm{gwce}(\Delta)$ for any Δ) or near-optimal (meaning $\mathrm{gwce}(\Delta^{\mathrm{opt}}) \le C\,\mathrm{gwce}(\Delta)$ for any Δ) is the ultimate goal of Optimal Recovery. One aims at such constructions in as many relevant situations as possible, always with the hope that Δ^{opt} shall be a 'simple' map, the epitome of which is a linear map. Along these lines, the following facts are well established:

(i) If the quantity of interest Γ is a linear functional and if the model set $\mathcal{K}$ is convex and symmetric about the origin, then there is a linear recovery map which is optimal; furthermore, if $\mathcal{K}$ is merely convex, then the conclusion holds with linear replaced by affine (this fact is due to [13] and is reproved in the appendix with a different argument).

(ii) If F is a Hilbert space and if the model set $\mathcal{K}$ is a centered hyperellipsoid, then there is a linear recovery map which is optimal (this fact is also reproved unconventionally in the appendix); furthermore, if $\mathcal{K}$ is the intersection of two (not more) centered hyperellipsoids, then the conclusion still holds (this fact was recently established in [5]).

(iii) Independently of F and Γ, if the model set $\mathcal{K}$ is convex and symmetric about the origin, then there is always a linear recovery map which is near-optimal with factor $C = 1 + \sqrt{M}$ (this fact is from [1]); furthermore, if $\mathcal{K}$ is merely convex, then the conclusion holds for affine maps with a near-optimality factor $C = 2\min\{\sqrt{N}, 1 + \sqrt{M}\}$ (this fact is proved in the appendix).

For the prediction of multivalued functions, none of these results are very useful: (i) does not apply because, as already pointed out, the point evaluation $\delta_{x^{(0)}}$ maps into $\mathbb{R}^N$ and hence is not a linear

[2]The λ_m's are usually linear functionals, so that $Y_m = \mathbb{R}$, and one can always reduce to this case, but if f is an N-valued function and λ_m is a point evaluation, then $Y_m = \mathbb{R}^N$.

functional, (ii) is too restrictive on the choice of model set, and (iii) supplies near-optimality factors which are too large.

2.2 Simple solution in case of independent components

In this subsection, the object is not yet pinned down as a multivariate function, as it is only required to be of the form $\mathbf{f} = [f_1; \dots; f_N]$ where each component f_n belongs to a common vector space F. For each n, there is an *a priori* information of the type $f_n \in \mathcal{K}_n$ for some subset $\mathcal{K}_n$ of F. Thus, the overall model set for the whole object is

$$\boldsymbol{\mathcal{K}}^{\mathrm{ind}} := \big\{\mathbf{f} = [f_1; \dots; f_N] \in F^N : f_1 \in \mathcal{K}_1, \dots, f_N \in \mathcal{K}_N\big\} \subseteq F^N. \tag{2}$$

No dependence relation between the f_n's, such as $f_1 + \cdots + f_N = 1$, is assumed at this point—this explains the chosen terminology of independent components. As soon to be revealed, this situation is not too interesting, since optimal recovery maps can be constructed componentwise. In the following statement, the linear functionals $\lambda_1, \dots, \lambda_M \in F^*$ generalize the point evaluations $\delta_{x^{(1)}}, \dots, \delta_{x^{(M)}}$ and give rise to the observation map

$$\Lambda : \mathbf{f} \in F^N \mapsto [\lambda(f_1), \dots, \lambda(f_N)] \in \mathbb{R}^{M \times N}, \qquad \lambda(f_n) := [\lambda_1(f_n); \dots; \lambda_M(f_n)] \in \mathbb{R}^M, \tag{3}$$

while the quantity of interest

$$\Gamma : \mathbf{f} \in F^N \mapsto [\gamma_1(f_1); \dots; \gamma_N(f_N)] \in \mathbb{R}^N \tag{4}$$

should be thought of with linear functionals $\gamma_1, \dots, \gamma_N$ all being equal to $\delta_{x^{(0)}}$. With only minor modifications, the proof below would show that the optimal recovery map Δ^{aff} can be taken linear if the $\mathcal{K}_n$'s are not only convex but also symmetric about the origin.

Proposition 3. Given the model set (2) based on convex sets $\mathcal{K}_1, \dots, \mathcal{K}_N$, the observation map (3), and the quantity of interest (4), the worst-case error of a recovery map $\Delta : \mathbb{R}^{M \times N} \to \mathbb{R}^N$, measured in ℓ_p for $p \in [1, \infty]$ as

$$\mathrm{gwce}_p(\Delta) := \sup_{\mathbf{f} \in \boldsymbol{\mathcal{K}}^{\mathrm{ind}}} \big\|\Gamma(\mathbf{f}) - \Delta(\Lambda \mathbf{f})\big\|_{\ell_p^N},$$

is minimized for the affine map $\Delta^{\mathrm{aff}} : \mathbb{R}^{M \times N} \to \mathbb{R}^N$ given by

$$\Delta^{\mathrm{aff}}(y) = [\Delta_1^{\mathrm{aff}}(y_1); \dots; \Delta_N^{\mathrm{aff}}(y_N)] \in \mathbb{R}^N, \qquad y = [y_1, \dots, y_N] \in \mathbb{R}^{M \times N}, \tag{5}$$

each $\Delta_n^{\mathrm{aff}} : \mathbb{R}^M \to \mathbb{R}$ being an affine optimal recovery map for the estimation of γ_n relative to $\mathcal{K}_n$.

Proof. A well-known lower bound for the worst-case error of any recovery map Δ involves the symmetrized set $\boldsymbol{\mathcal{K}}_{\mathrm{sym}}^{\mathrm{ind}} := (\boldsymbol{\mathcal{K}}^{\mathrm{ind}} - \boldsymbol{\mathcal{K}}^{\mathrm{ind}})/2$. It is obtained by considering any $\mathbf{h} \in \ker(\Lambda) \cap \boldsymbol{\mathcal{K}}_{\mathrm{sym}}^{\mathrm{ind}}$, writing $\mathbf{h} = (\mathbf{f}^+ - \mathbf{f}^-)/2$ with $\mathbf{f}^+, \mathbf{f}^- \in \boldsymbol{\mathcal{K}}^{\mathrm{ind}}$ and $\Lambda(\mathbf{f}^+) = \Lambda(\mathbf{f}^-) =: y$, and deriving

$$\mathrm{gwce}_p(\Delta) \ge \frac{1}{2}\big\|\Gamma(\mathbf{f}^+) - \Delta(y)\big\|_{\ell_p^N} + \frac{1}{2}\big\|\Gamma(\mathbf{f}^-) - \Delta(y)\big\|_{\ell_p^N} \ge \frac{1}{2}\big\|\Gamma(\mathbf{f}^+) - \Gamma(\mathbf{f}^-)\big\|_{\ell_p^N} = \big\|\Gamma(\mathbf{h})\big\|_{\ell_p^N}.$$

Taking the supremum over $\mathbf{h}$ results in the said lower bound, which reads

$$\mathrm{gwce}_p(\Delta) \geq \sup_{\substack{\mathbf{h} \in \ker(\Lambda) \\ \mathbf{h} \in \boldsymbol{\mathcal{K}}_{\mathrm{sym}}^{\mathrm{ind}}}} \|\Gamma(\mathbf{h})\|_{\ell_p^N}.$$

Noticing that $\mathbf{h} \in \ker(\Lambda)$ if and only if $h_1 \in \ker(\lambda), \ldots, h_N \in \ker(\lambda)$, that $\mathbf{h} \in \boldsymbol{\mathcal{K}}_{\mathrm{sym}}^{\mathrm{ind}}$ if and only if $h_1 \in (\mathcal{K}_1 - \mathcal{K}_1)/2, \ldots, h_N \in (\mathcal{K}_N - \mathcal{K}_N)/2$, and that $\|\Gamma(\mathbf{h})\|_{\ell_p^N}^p = \sum_{n=1}^N |\gamma_n(h_n)|^p$, one arrives at

$$\mathrm{gwce}_p(\Delta)^p \geq \sup_{\substack{h_n \in \ker(\lambda), \text{ all } n \\ h_n \in (\mathcal{K}_n - \mathcal{K}_n)/2, \text{ all } n}} \sum_{n=1}^N |\gamma_n(h_n)|^p = \sum_{n=1}^N \sup_{\substack{h_n \in \ker(\lambda) \\ h_n \in (\mathcal{K}_n - \mathcal{K}_n)/2}} |\gamma_n(h_n)|^p.$$

According to (i-b) of Proposition 13 and its proof (the argument behind the fact that convex model sets ensures optimality of affine maps, which is found in the appendix), each summand—power p omitted—coincides with the worst-case error of an affine optimal recovery map Δ_n^{aff} for the estimation of γ_n relative to $\mathcal{K}_n$. It follows that

$$\begin{aligned}\mathrm{gwce}_p(\Delta)^p &\geq \sum_{n=1}^N \sup_{f_n \in \mathcal{K}_n} |\gamma_n(f_n) - \Delta_n^{\mathrm{aff}}(\lambda(f_n))|^p = \sup_{f_n \in \mathcal{K}_n, \text{ all } n} \sum_{n=1}^N |\gamma_n(f_n) - \Delta_n^{\mathrm{aff}}(\lambda(f_n))|^p \\ &= \sup_{\mathbf{f} \in \boldsymbol{\mathcal{K}}^{\mathrm{ind}}} \|\Gamma(\mathbf{f}) - \Delta^{\mathrm{aff}}(\Lambda \mathbf{f})\|_{\ell_p^N}^p = \mathrm{gwce}_p(\Delta^{\mathrm{aff}})^p.\end{aligned}$$

This inequality establishes the announced optimality of Δ^{aff} defined in (5). Of note, the input of each component Δ_n^{aff} consists of observations about f_n alone, not about $f_1, \ldots, f_{n-1}, f_{n+1}, \ldots, f_N$—as one would have anticipated from the independence of $f_1, \ldots, f_N$. □

2.3 Existence result in case of dependent components

The case of independent components having been brushed aside, this subsection is concerned with a model set that includes dependence between the components of a multiobject $\mathbf{f} \in F^N$—again it need not be pinned down as a multivalued function yet. Generalizing the relation $f_1 + \cdots + f_N = 1$, one introduces several dependence relations of the type

$$a_{\ell,1} f_1 + \cdots + a_{\ell,N} f_N = b_\ell, \qquad \ell \in [1:L],$$

which are summarized as $A\mathbf{f} = \mathbf{b}$ for fixed $A \in \mathbb{R}^{L \times N}$ and $\mathbf{b} \in F^L$. As such, the contemplated model set takes the form

$$\boldsymbol{\mathcal{K}}^{\mathrm{dep}} := \big\{\mathbf{f} = [f_1; \ldots; f_N] \in F^N : f_1 \in \mathcal{K}_1, \ldots, f_N \in \mathcal{K}_N, A\mathbf{f} = \mathbf{b}\big\} \subseteq F^N, \tag{6}$$

or in short $\boldsymbol{\mathcal{K}}^{\mathrm{dep}} = \boldsymbol{\mathcal{K}}^{\mathrm{ind}} \cap A^{-1}(\{\mathbf{b}\})$. Note that this model set is convex as soon as the sets $\mathcal{K}_1, \ldots, \mathcal{K}_N$ are themselves convex. According to (i) of Proposition 13 in the appendix, with the same setting as Proposition 3, one can already guarantee that affine maps are near optimal with a factor $2\sqrt{N}$ (and even better when $p > 2$). For $p = \infty$, genuine optimality can actually be achieved—although not constructively at this point—as established below.

Proposition 4. Given the model set (6) based on convex sets $\mathcal{K}_1, \ldots, \mathcal{K}_N$, the observation map (3), and the quantity of interest (4), the worst-case error of a recovery map $\Delta : \mathbb{R}^{M\times N} \to \mathbb{R}^N$, measured in ℓ_∞, is minimized for an affine recovery map.

Proof. The worst-case error of any recovery map $\Delta = [\Delta_1; \ldots; \Delta_N]$ is

$$\begin{aligned}\mathrm{gwce}_\infty(\Delta) &= \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} \|\Gamma(\mathbf{f}) - \Delta(\Lambda\mathbf{f})\|_{\ell_\infty^N} = \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} \max_{n\in[1:N]} |\gamma_n(f_n) - \Delta_n(\Lambda\mathbf{f})| \\ &= \max_{n\in[1:N]} \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} |\gamma_n(f_n) - \Delta_n(\Lambda\mathbf{f})| = \max_{n\in[1:N]} \mathrm{gwce}(\Delta_n).\end{aligned}$$

If $\Delta_n^{\mathrm{aff}} : \mathbb{R}^{M\times N} \to \mathbb{R}$ denotes an optimal recovery map for the estimation of the linear functional $\mathbf{f} \in F^N \mapsto \gamma_n(f_n) \in \mathbb{R}$ relative to the convex set $\boldsymbol{\mathcal{K}}^{\mathrm{dep}}$ as a whole, recalling that Δ_n^{aff} can be taken as an affine map, then one has

$$\mathrm{gwce}_\infty(\Delta) \geq \max_{n\in[1:N]} \mathrm{gwce}(\Delta_n^{\mathrm{aff}}) = \mathrm{gwce}_\infty(\Delta^{\mathrm{aff}}),$$

where $\Delta^{\mathrm{aff}} : \mathbb{R}^{M\times N} \to \mathbb{R}^N$ is simply defined as $\Delta^{\mathrm{aff}} := [\Delta_1^{\mathrm{aff}}; \ldots; \Delta_N^{\mathrm{aff}}]$. Since the latter is an affine map, the result is now justified. But, contrary to Proposition 3, each Δ_n^{aff} is defined on all of $\mathbb{R}^{M\times N}$, i.e., it could take inputs obtained from observing components other than f_n. □

3 Constructive Results in Case of Dependent Components

A constructive version of the above result is derived in this section, specifically in Subsection 3.2. Before that, one isolates in Subsection 3.1 two useful observations: one is needed for the proof of the main theorem and the other for the proof of one of two consequences presented in Subsection 3.3. The results throughout this section are still rather abstract—they will be instantiated to more practical situations in the next section.

3.1 The key lemmas

The upcoming Theorem 8 and Theorem 10 rely on Lemma 5 and Lemma 6, respectively. These two lemmas, stated right below, are in fact byproducts of an encompassing theorem, so their proofs are deferred after the statement and justification of the theorem.

Lemma 5. Given a convex subset $\mathcal{C}$ of a vector space F, a linear map $\mathcal{A}$ from F into another vector space G, an element $b \in G$, and a linear functional $\eta \in F^*$, one has

$$\sup_{f\in F} \Big\{\eta(f) : f \in \mathcal{C} \text{ and } \mathcal{A}(f) = b\Big\} = \min_{\mu\in G^*} \Big\{\mu(b) + \sup_{f\in\mathcal{C}} (\eta - \mu\circ\mathcal{A})(f)\Big\}.$$

Lemma 6. Given a convex set $\mathcal{K}$ and a convex cone $\mathcal{C}$ in a vector space F and given a linear functional $\eta \in F^*$, one has

$$\sup_{f \in F} \Big\{ \eta(f) : f \in \mathcal{K} \cap \mathcal{C} \Big\} = \min_{\nu \in F^*} \Big\{ \sup_{f \in \mathcal{K}} \nu(f) : \ \nu - \eta \in \mathcal{C}^{\mathrm{dual}} \Big\},$$

where $\mathcal{C}^{\mathrm{dual}} := \{\mu \in F^* : \mu(f) \geq 0 \text{ for all } f \in F\}$ denotes the dual cone of $\mathcal{C}$.

The encompassing functional analytic result behind these lemmas is stated as Theorem 7 below. It is likely known—the arguments are certainly standard—but a proof is included for completeness.

Theorem 7. Given convex subsets $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of a vector space F and a linear functional $\eta \in F^*$, one has

$$\sup_{f \in F} \Big\{ \eta(f) : \ f \in \bigcap\nolimits_{k=1}^K \mathcal{C}_k \Big\} = \min_{\mu_1, \ldots, \mu_K \in F^*} \Big\{ \sum\nolimits_{k=1}^K \sup_{f_k \in \mathcal{C}_k} \mu_k(f_k) : \sum\nolimits_{k=1}^K \mu_k = \eta \Big\}. \tag{7}$$

Proof. Let lhs denote the supremum on the left-hand side of (7) and let rhs denote the infimum (proved to be a minimum later) on its right-hand side. The inequality rhs $\geq$ lhs is easily verified. Indeed, it suffices to observe that, for feasible $\mu_1, \ldots, \mu_K \in F^*$ and $f \in F$,

$$\sum\nolimits_{k=1}^K \sup_{f_k \in \mathcal{K}} \mu_k(f_k) \geq \sum\nolimits_{k=1}^K \mu_k(f) = \eta(f).$$

The reverse inequality requires more work. One starts by recalling that the Minkowski functional (aka gauge) associated with a subset $\mathcal{C}$ of F is defined by $\rho_{\mathcal{C}}(f) := \inf\{t > 0 : f \in t\,\mathcal{C}\} \in [0, +\infty]$ for any $f \in F$. Recall also that, if $\mathcal{C}$ is convex and $0 \in \mathcal{C}$, then $\rho_{\mathcal{C}}$ is a sublinear functional, meaning that $\rho_{\mathcal{C}}(tf) = t\,\rho_{\mathcal{C}}(f)$ and $\rho_{\mathcal{C}}(f+g) \leq \rho_{\mathcal{C}}(f) + \rho_{\mathcal{C}}(g)$ for all $t > 0$ and all $f, g \in F$. In particular, selecting an arbitrary $\overline{f} \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_K$, the Minkowski functional associated with the set $\mathcal{C} := (\mathcal{C}_1 - \overline{f}) \cap \cdots \cap (\mathcal{C}_K - \overline{f})$ is sublinear. The latter is (easily verified to be) given, for any $f \in F$, by $\rho_{\mathcal{C}}(f) = \max\{\rho_{\mathcal{C}_1 - \overline{f}}(f), \ldots, \rho_{\mathcal{C}_K - \overline{f}}(f)\}$. With lhs still denoting the supremum on the left-hand side of (7), one notices that $\eta(\overline{f} + h) \leq$ lhs, or $\eta(h) \leq \text{lhs} - \eta(\overline{f})$, for all $h \in \mathcal{C}$. This implies that $\eta(h) \leq [\text{lhs} - \eta(\overline{f})]\,\rho_{\mathcal{C}}(h)$ for all $h \in H$. Therefore, the linear functional ν defined on the subspace $S = \{[h; \ldots; h], h \in F\}$ of F^K by $\nu([h; \ldots; h]) = \eta(h)$ is dominated on S by the sublinear functional ρ defined on F^K by $\rho([h_1; \ldots; h_K]) = [\text{lhs} - \eta(\overline{f})]\, \max\{\rho_{\mathcal{C}_1 - \overline{f}}(h_1), \ldots, \rho_{\mathcal{C}_K - \overline{f}}(h_K)\}$. The dominated extension theorem now ensures the existence of a linear functional μ defined on F^K such that $\mu_{|S} = \nu$ and $\mu \leq \rho$. Such a linear functional takes the form $\mu([h_1; \ldots; h_K]) = \mu_1(h_1) + \cdots + \mu_K(h_K)$ for some linear functionals $\mu_1, \ldots, \mu_K \in F^*$. The fact that $\mu_{|S} = \nu$ implies that $\mu_1 + \cdots + \mu_K = \eta$. As for the fact that $\mu \leq \rho$, it implies that $\mu_1(h_1) + \cdots + \mu_K(h_K) \leq [\text{lhs} - \eta(\overline{f})]\, \max\{\rho_{\mathcal{C}_1 - \overline{f}}(h_1), \ldots, \rho_{\mathcal{C}_K - \overline{f}}(h_K)\}$ for all $[h_1; \ldots; h_K] \in F^K$. Thus, given $f_k \in \mathcal{C}_k$ for $k \in [1:K]$, since $h_k := f_k - \overline{f} \in \mathcal{C}_k - \overline{f}$ satisfies $\rho_{\mathcal{C}_k - \overline{f}}(h_k) \leq 1$, one obtains $\mu_1(f_1 - \overline{f}) + \cdots + \mu_K(f_K - \overline{f}) \leq [\text{lhs} - \eta(\overline{f})]$. After simplifying and taking the suprema over $f_1, \ldots, f_K$, one arrives at

$$\sup_{f_1 \in \mathcal{C}_1} \mu_1(f_1) + \cdots + \sup_{f_K \in \mathcal{C}_K} \mu_K(f_K) \leq \text{lhs},$$

which is the inequality required to complete the proof. □

With Theorem 7 now justified, it remains to deduce Lemmas 5 and 6 as consequences.

Proof of Lemma 5. Applying Theorem 7 with $K = 2$, $\mathcal{C}_1 = \mathcal{A}^{-1}(\{b\})$, and $\mathcal{C}_2 = \mathcal{C}$ shows that the sought-after supremum equals

$$\min_{\nu \in F^*} \left\{ \sup_{f_1 \in \mathcal{A}^{-1}(\{b\})} \nu(f_1) + \sup_{f_2 \in \mathcal{C}} (\eta - \nu)(f_2) \right\}.$$

Selecting some $\overline{f} \in \mathcal{A}^{-1}(\{b\})$, one observes that

$$\sup_{f_1 \in \mathcal{A}^{-1}(\{b\})} \nu(f_1) = \sup_{h \in \ker(\mathcal{A})} \nu(\overline{f} + h) = \begin{cases} \nu(\overline{f}) & \text{if } \ker(\mathcal{A}) \subseteq \ker(\nu), \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, the minimum over $\nu \in F^*$ will be achieved by imposing $\ker(\mathcal{A}) \subseteq \ker(\nu)$, which is equivalent to the existence of $\mu \in G^*$ such that $\nu = \mu \circ \mathcal{A}$, in which case $\nu(\overline{f}) = \mu(b)$. The announced result immediately follows. □

Proof of Lemma 6. Again applying Theorem 7 with $K = 2$, and this time with $\mathcal{C}_1 = \mathcal{K}$ and $\mathcal{C}_2 = \mathcal{C}$, shows that the sought-after supremum equals

$$\min_{\nu \in F^*} \left\{ \sup_{f_1 \in \mathcal{K}} \nu(f_1) + \sup_{f_2 \in \mathcal{C}} (\eta - \nu)(f_2) \right\}.$$

Using the fact that $\mathcal{C}$ is a cone, i.e. that if $f \in \mathcal{C}$, then $tf \in \mathcal{C}$ for all $t > 0$, one observes that

$$\sup_{f_2 \in \mathcal{C}} (\eta - \nu)(f_2) = \begin{cases} 0 & \text{if } -(\eta - \nu) \in \mathcal{C}^{\text{dual}}, \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, the minimum over $\nu \in F^*$ will be achieved by imposing $\nu - \eta \in \mathcal{C}^{\text{dual}}$. The announced result immediately follows. □

3.2 Abstract formulation of the main result

It is now time to state the main theorem about optimal recovery of multiobjects whose dependent components belong to convex model sets. Its formulation involves the support function of a set $\mathcal{C} \subseteq F$, as defined by

$$\|\eta\|_{\mathcal{C}} = \sup_{f \in \mathcal{C}} \eta(f), \qquad \eta \in F^*.$$

Beware that the notation may be misleading: the support function is not necessarily a norm, as $\|-\eta\|_{\mathcal{C}} = \|\eta\|_{\mathcal{C}}$ might not hold in general. Nonetheless, it is a sublinear functional—indeed, it is straightforward to verify that it coincides with the Minkowski functional associated with the polar

$\mathcal{C}^\circ := \{\eta \in F^* : \eta(f) \le 1 \text{ for all } f \in \mathcal{C}\}$ of $\mathcal{C}$, which is always convex and contains the origin. Despite its abstract form, the result is constructive, in the sense that an affine optimal recovery map is obtained after solving several convex optimization programs. Note that formally taking $L = 0$ (hence discarding $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$) also gives a constructive version of Proposition 3 for the case of independent components.

Theorem 8. In the setting of Proposition 4, one considers for each $j \in [1 : N]$ some solutions $\widehat{c}_j = \left[\widehat{c}_j^{(m,n)}\right]_{\substack{m=1,\dots,M\\ n=1,\dots,N}} \in \mathbb{R}^{M\times N}$ and $\widehat{d}_j \in \mathbb{R}$ to the convex optimization program

$$\underset{\substack{\boldsymbol{\mu}^+,\boldsymbol{\mu}^-\in(F^*)^L\\ c\in\mathbb{R}^{M\times N},\, d,e\in\mathbb{R}}}{\text{minimize}} \quad e \qquad \text{s.to} \quad \sum_{\ell=1}^{L} \mu_\ell^\pm(b_\ell) + \sum_{n=1}^{N} \Big\| \pm\delta_{j,n}\gamma_j \mp \sum_{m=1}^{M} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^\pm \Big\|_{\mathcal{K}_n} \le e \pm d. \tag{8}$$

Then an optimal recovery map is given by the affine map

$$\Delta^{\mathrm{aff}} : y \in \mathbb{R}^{M\times N} \mapsto \sum_{m=1}^{M}\sum_{n=1}^{N} y_{m,n}\widehat{c}^{(m,n)} + \widehat{d} \in \mathbb{R}^N. \tag{9}$$

Proof. According to Proposition 4 (and its proof), one can look for an optimal recovery map in the form $\Delta^{\mathrm{aff}} = [\Delta_1^{\mathrm{aff}}; \dots; \Delta_N^{\mathrm{aff}}]$. Fixing $j \in [1 : N]$, the affine map $\Delta_j^{\mathrm{aff}} : \mathbb{R}^{M\times N} \to \mathbb{R}$ can be written as

$$\Delta_j^{\mathrm{aff}}(y) = \sum_{m=1}^{M}\sum_{n=1}^{N} y_{m,n}c_{m,n} + d$$

for some $c \in \mathbb{R}^{M\times N}$ and $d \in \mathbb{R}$. These coefficients are sought to minimize the worst-case error

$$\mathrm{gwce}(\Delta_j^{\mathrm{aff}}) = \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} \Big|\gamma_j(f_j) - \sum_{m=1}^{M}\sum_{n=1}^{N} c_{m,n}\lambda_m(f_n) - d\Big| = \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} \big|\boldsymbol{\eta}^c(\mathbf{f}) - d\big|,$$

with a linear functional $\boldsymbol{\eta}^c$ implicitly defined on F^N by $\boldsymbol{\eta}^c(\mathbf{f}) = \gamma_j(f_j) - \sum_{m=1}^{M}\sum_{n=1}^{N} c_{m,n}\lambda_m(f_n)$. At this point, noticing that

$$\begin{aligned} \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} \big|\boldsymbol{\eta}^c(\mathbf{f}) - d\big| &= \inf_{e\in\mathbb{R}} e \quad \text{s.to} \ -e \le \boldsymbol{\eta}^c(\mathbf{f}) - d \le e \text{ for all } \mathbf{f} \in \boldsymbol{\mathcal{K}}^{\mathrm{dep}} \\ &= \inf_{e\in\mathbb{R}} e \quad \text{s.to} \begin{cases} \sup\limits_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} (+\boldsymbol{\eta}^c)(\mathbf{f}) \le e + d, \\ \sup\limits_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{dep}}} (-\boldsymbol{\eta}^c)(\mathbf{f}) \le e - d, \end{cases} \end{aligned} \tag{10}$$

it remains to invoke Lemma 5 to transform these two suprema into minima over $\boldsymbol{\mu}^+, \boldsymbol{\mu}^- \in (F^L)^*$. More precisely, the first supremum is

$$\sup_{\mathbf{f}\in F^N} \Big\{(+\boldsymbol{\eta}^c)(\mathbf{f}) : \mathbf{f} \in \boldsymbol{\mathcal{K}}^{\mathrm{ind}} \text{ and } A\mathbf{f} = \mathbf{b}\Big\} = \min_{\boldsymbol{\mu}^+\in(F^L)^*} \Big\{\boldsymbol{\mu}^+(\mathbf{b}) + \sup_{\mathbf{f}\in\boldsymbol{\mathcal{K}}^{\mathrm{ind}}} (+\boldsymbol{\eta}^c - \boldsymbol{\mu}^+ \circ A)\mathbf{f}\Big\}.$$

The linear functional $\boldsymbol{\mu}^+ \in (F^L)^*$ decomposes through $\mu_1^+, \dots, \mu_L^+ \in F^*$ as $\boldsymbol{\mu}(\mathbf{b}) = \sum_{\ell=1}^L \mu_\ell^+(b_\ell)$, so $\boldsymbol{\mu}^+(A\mathbf{f}) = \sum_{\ell=1}^L \mu_\ell^+\Big(\sum_{n=1}^N a_{\ell,n} f_n\Big) = \sum_{n=1}^N \Big(\sum_{\ell=1}^L a_{\ell,n}\mu_\ell^+\Big)(f_n)$, and similarly the linear functional $\boldsymbol{\eta}^c \in (F^N)^*$ decomposes through $\eta_1^c, \dots, \eta_N^c \in F^*$ as $\boldsymbol{\eta}^c(\mathbf{f}) = \sum_{n=1}^N \eta_n^c(f_n)$. Thus, the first constraint in (10)—the one with the + sign—reads: there exist $\mu_1^+, \dots, \mu_L^+ \in F^*$ such that

$$\sum_{\ell=1}^L \mu_\ell^+(b_\ell) + \sup_{f_1 \in \mathcal{K}_1, \dots, f_N \in \mathcal{K}_N} \sum_{n=1}^N \Big(+ \eta_n^c - \sum_{\ell=1}^L a_{\ell,n}\mu_n^+\Big)(f_n) \le e + d,$$

in other words, taking the expression of η_n^c into account,

$$\sum_{\ell=1}^L \mu_\ell^+(b_\ell) + \sum_{n=1}^N \sup_{f_n \in \mathcal{K}_n} \Big(+ \delta_{j,n}\gamma_j - \sum_{m=1}^M c_{m,n}\lambda_m - \sum_{\ell=1}^L a_{\ell,n}\mu_n^+\Big)(f_n) \le e + d.$$

This is the constraint appearing in (8) for the choice $\pm = +$. Likewise, the second constraint in (10) reduces to the constraint appearing in (8) for the choice $\pm = -$. Incorporating the $\mu_\ell^\pm$'s as optimization variables exposes $\mathrm{gwce}(\Delta_j^{\mathrm{aff}})$ as the minimal value of a convex program. Further minimizing over $c \in \mathbb{R}^{M \times N}$ and $d \in \mathbb{R}$ leads to the convex program (8). Its minimizers provide an optimal recovery map via (9), but keep also in mind that its minimal value is equal to the minimal worst-case error over all recovery maps. □

3.3 Two particular cases of the main result

The practical solvability of the convex optimization program (8) depends on the amenability to computations of the support functions of the model sets $\mathcal{K}_1, \dots, \mathcal{K}_N$. As such, it is worth looking at two types of particularly relevant sets, still without specifying the vector space F at this point.

Approximability sets. As alluded to in the introduction, model sets based on approximation capabilities are quite pertinent, since some assumptions about the approximation power of certain hypothesis classes are often made implicitly. Thus, it is natural to consider model sets of the type $\{f \in F : \mathrm{dist}_F(f, V) \le \varepsilon\}$ relative to a linear subspace V of F and a parameter $\varepsilon \ge 0$. Actually, one considers a more general model set involving an invertible operator T on F, namely

$$\mathcal{K}_{V,T} = \{f \in F : \mathrm{dist}_F(Tf, V) \le 1\}. \tag{11}$$

Applying Theorem 8 then entails to determining the support function of such a set, which is done as follows:

$$\begin{aligned} \|\eta\|_{\mathcal{K}_{V,T}} &= \sup_{f \in F}\{\eta(f) : \text{there exists } v \in V \text{ such that } \|Tf - v\|_F \le 1\} \\ &= \sup_{\substack{f \in F \\ v \in V}}\{\eta \circ T^{-1}(Tf) : \|Tf - v\|_F \le \varepsilon\} = \sup_{\substack{g \in F \\ v \in V}}\{\eta \circ T^{-1}(g) + \eta \circ T^{-1}(v) : \|g\|_F \le 1\} \\ &= \sup_{g \in F}\{\eta \circ T^{-1}(g) : \|g\|_F \le 1\} + \sup_{v \in V}\{\eta \circ T^{-1}(v)\} = \begin{cases} \|\eta \circ T^{-1}\|_{F^*} & \text{if } (\eta \circ T^{-1})_{|V} = 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

This observation leads to the following consequence of Theorem 8 (where the change $\boldsymbol{\mu}^- \leftrightarrow -\boldsymbol{\mu}^-$ of optimization variable is applied). It should be viewed as a generalization to $L \geq 1$ and $T \neq (1/\varepsilon)\mathrm{Id}_F$ of [2, Theorem 3.1].

Theorem 9. For model sets $\mathcal{K}_n = \mathcal{K}_{V_n,T_n}$, the affine optimal recovery map of Theorem 8 is constructed by solving, for each $j \in [1:N]$, the constrained convex optimization program

$$\underset{\substack{\boldsymbol{\mu}^+,\boldsymbol{\mu}^- \in (F^*)^L \\ c \in \mathbb{R}^{M\times N},\, d,e \in \mathbb{R}}}{\text{minimize}} \quad e \quad \text{s.to} \quad \pm \sum_{\ell=1}^{L} \mu_\ell^\pm(b_\ell) + \sum_{n=1}^{N} \left\| \left(\delta_{j,n}\gamma_j - \sum_{m=1}^{M} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^\pm \right) \circ T_n^{-1} \right\|_{F^*} \leq e \pm d$$

$$\text{and} \quad \left(\delta_{j,n}\gamma_j - \sum_{m=1}^{N} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^\pm \right)(T_n^{-1}(v)) = 0 \ \text{ for all } v \in V_n, n \in [1:N].$$

Model sets intersected with convex cones. The above result applies to the recovery of genuine multivalued functions satisfying $f_1 + \cdots + f_N = 1$, but does not yet incorporate the nonnegativity constraints $f_n \geq 0$. To this end, one considers more generally model sets of the type $\mathcal{K} \cap \mathcal{C}$, i.e., intersections of a convex model set $\mathcal{K}$ (e.g. an approximability set) with a convex cone $\mathcal{C}$ (e.g. the set of nonnegative functions). Thanks to Lemma 6, as soon as $\mathcal{K}$ and $\mathcal{C}$ are amenable to computations, the program (8) turns into the manageable convex program presented below.

Theorem 10. For model sets of the form $\mathcal{K}_n \cap \mathcal{C}_n$, where both $\mathcal{K}_n$ and $\mathcal{C}_n$ are convex and $\mathcal{C}_n$ is a cone, the affine optimal recovery map of Theorem 8 is constructed by solving, for each $j \in [1:N]$, the constrained convex optimization program

$$\underset{\substack{\boldsymbol{\mu}^+,\boldsymbol{\mu}^- \in (F^*)^L \\ \boldsymbol{\nu}^+,\boldsymbol{\nu}^- \in (F^*)^N \\ c \in \mathbb{R}^{M\times N},\, d,e \in \mathbb{R}}}{\text{minimize}} \quad e \quad \text{s.to} \quad \pm \sum_{\ell=1}^{L} \mu_\ell^\pm(b_\ell) + \sum_{n=1}^{N} \|\nu_n^\pm\|_{\mathcal{K}_n} \leq e \pm d$$

$$\text{and} \quad \nu_n^\pm \mp \left(\delta_{j,n}\gamma_j - \sum_{m=1}^{M} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^\pm \right) \in \mathcal{C}_n^{\mathrm{dual}} \quad \text{for all } n \in [1:N].$$

Proof. For each $\eta_n^\pm := \pm\delta_{j,n}\gamma_j \mp \sum_{m=1}^{M} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^\pm$ appearing in Theorem 8, the support function of $\mathcal{K}_n \cap \mathcal{C}_n$ evaluated at $\eta_n^\pm$ is transformed, according to Lemma 6, into

$$\begin{aligned} \|\eta_n^\pm\|_{\mathcal{K}_n\cap\mathcal{C}_n} &= \sup_{f_n \in \mathcal{K}_n\cap\mathcal{C}_n} \eta_n^\pm(f_n) = \min_{\nu_n^\pm \in F^*} \left\{ \sup_{f_n \in \mathcal{K}_n} \nu_n^\pm(f_n) : \nu_n^\pm - \eta_n^\pm \in \mathcal{C}_n^{\mathrm{dual}} \right\} \\ &= \min_{\nu_n^\pm \in F^*} \left\{ \|\nu_n^\pm\|_{\mathcal{K}_n} : \nu_n^\pm - \eta_n^\pm \in \mathcal{C}_n^{\mathrm{dual}} \right\}. \end{aligned}$$

Incorporating the $\nu_n^\pm$'s as optimization variables in (8) while also making the change $\boldsymbol{\mu}^- \leftrightarrow -\boldsymbol{\mu}^-$ leads to the announced optimization program. □

4 Instantiation of the Main Results

In this subsection, the abstract formalism adopted so far is specified in two situations of practical interest, leading to the two theorems highlighted in the introduction. These situations are that of a (reproducing kernel) Hilbert space and of a space of continuous functions.

4.1 Hilbert spaces

One assumes in this subsection that F is a Hilbert space—hence the notation H is used instead. In this situation, any linear functional $\mu \in H^*$ is identified with its Riesz representer $u \in H$ via $\mu(f) = \langle u, f \rangle$ for all $f \in H$. In particular, the linear functionals $\lambda_m \in H^*$ and $\gamma_j \in H^*$ are identified with vectors $w_m \in H$ and $g_j \in H$. Thus, when $\lambda_m = \delta_{x^{(m)}}$ and $\gamma_j = \delta_{x^{(0)}}$ are point evaluations in a reproducing kernel Hilbert space with kernel K, one has $w_m = K(x^{(m)}, \cdot)$ and $g_j = K(x^{(0)}, \cdot)$. Furthermore, the determination (12) of the support function of the set $\mathcal{K}_{V,T}$ defined in (11), evaluated at a linear functional identified with $u \in H$, easily reduces to

$$\|\!|u|\!\|_{\mathcal{K}_{V,T}} = \begin{cases} \|T^{-*}u\|_H & \text{if } T^{-*}u \in V^\perp, \\ +\infty & \text{otherwise.} \end{cases}$$

Taking this observation into account, Theorem 9 yields the corollary below, of which Theorem 1 is the special case $T_n = (1/\varepsilon_n)\mathrm{Id}_H$, $L = 1$, $A = [1; \dots; 1]$, and $b = 1$ (again assuming that H is a reproducing kernel Hilbert space containing the constant function 1).

Corollary 11. In a Hilbert space H, if the model sets are $\mathcal{K}_{V_n,T_n} = \{f \in H : \mathrm{dist}_H(T_n f, V_n) \le 1\}$, then the affine optimal recovery map of Theorem 8 is constructed by solving, for each $j \in [1:N]$, the convex optimization program

$$\begin{aligned} \underset{\substack{\mathbf{u}^+,\mathbf{u}^- \in H^L \\ c \in \mathbb{R}^{M\times N},\, d,e \in \mathbb{R}}}{\text{minimize}} \quad & e \quad \text{s.to} \quad \pm \sum_{\ell=1}^{L} \langle u_\ell^\pm, b_\ell \rangle + \sum_{n=1}^{N} \Big\| T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^{M} c_{m,n} w_m - \sum_{\ell=1}^{L} a_{\ell,n} u_\ell^\pm\Big)\Big\|_H \le e \pm d \\ & \text{and} \quad \Big\langle T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^{M} c_{m,n} w_m - \sum_{\ell=1}^{L} a_{\ell,n} u_\ell^\pm\Big), v \Big\rangle = 0 \text{ for all } v \in V_n, n \in [1:N]. \end{aligned} \tag{13}$$

Should the norms be squared in the above, the optimization program would have a closed-form solution. Such a favorable outcome cannot be hoped for here, since (13) includes the geometric median problem as a special case and the latter is not known to possess closed-form solutions. Regardless, there is no difficulty in solving (13) computationally—it is a second-order cone program. However, it is not immediately clear how to handle an infinite-dimensional Hilbert space H, but fortunately a representer theorem holds, at least in some cases. Precisely, it is possible to replace the minimization over H by a minimization over a finite-dimensional subspace of H when the T_n's are of the form $T_n = (1/\varepsilon_n)T$ for a common operator T. To justify this, it is enough to remark that

if $u_1^\pm, \dots, u_L^\pm, c, d, e$ are feasible for (13), then so are $\widetilde{u}_1^\pm, \dots, \widetilde{u}_L^\pm, c, d, e$, where $\widetilde{u}_\ell^\pm := T^* P_{\widetilde{H}} T^{-*} u_\ell^\pm$ for $\ell \in [1:L]$ and where $P_{\widetilde{H}}$ denotes the orthogonal projector onto the finite-dimensional space

$$\widetilde{H} := \operatorname{span}\{Tb_1, \dots, Tb_L, T^{-*}g_j, T^{-*}w_1, \dots, T^{-*}w_M\} + V_1 + \cdots + V_N.$$

This remark follows from the fact that, since the Tb_ℓ's belong to $\widetilde{H}$,

$$\langle \widetilde{u}_\ell^\pm, b_\ell \rangle = \langle T^{-*}u_\ell^\pm, P_{\widetilde{H}} Tb_\ell \rangle = \langle T^{-*}u_\ell^\pm, Tb_\ell \rangle = \langle u_\ell^\pm, b_\ell \rangle,$$

from the fact that, since $T^{-*}g_j$ and the $T^{-*}w_m$'s belong to $\widetilde{H}$,

$$\begin{aligned}
&\Big\| T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} \widetilde{u}_\ell^\pm\Big)\Big\|_H = \varepsilon_n \Big\| T^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m\Big) - \sum_{\ell=1}^L a_{\ell,n} P_{\widetilde{H}} T^{-*} u_\ell^\pm \Big\|_H \\
&= \varepsilon_n \Big\| P_{\widetilde{H}} T^{-*}\big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\big)\Big\|_H \le \varepsilon_n \Big\| T^{-*}\big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\big)\Big\|_H \\
&= \Big\| T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\Big)\Big\|_H,
\end{aligned}$$

and from the fact that, since all $v \in V_n$, $n \in [1:N]$, belong to $\widetilde{H}$,

$$\begin{aligned}
&\Big\langle T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} \widetilde{u}_\ell^\pm\Big), v\Big\rangle = \varepsilon_n \Big\langle P_{\widetilde{H}} T^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\Big), v\Big\rangle \\
&= \varepsilon_n \Big\langle T^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\Big), v\Big\rangle = \Big\langle T_n^{-*}\Big(\delta_{j,n} g_j - \sum_{m=1}^M c_{m,n} w_m - \sum_{\ell=1}^L a_{\ell,n} u_\ell^\pm\Big), v\Big\rangle.
\end{aligned}$$

4.2 Spaces of continuous functions

One assumes in this subsection that F is the space $C(\mathcal{X})$ of continuous functions on a compact set $\mathcal{X}$. In this situation, the linear functionals $\mu \in C(\mathcal{X})^*$ are identified—keeping the same notation—with signed Borel measures $\mu \in B(\mathcal{X})$ via $\mu(f) = \int_{\mathcal{X}} f \, d\mu$ for all $f \in C(\mathcal{X})$. The attention is put on approximability sets intersected with the cone $\mathcal{C}$ of nonnegative functions. Here, the crucial ingredients are the facts that the norm of $\mu \in C(\mathcal{X})^*$ is the total variation $\int_{\mathcal{X}} d|\mu|$ of $\mu \in B(\mathcal{X})$ and that the dual cone $\mathcal{C}^{\mathrm{dual}}$ of $\mathcal{C}$ is the set $B_+(\mathcal{X})$ of nonnegative Borel measures. Thus, Theorem 10 yields the following corollary, of which Theorem 2 is the special case $L = 1$, $A = [1; \dots; 1]$, $b = 1$.

Corollary 12. In a space $C(\mathcal{X})$ of continuous functions in a compact set $\mathcal{X}$, if the model sets are $\mathcal{K}_n = \{f \in C(\mathcal{X}) : \operatorname{dist}_{C(\mathcal{X})}(f, V_n) \le \varepsilon_n, f \ge 0\}$, then the affine optimal recovery map of Theorem 8 is constructed by solving, for each $j \in [1:N]$, the convex optimization program

$$\begin{array}{ll}\underset{\substack{\boldsymbol{\mu}^+,\boldsymbol{\mu}^-\in B(\mathcal{X})^L\\ \boldsymbol{\nu}^+,\boldsymbol{\nu}^-\in B(\mathcal{X})^N\\ c\in\mathbb{R}^{M\times N},\, d,e\in\mathbb{R}}}{\text{minimize}} & e \\ & \text{s.to } \pm\sum_{\ell=1}^{L}\int_{\mathcal{X}} b_\ell\, d\mu_\ell^{\pm} + \sum_{n=1}^{N}\varepsilon_n\int_{\mathcal{X}} d|\nu_n^{\pm}| \le e \pm d \\ & \text{and } \int_{\mathcal{X}} v\, d\nu_n^{\pm} = 0 \quad \text{for all } v\in V_n,\ n\in[1:N], \\ & \text{and } \nu_n^{\pm} \mp \Big(\delta_{j,n}\gamma_j - \sum_{m=1}^{M} c_{m,n}\lambda_m - \sum_{\ell=1}^{L} a_{\ell,n}\mu_\ell^{\pm}\Big) \in B_+(\mathcal{X}), \quad n\in[1:N].\end{array}$$

Despite the convexity of the above program, its infinite dimensionality calls for a justification of its computational implementation. In a nutshell, it is a minimization over measures and as such it can be attacked via the Moment-SoS hierarchy (see [9] for a survey). The guiding arguments are sketched below in the simpler univariate setting, e.g. considering each $f_n(x)$ as the concentration of the nth constituent as time x evolves in $[0,\pi]$, say. For convenience, it is assumed that each space $V_n \in C[0,\pi]$ is made of cosine polynomials of degree at most $K(n)$ and that each $b_\ell \in C[0,\pi]$ is also a cosine polynomial.

First, one decomposes each optimization variable $\eta \in B(\mathcal{X})$ as the difference of its positive part $\eta^\oplus \in B_+(\mathcal{X})$ and its negative part $\eta^\ominus \in B_+(\mathcal{X})$, so that $\eta = \eta^\oplus - \eta^\ominus$ and $|\eta| = \eta^\oplus + \eta^\ominus$. Next, one thinks of the nonnegative Borel measure $\eta^\oplus$ equivalently in terms of its (infinite) sequence $y^\oplus \in \mathbb{R}^{\mathbb{N}}$ of trigonometric moments defined by $y_k^\oplus = \int_0^\pi \cos(k\theta) d\eta^\oplus(\theta)$, $k \ge 0$, provided that the (infinite) symmetric Toeplitz matrix $\mathrm{Toep}_\infty(y^\oplus)$ built from it is positive semidefinite (see [6, Section 3] for details). Then, expressing each $b_\ell \in C[0,\pi]$ as $b_\ell(\theta) = \sum_k b_{\ell,k}\cos(k\theta)$—this is particularly neat for the dependence relation $f_1 + \cdots + f_N = 1$—one arrives at the equivalent infinite-dimensional semidefinite program

$$\begin{array}{ll}\underset{\substack{\mathbf{u}^{+,\oplus},\mathbf{u}^{-,\oplus}\in(\mathbb{R}^{\mathbb{N}})^L\\ \mathbf{v}^{+,\oplus},\mathbf{v}^{-,\oplus}\in(\mathbb{R}^{\mathbb{N}})^N\\ c\in\mathbb{R}^{M\times N},\, d,e\in\mathbb{R}}}{\text{minimize}} & e \\ & \text{s.to } \pm\sum_{\ell=1}^{L}\sum_{k} b_{k,\ell}(u_{\ell,k}^{\pm,\oplus} - u_{\ell,k}^{\pm,\ominus}) + \sum_{n=1}^{N}\varepsilon_n(v_{n,0}^{\pm,\oplus} + v_{n,0}^{\pm,\ominus}) \le e \pm d \\ & \text{and } v_{n,k}^{\pm,\oplus} - v_{n,k}^{\pm,\ominus} = 0 \quad \text{for all } k\in[0:K(n)],\ n\in[1:N], \\ & \text{and } \mathrm{Toep}_\infty\Big(v_n^{\pm,\oplus} - v_n^{\pm,\ominus} \mp \delta_{j,n}y_j \pm \sum_{m=1}^{M} c_{m,n}z_m \pm \sum_{\ell=1}^{L} a_{\ell,n}(u_\ell^{\pm,\oplus} - u_\ell^{\pm,\ominus})\Big) \succeq 0, \\ & \text{and } \mathrm{Toep}_\infty\big(u_\ell^{\pm,\oplus}\big) \succeq 0, \quad \ell\in[1:L], \qquad \mathrm{Toep}_\infty\big(v_n^{\pm,\oplus}\big) \succeq 0, \quad n\in[1:N].\end{array} \tag{14}$$

Above, the fixed $y_j, z_1, \ldots, z_M \in \mathbb{R}^{\mathbb{N}}$ represented the (infinite) sequences of trigonometric moments stemmed from $\gamma_j, \lambda_1, \ldots, \lambda_M$. In the case $\gamma_j = \delta_{x^{(0)}}$ and $\lambda_m = \delta_{x^{(m)}}$, these moments are given by $y_{j,k} = \cos(kx^{(0)})$ and $z_{m,k} = \cos(kx^{(m)})$, $k \ge 0$. Finally, in order to deal with manageable semidefinite programs, one truncates the optimization variables $u_\ell^{\pm,\oplus}, v_n^{\pm,\oplus}$ at a level r, so that they belong to $\mathbb{R}^r$ rather than $\mathbb{R}^{\mathbb{N}}$, and one concurrently replaces the infinite Toeplitz matrices by their $r \times r$ upper-left sections. It can be shown that the minimizers of these truncated programs

converge to minimizers of the infinite-dimensional program as $r \to \infty$. This is the essence of the semidefinite hierarchy.

But even without being ensured of the convergence, the above procedure supplies a lower bound lb_r for the optimal value of the original program at each level r, since truncating an infinite sequence feasible for (14) yields a finite sequence feasible for the truncated program. In other words, one can compute a value $\mathrm{lb}_r \le \inf\{\mathrm{gwce}_\infty(\Delta),\, \Delta : \mathbb{R}^{M\times N} \to \mathbb{R}^N\}$. The associated minimizers $c^{[r,j]} \in \mathbb{R}^{M\times N}$ and $d^{[r,j]} \in \mathbb{R}$ allow one to construct an affine recovery map $\Delta^{[r]} = [\Delta_1^{[r]}; \dots; \Delta_N^{[r]}]$ via $\Delta_j^{[r]} : y \in \mathbb{R}^{M\times N} \mapsto \sum_{m=1}^M \sum_{n=1}^N c_{m,n}^{[r,j]} y_{m,n} + d^{[r,j]}$. Its worst-case error is the maximal optimal value of N convex optimization programs involving measures (stemming from the proof of Theorem 8 and left for the reader to spell out). An upper bound $\mathrm{ub}_{r,s}$ on $\mathrm{gwce}_\infty(\Delta^{[r]})$ can then be obtained by imposing these measures to be atomic measures on an s-point grid—hence turning the convex programs into a linear programs. In other words, one can compute a value $\mathrm{ub}_{r,s} \ge \mathrm{gwce}(\Delta^{[r]})$. All in all, one can now certify that

$$\mathrm{gwce}(\Delta^{[r]}) \le \frac{\mathrm{ub}_{r,s}}{\mathrm{lb}_r} \inf\{\mathrm{gwce}_\infty(\Delta),\, \Delta : \mathbb{R}^{M\times N} \to \mathbb{R}^N\},$$

i.e., that $\Delta^{[r]}$ is a near-optimal recovery map with factor $\mathrm{ub}_{r,s}/\mathrm{lb}_r$ that can be estimated *a posteriori* (and that should approach one as r and s grow).

5 Parting Thoughts

As far as the author is aware, the practical reconstruction of multivalued functions was not treated in the Optimal Recovery literature before. This article starts to fill the gap, but there are several open questions still to be investigated, some of them discussed below.

Other measures of performance. The main results of this article dealt with worst-case errors using the ℓ_∞^N-norm to quantify the prediction of $\mathbf{f}(x^{(0)}) \in \mathbb{R}^N$. What about ℓ_p^N for other $p \in [1,\infty)$, say for $p = 1$? By virtue of the inequalities $\|z\|_{\ell_\infty^N} \le \|z\|_{\ell_1^N} \le N\|z\|_{\ell_\infty^N}$ for $z \in \mathbb{R}^N$, one easily sees that the affine optimal recovery map relative to ℓ_∞^N is near-optimal relative to ℓ_1^N with factor N. But this can be improved: according to Proposition (i) of 13 in the appendix, and in view of the estimation $\mathrm{proj}(\ell_1^N) \lesssim \sqrt{2/\pi}\sqrt{N}$ for the projection constant of ℓ_1^N originally due to [8], the near-optimality factor for affine maps is of order at most $\sqrt{N}$. Can this be further reduced, say to a constant? Importantly, can one construct a genuinely optimal recovery map, affine or not?

Full Recovery. The article mostly tackled the prediction of multivalued functions at a fixed x, i.e., the recovery of $\Gamma = \delta_x$. The ultimate goal would be the recovery of $\Gamma = \mathrm{Id}_F$. It can be seen that assembling affine optimal maps for the recovery of each δ_x, $x \in \mathcal{X}$, provides an affine optimal map for the full recovery of $\Gamma = \mathrm{Id}_F$ if one works within $F = L_\infty(\mathcal{X}, \ell_\infty^N)$. It is conceivable that this assembly remains optimal within $F = C(\mathcal{X}, \ell_\infty^N)$—this was established, not without efforts, for $N = 1$ and approximability sets in [2]. Such an assembly is not a practical construction, though,

and more efforts will be needed to achieve one—in the case $N = 1$, this was done for specific approximability sets in [4]. As for full recovery within a Hilbert space H, the problem seems even more intricate, because it includes (with the choice $A\mathbf{f} = [f_1 - f_2; \cdots ; f_N - f_{N-1}]$ and $\mathbf{b} = 0$) the full recovery problem relative to a model set equal to the intersection of N hyperellipsoids, which has only been solved for $N = 2$ in [5].

Observation Errors. The article only studied exact observations of the type $y_{m,n} = f_n(x^{(m)})$, but in realistic situations, these observations are inaccurate, i.e., of the type $y_{m,n} = f_n(x^{(m)}) + e_{n,m}$. The error vectors $e_1, \ldots, e_N \in \mathbb{R}^M$ can be modeled stochastically as random vectors. In this case, considering the prediction problem, one can optimistically hope for near-optimality of affine maps, as obtained in [3] for Gaussian noise and in [7] even for log-concave noise with an altered measure of performance. The errors vectors $e_1, \ldots, e_N \in \mathbb{R}^M$ can also be modeled deterministically as belonging to uncertainty sets $\mathcal{E}_1, \ldots, \mathcal{E}_N$. In this case, there is a classical reduction to the accurate scenario by introducing the compound objects $[f_n; e_n] \in F \times \mathbb{R}^M$, about which one has the *a priori* information $[f_n; e_n] \in \mathcal{K}_n \times \mathcal{E}_n$ and $A\mathbf{f} = \mathbf{b}$, as well as the *a posteriori* information $\widetilde{\lambda}([\mathbf{f}; \mathbf{e}]) := \Lambda(\mathbf{f}) + \mathbf{e}$, and one aims at recovering the quantity of interest $\widetilde{\Gamma}([\mathbf{f}; \mathbf{e}]) := \Gamma(\mathbf{f})$. The abstract theory developed here provides constructions of optimal recovery maps, but may require additional work to turn them into practical constructions, for instance when the model and uncertainty sets are not independent—e.g., when the functions f_n represent concentrations, the observations $y_{m,n}$ are likely to be measured as concentrations, too, hence assumptions $0 \le f_n(x) \le 1$ and $|e_{n,m}| \le \eta$, say, need to be augmented with the $0 \le f_n(x^{(m)}) + e_{n,m} \le 1$.

Appendix

This section goes back one step and reconsiders the general Optimal Recovery framework described in Subsection 2.1. Proposition 13 below amalgamates, through one fresh take, some old and new results about near-optimality of affine recovery maps relative to convex model sets. Since it involves the notion of projection constants, one recalls that the relative projection constant of a subspace V of a norm space W is defined by

$$\mathrm{proj}(V, W) := \inf \big\{ \|P\| : \ P \text{ is a projection from } W \text{ onto } V \big\},$$

where P being a projection from W onto V means that P is a linear map from W into V such that $Pv = v$ for all $v \in V$. As for the absolute projection constant of a normed space U, it is defined by

$$\mathrm{proj}(U) := \sup\{\mathrm{proj}(i(U), W), \ i \text{ is an isometric embedding from } U \text{ into } W\}.$$

The absolute projection constant equals the extension constant (see e.g. [14, Theorem 5]), i.e.,

$$\begin{aligned} \mathrm{proj}(U) = \inf \big\{ c : \text{for any } V \subseteq W \text{ and any linear map } T : V \to U, \\ \text{there exists a linear map } \widetilde{T} : W \to U \text{ such that } T_{|V} = T \text{ and } \|\widetilde{T}\| \le c\|T\| \big\}. \end{aligned}$$

To retrieve the results stated in the text, one needs Kadec–Snobar estimate $\mathrm{proj}(U) \le \sqrt{\dim(U)}$ (see [14, Theorem 10]) and its corollary $\mathrm{proj}(U) \le 1 + \sqrt{\mathrm{codim}(U)}$ (see [14, Corollary 11]).

Proposition 13. Given a vector space F and a model set $\mathcal{K} \subseteq F$, given a linear observation map $\Lambda : F \to \mathbb{R}^M$, and given a linear quantity of interest $\Gamma : F \to Z$,

(i) If $\mathcal{K}$ is convex,

(i-a) then there exists an affine recovery map which is near optimal with factor $2C$, where $C := \min\{\mathrm{proj}(Z), 1 + \sqrt{M}\}$;

(i-b) and if $\Gamma : F \to \mathbb{R}$ is furthermore a linear functional, then there exists an affine recovery map which is genuinely optimal.

(ii) If $\mathcal{K}$ is convex and contains the origin, then affine can be replaced by linear.

(iii) If $\mathcal{K}$ is convex and symmetric about the origin, then the factor $2C$ can be replaced by C.

Proof. Central to the forthcoming arguments is the lower bound on the worst-case error of a recovery map $\Delta : \mathbb{R}^M \to Z$ by half of the so-called diameter of information, i.e.,

$$\mathrm{gwce}(\Delta) := \sup_{f \in \mathcal{K}} \|\Gamma(f) - \Delta(\Lambda f)\|_Z \ge \mathrm{lb} := \sup_{\substack{h \in \ker(\Lambda) \\ h \in (\mathcal{K}-\mathcal{K})/2}} \|\Gamma(h)\|_Z.$$

This can be easily justified in exactly the same way as in the beginning of the proof of Proposition 3. Since $\mathcal{K}$ is convex, the set $\mathcal{K}_{\mathrm{sym}} := (\mathcal{K} - \mathcal{K})/2$ is convex and symmetric about the origin, so its Minkowski functional is a seminorm, denoted by $|\cdot|$ here. The defining expression of the lower bound yields

$$\|\Gamma(h)\|_Z \le \mathrm{lb}\, |h| \qquad \text{for all } h \in \ker(\Lambda). \tag{15}$$

It is implicitly assumed that the lower bound is a finite quantity for any Γ, in particular for $\Gamma = \mathrm{Id}_F$, which implies that $\ker(\Lambda) \cap G' = \{0\}$ where $G' := \{g' \in F : |g'| = 0\}$ is a linear subspace of F. Now consider another linear subspace G'' containing $\ker(\Lambda)$ such that $G' \oplus G'' = F$ and notice that $|\cdot|$ induces a norm on G''. On the one hand, by the definition of the extension constant of Z, there exists a linear map $\widehat{\Gamma} : G'' \to Z$ such that $\widehat{\Gamma}_{|\ker(\Lambda)} = \Gamma_{|\ker(\Lambda)}$ and

$$\|\widehat{\Gamma}(g'')\|_Z \le \mathrm{proj}(Z)\, \mathrm{lb}\, |g''| \qquad \text{for all } g'' \in G''.$$

On the other hand, by selecting a projection P from G'' onto $\ker(\Lambda)$ with operator norm at most $1 + \sqrt{M}$ and by defining a linear map $\widehat{\Gamma} : G'' \to Z$ via $\widehat{\Gamma}(g'') = \Gamma(P(g''))$ for $g'' \in G''$, one obtains

$$\|\widehat{\Gamma}(g'')\|_Z \le \mathrm{lb}\, |P(g'')| \le (1 + \sqrt{M})\, \mathrm{lb}\, |g''| \qquad \text{for all } g'' \in G''.$$

Therefore, one always has $\|\widehat{\Gamma}(g'')\|_Z \le C\, \mathrm{lb}\, |g''|$ for all $g'' \in G''$ for some $\widehat{\Gamma} : G'' \to Z$ satisfying $\widehat{\Gamma}_{|\ker(\Lambda)} = \Gamma_{|\ker(\Lambda)}$. Next, one introduces the linear map $\widetilde{\Gamma} : F \to Z$ defined by $\widetilde{\Gamma}(g') = 0$ for $g' \in G'$

and $\widetilde{\Gamma}(g'') = \widehat{\Gamma}(g'')$ for $g'' \in G''$. Noticing that $\widetilde{\Gamma}(g) = \widehat{\Gamma}(g'')$ and $|g| = |g''|$ for any $g \in F$ written as $g = g' + g''$ with $g' \in G'$ and $g'' \in G''$, one arrives at

$$\|\widetilde{\Gamma}(g)\|_Z \le C \operatorname{lb} |g| \qquad \text{for all } g \in F.$$

It also holds that $(\Gamma - \widetilde{\Gamma})_{|\ker(\Lambda)} = 0$, which implies the existence of $c_1, \dots, c_M \in Z$ such that $\Gamma - \widetilde{\Gamma} = \sum_{m=1}^M c_m \lambda_m$. Thus, the previous estimate reads

$$\Big\| \Big(\Gamma - \sum_{m=1}^M c_m \lambda_m \Big)(g) \Big\| \le C \operatorname{lb} |g| \qquad \text{for all } g \in F. \tag{16}$$

Now, given $f \in \mathcal{K}$ and picking an arbitrary $\overline{f} \in \mathcal{K}$, applying the above to $g := (f - \overline{f})/2 \in \mathcal{K}_{\mathrm{sym}}$, which satisfies $|g| \le 1$, yields

$$\Big\| \Big(\Gamma - \sum_{m=1}^M c_m \lambda_m \Big) f - \Big(\Gamma - \sum_{m=1}^M c_m \lambda_m \Big) \overline{f} \Big\| \le 2C \operatorname{lb},$$

i.e., $\|\Gamma(f) - \Delta^{\mathrm{aff}}(\Lambda f)\|_Z \le 2C \operatorname{lb}$, where the affine recovery map Δ^{aff} is defined for $y \in \mathbb{R}^M$ by $\Delta^{\mathrm{aff}}(y) = \sum_{m=1}^M c_m y_m + d$ with $d := \Gamma(\overline{f}) - \sum_{m=1}^M c_m \lambda_m(\overline{f})$. Taking the supremum over $f \in \mathcal{K}$ reveals that $\operatorname{gwce}(\Delta^{\mathrm{aff}}) \le 2C \operatorname{lb} \le 2C \inf\{\operatorname{gwce}(\Delta), \Delta : \mathbb{R}^M \to Z\}$, i.e., that the recovery map Δ^{aff} is near optimal with factor $2C$. This establishes (i-a). For (ii), note that, if $0 \in \mathcal{K}$, then one can pick $\overline{f} = 0$ to turn Δ^{aff} into a linear map. For (iii), note that, if $\mathcal{K}$ is symmetric about the origin, then $\mathcal{K}_{\mathrm{sym}}$ coincides with $\mathcal{K}$ and one can apply (16) directly to $g = f \in \mathcal{K}$, so that $|g| \le 1$, and derive near optimality with factor C instead of $2C$.

It remains to justify (i-b) when Γ is a linear functional—this is the result of [13], proved in a different manner here. One comes back to (15) written for $Z = \mathbb{R}$, i.e.,

$$\Gamma(h) \le \operatorname{lb} |h| \qquad \text{for all } h \in \ker(\Lambda).$$

The dominated extension theorem guarantees the existence of an extension $\widetilde{\Gamma}$ to $\Gamma_{|\ker(\Lambda)}$ to the whole F such that

$$\widetilde{\Gamma}(g) \le \operatorname{lb} |g| \qquad \text{for all } g \in F.$$

Therefore, for all $f^+, f^- \in \mathcal{K}$, one has $(\widetilde{\Gamma}(f^+) - \widetilde{\Gamma}(f^-))/2 \le \operatorname{lb}$, i.e., $(S - I)/2 \le \operatorname{lb}$, where $S := \sup\{\widetilde{\Gamma}(f), f \in \mathcal{K}\}$ and $I := \inf\{\widetilde{\Gamma}(f), f \in \mathcal{K}\}$. From $\widetilde{\Gamma}(f) \in [I, S]$ for all $f \in \mathcal{K}$, one obtains $\widetilde{\Gamma}(f) - d \in [-(S-I)/2, (S-I)/2] \subseteq [-\operatorname{lb}, \operatorname{lb}]$ where $d := (I + S)/2$. Since the extension $\widetilde{\Gamma}$ can be written as $\widetilde{\Gamma} = \Gamma - \sum_{m=1}^M c_m \lambda_m$ for some $c_1, \dots, c_M \in \mathbb{R}$, one derives

$$\Big| \Gamma(f) - \Big(\sum_{m=1}^M c_m \lambda_m(f) + d \Big) \Big| \le \operatorname{lb}.$$

Taking the supremum over $f \in \mathcal{K}$ yields $\operatorname{gwce}(\Delta^{\mathrm{aff}}) \le \operatorname{lb} \le \inf\{\operatorname{gwce}(\Delta), \Delta : \mathbb{R}^M \to \mathbb{R}\}$, which shows that the affine recovery map $\Delta^{\mathrm{aff}} : y \in \mathbb{R}^M \mapsto \sum_{m=1}^M c_m y_m + d$ is genuinely optimal. □

References

[1] Creutzig, J. and Wojtaszczyk, P., 2004. *Linear vs. nonlinear algorithms for linear problems.* Journal of Complexity, 20(6), 807–820.

[2] DeVore, R., Foucart, S., Petrova, G. and Wojtaszczyk, P., 2019. *Computing a quantity of interest from observational data.* Constructive Approximation, 49(3), 461–508.

[3] Donoho, D. L., 1994. *Statistical estimation and optimal recovery.* The Annals of Statistics, 22(1), 238–270.

[4] Foucart, S., 2023. *Full recovery from point values: an optimal algorithm for Chebyshev approximability prior.* Advances in Computational Mathematics, 49(4), 57.

[5] Foucart, S. and Liao, C., 2024. *Radius of information for two intersected centered hyperellipsoids and implications in optimal recovery from inaccurate data.* Journal of Complexity, 83, 101841.

[6] Foucart, S. and Lasserre, J. B., 2018. *Determining projection constants of univariate polynomial spaces.* Journal of Approximation Theory, 235, 74–91.

[7] Foucart, S. and Paouris, G., 2023. *Near-optimal estimation of linear functionals with log-concave observation errors.* Information and Inference, 12(4), 2546–2561.

[8] Grünbaum, B., 1960. *Projection constants.* Transactions of the American Mathematical Society, 95(3), 451–465.

[9] Lasserre, J. B., 2019. *The Moment-SOS Hierarchy.* In: Proceedings of the International Congress of Mathematicians (ICM 2018), 3773–3794.

[10] Micchelli, C. A. and Rivlin, T. J., 1985. *Lectures on optimal recovery.* In: Numerical Analysis Lancaster 1984, 21–93. Springer.

[11] Minh, H., 2010. *Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory.* Constructive Approximation, 32, 307–338.

[12] Shalev-Shwartz, S. and Ben-David, S., 2014. Understanding Machine Learning: from Theory to Algorithms. Cambridge University Press.

[13] Sukharev, A. G., 1986. *On the existence of optimal affine methods for approximating linear functionals.* Journal of Complexity, 2(4), 317–322.

[14] Wojtaszczyk, P., 1996. Banach Spaces for Analysts. Cambridge University Press.